Federated Compositional Deep AUC Maximization

Xinwen Zhang ∗
Temple University
Philadelphia, PA, USA
ellenz@temple.edu

Yihan Zhang ∗
Temple University
Philadelphia, PA, USA
yihan.zhang0002@temple.edu

Tianbao Yang
Texas A&M University
College Station, TX, USA
ianbao-yang@tamu.edu

Richard Souvenir
Temple University
Philadelphia, PA, USA
souvenir@temple.edu

Hongchang Gao
Temple University
Philadelphia, PA, USA
hongchang.gao@temple.edu

Abstract

Federated learning has attracted increasing attention due to the promise of balancing privacy and large-scale learning; numerous approaches have been proposed. However, most existing approaches focus on problems with balanced data, and prediction performance is far from satisfactory for many real-world applications where the number of samples in different classes is highly imbalanced. To address this challenging problem, we developed a novel federated learning method for imbalanced data by directly optimizing the area under curve (AUC) score. In particular, we formulate the AUC maximization problem as a federated compositional minimax optimization problem, develop a local stochastic compositional gradient descent ascent with momentum algorithm, and provide bounds on the computational and communication complexities of our algorithm. To the best of our knowledge, this is the first work to achieve such favorable theoretical results. Finally, extensive experimental results confirm the efficacy of our method.

1 Introduction

Federated learning [19, 34] is a paradigm for training a machine learning model across multiple devices without sharing the raw data from each device. Practically, models are trained on each device, and, periodically, model parameters are exchanged between these devices. By not sharing the data itself, federated learning allows private information in the raw data to be preserved to some extent. This property has allowed federated learning to be proposed for numerous real-world computer vision and machine learning tasks.

Currently, one main drawback of existing federated learning methodologies is the assumption of balanced data, where the number of samples across classes is essentially the same. Most real-world data is imbalanced, even highly imbalanced. For example, in the healthcare domain, it is common to encounter problems where the amount of data from one class (e.g., patients with a rare disease) is significantly lower than the other class(es), leading to a distribution that is highly imbalanced. Traditional federated learning methods do not handle such imbalanced data scenarios very well. Specifically, training the classifier typically requires minimizing a classification-error induced loss function (e.g., cross-entropy). As a result, the resulting classifier may excel at classifying the majority, while failing to classify the minority.

To handle imbalanced data classification, the most common approach is to train the classifier by optimizing metrics designed for imbalanced data distributions. For instance, under the single-machine

*Equal contributions

setting, Ying et al. [35] proposed to train the classifier by maximizing the Area under the ROC curve (AUC) score. Since the AUC score can be affected by performance on both the majority and minority classes, the classifier is less prone to favoring one class above the rest. Later, [8, 40] extended this approach to federated learning. However, optimizing the AUC score introduces some new challenges, since maximizing the AUC score requires solving a minimax optimization problem, which is more challenging to optimize than conventional minimization problems. More specifically, when the classifier is a deep neural network, recent work [39] has demonstrated empirically that training a deep classifier from scratch with the AUC objective function cannot learn discriminative features; the resulting classifier sometimes fails to achieve satisfactory performance. To address this issue, [39] developed a compositional deep AUC maximization model under the single-machine setting, which combines the AUC loss function and the traditional cross-entropy loss function, leading to a stochastic compositional minimax optimization problem. This compositional deep AUC maximization model can learn discriminative features, achieving superior performance over traditional models consistently.

Considering its remarkable performance under the single-machine setting, a natural question is: **How can a compositional deep AUC maximization model be applied to federated learning?** The challenge is that the loss function of the compositional model involves two levels of distributed functions. Moreover, the stochastic compositional gradient is a biased estimation of the full gradient. Therefore, on the algorithmic design side, it is unclear what variables should be communicated when estimating the stochastic compositional gradient. On the theoretical analysis side, it is unclear if the convergence rate can achieve the linear speedup with respect to the number of devices in the presence of a biased stochastic compositional gradient, two levels of distributed functions, and the minimax structure of the loss function.

To address the aforementioned challenges, in this paper, we developed a novel local stochastic compositional gradient descent ascent with momentum (LocalSCGDAM) algorithm for federated compositional deep AUC maximization. In particular, we demonstrated which variables should be communicated to address the issue of two levels of distributed functions. Moreover, for this nonconvex-strongly-concave problem, we established the convergence rate of our algorithm, disclosing how the communication period and the number of devices affect the computation and communication complexities. Specifically, with theoretical guarantees, the communication period can be as large as \(O(T^{3/4}/K^{3/4})\) so that our algorithm can achieve \(O(1/\sqrt{KT})\) convergence rate and \(O(T^{3/4}/K^{3/4})\) communication complexity, where \(K\) is the number of devices and \(T\) is the number of iterations. To the best of our knowledge, this is the first work to achieve such favorable theoretical results for the federated compositional minimax problem. Finally, we conduct extensive experiments on multiple image classification benchmark datasets, and the experimental results confirm the efficacy of our algorithm.

In summary, we made the following important contributions in our work.

- We developed a novel federated optimization algorithm, which enables compositional deep AUC maximization for federated learning.
- We established the theoretical convergence rate of our algorithm, demonstrating how it is affected by the communication period and the number of devices.
- We conducted extensive experiments on multiple imbalanced benchmark datasets, confirming the efficacy of our algorithm.

## 2 Related Work

**Imbalanced Data Classification.** In the field of machine learning, there has been a fair amount of work addressing imbalanced data classification. Instead of using conventional cross-entropy loss functions, which are not suitable for imbalanced datasets, optimizing the AUC score has been proposed. For instance, Ying et al. [35] have proposed the minimax loss function to optimize the AUC score for learning linear classifiers. Liu et al. [18] extended this minimax method to deep neural networks and developed the nonconvex-strongly-concave loss function. Yuan et al. [39] have proposed a compositional training framework for end-to-end deep AUC maximization, which minimizes a compositional loss function, where the outer-level function is an AUC loss, and the inner-level function substitutes a gradient descent step for minimizing a traditional loss. Based on empirical results, this approach improved the classification performance by a large degree.

To address stochastic minimax optimization problems, there have been a number of diverse efforts launched in recent years. In particular, numerous stochastic gradient descent ascent (SGDA) algo-
rithms \cite{42,17,21,33} have been proposed. However, most of them focus on non-compositional optimization problems. On the other hand, to solve compositional optimization problems, existing work \cite{31,41,38,7} tends to only focus on the minimization problem. Only two recent works \cite{5,39} studied how to optimize the compositional minimax optimization problem, but they focused on the single-machine setting.

**Federated Learning.** In recent years, federated learning has shown promise with several empirical studies in the field of large-scale deep learning \cite{19,20,26}. The FedAvg \cite{19} algorithm has spawned a number of variants \cite{25,37,36} designed to address the minimization problem. For instance, by maintaining a local momentum, Yu et al. \cite{36} have provided rigorous theoretical studies for the convergence of the local stochastic gradient descent with momentum (LocalSGDM) algorithm. These algorithms are often applied to balanced datasets, and their performance in the imbalanced imbalanced regime is lacking.

To address minimax optimization for federated learning, Deng et al. \cite{2} proposed local stochastic gradient descent ascent (LocalSGDA) algorithms to provably optimize federated minimax problems. However, their theoretical convergence rate was suboptimal and later improved by \cite{27}. However, neither method could achieve a linear speedup with respect to the number of devices. Recently, Sharma et al. \cite{23} developed the local stochastic gradient descent ascent with momentum (LocalSGDAM) algorithm, whose convergence rate is able to achieve a linear speedup for nonconvex-strongly-concave optimization problems. Guo et al. \cite{8} proposed and analyzed a communication-efficient distributed optimization algorithm (CoDA) for the minimax AUC loss function under the assumption of PL-condition, which can also achieve a linear speedup, in theory. Yuan et al. \cite{40} extended CoDA to heterogeneous data distributions and established its convergence rate. Shen et al. \cite{24} proposed to handle the imbalance issue via introducing a constrained optimization problem and then formulated it as an unconstrained minimax problem. While these algorithms are designed for federated minimax problems, none can deal with the federated compositional minimax problems.

To handle the compositional optimization problem under the distributed setting, Gao et al. \cite{3} developed the first parallel stochastic compositional gradient descent algorithm and established its convergence rate for nonconvex problems, inspiring many federated learning methods \cite{12,28,4,27,7} in the past few years. For instance, \cite{12} directly used a biased stochastic gradient to do local updates, suffering from large sample and communication complexities. \cite{4} employed the stochastic compositional gradient and the momentum technique, which can achieve much better sample and communication complexities than \cite{12}. \cite{7} considered the setting where the inner-level function is distributed on different devices, which shares similar sample and communication complexities as \cite{4}. However, all these works restrict their focus on the compositional minimization problem.

It is worth noting that our method is significantly different from the heterogeneous federated learning approaches \cite{14,16,29}. Specifically, most existing heterogeneous federated learning approaches consider a setting where the local distribution is imbalanced but the global distribution is balanced. For example, Scaffold \cite{14} method uses the global gradient to correct the local gradient because it assumes the global gradient is computed on a balanced distribution. On the contrary, our work considers a setting where both the local and global distributions are imbalanced, which is much more challenging than existing heterogeneous federating learning methods.

### 3 Preliminaries

In this section, we first introduce the compositional deep AUC maximization model under the single-machine setting and then provide the problem setup in federated learning.

#### 3.1 Compositional Deep AUC Maximization

Training classifiers by optimizing AUC \cite{9,11} is an effective way to handle highly imbalanced datasets. However, traditional AUC maximization models typically depend on pairwise sample input, limiting the application to large-scale data. Recently, Ying et al. \cite{35} formulated AUC maximization model as a minimax optimization problem, defined as follows:

$$
\min_{w, \bar{w}_1, \bar{w}_2, \bar{w}_3} \max_{a, b} \mathcal{L}_{AUC}(w, \bar{w}_1, \bar{w}_2, \bar{w}_3; a, b) \triangleq (1 - p)(h(w; a) - \bar{w}_1)^2 I_{\{b=1\}} - p(1 - p)\bar{w}_3^2 + p(h(w; a) - \bar{w}_2)^2 I_{\{b=-1\}} + 2(1 + \bar{w}_3)(ph(w; a)I_{\{b=-1\}} - (1 - p)h(w; a)I_{\{b=1\}})
$$

(1)
where \( h \) denotes the classifier parameterized by \( w \in \mathbb{R}^d, \tilde{w}_1 \in \mathbb{R}, \tilde{w}_2 \in \mathbb{R}, \tilde{w}_3 \in \mathbb{R} \) are the parameters for measuring AUC score, \((a, b)\) substitutes the sample’s feature and label, \( p \) is the prior probability of the positive class, and \( \mathbb{I} \) is an indicator function that takes value 1 if the argument is true and 0 otherwise. Such a minimax objective function decouples the dependence of pairwise samples so that it can be applied to large-scale data.

Since training a deep classifier from scratch with \( L_{AUC} \) loss function did not yield satisfactory performance, Yuan et al. [39] developed the compositional deep AUC maximization model, which is defined as follows:

\[
\min_{\mathbf{w}, \tilde{w}_1, \tilde{w}_2, \tilde{w}_3} \max_{a, b} L_{AUC}(\mathbf{w}, \tilde{w}_1, \tilde{w}_2, \tilde{w}_3; a, b) \quad \text{s.t.} \quad \tilde{w} = \mathbf{w} - \rho \nabla_{\mathbf{w}} L_{CE}(\mathbf{w}; a, b).
\]  

(2)

Here, \( L_{CE} \) denotes the cross-entropy loss function, \( \mathbf{w} - \rho \nabla_{\mathbf{w}} L_{CE} \) indicates using the gradient descent method to minimize the cross-entropy loss function, where \( \rho > 0 \) is the learning rate. Then, for the obtained model parameter \( \tilde{w} \), one can optimize it through optimizing the AUC loss function.

By denoting \( g(x) = x - \rho \Delta(x) \) and \( y = \tilde{w}_3 \), where \( x = [\mathbf{w}^T, \tilde{w}_1, \tilde{w}_2]^T \), \( \Delta(x) = [\nabla_{\mathbf{w}} L_{CE}(\mathbf{w}; a, b)^T, 0, 0]^T \), and \( f = L_{AUC} \). Eq. (2) can be represented as a generic compositional minimax optimization problem as follows:

\[
\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^d} f(g(x), y),
\]

(3)

where \( g \) is the inner-level function and \( f \) is the outer-level function. It is worth noting that when \( f \) is a nonlinear function, the stochastic gradient regarding \( x \) is a biased estimation of the full gradient. As such, the stochastic compositional gradient [31] is typically used to optimize this kind of problem. We will demonstrate how to adapt this compositional minimax optimization problem to federated learning and address the unique challenges.

### 3.2 Problem Setup

In this paper, to optimize the deep compositional AUC maximization problem under the cross-silo federated learning setting, we will concentrate on developing an efficient optimization algorithm to solve the following generic federated stochastic compositional minimax optimization problem:

\[
\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^d} \frac{1}{K} \sum_{k=1}^{K} f^{(k)} \left( \frac{1}{K} \sum_{k' = 1}^{K} g^{(k')} (x), y \right),
\]

(4)

where \( K \) is the number of devices, \( g^{(k)}(\cdot) = \mathbb{E}_{\xi \sim D^{(k)}} [g^{(k)}(\cdot; \xi)] \in \mathbb{R}^{d_y} \) denotes the inner-level function for the data distribution \( D^{(k)} \) of the \( k \)-th device, \( f^{(k)}(\cdot, \cdot) = \mathbb{E}_{\xi \sim D^{(k)}} [f^{(k)}(\cdot, \cdot; \xi)] \) represents the outer-level function for the data distribution \( D^{(k)} \) of the \( k \)-th device. It is worth noting that both the inner-level function and the outer-level function are distributed on different devices, which is significantly different from traditional federated learning models. Therefore, we need to design a new federated optimization algorithm to address this unique challenge.

Here, we introduce the commonly-used assumptions from existing work [6][41][39][5] for investigating the convergence rate of our algorithm.

**Assumption 1.** The gradient of the outer-level function \( f^{(k)}(\cdot, \cdot) \) is \( L_f \)-Lipschitz continuous where \( L_f > 0 \), i.e.,

\[
\| \nabla_x f^{(k)}(z_1, y_1) - \nabla_x f^{(k)}(z_2, y_2) \|^2 \leq L_f^2 \| (z_1, y_1) - (z_2, y_2) \|^2,
\]

\[
\| \nabla_y f^{(k)}(z_1, y_1) - \nabla_y f^{(k)}(z_2, y_2) \|^2 \leq L_f^2 \| (z_1, y_1) - (z_2, y_2) \|^2,
\]

hold for \( \forall (z_1, y_1), (z_2, y_2) \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \). The gradient of the inner-level function \( g^{(k)}(\cdot) \) is \( L_g \)-Lipschitz continuous where \( L_g > 0 \), i.e.,

\[
\| \nabla g^{(k)}(x_1) - \nabla g^{(k)}(x_2) \|^2 \leq L_g^2 \| x_1 - x_2 \|^2,
\]

holds for \( \forall x_1, x_2 \in \mathbb{R}^{d_x} \).

**Assumption 2.** The second moment of \( \nabla_x f^{(k)}(z, y; \xi) \) and \( \nabla_y g^{(k)}(x; \xi) \) satisfies:

\[
\mathbb{E}\| \nabla_x f^{(k)}(z, y; \xi) \|^2 \leq C_f^2, \quad \mathbb{E}\| \nabla_y g^{(k)}(x; \xi) \|^2 \leq C_g^2,
\]

(7)

for \( \forall (z, y) \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \) and \( \forall x \in \mathbb{R}^{d_x} \), where \( C_f > 0 \) and \( C_g > 0 \). Meanwhile, the second moment of the full gradient is assumed to have the same upper bound.
Assumption 3. The variance of the stochastic gradient of the outer-level function $f^{(k)}(\cdot, \cdot)$ satisfies:
\[
\mathbb{E}[\|\nabla_y f^{(k)}(z, y; \zeta) - \nabla_y f^{(k)}(z, y)\|^2] \leq \sigma_y^2, \quad \mathbb{E}[\|\nabla_y f^{(k)}(z, y; \zeta) - \nabla_y f^{(k)}(z, y)\|^2] \leq \sigma_y^2,
\]  
for $\forall (z, y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_2}$, where $\sigma_y > 0$. Additionally, the variance of the stochastic gradient and the stochastic function value of $g^{(k)}(\cdot)$ satisfies:
\[
\mathbb{E}[\|\nabla g^{(k)}(x; \xi) - \nabla g^{(k)}(x)\|^2] \leq \sigma_g^2, \quad \mathbb{E}[\|g^{(k)}(x; \xi) - g^{(k)}(x)\|^2] \leq \sigma_g^2,
\]  
for $\forall x \in \mathbb{R}^{d_1}$, where $\sigma_g > 0$ and $\sigma_g > 0$.

Assumption 4. The outer-level function $f^{(k)}(z, y)$ is $\mu$-strongly-concave with respect to $y$ for any fixed $z \in \mathbb{R}^{d_3}$, where $\mu > 0$, i.e.,
\[
f^{(k)}(z, y_1) \leq f^{(k)}(z, y_2) + \langle \nabla_y f^{(k)}(z, y_2), y_1 - y_2 \rangle - \frac{\mu}{2} \|y_1 - y_2\|^2.
\]

Notation: Throughout this paper, $a_i^{(k)}$ denotes the variable of the $k$-th device in the $t$-th iteration and $\bar{a}_t$ denotes the averaged variable across all devices, where $a$ denotes any variables used in this paper. $x_*$ denotes the optimal solution.

4 Methodology

In this section, we present the details of our algorithm for the federated compositional deep AUC maximization problem defined in Eq. (4).

Algorithm 1 LocalSCGDAM

Input: $x_0, y_0, \eta \in (0, 1), \gamma_x > 0, \gamma_y > 0, \beta_x > 0, \beta_y > 0, \alpha > 0, \alpha \eta \in (0, 1), \beta_y \eta \in (0, 1)$.

All workers conduct the steps below to update $x, y$: $x_0 = x_0, y_0 = y_0$.

1. for $t = 0, \ldots, T - 1$ do  
2. Update $x$ and $y$:  
3. Estimate the inner-level function: $h_t^{(k)} = (1 - \alpha \eta)h_t^{(k)} + \alpha \eta g^{(k)}(x_t^{(k)}; \xi_t^{(k)})$,  
4. Update momentum: $u_t^{(k)} = (1 - \beta \eta)u_t^{(k)} + \beta \eta \nabla g^{(k)}(x_t^{(k)}; \xi_t^{(k)}) = v_t^{(k)}(h_t^{(k)}; y_t^{(k)}; \zeta_t^{(k)})$,  
5. if mod$(t + 1, p) = 0$ then
6. $h_{t+1}^{(k)} = h_{t+1}^{(k)} + \frac{1}{K} \sum_{k'=1}^{K} v_{t+1}^{(k')}$, $u_{t+1}^{(k)} = u_{t+1}^{(k)} + \frac{1}{K} \sum_{k'=1}^{K} u_{t+1}^{(k')}$, $v_{t+1}^{(k)} = v_{t+1}^{(k)} + \frac{1}{K} \sum_{k'=1}^{K} v_{t+1}^{(k')}$, $x_{t+1}^{(k)} = \bar{x}_{t+1}^{(k)} + \frac{1}{K} \sum_{k'=1}^{K} x_{t+1}^{(k')}$, $y_{t+1}^{(k)} = \bar{y}_{t+1}^{(k)} + \frac{1}{K} \sum_{k'=1}^{K} y_{t+1}^{(k')}$,  
7. end if  
8. end for

To optimize Eq. (4), we developed a novel local stochastic compositional gradient descent ascent with momentum algorithm, shown in Algorithm 1. Generally speaking, in the $t$-th iteration, we employ the local stochastic (compositional) gradient with momentum to update the local model parameters $x_t^{(k)}$ and $y_t^{(k)}$ on the $k$-th device. There exists a unique challenge when computing the local stochastic compositional gradient compared to traditional federated learning models. Specifically, as shown in Eq. (4), the objective function depends on the global inner-level function. However, it is not feasible to communicate the inner-level function in every iteration. To address this challenge, we propose to employ the local inner-level function to compute the stochastic compositional gradient at each iteration and then communicate the estimation of this function periodically to obtain the global inner-level function.
As for the model parameters and momentum, we employ the same strategy as traditional federated.
In this way, each device is able to obtain the estimate of the global inner-level function periodically.

Then, based on Assumptions 1-4, we can obtain that
\[ \eta \]

Based on the obtained stochastic (compositional) gradient, we compute the momentum as follows:
\[ \bar{h}_{t+1} = \bar{h}_{t+1} = \frac{1}{K} \sum_{k'=1}^{K} h_{t+1}^{(k')} \] (14)

In this way, each device is able to obtain the estimate of the global inner-level function periodically. As for the model parameters and momentum, we employ the same strategy as traditional federated learning methods to communicate them periodically with the central server, which is shown in Step 6 in Algorithm 1.

In summary, we developed a novel local stochastic compositional gradient descent ascent with momentum algorithm for the compositional minimax problem, which shows how to deal with two distributed functions in federated learning. With our algorithm, we can enable federated learning for the compositional deep AUC maximization model, benefiting imbalanced data classification tasks.

5 Theoretical Analysis

In this section, we provide the convergence rate of our algorithm to show how it is affected by the number of devices and communication period.

To investigate the convergence rate of our algorithm, we introduce the following auxiliary functions:
\[ y_*(x) = \arg \max_{y \in \mathbb{R}^d} \frac{1}{K} \sum_{k=1}^{K} f(k) \left( \frac{1}{K} \sum_{k'=1}^{K} g(k') (x, y) \right) \text{,} \quad \Phi(x) = \frac{1}{K} \sum_{k=1}^{K} f(k) \left( \frac{1}{K} \sum_{k'=1}^{K} g(k') (x, y_*(x)) \right) \text{.} \] (15)

Then, based on Assumptions 1-4, we can obtain that \( \Phi(k) \) is \( L_{\Phi} \)-smooth, where \( L_{\Phi} = \frac{2C^2_\Phi L_1^2}{\mu} + C_f L_2 \).
The proof can be found in Lemma 2 of Appendix A. In terms of these auxiliary functions, we establish the convergence rate of our algorithm.

**Theorem 1.** Given Assumption 7-4 by setting \( \alpha > 0 \), \( \beta_x > 0 \), \( \beta_y > 0 \), \( \eta \), \( \gamma_x \leq \min \left\{ \frac{1}{100C^2_\gamma L_1^2 \sqrt{1+6L_1^2}}, \frac{1}{32C^2_\gamma L_1^2 + C^2_f L_2^2}, \frac{1}{144C^2_\gamma L_1^2}, \frac{24C_\gamma \sqrt{100C^2_\gamma L_1^2 + 2C_f^2 L_2^2}}{20C^2_\gamma L_1^2} \right\} \), and \( \gamma_y \leq \min \left\{ \frac{1}{6L_1}, \frac{3\beta^2_\gamma}{1000C_\gamma^2 L_1^2}, \frac{3\eta^2}{160} \right\} \),
Algorithm 1 has the following convergence rate:

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla \Phi(x_t)\|^2] \leq \frac{2(\Phi(x_0) - \Phi(x^*))}{\gamma_2 \eta T} + \frac{24C^2_x L^2}{\gamma_2 \eta \mu T} \|y_0 - y^*(x_0)\|^2 + O\left(\frac{\eta}{K}\right) + O\left(\frac{1}{\eta T}\right) \\
+ O(p^2 \eta^2) + O(p^4 \eta^4) + O(p^6 \eta^6) + O(p^8 \eta^8) + O(p^{10} \eta^{10})
\]

(16)

Remark 1. In terms of Theorem 7, for sufficiently large \( T \), by setting the learning rate \( \eta = O(K^{1/2}/T^{1/2}) \), \( p = O(T^{1/4}/K^{3/4}) \), Algorithm 1 can achieve \( O(1/\sqrt{T}) \) convergence rate, which indicates a linear speedup with respect to the number of devices \( K \). In addition, it is straightforward to show that the communication complexity of our algorithm is \( T/p = O(K^{3/4}T^{3/4}) \). Moreover, to achieve the \( \epsilon \)-accuracy solution, i.e., \( \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla \Phi(x_t)\|^2] \leq \epsilon^2 \), by setting \( \eta = O(K^\epsilon^2) \) and \( p = O(1/(K \epsilon)) \), then the sample complexity on each device is \( O(1/(K \epsilon^3)) \) and the communication complexity is \( O(1/\epsilon^3) \).

Challenges. The compositional structure in the loss function, especially when the inner-level functions are distributed on different devices, makes the convergence analysis challenging. In fact, the existing federated compositional minimization algorithms [4, 27, 12, 28] only consider a much simpler case where the inner-level functions are not distributed across devices. Therefore, our setting is much more challenging than existing works. On the other hand, all existing federated compositional minimization algorithms [4, 27, 12, 28] fail to achieve linear speedup with respect to the number of devices. Thus, it is still unclear whether the linear speedup is achievable for federated compositional optimization algorithm. In this paper, we successfully addressed these challenges with novel theoretical analysis strategies and achieved the linear speedup for the first time for federated compositional minimax optimization algorithms. We believe our approaches, e.g., that for bounding consensus errors, can be applied to the minimization algorithms to achieve linear speedup.

6 Experiments

In this section, we present the experimental results to demonstrate the performance of our algorithm.
Figure 2: Testing performance with AUC score versus the number of iterations when the communication period $p = 8$.

Figure 3: Testing performance with AUC score versus the number of iterations when the communication period $p = 16$.

6.1 Experimental Setup

Datasets. In our experiments, we employ six image classification datasets, including CIFAR10 [15], CIFAR100 [15], STL10 [1], FashionMNIST [32], CATvsDOG and Melanoma [22]. For the first four datasets, following [39], we consider the first half of classes to be the positive class, and the second half as the negative class. Then, in order to construct highly imbalanced data, we randomly drop some samples of the positive class in the training set. Specifically, the ratio between positive samples and all samples is set to 0.1. For the two-class dataset, CATvsDOG, we employ the same
Table 1: The comparison between the test AUC score of different methods on all datasets. Here, \( p \) denotes the communication period.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Methods</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( p = 4 )</td>
</tr>
<tr>
<td>CATvsDOG</td>
<td>LocalSCGDAM</td>
<td>0.933±0.000</td>
</tr>
<tr>
<td></td>
<td>CoDA</td>
<td>0.895±0.000</td>
</tr>
<tr>
<td></td>
<td>LocalSGDAM</td>
<td>0.899±0.000</td>
</tr>
<tr>
<td></td>
<td>LocalSGDM</td>
<td>0.888±0.001</td>
</tr>
<tr>
<td></td>
<td>LocalSGDM_RL</td>
<td>0.909±0.000</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>LocalSCGDAM</td>
<td>0.914±0.000</td>
</tr>
<tr>
<td></td>
<td>CoDA</td>
<td>0.890±0.000</td>
</tr>
<tr>
<td></td>
<td>LocalSGDAM</td>
<td>0.893±0.000</td>
</tr>
<tr>
<td></td>
<td>LocalSGDM</td>
<td>0.883±0.001</td>
</tr>
<tr>
<td></td>
<td>LocalSGDM_RL</td>
<td>0.890±0.000</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>LocalSCGDAM</td>
<td>0.702±0.000</td>
</tr>
<tr>
<td></td>
<td>CoDA</td>
<td>0.694±0.001</td>
</tr>
<tr>
<td></td>
<td>LocalSGDAM</td>
<td>0.692±0.000</td>
</tr>
<tr>
<td></td>
<td>LocalSGDM</td>
<td>0.675±0.001</td>
</tr>
<tr>
<td></td>
<td>LocalSGDM_RL</td>
<td>0.676±0.000</td>
</tr>
<tr>
<td>STL10</td>
<td>LocalSCGDAM</td>
<td>0.820±0.001</td>
</tr>
<tr>
<td></td>
<td>CoDA</td>
<td>0.801±0.000</td>
</tr>
<tr>
<td></td>
<td>LocalSGDAM</td>
<td>0.792±0.000</td>
</tr>
<tr>
<td></td>
<td>LocalSGDM</td>
<td>0.760±0.001</td>
</tr>
<tr>
<td></td>
<td>LocalSGDM_RL</td>
<td>0.773±0.001</td>
</tr>
<tr>
<td>FashionMNIST</td>
<td>LocalSCGDAM</td>
<td>0.980±0.000</td>
</tr>
<tr>
<td></td>
<td>CoDA</td>
<td>0.976±0.000</td>
</tr>
<tr>
<td></td>
<td>LocalSGDAM</td>
<td>0.977±0.000</td>
</tr>
<tr>
<td></td>
<td>LocalSGDM</td>
<td>0.963±0.000</td>
</tr>
<tr>
<td></td>
<td>LocalSGDM_RL</td>
<td>0.980±0.000</td>
</tr>
<tr>
<td>Melanoma</td>
<td>LocalSCGDAM</td>
<td>0.876±0.000</td>
</tr>
<tr>
<td></td>
<td>CoDA</td>
<td>0.734±0.000</td>
</tr>
<tr>
<td></td>
<td>LocalSGDAM</td>
<td>0.730±0.000</td>
</tr>
<tr>
<td></td>
<td>LocalSGDM</td>
<td>0.774±0.001</td>
</tr>
<tr>
<td></td>
<td>LocalSGDM_RL</td>
<td>0.737±0.001</td>
</tr>
</tbody>
</table>

strategy to construct the imbalanced training data. For these synthetic imbalanced datasets, the testing set is balanced. Melanoma is an intrinsically imbalanced medical image classification dataset, which we do not modify. The details about these benchmark datasets are summarized in Table 4.

**Experimental Settings.** For Melanoma, we use DenseNet121 [13] where the dimensionality of the last layer is set to 1 for binary classification. The details for the classifier for FashionMNIST can be found in Appendix B. For the other datasets, we use ResNet20 [10], where the last layer is also set to 1.

To demonstrate the performance of our algorithm, we compare it with three state-of-the-art methods: LocalSGD [36], LocalSGD_RL [30], CoDA [8], LocalSGDAM [23]. Specifically, LocalSGDAM uses momentum SGD to optimize the standard cross-entropy loss function. LocalSGDAM_RL employs momentum SGD to optimize a Ratio Loss function, which is to add a regularization term to the standard cross-entropy loss function to address the imbalance distribution issue. CoDA leverages SGDA to optimize AUC loss, while LocalSGDAM exploits momentum SGDA to optimize AUC loss. For a fair comparison, we use similar learning rates for all algorithms. The details can be found in Appendix B. We use 4 devices (i.e., GPUs) in our experiment. The batch size on each device is set to 8 for STL10, 16 for Melanoma, and 32 for the others.

6.2 Experimental Results

In Table 1 we report the AUC score of the test set for all methods, where we show the average and variance computed across all devices. Here, the communication period is set to 4, 8, and 16, respectively. It can be observed that our LocalSCGDAM algorithm outperforms all competing methods.
for all cases. For instance, our LocalSCGDAM can beat baseline methods on CATvsDOG dataset with a large margin for all communication periods. These observations confirm the effectiveness of our algorithm. In addition, we plot the average AUC score of the test set versus the number of iterations in Figures 1, 2, 3. It can also be observed that our algorithm outperforms baseline methods consistently, which further confirms the efficacy of our algorithm.

To further demonstrate the performance of our algorithm, we apply these algorithms to the dataset with different imbalance ratios. Using the CATvsDOG dataset, we set the imbalance ratio to 0.01, 0.05, and 0.2 to construct three imbalanced training sets. The averaged testing AUC score of these three datasets versus the number of iterations is shown in Figure 4. It can be observed that our algorithm outperforms competing methods consistently and is robust to large imbalances in the training data. Especially when the training set is highly imbalanced, e.g., the imbalance ratio is 0.01, all AUC based methods outperform the cross-entropy loss based method significantly, and our LocalSCGDAM beats other AUC based methods with a large margin.

Figure 4: The test AUC score versus the number of iterations when using different imbalance ratios for CATvsDOG.

Finally, we compare our algorithm with two additional baseline methods: SCAFFOLD [14] and FedProx [16]. These two methods assume the local data distribution is imbalanced but the global one is balanced. Then, they use the global gradient to correct the local one. However, this kind of methods do not work when the global data distribution is imbalanced. Specifically, when the global gradient itself is computed on the imbalanced data, rather than the balanced one, it cannot alleviate the imbalance issue in the local gradient. In Figure 5, we show the test AUC score of the STL10 dataset, where we use the same experimental setting as that of Figure 1, i.e., both the local and global data distributions are imbalanced. It can be observed that our algorithm outperforms those two baselines with a large margin, which confirms the effectiveness of our algorithm in handling the global imbalanced data distribution.

Figure 5: The test AUROC score for STL10.

7 Conclusion

In this paper, we developed a novel local stochastic compositional gradient descent ascent algorithm to solve the federated compositional deep AUC maximization problem. On the theoretical side, we established the convergence rate of our algorithm, which enjoys a linear speedup with respect to the number devices. On the empirical side, extensive experimental results on multiple imbalanced image classification tasks confirm the effectiveness of our algorithm.

Acknowledgements

We thank anonymous reviewers for constructive comments. T. Yang was partially supported by NSF Career Award 2246753, NSF Grant 2246757.
References


[38] H. Yuan, X. Lian, and J. Liu. Stochastic recursive variance reduction for efficient smooth


[40] Z. Yuan, Z. Guo, Y. Xu, Y. Ying, and T. Yang. Federated deep auc maximization for heterogeneous
data with a constant communication complexity. In International Conference on Machine


for nonconvex-concave min-max problems. Advances in Neural Information Processing Systems,
A Proof

To investigate the convergence rate of our algorithm, we first introduce the following auxiliary functions:

\[ y_*^{(k)}(x) = \arg \max_{y \in \mathbb{R}^d} f^{(k)}\left( \frac{1}{K} \sum_{k' = 1}^{K} g^{(k')}(x), y \right), \]

\[ \Phi^{(k)}(x) = f^{(k)}\left( \frac{1}{K} \sum_{k' = 1}^{K} g^{(k')}(x), y_*^{(k)}(x) \right), \]

\[ g(x) = \frac{1}{K} \sum_{k = 1}^{K} g^{(k)}(x). \]

Based on these auxiliary functions, we provide the following lemmas to complete the proof.

Lemma 1. Given Assumptions 1-4, the function \( y_*^{(k)}(x) \) is \( L_{y_*} \)-Lipschitz continuous, where \( L_{y_*} = \frac{C_g L_f}{\mu} \).

Proof. Since \( y_*^{(k)}(x) = \arg \max_{y \in \mathbb{R}^d} f^{(k)}\left( \frac{1}{K} \sum_{k' = 1}^{K} g^{(k')}\left( x \right), y \right) \), according to the optimality condition, for any \( x_1, x_2 \in \mathbb{R}^d \), we can get

\[ \langle \nabla_y f^{(k)} \left( \frac{1}{K} \sum_{k' = 1}^{K} g^{(k')}\left( x_1 \right), y_*^{(k)}\left( x_1 \right) \right), y_*^{(k)}\left( x_2 \right) - y_*^{(k)}\left( x_1 \right) \rangle \leq 0, \]

\[ \langle \nabla_y f^{(k)} \left( \frac{1}{K} \sum_{k' = 1}^{K} g^{(k')}\left( x_2 \right), y_*^{(k)}\left( x_2 \right) \right), y_*^{(k)}\left( x_1 \right) - y_*^{(k)}\left( x_2 \right) \rangle \leq 0. \]

As a result, we can get

\[ \langle y_*^{(k)}\left( x_2 \right) - y_*^{(k)}\left( x_1 \right), \nabla_y f^{(k)} \left( \frac{1}{K} \sum_{k' = 1}^{K} g^{(k')}\left( x_1 \right), y_*^{(k)}\left( x_1 \right) \right) - \nabla_y f^{(k)} \left( \frac{1}{K} \sum_{k' = 1}^{K} g^{(k')}\left( x_2 \right), y_*^{(k)}\left( x_2 \right) \right) \rangle \leq 0. \]

Meanwhile, due to the strong monotonicity of the gradient with respect to \( y \), we can get

\[ \langle y_*^{(k)}\left( x_2 \right) - y_*^{(k)}\left( x_1 \right), \nabla_y f^{(k)} \left( \frac{1}{K} \sum_{k' = 1}^{K} g^{(k')}\left( x_1 \right), y_*^{(k)}\left( x_1 \right) \right) - \nabla_y f^{(k)} \left( \frac{1}{K} \sum_{k' = 1}^{K} g^{(k')}\left( x_2 \right), y_*^{(k)}\left( x_2 \right) \right) \rangle + \mu \| y_*^{(k)}\left( x_2 \right) - y_*^{(k)}\left( x_1 \right) \|^2 \leq 0. \]

By adding above two inequalities together, we can get

\[ \mu \| y_*^{(k)}\left( x_2 \right) - y_*^{(k)}\left( x_1 \right) \|^2 \]

\[ \leq \langle y_*^{(k)}\left( x_2 \right) - y_*^{(k)}\left( x_1 \right), \nabla_y f^{(k)} \left( \frac{1}{K} \sum_{k' = 1}^{K} g^{(k')}\left( x_2 \right), y_*^{(k)}\left( x_2 \right) \right) - \nabla_y f^{(k)} \left( \frac{1}{K} \sum_{k' = 1}^{K} g^{(k')}\left( x_1 \right), y_*^{(k)}\left( x_1 \right) \right) \rangle \]

\[ \leq \| y_*^{(k)}\left( x_2 \right) - y_*^{(k)}\left( x_1 \right) \| \| \nabla_y f^{(k)} \left( \frac{1}{K} \sum_{k' = 1}^{K} g^{(k')}\left( x_2 \right), y_*^{(k)}\left( x_2 \right) \right) - \nabla_y f^{(k)} \left( \frac{1}{K} \sum_{k' = 1}^{K} g^{(k')}\left( x_1 \right), y_*^{(k)}\left( x_1 \right) \right) \| \]

\[ \leq L_f \| y_*^{(k)}\left( x_2 \right) - y_*^{(k)}\left( x_1 \right) \| \| \frac{1}{K} \sum_{k' = 1}^{K} g^{(k')}\left( x_2 \right) - \frac{1}{K} \sum_{k' = 1}^{K} g^{(k')}\left( x_1 \right) \| \]

\[ \leq C_g L_f \| y_*^{(k)}\left( x_2 \right) - y_*^{(k)}\left( x_1 \right) \| \| x_2 - x_1 \| . \]

Therefore, we can get

\[ \| y_*^{(k)}\left( x_2 \right) - y_*^{(k)}\left( x_1 \right) \| \leq \frac{C_g L_f}{\mu} \| x_2 - x_1 \|. \]
Lemma 2. Given Assumptions 1-4, \( \Phi^{(k)} \) is \( L_\Phi \)-smooth, where \( L_\Phi = \frac{2C^2\mathcal{L}^2}{\mu} + C_f L_g \).

Proof. Since \( \Phi^{(k)}(x) = f^{(k)}(\frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x), y_*(x)) \), we can get
\[
\begin{align*}
\| \nabla \Phi^{(k)}(x_1) - \nabla \Phi^{(k)}(x_2) \| &= \left\| \left( \frac{1}{K} \sum_{k'=1}^{K} \nabla g^{(k')}(x_1) \right) - \nabla \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_1) \right\| \\
&= \left\| \left( \frac{1}{K} \sum_{k'=1}^{K} \nabla g^{(k')}(x_1) \right) - \nabla \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_2) \right\| \\
&\leq \frac{\mathcal{L}}{\mu} \left\| x_1 - x_2 \right\|.
\end{align*}
\]

where the second inequality holds due to Assumption 2, the third inequality holds due to Assumption 1, and Lemma 1, the last step holds due to \( L_f / \mu \geq 1 \).

Lemma 3. Given Assumptions 1-4 and \( \eta \leq \frac{1}{2\gamma_x L_g} \), we can get
\[
\begin{align*}
\mathbb{E}[\Phi(\tilde{x}_{t+1})] &\leq \mathbb{E}[\Phi(\tilde{x}_t)] - \frac{\gamma_x \eta}{2} \mathbb{E}[\| \nabla \Phi(\tilde{x}_t) \|^2] - \frac{\gamma_x \eta}{4} \mathbb{E}[\| \tilde{u}_t \|^2] + 3\gamma_x \eta C_g^2 \mathcal{L}^2 \mathbb{E}[\| y_*(\tilde{x}_t) - \tilde{y}_t \|^2] \\
&+ 6\gamma_x \eta (C_g^2 \mathcal{L}^2 + C_f^2 \mathcal{L}^2) \frac{1}{K} \sum_{k=1}^{K} \left[ \| \tilde{x}_t - x^{(k)}_t \|^2 \right] + 3\gamma_x \eta C_g^2 \mathcal{L}^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\| \tilde{y}_t - y^{(k)}_t \|^2] \\
&+ 3\gamma_x \eta \mathbb{E}[\| \frac{1}{K} \sum_{k=1}^{K} \nabla_x f^{(k)}(g^{(k)}(x^{(k)}_t), y^{(k)}_t) - \tilde{u}_t \|^2].
\end{align*}
\]

(24)
Proof. Because $\Phi(x_t)$ is $L_\theta$-smooth, we can get

$$
\mathbb{E}[\Phi(x_{t+1})] \leq \mathbb{E}[\Phi(x_t)] + \mathbb{E}[(\nabla \Phi(x_t), x_{t+1} - x_t)] + \frac{L_\theta}{2} \mathbb{E}[\|x_{t+1} - x_t\|^2]
$$

$$
= \mathbb{E}[\Phi(x_t)] - \gamma_x \eta \mathbb{E}[(\nabla \Phi(x_t), \tilde{u}_t)] + \frac{\gamma_x^2 \eta^2 L_\theta}{2} \mathbb{E}[\|\tilde{u}_t\|^2]
$$

$$
= \mathbb{E}[\Phi(x_t)] - \frac{\gamma_x \eta}{2} \mathbb{E}[\|\nabla \Phi(x_t)\|^2] + \left(\frac{\gamma_x^2 \eta^2 L_\theta}{2} - \frac{\gamma_x \eta}{2}\right) \mathbb{E}[\|\tilde{u}_t\|^2] + \frac{\gamma_x \eta}{2} \mathbb{E}[\|\nabla \Phi(x_t) - \tilde{u}_t\|^2]
$$

$$
\leq \mathbb{E}[\Phi(x_t)] - \frac{\gamma_x \eta}{2} \mathbb{E}[\|\nabla \Phi(x_t)\|^2] - \frac{\gamma_x \eta}{4} \mathbb{E}[\|\tilde{u}_t\|^2]
$$

$$
+ 3\gamma_x \eta \mathbb{E}\left[\left\|\nabla \Phi(x_t) - \frac{1}{K} \sum_{k=1}^{K} \nabla_x f(k)(\frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_t), \bar{y}_t)\right\|^2\right]
$$

$$
+ 3\gamma_x \eta \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^{K} \nabla_x f(k)(g^{(k)}(x_t), y^{(k)}_t) - \frac{1}{K} \sum_{k=1}^{K} \nabla_x f(k)(g^{(k)}(x_t^k), y^{(k)}_t)\right\|^2\right]
$$

$$
+ 3\gamma_x \eta \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^{K} \nabla_x f(k)(g^{(k)}(x_t^k), y^{(k)}_t) - \bar{u}_t\right\|^2\right],
$$

(25)

where the last inequality holds due to $\eta \leq \frac{1}{2\gamma_x L_\theta}$. As for $T_1$, we can get

$$
T_1 = \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^{K} \nabla \Phi(k)(x_t) - \frac{1}{K} \sum_{k=1}^{K} \nabla_x f(k)(\frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_t), \bar{y}_t)\right\|^2\right]
$$

$$
\leq \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\left[\left\|\left(\frac{1}{K} \sum_{k'=1}^{K} \nabla g^{(k')}(x_t)\right) + \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_t)\right\|^2\right]
$$

$$
- \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_t), y_s(x_t)\right\|^2\right]
$$

$$
\leq \frac{C_g^2 L_\theta^2}{K} \mathbb{E}[\|y_s(x_t) - \bar{y}_t\|^2],
$$

(26)

where the last step holds due to Assumptions [1,2]. As for $T_2$, we can get

$$
T_2 = \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^{K} \nabla_x f(k)(\frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_t), \bar{y}_t) - \frac{1}{K} \sum_{k=1}^{K} \nabla_x f(k)(g^{(k)}(x_t), y_t^{(k)})\right\|^2\right]
$$

$$
+ \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\left[\left\|\nabla_x f(k)(g^{(k)}(x_t), y^{(k)}_t) - \frac{1}{K} \sum_{k'=1}^{K} \nabla_x f(k)(g^{(k')}(x_t), y^{(k)}_t)\right\|^2\right]
$$

$$
\leq 2 \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^{K} \nabla_x f(k)(\frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_t), \bar{y}_t) - \frac{1}{K} \sum_{k=1}^{K} \nabla_x f(k)(g^{(k)}(x_t), y^{(k)}_t)\right\|^2\right]
$$

$$
+ 2 \mathbb{E}\left[\left\|\frac{1}{K} \sum_{k=1}^{K} \nabla_x f(k)(g^{(k)}(x_t), y^{(k)}_t) - \frac{1}{K} \sum_{k=1}^{K} \nabla_x f(k)(g^{(k)}(x_t), y^{(k)}_t)\right\|^2\right].
$$

(27)
Then, as for $T_3$, we can get
\begin{align*}
T_3 & \leq \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \nabla_x f^{(k)} \left( g^{(k)}(\bar{x}_t), \bar{y}_t \right) - \nabla_x f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) \right\|^2 \right] \\
& = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \left( \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(\bar{x}_t) \right)^T \nabla_g f^{(k)} \left( \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(\bar{x}_t), \bar{y}_t \right) - \nabla_g f^{(k)}(g^{(k)}(\bar{x}_t), y_t^{(k)}) \right\|^2 \right] \\
& \leq C_g \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \nabla_g f^{(k)}(g^{(k)}(\bar{x}_t), y_t^{(k)}) \right\|^2 \right] \\
& \leq C_g^2 L_g^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \bar{y}_t - y_t^{(k)} \right\|^2 \right],
\end{align*}
where the third step holds due to the homogeneous data distribution assumption and Assumption 2 and the last step also holds due to the homogeneous data distribution assumption and Assumption 1.

As for $T_4$, we can get
\begin{align*}
T_4 & \leq \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \nabla_x f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) - \nabla_x f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) \right\|^2 \right] \\
& = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \nabla g^{(k)}(\bar{x}_t)^T \nabla_g f^{(k)}(g^{(k)}(\bar{x}_t), y_t^{(k)}) - \nabla g^{(k)}(\bar{x}_t)^T \nabla_g f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) \right\|^2 \right] \\
& \leq 2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \nabla g^{(k)}(\bar{x}_t)^T \nabla_g f^{(k)}(g^{(k)}(\bar{x}_t), y_t^{(k)}) - \nabla g^{(k)}(\bar{x}_t)^T \nabla_g f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) \right\|^2 \right] \\
& \quad + \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \nabla g^{(k)}(\bar{x}_t)^T \nabla_g f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) - \nabla g^{(k)}(\bar{x}_t)^T \nabla_g f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) \right\|^2 \right] \\
& \leq 2C_g^2 L_g^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| g^{(k)}(\bar{x}_t) - g^{(k)}(x_t^{(k)}) \right\|^2 \right] + 2C_g^2 L_g^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| x_t^{(k)} \right\|^2 \right] \\
& \leq 2C_g^2 L_g^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| x_t^{(k)} \right\|^2 \right] + 2C_g^2 L_g^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| x_t^{(k)} \right\|^2 \right] \\
& = 2(C_g^4 L_g^4 + C_g^2 L_g^2) \frac{1}{K} \sum_{k=1}^{K} \left[ \left\| x_t^{(k)} \right\|^2 \right].
\end{align*}
where fourth and fifth steps hold due to Assumptions 1 and 2. By combining $T_3$ and $T_4$, we can get
\begin{align*}
T_2 & \leq 2C_g^2 L_g^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| y_t - y_t^{(k)} \right\|^2 \right] + 4(C_g^4 L_g^4 + C_g^2 L_g^2) \frac{1}{K} \sum_{k=1}^{K} \left[ \left\| x_t^{(k)} \right\|^2 \right].
\end{align*}
By combining $T_1$ and $T_2$, we can get
\begin{align*}
\mathbb{E}\left[ \Phi(\bar{x}_{t+1}) \right] & \leq \mathbb{E}\left[ \Phi(\bar{x}_t) \right] - \frac{\gamma_x \eta}{2} \mathbb{E} \left[ \left\| \nabla \Phi(\bar{x}_t) \right\|^2 \right] - \frac{\gamma_x \eta}{4} \mathbb{E} \left[ \left\| \bar{u}_t \right\|^2 \right] + 3\gamma_x \eta \mathbb{E} \left[ \left\| y_*(\bar{x}_t) - \bar{y}_t \right\|^2 \right] \\
& \quad + 12\gamma_x \eta (C_g^4 L_g^4 + C_g^2 L_g^2) \frac{1}{K} \sum_{k=1}^{K} \left[ \left\| x_t^{(k)} \right\|^2 \right] + 6\gamma_x \eta C_g^2 L_g^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \bar{y}_t - y_t^{(k)} \right\|^2 \right] \\
& \quad + 3\gamma_x \eta \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_x f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) - \bar{u}_t \right\|^2 \right],
\end{align*}
which completes the proof. \qed
Lemma 4. Given Assumption 1 and $\beta, \eta \in (0, 1)$, we have

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|u_{t+1}^{(k)} - \bar{u}_{t+1}\|^2] \leq 6p^2\beta_x^2\eta^2C_g^2C_f^2.$$  \hfill (32)

Proof.

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|u_{t+1}^{(k)} - \bar{u}_{t+1}\|^2] = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[(1 - \beta_x \eta)u_t^{(k)} + \beta_x \eta \nabla g^{(k)}(x_{t+1}^{(k)}, \xi_{t+1}^{(k)})^T \nabla g^{(k)}(h_{t+1}^{(k)}, y_{t+1}^{(k)}, \zeta_{t+1}^{(k)})] - (1 - \beta_x \eta)u_t - \beta_x \eta \frac{1}{K} \sum_{k'=1}^{K} \nabla g^{(k')}(x_{t+1}^{(k')}, \xi_{t+1}^{(k')})^T \nabla g^{(k')}(h_{t+1}^{(k')}, y_{t+1}^{(k')}, \zeta_{t+1}^{(k')})\|^2]$$

$$\leq (1 - \beta_x \eta)^2(1 + \frac{1}{p}) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|u_t^{(k)} - \bar{u}_t\|^2] + (1 + p)\beta_x^2\eta^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|\nabla g^{(k)}(x_{t+1}^{(k)}; \xi_{t+1}^{(k)})^T \nabla g^{(k)}(h_{t+1}^{(k)}, y_{t+1}^{(k)}, \zeta_{t+1}^{(k)}) - \nabla g^{(k)}(x_{t+1}^{(k)}, \xi_{t+1}^{(k)})^T \nabla g^{(k)}(h_{t+1}^{(k)}; y_{t+1}^{(k)}, \zeta_{t+1}^{(k)})\|^2]$$

$$\leq (1 - \beta_x \eta)^2(1 + \frac{1}{p}) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|u_t^{(k)} - \bar{u}_t\|^2] + 2p\beta_x^2\eta^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|\nabla g^{(k)}(x_{t+1}^{(k)}; \xi_{t+1}^{(k)})^T \nabla g^{(k)}(h_{t+1}^{(k)}; y_{t+1}^{(k)}, \zeta_{t+1}^{(k)}) - \nabla g^{(k)}(x_{t+1}^{(k)}, \xi_{t+1}^{(k)})^T \nabla g^{(k)}(h_{t+1}^{(k)}; y_{t+1}^{(k)}, \zeta_{t+1}^{(k)})\|^2]$$

$$\leq (1 - \beta_x \eta)^2(1 + \frac{1}{p}) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|u_t^{(k)} - \bar{u}_t\|^2] + 2p\beta_x^2\eta^2 C_g^2C_f^2$$

$$\leq 2p\beta_x^2\eta^2 C_g^2C_f^2 \sum_{t'=s_t+1}^{t} (1 + \frac{1}{p})^{t-t'} \leq 6p^2\beta_x^2\eta^2C_g^2C_f^2,$$

where $s_t = \lfloor (t + 1)/p \rfloor$, the third step holds due to $\beta_x \eta \in (0, 1)$ and $1 + p < 2p$, the fifth step holds due to Assumption 2, the last step holds due to $(1 + \frac{1}{p})^p < 3$. \hfill \Box

Lemma 5. Given Assumption 1, $\beta \eta \in (0, 1)$, and $\eta \leq \frac{1}{10p^2\gamma \sigma L_f}$, we have

$$\frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|v_{t+1}^{(k)} - \bar{v}_t\|^2] \leq 3456\beta_x^2\gamma A^2 C_g^2 L_f^2 \sigma_g^2 + 41472\beta_x^2\gamma^2 A^2 \beta_x^2 p^2 \gamma^2 C_g^2 C_f^2 \sigma_f^2 + 96\beta_x^2 p^2 \eta^2 \sigma_f^2.$$  \hfill (34)
where the third step holds due to \( \beta \eta \in (0, 1) \) and \( 1 + p < 2p \).

The last term can be bounded as follows:

\[
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|v_{t+1}^{(k)} - \tilde{v}_{t+1}\|^2] \\
= \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|1 - \beta \eta \|v_t^{(k)} + \beta \eta \nabla f^{(k)}(h_{t+1}^{(k)}, y_{t+1}; \zeta_{t+1}^{(k)}) - (1 - \beta \eta) v_t - \beta \eta \|v_t^{(k)} - v_{t+1}\|^2] \\
\leq (1 - \beta \eta)^2 (1 + \frac{1}{p}) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|v_t^{(k)} - \tilde{v}_t\|^2] + (1 + p) \beta^2 \eta^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|\nabla f^{(k)}(h_{t+1}^{(k)}, y_{t+1}; \zeta_{t+1}^{(k)}) - \frac{1}{K} \sum_{k'=1}^{K} \nabla f^{(k')}(h_{t+1}^{(k')}, y_{t+1}; \zeta_{t+1}^{(k')})\|^2] \\
\leq (1 - \beta \eta)^2 (1 + \frac{1}{p}) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|v_t^{(k)} - \tilde{v}_t\|^2] + 2p \beta^2 \eta^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|\nabla f^{(k)}(h_{t+1}^{(k)}, y_{t+1}; \zeta_{t+1}^{(k)}) - \frac{1}{K} \sum_{k'=1}^{K} \nabla f^{(k')}(h_{t+1}^{(k')}, y_{t+1}; \zeta_{t+1}^{(k')})\|^2] \\
+ \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|\nabla f^{(k)}(h_{t+1}^{(k)}, y_{t+1}; \zeta_{t+1}^{(k)}) - \nabla f^{(k)}(h_{t+1}^{(k)}, y_{t+1})\|^2] \\
+ \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|\nabla f^{(k')}(h_{t+1}^{(k')}, y_{t+1}; \zeta_{t+1}^{(k')}) - \nabla f^{(k')}(h_{t+1}^{(k')}, y_{t+1}; \zeta_{t+1}^{(k')})\|^2] \\
\leq 8\sigma_f^2 + 8L_f^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|h_{t+1}^{(k)} - \tilde{h}_{t+1}\|^2] + 8L_f^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|y_{t+1}^{(k)} - y_{t+1}\|^2].
\]

Then, by combining these two inequalities, we can get

\[
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|v_{t+1}^{(k)} - \tilde{v}_{t+1}\|^2] \leq (1 + \frac{1}{p}) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|v_t^{(k)} - \tilde{v}_t\|^2] + 16p \beta^2 \eta^2 \sigma_f^2 \\
+ 16p \beta^2 \eta^2 L_f^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|h_{t+1}^{(k)} - \tilde{h}_{t+1}\|^2] + 16p \beta^2 \eta^2 L_f^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|y_{t+1}^{(k)} - y_{t+1}\|^2].
\]
In addition, we have
\[
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} [ \| y_{t+1}^{(k)} - y_{t+1} \|^{2} ] = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} [ \| y_{s_{t+1}^{(k)}} + \gamma_{y} \eta \sum_{t' = s_{t+1}^{(k)}}^{t} v_{t'} - \bar{y}_{s_{t+1}^{(k)}} - \gamma_{y} \eta \sum_{t' = s_{t+1}^{(k)}}^{t} \bar{v}_{t'} \|^{2} ]
\]
\[
\leq P^{2} \gamma_{y} \eta^{2} \frac{1}{K} \sum_{k=1}^{K} \sum_{t' = s_{t+1}^{(k)}}^{t} \mathbb{E} [ \| v_{t'}^{(k)} - \bar{v}_{t'} \|^{2} ] ,
\]
where \( s_{t} = \lfloor (t + 1)/p \rfloor \). Thus, we can get
\[
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} [ \| v_{t+1}^{(k)} - \bar{v}_{t+1} \|^{2} ] \leq (1 + \frac{1}{p}) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} [ \| v_{t}^{(k)} - \bar{v}_{t} \|^{2} ] + 16p^{2} \beta_{y}^{2} \eta^{2} \sigma_{f}^{2}
\]
\[
+ 16p^{2} \gamma_{y}^{2} \eta^{2} L_{f}^{2} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} [ \| h_{t+1}^{(k)} - \bar{h}_{t+1} \|^{2} ] + 16p^{2} \gamma_{y}^{2} \beta_{y}^{2} \eta^{4} L_{f}^{2} \frac{1}{K} \sum_{k=1}^{K} \sum_{t' = s_{t+1}^{(k)}}^{t} \mathbb{E} [ \| v_{t'}^{(k)} - \bar{v}_{t'} \|^{2} ]
\]
\[
\leq 16p^{2} \gamma_{y}^{2} \beta_{y}^{2} \eta^{4} L_{f}^{2} \frac{1}{K} \sum_{k=1}^{K} \sum_{t' = s_{t+1}^{(k)}}^{t} \frac{1}{t' - s_{t+1}^{(k)}} \mathbb{E} [ \| v_{t'}^{(k)} - \bar{v}_{t'} \|^{2} ]
\]
\[
+ 16p^{2} \beta_{y}^{2} \eta^{2} L_{f}^{2} \frac{1}{K} \sum_{k=1}^{K} \sum_{t' = s_{t+1}^{(k)}}^{t} \mathbb{E} [ \| h_{t+1}^{(k)} - \bar{h}_{t+1} \|^{2} ] + 16p^{2} \gamma_{y}^{2} \beta_{y}^{2} \eta^{2} \sigma_{f}^{2}
\]
\[
\leq 48p^{3} \gamma_{y}^{2} \beta_{y}^{2} \eta^{4} L_{f}^{2} \frac{1}{K} \sum_{k=1}^{K} \sum_{t' = s_{t+1}^{(k)}}^{t} \mathbb{E} [ \| v_{t'}^{(k)} - \bar{v}_{t'} \|^{2} ]
\]
\[
+ 48p^{2} \beta_{y}^{2} \eta^{2} L_{f}^{2} \frac{1}{K} \sum_{k=1}^{K} \sum_{t' = s_{t+1}^{(k)}}^{t} \mathbb{E} [ \| h_{t+1}^{(k)} - \bar{h}_{t+1} \|^{2} ] + 48p^{2} \beta_{y}^{2} \eta^{2} \sigma_{f}^{2} ,
\]
where the last step holds due to \((1 + \frac{1}{p})p < 3\). Then, we can get
\[
\frac{1}{T} \sum_{t=1}^{T-1} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} [ \| v_{t}^{(k)} - \bar{v}_{t} \|^{2} ] \leq 48p^{3} \gamma_{y}^{2} \beta_{y}^{2} \eta^{4} L_{f}^{2} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} [ \| v_{t}^{(k)} - \bar{v}_{t} \|^{2} ]
\]
\[
+ 48p^{2} \beta_{y}^{2} \eta^{2} L_{f}^{2} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} [ \| h_{t}^{(k)} - \bar{h}_{t} \|^{2} ] + 48p^{2} \beta_{y}^{2} \eta^{2} \sigma_{f}^{2}
\]
\[
\leq 48p^{3} \gamma_{y}^{2} \beta_{y}^{2} \eta^{4} L_{f}^{2} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} [ \| v_{t}^{(k)} - \bar{v}_{t} \|^{2} ] + 48p^{2} \beta_{y}^{2} \eta^{2} L_{f}^{2} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} [ \| h_{t}^{(k)} - \bar{h}_{t} \|^{2} ]
\]
\[
+ 48p^{2} \beta_{y}^{2} \eta^{2} \sigma_{f}^{2} ,
\]
where the last step holds due to \( \beta_{y} \eta \leq 1 \). By setting \( \eta \leq \frac{1}{4p^{2} \gamma_{y} L_{f}} \) such that \( 1 - 48p^{3} \gamma_{y} \beta_{y}^{2} L_{f}^{2} \geq \frac{1}{2} \), we can get
\[
\frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} [ \| v_{t}^{(k)} - \bar{v}_{t} \|^{2} ] \leq 96p^{2} \beta_{y}^{2} \eta^{2} L_{f}^{2} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} [ \| h_{t}^{(k)} - \bar{h}_{t} \|^{2} ] + 96p^{2} \beta_{y}^{2} \eta^{2} \sigma_{f}^{2}
\]
\[
\leq 3456 \gamma_{y}^{2} \beta_{y}^{2} \eta^{4} \sum_{j}^{C_{g}^{2} L_{f}^{2} \sigma_{g}^{2}} + 41472 \beta_{y}^{2} \gamma_{y}^{2} \beta_{y}^{2} \eta^{8} \sum_{j}^{C_{g}^{4} C_{f}^{2} L_{f}^{2}} + 96 \beta_{y}^{2} \eta^{2} \sigma_{f}^{2} .
\]
where the last step holds due to Lemma 9.

**Lemma 6.** Given Assumption \[14\] we have
\[
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} [ \| x_{t+1}^{(k)} - x_{t+1} \|^{2} ] \leq 6 \gamma_{x}^{2} \beta_{x}^{2} p^{4} \eta^{4} \sum_{j}^{C_{g}^{2} C_{f}^{2}} .
\]

\[
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|x_t^{(k)} - \bar{x}_{t+1}\|^2] = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|x_{s,t}^{(k)} - \gamma_{x,t} \sum_{t'=s,t}^{t} u_{t'}^{(k)} - \bar{x}_{s,t} + \gamma_{x,t} \sum_{t'=s,t}^{t} \bar{u}_{t'}\|^2] 
\]

\[
\leq p\gamma_x^2 \eta_x^2 \frac{1}{K} \sum_{k=1}^{K} \sum_{t'=s,t}^{t} \mathbb{E}[\|u_{t'}^{(k)} - \bar{u}_{t'}\|^2] \leq 6\gamma_x^2 \beta_x^2 p^4 \eta^4 C_g^2 C_f^2 ,
\]

where \(s_t = \lfloor (t + 1)/p \rfloor\), the last step holds due to Lemma 4.

Lemma 7. Given Assumption 14 we have

\[
\frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|y_{t}^{(k)} - \bar{y}_t\|^2] 
\]

\[
\leq 3456 \gamma_y^2 \beta_y^2 \alpha^2 \eta_y^6 C_g^4 L_f^2 \sigma_g^2 + 41472 \gamma_y^2 \beta_y^2 \alpha^2 \gamma_x^2 \beta_x^2 p^{10} \eta^4 C_g^2 C_f^2 L_f^2 + 96 \gamma_y^2 \beta_y^2 p^4 \eta^4 \sigma_f^2 .
\]

Proof.  

\[
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|y_{t+1}^{(k)} - \bar{y}_{t+1}\|^2] = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|y_{s,t}^{(k)} + \gamma_{y,t} \sum_{t'=s,t}^{t} v_{t'}^{(k)} - \bar{y}_{s,t} + \gamma_{y,t} \sum_{t'=s,t}^{t} \bar{v}_{t'}\|^2] 
\]

\[
\leq p\gamma_y^2 \eta_y^2 \frac{1}{K} \sum_{k=1}^{K} \sum_{t'=s,t}^{t} \mathbb{E}[\|v_{t'}^{(k)} - \bar{v}_{t'}\|^2] ,
\]

where \(s_t = \lfloor (t + 1)/p \rfloor\). Then, it is easy to know

\[
\frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|y_{t}^{(k)} - \bar{y}_t\|^2] \leq p^2 \gamma_y^2 \eta_y^2 \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|v_{t}^{(k)} - \bar{v}_t\|^2] 
\]

\[
\leq 3456 \gamma_y^2 \beta_y^2 \alpha^2 \eta_y^6 C_g^4 L_f^2 \sigma_g^2 + 41472 \gamma_y^2 \beta_y^2 \alpha^2 \gamma_x^2 \beta_x^2 p^{10} \eta^4 C_g^2 C_f^2 L_f^2 + 96 \gamma_y^2 \beta_y^2 p^4 \eta^4 \sigma_f^2 .
\]

where the last step holds due to Lemma 5.

Lemma 8. Given Assumptions 14 we can get

\[
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| g^{(k)}(x_t^{(k)}) - \frac{1}{K} \sum_{k'=1}^{K} g^{(k')} (x_t^{(k')}) \right\| ^2 \right] \leq 24 \gamma_x^2 \beta_x^2 \omega_p^4 \eta^4 C_g^2 C_f^2 .
\]

Proof.  

\[
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| g^{(k)}(x_t^{(k)}) - \frac{1}{K} \sum_{k'=1}^{K} g^{(k')} (x_t^{(k')}) \right\| ^2 \right] 
\]

\[
\leq \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| g^{(k)}(x_t^{(k)}) - g^{(k)}(\bar{x}_t) + g(\bar{x}_t) - \frac{1}{K} \sum_{k'=1}^{K} g^{(k')} (x_t^{(k')}) \right\| ^2 \right] 
\]

\[
\leq 2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| g^{(k)}(x_t^{(k)}) - g^{(k)}(\bar{x}_t) \right\| ^2 \right] + 2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| g(\bar{x}_t) - \frac{1}{K} \sum_{k'=1}^{K} g^{(k')} (x_t^{(k')}) \right\| ^2 \right] 
\]

\[
\leq 4C_g^2 \omega_p^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|x_t^{(k)} - \bar{x}_t\|^2] 
\]

\[
\leq 24 \gamma_x^2 \beta_x^2 \omega_p^4 \eta^4 C_g^4 C_f^2 ,
\]

where the second step holds due to \(g^{(k)}(\bar{x}_t) = g(\bar{x}_t)\) for the homogeneous data distribution, the third step holds due to Assumption 14 the last step holds due to Lemma 6.
Lemma 9. Given Assumptions[14], we can get

$$
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| h_{t+1}^{(k)} - \frac{1}{K} \sum_{k'=1}^{K} h_{t+1}^{(k')} \right\|^2 \right] \leq 36\alpha^2p^2\eta^2C_g^2\sigma_g^2 + 432\alpha^2\gamma_2^2\beta_2^2p^6 \eta^6C_g^6C_f^2 .
$$

(49)

Proof.

$$
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| h_{t+1}^{(k)} - \frac{1}{K} \sum_{k'=1}^{K} h_{t+1}^{(k')} \right\|^2 \right]
= \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| (1 - \alpha\eta)h_{t}^{(k)} + \alpha\eta g^{(k)}(x_{t+1}^{(k)};\xi_{t+1}^{(k)}) - (1 - \alpha\eta) \frac{1}{K} \sum_{k'=1}^{K} h_{t}^{(k')} - \alpha\eta \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_{t+1}^{(k')};\xi_{t+1}^{(k')}) \right\|^2 \right]
\leq (1 - \alpha\eta)^2(1 + \frac{1}{p}) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| h_{t}^{(k)} - \frac{1}{K} \sum_{k'=1}^{K} h_{t}^{(k')} \right\|^2 \right]
+ \alpha^2\eta^2(1 + p) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \left(1 + \frac{1}{p}\right) \frac{1}{K} \sum_{k'=1}^{K} h_{t}^{(k')} \right\|^2 \right]
+ \frac{2p\alpha^2\eta^2}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| g^{(k)}(x_{t+1}^{(k)};\xi_{t+1}^{(k)}) - \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_{t+1}^{(k')};\xi_{t+1}^{(k')}) \right\|^2 \right]
\leq (1 + \frac{1}{p}) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| h_{t}^{(k)} - \frac{1}{K} \sum_{k'=1}^{K} h_{t}^{(k')} \right\|^2 \right]
+ \frac{2\alpha^2\eta^2C_g^2}{K} \sum_{t'=t+1}^{t} (1 + \frac{1}{p}) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| g^{(k)}(x_{t'}^{(k)};\xi_{t'}^{(k)}) - \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_{t'}^{(k')};\xi_{t'}^{(k')}) \right\|^2 \right]
,$$

(50)

where $s_t = \lfloor (t + 1)/p \rfloor$, the third step holds due to $\alpha\eta \in (0, 1)$ and $1 + p < 2p$. In addition, we can get

$$
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| g^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_t^{(k')};\xi_t^{(k')}) \right\|^2 \right]
= \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ g^{(k)}(x_t^{(k)};\xi_t^{(k)}) - g^{(k)}(x_t^{(k)}) + g^{(k)}(x_t^{(k)}) - \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_t^{(k')};\xi_t^{(k')}) \right]^2
\leq 6\sigma_g^2 + \frac{3}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| g^{(k)}(x_t^{(k)}) - \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_t^{(k')}) \right\|^2 \right]
\leq 6\sigma_g^2 + 72\gamma_2^2\beta_2^2p^4 \eta^4C_g^4C_f^2 ,
$$

(51)

where the last step holds due to Lemma[8]. Therefore, we can get

$$
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| h_{t+1}^{(k)} - \frac{1}{K} \sum_{k'=1}^{K} h_{t+1}^{(k')} \right\|^2 \right]
\leq 2\alpha^2\eta^2C_g^2 \sum_{t'=t+1}^{t} (1 + \frac{1}{p}) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| g^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_t^{(k')};\xi_t^{(k')}) \right\|^2 \right]
\leq 2\alpha^2\eta^2C_g^2 (6\sigma_g^2 + 72\gamma_2^2\beta_2^2p^4 \eta^4C_g^4C_f^2) \sum_{t'=t+1}^{t} (1 + \frac{1}{p}) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| g^{(k)}(x_t^{(k)};\xi_t^{(k)}) - \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_t^{(k')};\xi_t^{(k')}) \right\|^2 \right]
\leq 36\alpha^2p^2\eta^2C_g^2 \sigma_g^2 + 432\alpha^2\gamma_2^2\beta_2^2p^6 \eta^6C_g^6C_f^2 .
$$

(52)

where the last step holds due to $(1 + \frac{1}{p})^p < 3$. □
Lemma 10. Given Assumptions \[7\] \[2\] we can get

\[
\frac{1}{K} \sum_{k=1}^{K} E \left[ \left\| \nabla_x f^{(k)}(g^{(k)}(x_t^{(k)}),y_t^{(k)}) - \nabla_x f^{(k)}(g^{(k)}(x_{t-1}^{(k)}),y_{t-1}) \right\|^2 \right] 
\leq 4\gamma_2^2\eta^2 (C_g^4 L_f^2 + C_f^2 L_g^2) E \left[ \left\| \bar{u}_{t-1} \right\|^2 \right] 
+ 4\gamma_2^2\eta^2 C_g^2 L_f^2 E \left[ \left\| \bar{v}_{t-1} \right\|^2 \right] 
+ 24\beta_2^2\gamma_2 p^2 \eta^4 (C_g^4 L_f^2 + C_f^2 L_g^2) C_g^2 C_f^2 
+ 13824\beta_2^2\gamma_2^2 \alpha^2 p^2 \eta^6 C_g^4 L_f^4 \sigma_g^2 + 165888\beta_2^2\gamma_2^2 \alpha^2 \beta_2^2 p^8 \eta^{10} C_g^8 C_f^2 L_f^4 
+ 384\beta_2^2\gamma_2 p^2 \eta^4 C_g^2 L_f^5 \sigma_f^2 .
\]

(53)

Proof.

\[
\frac{1}{K} \sum_{k=1}^{K} E \left[ \left\| \nabla_x f^{(k)}(g^{(k)}(x_t^{(k)}),y_t^{(k)}) - \nabla_x f^{(k)}(g^{(k)}(x_{t-1}^{(k)}),y_{t-1}) \right\|^2 \right] 
= \frac{1}{K} \sum_{k=1}^{K} E \left[ \left\| \nabla g^{(k)}(x_t^{(k)})^T \nabla g^{(k)}(g^{(k)}(x_t^{(k)}),y_t^{(k)}) - \nabla g^{(k)}(x_t^{(k)})^T \nabla g^{(k)}(g^{(k)}(x_{t-1}^{(k)}),y_{t-1}) \right\|^2 \right] 
+ \frac{1}{K} \sum_{k=1}^{K} E \left[ \left\| \nabla g^{(k)}(x_t^{(k)})^T \nabla f^{(k)}(g^{(k)}(x_t^{(k)}),y_t^{(k)}) - \nabla g^{(k)}(x_t^{(k)})^T \nabla f^{(k)}(g^{(k)}(x_{t-1}^{(k)}),y_{t-1}) \right\|^2 \right] 
\leq \frac{2C^2_g L_f^2 \left( C_g^2 \frac{1}{K} \sum_{k=1}^{K} E \left[ \left\| x_t^{(k)} - x_{t-1}^{(k)} \right\|^2 \right] + \frac{1}{K} \sum_{k=1}^{K} E \left[ \left\| y_t^{(k)} - y_{t-1}^{(k)} \right\|^2 \right] \right) + 2C^2_g L_f^2 \frac{1}{K} \sum_{k=1}^{K} E \left[ \left\| x_t^{(k)} - x_{t-1}^{(k)} \right\|^2 \right] 
= 2(C_g^4 L_f^2 + C_f^2 L_g^2) \frac{1}{K} \sum_{k=1}^{K} E \left[ \left\| x_t^{(k)} - x_{t-1}^{(k)} \right\|^2 \right] 
+ 2C^2_g L_f^2 \frac{1}{K} \sum_{k=1}^{K} E \left[ \left\| y_t^{(k)} - y_{t-1}^{(k)} \right\|^2 \right] 
\leq 4\gamma_2^2\eta^2 (C_g^4 L_f^2 + C_f^2 L_g^2) \frac{1}{K} \sum_{k=1}^{K} E \left[ \left\| u_{t-1}^{(k)} - \bar{u}_{t-1} \right\|^2 \right] 
+ 4\gamma_2^2\eta^2 (C_g^4 L_f^2 + C_f^2 L_g^2) E \left[ \left\| \bar{u}_{t-1} \right\|^2 \right] 
+ 4\gamma_2^2\eta^2 C_g^2 L_f^2 \frac{1}{K} \sum_{k=1}^{K} E \left[ \left\| v_{t-1}^{(k)} - \bar{v}_{t-1} \right\|^2 \right] 
+ 4\gamma_2^2\eta^2 C_g^2 L_f^2 \frac{1}{K} \sum_{k=1}^{K} E \left[ \left\| \bar{v}_{t-1} \right\|^2 \right] 
\leq 4\gamma_2^2\eta^2 (C_g^4 L_f^2 + C_f^2 L_g^2) E \left[ \left\| \bar{u}_{t-1} \right\|^2 \right] 
+ 4\gamma_2^2\eta^2 C_g^2 L_f^2 E \left[ \left\| \bar{v}_{t-1} \right\|^2 \right] 
+ 24\beta_2^2\gamma_2 p^2 \eta^4 (C_g^4 L_f^2 + C_f^2 L_g^2) C_g^2 C_f^2 
+ 13824\beta_2^2\gamma_2^2 \alpha^2 p^2 \eta^6 C_g^4 L_f^4 \sigma_g^2 + 165888\beta_2^2\gamma_2^2 \alpha^2 \beta_2^2 p^8 \eta^{10} C_g^8 C_f^2 L_f^4 
+ 384\beta_2^2\gamma_2 p^2 \eta^4 C_g^2 L_f^5 \sigma_f^2 .
\]

(54)

where the third step holds due to Assumptions \[1\] \[2\] the last step holds due to Lemma \[4\] and Lemma \[5\]. □

Lemma 11. Given Assumptions \[7\] \[2\] and \[\alpha \eta \in (0, 1)\], we can get

\[
\frac{1}{T} \sum_{t=0}^{T-1} E \left[ \left\| h_t - \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(x_t^{(k)}) \right\|^2 \right] \leq \frac{2\gamma_2^2 C_g^2}{\alpha^2} \frac{1}{T} \sum_{t=0}^{T-1} E \left[ \left\| u_t \right\|^2 \right] + \frac{\sigma_g^2}{\alpha \eta T K} + \frac{12\beta_2^2\gamma_2 p^2 \eta^4 C_g^4 L_f^2 + \alpha \eta \sigma_g^2}{\alpha^2} .
\]

(55)
Proof.

\[
\begin{align*}
\mathbb{E} \left[ \left\| \mathbf{h}_{t+1} - \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(\mathbf{x}_{t+1}^{(k)}) \right\|^2 \right] &= \mathbb{E} \left[ \left\| (1 - \alpha \eta) \frac{1}{K} \sum_{k=1}^{K} \mathbf{h}^{(k)}_t + \alpha \eta \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(\mathbf{x}_{t+1}^{(k)}; \zeta_t^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(\mathbf{x}_{t+1}^{(k)}) \right\|^2 \right] \\
&= \mathbb{E} \left[ \left\| (1 - \alpha \eta) \frac{1}{K} \sum_{k=1}^{K} \left( \mathbf{h}^{(k)}_t - g^{(k)}(\mathbf{x}_t^{(k)}) \right) + (1 - \alpha \eta) \frac{1}{K} \sum_{k=1}^{K} \left( g^{(k)}(\mathbf{x}_{t+1}^{(k)}) - g^{(k)}(\mathbf{x}_t^{(k)}) \right) \right\|^2 \right] \\
&\quad + \alpha \eta \frac{1}{K} \sum_{k=1}^{K} \left( g^{(k)}(\mathbf{x}_{t+1}^{(k)}; \zeta_t^{(k)}) - g^{(k)}(\mathbf{x}_t^{(k)}) \right) \right\|^2 \right] \\
&\quad + \mathbb{E} \left[ \left\| \alpha \eta \frac{1}{K} \sum_{k=1}^{K} \left( g^{(k)}(\mathbf{x}_{t+1}^{(k)}; \zeta_t^{(k)}) - g^{(k)}(\mathbf{x}_t^{(k)}) \right) \right\|^2 \right] \\
&\leq (1 - \alpha \eta)^2 (1 + \frac{1}{a}) \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \left( \mathbf{h}^{(k)}_t - g^{(k)}(\mathbf{x}_t^{(k)}) \right) \right\|^2 \right] + \frac{\alpha^2 \eta^2 \sigma^2}{K} \\
&\quad + (1 - \alpha \eta)^2 (1 + a) \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \left( g^{(k)}(\mathbf{x}_{t+1}^{(k)}) - g^{(k)}(\mathbf{x}_t^{(k)}) \right) \right\|^2 \right] \\
&\leq (1 - \alpha \eta) \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \left( \mathbf{h}^{(k)}_t - g^{(k)}(\mathbf{x}_t^{(k)}) \right) \right\|^2 \right] + C_0^2 \frac{\alpha}{\alpha \eta} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \mathbf{x}_{t+1}^{(k)} - \mathbf{x}_t^{(k)} \right\|^2 \right] + \frac{\alpha^2 \eta^2 \sigma^2}{K} \\
&\leq (1 - \alpha \eta) \mathbb{E} \left[ \left\| \mathbf{h}_t - \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(\mathbf{x}_t^{(k)}) \right\|^2 \right] + 2 \gamma \frac{\sigma^2}{\alpha} \mathbb{E} \left[ \left\| \mathbf{u}_t \right\|^2 \right] + 12 \frac{\sigma^2 \beta^2 \eta^2 C_0^4 C_0^2}{\alpha^2} + \frac{\alpha^2 \eta^2 \sigma^2}{K}, \\
\end{align*}
\]

where the fifth step holds due to \( a = \frac{1 - \alpha \eta}{\alpha \eta} \) and \( \alpha \eta < 1 \), the second to last step holds due to Assumption 3, the second to last step holds due to Lemma 4. The inequality (56).

It can be reformulated as follows:

\[
\begin{align*}
\alpha \eta \mathbb{E} \left[ \left\| \mathbf{h}_t - \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(\mathbf{x}_t^{(k)}) \right\|^2 \right] \leq \mathbb{E} \left[ \left\| \mathbf{h}_t - \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(\mathbf{x}_t^{(k)}) \right\|^2 \right] \\
- \mathbb{E} \left[ \left\| \mathbf{h}_{t+1} - \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(\mathbf{x}_{t+1}^{(k)}) \right\|^2 \right] + 2 \gamma \frac{\sigma^2}{\alpha} \mathbb{E} \left[ \left\| \mathbf{u}_t \right\|^2 \right] + 12 \frac{\sigma^2 \beta^2 \eta^2 C_0^4 C_0^2}{\alpha^2} + \frac{\alpha^2 \eta^2 \sigma^2}{K}. \\
\end{align*}
\]

By summing over \( t \) from 0 to \( T - 1 \), we can get

\[
\begin{align*}
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \left\| \mathbf{h}_t - \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(\mathbf{x}_t^{(k)}) \right\|^2 \right] \\
\leq \frac{1}{\alpha \eta T} \mathbb{E} \left[ \left\| \mathbf{h}_0 - \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(\mathbf{x}_0^{(k)}) \right\|^2 \right] + 2 \gamma \frac{\sigma^2}{\alpha} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \left\| \mathbf{u}_t \right\|^2 \right] + 12 \frac{\sigma^2 \beta^2 \eta^2 C_0^4 C_0^2}{\alpha^2} + \frac{\alpha \eta \sigma^2}{K} \\
\leq 2 \gamma \frac{\sigma^2}{\alpha^2} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \left\| \mathbf{u}_t \right\|^2 \right] + \frac{\sigma^2}{\alpha \eta T K} + 12 \frac{\sigma^2 \beta^2 \eta^2 C_0^4 C_0^2}{\alpha^2} + \frac{\alpha \eta \sigma^2}{K},
\end{align*}
\]

where the last step holds due to the following inequality:

\[
\begin{align*}
\mathbb{E} \left[ \left\| \mathbf{h}_0 - \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(\mathbf{x}_0^{(k)}) \right\|^2 \right] = \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(\mathbf{x}_0^{(k)}; \zeta_0^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(\mathbf{x}_0^{(k)}) \right\|^2 \right] \leq \frac{\sigma^2}{K}. \\
\end{align*}
\]
Lemma 12. Given Assumptions 1, 2, we can get

\[
E \left[ \frac{1}{K} \sum_{k=1}^{K} \left( \nabla g^{(k)}(\mathbf{x}^{(k)}_t; \xi_t^{(k)})^T \nabla g f^{(k)}(\mathbf{h}^{(k)}_t; \mathbf{y}^{(k)}_t) - \nabla g^{(k)}(\mathbf{x}^{(k)}_t; \xi_t^{(k)})^T \nabla g f^{(k)}(\mathbf{h}^{(k)}_t; \mathbf{y}^{(k)}_t) \right) \right] \leq \frac{C^2 \sigma^2}{K},
\]

for all random factors except the sampling operation in the t-th iteration, \( \xi \) and \( \zeta \) are independent random vectors regarding \( \xi \) and \( \zeta \), respectively. The last step holds due to Assumptions 1, 2.

Proof.

\[
E \left[ \frac{1}{K} \sum_{k=1}^{K} \left( \nabla g^{(k)}(\mathbf{x}^{(k)}_t; \xi_t^{(k)})^T \nabla g f^{(k)}(\mathbf{h}^{(k)}_t; \mathbf{y}^{(k)}_t) - \nabla g^{(k)}(\mathbf{x}^{(k)}_t; \xi_t^{(k)})^T \nabla g f^{(k)}(\mathbf{h}^{(k)}_t; \mathbf{y}^{(k)}_t) \right) \right] 
\]

\[
= \mathbb{E}_\mathcal{F} \mathbb{E}_{\xi | \mathcal{F}} \mathbb{E}_{\zeta | \mathcal{F}} \left[ \frac{1}{K} \sum_{k=1}^{K} \left( \nabla g^{(k)}(\mathbf{x}^{(k)}_t; \xi_t^{(k)})^T \nabla g f^{(k)}(\mathbf{h}^{(k)}_t; \mathbf{y}^{(k)}_t) - \nabla g^{(k)}(\mathbf{x}^{(k)}_t; \xi_t^{(k)})^T \nabla g f^{(k)}(\mathbf{h}^{(k)}_t; \mathbf{y}^{(k)}_t) \right) \right] 
\]

\[
= \mathbb{E}_\mathcal{F} \mathbb{E}_{\xi | \mathcal{F}} \mathbb{E}_{\zeta | \mathcal{F}} \left[ \frac{1}{K} \sum_{k=1}^{K} \left( \nabla g^{(k)}(\mathbf{x}^{(k)}_t; \xi_t^{(k)})^T \nabla g f^{(k)}(\mathbf{h}^{(k)}_t; \mathbf{y}^{(k)}_t) - \nabla g^{(k)}(\mathbf{x}^{(k)}_t; \xi_t^{(k)})^T \nabla g f^{(k)}(\mathbf{h}^{(k)}_t; \mathbf{y}^{(k)}_t) \right) \right] 
\]

\[
\leq \frac{\sigma^2 C^2}{K},
\]

where \( \mathcal{F} \) denotes all random factors except the sampling operation in the t-th iteration, \( \xi \) and \( \zeta \) are independent random vectors regarding \( \xi \) across workers and the mean is zero, the last step holds due to Assumptions 1, 2.

Similarly, we can get

\[
E \left[ \frac{1}{K} \sum_{k=1}^{K} \left( \nabla g^{(k)}(\mathbf{x}^{(k)}_t; \xi_t^{(k)})^T \nabla g f^{(k)}(\mathbf{h}^{(k)}_t; \mathbf{y}^{(k)}_t) - \nabla g^{(k)}(\mathbf{x}^{(k)}_t; \xi_t^{(k)})^T \nabla g f^{(k)}(\mathbf{h}^{(k)}_t; \mathbf{y}^{(k)}_t) \right) \right] 
\]

\[
= \mathbb{E}_\mathcal{F} \mathbb{E}_{\xi | \mathcal{F}} \mathbb{E}_{\zeta | \mathcal{F}} \left[ \frac{1}{K} \sum_{k=1}^{K} \left( \nabla g^{(k)}(\mathbf{x}^{(k)}_t; \xi_t^{(k)})^T \nabla g f^{(k)}(\mathbf{h}^{(k)}_t; \mathbf{y}^{(k)}_t) - \nabla g^{(k)}(\mathbf{x}^{(k)}_t; \xi_t^{(k)})^T \nabla g f^{(k)}(\mathbf{h}^{(k)}_t; \mathbf{y}^{(k)}_t) \right) \right] 
\]

\[
= \mathbb{E}_\mathcal{F} \mathbb{E}_{\xi | \mathcal{F}} \mathbb{E}_{\zeta | \mathcal{F}} \left[ \frac{1}{K} \sum_{k=1}^{K} \left( \nabla g^{(k)}(\mathbf{x}^{(k)}_t; \xi_t^{(k)})^T \nabla g f^{(k)}(\mathbf{h}^{(k)}_t; \mathbf{y}^{(k)}_t) - \nabla g^{(k)}(\mathbf{x}^{(k)}_t; \xi_t^{(k)})^T \nabla g f^{(k)}(\mathbf{h}^{(k)}_t; \mathbf{y}^{(k)}_t) \right) \right] 
\]

\[
\leq \frac{C^2 \sigma^2}{K},
\]

where \( \mathcal{F} \) denotes all random factors except the sampling operation in the t-th iteration, \( \xi \) and \( \zeta \) are independent random vectors regarding \( \xi \) across workers and the mean is zero, the last step holds due to Assumptions 1, 2.
Lemma 13. Given Assumption 4, \( \beta_x \eta \in (0, 1) \), and \( \eta < 1 \), we can get

\[
\begin{align*}
&\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_x f^{(k)}(g^{(k)}(x_t^{(k)}), y^{(k)}_t) - \frac{1}{K} \sum_{k=1}^{K} u^{(k)}_t \right\|^2 \right] \\
&\leq \frac{3 C_g^2 L_f^2 \sigma_g^2 + 3 C_g^2 \sigma_g^2 + 3 C_g^2 \sigma_f^2}{\beta_x \eta T} + 6 C_g^2 \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \left\| \hat{h}_t - \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(x^{(k)}_t) \right\|^2 \right] \\
&+ \frac{8 \gamma_2^2 (C_g^2 L_f^2 + C_g^2 \sigma_g^2) + 144 \alpha^2 \eta^2 \sigma_g^2 \sigma_f^2}{\beta_x^2} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\| \hat{u}_t \|^2] + \frac{12 \gamma_2^2 C_g}{\alpha} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\| \hat{u}_t \|^2] \\
&+ \frac{2 \beta_x \eta^2 C_g^2 \sigma_f^2}{K} + \frac{2 \beta_x \eta C_g^2 \sigma_f}{K} + \frac{12 \alpha^2 \eta C_g^2 \sigma_f^2}{K}
\end{align*}
\]

(63)

Proof.

\[
\begin{align*}
\mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_x f^{(k)}(g^{(k)}(x_t^{(k)}), y^{(k)}_t) - \frac{1}{K} \sum_{k=1}^{K} u^{(k)}_t \right\|^2 \right] \\
&= \mathbb{E} \left[ \left\| (1 - \beta_x \eta) \frac{1}{K} \sum_{k=1}^{K} \left( \nabla_x f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) - u_{t-1}^{(k)} \right) \\
+ (1 - \beta_x \eta) \frac{1}{K} \sum_{k=1}^{K} \left( \nabla_x f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) - \nabla_x f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) \right) \right) \\
+ \beta_x \eta \frac{1}{K} \sum_{k=1}^{K} \left( \nabla g^{(k)}(x_t^{(k)})^T \nabla g f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) - \nabla g^{(k)}(x_t^{(k)})^T \nabla g f^{(k)}(h_t^{(k)}, y_t^{(k)}) \right) \\
+ \nabla g^{(k)}(x_t^{(k)})^T \nabla g f^{(k)}(h_t^{(k)}, y_t^{(k)}) - \nabla g^{(k)}(x_t^{(k)}; \xi_t^{(k)})^T \nabla g f^{(k)}(h_t^{(k)}, y_t^{(k)}; \zeta_t^{(k)}) \right)^2 \right] \\
&= \mathbb{E} \left[ \left\| (1 - \beta_x \eta) \frac{1}{K} \sum_{k=1}^{K} \left( \nabla_x f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) - u_{t-1}^{(k)} \right) \\
+ (1 - \beta_x \eta) \frac{1}{K} \sum_{k=1}^{K} \left( \nabla_x f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) - \nabla_x f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) \right) \right) \\
+ \beta_x \eta \frac{1}{K} \sum_{k=1}^{K} \left( \nabla g^{(k)}(x_t^{(k)})^T \nabla g f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) - \nabla g^{(k)}(x_t^{(k)})^T \nabla g f^{(k)}(h_t^{(k)}, y_t^{(k)}) \right) \right]^2 \right] \\
&+ \beta_x \eta^2 \mathbb{E} \left[ \left\| (1 - \beta_x \eta) \frac{1}{K} \sum_{k=1}^{K} \left( \nabla g^{(k)}(x_t^{(k)})^T \nabla g f^{(k)}(h_t^{(k)}, y_t^{(k)}) - \nabla g^{(k)}(x_t^{(k)}; \xi_t^{(k)})^T \nabla g f^{(k)}(h_t^{(k)}, y_t^{(k)}; \zeta_t^{(k)}) \right) \\
+ \nabla g^{(k)}(x_t^{(k)}; \xi_t^{(k)})^T \nabla g f^{(k)}(h_t^{(k)}, y_t^{(k)}; \zeta_t^{(k)}) \right)^2 \right] \\
&\triangleq T_1 + \beta_x^2 \eta^2 T_2
\end{align*}
\]

where the second step holds due to \( \mathbb{E}[\nabla g^{(k)}(x_t^{(k)}; \xi_t^{(k)})] = \nabla g^{(k)}(x_t^{(k)}) \) and \( \mathbb{E}[f^{(k)}(h_t^{(k)}, y_t^{(k)}; \zeta_t^{(k)})] = f^{(k)}(h_t^{(k)}, y_t^{(k)}) \), \( T_1 \) denotes the first expectation and \( T_2 \) denotes
the second expectation. Then, $T_1$ can be bounded as follows:

$$
T_1 \leq (1 - \beta_x \eta)^2 (1 + a) \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_x f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} u_{t-1}^{(k)} \right\|^2 \right] + 2(1 - \beta_x \eta)^2 (1 + \frac{1}{a}) \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_x f^{(k)}(g^{(k)}(x_{t}^{(k)}), y_{t}^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla_x f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) \right\|^2 \right] + 2\beta_x^2 \eta^2 (1 + \frac{1}{a}) \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla g^{(k)}(x_{t}^{(k)}) \nabla g f^{(k)}(g^{(k)}(x_{t}^{(k)}), y_{t}^{(k)}) \right\|^2 \right] - \frac{1}{K} \sum_{k=1}^{K} \nabla g^{(k)}(x_{t}^{(k)}) \nabla g f^{(k)}(h_{t}^{(k)}, y_{t}^{(k)}) \right\|^2 \right]

\leq (1 - \beta_x \eta)^2 \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_x f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} u_{t-1}^{(k)} \right\|^2 \right] + 2 \frac{\beta_x \eta}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \nabla_x f^{(k)}(g^{(k)}(x_{t}^{(k)}), y_{t}^{(k)}) - \nabla_x f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) \right\|^2 \right] + 2 \frac{\beta_x \eta}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \nabla g^{(k)}(x_{t}^{(k)}) \nabla g f^{(k)}(g^{(k)}(x_{t}^{(k)}), y_{t}^{(k)}) - \nabla g^{(k)}(x_{t}^{(k)}) \nabla g f^{(k)}(h_{t}^{(k)}, y_{t}^{(k)}) \right\|^2 \right]

\leq (1 - \beta_x \eta)^2 \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_x f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} u_{t-1}^{(k)} \right\|^2 \right] + \frac{2}{K} \beta_x \gamma \mathbb{E} \left[ \left\| \nabla_x f^{(k)}(g^{(k)}(x_{t}^{(k)}), y_{t}^{(k)}) - \nabla_x f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) \right\|^2 \right] + \frac{2}{K} \beta_x \gamma \mathbb{E} \left[ \left\| \nabla g^{(k)}(x_{t}^{(k)}) \nabla g f^{(k)}(g^{(k)}(x_{t}^{(k)}), y_{t}^{(k)}) - \nabla g^{(k)}(x_{t}^{(k)}) \nabla g f^{(k)}(h_{t}^{(k)}, y_{t}^{(k)}) \right\|^2 \right]

\leq (1 - \beta_x \eta)^2 \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_x f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} u_{t-1}^{(k)} \right\|^2 \right] + \frac{27648\beta_x^2 \gamma^2 \eta^2 \alpha^2 p^2 \eta \gamma C_g^4 L_f^2\sigma_g^2}{\beta_x} + \frac{331776\beta_x^2 \gamma^2 \eta^2 \alpha^2 \beta_x^2 \beta_x^2 \eta \gamma C_g^4 L_f^2 \sigma_g^2}{\beta_x} + \frac{786\beta_x^2 \gamma^2 \eta^2 \alpha^2 \beta_x^2 \beta_x^2 \eta \gamma C_g^4 L_f^2 \sigma_g^2}{\beta_x} + \frac{8\gamma^2 \eta C_g^2 L_f^2}{\beta_x} \mathbb{E} \left[ \left\| \nabla g f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) \right\|^2 \right] + \frac{8\gamma^2 \eta C_g^2 L_f^2}{\beta_x} \mathbb{E} \left[ \left\| \nabla g f^{(k)}(h_{t}^{(k)}, y_{t}^{(k)}) \right\|^2 \right] + \frac{12\beta_x \gamma \eta C_g^4}{\alpha} \mathbb{E} \left[ \left\| \nabla g f^{(k)}(h_{t}^{(k)}, y_{t}^{(k)}) \right\|^2 \right] + 6\beta_x \eta C_g^2 \mathbb{E} \left[ \left\| h_{t-1} - \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(x_{t-1}^{(k)}) \right\|^2 \right] + \frac{72\gamma^2 \beta_x^2 \beta_x^2 \beta_x^2 \eta^3 C_g^3 L_f^2}{\alpha} + \frac{6\alpha^2 \eta^2 \beta_x C_g^2 \sigma_g^2}{K},

(65)
where the second step holds due to $\alpha = \frac{\beta \eta^2}{1 - \beta \eta}$ and $0 < \beta \eta < 1$, the last step holds due to Lemma 10 and the following inequality.

$$\begin{align*}
T_3 &= \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \nabla g(k) (x_t^{(k)})^T \nabla g f(k) (g(k)(x_t^{(k)}), y_t^{(k)}) \right\|^2 \right] \\
&= - \nabla g(k)(x_t^{(k)})^T \nabla g f(k) \left( \frac{1}{K} \sum_{k' = 1}^{K} g(k') (x_t^{(k')}, y_t^{(k)}) \right) \\
&+ \nabla g(k)(x_t^{(k)})^T \nabla g f(k) \left( \frac{1}{K} \sum_{k' = 1}^{K} g(k') (x_t^{(k')}, y_t^{(k)}) \right) \\
&- \nabla g(k)(x_t^{(k)})^T \nabla g f(k) \left( \frac{1}{K} \sum_{k' = 1}^{K} h_t^{(k')}, y_t^{(k)} \right) \\
&+ \nabla g(k)(x_t^{(k)})^T \nabla g f(k) \left( \frac{1}{K} \sum_{k' = 1}^{K} h_t^{(k')}, y_t^{(k)} \right) - \nabla g(k)(x_t^{(k)})^T \nabla g f(k) (h_t^{(k')}, y_t^{(k)}) \right\|^2 \right] \\
&\leq 3 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \nabla g(k)(x_t^{(k)})^T \nabla g f(k) (g(k)(x_t^{(k)}), y_t^{(k)}) \right\|^2 \right] \\
&- \nabla g(k)(x_t^{(k)})^T \nabla g f(k) \left( \frac{1}{K} \sum_{k' = 1}^{K} g(k') (x_t^{(k')}, y_t^{(k)}) \right) \\
&+ \nabla g(k)(x_t^{(k)})^T \nabla g f(k) \left( \frac{1}{K} \sum_{k' = 1}^{K} g(k') (x_t^{(k')}, y_t^{(k)}) \right) \\
&- \nabla g(k)(x_t^{(k)})^T \nabla g f(k) \left( \frac{1}{K} \sum_{k' = 1}^{K} h_t^{(k')}, y_t^{(k)} \right) \\
&+ \nabla g(k)(x_t^{(k)})^T \nabla g f(k) \left( \frac{1}{K} \sum_{k' = 1}^{K} h_t^{(k')}, y_t^{(k)} \right) - \nabla g(k)(x_t^{(k)})^T \nabla g f(k) (h_t^{(k')}, y_t^{(k)}) \right\|^2 \right] \\
&\leq 3C_g^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \nabla g f(k) (g(k)(x_t^{(k)}), y_t^{(k)}) - \nabla g f(k) \left( \frac{1}{K} \sum_{k' = 1}^{K} g(k') (x_t^{(k')}, y_t^{(k)}) \right) \right\|^2 \right] \\
&+ 3C_g^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \nabla g f(k) \left( \frac{1}{K} \sum_{k' = 1}^{K} g(k') (x_t^{(k')}, y_t^{(k)}) \right) - \nabla g f(k) \left( \frac{1}{K} \sum_{k' = 1}^{K} h_t^{(k')}, y_t^{(k)} \right) \right\|^2 \right] \\
&+ 3C_g^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \nabla g f(k) \left( \frac{1}{K} \sum_{k' = 1}^{K} h_t^{(k')}, y_t^{(k)} \right) - \nabla g f(k) (h_t^{(k')}, y_t^{(k)}) \right\|^2 \right] \\
&\leq 3C_g^2 L^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| g(k)(x_t^{(k)}) - \frac{1}{K} \sum_{k' = 1}^{K} g(k') (x_t^{(k')}) \right\|^2 \right] + 3C_g^2 \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} g(k)(x_t^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} h_t^{(k)} \right\|^2 \right] \\
&+ 3C_g^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| h_t^{(k)} - h_t^{(k')} \right\|^2 \right] \\
&\leq 72 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| h_t^{(k)} \right\|^4 \right] + 108 \alpha^2 \rho^2 \eta^2 C_4^3 \sigma^4_g + 1296 \alpha^2 \beta^2 \rho^2 \eta^2 C_4^3 C_g^2 \\
&+ 3C_g^2 \mathbb{E} \left[ \left\| h_{t-1} - \frac{1}{K} \sum_{k=1}^{K} g(k)(x_{t-1}^{(k)}) \right\|^2 \right] + \frac{6 \rho^2 \eta^2 C_4^4}{\alpha} \mathbb{E} \left[ \left\| h_{t-1} \right\|^2 \right] + \frac{36 \rho^2 \beta^2 \rho^2 \eta^2 C_4^3 C_g^2}{\alpha K} + \frac{3 \alpha \eta \rho^2 C_4^2 C_g^2}{K}.
\end{align*}
where the second to last step holds due to Lemma 8, Lemma 9, and Eq. (56), the last step holds due to $\eta < 1$. As for $T_2$, we can get

\[
T_2 \leq 2E\left[\frac{1}{K} \sum_{k=1}^{K} \left( \nabla g(x^{(k)}(x^{(k)}_t), y^{(k)}_t) \cdot \nabla f^{(k)}(h^{(k)}_t, y^{(k)}_t) - \nabla g^{(k)}(x^{(k)}_t, \xi^{(k)}_t) \cdot \nabla f^{(k)}(h^{(k)}_t, y^{(k)}_t) \right)^2 \right]
\]

\[
+ 2E\left[\frac{1}{K} \sum_{k=1}^{K} \left( \nabla g^{(k)}(x^{(k)}_t, \xi^{(k)}_t) \cdot \nabla f^{(k)}(h^{(k)}_t, y^{(k)}_t) - \nabla g^{(k)}(x^{(k)}_t, \xi^{(k)}_t) \cdot \nabla f^{(k)}(h^{(k)}_t, y^{(k)}_t) \right)^2 \right]
\]

\[
\leq \frac{2C_1^2 + 2C_2^2 \sigma_f^2}{K}
\]

where the last step holds due to Lemma 12. By combining $T_1$ and $T_2$, we can get

\[
E\left[\frac{1}{K} \sum_{k=1}^{K} \nabla f^{(k)}(g^{(k)}(x^{(k)}_t), y^{(k)}_t) - \frac{1}{K} \sum_{k=1}^{K} u^{(k)} \right]^2
\]

\[
\leq (1 - \beta_x \eta)E\left[\frac{1}{K} \sum_{k=1}^{K} \nabla f^{(k)}(g^{(k)}(x^{(k)}_t), y^{(k)}_t) - \frac{1}{K} \sum_{k=1}^{K} u^{(k)} \right]^2 + 48\beta_x \gamma^2 \eta^2 \alpha^3 (C_1^4 L_1^2 + C_2^4 L_2^2) C_3^2 C_4^2
\]

\[
+ 27648 \beta_x \gamma \eta C_3^2 C_4^2 \alpha^2 p^2 \eta^2 C_1^2 L_1^2 \epsilon_2^2
\]

\[
+ 33176 \beta_x \gamma \eta C_3^2 C_4^2 \alpha^2 p^2 \eta^2 C_1^2 L_1^2 \epsilon_2^2 + 786 \beta_x \gamma \eta C_3^2 C_4^2 \alpha \eta
\]

\[
+ 144 \gamma \eta^2 C_3^2 C_4^2 \alpha \eta^2 C_1^2 L_1^2 \epsilon_2^2 + 216 \beta_x \eta \eta C_1^2 L_1^2 \epsilon_2^2 + 2592 \alpha \gamma \eta C_3^2 C_4^2 \alpha \eta
\]

\[
+ 8 \gamma \eta C_3^2 C_4^2 \alpha \eta + 8 \gamma \eta C_3^2 C_4^2 \alpha \eta + 12 \beta_x \eta \gamma C_4^2 \alpha \eta + 6 \beta_x \eta \eta C_3^2 C_4^2 \alpha \eta
\]

\[
+ \frac{2 \beta_x \eta^2 C_3^2 C_4^2 \alpha \eta}{K} + \frac{2 \beta_x \eta^2 C_3^2 C_4^2 \alpha \eta}{K}
\]

It can be reformulated as follows:

\[
\beta_x \eta E\left[\frac{1}{K} \sum_{k=1}^{K} \nabla f^{(k)}(g^{(k)}(x^{(k)}_t), y^{(k)}_t) - \frac{1}{K} \sum_{k=1}^{K} u^{(k)} \right]^2
\]

\[
\leq E\left[\frac{1}{K} \sum_{k=1}^{K} \nabla f^{(k)}(g^{(k)}(x^{(k)}_t), y^{(k)}_t) - \frac{1}{K} \sum_{k=1}^{K} u^{(k)} \right]^2
\]

\[
- E\left[\frac{1}{K} \sum_{k=1}^{K} \nabla f^{(k)}(g^{(k)}(x^{(k)}_t), y^{(k)}_t) - \frac{1}{K} \sum_{k=1}^{K} u^{(k)} \right]^2 + 8 \gamma \eta C_3^2 C_4^2 \alpha \eta
\]

\[
+ \frac{8 \gamma \eta C_3^2 C_4^2 \alpha \eta}{K} + \frac{12 \beta_x \eta \gamma C_3^2 C_4^2 \alpha \eta}{K}
\]

\[
+ 27648 \beta_x \gamma \eta C_3^2 C_4^2 \alpha^2 p^2 \eta^2 C_1^2 L_1^2 \epsilon_2^2 + 33176 \beta_x \gamma \eta C_3^2 C_4^2 \alpha^2 p^2 \eta^2 C_1^2 L_1^2 \epsilon_2^2
\]

\[
+ 786 \beta_x \gamma \eta C_3^2 C_4^2 \alpha \eta + 144 \gamma \eta^2 C_3^2 C_4^2 \alpha \eta^2 C_1^2 L_1^2 \epsilon_2^2 + 216 \beta_x \eta \eta C_1^2 L_1^2 \epsilon_2^2
\]

\[
+ 2592 \alpha \gamma \eta C_3^2 C_4^2 \alpha \eta + 8 \gamma \eta C_3^2 C_4^2 \alpha \eta + 8 \gamma \eta C_3^2 C_4^2 \alpha \eta
\]

\[
+ \frac{2 \beta_x \eta^2 C_3^2 C_4^2 \alpha \eta}{K} + \frac{2 \beta_x \eta^2 C_3^2 C_4^2 \alpha \eta}{K}
\]

\[
(69)
\]
By summing over $t$ from 0 to $T - 1$, we can get

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \mathbf{u}_t^{(k)} \right\|^2 \right] \\
\leq \frac{1}{\beta_s \eta T} \mathbb{E}\left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_0^{(k)}), \mathbf{y}_0^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \mathbf{u}_0^{(k)} \right\|^2 \right] + 6C_g \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[ \left\| \mathbf{h}_t - \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(\mathbf{x}_t^{(k)}) \right\|^2 \right] \\
+ 8\gamma_2^2 C_1^4 L^2_{\gamma} + 2\gamma_2^2 C_1^4 L^2_{\gamma} \frac{1}{\alpha} \sum_{t=0}^{T-1} \mathbb{E}\left[ \left\| \mathbf{h}_t \right\|^2 \right] + 12\gamma_2^2 C_1^4 \frac{1}{\alpha} \sum_{t=0}^{T-1} \mathbb{E}\left[ \left\| \mathbf{h}_t \right\|^2 \right] + \frac{2\beta_s \eta C_1^2 \sigma_f^2}{K} + \frac{2\beta_s \eta C_1^2 \sigma_f^2}{K} + \frac{72\gamma_2^2 \sigma_f^2 \eta^2 C_1^2 C_f^2}{\alpha} + \frac{6\alpha^2 \eta C_1^2 \sigma_f^2}{K} \\
\leq \frac{3C_g L^2_{\gamma} \sigma_f^2 + 3C_g^2 \sigma_f^2 + 3C_g^2 \sigma_f^2}{\beta_s \eta T} + 6C_g \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[ \left\| \mathbf{h}_t - \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(\mathbf{x}_t^{(k)}) \right\|^2 \right] \\
+ 8\gamma_2^2 C_1^4 L^2_{\gamma} + 2\gamma_2^2 C_1^4 L^2_{\gamma} \frac{1}{\alpha} \sum_{t=0}^{T-1} \mathbb{E}\left[ \left\| \mathbf{h}_t \right\|^2 \right] + 12\gamma_2^2 C_1^4 \frac{1}{\alpha} \sum_{t=0}^{T-1} \mathbb{E}\left[ \left\| \mathbf{h}_t \right\|^2 \right] + \frac{2\beta_s \eta C_1^2 \sigma_f^2}{K} + \frac{2\beta_s \eta C_1^2 \sigma_f^2}{K} + \frac{72\gamma_2^2 \sigma_f^2 \eta^2 C_1^2 C_f^2}{\alpha} + \frac{6\alpha^2 \eta C_1^2 \sigma_f^2}{K},
\]

where the second step holds due to the following inequality:

\[
\mathbb{E}\left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_0^{(k)}), \mathbf{y}_0^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \mathbf{u}_0^{(k)} \right\|^2 \right] \\
= \mathbb{E}\left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_x g^{(k)}(\mathbf{x}_0^{(k)})^T \nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_0^{(k)}), \mathbf{y}_0^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla_x g^{(k)}(\mathbf{x}_0^{(k)})^T \nabla_x f^{(k)}(h_0^{(k)}, y_0^{(k)}, \xi_0^{(k)}) \right\|^2 \right] \\
= \mathbb{E}\left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_x g^{(k)}(\mathbf{x}_0^{(k)})^T \nabla_x f^{(k)}(h_0^{(k)}, y_0^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla_x g^{(k)}(\mathbf{x}_0^{(k)})^T \nabla_x f^{(k)}(h_0^{(k)}, y_0^{(k)}) \right\|^2 \right] \\
+ \frac{1}{K} \sum_{k=1}^{K} \nabla g^{(k)}(\mathbf{x}_0^{(k)})^T \nabla_x f^{(k)}(h_0^{(k)}, y_0^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla g^{(k)}(\mathbf{x}_0^{(k)})^T \nabla_x f^{(k)}(h_0^{(k)}, y_0^{(k)}) \\
+ \frac{1}{K} \sum_{k=1}^{K} \nabla g^{(k)}(\mathbf{x}_0^{(k)})^T \nabla_x f^{(k)}(h_0^{(k)}, y_0^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla g^{(k)}(\mathbf{x}_0^{(k)})^T \nabla_x f^{(k)}(h_0^{(k)}, y_0^{(k)}) \\
\leq 3C_g^2 L_{\gamma}^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\left[ \left\| g^{(k)}(\mathbf{x}_0^{(k)}) - h_0^{(k)} \right\|^2 \right] + 3C_f^2 \sigma_g^2 + 3C_f^2 \sigma_f^2 \\
\leq 3C_g^2 L_{\gamma}^2 \sigma_g^2 + 3C_f^2 \sigma_g^2 + 3C_f^2 \sigma_f^2.
\]
Lemma 14. Given Assumption \([\beta, \eta] \in (0, 1), \) and \(\eta < 1,\) we can get

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} v_t^{(k)} \right\|^2 \right] \\
\leq \frac{2L^2 \sigma_y^2 + 2\sigma_y^2}{\beta_y \eta T} + 6L_7^2 \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \frac{1}{K} \sum_{k=1}^{K} x_t^{(k)} \right\|^2 \right] \\
+ \frac{12L^2 \gamma^2 C_1^2}{\alpha T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| \frac{1}{K} \sum_{k=1}^{K} v_t^{(k)} \right\|^2 \right]
+ \frac{4 \gamma^2 L_7^2}{\beta_y \eta^2} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| v_t \right\|^2 \right] + \frac{4 \gamma^2 L_7^2}{\beta_y \eta^2} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\| v_t \right\|^2 \right]
+ \frac{144 \gamma^2 \beta_p^2 \eta^4 C_1^2 L_7^2 + 216 \alpha^2 \sigma_y^2 \eta^2 C_1^2 L_7^2 \sigma_y^2 + 2592 \alpha^2 \beta_p^2 \sigma_y^2 \eta^2 C_1^2 L_7^2 + 13824 \gamma^2 \beta_p^2 \eta^4 C_1^2 L_7^2 \sigma_y^2 + 64 \alpha^2 \sigma_y^2 \eta^2 L_7^2 C_1^2 + 24 \gamma^2 \beta_p^2 \eta^2 L_7^2 C_1^2}{\alpha}.
\]  

Proof.

\[
\mathbb{E}\left[\left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} v_t^{(k)} \right\|^2 \right] \\
= \mathbb{E}\left[\left\| (1 - \beta_y \eta) \left( \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} v_{t-1}^{(k)} \right) \\
+ (1 - \beta_y \eta) \left( \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) \right) \\
+ \beta_y \eta \left( \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(h_t^{(k)}; y_t^{(k)}) \right) \\
+ \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(h_t^{(k)}; y_t^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(h_t^{(k)}; y_t^{(k)}; \zeta_t^{(k)}) \right\|^2 \right] \\
\leq (1 - \beta_y \eta)^2 (1 + a) \mathbb{E}\left[\left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} v_{t-1}^{(k)} \right\|^2 \right] \\
+ 2(1 - \beta_y \eta)^2 (1 + a) \mathbb{E}\left[\left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) \right\|^2 \right]
+ 2 \beta_y^2 \eta^2 (1 + a) \mathbb{E}\left[\left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(h_t^{(k)}; y_t^{(k)}) \right\|^2 \right]
+ \beta_y^2 \eta^2 \mathbb{E}\left[\left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(h_t^{(k)}; y_t^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(h_t^{(k)}; y_t^{(k)}; \zeta_t^{(k)}) \right\|^2 \right]
\leq (1 - \beta_y \eta) \mathbb{E}\left[\left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} v_{t-1}^{(k)} \right\|^2 \right] \\
+ \frac{2}{\beta_y \eta} \mathbb{E}\left[\left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(g^{(k)}(x_{t-1}^{(k)}), y_{t-1}^{(k)}) \right\|^2 \right] \\
+ 2 \beta_y \eta \mathbb{E}\left[\left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(h_t^{(k)}; y_t^{(k)}) \right\|^2 \right] \\
+ \frac{2 \beta_y^2 \eta^2 \sigma_f^2}{K}.
\]  

(73)
Then, for $T_1$, we can get

$$T_1 \leq L_f^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| g^{(k)}(x_t^{(k)}) - g^{(k)}(x_{t-1}^{(k)}) \right\|^2 \right] + L_f^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| y_t^{(k)} - y_{t-1}^{(k)} \right\|^2 \right]$$

$$\leq L_f^2 C_g \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \tilde{u}_t^{(k)} - \tilde{u}_{t-1}^{(k)} \right\|^2 \right] + L_f^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| y_t^{(k)} - y_{t-1}^{(k)} \right\|^2 \right]$$

$$\leq 2\gamma^2 \eta^2 L_f^2 C_g \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \tilde{v}_t^{(k)} - \tilde{v}_{t-1}^{(k)} \right\|^2 \right] + 2\gamma^2 \eta^2 L_f^2 C_g \mathbb{E} \left[ \left\| \tilde{u}_t^{(k)} - \tilde{u}_{t-1}^{(k)} \right\|^2 \right]$$

$$+ 2\gamma^2 \eta^2 L_f^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \tilde{v}_t^{(k)} - \tilde{v}_{t-1}^{(k)} \right\|^2 \right] + 2\gamma^2 \eta^2 L_f^2 \mathbb{E} \left[ \left\| \tilde{u}_t^{(k)} - \tilde{u}_{t-1}^{(k)} \right\|^2 \right],$$

where these inequalities hold due to Assumptions [1][2].

As for $T_2$, we can get

$$T_2 = \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(h_t^{(k)}, y_t^{(k)}) \right\|^2 \right]$$

$$+ \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)} \left( \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_t^{(k')}, y_t^{(k')}) - \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_t^{(k')}, y_t^{(k)}) \right)$$

$$+ \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)} \left( \frac{1}{K} \sum_{k'=1}^{K} h_t^{(k')}, y_t^{(k')} \right) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)} \left( h_t^{(k)}, y_t^{(k)} \right) \right]$$

$$\leq 3\mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(g^{(k)}(x_t^{(k)}), y_t^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(h_t^{(k)}, y_t^{(k)}) \right\|^2 \right]$$

$$+ 3\mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)} \left( \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_t^{(k')}, y_t^{(k')}) - \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_t^{(k')}, y_t^{(k)}) \right) \right\|^2 \right]$$

$$+ 3\mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)} \left( \frac{1}{K} \sum_{k'=1}^{K} h_t^{(k')}, y_t^{(k')} \right) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f^{(k)}(h_t^{(k)}, y_t^{(k)}) \right\|^2 \right]$$

$$\leq 3L_f^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| g^{(k)}(x_t^{(k)}) - \frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(x_t^{(k')}) \right\|^2 \right] + 3L_f^2 \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(x_t^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} h_t^{(k)} \right\|^2 \right]$$

$$+ 3L_f^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} h_t^{(k)} - h_t^{(k)} \right\|^2 \right]$$

$$\leq 72\gamma^2 \eta^2 L_f^2 C_g^2 C_f^2 L_f^2 + 108\alpha^2 p^2 \eta^2 C_g^2 L_f^2 \sigma_f^2 + 1296\alpha^2 \gamma^2 \beta^2 \eta^2 C_g^2 C_f^2 L_f^2$$

$$+ 3L_f^2 \mathbb{E} \left[ \left\| \tilde{h}_{t-1}^{(k)} - \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(x_{t-1}^{(k)}) \right\|^2 \right] + \frac{6\gamma^2 \beta^2 \eta^2 C_g^2 C_f^2 L_f^2}{\alpha} \mathbb{E} \left[ \left\| \tilde{u}_{t-1}^{(k)} \right\|^2 \right]$$

$$+ \frac{3\gamma^2 \sigma_f^2 L_f^2}{K} \mathbb{E} \left[ \left\| \tilde{u}_{t-1}^{(k)} \right\|^2 \right]$$

$$\leq 72\gamma^2 \eta^2 L_f^2 C_g^2 C_f^2 L_f^2 + 108\alpha^2 p^2 \eta^2 C_g^2 L_f^2 \sigma_f^2 + 1296\alpha^2 \gamma^2 \beta^2 \eta^2 C_g^2 C_f^2 L_f^2$$

$$+ 3L_f^2 \mathbb{E} \left[ \left\| \tilde{h}_{t-1}^{(k)} - \frac{1}{K} \sum_{k=1}^{K} g^{(k)}(x_{t-1}^{(k)}) \right\|^2 \right] + \frac{6\gamma^2 \beta^2 \eta^2 C_g^2 C_f^2 L_f^2}{\alpha} \mathbb{E} \left[ \left\| \tilde{u}_{t-1}^{(k)} \right\|^2 \right]$$

$$+ \frac{3\gamma^2 \sigma_f^2 L_f^2}{K} \mathbb{E} \left[ \left\| \tilde{u}_{t-1}^{(k)} \right\|^2 \right] + \frac{3\gamma^2 \sigma_f^2 L_f^2}{K},$$

where the last step holds due to Lemmas [3][4] and Eq. [56], the last step holds due to $\eta < 1$. 

32
Then, combining $T_1, T_2$, Lemma 4 and Lemma 5, we can get

$$
\begin{align*}
& \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_f f^{(k)}(g^{(k)}(\mathbf{x}^{(k)}), y^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \mathbf{v}_t^{(k)} \right\|^2 \right] \\
& \leq (1 - \beta_t \eta) \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_f f^{(k)}(g^{(k)}(\mathbf{x}^{(k-1)}), y^{(k-1)}) - \frac{1}{K} \sum_{k=1}^{K} \mathbf{v}_{t-1}^{(k)} \right\|^2 \right] + 12 \beta_t \eta L_f^2 C^2 \mathbb{E} \left[ \left\| \mathbf{h}_{t-1} \right\|^2 \right] + 4 \gamma_t^2 \eta L_f^2 C^2 \mathbb{E} \left[ \left\| \mathbf{u}_{t-1} \right\|^2 \right] + 4 \gamma_t^2 \eta L_f^2 C^2 \mathbb{E} \left[ \left\| \mathbf{v}_{t-1} \right\|^2 \right] \\
& \quad + 14 \beta_t \eta L_f^2 C^2 + 216 \beta_t \eta L_f^2 C^2 \bar{L}_f \sigma^2 + 2592 \bar{L}_f \sigma^2 + 13824 \beta_t \eta L_f^2 C^2 \bar{L}_f \sigma^2 + 16588 \bar{\gamma}_t^2 \eta L_f^2 C^2 \bar{L}_f \sigma^2 + 384 \beta_t \eta L_f^2 C^2 \sigma^2 \\
& \quad + 72 \beta_t \eta L_f^2 C^2 \sigma^2 + 6 \beta_t \eta \sigma^2 + \frac{24 \gamma_t^2 \eta L_f^2 C^2 \sigma^2}{\beta_t}.
\end{align*}
$$

(76)

It can be reformulated as

$$
\begin{align*}
& \beta_t \eta \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_f f^{(k)}(g^{(k)}(\mathbf{x}^{(k)}), y^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \mathbf{v}_t^{(k)} \right\|^2 \right] \\
& \leq \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_f f^{(k)}(g^{(k)}(\mathbf{x}^{(k)}), y^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \mathbf{v}_t^{(k)} \right\|^2 \right] - \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_f f^{(k)}(g^{(k)}(\mathbf{x}^{(k)}), y^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \mathbf{v}_t^{(k)} \right\|^2 \right] \\
& \quad + 6 \beta_t \eta \mathbb{E} \left[ \left\| \mathbf{h}_t \right\|^2 \right] + \frac{12 \beta_t \eta L_f^2 C^2 \mathbb{E} \left[ \left\| \mathbf{u}_t \right\|^2 \right] + 4 \gamma_t^2 \eta L_f^2 C^2 \mathbb{E} \left[ \left\| \mathbf{u}_t \right\|^2 \right] + 4 \gamma_t^2 \eta L_f^2 C^2 \mathbb{E} \left[ \left\| \mathbf{v}_t \right\|^2 \right] \\
& \quad + 14 \beta_t \eta L_f^2 C^2 + 216 \beta_t \eta L_f^2 C^2 \bar{L}_f \sigma^2 + 2592 \bar{L}_f \sigma^2 + 13824 \beta_t \eta L_f^2 C^2 \bar{L}_f \sigma^2 + 16588 \bar{\gamma}_t^2 \eta L_f^2 C^2 \bar{L}_f \sigma^2 + 384 \beta_t \eta L_f^2 C^2 \sigma^2 \\
& \quad + 72 \beta_t \eta L_f^2 C^2 \sigma^2 + 6 \beta_t \eta \sigma^2 + \frac{24 \gamma_t^2 \eta L_f^2 C^2 \sigma^2}{\beta_t}.
\end{align*}
$$

(77)

By summing over $t$ from 0 to $T - 1$, we can get

$$
\begin{align*}
& \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_f f^{(k)}(g^{(k)}(\mathbf{x}^{(k)}), y^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \mathbf{v}_t^{(k)} \right\|^2 \right] \\
& \leq \frac{1}{\beta_t \eta T} \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_f f^{(k)}(g^{(k)}(\mathbf{x}_0), y_0) - \frac{1}{K} \sum_{k=1}^{K} \mathbf{v}_0^{(k)} \right\|^2 \right] + 6 L_f^2 \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \left\| \mathbf{h}_t \right\|^2 \right] + \frac{12 \beta_t \eta L_f^2 C^2 \mathbb{E} \left[ \left\| \mathbf{u}_t \right\|^2 \right] + 4 \gamma_t^2 \eta L_f^2 C^2 \mathbb{E} \left[ \left\| \mathbf{u}_t \right\|^2 \right] + 4 \gamma_t^2 \eta L_f^2 C^2 \mathbb{E} \left[ \left\| \mathbf{v}_t \right\|^2 \right] \\
& \quad + 14 \beta_t \eta L_f^2 C^2 + 216 \beta_t \eta L_f^2 C^2 \bar{L}_f \sigma^2 + 2592 \bar{L}_f \sigma^2 + 13824 \beta_t \eta L_f^2 C^2 \bar{L}_f \sigma^2 + 16588 \bar{\gamma}_t^2 \eta L_f^2 C^2 \bar{L}_f \sigma^2 + 384 \beta_t \eta L_f^2 C^2 \sigma^2 \\
& \quad + 72 \beta_t \eta L_f^2 C^2 \sigma^2 + 6 \beta_t \eta \sigma^2 + \frac{24 \gamma_t^2 \eta L_f^2 C^2 \sigma^2}{\beta_t} \\
& \quad \leq \frac{2 L_f^2 \sigma^2 + 2 \gamma_t^2 \eta}{\beta_t \eta} + 6 L_f^2 \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \left\| \mathbf{h}_t \right\|^2 \right] + \frac{12 \beta_t \eta L_f^2 C^2 \mathbb{E} \left[ \left\| \mathbf{u}_t \right\|^2 \right] + 4 \gamma_t^2 \eta L_f^2 C^2 \mathbb{E} \left[ \left\| \mathbf{u}_t \right\|^2 \right] + 4 \gamma_t^2 \eta L_f^2 C^2 \mathbb{E} \left[ \left\| \mathbf{v}_t \right\|^2 \right] \\
& \quad + 14 \beta_t \eta L_f^2 C^2 + 216 \beta_t \eta L_f^2 C^2 \bar{L}_f \sigma^2 + 2592 \bar{L}_f \sigma^2 + 13824 \beta_t \eta L_f^2 C^2 \bar{L}_f \sigma^2 + 16588 \bar{\gamma}_t^2 \eta L_f^2 C^2 \bar{L}_f \sigma^2 + 384 \beta_t \eta L_f^2 C^2 \sigma^2 \\
& \quad + 72 \beta_t \eta L_f^2 C^2 \sigma^2 + 6 \beta_t \eta \sigma^2 + \frac{24 \gamma_t^2 \eta L_f^2 C^2 \sigma^2}{\beta_t}.
\end{align*}
$$

(78)
In addition, we have

\[
\mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla y_f^{(k)}(g^{(k)}(x_0^{(k)}), y_0^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} v_0^{(k)} \right\|^2 \right] \\
eq \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla y_f^{(k)}(g^{(k)}(x_0^{(k)}), y_0^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla y_f^{(k)}(g^{(k)}(x_0^{(k)}, \xi_0^{(k)}), y_0^{(k)}, \xi_0^{(k)}) \right\|^2 \right] \\
eq \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla y_f^{(k)}(g^{(k)}(x_0^{(k)}), y_0^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla y_f^{(k)}(g^{(k)}(x_0^{(k)}, \xi_0^{(k)}), y_0^{(k)}) \right\|^2 \right] \\
+ \frac{1}{K} \sum_{k=1}^{K} \left\| \nabla y_f^{(k)}(g^{(k)}(x_0^{(k)}), \xi_0^{(k)}), y_0^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} \nabla y_f^{(k)}(g^{(k)}(x_0^{(k)}, \xi_0^{(k)}), y_0^{(k)}, \xi_0^{(k)}) \right\|^2 \right] \\
\leq 2L_f^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \left\| g^{(k)}(x_0^{(k)}) - g^{(k)}(x_0^{(k)}, \xi_0^{(k)}) \right\|^2 \right] + 2\sigma_f^2 \\
\leq 2L_f^2 \sigma_g^2 + 2\sigma_f^2.
\]

which completes the proof.

\[\square\]

**Lemma 15.** Given Assumption 4 and if \( \gamma \leq \frac{1}{5T} \) and \( \eta \leq 1 \), we have

\[
\frac{1}{T} \sum_{t=0}^{T-1} \left\| y_t - y^*(x_t) \right\|^2 \leq \frac{4}{3\gamma g \mu T} \left\| y_0 - y^*(x_0) \right\|^2 - \frac{3\gamma g}{\mu T} \sum_{t=0}^{T-1} \left\| v_t \right\|^2 + \frac{50\gamma^2 C_g^2 L_f^2}{3\gamma^2 g \mu^2} \frac{1}{T} \sum_{t=0}^{T-1} \left\| u_t \right\|^2 \\
+ \frac{100}{\mu^2} \left( 1728\gamma^2 \beta_2^2 \beta_3^2 \eta^6 C_g^2 L_f^4 \sigma_g^2 + 20736\gamma^2 \beta_2^2 \beta_3^2 \eta^6 \sigma_g^2 \right) + 48\gamma_2^3 \beta_2^2 p^4 \eta^4 L_f^2 \sigma_f^2 \\
+ 1280\gamma_2^3 \beta_2^2 p^4 \eta^4 L_f^2 \sigma_f^2.
\]

(80)
Proof. 

\[ \|\bar{y}_{t+1} - y^*(\bar{x}_{t+1})\|^2 \]

\[ \leq (1 - \frac{\eta y^H}{4}) \|\bar{y}_t - y^*(\bar{x}_t)\|^2 - \frac{3\gamma^2_y}{4} \|\bar{v}_t\|^2 + \frac{25\gamma^2_y C^2 L^2}{6\gamma y^3} \|\bar{u}_t\|^2 + \frac{25\gamma^2_y}{6\mu} \|\nabla_y f(g(\bar{x}_t), \bar{y}_t) - \bar{v}_t\|^2 \]

\[ \leq (1 - \frac{\eta y^H}{4}) \|\bar{y}_t - y^*(\bar{x}_t)\|^2 - \frac{3\gamma^2_y}{4} \|\bar{v}_t\|^2 + \frac{25\gamma^2_y C^2 L^2}{6\gamma y^3} \|\bar{u}_t\|^2 \]

\[ + \frac{25\gamma^2_y}{6\mu} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f(k)(\bar{y}_t(k)) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f(k)(\bar{x}_t(k), y_t(k)) \right\|^2 \right] \]

\[ + \frac{1}{K} \sum_{k=1}^{K} \nabla_y f(k)(\bar{x}_t(k), y_t(k)) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f(k)(\bar{x}_t(k), y_t(k)) \]

\[ + \frac{1}{K} \sum_{k=1}^{K} \nabla_y f(k)(\bar{x}_t(k), y_t(k)) - \frac{1}{K} \sum_{k=1}^{K} v_t(k) \right\|^2 \]

\[ \leq (1 - \frac{\eta y^H}{4}) \|\bar{y}_t - y^*(\bar{x}_t)\|^2 - \frac{3\gamma^2_y}{4} \|\bar{v}_t\|^2 + \frac{25\gamma^2_y C^2 L^2}{6\gamma y^3} \|\bar{u}_t\|^2 \]

\[ + \frac{25\gamma^2_y}{2\mu} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f(k)(\bar{y}_t(k)) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f(k)(\bar{x}_t(k), y_t(k)) \right\|^2 \right] \]

\[ + \frac{25\gamma^2_y}{2\mu} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f(k)(\bar{x}_t(k), y_t(k)) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f(k)(\bar{x}_t(k), y_t(k)) \right\|^2 \right] \]

\[ + \frac{25\gamma^2_y}{2\mu} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f(k)(\bar{x}_t(k), y_t(k)) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f(k)(\bar{x}_t(k), y_t(k)) \right\|^2 \right] \]

where the first step holds due to Lemma 5 in [5], the second step holds due to the homogeneous data distribution assumption, the second to last step holds due to Assumption 1 and Assumption 2; the last step holds due to Lemmas 3, 7. We further reformulate it as follows:

\[ \frac{\eta y^H}{4} \|\bar{y}_t - y^*(\bar{x}_t)\|^2 \leq \|\bar{y}_t - y^*(\bar{x}_t)\|^2 - \|\bar{y}_{t+1} - y^*(\bar{x}_{t+1})\|^2 - \frac{3\gamma^2_y}{4} \|\bar{v}_t\|^2 + \frac{25\gamma^2_y C^2 L^2}{6\gamma y^3} \|\bar{u}_t\|^2 \]

\[ + \frac{25\gamma^2_y}{2\mu} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f(k)(\bar{x}_t(k), y_t(k)) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f(k)(\bar{x}_t(k), y_t(k)) \right\|^2 \right] + \frac{25\gamma^2_y}{2\mu} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f(k)(\bar{x}_t(k), y_t(k)) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f(k)(\bar{x}_t(k), y_t(k)) \right\|^2 \right] \]

\[ + \frac{25\gamma^2_y}{2\mu} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_y f(k)(\bar{x}_t(k), y_t(k)) - \frac{1}{K} \sum_{k=1}^{K} \nabla_y f(k)(\bar{x}_t(k), y_t(k)) \right\|^2 \right] \]

By summing over \( t \) from 0 to \( T - 1 \), we can complete the proof.
Based on the aforementioned lemmas, we are ready to prove Theorem 1.

**Proof.** At first, from Lemmas [3] we can get

\[
\frac{\gamma_x \eta}{2} T \sum_{t=0}^{T-1} \mathbb{E} \| \nabla \Phi(\bar{x}_t) \|^2 \leq \mathbb{E}[\Phi(\bar{x}_0) - \Phi(\bar{x}_T)] - \frac{\gamma_x \eta}{4} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \| \bar{u}_t \|^2 + 3 \gamma_x \eta C_g^2 L_f^2 \mathbb{E} \| y_*(\bar{x}_t) - \bar{y}_t \|^2
\]

\[
+ 12 \gamma_x \eta (C_g^2 L_f^2 + C_g^2 L_g^2) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \| \bar{x}_t - \bar{x}_t^{(k)} \|^2 + 6 \gamma_x \eta \frac{C_g^2 L_f^2}{L_f} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \| \bar{y}_t - \bar{y}_t^{(k)} \|^2
\]

\[
+ 3 \gamma_x \eta \frac{1}{K} \sum_{k=1}^{K} \nabla_x f^{(k)}(g^{(k)}(\bar{x}_t^{(k)}), \bar{y}_t^{(k)}) - \bar{u}_t \| ^2 .
\]

(83)

By summing it over \( t \) from 0 to \( T - 1 \), we can get

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \| \nabla \Phi(\bar{x}_t) \|^2 \leq \frac{2(\Phi(\bar{x}_0) - \Phi(\bar{x}_T))}{\gamma_x \eta T} - \frac{1}{2T} \sum_{t=0}^{T-1} \mathbb{E} \| \bar{u}_t \|^2
\]

\[
+ 24 (C_g^2 L_f^2 + C_g^2 L_g^2) \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \| \bar{x}_t - \bar{x}_t^{(k)} \|^2 + 12 C_g^2 L_f^2 \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \| \bar{y}_t - \bar{y}_t^{(k)} \|^2
\]

\[
+ 6 \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \left\| \frac{1}{K} \sum_{k=1}^{K} \nabla_x f^{(k)}(g^{(k)}(\bar{x}_t^{(k)}), \bar{y}_t^{(k)}) - \bar{u}_t \right\|^2 \right] + 6 C_g^2 L_f^2 \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \| y_*(\bar{x}_t) - \bar{y}_t \|^2
\]

\[
\leq \frac{2(\Phi(\bar{x}_0) - \Phi(\bar{x}_T))}{\gamma_x \eta T} + 24 C_g^2 L_f^2 \frac{1}{\gamma_x \eta T} \| \bar{y}_0 - y^*(\bar{x}_0) \|^2
\]

\[
+ \left( \frac{300 C_g^2 L_f^2 (1 + 6L_f^2)}{\mu^2} + \frac{72 \gamma_x \eta C_g^4 L_f^2}{\beta_\gamma^2} + \frac{48 \gamma_x \eta C_g^4 L_f^2}{\beta_\gamma^2} + \frac{300 C_g^2 L_f^2 (1 + 6L_f^2)}{\beta_\gamma^2} + \frac{100 \gamma_x \eta C_g^4 L_f^2}{\gamma_x \eta T} \right) \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \| \bar{u}_t \|^2
\]

\[
+ \left( \frac{300 C_g^2 L_f^2 (1 + 6L_f^2)}{\mu^2} + \frac{300 C_g^2 L_f^2 (1 + 6L_f^2)}{\beta_\gamma^2} + \frac{18 \gamma_x \eta C_g^4 L_f^2}{\beta_\gamma^2} + \frac{300 C_g^2 L_f^2 (1 + 6L_f^2)}{\mu^2} \right) \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \| \bar{u}_t \|^2
\]

\[
+ \frac{600 (L_f^2 \gamma_x^2 + 1 + 6L_f^2)^2}{\alpha \mu \gamma_x \eta T} + \frac{43200 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K} + \frac{648000 \alpha^2 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K} + \frac{7766000 \alpha^2 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K}
\]

\[
+ \frac{4147200 \alpha^2 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K} + \frac{4976640 \alpha^2 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K} + \frac{115200 \alpha^2 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K}
\]

\[
+ \frac{21600 \alpha^2 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K} + \frac{1800 \alpha^2 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K} + \frac{300 \alpha \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K} + \frac{7200 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K}
\]

\[
+ \frac{144 (C_g^4 L_f^2 + C_g^2 L_g^2)^2 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4 + 14172 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4 + 497664 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K}
\]

\[
+ \frac{1152 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K} + \frac{6 (3 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4 + 3 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4 + 3 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4)}{K}
\]

\[
+ \frac{288 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4 + 288 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K}
\]

\[
+ \frac{1296 \alpha^2 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K} + \frac{15552 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K} + \frac{12 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K}
\]

\[
+ \frac{165888 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K} + \frac{1990656 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K} + \frac{4716 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K}
\]

\[
+ \frac{432 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K} + \frac{36 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K} + \frac{1800 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K}
\]

\[
+ \frac{600 \gamma_x \eta \beta_\gamma^2 \varphi^p \eta^4 C_g^8 C_f^6 L_f^4}{K} \right)
\]

(84)

where the second step holds due to Lemmas [6, 7, 13, 14, 15]. Then, we enforce the coefficient of \( \sum_{t=0}^{T-1} \mathbb{E} \| \bar{u}_t \|^2 \) to be non-positive in the following. In particular, it can be done by solving the following
As a result, by setting \( \alpha > 0 \), we can get
\[
\frac{100\gamma_2^2 C_y^4 L^2_f}{\gamma_0^2 \mu^4} - \frac{1}{2} \leq -\frac{1}{4}.
\]
Furthermore, we have the following inequalities:
\[
\frac{100\gamma_2^2 C_y^4 L^2_f}{\gamma_0^2 \mu^4} - \frac{1}{2} \leq -\frac{1}{4},
\]
\[
\frac{\gamma_2^2 600C_y^4 L^2_f (1 + 6L^2_f)}{\alpha^2 \mu^2} \leq \frac{1}{16},
\]
\[
\frac{\gamma_2^2 (3600C_y^4 L^2_f)}{\alpha \mu^2} + 72C_y^4 \leq \frac{1}{16},
\]
\[
\frac{\gamma_2^2 1200C_y^4 L^2_f}{\beta_0^2 \mu^2} \leq \frac{1}{16},
\]
\[
\frac{\gamma_2^2 48(C_y^4 L^2_f + C_y^2 L^2_f)}{x^2} \leq \frac{1}{16}.
\]
By solving these inequalities, we can get
\[
\gamma_x \leq \min \left\{ \frac{\gamma_2 \mu^2}{20C_y^2 L^2_f}, \frac{\alpha \mu}{100C_y^2 L_f \sqrt{1 + 6L^2_f}}, \frac{\sqrt{\alpha \mu}}{24C_y \sqrt{100C_y^2 L^4_f + 2C_y^2 \mu^2}}, \frac{\beta_0 \mu}{144C_y^2 L^2_f}, \frac{\beta_x}{32 \sqrt{C_y^4 L^4_f + C_y^2 L^2_f}} \right\}.
\]
Similarly, we enforce the coefficient of \( \frac{1}{\mu} \sum_{t=0}^{T-1} \mathbb{E}[\|v_t\|^2] \) to be non-positive as follows:
\[
\frac{300C_y^2 L^2_f}{\mu^2} \frac{4\gamma_2^2 C_y^2 L^2_f}{\beta_0^2} + \frac{48\gamma_2^2 C_y^2 L^2_f}{\beta_x^2} - \frac{18\gamma_2 C_y^2 L^2_f}{\mu^2} \leq 0.
\]
Then, it can be done by solving the following inequalities:
\[
\frac{1200\gamma_2 L^2_f}{\mu^2 \beta_0^2} \leq \frac{9}{\mu},
\]
\[
\frac{48\gamma_2}{\beta_x^2} \leq \frac{9}{\mu}.
\]
Therefore, we can get
\[
\gamma_y \leq \min \left\{ \frac{3\beta_0^2}{400L^2_f}, \frac{3\beta_x^2}{16\mu} \right\}.
\]
As a result, by setting \( \alpha > 0, \beta_x > 0, \beta_y > 0, \eta \leq \min \left\{ \frac{1}{2\gamma_x \gamma_y}, \frac{1}{\alpha}, \frac{1}{\beta_x}, \frac{1}{\beta_y}, \frac{1}{\eta}, 1 \right\} \), and
\[
\gamma_x \leq \min \left\{ \frac{\gamma_2 \mu^2}{20C_y^2 L^2_f}, \frac{\alpha \mu}{100C_y^2 L_f \sqrt{1 + 6L^2_f}}, \frac{\beta_0 \mu}{144C_y^2 L^2_f}, \frac{\beta_x}{32 \sqrt{C_y^4 L^4_f + C_y^2 L^2_f}}, \frac{\sqrt{\alpha \mu}}{24C_y \sqrt{100C_y^2 L^4_f + 2C_y^2 \mu^2}} \right\},
\]
\[
\gamma_y \leq \min \left\{ \frac{1}{6L^2_f}, \frac{3\beta_0^2}{400L^2_f}, \frac{3\beta_x^2}{16\mu} \right\}.
\]
we can get

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla \Phi(x_t)\|^2] \leq \frac{2(\Phi(\tilde{x}_0) - \Phi(x_T))}{\gamma_0 \eta T} + \frac{24C^2 L_f^2}{\gamma_0 \eta \mu T} \|y_0 - y^*(\tilde{x}_0)\|^2
\]

\[
+ \frac{300C_f^2 L_f^2 (1 + 6L_f^2)}{\alpha \eta T \mu^2} + \frac{300C_f^2 L_f^2 (1 + 6L_f^2) \alpha \gamma^2 }{\alpha^4 T \mu^2} + \frac{300C_f^2 L_f^2 (1 + 6L_f^2) \alpha \gamma^2 }{T \mu^2}
\]

\[
+ \frac{600(L_f^2 \sigma^2 + \sigma^2 C_f^2 L_f^2 \sigma^2_g)}{\beta \eta T \mu^2} + \frac{43200 \alpha \beta^2 \eta^2 \mu \eta^4 C_f^2 L_f^2 \sigma_g^2}{\alpha^4 T \mu^2} + \frac{64800 \alpha \beta^2 \eta^2 \mu \eta^4 C_f^2 L_f^2 \sigma_g^2}{T \mu^2} + \frac{777600 \alpha \beta^2 \eta^2 \mu \eta^4 C_f^2 L_f^2 \sigma_g^2}{\mu^2}
\]

\[
+ \frac{1417200 \alpha \beta^2 \eta^2 \mu \eta^4 C_f^2 L_f^2 \sigma_g^2}{T \mu^2} + \frac{497664 \alpha \beta^2 \eta^2 \mu \eta^4 C_f^2 L_f^2 \sigma_g^2}{T \mu^2} + \frac{115200 \alpha \beta^2 \eta^2 \mu \eta^4 C_f^2 L_f^2 \sigma_g^2}{\mu^2}
\]

\[
+ \frac{21600 \alpha \beta^2 \eta^2 \mu \eta^4 C_f^2 L_f^2 \sigma_g^2}{T \mu^2} + \frac{1800 \alpha \beta^2 \eta^2 \mu \eta^4 C_f^2 L_f^2 \sigma_g^2}{T \mu^2} + \frac{300 \beta \eta \mu \eta^4 C_f^2 L_f^2 \sigma_g^2}{\mu^2} + \frac{720 \gamma \beta^2 \eta^2 \mu \eta^4 C_f^2 L_f^2 \sigma_g^2}{\mu^2}
\]

\[
+ \frac{144(C_f^2 L_f^2 + C_f^2 L_f^2) \gamma \beta \eta \mu \eta^4 C_f^2 L_f^2 \sigma_g^2 + 41472 \gamma \beta \eta \mu \eta^4 C_f^2 L_f^2 \sigma_g^2 + 497664 \gamma \beta \eta \mu \eta^4 C_f^2 L_f^2 \sigma_g^2}{\mu^2}
\]

\[
+ \frac{1152 \gamma \beta \eta \mu \eta^4 C_f^2 L_f^2 \sigma_g^2}{\mu^2} + \frac{6(3 \beta \eta \mu \eta^4 C_f^2 L_f^2 \sigma_g^2 + 3 \beta \eta \mu \eta^4 C_f^2 L_f^2 \sigma_g^2)}{\mu^2} + \frac{288 \gamma \beta \eta \mu \eta^4 C_f^2 L_f^2 \sigma_g^2}{\mu^2}
\]

\[
+ \frac{1296 \alpha \beta \eta \mu \eta^4 C_f^2 \sigma_g^2}{\mu^2} + 15552 \alpha \beta \eta \mu \eta^4 C_f^2 \sigma_g^2 + \frac{12 \beta \eta \mu \eta^4 C_f^2 \sigma_g^2}{K} + \frac{12 \beta \eta \mu \eta^4 C_f^2 \sigma_g^2}{K}
\]

\[
+ \frac{16588 \beta \eta \mu \eta^4 C_f^2 \sigma_g^2}{\mu^2} + \frac{199065 \beta \eta \mu \eta^4 C_f^2 \sigma_g^2}{\mu^2} + \frac{4716 \beta \eta \mu \eta^4 C_f^2 \sigma_g^2}{\mu^2}
\]

\[
+ \frac{432 \beta \eta \mu \eta^4 C_f^2 \sigma_g^2}{\mu^2} + \frac{36 \beta \eta \mu \eta^4 C_f^2 \sigma_g^2}{\mu^2} + \frac{180 \beta \eta \mu \eta^4 C_f^2 \sigma_g^2}{\mu^2}
\]

\[
+ \frac{600 \mu (1728 \gamma \beta \eta \mu \eta^4 C_f^2 \sigma_g^2 + 20736 \gamma \beta \eta \mu \eta^4 C_f^2 \sigma_g^2 + 48 \gamma \beta \eta \mu \eta^4 C_f^2 \sigma_g^2)}{\mu^2}
\]

Since \( \alpha, \beta, \gamma, \) and \( \eta \) can be set as free hyperparameters, i.e., they are independent of the number of iterations, we can obtain

\[
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla \Phi(x_t)\|^2] \leq \frac{2(\Phi(\tilde{x}_0) - \Phi(x_T))}{\gamma_0 \eta T} + \frac{24C^2 L_f^2}{\gamma_0 \eta \mu T} \|y_0 - y^*(\tilde{x}_0)\|^2 + O(\frac{\eta}{K}) + O(\frac{1}{\eta T})
\]

\[
+ O(p^2 \eta^2) + O(p^4 \eta^4) + O(p^6 \eta^6) + O(p^8 \eta^8) + O(p^{10} \eta^{10})
\]

where \( x_* \) denotes the optimal solution.
### B Experimental Details

In Table 2, we summarize the hyperparameters for all methods. For a fair comparison, we use similar learning rates for all algorithms. For instance, the learning rate of LocalSGDAM and LocalSCGDAM is $\eta \gamma_x = 0.099$, which is very close to that of LocalSGDM and CoDA. In addition, the learning rate is decayed by 10 at 50% and 75% epochs for all methods. As for the number of epochs, we set it to 16 for Melanoma, 50 for FashionMNIST, and 100 for the others. Additionally, since CoDA is a stage-wise method, we use the same stage as that for learning rate decay.

**Table 2: The hyperparameters of different methods.**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Hyperparameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>LocalSGDM</td>
<td>learning rate</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>momentum coefficient</td>
<td>0.1</td>
</tr>
<tr>
<td>CoDA</td>
<td>learning rate $\eta$</td>
<td>0.3</td>
</tr>
<tr>
<td>LocalSGDAM</td>
<td>learning rate coefficient $\gamma_x$</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>momentum coefficient $\beta_x$</td>
<td>3.3</td>
</tr>
<tr>
<td>LocalSCGDAM (Ours)</td>
<td>learning rate coefficient $\eta$</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>learning rate coefficient $\gamma_x$</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>momentum coefficient $\beta_x$</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>coefficient $\alpha$</td>
<td>3.0</td>
</tr>
</tbody>
</table>

The classifier for FashionMNIST is summarized in Table 3.

**Table 3: The classifier for FashionMNIST.**

<table>
<thead>
<tr>
<th>Layers</th>
<th>Operators</th>
<th>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer 1</td>
<td>CNN</td>
<td>output channels: 32</td>
</tr>
<tr>
<td></td>
<td>Batchnorm</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ReLU</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maxpooling</td>
<td>kernel size: 2, stride: 2</td>
</tr>
<tr>
<td>Layer 2</td>
<td>CNN</td>
<td>output channels: 64</td>
</tr>
<tr>
<td></td>
<td>Batchnorm</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ReLU</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maxpooling</td>
<td>kernel size: 2, stride: 2</td>
</tr>
<tr>
<td>Layer 3</td>
<td>FC</td>
<td>output features: 600</td>
</tr>
<tr>
<td>Layer 4</td>
<td>FC</td>
<td>output features: 120</td>
</tr>
<tr>
<td>Layer 5</td>
<td>FC</td>
<td>output features: 1</td>
</tr>
</tbody>
</table>

**Table 4: Description of benchmark datasets.** Here, #pos denotes the number of positive samples, and #neg denotes the number of negative samples.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Training set</th>
<th>Testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#pos</td>
<td>#neg</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>2,777</td>
<td>25,000</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>2,777</td>
<td>25,000</td>
</tr>
<tr>
<td>STL10</td>
<td>277</td>
<td>2,500</td>
</tr>
<tr>
<td>FashionMNIST</td>
<td>3,333</td>
<td>30,000</td>
</tr>
<tr>
<td>CATvsDOG</td>
<td>1,112</td>
<td>10,016</td>
</tr>
<tr>
<td>Melanoma</td>
<td>868</td>
<td>25,670</td>
</tr>
</tbody>
</table>