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

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

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 $O(T^{1/4}/K^{3/4})$ 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) algorithms [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 [31, 41, 38, 7] tends to only focus on the minimization problem. Only two recent works [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 [19, 20, 26]. The FedAvg [19] algorithm has spawned a number of variants [25, 37, 36] designed to address the minimization problem. For instance, by maintaining a local momentum, Yu et al. [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.[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 [27]. However, neither method could achieve a linear speedup with respect to the number of devices. Recently, Sharma et al. [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. [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. [40] extended CoDA to hetereogneous data distributions and established its convergence rate. Shen et al. [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. [3] developed the *first* parallel stochastic compositional gradient descent algorithm and established its convergence rate for nonconvex problems, inspiring many federated learning methods [12, 28, 4, 27, 7] in the past few years. For instance, [12] directly used a biased stochastic gradient to do local updates, suffering from large sample and communication complexities. [4] employed the stochastic compositional gradient and the momentum technique, which can achieve much better sample and communication complexities than [12]. [7] considered the setting where the inner-level function is distributed on different devices, which shares similar sample and communication complexities as [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 [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 [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 federated 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 ([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. [35] formulated AUC maximization model as a minimax optimization problem, defined as follows:

$$\begin{aligned}\min_{\mathbf{w},\tilde{w}_1,\tilde{w}_2} \max_{\tilde{w}_3} \mathcal{L}_{AUC}(\mathbf{w},\tilde{w}_1,\tilde{w}_2,\tilde{w}_3;a,b) &\triangleq (1-p)(h(\mathbf{w};a)-\tilde{w}_1)^2\mathbb{I}_{[b=1]} - p(1-p)\tilde{w}_3^2 \\ &+ p(h(\mathbf{w};a)-\tilde{w}_2)^2\mathbb{I}_{[b=-1]} + 2(1+\tilde{w}_3)(ph(\mathbf{w};a)\mathbb{I}_{[b=-1]} - (1-p)h(\mathbf{w};a)\mathbb{I}_{[b=1]})\ ,\end{aligned} \tag{1}$$

where h denotes the classifier parameterized by $\mathbf{w} \in \mathbb{R}^d$, $\tilde{w}_1 \in \mathbb{R}, \tilde{w}_2 \in \mathbb{R}, \tilde{w}_3 \in \mathbb{R}$ 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 $\mathcal{L}_{AUC}$ loss function did not yield satisfactory performance, Yuan et al. [39] developed the compositional deep AUC maximization model, which is defined as follows:

$$\min_{\tilde{\mathbf{w}}, \tilde{w}_1, \tilde{w}_2} \max_{\tilde{w}_3} \mathcal{L}_{AUC}(\tilde{\mathbf{w}}, \tilde{w}_1, \tilde{w}_2, \tilde{w}_3; a, b) \qquad s.t. \quad \tilde{\mathbf{w}} = \mathbf{w} - \rho \nabla_{\mathbf{w}} \mathcal{L}_{CE}(\mathbf{w}; a, b) \ . \tag{2}$$

Here, $\mathcal{L}_{CE}$ denotes the cross-entropy loss function, $\mathbf{w} - \rho \nabla_{\mathbf{w}} \mathcal{L}_{CE}$ 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 $\tilde{\mathbf{w}}$, one can optimize it through optimizing the AUC loss function.

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

$$\min_{\mathbf{x} \in \mathbb{R}^{d_1}} \max_{\mathbf{y} \in \mathbb{R}^{d_2}} f(g(\mathbf{x}), \mathbf{y}) \ , \tag{3}$$

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 $\mathbf{x}$ 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:

$$\min_{\mathbf{x} \in \mathbb{R}^{d_1}} \max_{\mathbf{y} \in \mathbb{R}^{d_2}} \frac{1}{K} \sum_{k=1}^{K} f^{(k)}\Big(\frac{1}{K} \sum_{k'=1}^{K} g^{(k')}(\mathbf{x}), \mathbf{y} \Big) \ , \tag{4}$$

where K is the number of devices, $g^{(k)}(\cdot) = \mathbb{E}_{\xi \sim \mathcal{D}_g^{(k)}}[g^{(k)}(\cdot; \xi)] \in \mathbb{R}^{d_g}$ denotes the inner-level function for the data distribution $\mathcal{D}_g^{(k)}$ of the k-th device, $f^{(k)}(\cdot, \cdot) = \mathbb{E}_{\zeta \sim \mathcal{D}_f^{(k)}}[f^{(k)}(\cdot, \cdot; \zeta)]$ represents the outer-level function for the data distribution $\mathcal{D}_f^{(k)}$ 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.,*

$$\begin{aligned} \|\nabla_g f^{(k)}(\mathbf{z}_1, \mathbf{y}_1) - \nabla_g f^{(k)}(\mathbf{z}_2, \mathbf{y}_2)\|^2 \leq L_f^2 \|(\mathbf{z}_1, \mathbf{y}_1) - (\mathbf{z}_2, \mathbf{y}_2)\|^2 \ , \\ \|\nabla_{\mathbf{y}} f^{(k)}(\mathbf{z}_1, \mathbf{y}_1) - \nabla_{\mathbf{y}} f^{(k)}(\mathbf{z}_2, \mathbf{y}_2)\|^2 \leq L_f^2 \|(\mathbf{z}_1, \mathbf{y}_1) - (\mathbf{z}_2, \mathbf{y}_2)\|^2 \ , \end{aligned} \tag{5}$$

hold for $\forall (\mathbf{z}_1, \mathbf{y}_1), (\mathbf{z}_2, \mathbf{y}_2) \in \mathbb{R}^{d_g} \times \mathbb{R}^{d_2}$. The gradient of the inner-level function $g^{(k)}(\cdot)$ is L_g-Lipschitz continuous where $L_g > 0$, i.e.,

$$\|\nabla g^{(k)}(\mathbf{x}_1) - \nabla g^{(k)}(\mathbf{x}_2)\|^2 \leq L_g^2 \|\mathbf{x}_1 - \mathbf{x}_2\|^2 \ , \tag{6}$$

holds for $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^{d_1}$.

Assumption 2. *The second moment of $\nabla_g f^{(k)}(\mathbf{z}, \mathbf{y}; \zeta)$ and $\nabla g^{(k)}(\mathbf{x}; \xi)$ satisfies:*

$$\mathbb{E}[\|\nabla_g f^{(k)}(\mathbf{z}, \mathbf{y}; \zeta)\|^2] \leq C_f^2 \ , \mathbb{E}[\|\nabla g^{(k)}(\mathbf{x}; \xi)\|^2] \leq C_g^2 \ , \tag{7}$$

for $\forall (\mathbf{z}, \mathbf{y}) \in \mathbb{R}^{d_g} \times \mathbb{R}^{d_2}$ and $\forall \mathbf{x} \in \mathbb{R}^{d_1}$, 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:*

$$\mathbb{E}[\|\nabla_g f^{(k)}(\mathbf{z},\mathbf{y};\zeta) - \nabla_g f^{(k)}(\mathbf{z},\mathbf{y})\|^2] \leq \sigma_f^2, \mathbb{E}[\|\nabla_{\mathbf{y}} f^{(k)}(\mathbf{z},\mathbf{y};\zeta) - \nabla_{\mathbf{y}} f^{(k)}(\mathbf{z},\mathbf{y})\|^2] \leq \sigma_f^2 \ , \quad (8)$$

for $\forall(\mathbf{z},\mathbf{y}) \in \mathbb{R}^{d_g} \times \mathbb{R}^{d_2}$, where $\sigma_f > 0$. Additionally, the variance of the stochastic gradient and the stochastic function value of $g^{(k)}(\cdot)$ satisfies:

$$\mathbb{E}[\|\nabla g^{(k)}(\mathbf{x};\xi) - \nabla g^{(k)}(\mathbf{x})\|^2] \leq \sigma_{g'}^2 \ , \mathbb{E}[\|g^{(k)}(\mathbf{x};\xi) - g^{(k)}(\mathbf{x})\|^2] \leq \sigma_g^2 \ , \quad (9)$$

for $\forall \mathbf{x} \in \mathbb{R}^{d_1}$, where $\sigma_g > 0$ and $\sigma_{g'} > 0$.

Assumption 4. *The outer-level function $f^{(k)}(\mathbf{z},\mathbf{y})$ is μ-strongly-concave with respect to $\mathbf{y}$ for any fixed $\mathbf{z} \in \mathbb{R}^{d_g}$, where $\mu > 0$, i.e.,*

$$f^{(k)}(\mathbf{z},\mathbf{y_1}) \leq f^{(k)}(\mathbf{z},\mathbf{y_2}) + \langle \nabla_y f^{(k)}(\mathbf{z},\mathbf{y_2}), \mathbf{y}_1 - \mathbf{y}_2 \rangle - \frac{\mu}{2}\|\mathbf{y}_1 - \mathbf{y}_2\|^2 \ . \quad (10)$$

Notation: Throughout this paper, $\mathbf{a}_t^{(k)}$ denotes the variable of the k-th device in the t-th iteration and $\bar{\mathbf{a}}_t = \frac{1}{K}\sum_{k=1}^{K} \mathbf{a}_t^{(k)}$ denotes the averaged variable across all devices, where a denotes any variables used in this paper. $\mathbf{x}_*$ 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: $\mathbf{x}_0$, $\mathbf{y}_0$, $\eta \in (0,1)$, $\gamma_x > 0$, $\gamma_y > 0$, $\beta_x > 0$, $\beta_y > 0$, $\alpha > 0$, $\alpha\eta \in (0,1)$, $\beta_x\eta \in (0,1)$, $\beta_y\eta \in (0,1)$.

All workers conduct the steps below to update $\mathbf{x}$, $\mathbf{y}$. $\mathbf{x}_0^{(k)} = \mathbf{x}_0$, $\mathbf{y}_0^{(k)} = \mathbf{y}_0$,
$\mathbf{h}_0^{(k)} = g^{(k)}(\mathbf{x}_0^{(k)};\xi_0^{(k)})$,
$\mathbf{u}_0^{(k)} = \nabla g^{(k)}(\mathbf{x}_0^{(k)};\xi_0^{(k)})^T \nabla_g f^{(k)}(\mathbf{h}_0^{(k)},\mathbf{y}_0^{(k)};\zeta_0^{(k)})$, $\quad \mathbf{v}_0^{(k)} = \nabla_y f^{(k)}(\mathbf{h}_0^{(k)},\mathbf{y}_0^{(k)};\zeta_0^{(k)})$,

for $t = 0, \cdots, T-1$ **do**
 Update $\mathbf{x}$ and $\mathbf{y}$:
 $\mathbf{x}_{t+1}^{(k)} = \mathbf{x}_t^{(k)} - \gamma_x\eta\mathbf{u}_t^{(k)}$, $\quad \mathbf{y}_{t+1}^{(k)} = \mathbf{y}_t^{(k)} + \gamma_y\eta\mathbf{v}_t^{(k)}$,
 Estimate the inner-level function:
 $\mathbf{h}_{t+1}^{(k)} = (1-\alpha\eta)\mathbf{h}_t^{(k)} + \alpha\eta g^{(k)}(\mathbf{x}_{t+1}^{(k)};\xi_{t+1}^{(k)})$,
 Update momentum:
 $\mathbf{u}_{t+1}^{(k)} = (1-\beta_x\eta)\mathbf{u}_t^{(k)} + \beta_x\eta\nabla g^{(k)}(\mathbf{x}_{t+1}^{(k)};\xi_{t+1}^{(k)})^T \nabla_g f^{(k)}(\mathbf{h}_{t+1}^{(k)},\mathbf{y}_{t+1}^{(k)};\zeta_{t+1}^{(k)})$,
 $\mathbf{v}_{t+1}^{(k)} = (1-\beta_y\eta)\mathbf{v}_t^{(k)} + \beta_y\eta\nabla_y f^{(k)}(\mathbf{h}_{t+1}^{(k)},\mathbf{y}_{t+1}^{(k)};\zeta_{t+1}^{(k)})$,
 if mod$(t+1,p) == 0$ **then**
  $\mathbf{h}_{t+1}^{(k)} = \bar{\mathbf{h}}_{t+1} \triangleq \frac{1}{K}\sum_{k'=1}^{K}\mathbf{h}_{t+1}^{(k')}$,
  $\mathbf{u}_{t+1}^{(k)} = \bar{\mathbf{u}}_{t+1} \triangleq \frac{1}{K}\sum_{k'=1}^{K}\mathbf{u}_{t+1}^{(k')}$, $\quad \mathbf{v}_{t+1}^{(k)} = \bar{\mathbf{v}}_{t+1} \triangleq \frac{1}{K}\sum_{k'=1}^{K}\mathbf{v}_{t+1}^{(k')}$,
  $\mathbf{x}_{t+1}^{(k)} = \bar{\mathbf{x}}_{t+1} \triangleq \frac{1}{K}\sum_{k'=1}^{K}\mathbf{x}_{t+1}^{(k')}$, $\quad \mathbf{y}_{t+1}^{(k)} = \bar{\mathbf{y}}_{t+1} \triangleq \frac{1}{K}\sum_{k'=1}^{K}\mathbf{y}_{t+1}^{(k')}$,
 end if
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 $\mathbf{x}_t^{(k)}$ and $\mathbf{y}_t^{(k)}$ on the k-th device. There exists an 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.

In detail, since the objective function in Eq. (4) is a compositional function whose stochastic gradient regarding $\mathbf{x}$, i.e., $\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T \nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)}), \mathbf{y}_t^{(k)};\zeta_t^{(k)})$, is a biased estimation for the full gradient, we employ the stochastic compositional gradient $\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T \nabla_g f^{(k)}(\mathbf{h}_t^{(k)}, \mathbf{y}_t^{(k)};\zeta_t^{(k)})$ to update the model parameter $\mathbf{x}$, where $\mathbf{h}_t^{(k)}$ is the moving-average estimation of the inner-level function $g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})$ on the k-th device, which is defined as follows:

$$\mathbf{h}_t^{(k)} = (1-\alpha\eta)\mathbf{h}_{t-1}^{(k)} + \alpha\eta g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)}) \ , \tag{11}$$

where $\alpha > 0$ and $\eta > 0$ are two hyperparameters, and $\alpha\eta \in (0, 1)$. The objective function in Eq. (4) is not compositional regarding $\mathbf{y}$, thus we can directly leverage its stochastic gradient to perform an update. Then, based on the obtained stochastic (compositional) gradient, we compute the momentum as follows:

$$\begin{aligned}
\mathbf{u}_t^{(k)} &= (1-\beta_x\eta)\mathbf{u}_{t-1}^{(k)} + \beta_x\eta \nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T \nabla_g f^{(k)}(\mathbf{h}_t^{(k)}, \mathbf{y}_t^{(k)};\zeta_t^{(k)}) \ , \\
\mathbf{v}_t^{(k)} &= (1-\beta_y\eta)\mathbf{v}_{t-1}^{(k)} + \beta_y\eta \nabla_y f^{(k)}(\mathbf{h}_t^{(k)}, \mathbf{y}_t^{(k)};\zeta_t^{(k)}) \ ,
\end{aligned} \tag{12}$$

where $\beta_x > 0$ and $\beta_y > 0$ are two hyperparameters, $\beta_x\eta \in (0, 1)$, and $\beta_y\eta \in (0, 1)$. Based on the obtained momentum, each device updates its local model parameters as follows:

$$\mathbf{x}_{t+1}^{(k)} = \mathbf{x}_t^{(k)} - \gamma_x\eta\mathbf{u}_t^{(k)} \ , \mathbf{y}_{t+1}^{(k)} = \mathbf{y}_t^{(k)} + \gamma_y\eta\mathbf{v}_t^{(k)} \ , \tag{13}$$

where $\gamma_x > 0$ and $\gamma_y > 0$.

As we mentioned before, to obtain the global inner-level function, our algorithm periodically communicates the moving-average estimation of the inner-level function, i.e., $\mathbf{h}_{t+1}^{(k)}$. In particular, at every p iterations, i.e., $\text{mod}(t+1, p) == 0$ where $p > 1$ is the *communication period*, each device uploads $\mathbf{h}_{t+1}^{(k)}$ to the central server and the central server computes the average of all received variables, which will be further broadcast to all devices as follows:

$$\mathbf{h}_{t+1}^{(k)} = \bar{\mathbf{h}}_{t+1} \triangleq \frac{1}{K}\sum_{k'=1}^{K} \mathbf{h}_{t+1}^{(k')} \ . \tag{14}$$

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 [23, 36] 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:

$$\mathbf{y}_*(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathbb{R}^{d_2}} \frac{1}{K}\sum_{k=1}^{K} f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}), \mathbf{y}) \ , \quad \Phi(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}), \mathbf{y}_*(\mathbf{x})) \ . \tag{15}$$

Then, based on Assumptions 1-4, we can obtain that $\Phi^{(k)}$ is L_Φ-smooth, where $L_\Phi = \frac{2C_g^2L_f^2}{\mu} + C_fL_g$. 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 1-4, by setting* $\alpha > 0$, $\beta_x > 0$, $\beta_y > 0$, $\eta \leq \min\{\frac{1}{2\gamma_x L_\phi}, \frac{1}{10p^2\gamma_y L_f}, \frac{1}{\alpha}, \frac{1}{\beta_x}, \frac{1}{\beta_y}, 1\}$, $\gamma_y \leq \min\left\{\frac{1}{6L_f}, \frac{3\mu\beta_y^2}{400L_f^2}, \frac{3\beta_x^2}{16\mu}\right\}$, *and* $\gamma_x \leq \min\left\{\frac{\alpha\mu}{100C_g^2L_f\sqrt{1+6L_f^2}}, \frac{\beta_x}{32\sqrt{C_g^4L_f^2+C_f^2L_g^2}}, \frac{\beta_y\mu}{144C_g^2L_f^2}, \frac{\sqrt{\alpha}\mu}{24C_g\sqrt{100C_g^2L_f^4+2C_g^2\mu^2}}, \frac{\gamma_y\mu^2}{20C_g^2L_f^2}\right\}$, *Algorithm 1 has the following convergence rate*

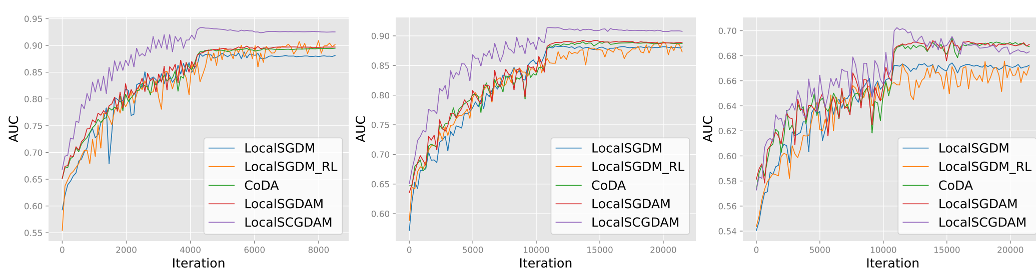
$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\bar{\mathbf{x}}_t)\|^2] \leq \frac{2(\Phi(\mathbf{x}_0)-\Phi(\mathbf{x}_*)}{\gamma_x\eta T} + \frac{24C_g^2L_f^2}{\gamma_y\eta\mu T}\|\mathbf{y}_0-\mathbf{y}^*(\mathbf{x}_0)\|^2 + O(\frac{\eta}{K}) + O(\frac{1}{\eta T})$$
$$+ O(p^2\eta^2) + O(p^4\eta^4) + O(p^6\eta^6) + O(p^8\eta^8) + O(p^{10}\eta^{10})\ . \tag{16}$$

Remark 1. *In terms of Theorem 1, 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 $O(1/\sqrt{KT})$ 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 ϵ-accuracy solution, i.e., $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\bar{\mathbf{x}}_t)\|^2] \leq \epsilon^2$, by setting $\eta = O(K\epsilon^2)$ and $p = O(1/(K\epsilon))$, then the sample complexity on each device is $O(1/(K\epsilon^4))$ and the communication complexity is $O(1/\epsilon^3)$.*

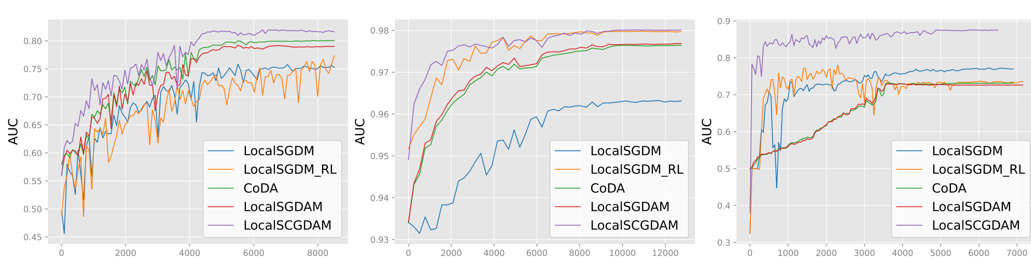
Challenges. The compositional structure in the loss function, especially when the inner-level functions are distributed on different devices, makes the convergence analysis challenging. In fact, the existing federated compositional *minimization* algorithms [4, 27, 12, 28] only consider a much simpler 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 numbe 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.

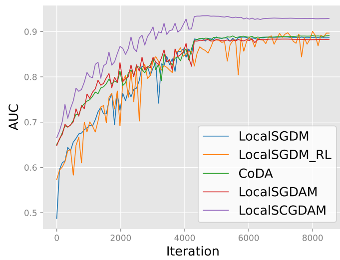


(a) CATvsDOG (b) CIFAR10 (c) CIFAR100

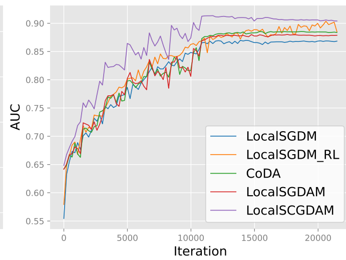


(d) STL10 (e) FashionMNIST (f) Melanoma

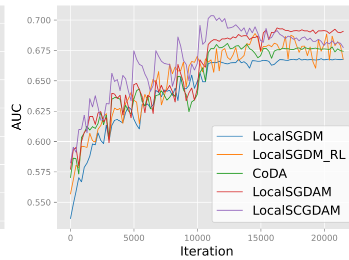
Figure 1: Testing performance with AUC score versus the number of iterations when the communication period $p = 4$.



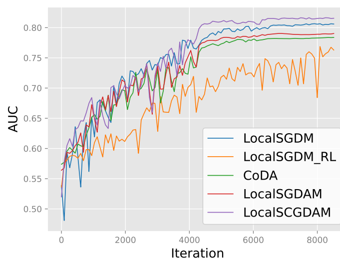
(a) CATvsDOG



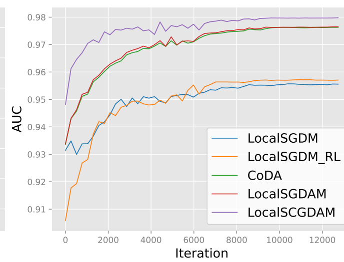
(b) CIFAR10



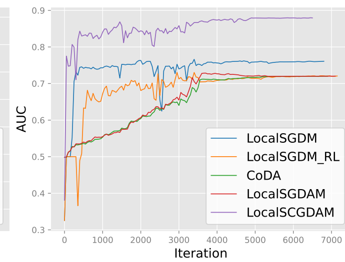
(c) CIFAR100



(d) STL10

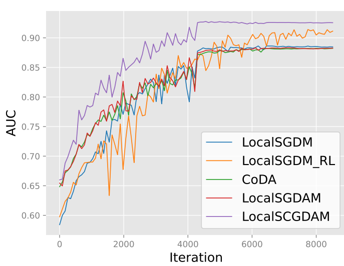


(e) FashionMNIST

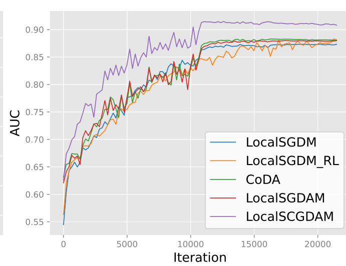


(f) Melanoma

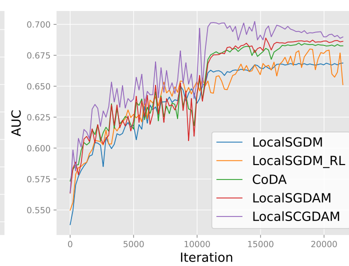
Figure 2: Testing performance with AUC score versus the number of iterations when the communication period $p = 8$.



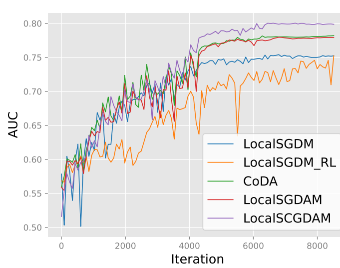
(a) CATvsDOG



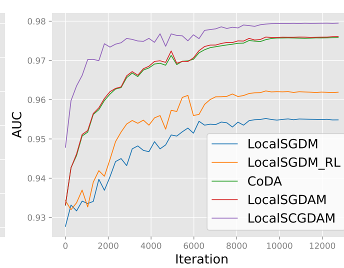
(b) CIFAR10



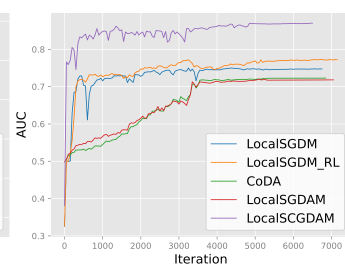
(c) CIFAR100



(d) STL10



(e) FashionMNIST



(f) Melanoma

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 [2], 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

[2] https://www.kaggle.com/c/dogs-vs-cats

Table 1: The comparison between the test AUC score of different methods on all datasets. Here, p denotes the communication period.

Datasets	Methods	AUC $p=4$	AUC $p=8$	AUC $p=16$
CATvsDOG	**LocalSCGDAM**	**0.933±0.000**	**0.936±0.000**	**0.928±0.000**
	CoDA	0.895±0.000	0.892±0.000	0.883±0.001
	LocalSGDAM	0.899±0.000	0.884±0.000	0.884±0.001
	LocalSGDM	0.888±0.001	0.889±0.000	0.887±0.000
	LocalSGDM_RL	0.909±0.000	0.901±0.001	0.917±0.001
CIFAR10	**LocalSCGDAM**	**0.914±0.000**	**0.914±0.000**	**0.916±0.000**
	CoDA	0.890±0.000	0.886±0.000	0.883±0.000
	LocalSGDAM	0.893±0.000	0.880±0.000	0.880±0.000
	LocalSGDM	0.883±0.001	0.871±0.000	0.874±0.000
	LocalSGDM_RL	0.890±0.000	0.904±0.001	0.883±0.001
CIFAR100	**LocalSCGDAM**	**0.702±0.000**	**0.704±0.001**	**0.703±0.001**
	CoDA	0.694±0.001	0.681±0.001	0.685±0.000
	LocalSGDAM	0.692±0.000	0.694±0.000	0.689±0.001
	LocalSGDM	0.675±0.001	0.669±0.000	0.669±0.000
	LocalSGDM_RL	0.676±0.000	0.690±0.001	0.682±0.001
STL10	**LocalSCGDAM**	**0.820±0.001**	**0.817±0.000**	**0.801±0.000**
	CoDA	0.801±0.000	0.784±0.000	0.783±0.000
	LocalSGDAM	0.792±0.000	0.790±0.000	0.780±0.000
	LocalSGDM	0.760±0.001	0.808±0.000	0.757±0.001
	LocalSGDM_RL	0.773±0.001	0.771±0.003	0.752±0.003
FashionMNIST	**LocalSCGDAM**	**0.980±0.000**	**0.980±0.000**	**0.980±0.000**
	CoDA	0.976±0.000	0.976±0.000	0.976±0.000
	LocalSGDAM	0.977±0.000	0.977±0.000	0.976±0.000
	LocalSGDM	0.963±0.000	0.956±0.000	0.955±0.000
	LocalSGDM_RL	0.980±0.000	0.957±0.000	0.962±0.000
Melanoma	**LocalSCGDAM**	**0.876±0.000**	**0.880±0.000**	**0.870±0.000**
	CoDA	0.734±0.002	0.721±0.000	0.725±0.003
	LocalSGDAM	0.730±0.000	0.729±0.000	0.721±0.003
	LocalSGDM	0.774±0.001	0.766±0.001	0.750±0.000
	LocalSGDM_RL	0.737±0.001	0.721±0.000	0.773±0.001

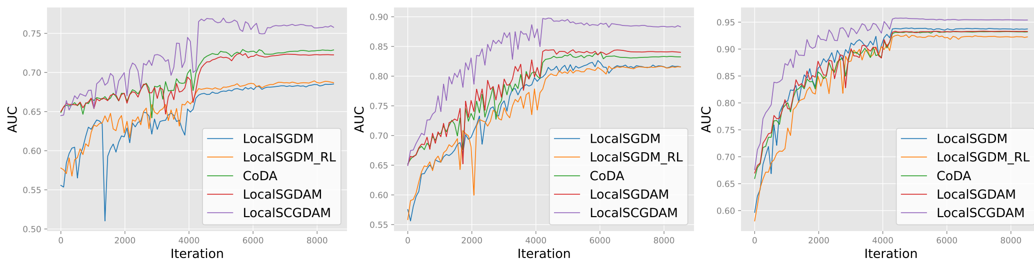
strategy to construct the imbalanced training data. For these synthetic imbalanced datasets, the testing set is balanced. Melanoma is an intrinsically imbalanced medical image classification dataset, which we do not modify. The details about these benchmark datasets are summarized in Table 4.

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: LocalSGDM [36], LocalSGDM_RL [30], CoDA [8], LocalSGDAM [23]. Specifically, LocalSGDM uses momentum SGD to optimize the standard cross-entropy loss function. LocalSGDM_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.



(a) Imbalance Ratio:0.01 (b) Imbalance Ratio:0.05 (c) Imbalance Ratio:0.2

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.

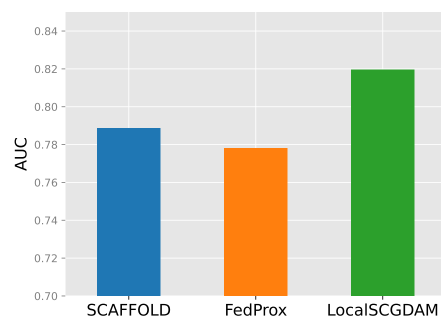


Figure 5: The test AUROC score for STL10.

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

[1] A. Coates, A. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[2] Y. Deng and M. Mahdavi. Local stochastic gradient descent ascent: Convergence analysis and communication efficiency. In *International Conference on Artificial Intelligence and Statistics*, pages 1387–1395. PMLR, 2021.

[3] H. Gao and H. Huang. Fast training method for stochastic compositional optimization problems. *Advances in Neural Information Processing Systems*, 34:25334–25345, 2021.

[4] H. Gao, J. Li, and H. Huang. On the convergence of local stochastic compositional gradient descent with momentum. In *International Conference on Machine Learning*, pages 7017–7035. PMLR, 2022.

[5] H. Gao, X. Wang, L. Luo, and X. Shi. On the convergence of stochastic compositional gradient descent ascent method. In *Thirtieth International Joint Conference on Artificial Intelligence*, 2021.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[7] Z. Guo, R. Jin, J. Luo, and T. Yang. Fedxl: Provable federated learning for deep x-risk optimization. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 11934–11966. PMLR, 2023.

[8] Z. Guo, M. Liu, Z. Yuan, L. Shen, W. Liu, and T. Yang. Communication-efficient distributed stochastic auc maximization with deep neural networks. In *International Conference on Machine Learning*, pages 3864–3874. PMLR, 2020.

[9] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] A. Herschtal and B. Raskutti. Optimising area under the roc curve using gradient descent. In *Proceedings of the twenty-first international conference on Machine learning*, page 49, 2004.

[12] F. Huang, J. Li, and H. Huang. Compositional federated learning: Applications in distributionally robust averaging and meta learning. *arXiv preprint arXiv:2106.11264*, 2021.

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

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

[15] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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

[17] T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020.

[18] M. Liu, Z. Yuan, Y. Ying, and T. Yang. Stochastic auc maximization with deep neural networks. *arXiv preprint arXiv:1908.10831*, 2019.

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

[20] D. Povey, X. Zhang, and S. Khudanpur. Parallel training of dnns with natural gradient and parameter averaging. *arXiv preprint arXiv:1410.7455*, 2014.

[21] S. Qiu, Z. Yang, X. Wei, J. Ye, and Z. Wang. Single-timescale stochastic nonconvex-concave optimization for smooth nonlinear td learning. *arXiv preprint arXiv:2008.10103*, 2020.

[22] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):1–8, 2021.

[23] P. Sharma, R. Panda, G. Joshi, and P. Varshney. Federated minimax optimization: Improved convergence analyses and algorithms. In *International Conference on Machine Learning*, pages 19683–19730. PMLR, 2022.

[24] Z. Shen, J. Cervino, H. Hassani, and A. Ribeiro. An agnostic approach to federated learning with class imbalance. In *International Conference on Learning Representations*, 2021.

[25] S. U. Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

[26] H. Su and H. Chen. Experiments on parallel training of deep neural network using model averaging. *arXiv preprint arXiv:1507.01239*, 2015.

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

[28] B. Wang, Z. Yuan, Y. Ying, and T. Yang. Memory-based optimization methods for model-agnostic meta-learning. *arXiv preprint arXiv:2106.04911*, 2021.

[29] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

[30] L. Wang, S. Xu, X. Wang, and Q. Zhu. Addressing class imbalance in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10165–10173, 2021.

[31] M. Wang, E. X. Fang, and H. Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1):419–449, 2017.

[32] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[33] Y. Yan, Y. Xu, Q. Lin, W. Liu, and T. Yang. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. *Advances in Neural Information Processing Systems*, 33:5789–5800, 2020.

[34] T. Yang. Trading computation for communication: Distributed stochastic dual coordinate ascent. *Advances in neural information processing systems*, 26, 2013.

[35] Y. Ying, L. Wen, and S. Lyu. Stochastic online auc maximization. *Advances in neural information processing systems*, 29, 2016.

[36] H. Yu, R. Jin, and S. Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pages 7184–7193. PMLR, 2019.

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

[38] H. Yuan, X. Lian, and J. Liu. Stochastic recursive variance reduction for efficient smooth non-convex compositional optimization. *arXiv preprint arXiv:1912.13515*, 2019.

[39] Z. Yuan, Z. Guo, N. Chawla, and T. Yang. Compositional training for end-to-end deep auc maximization. In *International Conference on Learning Representations*, 2021.

[40] Z. Yuan, Z. Guo, Y. Xu, Y. Ying, and T. Yang. Federated deep auc maximization for hetergeneous data with a constant communication complexity. In *International Conference on Machine Learning*, pages 12219–12229. PMLR, 2021.

[41] J. Zhang and L. Xiao. A composite randomized incremental gradient method. In *International Conference on Machine Learning*, pages 7454–7462, 2019.

[42] J. Zhang, P. Xiao, R. Sun, and Z. Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in Neural Information Processing Systems*, 33:7377–7389, 2020.

A Proof

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

$$
\begin{aligned}
&\mathbf{y}_*^{(k)}(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathbb{R}^{d_2}} f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}), \mathbf{y}) \;, \mathbf{y}_*(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathbb{R}^{d_2}} \frac{1}{K}\sum_{k=1}^{K} f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}), \mathbf{y}) \;, \\
&\Phi^{(k)}(\mathbf{x}) = f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}), \mathbf{y}_*(\mathbf{x})) \;, \Phi(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}), \mathbf{y}_*(\mathbf{x})) \;, \\
&g(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} g^{(k)}(\mathbf{x}) \;.
\end{aligned}
\tag{17}
$$

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

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

Proof. Since $\mathbf{y}_*^{(k)}(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathbb{R}^{d_2}} f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}), \mathbf{y})$, according to the optimality condition, for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^{d_1}$, we can get

$$
\begin{aligned}
&\langle \nabla_{\mathbf{y}} f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_1), \mathbf{y}_*^{(k)}(\mathbf{x}_1)), \mathbf{y}_*^{(k)}(\mathbf{x}_2) - \mathbf{y}_*^{(k)}(\mathbf{x}_1)\rangle \leq 0 \;, \\
&\langle \nabla_{\mathbf{y}} f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_2), \mathbf{y}_*^{(k)}(\mathbf{x}_2)), \mathbf{y}_*^{(k)}(\mathbf{x}_1) - \mathbf{y}_*^{(k)}(\mathbf{x}_2)\rangle \leq 0 \;,
\end{aligned}
\tag{18}
$$

As a result, we can get

$$
\langle \mathbf{y}_*^{(k)}(\mathbf{x}_2) - \mathbf{y}_*^{(k)}(\mathbf{x}_1), \nabla_{\mathbf{y}} f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_1), \mathbf{y}_*^{(k)}(\mathbf{x}_1)) - \nabla_{\mathbf{y}} f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_2), \mathbf{y}_*^{(k)}(\mathbf{x}_2))\rangle \leq 0 \;.
\tag{19}
$$

Meanwhile, due to the strong monotonicity of the gradient with respect to $\mathbf{y}$, we can get

$$
\begin{aligned}
&\langle \mathbf{y}_*^{(k)}(\mathbf{x}_2) - \mathbf{y}_*^{(k)}(\mathbf{x}_1), \nabla_{\mathbf{y}} f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_1), \mathbf{y}_*^{(k)}(\mathbf{x}_2)) - \nabla_{\mathbf{y}} f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_1), \mathbf{y}_*^{(k)}(\mathbf{x}_1))\rangle \\
&\quad + \mu\|\mathbf{y}_*^{(k)}(\mathbf{x}_2) - \mathbf{y}_*^{(k)}(\mathbf{x}_1)\|^2 \leq 0 \;.
\end{aligned}
\tag{20}
$$

By adding above two inequalities together, we can get

$$
\begin{aligned}
&\mu\|\mathbf{y}_*^{(k)}(\mathbf{x}_2) - \mathbf{y}_*^{(k)}(\mathbf{x}_1)\|^2 \\
&\leq \langle \mathbf{y}_*^{(k)}(\mathbf{x}_2) - \mathbf{y}_*^{(k)}(\mathbf{x}_1), \nabla_{\mathbf{y}} f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_2), \mathbf{y}_*^{(k)}(\mathbf{x}_2)) - \nabla_{\mathbf{y}} f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_1), \mathbf{y}_*^{(k)}(\mathbf{x}_2))\rangle \\
&\leq \|\mathbf{y}_*^{(k)}(\mathbf{x}_2) - \mathbf{y}_*^{(k)}(\mathbf{x}_1)\|\|\nabla_{\mathbf{y}} f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_2), \mathbf{y}_*^{(k)}(\mathbf{x}_2)) - \nabla_{\mathbf{y}} f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_1), \mathbf{y}_*^{(k)}(\mathbf{x}_2))\| \\
&\leq L_f\|\mathbf{y}_*^{(k)}(\mathbf{x}_2) - \mathbf{y}_*^{(k)}(\mathbf{x}_1)\|\|\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_2) - \frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_1)\| \\
&\leq C_g L_f\|\mathbf{y}_*^{(k)}(\mathbf{x}_2) - \mathbf{y}_*^{(k)}(\mathbf{x}_1)\|\|\mathbf{x}_2 - \mathbf{x}_1\| \;.
\end{aligned}
\tag{21}
$$

where the third step holds due to Assumption 1 and the last step holds due to Assumption 2. Therefore, we can get

$$
\|\mathbf{y}_*^{(k)}(\mathbf{x}_2) - \mathbf{y}_*^{(k)}(\mathbf{x}_1)\| \leq \frac{C_g L_f}{\mu}\|\mathbf{x}_2 - \mathbf{x}_1\| \;.
\tag{22}
$$

$\square$

Lemma 2. *Given Assumptions 1-4, $\Phi^{(k)}$ is L_Φ-smooth, where $L_\Phi = \frac{2C_g^2 L_f^2}{\mu} + C_f L_g$.*

Proof. Since $\Phi^{(k)}(\mathbf{x}) = f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}), \mathbf{y}_*(\mathbf{x}))$, we can get

$$
\begin{aligned}
&\left\|\nabla\Phi^{(k)}(\mathbf{x}_1) - \nabla\Phi^{(k)}(\mathbf{x}_2)\right\| \\
&= \Big\|\Big(\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\mathbf{x}_1)\Big)^T \nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_1), \mathbf{y}_*(\mathbf{x}_1)) \\
&\quad - \Big(\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\mathbf{x}_1)\Big)^T \nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_2), \mathbf{y}_*(\mathbf{x}_2)) \\
&\quad + \Big(\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\mathbf{x}_1)\Big)^T \nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_2), \mathbf{y}_*(\mathbf{x}_2)) \\
&\quad - \Big(\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\mathbf{x}_2)\Big)^T \nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_2), \mathbf{y}_*(\mathbf{x}_2))\Big\| \\
&\leq \Big\|\Big(\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\mathbf{x}_1)\Big)^T \nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_1), \mathbf{y}_*(\mathbf{x}_1)) \\
&\quad - \Big(\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\mathbf{x}_1)\Big)^T \nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_2), \mathbf{y}_*(\mathbf{x}_2))\Big\| \\
&\quad + \Big\|\Big(\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\mathbf{x}_1)\Big)^T \nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_2), \mathbf{y}_*(\mathbf{x}_2)) \\
&\quad - \Big(\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\mathbf{x}_2)\Big)^T \nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_2), \mathbf{y}_*(\mathbf{x}_2))\Big\| \\
&\leq C_g\Big\|\nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_1), \mathbf{y}_*(\mathbf{x}_1)) - \nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_2), \mathbf{y}_*(\mathbf{x}_2))\Big\| \\
&\quad + C_f\Big\|\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\mathbf{x}_1) - \frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\mathbf{x}_2)\Big\| \\
&\leq C_g L_f\Big\|\Big(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_1), \mathbf{y}_*(\mathbf{x}_1)\Big) - \Big(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_2), \mathbf{y}_*(\mathbf{x}_2)\Big)\Big\| + C_f L_g\|\mathbf{x}_1 - \mathbf{x}_2\| \\
&\leq C_g L_f(C_g\|\mathbf{x}_1 - \mathbf{x}_2\| + \frac{C_g L_f}{\mu}\|\mathbf{x}_1 - \mathbf{x}_2\|) + C_f L_g\|\mathbf{x}_1 - \mathbf{x}_2\| \\
&\leq \Big(\frac{2C_g^2 L_f^2}{\mu} + C_f L_g\Big)\|\mathbf{x}_1 - \mathbf{x}_2\| ,
\end{aligned} \tag{23}
$$

where the second inequality holds due to Assumption 2, the third inequality holds due to Assumption 1, the fourth inequality holds due to Assumption 2 and Lemma 1, the last step holds due to $L_f/\mu > 1$. $\square$

Lemma 3. *Given Assumption 1-4 and $\eta \leq \frac{1}{2\gamma_x L_\phi}$, we can get*

$$
\begin{aligned}
&\mathbb{E}[\Phi(\bar{\mathbf{x}}_{t+1})] \leq \mathbb{E}[\Phi(\bar{\mathbf{x}}_t)] - \frac{\gamma_x\eta}{2}\mathbb{E}[\|\nabla\Phi(\bar{\mathbf{x}}_t)\|^2] - \frac{\gamma_x\eta}{4}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2] + 3\gamma_x\eta C_g^2 L_f^2\mathbb{E}[\|\mathbf{y}_*(\bar{\mathbf{x}}_t) - \bar{\mathbf{y}}_t\|^2] \\
&\quad + 6\gamma_x\eta(C_g^4 L_f^2 + C_f^2 L_g^2)\frac{1}{K}\sum_{k=1}^{K}\Big[\Big\|\bar{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\Big\|^2\Big] + 3\gamma_x\eta C_g^2 L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\bar{\mathbf{y}}_t - \mathbf{y}_t^{(k)}\Big\|^2\Big] \\
&\quad + 3\gamma_x\eta\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{x}} f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)}) - \bar{\mathbf{u}}_t\Big\|^2\Big] .
\end{aligned} \tag{24}
$$

Proof. Because $\Phi(\bar{\mathbf{x}}_t)$ is L_ϕ-smooth, we can get

$$
\begin{aligned}
&\mathbb{E}[\Phi(\bar{\mathbf{x}}_{t+1})] \leq \mathbb{E}[\Phi(\bar{\mathbf{x}}_t)] + \mathbb{E}[\langle \nabla\Phi(\bar{\mathbf{x}}_t), \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t\rangle] + \frac{L_\Phi}{2}\mathbb{E}[\|\bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t\|^2] \\
&= \mathbb{E}[\Phi(\bar{\mathbf{x}}_t)] - \gamma_x\eta\mathbb{E}[\langle \nabla\Phi(\bar{\mathbf{x}}_t), \bar{\mathbf{u}}_t\rangle] + \frac{\gamma_x^2\eta^2 L_\Phi}{2}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2] \\
&= \mathbb{E}[\Phi(\bar{\mathbf{x}}_t)] - \frac{\gamma_x\eta}{2}\mathbb{E}[\|\nabla\Phi(\bar{\mathbf{x}}_t)\|^2] + \Big(\frac{\gamma_x^2\eta^2 L_\Phi}{2} - \frac{\gamma_x\eta}{2}\Big)\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2] + \frac{\gamma_x\eta}{2}\mathbb{E}[\|\nabla\Phi(\bar{\mathbf{x}}_t) - \bar{\mathbf{u}}_t\|^2] \\
&\leq \mathbb{E}[\Phi(\bar{\mathbf{x}}_t)] - \frac{\gamma_x\eta}{2}\mathbb{E}[\|\nabla\Phi(\bar{\mathbf{x}}_t)\|^2] - \frac{\gamma_x\eta}{4}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2] \\
&\quad + 3\gamma_x\eta\,\underbrace{\mathbb{E}\Big[\Big\|\nabla\Phi(\bar{\mathbf{x}}_t) - \frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{x}}f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\bar{\mathbf{x}}_t), \bar{\mathbf{y}}_t)\Big\|^2\Big]}_{T_1} \\
&\quad + 3\gamma_x\eta\,\underbrace{\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{x}}f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\bar{\mathbf{x}}_t), \bar{\mathbf{y}}_t) - \frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{x}}f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)})\Big\|^2\Big]}_{T_2} \\
&\quad + 3\gamma_x\eta\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{x}}f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)}) - \bar{\mathbf{u}}_t\Big\|^2\Big] ,
\end{aligned}
\tag{25}
$$

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

$$
\begin{aligned}
&T_1 = \mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla\Phi^{(k)}(\bar{\mathbf{x}}_t) - \frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{x}}f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\bar{\mathbf{x}}_t), \bar{\mathbf{y}}_t)\Big\|^2\Big] \\
&\leq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\Big(\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\bar{\mathbf{x}}_t)\Big)^T\nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\bar{\mathbf{x}}_t), \mathbf{y}_*(\bar{\mathbf{x}}_t)) \\
&\quad - \Big(\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\bar{\mathbf{x}}_t)\Big)^T\nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\bar{\mathbf{x}}_t), \bar{\mathbf{y}}_t)\Big\|^2\Big] \\
&\leq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\bar{\mathbf{x}}_t)\Big\|^2\Big\|\nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\bar{\mathbf{x}}_t), \mathbf{y}_*(\bar{\mathbf{x}}_t)) - \nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\bar{\mathbf{x}}_t), \bar{\mathbf{y}}_t)\Big\|^2\Big] \\
&\leq C_g^2 L_f^2\mathbb{E}[\|\mathbf{y}_*(\bar{\mathbf{x}}_t) - \bar{\mathbf{y}}_t\|^2] ,
\end{aligned}
\tag{26}
$$

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

$$
\begin{aligned}
T_2 &= \mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{x}}f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\bar{\mathbf{x}}_t), \bar{\mathbf{y}}_t) - \frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{x}}f^{(k)}(g^{(k)}(\bar{\mathbf{x}}_t), \mathbf{y}_t^{(k)}) \\
&\quad + \frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{x}}f^{(k)}(g^{(k)}(\bar{\mathbf{x}}_t), \mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{x}}f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\leq 2\,\underbrace{\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{x}}f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\bar{\mathbf{x}}_t), \bar{\mathbf{y}}_t) - \frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{x}}f^{(k)}(g^{(k)}(\bar{\mathbf{x}}_t), \mathbf{y}_t^{(k)})\Big\|^2\Big]}_{T_3} \\
&\quad + 2\,\underbrace{\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{x}}f^{(k)}(g^{(k)}(\bar{\mathbf{x}}_t), \mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{x}}f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)})\Big\|^2\Big]}_{T_4} .
\end{aligned}
\tag{27}
$$

Then, as for T_3, we can get

$$
\begin{aligned}
T_3 &\leq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla_{\mathbf{x}} f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\bar{\mathbf{x}}_t), \bar{\mathbf{y}}_t) - \nabla_{\mathbf{x}} f^{(k)}(g^{(k)}(\bar{\mathbf{x}}_t), \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\Big(\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\bar{\mathbf{x}}_t)\Big)^T \nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\bar{\mathbf{x}}_t), \bar{\mathbf{y}}_t) - \nabla g^{(k)}(\bar{\mathbf{x}}_t)^T \nabla_g f^{(k)}(g^{(k)}(\bar{\mathbf{x}}_t), \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\leq C_g^2 \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\bar{\mathbf{x}}_t), \bar{\mathbf{y}}_t) - \nabla_g f^{(k)}(g^{(k)}(\bar{\mathbf{x}}_t), \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\leq C_g^2 L_f^2 \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\bar{\mathbf{y}}_t - \mathbf{y}_t^{(k)}\Big\|^2\Big] ,
\end{aligned} \tag{28}
$$

where the third step holds due to the homogeneous data distribution assumption and Assumption 2, the last step also holds due to the homogeneous data distribution assumption and Assumption 1.

As for T_4, we can get

$$
\begin{aligned}
T_4 &\leq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla_{\mathbf{x}} f^{(k)}(g^{(k)}(\bar{\mathbf{x}}_t), \mathbf{y}_t^{(k)}) - \nabla_{\mathbf{x}} f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla g^{(k)}(\bar{\mathbf{x}}_t)^T \nabla_g f^{(k)}(g^{(k)}(\bar{\mathbf{x}}_t), \mathbf{y}_t^{(k)}) - \nabla g^{(k)}(\mathbf{x}_t^{(k)})^T \nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\leq 2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla g^{(k)}(\bar{\mathbf{x}}_t)^T \nabla_g f^{(k)}(g^{(k)}(\bar{\mathbf{x}}_t), \mathbf{y}_t^{(k)}) - \nabla g^{(k)}(\bar{\mathbf{x}}_t)^T \nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\quad + 2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla g^{(k)}(\bar{\mathbf{x}}_t)^T \nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)}) - \nabla g^{(k)}(\mathbf{x}_t^{(k)})^T \nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\leq 2C_g^2 L_f^2 \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g^{(k)}(\bar{\mathbf{x}}_t) - g^{(k)}(\mathbf{x}_t^{(k)})\Big\|^2\Big] + 2C_f^2 \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla g^{(k)}(\bar{\mathbf{x}}_t) - \nabla g^{(k)}(\mathbf{x}_t^{(k)})\Big\|^2\Big] \\
&\leq 2C_g^4 L_f^2 \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\bar{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\Big\|^2\Big] + 2C_f^2 L_g^2 \frac{1}{K}\sum_{k=1}^{K}\Big[\Big\|\bar{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\Big\|^2\Big] \\
&= 2(C_g^4 L_f^2 + C_f^2 L_g^2)\frac{1}{K}\sum_{k=1}^{K}\Big[\Big\|\bar{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\Big\|^2\Big] ,
\end{aligned} \tag{29}
$$

where fourth and fifth steps hold due to Assumptions 1 and 2. By combining T_3 and T_4, we can get

$$
T_2 \leq 2C_g^2 L_f^2 \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\bar{\mathbf{y}}_t - \mathbf{y}_t^{(k)}\Big\|^2\Big] + 4(C_g^4 L_f^2 + C_f^2 L_g^2)\frac{1}{K}\sum_{k=1}^{K}\Big[\Big\|\bar{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\Big\|^2\Big] . \tag{30}
$$

By combining T_1 and T_2, we can get

$$
\begin{aligned}
\mathbb{E}[\Phi(\bar{\mathbf{x}}_{t+1})] &\leq \mathbb{E}[\Phi(\bar{\mathbf{x}}_t)] - \frac{\gamma_x\eta}{2}\mathbb{E}[\|\nabla\Phi(\bar{\mathbf{x}}_t)\|^2] - \frac{\gamma_x\eta}{4}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2] + 3\gamma_x\eta C_g^2 L_f^2 \mathbb{E}[\|\mathbf{y}_*(\bar{\mathbf{x}}_t) - \bar{\mathbf{y}}_t\|^2] \\
&\quad + 12\gamma_x\eta(C_g^4 L_f^2 + C_f^2 L_g^2)\frac{1}{K}\sum_{k=1}^{K}\Big[\Big\|\bar{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\Big\|^2\Big] + 6\gamma_x\eta C_g^2 L_f^2 \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\bar{\mathbf{y}}_t - \mathbf{y}_t^{(k)}\Big\|^2\Big] \\
&\quad + 3\gamma_x\eta\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{x}} f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)}) - \bar{\mathbf{u}}_t\Big\|^2\Big] ,
\end{aligned} \tag{31}
$$

which completes the proof. □

Lemma 4. *Given Assumption 1-4 and $\beta_x\eta \in (0,1)$, we have*

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

Proof.

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{u}_{t+1}^{(k)} - \bar{\mathbf{u}}_{t+1}\|^2] \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|(1-\beta_x\eta)\mathbf{u}_t^{(k)} + \beta_x\eta\nabla g^{(k)}(\mathbf{x}_{t+1}^{(k)};\xi_{t+1}^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_{t+1}^{(k)},\mathbf{y}_{t+1}^{(k)};\zeta_{t+1}^{(k)}) \\
&\quad - (1-\beta_x\eta)\bar{\mathbf{u}}_t - \beta_x\eta\frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\mathbf{x}_{t+1}^{(k')};\xi_{t+1}^{(k')})^T\nabla_g f^{(k')}(\mathbf{h}_{t+1}^{(k')},\mathbf{y}_{t+1}^{(k')};\zeta_{t+1}^{(k')})\Big\|^2\Big] \\
&\leq (1-\beta_x\eta)^2(1+\frac{1}{p})\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{u}_t^{(k)} - \bar{\mathbf{u}}_t\|^2] \\
&\quad + (1+p)\beta_x^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla g^{(k)}(\mathbf{x}_{t+1}^{(k)};\xi_{t+1}^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_{t+1}^{(k)},\mathbf{y}_{t+1}^{(k)};\zeta_{t+1}^{(k)}) \\
&\quad - \frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\mathbf{x}_{t+1}^{(k')};\xi_{t+1}^{(k')})^T\nabla_g f^{(k')}(\mathbf{h}_{t+1}^{(k')},\mathbf{y}_{t+1}^{(k')};\zeta_{t+1}^{(k')})\Big\|^2\Big] \\
&\leq (1+\frac{1}{p})\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{u}_t^{(k)} - \bar{\mathbf{u}}_t\|^2] + 2p\beta_x^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla g^{(k)}(\mathbf{x}_{t+1}^{(k)};\xi_{t+1}^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_{t+1}^{(k)},\mathbf{y}_{t+1}^{(k)};\zeta_{t+1}^{(k)}) \\
&\quad - \frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(\mathbf{x}_{t+1}^{(k')};\xi_{t+1}^{(k')})^T\nabla_g f^{(k')}(\mathbf{h}_{t+1}^{(k')},\mathbf{y}_{t+1}^{(k')};\zeta_{t+1}^{(k')})\Big\|^2\Big] \\
&\leq (1+\frac{1}{p})\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{u}_t^{(k)} - \bar{\mathbf{u}}_t\|^2] + 2p\beta_x^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla g^{(k)}(\mathbf{x}_{t+1}^{(k)};\xi_{t+1}^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_{t+1}^{(k)},\mathbf{y}_{t+1}^{(k)};\zeta_{t+1}^{(k)})\Big\|^2\Big] \\
&\leq (1+\frac{1}{p})\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{u}_t^{(k)} - \bar{\mathbf{u}}_t\|^2] + 2p\beta_x^2\eta^2 C_g^2 C_f^2 \\
&\leq 2p\beta_x^2\eta^2 C_g^2 C_f^2\sum_{t'=s_t p}^{t}(1+\frac{1}{p})^{t-t'} \\
&\leq 6p^2\beta_x^2\eta^2 C_g^2 C_f^2 \ ,
\end{aligned} \tag{33}$$

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 $(1+\frac{1}{p})^p < 3$. ☐

Lemma 5. *Given Assumption 1-4, $\beta_y\eta \in (0,1)$, and $\eta \leq \frac{1}{10p^2\gamma_y L_f}$, we have*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{v}_t^{(k)} - \bar{\mathbf{v}}_t\|^2] \\
&\leq 3456\beta_y^2\alpha^2p^4\eta^4 C_g^2 L_f^2\sigma_g^2 + 41472\beta_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^8 C_g^6 C_f^2 L_f^2 + 96\beta_y^2p^2\eta^2\sigma_f^2 \ .
\end{aligned} \tag{34}$$

Proof.

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{v}_{t+1}^{(k)}-\bar{\mathbf{v}}_{t+1}\|^2] \\
&=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|(1-\beta_y\eta)\mathbf{v}_t^{(k)}+\beta_y\eta\nabla_y f^{(k)}(\mathbf{h}_{t+1}^{(k)},\mathbf{y}_{t+1}^{(k)};\zeta_{t+1}^{(k)}) \\
&\quad-(1-\beta_y\eta)\bar{\mathbf{v}}_t-\beta_y\eta\frac{1}{K}\sum_{k'=1}^{K}\nabla_y f^{(k')}(\mathbf{h}_{t+1}^{(k')},\mathbf{y}_{t+1}^{(k')};\zeta_{t+1}^{(k')})\Big\|^2\Big] \\
&\leq(1-\beta_y\eta)^2(1+\frac{1}{p})\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{v}_t^{(k)}-\bar{\mathbf{v}}_t\|^2]+(1+p)\beta_y^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla_y f^{(k)}(\mathbf{h}_{t+1}^{(k)},\mathbf{y}_{t+1}^{(k)};\zeta_{t+1}^{(k)}) \\
&\quad-\frac{1}{K}\sum_{k'=1}^{K}\nabla_y f^{(k')}(\mathbf{h}_{t+1}^{(k')},\mathbf{y}_{t+1}^{(k')};\zeta_{t+1}^{(k')})\Big\|^2\Big] \\
&\leq(1+\frac{1}{p})\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{v}_t^{(k)}-\bar{\mathbf{v}}_t\|^2]+2p\beta_y^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla_y f^{(k)}(\mathbf{h}_{t+1}^{(k)},\mathbf{y}_{t+1}^{(k)};\zeta_{t+1}^{(k)}) \\
&\quad-\frac{1}{K}\sum_{k'=1}^{K}\nabla_y f^{(k')}(\mathbf{h}_{t+1}^{(k')},\mathbf{y}_{t+1}^{(k')};\zeta_{t+1}^{(k')})\Big\|^2\Big]\ ,
\end{aligned} \tag{35}$$

where the third step holds due to $\beta_y\eta\in(0,1)$ and $1+p<2p$.

The last term can be bounded as follows:

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla_y f^{(k)}(\mathbf{h}_{t+1}^{(k)},\mathbf{y}_{t+1}^{(k)};\zeta_{t+1}^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}\nabla_y f^{(k')}(\mathbf{h}_{t+1}^{(k')},\mathbf{y}_{t+1}^{(k')};\zeta_{t+1}^{(k')})\Big\|^2\Big] \\
&\leq 4\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla_y f^{(k)}(\mathbf{h}_{t+1}^{(k)},\mathbf{y}_{t+1}^{(k)};\zeta_{t+1}^{(k)})-\nabla_y f^{(k)}(\mathbf{h}_{t+1}^{(k)},\mathbf{y}_{t+1}^{(k)})\Big\|^2\Big] \\
&\quad+4\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla_y f^{(k)}(\mathbf{h}_{t+1}^{(k)},\mathbf{y}_{t+1}^{(k)})-\nabla_y f^{(k)}(\bar{\mathbf{h}}_{t+1},\bar{\mathbf{y}}_{t+1})\Big\|^2\Big] \\
&\quad+4\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla_y f^{(k)}(\bar{\mathbf{h}}_{t+1},\bar{\mathbf{y}}_{t+1})-\frac{1}{K}\sum_{k'=1}^{K}\nabla_y f^{(k')}(\mathbf{h}_{t+1}^{(k')},\mathbf{y}_{t+1}^{(k')})\Big\|^2\Big] \\
&\quad+4\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k'=1}^{K}\nabla_y f^{(k')}(\mathbf{h}_{t+1}^{(k')},\mathbf{y}_{t+1}^{(k')})-\frac{1}{K}\sum_{k'=1}^{K}\nabla_y f^{(k')}(\mathbf{h}_{t+1}^{(k')},\mathbf{y}_{t+1}^{(k')};\zeta_{t+1}^{(k')})\Big\|^2\Big] \\
&\leq 8\sigma_f^2+8L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{h}_{t+1}^{(k)}-\bar{\mathbf{h}}_{t+1}\|^2]+8L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{y}_{t+1}^{(k)}-\bar{\mathbf{y}}_{t+1}\|^2]\ .
\end{aligned} \tag{36}$$

Then, by combining these two inequalities, we can get

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{v}_{t+1}^{(k)}-\bar{\mathbf{v}}_{t+1}\|^2]\leq(1+\frac{1}{p})\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{v}_t^{(k)}-\bar{\mathbf{v}}_t\|^2]+16p\beta_y^2\eta^2\sigma_f^2 \\
&\quad+16p\beta_y^2\eta^2L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{h}_{t+1}^{(k)}-\bar{\mathbf{h}}_{t+1}\|^2]+16p\beta_y^2\eta^2L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{y}_{t+1}^{(k)}-\bar{\mathbf{y}}_{t+1}\|^2]\ .
\end{aligned} \tag{37}$$

In addition, we have

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{y}_{t+1}^{(k)}-\bar{\mathbf{y}}_{t+1}\|^2]=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{y}_{s_tp}^{(k)}+\gamma_y\eta\sum_{t'=s_tp}^{t}\mathbf{v}_{t'}^{(k)}-\bar{\mathbf{y}}_{s_tp}-\gamma_y\eta\sum_{t'=s_tp}^{t}\bar{\mathbf{v}}_{t'}\|^2]$$
$$\leq p\gamma_y^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\sum_{t'=s_tp}^{t}\mathbb{E}[\|\mathbf{v}_{t'}^{(k)}-\bar{\mathbf{v}}_{t'}\|^2]\ , \tag{38}$$

where $s_t=\lfloor (t+1)/p\rfloor$. Thus, we can get

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{v}_{t+1}^{(k)}-\bar{\mathbf{v}}_{t+1}\|^2]\leq(1+\frac{1}{p})\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{v}_{t}^{(k)}-\bar{\mathbf{v}}_{t}\|^2]+16p\beta_y^2\eta^2\sigma_f^2$$
$$+16p\beta_y^2\eta^2L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{h}_{t+1}^{(k)}-\bar{\mathbf{h}}_{t+1}\|^2]+16p^2\gamma_y^2\beta_y^2\eta^4L_f^2\frac{1}{K}\sum_{k=1}^{K}\sum_{t'=s_tp}^{t}\mathbb{E}[\|\mathbf{v}_{t'}^{(k)}-\bar{\mathbf{v}}_{t'}\|^2]$$
$$\leq 16p^2\gamma_y^2\beta_y^2\eta^4L_f^2\sum_{t'=s_tp}^{t}(1+\frac{1}{p})^{t-t'}\frac{1}{K}\sum_{k=1}^{K}\sum_{t''=s_tp}^{t'}\mathbb{E}[\|\mathbf{v}_{t''}^{(k)}-\bar{\mathbf{v}}_{t''}\|^2]$$
$$+16p\beta_y^2\eta^2L_f^2\sum_{t'=s_tp}^{t}(1+\frac{1}{p})^{t-t'}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{h}_{t'+1}^{(k)}-\bar{\mathbf{h}}_{t'+1}\|^2]+16p\beta_y^2\eta^2\sigma_f^2\sum_{t'=s_tp}^{t}(1+\frac{1}{p})^{t-t'}$$
$$\leq 48p^3\gamma_y^2\beta_y^2\eta^4L_f^2\frac{1}{K}\sum_{k=1}^{K}\sum_{t'=s_tp}^{t}\mathbb{E}[\|\mathbf{v}_{t'}^{(k)}-\bar{\mathbf{v}}_{t'}\|^2]$$
$$+48p\beta_y^2\eta^2L_f^2\sum_{t'=s_tp}^{t}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{h}_{t'+1}^{(k)}-\bar{\mathbf{h}}_{t'+1}\|^2]+48p^2\beta_y^2\eta^2\sigma_f^2\ , \tag{39}$$

where the last step holds due to $(1+\frac{1}{p})^p<3$. Then, we can get

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{v}_{t}^{(k)}-\bar{\mathbf{v}}_{t}\|^2]\leq 48p^4\gamma_y^2\beta_y^2\eta^4L_f^2\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{v}_{t}^{(k)}-\bar{\mathbf{v}}_{t}\|^2]$$
$$+48p^2\beta_y^2\eta^2L_f^2\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{h}_{t}^{(k)}-\bar{\mathbf{h}}_{t}\|^2]+48p^2\beta_y^2\eta^2\sigma_f^2$$
$$\leq 48p^4\gamma_y^2\eta^2L_f^2\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{v}_{t}^{(k)}-\bar{\mathbf{v}}_{t}\|^2]+48p^2\beta_y^2\eta^2L_f^2\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{h}_{t}^{(k)}-\bar{\mathbf{h}}_{t}\|^2]$$
$$+48p^2\beta_y^2\eta^2\sigma_f^2\ , \tag{40}$$

where the last step holds due to $\beta_y\eta\leq 1$. By setting $\eta\leq\frac{1}{10p^2\gamma_yL_f}$ such that $1-48p^4\gamma_y^2\eta^2L_f^2\geq\frac{1}{2}$, we can get

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{v}_{t}^{(k)}-\bar{\mathbf{v}}_{t}\|^2]\leq 96p^2\beta_y^2\eta^2L_f^2\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{h}_{t}^{(k)}-\bar{\mathbf{h}}_{t}\|^2]+96p^2\beta_y^2\eta^2\sigma_f^2$$
$$\leq 3456\beta_y^2\alpha^2p^4\eta^4C_g^2L_f^2\sigma_g^2+41472\beta_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^8C_g^6C_f^2L_f^2+96\beta_y^2p^2\eta^2\sigma_f^2\ . \tag{41}$$

where the last step holds due to Lemma 9.

□

Lemma 6. *Given Assumption 1-4, we have*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{x}_{t+1}^{(k)}-\bar{\mathbf{x}}_{t+1}\|^2]\leq 6\gamma_x^2\beta_x^2p^4\eta^4C_g^2C_f^2\ . \tag{42}$$

Proof.

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{x}_{t+1}^{(k)}-\bar{\mathbf{x}}_{t+1}\|^2]=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{x}_{s_tp}^{(k)}-\gamma_x\eta\sum_{t'=s_tp}^{t}\mathbf{u}_{t'}^{(k)}-\bar{\mathbf{x}}_{s_tp}+\gamma_x\eta\sum_{t'=s_tp}^{t}\bar{\mathbf{u}}_{t'}\|^2]\\
&\leq p\gamma_x^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\sum_{t'=s_tp}^{t}\mathbb{E}[\|\mathbf{u}_{t'}^{(k)}-\bar{\mathbf{u}}_{t'}\|^2]\leq 6\gamma_x^2\beta_x^2p^4\eta^4C_g^2C_f^2\ ,
\end{aligned}
\tag{43}
$$

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

Lemma 7. *Given Assumption 1-4, we have*

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{y}_t^{(k)}-\bar{\mathbf{y}}_t\|^2]\\
&\leq 3456\gamma_y^2\beta_y^2\alpha^2p^6\eta^6C_g^2L_f^2\sigma_g^2+41472\gamma_y^2\beta_y^2\alpha^2\gamma_x^2\beta_x^2p^{10}\eta^{10}C_g^6C_f^2L_f^2+96\gamma_y^2\beta_y^2p^4\eta^4\sigma_f^2\ .
\end{aligned}
\tag{44}
$$

Proof.

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{y}_{t+1}^{(k)}-\bar{\mathbf{y}}_{t+1}\|^2]=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{y}_{s_tp}^{(k)}+\gamma_y\eta\sum_{t'=s_tp}^{t}\mathbf{v}_{t'}^{(k)}-\bar{\mathbf{y}}_{s_tp}-\gamma_y\eta\sum_{t'=s_tp}^{t}\bar{\mathbf{v}}_{t'}\|^2]\\
&\leq p\gamma_y^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\sum_{t'=s_tp}^{t}\mathbb{E}[\|\mathbf{v}_{t'}^{(k)}-\bar{\mathbf{v}}_{t'}\|^2]\ ,
\end{aligned}
\tag{45}
$$

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

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{y}_t^{(k)}-\bar{\mathbf{y}}_t\|^2]\leq p^2\gamma_y^2\eta^2\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{v}_t^{(k)}-\bar{\mathbf{v}}_t\|^2]\\
&\leq 3456\gamma_y^2\beta_y^2\alpha^2p^6\eta^6C_g^2L_f^2\sigma_g^2+41472\gamma_y^2\beta_y^2\alpha^2\gamma_x^2\beta_x^2p^{10}\eta^{10}C_g^6C_f^2L_f^2+96\gamma_y^2\beta_y^2p^4\eta^4\sigma_f^2\ .
\end{aligned}
\tag{46}
$$

where the last step holds due to Lemma 5.

□

Lemma 8. *Given Assumptions 1-4, we can get*

$$
\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g^{(k)}(\mathbf{x}_t^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\mathbf{x}_t^{(k')})\Big\|^2\Big]\leq 24\gamma_x^2\beta_x^2p^4\eta^4C_g^4C_f^2\ . \tag{47}
$$

Proof.

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g^{(k)}(\mathbf{x}_t^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\mathbf{x}_t^{(k')})\Big\|^2\Big]\\
&\leq\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g^{(k)}(\mathbf{x}_t^{(k)})-g^{(k)}(\bar{\mathbf{x}}_t)+g(\bar{\mathbf{x}}_t)-\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\mathbf{x}_t^{(k')})\Big\|^2\Big]\\
&\leq 2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g^{(k)}(\mathbf{x}_t^{(k)})-g^{(k)}(\bar{\mathbf{x}}_t)\Big\|^2\Big]+2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g(\bar{\mathbf{x}}_t)-\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\mathbf{x}_t^{(k')})\Big\|^2\Big]\\
&\leq 4C_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\mathbf{x}_t^{(k)}-\bar{\mathbf{x}}_t\|^2]\\
&\leq 24\gamma_x^2\beta_x^2p^4\eta^4C_g^4C_f^2\ ,
\end{aligned}
\tag{48}
$$

where the second step holds due to $g^{(k)}(\bar{\mathbf{x}}_t)=g(\bar{\mathbf{x}}_t)$ for the homogeneous data distribution, the third step holds due to Assumption 2, the last step holds due to Lemma 6. □

Lemma 9. *Given Assumptions 1-4, we can get*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{h}_{t+1}^{(k)}-\frac{1}{K}\sum_{k'=1}^{K}\mathbf{h}_{t+1}^{(k')}\Big\|^2\Big]\leq 36\alpha^2p^2\eta^2C_g^2\sigma_g^2+432\alpha^2\gamma_x^2\beta_x^2p^6\eta^6C_g^6C_f^2\ . \tag{49}$$

Proof.

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{h}_{t+1}^{(k)}-\frac{1}{K}\sum_{k'=1}^{K}\mathbf{h}_{t+1}^{(k')}\Big\|^2\Big]\\
&=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|(1-\alpha\eta)\mathbf{h}_{t}^{(k)}+\alpha\eta g^{(k)}(\mathbf{x}_{t+1}^{(k)};\xi_{t+1}^{(k)})-(1-\alpha\eta)\frac{1}{K}\sum_{k'=1}^{K}\mathbf{h}_{t}^{(k')}-\alpha\eta\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\mathbf{x}_{t+1}^{(k')};\xi_{t+1}^{(k')})\Big\|^2\Big]\\
&\leq(1-\alpha\eta)^2(1+\frac{1}{p})\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{h}_{t}^{(k)}-\frac{1}{K}\sum_{k'=1}^{K}\mathbf{h}_{t}^{(k')}\Big\|^2\Big]\\
&\quad+\alpha^2\eta^2(1+p)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g^{(k)}(\mathbf{x}_{t+1}^{(k)};\xi_{t+1}^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\mathbf{x}_{t+1}^{(k')};\xi_{t+1}^{(k')})\Big\|^2\Big]\\
&\leq(1+\frac{1}{p})\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{h}_{t}^{(k)}-\frac{1}{K}\sum_{k'=1}^{K}\mathbf{h}_{t}^{(k')}\Big\|^2\Big]+2p\alpha^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g^{(k)}(\mathbf{x}_{t+1}^{(k)};\xi_{t+1}^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\mathbf{x}_{t+1}^{(k')};\xi_{t+1}^{(k')})\Big\|^2\Big]\\
&\leq 2p\alpha^2\eta^2C_g^2\sum_{t'=s_tp}^{t}(1+\frac{1}{p})^{t-t'}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g^{(k)}(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\mathbf{x}_{t'}^{(k')};\xi_{t'}^{(k')})\Big\|^2\Big]\ ,
\end{aligned}\tag{50}$$

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

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g^{(k)}(\mathbf{x}_{t}^{(k)};\xi_{t}^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\mathbf{x}_{t}^{(k')};\xi_{t}^{(k')})\Big\|^2\Big]\\
&=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g^{(k)}(\mathbf{x}_{t}^{(k)};\xi_{t}^{(k)})-g^{(k)}(\mathbf{x}_{t}^{(k)})+g^{(k)}(\mathbf{x}_{t}^{(k)})\\
&\quad-\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\mathbf{x}_{t}^{(k')})+\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\mathbf{x}_{t}^{(k')})-\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\mathbf{x}_{t}^{(k')};\xi_{t}^{(k')})\Big\|^2\Big]\\
&\leq 6\sigma_g^2+3\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g^{(k)}(\mathbf{x}_{t}^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\mathbf{x}_{t}^{(k')})\Big\|^2\Big]\\
&\leq 6\sigma_g^2+72\gamma_x^2\beta_x^2p^4\eta^4C_g^4C_f^2\ ,
\end{aligned}\tag{51}$$

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

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{h}_{t+1}^{(k)}-\frac{1}{K}\sum_{k'=1}^{K}\mathbf{h}_{t+1}^{(k')}\Big\|^2\Big]\\
&\leq 2p\alpha^2\eta^2C_g^2\sum_{t'=s_tp}^{t}(1+\frac{1}{p})^{t-t'}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g^{(k)}(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}g^{(k')}(\mathbf{x}_{t'}^{(k')};\xi_{t'}^{(k')})\Big\|^2\Big]\\
&\leq 2p\alpha^2\eta^2C_g^2(6\sigma_g^2+72\gamma_x^2\beta_x^2p^4\eta^4C_g^4C_f^2)\sum_{t'=s_tp}^{t}(1+\frac{1}{p})^{t-t'}\\
&\leq 36\alpha^2p^2\eta^2C_g^2\sigma_g^2+432\alpha^2\gamma_x^2\beta_x^2p^6\eta^6C_g^6C_f^2\ .
\end{aligned}\tag{52}$$

where the last step holds due to $(1+\frac{1}{p})^p<3$. □

Lemma 10. *Given Assumptions 1-4, we can get*

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})\Big\|^2\Big] \\
&\leq 4\gamma_x^2\eta^2(C_g^4L_f^2 + C_f^2L_g^2)\mathbb{E}\Big[\Big\|\bar{\mathbf{u}}_{t-1}\Big\|^2\Big] + 4\gamma_y^2\eta^2C_g^2L_f^2\mathbb{E}\Big[\Big\|\bar{\mathbf{v}}_{t-1}\Big\|^2\Big] + 24\beta_x^2\gamma_x^2p^2\eta^4(C_g^4L_f^2 + C_f^2L_g^2)C_g^2C_f^2 \\
&\quad + 13824\beta_y^2\gamma_y^2\alpha^2p^4\eta^6C_g^4L_f^4\sigma_g^2 + 165888\beta_y^2\gamma_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^{10}C_g^8C_f^2L_f^4 + 384\beta_y^2\gamma_y^2p^2\eta^4C_g^2L_f^2\sigma_f^2\,.
\end{aligned}
\tag{53}
$$

Proof.

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})\Big\|^2\Big] \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)}) \\
&\quad + \nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)}) - \nabla g^{(k)}(\mathbf{x}_{t-1}^{(k)})^T\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})\Big\|^2\Big] \\
&\leq 2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})\Big\|^2\Big] \\
&\quad + 2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)}) - \nabla g^{(k)}(\mathbf{x}_{t-1}^{(k)})^T\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})\Big\|^2\Big] \\
&\leq 2C_g^2L_f^2\Big(C_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{x}_t^{(k)} - \mathbf{x}_{t-1}^{(k)}\Big\|^2\Big] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{y}_t^{(k)} - \mathbf{y}_{t-1}^{(k)}\Big\|^2\Big]\Big) + 2C_f^2L_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{x}_t^{(k)} - \mathbf{x}_{t-1}^{(k)}\Big\|^2\Big] \\
&= 2(C_g^4L_f^2 + C_f^2L_g^2)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{x}_t^{(k)} - \mathbf{x}_{t-1}^{(k)}\Big\|^2\Big] + 2C_g^2L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{y}_t^{(k)} - \mathbf{y}_{t-1}^{(k)}\Big\|^2\Big] \\
&\leq 4\gamma_x^2\eta^2(C_g^4L_f^2 + C_f^2L_g^2)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{u}_{t-1}^{(k)} - \bar{\mathbf{u}}_{t-1}\Big\|^2\Big] + 4\gamma_x^2\eta^2(C_g^4L_f^2 + C_f^2L_g^2)\mathbb{E}\Big[\Big\|\bar{\mathbf{u}}_{t-1}\Big\|^2\Big] \\
&\quad + 4\gamma_y^2\eta^2C_g^2L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{v}_{t-1}^{(k)} - \bar{\mathbf{v}}_{t-1}\Big\|^2\Big] + 4\gamma_y^2\eta^2C_g^2L_f^2\mathbb{E}\Big[\Big\|\bar{\mathbf{v}}_{t-1}\Big\|^2\Big] \\
&\leq 4\gamma_x^2\eta^2(C_g^4L_f^2 + C_f^2L_g^2)\mathbb{E}\Big[\Big\|\bar{\mathbf{u}}_{t-1}\Big\|^2\Big] + 4\gamma_y^2\eta^2C_g^2L_f^2\mathbb{E}\Big[\Big\|\bar{\mathbf{v}}_{t-1}\Big\|^2\Big] + 24\beta_x^2\gamma_x^2p^2\eta^4(C_g^4L_f^2 + C_f^2L_g^2)C_g^2C_f^2 \\
&\quad + 13824\beta_y^2\gamma_y^2\alpha^2p^4\eta^6C_g^4L_f^4\sigma_g^2 + 165888\beta_y^2\gamma_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^{10}C_g^8C_f^2L_f^4 + 384\beta_y^2\gamma_y^2p^2\eta^4C_g^2L_f^2\sigma_f^2\,,
\end{aligned}
\tag{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 1-4 and $\alpha\eta \in (0, 1)$, we can get*

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_t - \frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_t^{(k)})\Big\|^2\Big] \quad \leq \frac{2\gamma_x^2C_g^2}{\alpha^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2] + \frac{\sigma_g^2}{\alpha\eta TK} + \frac{12\gamma_x^2\beta_x^2p^2\eta^2C_g^4C_f^2}{\alpha^2} + \frac{\alpha\eta\sigma_g^2}{K}\,.
\tag{55}
$$

Proof.

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_{t+1}-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_{t+1}^{(k)})\Big\|^2\Big]=\mathbb{E}\Big[\Big\|(1-\alpha\eta)\frac{1}{K}\sum_{k=1}^{K}\mathbf{h}_t^{(k)}+\alpha\eta\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_{t+1}^{(k)};\xi_{t+1}^{(k)})-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_{t+1}^{(k)})\Big\|^2\Big]\\
&=\mathbb{E}\Big[\Big\|(1-\alpha\eta)\frac{1}{K}\sum_{k=1}^{K}\Big(\mathbf{h}_t^{(k)}-g^{(k)}(\mathbf{x}_t^{(k)})\Big)+(1-\alpha\eta)\frac{1}{K}\sum_{k=1}^{K}\Big(g^{(k)}(\mathbf{x}_{t+1}^{(k)})-g^{(k)}(\mathbf{x}_t^{(k)})\Big)\\
&\quad+\alpha\eta\frac{1}{K}\sum_{k=1}^{K}\Big(g^{(k)}(\mathbf{x}_{t+1}^{(k)};\xi_{t+1}^{(k)})-g^{(k)}(\mathbf{x}_{t+1}^{(k)})\Big)\Big\|^2\Big]\\
&=\mathbb{E}\Big[\Big\|(1-\alpha\eta)\frac{1}{K}\sum_{k=1}^{K}\Big(\mathbf{h}_t^{(k)}-g^{(k)}(\mathbf{x}_t^{(k)})\Big)+(1-\alpha\eta)\frac{1}{K}\sum_{k=1}^{K}\Big(g^{(k)}(\mathbf{x}_{t+1}^{(k)})-g^{(k)}(\mathbf{x}_t^{(k)})\Big)\Big\|^2\Big]\\
&\quad+\mathbb{E}\Big[\Big\|\alpha\eta\frac{1}{K}\sum_{k=1}^{K}\Big(g^{(k)}(\mathbf{x}_{t+1}^{(k)};\xi_{t+1}^{(k)})-g^{(k)}(\mathbf{x}_{t+1}^{(k)})\Big)\Big\|^2\Big]\\
&\leq(1-\alpha\eta)^2(1+\frac{1}{a})\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\Big(\mathbf{h}_t^{(k)}-g^{(k)}(\mathbf{x}_t^{(k)})\Big)\Big\|^2\Big]\\
&\quad+(1-\alpha\eta)^2(1+a)\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\Big(g^{(k)}(\mathbf{x}_{t+1}^{(k)})-g^{(k)}(\mathbf{x}_t^{(k)})\Big)\Big\|^2\Big]+\frac{\alpha^2\eta^2\sigma_g^2}{K}\\
&\leq(1-\alpha\eta)\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\Big(\mathbf{h}_t^{(k)}-g^{(k)}(\mathbf{x}_t^{(k)})\Big)\Big\|^2\Big]+\frac{C_g^2}{\alpha\eta}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{x}_{t+1}^{(k)}-\mathbf{x}_t^{(k)}\Big\|^2\Big]+\frac{\alpha^2\eta^2\sigma_g^2}{K}\\
&\leq(1-\alpha\eta)\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_t-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_t^{(k)})\Big\|^2\Big]+\frac{2\eta\gamma_x^2C_g^2}{\alpha}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{u}_t^{(k)}-\bar{\mathbf{u}}_t\Big\|^2\Big]+\frac{2\eta\gamma_x^2C_g^2}{\alpha}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2]+\frac{\alpha^2\eta^2\sigma_g^2}{K}\\
&\leq(1-\alpha\eta)\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_t-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_t^{(k)})\Big\|^2\Big]+\frac{2\eta\gamma_x^2C_g^2}{\alpha}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2]+\frac{12\gamma_x^2\beta_x^2p^2\eta^3C_g^4C_f^2}{\alpha}+\frac{\alpha^2\eta^2\sigma_g^2}{K}\ ,
\end{aligned}
\tag{56}
$$

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.

It can be reformulated as follows:

$$
\begin{aligned}
&\alpha\eta\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_t-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_t^{(k)})\Big\|^2\Big]\leq\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_t-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_t^{(k)})\Big\|^2\Big]\\
&-\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_{t+1}-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_{t+1}^{(k)})\Big\|^2\Big]+\frac{2\eta\gamma_x^2C_g^2}{\alpha}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2]+\frac{12\gamma_x^2\beta_x^2p^2\eta^3C_g^4C_f^2}{\alpha}+\frac{\alpha^2\eta^2\sigma_g^2}{K}\ .
\end{aligned}
\tag{57}
$$

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

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_t-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_t^{(k)})\Big\|^2\Big]\\
&\leq\frac{1}{\alpha\eta T}\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_0-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_0^{(k)})\Big\|^2\Big]+\frac{2\gamma_x^2C_g^2}{\alpha^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2]+\frac{12\gamma_x^2\beta_x^2p^2\eta^2C_g^4C_f^2}{\alpha^2}+\frac{\alpha\eta\sigma_g^2}{K}\\
&\leq\frac{2\gamma_x^2C_g^2}{\alpha^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2]+\frac{\sigma_g^2}{\alpha\eta TK}+\frac{12\gamma_x^2\beta_x^2p^2\eta^2C_g^4C_f^2}{\alpha^2}+\frac{\alpha\eta\sigma_g^2}{K}\ ,
\end{aligned}
\tag{58}
$$

where the last step holds due to the following inequality:

$$
\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_0-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_0^{(k)})\Big\|^2\Big]=\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_0^{(k)};\xi_0^{(k)})-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_0^{(k)})\Big\|^2\Big]\leq\frac{\sigma_g^2}{K}\ .
\tag{59}
$$

□

Lemma 12. *Given Assumptions 1-4, we can get*

$$\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\Big(\nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})-\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})\Big)\Big\|^2\Big]\leq\frac{C_f^2\sigma_{g'}^2}{K}\ ,$$
$$\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\Big(\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})-\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)};\zeta_t^{(k)})\Big)\Big\|^2\Big]\leq\frac{C_g^2\sigma_f^2}{K}\ . \tag{60}$$

Proof.

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\Big(\nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})-\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})\Big)\Big\|^2\Big]\\
&=\mathbb{E}_{\mathcal{F}}\mathbb{E}_{\xi|\mathcal{F}}\mathbb{E}_{\zeta|\mathcal{F}}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\Big(\nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})-\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})\Big)\Big\|^2\Big]\\
&=\mathbb{E}_{\mathcal{F}}\mathbb{E}_{\xi|\mathcal{F}}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\Big(\nabla g^{(k)}(\mathbf{x}_t^{(k)})-\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})\Big)^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})\Big\|^2\Big]\\
&=\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}}\mathbb{E}_{\xi|\mathcal{F}}\Big[\Big\|\Big(\nabla g^{(k)}(\mathbf{x}_t^{(k)})-\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})\Big)^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})\Big\|^2\Big]\\
&\leq\frac{\sigma_{g'}^2C_f^2}{K}\ ,
\end{aligned}\tag{61}$$

where $\mathcal{F}$ denotes all random factors except the sampling operation in the t-th iteration, ξ and ζ are independent given $\mathcal{F}$, the third step holds since $(\nabla g^{(k)}(\mathbf{x}_t^{(k)})-\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)}))^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})$ are independent random vectors regarding ξ across workers and the mean is zero, the last step holds due to Assumptions 1, 2.

Similarly, we can get

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\Big(\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})-\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)};\zeta_t^{(k)})\Big)\Big\|^2\Big]\\
&=\mathbb{E}_{\mathcal{F}}\mathbb{E}_{\xi|\mathcal{F}}\mathbb{E}_{\zeta|\mathcal{F}}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\Big(\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})-\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)};\zeta_t^{(k)})\Big)\Big\|^2\Big]\\
&=\mathbb{E}_{\mathcal{F}}\mathbb{E}_{\xi|\mathcal{F}}\mathbb{E}_{\zeta|\mathcal{F}}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\Big(\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})-\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)};\zeta_t^{(k)})\Big)\Big\|^2\Big]\\
&=\mathbb{E}_{\mathcal{F}}\mathbb{E}_{\xi|\mathcal{F}}\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}_{\zeta|\mathcal{F}}\Big[\Big\|\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\Big(\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})-\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)};\zeta_t^{(k)})\Big)\Big\|^2\Big]\\
&=\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}_{\mathcal{F}}\mathbb{E}_{\xi|\mathcal{F}}\mathbb{E}_{\zeta|\mathcal{F}}\Big[\Big\|\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\Big(\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})-\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)};\zeta_t^{(k)})\Big)\Big\|^2\Big]\\
&\leq\frac{C_g^2\sigma_f^2}{K}\ ,
\end{aligned}\tag{62}$$

where $\mathcal{F}$ denotes all random factors except the sampling operation in the t-th iteration, ξ and ζ are independent given $\mathcal{F}$, the third step holds since $\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\Big(\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})-\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)};\zeta_t^{(k)})\Big)$ are independent random vectors regarding ζ across workers and the mean is zero, the last step holds due to Assumptions 1, 2. $\square$

Lemma 13. *Given Assumption 1-4, $\beta_x\eta \in (0,1)$, and $\eta < 1$, we can get*

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathbf{u}_t^{(k)}\Big\|^2\Big] \\
&\leq \frac{3C_g^2L_f^2\sigma_g^2 + 3C_f^2\sigma_{g'}^2 + 3C_g^2\sigma_f^2}{\beta_x\eta T} + 6C_g^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_t - \frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_t^{(k)})\Big\|^2\Big] \\
&+ \frac{8\gamma_x^2(C_g^4L_f^2 + C_f^2L_g^2)}{\beta_x^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2] + \frac{8\gamma_y^2C_g^2L_f^2}{\beta_x^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{v}}_t\|^2] + \frac{12\gamma_x^2C_g^4}{\alpha}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2] \\
&+ 48\gamma_x^2p^2\eta^2(C_g^4L_f^2 + C_f^2L_g^2)C_g^2C_f^2 + 144\gamma_x^2\beta_x^2p^4\eta^4C_g^6C_f^4 + 216\alpha^2p^2\eta^2C_g^4\sigma_g^2 + 2592\alpha^2\gamma_x^2\beta_x^2p^6\eta^6C_g^8C_f^2 \\
&+ \frac{27648\beta_y^2\gamma_y^2\alpha^2p^4\eta^4C_g^4L_f^4\sigma_g^2}{\beta_x^2} + \frac{331776\beta_y^2\gamma_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^8C_g^8C_f^2L_f^4}{\beta_x^2} + \frac{786\beta_y^2\gamma_y^2p^2\eta^2C_g^2L_f^2\sigma_f^2}{\beta_x^2} \\
&+ \frac{2\beta_x\eta C_f^2\sigma_{g'}^2}{K} + \frac{2\beta_x\eta C_g^2\sigma_f^2}{K} + \frac{72\gamma_x^2\beta_x^2p^2\eta^2C_g^6C_f^2}{\alpha} + \frac{6\alpha^2\eta C_g^2\sigma_g^2}{K}\ .
\end{aligned}
\tag{63}
$$

Proof.

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\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)}\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|(1-\beta_x\eta)\frac{1}{K}\sum_{k=1}^{K}(\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)}) - \mathbf{u}_{t-1}^{(k)}) \\
&+ (1-\beta_x\eta)\frac{1}{K}\sum_{k=1}^{K}(\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})) \\
&+ \beta_x\eta\frac{1}{K}\sum_{k=1}^{K}\Big(\nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)}) \\
&+ \nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)}) - \nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)}) \\
&+ \nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)}) - \nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)};\zeta_t^{(k)})\Big)\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|(1-\beta_x\eta)\frac{1}{K}\sum_{k=1}^{K}(\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)}) - \mathbf{u}_{t-1}^{(k)}) \\
&\quad + (1-\beta_x\eta)\frac{1}{K}\sum_{k=1}^{K}(\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})) \\
&\quad + \beta_x\eta\frac{1}{K}\sum_{k=1}^{K}(\nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)}))\Big\|^2\Big] \\
&+ \beta_x\eta^2\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\Big(\nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)}) - \nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)}) \\
&\quad + \nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)}) - \nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)};\zeta_t^{(k)})\Big)\Big\|^2\Big] \\
&\triangleq T_1 + \beta_x^2\eta^2T_2
\end{aligned}
\tag{64}
$$

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

$$
\begin{aligned}
T_1 &\leq (1-\beta_x\eta)^2(1+a)\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\mathbf{u}_{t-1}^{(k)}\Big\|^2\Big] \\
&\quad + 2(1-\beta_x\eta)^2(1+\frac{1}{a})\mathbb{E}\Big[\Big\|\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}\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})\Big\|^2\Big] \\
&\quad + 2\beta_x^2\eta^2(1+\frac{1}{a})\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla g^{(k)}(\mathbf{x}_{t}^{(k)})^T\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_{t}^{(k)}),\mathbf{y}_{t}^{(k)}) \\
&\quad - \frac{1}{K}\sum_{k=1}^{K}\nabla g^{(k)}(\mathbf{x}_{t}^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_{t}^{(k)},\mathbf{y}_{t}^{(k)})\Big\|^2\Big] \\
&\leq (1-\beta_x\eta)\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\mathbf{u}_{t-1}^{(k)}\Big\|^2\Big] \\
&\quad + \frac{2}{\beta_x\eta}\mathbb{E}\Big[\Big\|\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}\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})\Big\|^2\Big] \\
&\quad + 2\beta_x\eta\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla g^{(k)}(\mathbf{x}_{t}^{(k)})^T\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_{t}^{(k)}),\mathbf{y}_{t}^{(k)}) \\
&\quad - \frac{1}{K}\sum_{k=1}^{K}\nabla g^{(k)}(\mathbf{x}_{t}^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_{t}^{(k)},\mathbf{y}_{t}^{(k)})\Big\|^2\Big] \\
&\leq (1-\beta_x\eta)\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\mathbf{u}_{t-1}^{(k)}\Big\|^2\Big] \\
&\quad + \frac{2}{\beta_x\eta}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_{t}^{(k)}),\mathbf{y}_{t}^{(k)})-\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})\Big\|^2\Big] \\
&\quad + 2\beta_x\eta\underbrace{\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla g^{(k)}(\mathbf{x}_{t}^{(k)})^T\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_{t}^{(k)}),\mathbf{y}_{t}^{(k)})-\nabla g^{(k)}(\mathbf{x}_{t}^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_{t}^{(k)},\mathbf{y}_{t}^{(k)})\Big\|^2\Big]}_{T_3} \\
&\leq (1-\beta_x\eta)\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\mathbf{u}_{t-1}^{(k)}\Big\|^2\Big] + 48\beta_x\gamma_x^2p^2\eta^3(C_g^4L_f^2+C_f^2L_g^2)C_g^2C_f^2 \\
&\quad + \frac{27648\beta_y^2\gamma_y^2\alpha^2p^4\eta^5C_g^4L_f^4\sigma_g^2}{\beta_x} + \frac{331776\beta_y^2\gamma_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^9C_g^8C_f^2L_f^4}{\beta_x} + \frac{786\beta_y^2\gamma_y^2p^2\eta^3C_g^2L_f^2\sigma_f^2}{\beta_x} \\
&\quad + 144\gamma_x^2\beta_x^3p^4\eta^5C_g^6C_f^4 + 216\beta_x\alpha^2p^2\eta^3C_g^4\sigma_g^2 + 2592\alpha^2\gamma_x^2\beta_x^3p^6\eta^7C_g^8C_f^2 \\
&\quad + \frac{8\gamma_x^2\eta(C_g^4L_f^2+C_f^2L_g^2)}{\beta_x}\mathbb{E}[\|\bar{\mathbf{u}}_{t-1}\|^2] + \frac{8\gamma_y^2\eta C_g^2L_f^2}{\beta_x}\mathbb{E}[\|\bar{\mathbf{v}}_{t-1}\|^2] + \frac{12\beta_x\eta\gamma_x^2C_g^4}{\alpha}\mathbb{E}[\|\bar{\mathbf{u}}_{t-1}\|^2] \\
&\quad + 6\beta_x\eta C_g^2\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_{t-1}-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_{t-1}^{(k)})\Big\|^2\Big] + \frac{72\gamma_x^2\beta_x^3p^2\eta^3C_g^6C_f^2}{\alpha} + \frac{6\alpha^2\eta^2\beta_xC_g^2\sigma_g^2}{K}\ ,
\end{aligned} \tag{65}
$$

where the second step holds due to $a = \frac{\beta_x \eta}{1-\beta_x \eta}$ and $0 < \beta_x \eta < 1$, the last step holds due to Lemma 10 and the following inequality.

$$
\begin{aligned}
T_3 &= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)}) \\
&\quad - \nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_t^{(k')}), \mathbf{y}_t^{(k)}) \\
&\quad + \nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_t^{(k')}), \mathbf{y}_t^{(k)}) \\
&\quad - \nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} \mathbf{h}_t^{(k')}, \mathbf{y}_t^{(k)}) \\
&\quad + \nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} \mathbf{h}_t^{(k')}, \mathbf{y}_t^{(k)}) - \nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)}, \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\le 3\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)}) \\
&\quad - \nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_t^{(k')}), \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\quad + 3\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_t^{(k')}), \mathbf{y}_t^{(k)}) \\
&\quad - \nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} \mathbf{h}_t^{(k')}, \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\quad + 3\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} \mathbf{h}_t^{(k')}, \mathbf{y}_t^{(k)}) - \nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)}, \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\le 3C_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)}) - \nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_t^{(k')}), \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\quad + 3C_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_t^{(k')}), \mathbf{y}_t^{(k)}) - \nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} \mathbf{h}_t^{(k')}, \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\quad + 3C_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\nabla_g f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} \mathbf{h}_t^{(k')}, \mathbf{y}_t^{(k)}) - \nabla_g f^{(k)}(\mathbf{h}_t^{(k)}, \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\le 3C_g^2L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g^{(k)}(\mathbf{x}_t^{(k)}) - \frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_t^{(k')})\Big\|^2\Big] + 3C_g^2\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K} g^{(k)}(\mathbf{x}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathbf{h}_t^{(k)}\Big\|^2\Big] \\
&\quad + 3C_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k'=1}^{K}\mathbf{h}_t^{(k')} - \mathbf{h}_t^{(k)}\Big\|^2\Big] \\
&\le 72\gamma_x^2\beta_x^2p^4\eta^4C_g^6C_f^4 + 108\alpha^2p^2\eta^2C_g^4\sigma_g^2 + 1296\alpha^2\gamma_x^2\beta_x^2p^6\eta^6C_g^8C_f^2 \\
&\quad + 3C_g^2\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_{t-1} - \frac{1}{K}\sum_{k=1}^{K} g^{(k)}(\mathbf{x}_{t-1}^{(k)})\Big\|^2\Big] + \frac{6\eta\gamma_x^2C_g^4}{\alpha}\mathbb{E}[\|\bar{\mathbf{u}}_{t-1}\|^2] + \frac{36\gamma_x^2\beta_x^2p^2\eta^3C_g^6C_f^2}{\alpha} + \frac{3\alpha^2\eta^2C_g^2\sigma_g^2}{K}\ , \\
&\le 72\gamma_x^2\beta_x^2p^4\eta^4C_g^6C_f^4 + 108\alpha^2p^2\eta^2C_g^4\sigma_g^2 + 1296\alpha^2\gamma_x^2\beta_x^2p^6\eta^6C_g^8C_f^2 \\
&\quad + 3C_g^2\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_{t-1} - \frac{1}{K}\sum_{k=1}^{K} g^{(k)}(\mathbf{x}_{t-1}^{(k)})\Big\|^2\Big] + \frac{6\gamma_x^2C_g^4}{\alpha}\mathbb{E}[\|\bar{\mathbf{u}}_{t-1}\|^2] + \frac{36\gamma_x^2\beta_x^2p^2\eta^2C_g^6C_f^2}{\alpha} + \frac{3\alpha^2\eta C_g^2\sigma_g^2}{K}\ ,
\end{aligned} \tag{66}
$$

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

$$
\begin{aligned}
T_2 &\leq 2\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\Big(\nabla g^{(k)}(\mathbf{x}_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)}) - \nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})\Big)\Big\|^2\Big] \\
&\quad + 2\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\Big(\nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)}) - \nabla g^{(k)}(\mathbf{x}_t^{(k)};\xi_t^{(k)})^T\nabla_g f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)};\zeta_t^{(k)})\Big)\Big\|^2\Big] \\
&\leq \frac{2C_f^2\sigma_{g'}^2}{K} + \frac{2C_g^2\sigma_f^2}{K}\ ,
\end{aligned}
\tag{67}
$$

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

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathbf{u}_t^{(k)}\Big\|^2\Big] \\
&\leq (1-\beta_x\eta)\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathbf{u}_{t-1}^{(k)}\Big\|^2\Big] + 48\beta_x\gamma_x^2p^2\eta^3(C_g^4L_f^2+C_f^2L_g^2)C_g^2C_f^2 \\
&\quad + \frac{27648\beta_y^2\gamma_y^2\alpha^2p^4\eta^5C_g^4L_f^4\sigma_g^2}{\beta_x} + \frac{331776\beta_y^2\gamma_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^9C_g^8C_f^2L_f^4}{\beta_x} + \frac{786\beta_y^2\gamma_y^2p^2\eta^3C_g^2L_f^2\sigma_f^2}{\beta_x} \\
&\quad + 144\gamma_x^2\beta_x^3p^4\eta^5C_g^6C_f^4 + 216\beta_x\alpha^2p^2\eta^3C_g^4\sigma_g^2 + 2592\alpha^2\gamma_x^2\beta_x^3p^6\eta^7C_g^8C_f^2 \\
&\quad + \frac{8\gamma_x^2\eta(C_g^4L_f^2+C_f^2L_g^2)}{\beta_x}\mathbb{E}[\|\bar{\mathbf{u}}_{t-1}\|^2] + \frac{8\gamma_y^2\eta C_g^2L_f^2}{\beta_x}\mathbb{E}[\|\bar{\mathbf{v}}_{t-1}\|^2] + \frac{12\beta_x\eta\gamma_x^2C_g^4}{\alpha}\mathbb{E}[\|\bar{\mathbf{u}}_{t-1}\|^2] \\
&\quad + 6\beta_x\eta C_g^2\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_{t-1} - \frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_{t-1}^{(k)})\Big\|^2\Big] + \frac{72\gamma_x^2\beta_x^3p^2\eta^3C_g^6C_f^2}{\alpha} + \frac{6\alpha^2\eta^2\beta_xC_g^2\sigma_g^2}{K} \\
&\quad + \frac{2\beta_x^2\eta^2C_f^2\sigma_{g'}^2}{K} + \frac{2\beta_x^2\eta^2C_g^2\sigma_f^2}{K}\ ,
\end{aligned}
\tag{68}
$$

It can be reformulated as follows:

$$
\begin{aligned}
&\beta_x\eta\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathbf{u}_t^{(k)}\Big\|^2\Big] \\
&\leq \mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathbf{u}_t^{(k)}\Big\|^2\Big] \\
&\quad - \mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_{t+1}^{(k)}),\mathbf{y}_{t+1}^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathbf{u}_{t+1}^{(k)}\Big\|^2\Big] + \frac{8\gamma_x^2\eta(C_g^4L_f^2+C_f^2L_g^2)}{\beta_x}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2] \\
&\quad + \frac{8\gamma_y^2\eta C_g^2L_f^2}{\beta_x}\mathbb{E}[\|\bar{\mathbf{v}}_t\|^2] + \frac{12\beta_x\eta\gamma_x^2C_g^4}{\alpha}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2] + 6\beta_x\eta C_g^2\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_t - \frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_t^{(k)})\Big\|^2\Big] \\
&\quad + 48\beta_x\gamma_x^2p^2\eta^3(C_g^4L_f^2+C_f^2L_g^2)C_g^2C_f^2 + 144\gamma_x^2\beta_x^3p^4\eta^5C_g^6C_f^4 + 216\beta_x\alpha^2p^2\eta^3C_g^4\sigma_g^2 + 2592\alpha^2\gamma_x^2\beta_x^3p^6\eta^7C_g^8C_f^2 \\
&\quad + \frac{27648\beta_y^2\gamma_y^2\alpha^2p^4\eta^5C_g^4L_f^4\sigma_g^2}{\beta_x} + \frac{331776\beta_y^2\gamma_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^9C_g^8C_f^2L_f^4}{\beta_x} + \frac{786\beta_y^2\gamma_y^2p^2\eta^3C_g^2L_f^2\sigma_f^2}{\beta_x} \\
&\quad + \frac{2\beta_x^2\eta^2C_f^2\sigma_{g'}^2}{K} + \frac{2\beta_x^2\eta^2C_g^2\sigma_f^2}{K} + \frac{72\gamma_x^2\beta_x^3p^2\eta^3C_g^6C_f^2}{\alpha} + \frac{6\alpha^2\eta^2\beta_xC_g^2\sigma_g^2}{K}\ .
\end{aligned}
\tag{69}
$$

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

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\Big\|\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)}\Big\|^2\Big] \\
&\leq \frac{1}{\beta_x\eta T}\mathbb{E}\Big[\Big\|\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)}\Big\|^2\Big]+6C_g^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_t-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_t^{(k)})\Big\|^2\Big] \\
&\quad+\frac{8\gamma_x^2(C_g^4L_f^2+C_f^2L_g^2)}{\beta_x^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2]+\frac{8\gamma_y^2C_g^2L_f^2}{\beta_x^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{v}}_t\|^2]+\frac{12\gamma_x^2C_g^4}{\alpha}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2] \\
&\quad+48\gamma_x^2p^2\eta^2(C_g^4L_f^2+C_f^2L_g^2)C_g^2C_f^2+144\gamma_x^2\beta_x^2p^4\eta^4C_g^6C_f^4+216\alpha^2p^2\eta^2C_g^4\sigma_g^2+2592\alpha^2\gamma_x^2\beta_x^2p^6\eta^6C_g^8C_f^2 \\
&\quad+\frac{27648\beta_y^2\gamma_y^2\alpha^2p^4\eta^4C_g^4L_f^4\sigma_g^2}{\beta_x^2}+\frac{331776\beta_y^2\gamma_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^8C_g^8C_f^2L_f^4}{\beta_x^2}+\frac{786\beta_y^2\gamma_y^2p^2\eta^2C_g^2L_f^2\sigma_f^2}{\beta_x^2} \\
&\quad+\frac{2\beta_x\eta C_f^2\sigma_{g'}^2}{K}+\frac{2\beta_x\eta C_g^2\sigma_f^2}{K}+\frac{72\gamma_x^2\beta_x^2p^2\eta^2C_g^6C_f^2}{\alpha}+\frac{6\alpha^2\eta C_g^2\sigma_g^2}{K} \\
&\leq \frac{3C_g^2L_f^2\sigma_g^2+3C_f^2\sigma_{g'}^2+3C_g^2\sigma_f^2}{\beta_x\eta T}+6C_g^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_t-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_t^{(k)})\Big\|^2\Big] \\
&\quad+\frac{8\gamma_x^2(C_g^4L_f^2+C_f^2L_g^2)}{\beta_x^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2]+\frac{8\gamma_y^2C_g^2L_f^2}{\beta_x^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{v}}_t\|^2]+\frac{12\gamma_x^2C_g^4}{\alpha}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2] \\
&\quad+48\gamma_x^2p^2\eta^2(C_g^4L_f^2+C_f^2L_g^2)C_g^2C_f^2+144\gamma_x^2\beta_x^2p^4\eta^4C_g^6C_f^4+216\alpha^2p^2\eta^2C_g^4\sigma_g^2+2592\alpha^2\gamma_x^2\beta_x^2p^6\eta^6C_g^8C_f^2 \\
&\quad+\frac{27648\beta_y^2\gamma_y^2\alpha^2p^4\eta^4C_g^4L_f^4\sigma_g^2}{\beta_x^2}+\frac{331776\beta_y^2\gamma_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^8C_g^8C_f^2L_f^4}{\beta_x^2}+\frac{786\beta_y^2\gamma_y^2p^2\eta^2C_g^2L_f^2\sigma_f^2}{\beta_x^2} \\
&\quad+\frac{2\beta_x\eta C_f^2\sigma_{g'}^2}{K}+\frac{2\beta_x\eta C_g^2\sigma_f^2}{K}+\frac{72\gamma_x^2\beta_x^2p^2\eta^2C_g^6C_f^2}{\alpha}+\frac{6\alpha^2\eta C_g^2\sigma_g^2}{K}\ ,
\end{aligned}
\tag{70}
$$

where the second step holds due to the following inequality:

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^{(k)}(g^{(k)}(\mathbf{x}_0^{(k)}),\mathbf{y}_0^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\mathbf{u}_0^{(k)}\Big\|^2\Big] \\
&=\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_x g^{(k)}(x^{(k)})^T\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_0^{(k)}),\mathbf{y}_0^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\nabla g^{(k)}(x_0^{(k)};\xi_0^{(k)})^T\nabla_g f^{(k)}(h_0^{(k)},y_0^{(k)};\zeta_0^{(k)})\Big\|^2\Big] \\
&=\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_x g^{(k)}(x_0^{(k)})^T\nabla_g f^{(k)}(g^{(k)}(\mathbf{x}_0^{(k)}),\mathbf{y}_0^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\nabla g^{(k)}(x_0^{(k)})^T\nabla_g f^{(k)}(h_0^{(k)}),y_0^{(k)}) \\
&\quad+\frac{1}{K}\sum_{k=1}^{K}\nabla g^{(k)}(x_0^{(k)})^T\nabla_g f^{(k)}(h_0^{(k)}),y_0^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\nabla g^{(k)}(x_0^{(k)};\xi_0^{(k)})^T\nabla_g f^{(k)}(h_0^{(k)}),y_0^{(k)}) \\
&\quad+\frac{1}{K}\sum_{k=1}^{K}\nabla g^{(k)}(x_0^{(k)};\xi_0^{(k)})^T\nabla_g f^{(k)}(h_0^{(k)}),y_0^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\nabla g^{(k)}(x_0^{(k)};\xi_0^{(k)})^T\nabla_g f^{(k)}(h_0^{(k)},y_0^{(k)};\zeta_0^{(k)})\Big\|^2\Big] \\
&\leq 3C_g^2L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g^{(k)}(x_0^{(k)})-h_0^{(k)}\Big\|^2\Big]+3C_f^2\sigma_{g'}^2+3C_g^2\sigma_f^2 \\
&\leq 3C_g^2L_f^2\sigma_g^2+3C_f^2\sigma_{g'}^2+3C_g^2\sigma_f^2\ .
\end{aligned}
\tag{71}
$$

□

Lemma 14. *Given Assumption 1-4, $\beta_y\eta \in (0,1)$, and $\eta < 1$, we can get*

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathbf{v}_t^{(k)}\Big\|^2\Big] \\
&\leq \frac{2L_f^2\sigma_g^2 + 2\sigma_f^2}{\beta_y\eta T} + 6L_f^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_t - \frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_t^{(k)})\Big\|^2\Big] \\
&+ \frac{12L_f^2\gamma_x^2C_g^2}{\alpha}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2] + \frac{4\gamma_x^2L_f^2C_g^2}{\beta_y^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2] + \frac{4\gamma_y^2L_f^2}{\beta_y^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{v}}_t\|^2] \\
&+ 144\gamma_x^2\beta_x^2p^4\eta^4C_g^4C_f^2L_f^2 + 216\alpha^2p^2\eta^2C_g^2L_f^2\sigma_g^2 + 2592\alpha^2\gamma_x^2\beta_x^2p^6\eta^6C_g^6C_f^2L_f^2 \\
&+ 13824\gamma_y^2\alpha^2p^4\eta^4C_g^2L_f^4\sigma_g^2 + 165888\gamma_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^8C_g^6C_f^2L_f^4 + 384\gamma_y^2p^2\eta^2L_f^2\sigma_f^2 \\
&+ \frac{72\gamma_x^2\beta_x^2p^2\eta^2C_g^4C_f^2L_f^2}{\alpha} + \frac{6\alpha^2\eta\sigma_g^2L_f^2}{K} + \frac{\beta_y\eta\sigma_f^2}{K} + \frac{24\gamma_x^2p^2\beta_x^2\eta^2L_f^2C_g^4C_f^2}{\beta_y^2}\,.
\end{aligned} \tag{72}
$$

Proof.

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathbf{v}_t^{(k)}\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|(1-\beta_y\eta)\Big(\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathbf{v}_{t-1}^{(k)}\Big) \\
&\quad + (1-\beta_y\eta)\Big(\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})\Big) \\
&\quad + \beta_y\eta\Big(\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)}) \\
&\qquad + \frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)};\zeta_t^{(k)})\Big)\Big\|^2\Big] \\
&\leq (1-\beta_y\eta)^2(1+a)\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathbf{v}_{t-1}^{(k)}\Big\|^2\Big] \\
&\quad + 2(1-\beta_y\eta)^2(1+\frac{1}{a})\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})\Big\|^2\Big] \\
&\quad + 2\beta_y^2\eta^2(1+\frac{1}{a})\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\quad + \beta_y^2\eta^2\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)};\zeta_t^{(k)})\Big\|^2\Big] \\
&\leq (1-\beta_y\eta)\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathbf{v}_{t-1}^{(k)}\Big\|^2\Big] \\
&\quad + \frac{2}{\beta_y\eta}\underbrace{\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})\Big\|^2\Big]}_{T_1} \\
&\quad + 2\beta_y\eta\underbrace{\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(\mathbf{h}_t^{(k)},\mathbf{y}_t^{(k)})\Big\|^2\Big]}_{T_2} + \frac{\beta_y^2\eta^2\sigma_f^2}{K}\,.
\end{aligned} \tag{73}
$$

Then, for T_1, we can get

$$
\begin{aligned}
T_1 &\leq L_f^2 \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g^{(k)}(\mathbf{x}_t^{(k)}) - g^{(k)}(\mathbf{x}_{t-1}^{(k)})\Big\|^2\Big] + L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{y}_t^{(k)} - \mathbf{y}_{t-1}^{(k)}\Big\|^2\Big] \\
&\leq L_f^2 C_g^2 \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{x}_t^{(k)} - \mathbf{x}_{t-1}^{(k)}\Big\|^2\Big] + L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{y}_t^{(k)} - \mathbf{y}_{t-1}^{(k)}\Big\|^2\Big] \\
&\leq 2\gamma_x^2\eta^2 L_f^2 C_g^2 \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{u}_{t-1}^{(k)} - \bar{\mathbf{u}}_{t-1}\Big\|^2\Big] + 2\gamma_x^2\eta^2 L_f^2 C_g^2\mathbb{E}[\|\bar{\mathbf{u}}_{t-1}\|^2] \\
&\quad + 2\gamma_y^2\eta^2 L_f^2 \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\mathbf{v}_{t-1}^{(k)} - \bar{\mathbf{v}}_{t-1}\Big\|^2\Big] + 2\gamma_y^2\eta^2 L_f^2\mathbb{E}[\|\bar{\mathbf{v}}_{t-1}\|^2] , \\
&\leq 12\gamma_x^2 p^2\beta_x^2\eta^4 L_f^2 C_g^4 C_f^2 + 2\gamma_x^2\eta^2 L_f^2 C_g^2\mathbb{E}[\|\bar{\mathbf{u}}_{t-1}\|^2] \\
&\quad + 6912\beta_y^2\gamma_y^2\alpha^2 p^4\eta^6 C_g^2 L_f^4\sigma_g^2 + 82944\beta_y^2\gamma_y^2\alpha^2\gamma_x^2\beta_x^2 p^8\eta^{10} C_g^6 C_f^2 L_f^4 + 192\beta_y^2\gamma_y^2 p^2\eta^4 L_f^2\sigma_f^2 + 2\gamma_y^2\eta^4 L_f^2\mathbb{E}[\|\bar{\mathbf{v}}_{t-1}\|^2] ,
\end{aligned} \tag{74}
$$

where these inequalities hold due to Assumptions 1-2.

As for T_2, we can get

$$
\begin{aligned}
T_2 &= \mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_t^{(k')}), \mathbf{y}_t^{(k)}) \\
&\quad + \frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_t^{(k')}), \mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K}\mathbf{h}_t^{(k')}, \mathbf{y}_t^{(k)}) \\
&\quad + \frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K}\mathbf{h}_t^{(k')}, \mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(\mathbf{h}_t^{(k)}, \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\leq 3\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_t^{(k')}), \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\quad + 3\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_t^{(k')}), \mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K}\mathbf{h}_t^{(k')}, \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\quad + 3\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(\frac{1}{K}\sum_{k'=1}^{K}\mathbf{h}_t^{(k')}, \mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(\mathbf{h}_t^{(k)}, \mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\leq 3L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g^{(k)}(\mathbf{x}_t^{(k)}) - \frac{1}{K}\sum_{k'=1}^{K} g^{(k')}(\mathbf{x}_t^{(k')})\Big\|^2\Big] + 3L_f^2\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K} g^{(k)}(\mathbf{x}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathbf{h}_t^{(k)}\Big\|^2\Big] \\
&\quad + 3L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k'=1}^{K}\mathbf{h}_t^{(k')} - \mathbf{h}_t^{(k)}\Big\|^2\Big] \\
&\leq 72\gamma_x^2\beta_x^2 p^4\eta^4 C_g^4 C_f^2 L_f^2 + 108\alpha^2 p^2\eta^2 C_g^2 L_f^2\sigma_g^2 + 1296\alpha^2\gamma_x^2\beta_x^2 p^6\eta^6 C_g^6 C_f^2 L_f^2 \\
&\quad + 3L_f^2\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_{t-1} - \frac{1}{K}\sum_{k=1}^{K} g^{(k)}(\mathbf{x}_{t-1}^{(k)})\Big\|^2\Big] + \frac{6\eta L_f^2\gamma_x^2 C_g^2}{\alpha}\mathbb{E}[\|\bar{\mathbf{u}}_{t-1}\|^2] + \frac{36\gamma_x^2\beta_x^2 p^2\eta^3 C_g^4 C_f^2 L_f^2}{\alpha} + \frac{3\alpha^2\eta^2\sigma_g^2 L_f^2}{K} \\
&\leq 72\gamma_x^2\beta_x^2 p^4\eta^4 C_g^4 C_f^2 L_f^2 + 108\alpha^2 p^2\eta^2 C_g^2 L_f^2\sigma_g^2 + 1296\alpha^2\gamma_x^2\beta_x^2 p^6\eta^6 C_g^6 C_f^2 L_f^2 \\
&\quad + 3L_f^2\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_{t-1} - \frac{1}{K}\sum_{k=1}^{K} g^{(k)}(\mathbf{x}_{t-1}^{(k)})\Big\|^2\Big] + \frac{6L_f^2\gamma_x^2 C_g^2}{\alpha}\mathbb{E}[\|\bar{\mathbf{u}}_{t-1}\|^2] + \frac{36\gamma_x^2\beta_x^2 p^2\eta^2 C_g^4 C_f^2 L_f^2}{\alpha} + \frac{3\alpha^2\eta\sigma_g^2 L_f^2}{K} ,
\end{aligned} \tag{75}
$$

where the last step holds due to Lemmas 8, 9, and Eq. (56), the last step holds due to $\eta < 1$.

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

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\mathbf{v}_t^{(k)}\Big\|^2\Big]\\
&\leq (1-\beta_y\eta)\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_{t-1}^{(k)}),\mathbf{y}_{t-1}^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\mathbf{v}_{t-1}^{(k)}\Big\|^2\Big]+6\beta_y\eta L_f^2\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_{t-1}-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_{t-1}^{(k)})\Big\|^2\Big]\\
&\quad+\frac{12\beta_y\eta L_f^2\gamma_x^2C_g^2}{\alpha}\mathbb{E}[\|\bar{\mathbf{u}}_{t-1}\|^2]+\frac{4\gamma_x^2\eta L_f^2C_g^2}{\beta_y}\mathbb{E}[\|\bar{\mathbf{u}}_{t-1}\|^2]+\frac{4\gamma_y^2\eta L_f^2}{\beta_y}\mathbb{E}[\|\bar{\mathbf{v}}_{t-1}\|^2]\\
&\quad+144\beta_y\gamma_x^2\beta_x^2p^4\eta^5C_g^4C_f^2L_f^2+216\beta_y\alpha^2p^2\eta^3C_g^2L_f^2\sigma_g^2+2592\beta_y\alpha^2\gamma_x^2\beta_x^2p^6\eta^7C_g^6C_f^2L_f^2\\
&\quad+13824\beta_y\gamma_y^2\alpha^2p^4\eta^5C_g^2L_f^4\sigma_g^2+165888\beta_y\gamma_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^9C_g^6C_f^2L_f^4+384\beta_y\gamma_y^2p^2\eta^3L_f^2\sigma_f^2\\
&\quad+\frac{72\beta_y\gamma_x^2\beta_x^2p^2\eta^3C_g^4C_f^2L_f^2}{\alpha}+\frac{6\beta_y\alpha^2\eta^2\sigma_g^2L_f^2}{K}+\frac{\beta_y^2\eta^2\sigma_f^2}{K}+\frac{24\gamma_x^2p^2\beta_x^2\eta^3L_f^2C_g^4C_f^2}{\beta_y}\,.
\end{aligned}
\tag{76}
$$

It can be reformulated as

$$
\begin{aligned}
&\beta_y\eta\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\mathbf{v}_t^{(k)}\Big\|^2\Big]\\
&\leq \mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\mathbf{v}_t^{(k)}\Big\|^2\Big]-\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_{t+1}^{(k)}),\mathbf{y}_{t+1}^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\mathbf{v}_{t+1}^{(k)}\Big\|^2\Big]\\
&\quad+6\beta_y\eta L_f^2\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_{t}-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_{t}^{(k)})\Big\|^2\Big]+\frac{12\beta_y\eta L_f^2\gamma_x^2C_g^2}{\alpha}\mathbb{E}[\|\bar{\mathbf{u}}_{t}\|^2]+\frac{4\gamma_x^2\eta L_f^2C_g^2}{\beta_y}\mathbb{E}[\|\bar{\mathbf{u}}_{t}\|^2]+\frac{4\gamma_y^2\eta L_f^2}{\beta_y}\mathbb{E}[\|\bar{\mathbf{v}}_{t}\|^2]\\
&\quad+144\beta_y\gamma_x^2\beta_x^2p^4\eta^5C_g^4C_f^2L_f^2+216\beta_y\alpha^2p^2\eta^3C_g^2L_f^2\sigma_g^2+2592\beta_y\alpha^2\gamma_x^2\beta_x^2p^6\eta^7C_g^6C_f^2L_f^2\\
&\quad+13824\beta_y\gamma_y^2\alpha^2p^4\eta^5C_g^2L_f^4\sigma_g^2+165888\beta_y\gamma_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^9C_g^6C_f^2L_f^4+384\beta_y\gamma_y^2p^2\eta^3L_f^2\sigma_f^2\\
&\quad+\frac{72\beta_y\gamma_x^2\beta_x^2p^2\eta^3C_g^4C_f^2L_f^2}{\alpha}+\frac{6\beta_y\alpha^2\eta^2\sigma_g^2L_f^2}{K}+\frac{\beta_y^2\eta^2\sigma_f^2}{K}+\frac{24\gamma_x^2p^2\beta_x^2\eta^3L_f^2C_g^4C_f^2}{\beta_y}\,.
\end{aligned}
\tag{77}
$$

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

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\mathbf{v}_t^{(k)}\Big\|^2\Big]\\
&\leq \frac{1}{\beta_y\eta T}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_0^{(k)}),\mathbf{y}_0^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\mathbf{v}_0^{(k)}\Big\|^2\Big]+6L_f^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_{t}-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_{t}^{(k)})\Big\|^2\Big]\\
&\quad+\frac{12L_f^2\gamma_x^2C_g^2}{\alpha}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_{t}\|^2]+\frac{4\gamma_x^2L_f^2C_g^2}{\beta_y^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_{t}\|^2]+\frac{4\gamma_y^2L_f^2}{\beta_y^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{v}}_{t}\|^2]\\
&\quad+144\gamma_x^2\beta_x^2p^4\eta^4C_g^4C_f^2L_f^2+216\alpha^2p^2\eta^2C_g^2L_f^2\sigma_g^2+2592\alpha^2\gamma_x^2\beta_x^2p^6\eta^6C_g^6C_f^2L_f^2\\
&\quad+13824\gamma_y^2\alpha^2p^4\eta^4C_g^2L_f^4\sigma_g^2+165888\gamma_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^8C_g^6C_f^2L_f^4+384\gamma_y^2p^2\eta^2L_f^2\sigma_f^2\\
&\quad+\frac{72\gamma_x^2\beta_x^2p^2\eta^2C_g^4C_f^2L_f^2}{\alpha}+\frac{6\alpha^2\eta\sigma_g^2L_f^2}{K}+\frac{\beta_y\eta\sigma_f^2}{K}+\frac{24\gamma_x^2p^2\beta_x^2\eta^2L_f^2C_g^4C_f^2}{\beta_y^2}\\
&\leq \frac{2L_f^2\sigma_g^2+2\sigma_f^2}{\beta_y\eta T}+6L_f^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\Big\|\bar{\mathbf{h}}_{t}-\frac{1}{K}\sum_{k=1}^{K}g^{(k)}(\mathbf{x}_{t}^{(k)})\Big\|^2\Big]\\
&\quad+\frac{12L_f^2\gamma_x^2C_g^2}{\alpha}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_{t}\|^2]+\frac{4\gamma_x^2L_f^2C_g^2}{\beta_y^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_{t}\|^2]+\frac{4\gamma_y^2L_f^2}{\beta_y^2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{v}}_{t}\|^2]\\
&\quad+144\gamma_x^2\beta_x^2p^4\eta^4C_g^4C_f^2L_f^2+216\alpha^2p^2\eta^2C_g^2L_f^2\sigma_g^2+2592\alpha^2\gamma_x^2\beta_x^2p^6\eta^6C_g^6C_f^2L_f^2\\
&\quad+13824\gamma_y^2\alpha^2p^4\eta^4C_g^2L_f^4\sigma_g^2+165888\gamma_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^8C_g^6C_f^2L_f^4+384\gamma_y^2p^2\eta^2L_f^2\sigma_f^2\\
&\quad+\frac{72\gamma_x^2\beta_x^2p^2\eta^2C_g^4C_f^2L_f^2}{\alpha}+\frac{6\alpha^2\eta\sigma_g^2L_f^2}{K}+\frac{\beta_y\eta\sigma_f^2}{K}+\frac{24\gamma_x^2p^2\beta_x^2\eta^2L_f^2C_g^4C_f^2}{\beta_y^2}\,.
\end{aligned}
\tag{78}
$$

In addition, we have

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_0^{(k)}),\mathbf{y}_0^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\mathbf{v}_0^{(k)}\Big\|^2\Big]\\
&=\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_0^{(k)}),\mathbf{y}_0^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_0^{(k)};\xi_0^{(k)}),\mathbf{y}_0^{(k)};\zeta_0^{(k)})\Big\|^2\Big]\\
&=\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_0^{(k)}),\mathbf{y}_0^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_0^{(k)};\xi_0^{(k)}),\mathbf{y}_0^{(k)})\\
&\quad+\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_0^{(k)};\xi_0^{(k)}),\mathbf{y}_0^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_0^{(k)};\xi_0^{(k)}),\mathbf{y}_0^{(k)};\zeta_0^{(k)})\Big\|^2\Big]\\
&\leq 2L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|g^{(k)}(\mathbf{x}_0^{(k)})-g^{(k)}(\mathbf{x}_0^{(k)};\xi_0^{(k)})\Big\|^2\Big]+2\sigma_f^2\\
&\leq 2L_f^2\sigma_g^2+2\sigma_f^2\ ,
\end{aligned}
\tag{79}
$$

which completes the proof. □

Lemma 15. *Given Assumption 1-4 and if $\gamma_y \leq \frac{1}{6L_f}$ and $\eta \leq 1$, we have*

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\|\bar{\mathbf{y}}_t-\mathbf{y}^*(\bar{\mathbf{x}}_t)\|^2 \leq \frac{4}{\eta\gamma_y\mu T}\|\bar{\mathbf{y}}_0-\mathbf{y}^*(\bar{\mathbf{x}}_0)\|^2-\frac{3\gamma_y}{\mu}\frac{1}{T}\sum_{t=0}^{T-1}[\|\bar{\mathbf{v}}_t\|^2]+\frac{50\gamma_x^2C_g^2L_f^2}{3\gamma_y^2\mu^4}\frac{1}{T}\sum_{t=0}^{T-1}[\|\bar{\mathbf{u}}_t\|^2]\\
&\quad+\frac{50}{\mu^2}\frac{1}{T}\sum_{t=0}^{T-1}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}\mathbf{v}_t^{(k)}\Big\|^2\Big]+\frac{300\gamma_x^2\beta_x^2p^4\eta^4C_g^4C_f^2L_f^2}{\mu^2}\\
&\quad+\frac{100}{\mu^2}(1728\gamma_y^2\beta_y^2\alpha^2p^6\eta^6C_g^2L_f^4\sigma_g^2+20736\gamma_y^2\beta_y^2\alpha^2\gamma_x^2\beta_x^2p^{10}\eta^{10}C_g^6C_f^2L_f^4+48\gamma_y^2\beta_y^2p^4\eta^4L_f^2\sigma_f^2)\ .
\end{aligned}
\tag{80}
$$

Proof.

$$
\begin{aligned}
&\|\bar{\mathbf{y}}_{t+1}-\mathbf{y}^*(\bar{\mathbf{x}}_{t+1})\|^2 \\
&\leq (1-\frac{\eta\gamma_y\mu}{4})\|\bar{\mathbf{y}}_t-\mathbf{y}^*(\bar{\mathbf{x}}_t)\|^2 - \frac{3\eta\gamma_y^2}{4}\|\bar{\mathbf{v}}_t\|^2 + \frac{25\eta\gamma_x^2C_g^2L_f^2}{6\gamma_y\mu^3}\|\bar{\mathbf{u}}_t\|^2 + \frac{25\eta\gamma_y}{6\mu}\|\nabla_y f(g(\bar{\mathbf{x}}_t),\bar{\mathbf{y}}_t)-\bar{\mathbf{v}}_t\|^2 \\
&\leq (1-\frac{\eta\gamma_y\mu}{4})\|\bar{\mathbf{y}}_t-\mathbf{y}^*(\bar{\mathbf{x}}_t)\|^2 - \frac{3\eta\gamma_y^2}{4}\|\bar{\mathbf{v}}_t\|^2 + \frac{25\eta\gamma_x^2C_g^2L_f^2}{6\gamma_y\mu^3}\|\bar{\mathbf{u}}_t\|^2 \\
&\quad + \frac{25\eta\gamma_y}{6\mu}\Big[\Big\|\frac{1}{K}\sum_{k=1}^K\nabla_y f^{(k)}(\frac{1}{K}\sum_{k'=1}^K g^{(k')}(\bar{\mathbf{x}}_t),\bar{\mathbf{y}}_t) - \frac{1}{K}\sum_{k=1}^K\nabla_y f^{(k)}(g^{(k)}(\bar{\mathbf{x}}_t),\mathbf{y}_t^{(k)}) \\
&\qquad + \frac{1}{K}\sum_{k=1}^K\nabla_y f^{(k)}(g^{(k)}(\bar{\mathbf{x}}_t),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) \\
&\qquad + \frac{1}{K}\sum_{k=1}^K\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K\mathbf{v}_t^{(k)}\Big\|^2\Big] \\
&\leq (1-\frac{\eta\gamma_y\mu}{4})\|\bar{\mathbf{y}}_t-\mathbf{y}^*(\bar{\mathbf{x}}_t)\|^2 - \frac{3\eta\gamma_y^2}{4}\|\bar{\mathbf{v}}_t\|^2 + \frac{25\eta\gamma_x^2C_g^2L_f^2}{6\gamma_y\mu^3}\|\bar{\mathbf{u}}_t\|^2 \\
&\quad + \frac{25\eta\gamma_y}{2\mu}\Big[\Big\|\frac{1}{K}\sum_{k=1}^K\nabla_y f^{(k)}(\frac{1}{K}\sum_{k'=1}^K g^{(k')}(\bar{\mathbf{x}}_t),\bar{\mathbf{y}}_t) - \frac{1}{K}\sum_{k=1}^K\nabla_y f^{(k)}(g^{(k)}(\bar{\mathbf{x}}_t),\mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\quad + \frac{25\eta\gamma_y}{2\mu}\Big[\Big\|\frac{1}{K}\sum_{k=1}^K\nabla_y f^{(k)}(g^{(k)}(\bar{\mathbf{x}}_t),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)})\Big\|^2\Big] \\
&\quad + \frac{25\eta\gamma_y}{2\mu}\Big[\Big\|\frac{1}{K}\sum_{k=1}^K\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K\mathbf{v}_t^{(k)}\Big\|^2\Big] \\
&\leq (1-\frac{\eta\gamma_y\mu}{4})\|\bar{\mathbf{y}}_t-\mathbf{y}^*(\bar{\mathbf{x}}_t)\|^2 - \frac{3\eta\gamma_y^2}{4}\|\bar{\mathbf{v}}_t\|^2 + \frac{25\eta\gamma_x^2C_g^2L_f^2}{6\gamma_y\mu^3}\|\bar{\mathbf{u}}_t\|^2 + \frac{25\eta\gamma_yL_f^2}{2\mu}\frac{1}{K}\sum_{k=1}^K\Big[\Big\|\bar{\mathbf{y}}_t-\mathbf{y}_t^{(k)}\Big\|^2\Big] \\
&\quad + \frac{25\eta\gamma_yL_f^2C_g^2}{2\mu}\frac{1}{K}\sum_{k=1}^K\Big[\Big\|\bar{\mathbf{x}}_t-\mathbf{x}_t^{(k)}\Big\|^2\Big] + \frac{25\eta\gamma_y}{2\mu}\Big[\Big\|\frac{1}{K}\sum_{k=1}^K\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K\mathbf{v}_t^{(k)}\Big\|^2\Big] \\
&\leq (1-\frac{\eta\gamma_y\mu}{4})[\|\bar{\mathbf{y}}_t-\mathbf{y}^*(\bar{\mathbf{x}}_t)\|^2] - \frac{3\eta\gamma_y^2}{4}[\|\bar{\mathbf{v}}_t\|^2] + \frac{25\eta\gamma_x^2C_g^2L_f^2}{6\gamma_y\mu^3}[\|\bar{\mathbf{u}}_t\|^2] \\
&\quad + \frac{25}{\mu}(1728\gamma_y^3\beta_y^2\alpha^2p^6\eta^7C_g^2L_f^4\sigma_g^2 + 20736\gamma_y^3\beta_y^2\alpha^2\gamma_x^2\beta_x^2p^{10}\eta^{11}C_g^6C_f^2L_f^4 + 48\gamma_y^3\beta_y^2p^4\eta^5L_f^2\sigma_f^2) \\
&\quad + \frac{75\gamma_y\gamma_x^2\beta_x^2p^4\eta^5C_g^4C_f^2L_f^2}{\mu} + \frac{25\eta\gamma_y}{2\mu}\Big[\Big\|\frac{1}{K}\sum_{k=1}^K\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K\mathbf{v}_t^{(k)}\Big\|^2\Big] ,
\end{aligned}
\tag{81}
$$

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 6, 7. We further reformulate it as follows:

$$
\begin{aligned}
&\frac{\eta\gamma_y\mu}{4}\|\bar{\mathbf{y}}_t-\mathbf{y}^*(\bar{\mathbf{x}}_t)\|^2 \leq \|\bar{\mathbf{y}}_t-\mathbf{y}^*(\bar{\mathbf{x}}_t)\|^2 - \|\bar{\mathbf{y}}_{t+1}-\mathbf{y}^*(\bar{\mathbf{x}}_{t+1})\|^2 - \frac{3\eta\gamma_y^2}{4}[\|\bar{\mathbf{v}}_t\|^2] + \frac{25\eta\gamma_x^2C_g^2L_f^2}{6\gamma_y\mu^3}[\|\bar{\mathbf{u}}_t\|^2] \\
&\quad + \frac{25\eta\gamma_y}{2\mu}\Big[\Big\|\frac{1}{K}\sum_{k=1}^K\nabla_y f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}),\mathbf{y}_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K\mathbf{v}_t^{(k)}\Big\|^2\Big] + \frac{75\gamma_y\gamma_x^2\beta_x^2p^4\eta^5C_g^4C_f^2L_f^2}{\mu} \\
&\quad + \frac{25}{\mu}(1728\gamma_y^3\beta_y^2\alpha^2p^6\eta^7C_g^2L_f^4\sigma_g^2 + 20736\gamma_y^3\beta_y^2\alpha^2\gamma_x^2\beta_x^2p^{10}\eta^{11}C_g^6C_f^2L_f^4 + 48\gamma_y^3\beta_y^2p^4\eta^5L_f^2\sigma_f^2) .
\end{aligned}
\tag{82}
$$

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

$$
\begin{aligned}
\frac{\gamma_x\eta}{2}\mathbb{E}[\|\nabla\Phi(\bar{\mathbf{x}}_t)\|^2] &\leq \mathbb{E}[\Phi(\bar{\mathbf{x}}_t)] - \mathbb{E}[\Phi(\bar{\mathbf{x}}_{t+1})] - \frac{\gamma_x\eta}{4}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2] + 3\gamma_x\eta C_g^2L_f^2\mathbb{E}[\|\mathbf{y}_*(\bar{\mathbf{x}}_t) - \bar{\mathbf{y}}_t\|^2] \\
&\quad + 12\gamma_x\eta(C_g^4L_f^2 + C_f^2L_g^2)\frac{1}{K}\sum_{k=1}^{K}\Big[\Big\|\bar{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\Big\|^2\Big] + 6\gamma_x\eta C_g^2L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\bar{\mathbf{y}}_t - \mathbf{y}_t^{(k)}\Big\|^2\Big] \\
&\quad + 3\gamma_x\eta\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{x}}f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)}) - \bar{\mathbf{u}}_t\Big\|^2\Big] .
\end{aligned}
\tag{83}
$$

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

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\bar{\mathbf{x}}_t)\|^2] &\leq \frac{2(\Phi(\bar{\mathbf{x}}_0) - \Phi(\bar{\mathbf{x}}_T)}{\gamma_x\eta T} - \frac{1}{2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2] \\
&\quad + 24(C_g^4L_f^2 + C_f^2L_g^2)\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\Big[\Big\|\bar{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\Big\|^2\Big] + 12C_g^2L_f^2\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|\bar{\mathbf{y}}_t - \mathbf{y}_t^{(k)}\Big\|^2\Big] \\
&\quad + 6\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\Big\|\frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{x}}f^{(k)}(g^{(k)}(\mathbf{x}_t^{(k)}), \mathbf{y}_t^{(k)}) - \bar{\mathbf{u}}_t\Big\|^2\Big] + 6C_g^2L_f^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{y}_*(\bar{\mathbf{x}}_t) - \bar{\mathbf{y}}_t\|^2] \\
&\leq \frac{2(\Phi(\bar{\mathbf{x}}_0) - \Phi(\bar{\mathbf{x}}_T)}{\gamma_x\eta T} + \frac{24C_g^2L_f^2}{\gamma_y\eta\mu T}\|\bar{\mathbf{y}}_0 - \mathbf{y}^*(\bar{\mathbf{x}}_0)\|^2 \\
&\quad + (\frac{300C_g^2L_f^2(1+6L_f^2)}{\mu^2}\frac{2\gamma_x^2C_g^2}{\alpha^2} + \frac{300C_g^2L_f^2}{\mu^2}\frac{4\gamma_x^2L_f^2C_g^2}{\beta_y^2} + \frac{300C_g^2L_f^2}{\mu^2}\frac{12L_f^2\gamma_x^2C_g^2}{\alpha} + \frac{100\gamma_x^2C_g^4L_f^4}{\gamma_y^2\mu^4} \\
&\qquad + \frac{72\gamma_x^2C_g^4}{\alpha} + \frac{48\gamma_x^2(C_g^4L_f^2 + C_f^2L_g^2)}{\beta_x^2} - \frac{1}{2})\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2] \\
&\quad + (\frac{300C_g^2L_f^2}{\mu^2}\frac{4\gamma_y^2L_f^2}{\beta_y^2} + \frac{48\gamma_y^2C_g^2L_f^2}{\beta_x^2} - \frac{18\gamma_yC_g^2L_f^2}{\mu})\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{v}}_t\|^2] \\
&\quad + \frac{300C_g^2L_f^2(1+6L_f^2)}{\mu^2}\frac{\sigma_g^2}{\alpha\eta TK} + \frac{300C_g^2L_f^2(1+6L_f^2)}{\mu^2}\frac{12\gamma_x^2\beta_x^2p^2\eta^2C_g^4C_f^2}{\alpha^2} + \frac{300C_g^2L_f^2(1+6L_f^2)}{\mu^2}\frac{\alpha\eta\sigma_g^2}{K} \\
&\quad + \frac{600(L_f^2\sigma_g^2 + \sigma_f^2)C_g^2L_f^2}{\beta_y\eta T\mu^2} + \frac{43200\gamma_x^2\beta_x^2p^4\eta^4C_g^6C_f^2L_f^4}{\mu^2} + \frac{64800\alpha^2p^2\eta^2C_g^4L_f^6\sigma_g^2}{\mu^2} + \frac{7776000\alpha^2\gamma_x^2\beta_x^2p^6\eta^6C_g^8C_f^2L_f^4}{\mu^2} \\
&\quad + \frac{4147200\gamma_y^2\alpha^2p^4\eta^4C_g^4L_f^6\sigma_g^2}{\mu^2} + \frac{49766400\gamma_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^8C_g^8C_f^2L_f^6}{\mu^2} + \frac{115200\gamma_y^2p^2\eta^2C_g^2L_f^4\sigma_f^2}{\mu^2} \\
&\quad + \frac{21600\gamma_x^2\beta_x^2p^2\eta^2C_g^6C_f^2L_f^4}{\alpha\mu^2} + \frac{1800\alpha^2\eta\sigma_g^2C_g^2L_f^4}{K\mu^2} + \frac{300\beta_y\eta\sigma_f^2C_g^2L_f^2}{K\mu^2} + \frac{7200\gamma_x^2p^2\beta_x^2\eta^2L_f^4C_g^6C_f^2}{\beta_y^2\mu^2} \\
&\quad + 144(C_g^4L_f^2 + C_f^2L_g^2)\gamma_x^2\beta_x^2p^4\eta^4C_g^2C_f^2 + 41472\gamma_y^2\beta_y^2\alpha^2p^6\eta^6C_g^4L_f^4\sigma_g^2 + 497664\gamma_y^2\beta_y^2\alpha^2\gamma_x^2\beta_x^2p^{10}\eta^{10}C_g^8C_f^2L_f^4 \\
&\quad + 1152\gamma_y^2\beta_y^2p^4\eta^4C_g^2L_f^2\sigma_f^2 + \frac{6(3C_g^2L_f^2\sigma_g^2 + 3C_f^2\sigma_{g'}^2 + 3C_g^2\sigma_f^2)}{\beta_x\eta T} + 288\gamma_x^2p^2\eta^2(C_g^4L_f^2 + C_f^2L_g^2)C_g^2C_f^2 + 864\gamma_x^2\beta_x^2p^4\eta^4C_g^6C_f^4 \\
&\quad + 1296\alpha^2p^2\eta^2C_g^4\sigma_g^2 + 15552\alpha^2\gamma_x^2\beta_x^2p^6\eta^6C_g^8C_f^2 + \frac{12\beta_x\eta C_f^2\sigma_{g'}^2}{K} + \frac{12\beta_x\eta C_g^2\sigma_f^2}{K} \\
&\quad + \frac{165888\beta_y^2\gamma_y^2\alpha^2p^4\eta^4C_g^4L_f^4\sigma_g^2}{\beta_x^2} + \frac{1990656\beta_y^2\gamma_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^8C_g^8C_f^2L_f^4}{\beta_x^2} + \frac{4716\beta_y^2\gamma_y^2p^2\eta^2C_g^2L_f^2\sigma_f^2}{\beta_x^2} \\
&\quad + \frac{432\gamma_x^2\beta_x^2p^2\eta^2C_g^6C_f^2}{\alpha} + \frac{36\alpha^2\eta C_g^2\sigma_g^2}{K} + \frac{1800\gamma_x^2\beta_x^2p^4\eta^4C_g^6C_f^2L_f^4}{\mu^2} \\
&\quad + \frac{600}{\mu^2}(1728\gamma_y^2\beta_y^2\alpha^2p^6\eta^6C_g^4L_f^6\sigma_g^2 + 20736\gamma_y^2\beta_y^2\alpha^2\gamma_x^2\beta_x^2p^{10}\eta^{10}C_g^8C_f^2L_f^6 + 48\gamma_y^2\beta_y^2p^4\eta^4C_g^2L_f^4\sigma_f^2) .
\end{aligned}
\tag{84}
$$

where the second step holds due to Lemmas 6, 7, 13, 14, 11, 15. Then, we enforce the coefficient of $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{u}}_t\|^2]$ to be non-positive in the following. In particular, it can be done by solving the following inequalities:

$$
\begin{aligned}
&\frac{100\gamma_x^2 C_g^4 L_f^4}{\gamma_y^2 \mu^4} - \frac{1}{2} \leq -\frac{1}{4} \,, \\
&\frac{300C_g^2 L_f^2(1+6L_f^2)}{\mu^2}\frac{2\gamma_x^2 C_g^2}{\alpha^2} + \frac{300C_g^2 L_f^2}{\mu^2}\frac{4\gamma_x^2 L_f^2 C_g^2}{\beta_y^2} + \frac{300C_g^2 L_f^2}{\mu^2}\frac{12L_f^2\gamma_x^2 C_g^2}{\alpha} + \frac{72\gamma_x^2 C_g^4}{\alpha} \\
&\quad + \frac{48\gamma_x^2(C_g^4 L_f^2 + C_f^2 L_g^2)}{\beta_x^2} - \frac{1}{4} \leq 0 \,.
\end{aligned} \tag{85}
$$

Furthermore, we have the following inequalities:

$$
\begin{aligned}
&\frac{100\gamma_x^2 C_g^4 L_f^4}{\gamma_y^2 \mu^4} - \frac{1}{2} \leq -\frac{1}{4} \,, \\
&\frac{\gamma_x^2}{\alpha^2}\frac{600C_g^4 L_f^2(1+6L_f^2)}{\mu^2} \leq \frac{1}{16} \,, \\
&\frac{\gamma_x^2}{\alpha}(\frac{3600C_g^4 L_f^4}{\mu^2} + 72C_g^4) \leq \frac{1}{16} \,, \\
&\frac{\gamma_x^2}{\beta_y^2}\frac{1200C_g^4 L_f^4}{\mu^2} \leq \frac{1}{16} \,, \\
&\frac{\gamma_x^2}{\beta_x^2}48(C_g^4 L_f^2 + C_f^2 L_g^2) \leq \frac{1}{16} \,,
\end{aligned} \tag{86}
$$

By solving these inequalities, we can get

$$
\gamma_x \leq \min\Big\{\frac{\gamma_y \mu^2}{20C_g^2 L_f^2}, \frac{\alpha\mu}{100C_g^2 L_f\sqrt{1+6L_f^2}}, \frac{\sqrt{\alpha}\mu}{24C_g\sqrt{100C_g^2 L_f^4 + 2C_g^2\mu^2}}, \frac{\beta_y\mu}{144C_g^2 L_f^2}, \frac{\beta_x}{32\sqrt{C_g^4 L_f^2 + C_f^2 L_g^2}}\Big\} \,. \tag{87}
$$

Similarly, we enforce the coefficient of $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{v}}_t\|^2]$ to be non-positive as follows:

$$
\frac{300C_g^2 L_f^2}{\mu^2}\frac{4\gamma_y^2 L_f^2}{\beta_y^2} + \frac{48\gamma_y^2 C_g^2 L_f^2}{\beta_x^2} - \frac{18\gamma_y C_g^2 L_f^2}{\mu} \leq 0 \,. \tag{88}
$$

Then, it can be done by solving the following inequalities:

$$
\begin{aligned}
&\frac{1200\gamma_y L_f^2}{\mu^2\beta_y^2} \leq \frac{9}{\mu} \,, \\
&\frac{48\gamma_y}{\beta_x^2} \leq \frac{9}{\mu} \,.
\end{aligned} \tag{89}
$$

Therefore, we can get

$$
\gamma_y \leq \min\Big\{\frac{3\mu\beta_y^2}{400L_f^2}, \frac{3\beta_x^2}{16\mu}\Big\} \,. \tag{90}
$$

As a result, by setting $\alpha > 0$, $\beta_x > 0$, $\beta_y > 0$, $\eta \leq \min\{\frac{1}{2\gamma_x L_\phi}, \frac{1}{\alpha}, \frac{1}{\beta_x}, \frac{1}{\beta_y}, 1\}$, and

$$
\begin{aligned}
&\gamma_x \leq \min\Big\{\frac{\gamma_y \mu^2}{20C_g^2 L_f^2}, \frac{\alpha\mu}{100C_g^2 L_f\sqrt{1+6L_f^2}}, \frac{\beta_x}{32\sqrt{C_g^4 L_f^2 + C_f^2 L_g^2}}, \frac{\beta_y\mu}{144C_g^2 L_f^2}, \frac{\sqrt{\alpha}\mu}{24C_g\sqrt{100C_g^2 L_f^4 + 2C_g^2\mu^2}}\Big\} \,, \\
&\gamma_y \leq \min\Big\{\frac{1}{6L_f}, \frac{3\mu\beta_y^2}{400L_f^2}, \frac{3\beta_x^2}{16\mu}\Big\} \,,
\end{aligned} \tag{91}
$$

we can get

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\bar{\mathbf{x}}_t)\|^2] \le \frac{2(\Phi(\bar{\mathbf{x}}_0)-\Phi(\bar{\mathbf{x}}_T)}{\gamma_x\eta T} + \frac{24C_g^2L_f^2}{\gamma_y\eta\mu T}\|\bar{\mathbf{y}}_0-\mathbf{y}^*(\bar{\mathbf{x}}_0)\|^2 \\
&+ \frac{300C_g^2L_f^2(1+6L_f^2)}{\mu^2}\frac{\sigma_g^2}{\alpha\eta TK} + \frac{300C_g^2L_f^2(1+6L_f^2)}{\mu^2}\frac{12\gamma_x^2\beta_x^2p^2\eta^2C_g^4C_f^2}{\alpha^2} + \frac{300C_g^2L_f^2(1+6L_f^2)}{\mu^2}\frac{\alpha\eta\sigma_g^2}{K} \\
&+ \frac{600(L_f^2\sigma_g^2+\sigma_f^2)C_g^2L_f^2}{\beta_y\eta T\mu^2} + \frac{43200\gamma_x^2\beta_x^2p^4\eta^4C_g^6C_f^2L_f^4}{\mu^2} + \frac{64800\alpha^2p^2\eta^2C_g^4L_f^6\sigma_g^2}{\mu^2} + \frac{7776000\alpha^2\gamma_x^2\beta_x^2p^6\eta^6C_g^8C_f^2L_f^4}{\mu^2} \\
&+ \frac{4147200\gamma_y^2\alpha^2p^4\eta^4C_g^4L_f^6\sigma_g^2}{\mu^2} + \frac{49766400\gamma_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^8C_g^8C_f^2L_f^6}{\mu^2} + \frac{115200\gamma_y^2p^2\eta^2C_g^2L_f^4\sigma_f^2}{\mu^2} \\
&+ \frac{21600\gamma_x^2\beta_x^2p^2\eta^2C_g^6C_f^2L_f^4}{\alpha\mu^2} + \frac{1800\alpha^2\eta\sigma_g^2C_g^2L_f^4}{K\mu^2} + \frac{300\beta_y\eta\sigma_f^2C_g^2L_f^2}{K\mu^2} + \frac{7200\gamma_x^2p^2\beta_x^2\eta^2L_f^4C_g^6C_f^2}{\beta_y^2\mu^2} \\
&+ 144(C_g^4L_f^2+C_f^2L_g^2)\gamma_x^2\beta_x^2p^4\eta^4C_g^2C_f^2 + 41472\gamma_y^2\beta_y^2\alpha^2p^6\eta^6C_g^4L_f^4\sigma_g^2 + 497664\gamma_y^2\beta_y^2\alpha^2\gamma_x^2\beta_x^2p^{10}\eta^{10}C_g^8C_f^2L_f^4 \\
&+ 1152\gamma_y^2\beta_y^2p^4\eta^4C_g^2L_f^2\sigma_f^2 + \frac{6(3C_g^2L_f^2\sigma_g^2+3C_f^2\sigma_{g'}^2+3C_g^2\sigma_f^2)}{\beta_x\eta T} + 288\gamma_x^2p^2\eta^2(C_g^4L_f^2+C_f^2L_g^2)C_g^2C_f^2 + 864\gamma_x^2\beta_x^2p^4\eta^4C_g^6C_f^4 \\
&+ 1296\alpha^2p^2\eta^2C_g^4\sigma_g^2 + 15552\alpha^2\gamma_x^2\beta_x^2p^6\eta^6C_g^8C_f^2 + \frac{12\beta_x\eta C_f^2\sigma_{g'}^2}{K} + \frac{12\beta_x\eta C_g^2\sigma_f^2}{K} \\
&+ \frac{165888\beta_y^2\gamma_y^2\alpha^2p^4\eta^4C_g^4L_f^4\sigma_g^2}{\beta_x^2} + \frac{1990656\beta_y^2\gamma_y^2\alpha^2\gamma_x^2\beta_x^2p^8\eta^8C_g^8C_f^2L_f^4}{\beta_x^2} + \frac{4716\beta_y^2\gamma_y^2p^2\eta^2C_g^2L_f^2\sigma_f^2}{\beta_x^2} \\
&+ \frac{432\gamma_x^2\beta_x^2p^2\eta^2C_g^6C_f^2}{\alpha} + \frac{36\alpha^2\eta C_g^2\sigma_g^2}{K} + \frac{1800\gamma_x^2\beta_x^2p^4\eta^4C_g^6C_f^2L_f^4}{\mu^2} \\
&+ \frac{600}{\mu^2}(1728\gamma_y^2\beta_y^2\alpha^2p^6\eta^6C_g^4L_f^6\sigma_g^2 + 20736\gamma_y^2\beta_y^2\alpha^2\gamma_x^2\beta_x^2p^{10}\eta^{10}C_g^8C_f^2L_f^6 + 48\gamma_y^2\beta_y^2p^4\eta^4C_g^2L_f^4\sigma_f^2)\,.
\end{aligned}
\tag{92}
$$

Since α, β_x, β_y, γ_x, and γ_y can be set as free hyperparameters, i.e., they are independent of the number of iterations, we can obtain

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\bar{\mathbf{x}}_t)\|^2] \le \frac{2(\Phi(\mathbf{x}_0)-\Phi(\mathbf{x}_*)}{\gamma_x\eta T} + \frac{24C_g^2L_f^2}{\gamma_y\eta\mu T}\|\mathbf{y}_0-\mathbf{y}^*(\mathbf{x}_0)\|^2 + O(\frac{\eta}{K}) + O(\frac{1}{\eta T}) \\
+ O(p^2\eta^2) + O(p^4\eta^4) + O(p^6\eta^6) + O(p^8\eta^8) + O(p^{10}\eta^{10})\,,
\end{aligned}
\tag{93}
$$

where $\mathbf{x}_*$ 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.

Methods	Hyperparameters	Value
LocalSGDM	learning rate	0.1
	momentum coefficient	0.1
CoDA	learning rate	0.1
LocalSGDAM	learning rate η	0.3
	learning rate coefficient γ_x and γ_y	0.33
	momentum coefficient β_x and β_y	3.3
LocalSCGDAM (Ours)	learning rate η	0.3
	learning rate coefficient γ_x and γ_y	0.33
	momentum coefficient β_x and β_y	3.3
	coefficient α	3.0

The classifier for FashionMNIST is summarized in Table 3.

Table 3: The classifier for FashionMNIST.

Layers	Operators	Configuration
Layer 1	CNN	output channels: 32
	Batchnorm	-
	ReLU	-
	Maxpooling	kernel size: 2, stride: 2
Layer 2	CNN	output channels: 64
	Batchnorm	-
	ReLU	-
	Maxpooling	kernel size: 2, stride: 2
Layer 3	FC	output features: 600
Layer 4	FC	output features: 120
Layer 5	FC	output features: 1

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

Dataset	Training set		Testing set	
	#pos	#neg	#pos	#neg
CIFAR10	2,777	25,000	5,000	5,000
CIFAR100	2,777	25,000	5,000	5,000
STL10	277	2,500	8,000	8,000
FashionMNIST	3,333	30,000	5,000	5,000
CATvsDOG	1,112	10,016	2,516	2,888
Melanoma	868	25,670	117	6,881