A Federated Stochastic Multi-Level Compositional Minimax Algorithm for Deep AUC Maximization

Xinwen Zhang [1] **Ali Payani** [2] **Myungjin Lee** [2] **Richard Souvenir** [1] **Hongchang Gao** [1]

Abstract

AUC maximization is an effective approach to address the imbalanced data classification problem in federated learning. In the past few years, a couple of federated AUC maximization approaches have been developed based on the minimax optimization. However, directly solving a minimax optimization problem to maximize the AUC score cannot achieve satisfactory performance. To address this issue, we propose to maximize AUC via optimizing a federated multi-level compositional minimax problem. Specifically, we develop a novel federated multi-level compositional minimax algorithm with rigorous theoretical guarantees to solve this new learning paradigm in both algorithmic design and theoretical analysis. To the best of our knowledge, this is the first work studying the multi-level minimax optimization problem. Additionally, extensive empirical evaluations confirm the efficacy of our proposed approach.

1. Introduction

Federated AUC maximization has attracted increasing attention in federated learning (FL) due to its efficacy in handling imbalanced data classification. In FL, data are distributed across various devices, and information is periodically exchanged with a central server. Unlike most traditional heterogeneous federated learning approaches, which assume that **the local data distribution is imbalanced but the global one is balanced**, e.g., Scaffold (Karimireddy et al., 2020), federated AUC maximization approaches can handle the case where **both local and global data distributions are imbalanced**. Thus, federated AUC maximization approaches have demonstrated superior performance in practical applications, such as disease prediction, where the global distribution is imbalanced.

The traditional AUC maximization approach on a single machine depends on the sum of pairwise losses between examples from different classes, which requires storing all training data, making it impractical for large-scale online learning. To address this scalability issue, (Ying et al., 2016) successfully reformulated the original problem into a minimax decomposable formulation, allowing for stochastic algorithms based on mini-batch data without explicitly creating the pairs. Based on this minimax learning paradigm, (Guo et al., 2020) developed the federated AUC maximization approach by updating the primal-dual variables locally and averaging the global variable periodically. However, directly optimizing AUC can result in suboptimal results. Particularly in the early stages of training, updating the AUC loss cannot always lead to effective feature extraction when compared to updating the traditional cross-entropy loss (Yuan et al., 2021c). To overcome this limitation, (Yuan et al., 2021a) designed an end-to-end deep AUC maximization with the compositional framework, ensuring not only robust feature learning of lower layers through the minimization of the standard cross-entropy loss, but also robust classifier learning through the optimization of an AUC loss. Furthermore, (Zhang et al., 2023) developed a federated compositional gradient descent ascent algorithm to enable this compositional framework for federated learning, which has demonstrated superior performance over traditional federated AUC maximization approaches (Guo et al., 2020; Yuan et al., 2021b; Sharma et al., 2022).

However, one limitation of (Zhang et al., 2023) is that it can only handle the two-level compositional optimization problem, whose performance is not satisfactory. In particular, our preliminary study depicted in Figure 1 shows that **a multi-level compositional learning paradigm**, i.e., LocalSMCGDAM in Algorithm 1, where the weights of the backbone neural network are updated multiple times

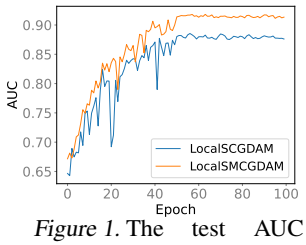


Figure 1. The test AUC score of CATvsDOG. The imbalance ratio is 0.05, the number of devices is eight, and the communication period is four.

[1]Department of Computer and Information Sciences, Temple University, Philadelphia, USA [2]Cisco Systems Inc.. Correspondence to: Hongchang Gao <hongchang.gao@temple.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

before optimizing the AUC loss, can achieve much better performance than the **two-level compositional learning paradigm**, i.e., LocalSCGDAM (Zhang et al., 2023), where the weights are only updated once.

Given the outstanding performance of the multi-level compositional learning paradigm, an intriguing question naturally arises: **How to enable the multi-level compositional minimax problem for federated AUC maximization with rigorous theoretical guarantees?** In this paper, we focus on the *federated multi-level compositional minimax optimization algorithm* for deep AUC maximization to address the unique computation and communication challenges.

The stochastic multi-level compositional optimization problem finds applications in many machine learning tasks, such as multi-step model-agnostic meta-learning (Finn et al., 2017), risk-averse portfolio optimization (Shapiro et al., 2021), and graph neural networks (Yu et al., 2022). Several works (Yang et al., 2019; Balasubramanian et al., 2022; Jiang et al., 2022) have provided insights into developing stochastic multi-level compositional algorithms under the single-machine setting. Moreover, (Gao, 2024) developed the distributed multi-level compositional optimization algorithm, which employed a decentralized communication manner. Nevertheless, all of these efforts only address the multi-level compositional *minimization* problem. Remarkably, the federated multi-level compositional *minimax* optimization remains unexplored in the existing literature.

The key challenge of the federated multi-level compositional minimax optimization problem lies in that *the multi-level compositional structure* coupled with *the distributed inner-level functions* leads to the biased stochastic gradient issue for *both primal and dual variables*, which is more challenging to solve than both the single-machine setting and minimization setting. As a result, from the view of algorithmic design, it remains unclear how to perform local updates regarding the multi-level inner functions and how to communicate them across different devices to alleviate the issue caused by the biased gradient. In particular, *it is unclear whether the additional bias caused by the multi-level compositional structure for the dual variable can be mitigated to achieve the level-independent sample complexity as the minimization algorithm.* In this study, a *level-dependent* convergence rate implies that the number of levels, represented by K, affects the convergence rate exponentially, such as $O(\epsilon^{-K})$. A *level-independent* convergence rate, on the other hand, means that K has no effect on the order of the convergence rate but may influence its coefficient, such as $O(K\epsilon^{-2})$. Therefore, from the point of view of theoretical analysis, it motivates us to investigate how this multi-level structure affects the convergence rate, particularly regarding both the primal and dual variables.

In response to these challenges, we develop a novel algorithm, named Local Stochastic Multi-Level Compositional Gradient Descent Ascent with Momentum (LocalSMCGDAM), for federated AUC maximization, which not only handles the difficulties of updating a biased multi-level compositional gradient, but it also obtains a level-independent convergence rate. In particular, given the non-convex strongly-concave problem, under our novel theoretical analysis, our algorithm achieves $O(K^2/\sqrt{NT})$ convergence rate, suggesting a linear speedup with respect to the number of devices N, while the number of levels K does not affect the order of the convergence rate.

In summary, this article makes the following significant contributions:

- We develop a novel federated stochastic multi-level compositional minimax framework for deep AUC maximization named LocalSMCGDAM, showing how to conduct local updates and global communication to mitigate the biased gradient issue. To the best of our knowledge, this is the first federated stochastic multi-level compositional minimax algorithm for deep AUC maximization.
- We investigate the theoretical convergence rate of our novel algorithm, which can achieve linear speedup and level-independent complexity. To be more specific, our LocalSMCGDAM can achieve the $O(1/\epsilon^4)$ sample complexity and $O(1/\epsilon^3)$ communication complexity, which is *the first time establishing the convergence rate for the multi-level compositional minimax problem.*
- We conduct extensive experiments on highly imbalanced benchmark datasets to validate the effectiveness of our proposed algorithm.

2. Related Work

2.1. Federated Deep AUC Maximization

AUC emerges as a robust metric (Elkan, 2001), especially in the face of imbalanced data, leveraging its capacity to capture the nuanced performance of models in situations where sparse but critical classifications carry significance. In the realm of single-machine settings, numerous efforts have been made to explore the complexities of AUC, ranging from reformulating it as a minimax problem (Ying et al., 2016) to introducing compositional frameworks for end-to-end DAM (Yuan et al., 2021a). (Ying et al., 2016) considered optimizing the pairwise AUC squared loss as an equivalent min-max optimization problem, successfully transforming the original non-decomposable objective into a decomposable objective over individual examples. Subsequently, other studies (Liu et al., 2019; Yuan et al., 2020) create a variety of realistic and verifiable stochastic algorithms for this min-max formulation in deep AUC maximization. However, due to deteriorated feature representations, directly

optimizing the AUC loss for training a deep neural network may not work effectively (Cao et al., 2019). Prior methods (Kang et al., 2019; Jamal et al., 2020) concentrated on two stages and found that fine-tuning some of the higher layers after the classifier layer in the second stage was advantageous. However, deciding when to move on to the second phase remains a challenge. Later, (Yuan et al., 2021a) suggested an end-to-end compositional framework to not only extract robust model features, but also improve the overall classification performance.

CoDA (Guo et al., 2020) is the first study to solve the deep AUC maximization problem in a distributed, communication-efficient manner, employing a proximal-point algorithmic technique to approximate solve the federated non-convex concave problem and achieving linear speedup. A couple of methods (Yuan et al., 2021b; Guo et al., 2023) were later proposed to further improve it under different settings. Meanwhile, there are some generic federated minimax optimization algorithms (Deng & Mahdavi, 2021; Sharma et al., 2022; Yang et al., 2022; Wu et al., 2024) that can be leveraged to solve the AUC maximization problem. More recently, (Zhang et al., 2023) introduced LocalSCGDAM, a method that combines the compositional framework with deep AUC maximization, highlighting the superiority of the compositional framework in handling highly imbalanced data and showcasing its potential to improve classification performance. However, a significant gap in the existing research remains unfilled: the investigation of multi-level compositional frameworks in federated conditional learning. The question is whether employing such multi-level techniques will result in additional improvements in imbalanced classification inside the federated setting.

2.2. Stochastic Compositional Optimization

Two-Level Stochastic Compositional Optimization Problems. The primary challenge to the stochastic compositional problem is that the standard stochastic gradient introduces bias in estimating the full gradient when the outer-level function is nonlinear. To overcome the problem of biased gradient estimation, many stochastic compositional gradient descent algorithms (Wang et al., 2017; Ghadimi et al., 2020) use the moving average technique to estimate the inner-level function, which provides a mechanism for regulating estimation errors. Furthermore, it has been demonstrated that using a variance-reduced estimator for both the inner-level function and its gradient improves the overall efficacy of stochastic compositional optimization (Yuan & Hu, 2020). While previous studies primarily focused on the single-machine scenario, (Gao et al., 2022) and (Zhang et al., 2023) extended the use of the moving average technique to handle the estimate of inner-level functions in the minimization and minimax problem inside federated compositional learning. Another investigation into federated nested learning is presented in (Tarzanagh et al., 2022). However, their focus is solely on the federated compositional problem and the federated minimax problem independently, considering only a two-level compositional framework.

Multi-Level Stochastic Compositional Optimization Problems. Even though most earlier algorithms were built within the limits of a basic two-level compositional framework, (Yang et al., 2019) introduced multi-level compositional optimization. In reality, there are numerous multi-level stochastic compositional learning applications, including multi-step model-agnostic meta-learning (Finn et al., 2017) and stochastic training of graph neural networks (Yu et al., 2022). However, the convergence rate of the earlier multi-level compositional optimization algorithm is exponentially affected by the number of function levels K. To be more specific, the algorithm proposed by (Yang et al., 2019) can only achieve the sample complexity of $O(\epsilon^{-(7+K)/2})$. Later, both the moving-average technique and variance-reduced techniques (Cutkosky & Orabona, 2019; Tran-Dinh et al., 2022) were employed for both each level function and its gradient to alleviate the exponential dependence issue (Balasubramanian et al., 2022; Zhang & Xiao, 2021; Jiang et al., 2022). For example, (Balasubramanian et al., 2022; Chen et al., 2021) can achieve the level-independent sample complexity of $O(\epsilon^{-4})$. However, it is noteworthy that these algorithms were primarily designed for the traditional *minimization* problem under a single-machine setting. No existing works have studied the stochastic multi-level compositional *minimax* problem under the single-machine or federated-learning setting.

3. Federated Multi-Level Compositional Minimax Optimization

In this section, we present the details of our federated multi-level compositional minimax optimization algorithm.

3.1. Preliminaries

Following (Zhang et al., 2023), the deep AUC maximization problem under the federated learning setting can be formulated as the following federated two-level compositional minimax problem:

$$\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathbb{R}^{d_2}} \frac{1}{N}\sum_{n=1}^{N} f_n\Big(\frac{1}{N}\sum_{n'=1}^{N} g_{n'}(x;\xi), y; \zeta\Big)\,, \tag{1}$$

where N is the number of devices. Generally speaking, the inner-level function $g_n(\cdot)$ is to *update the classifier parameters by minimizing the cross-entropy loss function with gradient descent*, and the outer-level function $f_n(\cdot,\cdot)$ is to *update the classifier parameters via optimizing the minimax AUC loss function*. Specifically, by denoting

$x = [w^T, \tilde{w}_1, \tilde{w}_2]^T$, where $w \in \mathbb{R}^d$ denotes the classifier's parameter, $\tilde{w}_1 \in \mathbb{R}$ and $\tilde{w}_2 \in \mathbb{R}$ are two parameters for computing the minimax AUC loss function, and the cross-entropy loss function $\mathcal{L}_n(w;\xi)$, the inner-level function on the n-th device is defined as $g_n(x;\xi) = x - \rho\Delta_n(x;\xi)$, where $\Delta_n(x;\xi) = [\nabla_w \mathcal{L}_n(w;\xi)^T, 0, 0]^T$, $\rho > 0$ is a constant, and ξ denotes random samples. The outer-level function $f_n(x, y; \zeta)$ is the minimax AUC loss function, which is a nonconvex-strongly-concave function (Ying et al., 2016; Yuan et al., 2021a). Specifically, $y \in \mathbb{R}$ is the parameter for computing the minimax AUC loss function, and ζ denotes the random samples. More details about the compositional AUC loss function are deferred to Appendix A.

3.2. Problem Definition

From Eq. (1), it can be observed that the inner-level function performs one-step gradient descent to minimize the cross-entropy loss function $\mathcal{L}_n(w;\xi)$. However, our preliminary study shows that performing multiple gradient descent steps can achieve a much better AUC score. This kind of multi-step update leads to a **multi-level** compositional minimax problem, rather than a two-level compositional minimax problem as Eq. (1). As such, the existing approach (Zhang et al., 2023) cannot be applied to the multi-level case. Specifically, it is not clear how to *compute* the multi-level inner functions within each device and how to *communicate* them across different devices. Furthermore, it is unclear how the number of levels affects the convergence rate in the federated learning setting.

This motivates us to develop a novel federated multi-level compositional minimax algorithm with theoretical guarantees to solve the following problem:

$$\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathbb{R}^{d_2}} \frac{1}{N}\sum_{n=1}^{N} f_n\Big(\frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(K)}(\cdots(g_{n'}^{(1)}(x))\cdots), y\Big) . \tag{2}$$

Here, $f_n(\cdot,\cdot) = \mathbb{E}[f_n(\cdot,\cdot;\zeta_n)]$ denotes the outermost level function on the n-th device where ζ_n denotes the data distribution on the n-th device, which actually is a nonconvex-strongly-concave AUC loss function. $G(x) \triangleq \frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(K)}(\cdots(g_{n'}^{(1)}(x))\cdots)$ is a K-level compositional function on the n-th device, which represents the multi-step gradient descent for minimizing the cross-entropy loss function. Specifically, for $k \in \{1,\cdots,K\}$, the k-th inner-level function is defined as follows:

$$g_n^{(k)}(\cdot) \triangleq \begin{cases} \mathbb{E}[g_n^{(1)}(x;\xi_n^{(1)})] = \mathbb{E}[x - \eta'\Delta_n(x;\xi_n^{(1)})], & k = 1, \\ \mathbb{E}[g_n^{(k)}(z;\xi_n^{(k)})] = \mathbb{E}[z - \eta'\Delta_n(z;\xi_n^{(k)})], & k \neq 1, \end{cases} \tag{3}$$

where $z \triangleq g_n^{(k-1)}(\cdot)$ for $k \in \{2,\cdots,K\}$, and $\xi_n^{(k)}$ denotes the data distribution for the k-th level function on the n-th worker, η' is the learning rate for inner-level functions. From Eq. (3), it can be observed that K-step gradient descent for minimizing the cross-entropy loss function leads to a K-level inner function $G(\cdot)$.

To demonstrate the challenges when optimizing the federated stochastic multi-level compositional minimax problem in Eq. (2), for $k \in \{1,\cdots,K\}$, we first introduce the following terminologies:

$$G_n^{(k)}(x) = \begin{cases} g_n^{(1)}(x), & k = 1, \\ g_n^{(k)}(G_n^{(k-1)}(x)), & k \neq 1. \end{cases} \tag{4}$$

It is easy to know that $G(x) = \frac{1}{N}\sum_{n=1}^{N} G_n^{(K)}(x)$. The gradient of $G_n^{(k)}(x)$ can be represented as follows:

$$\nabla G_n^{(k)}(x) = \begin{cases} \nabla g_n^{(1)}(x), & k = 1, \\ \nabla G_n^{(k-1)}(x)\nabla g_n^{(k)}(G_n^{(k-1)}(x)), & k \neq 1. \end{cases} \tag{5}$$

Similarly, we can know $\nabla G(x) = \frac{1}{N}\sum_{n=1}^{N} \nabla G_n^{(K)}(x)$.

Key Challenges. On the n-th device, the gradient regarding the primal variable x and the dual variable y is defined as follows:

$$\begin{aligned} \nabla_x f_n &= \Big(\frac{1}{N}\sum_{n'=1}^{N} \nabla G_{n'}^{(K)}(x)\Big)\nabla_G f_n\Big(\frac{1}{N}\sum_{n'=1}^{N} G_{n'}^{(K)}(x), y\Big) , \\ \nabla_y f_n &= \nabla_y f_n\Big(\frac{1}{N}\sum_{n'=1}^{N} G_{n'}^{(K)}(x), y\Big) . \end{aligned} \tag{6}$$

It can be observed that both $\nabla_x f_n$ and $\nabla_y f_n$ depend on the global K-level function $\frac{1}{N}\sum_{n'=1}^{N} G_{n'}^{(K)}(x)$ and $\nabla_x f_n$ depends on the gradient of the K-level compositional function: $\frac{1}{N}\sum_{n'=1}^{N} \nabla G_{n'}^{(K)}(x)$. This dependence leads to unique challenges when computing gradients on each device.

First, **the dependence on $\frac{1}{N}\sum_{n'=1}^{N} G_{n'}^{(K)}(x)$ introduces biases on each device**. In particular, for any $k \in \{2,\cdots,K\}$ and any device $n \in \{1,\cdots,N\}$, according to Eq. (4), when using random samples to compute the k-th level function, there exist biases as follows:

$$\mathbb{E}[g_n^{(k)}(g_n^{(k-1)}(\cdot;\xi_n^{(k-1)});\xi_n^{(k)})] \neq G_n^{(k)}(\cdot) , \tag{7}$$

where $g_n^{(k)}$ is non-linear for any k. This leads to biased gradient estimators for both $\nabla_x f_n$ and $\nabla_y f_n$.

Second, **the dependence on $\frac{1}{N}\sum_{n'=1}^{N} \nabla G_{n'}^{(K)}(x)$ introduces biases on each device**. Specifically, for any $k \in \{2,\cdots,K\}$ and any device $n \in \{1,\cdots,N\}$, according to Eq. (5), using random samples to compute the gradient of the k-th level function introduces biases as follows:

$$\begin{aligned} &\mathbb{E}[\nabla g_n^{(k-1)}(\cdot;\xi_n^{(k-1)})\nabla_g g_n^{(k)}(g_n^{(k-1)}(\cdot;\xi_n^{(k-1)});\xi_n^{(k)})] \\ &\neq \nabla G_n^{(k)}(\cdot) , \end{aligned} \tag{8}$$

which introduces additional biases when estimating $\nabla_x f_n$.

In summary, the *multi-level structure* introduces more challenges to solve than the two-level compositional problem because there exist biases in all levels when estimating the inner-level function value and its gradient. Moreover, the *minimax structure* makes it more challenging to solve Eq. (2) than the multi-level compositional minimization problem because the stochastic gradient for both primal and dual variables is a biased estimator of the full gradient.

3.3. Algorithm

To mitigate the problem of biased *function* and *gradient* estimators at each level, we propose the Local Stochastic Multi-level Compositional Gradient Descent Ascent with Momentum (LocalSMCGDM) algorithm, presented in Algorithm 1. In detail, we employ a STORM-like technique (Cutkosky & Orabona, 2019) to estimate the k-th inner level function for any $k \in \{1, \cdots, K\}$ in Step 6 as follows:

$$\begin{aligned} h_{n,t+1}^{(k)} =& (1-\alpha\eta^2)(h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)}; \xi_{n,t+1}^{(k)})) \\ &+ g_n^{(k)}(h_{n,t+1}^{(k-1)}; \xi_{n,t+1}^{(k)}) \,, \end{aligned} \tag{9}$$

where $\alpha > 0$ and $\eta > 0$ are hyperparameters satisfying $\alpha\eta^2 < 1$, and $h_{n,t}^{(k)}$ is the estimator of the k-th inner level function $G_n^{(k)}(x_{n,t})$ on the n-th device at the t-th iteration. By employing the STORM-like estimators, this approach is able to approximate the inner level function more accurately.

On the other hand, for the gradient, we utilize the momentum technique to update both primal and dual variables. Specifically, the stochastic compositional gradient regarding the primal variable and dual variable is computed as follows:

$$\begin{aligned} u_{n,t+1} &= \nabla g_n^{(1)}(h_{n,t+1}^{(0)}; \xi_{n,t+1}^{(1)}) \cdots \nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)}; \xi_{n,t+1}^{(K-1)}) \\ &\times \nabla g_n^{(K)}(h_{n,t+1}^{(K-1)}; \xi_{n,t+1}^{(K)}) \nabla_h f_n(h_{n,t+1}^{(K)}, y_{n,t+1}; \zeta_{n,t+1}) \,, \\ v_{n,t+1} &= \nabla_y f_n(h_{n,t+1}^{(K)}, y_{n,t+1}; \zeta_{n,t+1}) \,. \end{aligned} \tag{10}$$

We then employ the moving-average technique to update the momentum as follows:

$$\begin{aligned} p_{n,t+1} &= (1-\rho_x\eta)p_{n,t} + \rho_x\eta u_{n,t+1} \,, \\ q_{n,t+1} &= (1-\rho_y\eta)q_{n,t} + \rho_y\eta v_{n,t+1} \,, \end{aligned} \tag{11}$$

where $\rho_x > 0, \rho_y > 0$ is the momentum coefficient hyperparameters, and η is the learning rate, satisfying $\rho_x\eta < 1$ and $\rho_y\eta < 1$. The final step of the algorithm involves using gradient descent to update the primal variable and gradient ascent to update the dual variable, as shown in Step 3.

Regarding the communication procedure in Steps 10-14, the essential components, including the model parameters $x_{n,t}, y_{n,t}$, momentum $p_{n,t}, q_{n,t}$, and estimators $h_{n,t}^{(k)}$ of k-th inner level function, are synchronized across the federated learning system every τ iterations, where $\tau > 1$ is the communication period.

Algorithm 1 LocalSMCGDAM

Input: x_0, y_0, $\eta \in (0,1)$, $\gamma_x > 0$, $\gamma_y > 0$, $\rho_x > 0$, $\rho_y > 0$, $\alpha > 0$, $\alpha\eta^2 \in (0,1)$, $\rho_x\eta \in (0,1)$, $\rho_y\eta \in (0,1)$.

1: Initialize $p_{n,0}, q_{n,0}, h_{n,0}^{(k)}$ for $k = \{1, \cdots, K\}$ as Eq (12)
2: **for** $t = 0, \cdots, T-1$, each device n **do**
3: Update x and y:
$x_{n,t+1} = x_{n,t} - \gamma_x\eta p_{n,t}$,
$y_{n,t+1} = y_{n,t} + \gamma_y\eta q_{n,t}$,
4: $h_{n,t+1}^{(0)} = x_{n,t+1}$,
5: **for** $k = 1, \cdots, K$ **do**
6: Estimate the k-th inner-level function as Eq. (9),
7: **end for**
8: Compute the stochastic compositional gradient w.r.t. x and y as Eq. (10),
9: Update momentum as Eq. (11),
10: **if** $\text{mod}(t+1, \tau) == 0$ **then**
11: $x_{n,t+1} = \frac{1}{N}\sum_{n'=1}^{N} x_{n',t+1}$,
$y_{n,t+1} = \frac{1}{N}\sum_{n'=1}^{N} y_{n',t+1}$,
$p_{n,t+1} = \frac{1}{N}\sum_{n'=1}^{N} p_{n',t+1}$,
$q_{n,t+1} = \frac{1}{N}\sum_{n'=1}^{N} q_{n',t+1}$,
12: **for** $k = 1, \cdots, K$ **do**
13: $h_{n,t+1}^{(k)} = \frac{1}{N}\sum_{n'=1}^{N} h_{n',t+1}^{(k)}$,
14: **end for**
15: **end if**
16: **end for**

As for the initialization step in Algorithm 1, we select a mini-batch of samples to initialize $p_{n,0}$, $q_{n,0}$, and $h_{n,0}^{(k)}$ as:

$$\begin{aligned} p_{n,0} &= \nabla g_n^{(1)}(x_{n,0}; \xi_{n,0}^{(1)}) \cdots \nabla g_n^{(K-1)}(G_n^{(K-2)}(x_{n,0}); \xi_{n,0}^{(K-1)}) \\ &\times \nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,0}); \xi_{n,0}^{(K)}) \nabla_x f_n(G_n^{(K)}(x_{n,0}), y_{n,0}; \zeta_{n,0}) \,, \\ q_{n,0} &= \nabla_y f_n(G_n^{(K)}(x_{n,0}), y_{n,0}; \zeta_{n,0}), \\ h_{n,0}^{(k)} &= g_n^{(k)}(h_{n,0}^{(k-1)}; \xi_{n,0}^{(k)}), k = \{1, \cdots, K\}, \end{aligned} \tag{12}$$

where the batch size is M in the initialization step.

4. Theoretical Analysis

In this section, we give the theoretical analysis of the proposed algorithm.

Assumption 4.1. For any $k \in \{1, \cdots, K\}$ and $n \in \{1, \cdots, n\}$, $g_n^{(k)}(\cdot; \xi^{(k)})$ is C_g-Lipschitz continuous where $C_g > 0$, $\nabla g_n^{(k)}(\cdot; \xi^{(k)})$ is L_g-Lipschitz continuous where $L_g > 0$, $f_n(\cdot, \cdot; \zeta)$ is C_f-Lipschitz continuous with respect to the primal variable where $C_f > 0$, $\nabla f_n(\cdot, \cdot; \zeta)$ is L_f-Lipschitz continuous where $L_f > 0$.

Assumption 4.2. For any $k \in \{1, \cdots, K\}$ and $n \in \{1, \cdots, n\}$, the variance of the stochastic gradients $\nabla g_n^{(k)}(\cdot; \xi^{(k)})$ and $\nabla f_n(\cdot, \cdot; \zeta)$ is upper bounded by σ^2 where $\sigma > 0$ is a constant. Also, the variance of the stochastic function value of $g_n^{(k)}(\cdot; \xi^{(k)})$ is upper bounded by δ^2 where $\delta > 0$ is a constant.

Assumption 4.3. For any $n \in \{1, \cdots, n\}$, $f_n(\cdot, \cdot)$ is μ-strongly concave with respect to the dual variable where

$\mu > 0$ is a constant.

These assumptions are widely used in existing compositional minimization and minimax optimization problems (Yang et al., 2019; Jiang et al., 2022; Balasubramanian et al., 2022; Yuan et al., 2021a; Zhang et al., 2023).

Next, we introduce the following auxiliary functions for analyzing the convergence rate of our algorithm:

$$\begin{aligned} y^*(x) &= \arg\max_{y\in\mathbb{R}^{d_2}} \frac{1}{N}\sum_{n=1}^{N} f_n(G(x), y)\ , \\ \Phi_n(x) &\triangleq f_n(G(x), y^*)\ , \\ \Phi(x) &= \frac{1}{N}\sum_{n=1}^{N}\Phi_n(x) = \frac{1}{N}\sum_{n=1}^{N} f_n(G(x), y^*)\ , \end{aligned} \tag{13}$$

In this paper, $\bar{a}_t$ is utilized to denote the mean value across devices for any variable a_t.

Theorem 4.4. *Given Assumption 4.1-4.3, for any $k \in \{1, \cdots, K\}$, we denote*

$$\begin{aligned} \tilde{w}_k = \frac{\mu^2}{\alpha}16A_k + 32B_k\mu^2 + \frac{500C_G^2L_f^4}{\alpha}C_g^{2(K-k)} \\ + 1000\times 2^{K+1}L_f^4C_g^{2(2K-k)}\ , \end{aligned}$$

where $A_k = C_g^{2(K-1)}C_f^2L_g^2(\sum_{j=k}^{K-1}C_g^{j-k})^2 + L_f^2C_g^{2(2K-k)}$ and $B_k = \sum_{j=k+1}^{K}A_j(2C_g^2)^{j-k}$ are constant values regarding the Lipschitz constant C_g, C_f, L_g, and L_f. Then, by setting $\alpha > 0$, $\rho_x > 0$, $\rho_y > 0$, and

$$\begin{aligned} \eta \le& \min\left\{\frac{1}{2\gamma_x L_\Phi}, \frac{1}{\sqrt{\alpha}}, \frac{1}{\rho_x}, \frac{1}{\rho_y}, \frac{1}{4\sqrt{\tau\alpha C_g}}, \frac{1}{4\tau\sqrt{\rho_y\gamma_y L_f}},\right. \\ & \left.\frac{1}{2}\sqrt{\frac{\tilde{w}_k}{\alpha\sum_{j=k+1}^{K}\tilde{w}_j(2C_g^2)^{j-k}}}, 1\right\}, \\ \gamma_y \le& \min\left\{\frac{1}{6L_f}, \frac{3\rho_y^2\mu}{100L_f^2}, \frac{15\rho_x^2}{8\mu(K+1)}\right\}, \\ \gamma_x \le& \min\left\{\frac{\gamma_y\mu^2}{13C_g^{2K}L_f^2}, \frac{\rho_y\mu}{78C_g^{2K}L_f^2}, \frac{\mu}{\sqrt{24K\sum_{k=1}^{K}\tilde{w}_k(2C_g^2)^k}},\right. \\ & \left.\frac{\rho_x}{\sqrt{96(K+1)(C_g^{4K}L_f^2 + \sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2)}}\right\}, \end{aligned}$$

Algorithm 1 has the following convergence rate

$$\begin{aligned} &\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|^2] + C_G^2L_f^2\mathbb{E}\|y^*(\bar{x}_t) - \bar{y}_t\|^2]) \\ &\le \frac{2(\Phi(x_0) - \Phi^*)}{\gamma_x\eta T} + \frac{20C_G^2L_f^2}{\gamma_y\eta\mu T}\mathbb{E}[\|\bar{y}_0 - y^*(\bar{x}_0)\|^2] + O(\tau^2\eta^2) \\ &\quad + O(K\tau^4\eta^4) + O(K\tau^6\eta^6) + O(K\tau^8\eta^8) + O(K\tau^{10}\eta^{10}) \\ &\quad + O(\frac{K^2}{\eta T}) + O(\eta^2K^2) + O(\frac{\eta K^2}{N}) + O(\frac{K^2}{\eta^2 MT})\ , \end{aligned} \tag{14}$$

where M is the mini-batch size in initialization step, Φ^* denotes the function value at the optimal solution, L_Φ and C_G are defined in Lemma B.1.

Remark 4.5. For sufficiently large T, by setting the learning rate $\eta = O(N^{1/2}/T^{1/2}), \tau = O(T^{1/4}/N^{3/4})$, $M = O(T^{1/2}/N^{1/2})$, Algorithm 1 can achieve $O(1/\sqrt{NT})$ convergence rate when ignoring the number of levels K. It indicates a linear speedup with respect to the number of devices N. In addition, the communication complexity of our algorithm is $T/\tau = O(N^{3/4}T^{3/4})$. Furthermore, when considering the number of levels K, our convergence rate can be represented as $O(\tilde{C}^K K^2/\sqrt{NT})$ where $\tilde{C}$ is a constant depending on Lipschitz constants. Note that previous works (Yang et al., 2019; Balasubramanian et al., 2022; Zhang & Xiao, 2021; Jiang et al., 2022) also have such a constant factor $\tilde{C}^K$ depending on K exponentially. Following Remark 1 and 3 in (Balasubramanian et al., 2022), the factor $\tilde{C}^K$ can be ignored so that the number of levels K affects the convergence rate polynomially instead of exponentially, which can be viewed as a level-independent convergence rate as (Balasubramanian et al., 2022).

Remark 4.6. By setting $\eta = O(N\epsilon^2)$ and $\tau = O(1/(N\epsilon))$, we can achieve the ϵ-accuracy solution, i.e., $\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}[\|\nabla\Phi(\bar{\mathbf{x}}_t)\|^2] + C_G^2L_f^2\mathbb{E}[\|y^*(\bar{x}_t) - \bar{y}_t\|^2]) \le \epsilon^2$. Then, the sample complexity on each device is $O(1/(N\epsilon^4))$ and the communication complexity is $O(1/\epsilon^3)$.

Proof sketch of Theorem 4.4 By introducing the Lyapunov function based on Algorithm 1, we have

$$\begin{aligned} P_{t+1} =& \mathbb{E}[\Phi(\bar{x}_{t+1})] + w_0\mathbb{E}[\|\bar{y}_{t+1} - y^*(\bar{x}_{t+1})\|^2] \\ &+ \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} w_k\mathbb{E}[\|h_{t+1}^{(k)} - g_n^{(k)}(h_{n,t+1}^{(k-1)})\|^2] \\ &+ w_{K+1}\mathbb{E}\Big[\Big\|\bar{p}_{t+1} - \frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1})) \\ &\quad \cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_G f_n(G_n^{(K)}(x_{n,t+1}), y_{n,t+1})\Big\|^2\Big] \\ &+ w_{K+2}\mathbb{E}\Big[\Big\|\bar{q}_{t+1} - \frac{1}{N}\sum_{n=1}^{N}\nabla_y f_n(G_n(x_{n,t+1}), y_{n,t+1})\Big\|^2\Big] \end{aligned} \tag{15}$$

where $w_0 = \frac{10\gamma_x C_G^2L_f^2}{\gamma_y\mu}$, $w_{K+1} = \frac{2\gamma_x}{\rho_x}$, $w_{K+2} = \frac{125\gamma_x C_G^2L_f^2}{\rho_y\mu^2}$ and $w_k = \frac{\gamma_x K}{\eta\mu^2}\tilde{w}_k$. Here, the third item on the right-hand side quantifies the estimator error of the inner-level function $g_n^{(k)}(h_{n,t+1}^{(k-1)})$, the fourth item bounds the estimation error between the primal momentum and the partial gradient regarding the primal variable using the chain rule, the fifth item represents the approximation error between the dual momentum and the partial gradient regarding the dual variable. Then, by bounding $P_{t+1} - P_t$ and summing t over $0, \cdots, T-1$, we can complete the proof.

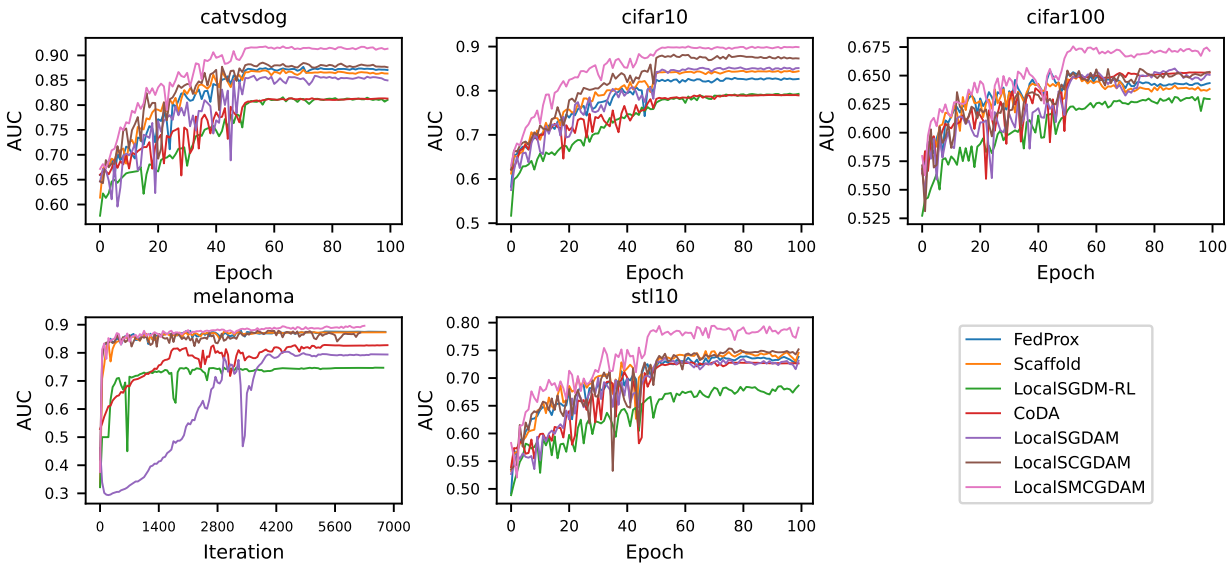


Figure 2. Testing performance with AUC score versus the number of epochs when the communication period $\tau = 4$ and imratio$= 0.05$.

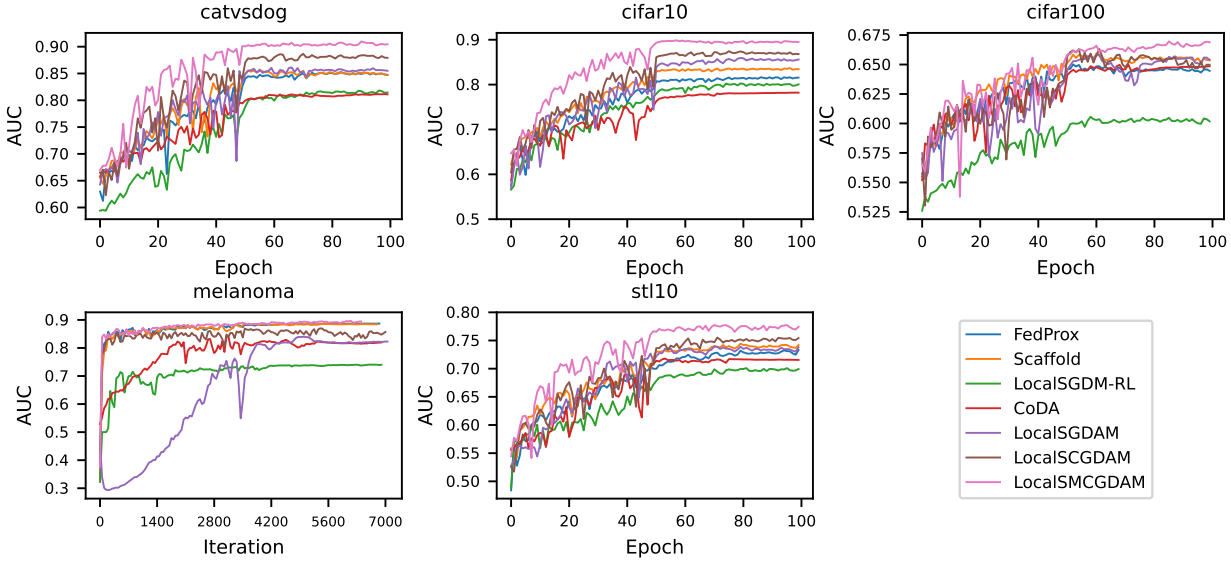


Figure 3. Testing performance with AUC score versus the number of epochs when the communication period $\tau = 8$ and imratio$= 0.05$.

Unique Challenge. Due to the multi-level structure coupled with the federated minimax optimization setting, investigating the convergence rate of our algorithm is challenging. Specifically, as shown in Lemma B.8 and Lemma B.9, the consensus error with respect to the momentum of the dual variable $\mathbb{E}[\|q_{n,t} - \bar{q}_t\|^2]$ depends on the consensus error $\mathbb{E}[\|h_{n,t}^{(k)} - \bar{h}_t^{(k)}\|^2]$ and the estimation error $\mathbb{E}[\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\|^2]$ of the inner-level function for all $k \in \{1, \cdots, K\}$, both of which further recursively depends on their corresponding error in lower levels. Such kinds of *accumulated* and *recursive* errors across levels make it more challenging to bound the consensus error than the two-level compositional minimax problem. Moreover, the accumulated consensus and estimation errors across levels make achieving the linear speedup convergence rate more challenging. In particular, those accumulated consensus and estimation errors could prevent the algorithm from achieving linear speedup. For instance, for the multi-level case, we have to bound the estimation error $\mathbb{E}[\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\|^2]$ for each device n, rather than the mean value across devices. Such an individual estimation error introduces a large variance because it cannot be scaled by the number of devices N, presenting new challenges for linear speedup. Our convergence analysis addressed these unique challenges, achieving both the linear speedup and the level-independent convergence rate for the federated multi-level compositional minimax problem. To the best of our knowledge, this is the first work achieving such fa-

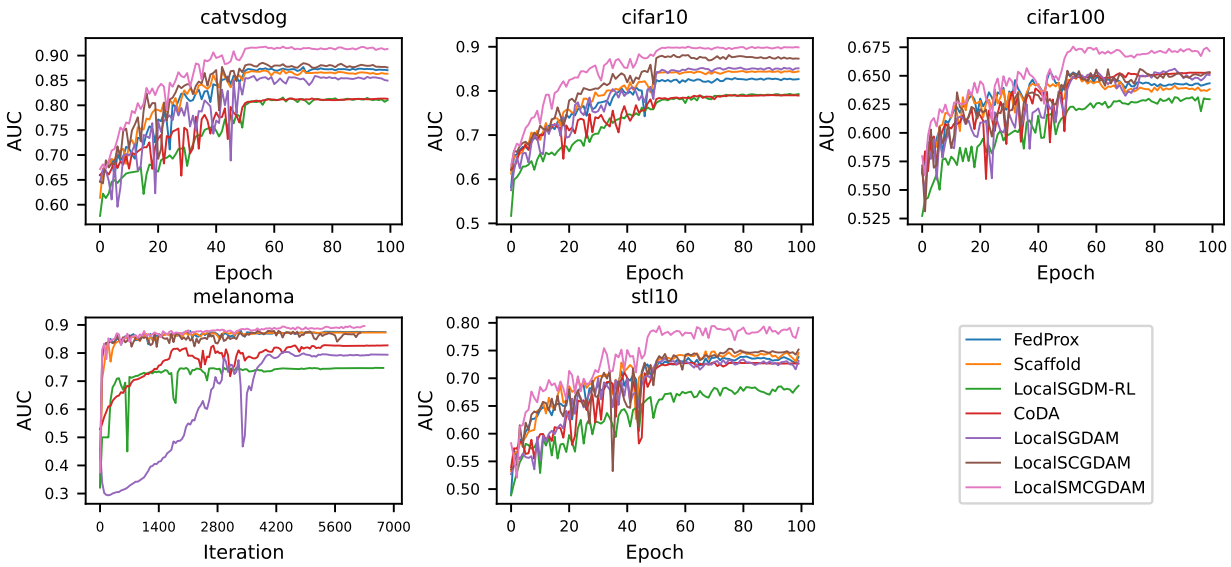


Figure 4. Testing performance with AUC score versus the number of epochs when the communication period $\tau = 16$ and imratio$= 0.05$.

vorable results. In fact, such theoretical results do not even exist under the single-machine setting. In summary, our convergence analysis is novel and has important theoretical contributions to compositional minimax optimization.

5. Experiment

In this section, we validate the practical performance of our algorithm on highly imbalanced classification tasks in the federated learning setting.

Baselines. We compare our method with six state-of-the-art federated learning approaches: FedProx (Li et al., 2020), Scaffold (Karimireddy et al., 2020), LocalSGDM-RL (Wang et al., 2021), CoDA (Guo et al., 2020), LocalSGDAM (Sharma et al., 2022) and LocalSCGDAM (Zhang et al., 2023). The first two methods address heterogeneous scenarios, assuming that the local data distribution is imbalanced, but the global one is balanced. Scaffold (Karimireddy et al., 2020) utilizes SGD to optimize the traditional cross-entropy loss function, employing a control variate to correct for the stochastic gradient in local updates. FedProx (Li et al., 2020), on the other hand, optimizes the traditional cross-entropy loss function with a proximal term using SGD. LocalSGDM-RL (Wang et al., 2021) leverages momentum SGD to optimize a Ratio Loss function, particularly designed for imbalanced distributions by adding a regularization term to the standard cross-entropy loss function. CoDA (Guo et al., 2020) employs SGDA to optimize AUC loss in a stage-wise manner, obtaining the primal variable through a communication-efficient algorithm and estimating the corresponding dual variable by sampling data from local machines. For LocalSGDAM (Sharma et al., 2022), we optimize the AUC loss using momentum SGDA. LocalSCGDAM (Zhang et al., 2023) introduces the compositional structure in the AUC loss function and proposes a federated compositional minimax optimization problem, utilizing momentum SCGDA to optimize AUC loss.

Dataset and Hyperparameter Setting. We employ five image datasets: CatvsDog [1], CIFAR10 (Krizhevsky et al., 2009), CIFAR100 (Krizhevsky et al., 2009), STL10 (Coates et al., 2011), and Melanoma (Rotemberg et al., 2021), for binary classification [2]. We create imbalanced datasets by randomly dropping samples from the positive class in the training set, while keeping the testing set balanced. The imbalance ratio signifies the ratio of positive samples to the total number of samples. The Melanoma dataset maintains its inherent uneven distribution. Note that The batch size on each device is 8 for STL10 and 16 for the remaining datasets. Across all algorithms, the learning rate is fixed at 0.1 and decreased by a factor of 10 at the 50% and 75% epochs for all methods during the total training duration. Specifically, for our multi-level method, the learning rate for each inner-level is decreased by 100 to avoid overfitting. Additionally, the number of levels K is set to 3 in our experiments. We leverage eight workers ($N = 8$), simulated on four V100-GPUs, for the first four datasets. However, for Melanoma, we only implement four workers ($N = 4$) due to the GPU memory constraints. More experimental details can be found in Appendix A.

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

[2] Note that our algorithm can be easily adapted to the multi-class classification problem by following (Liu et al., 2019).

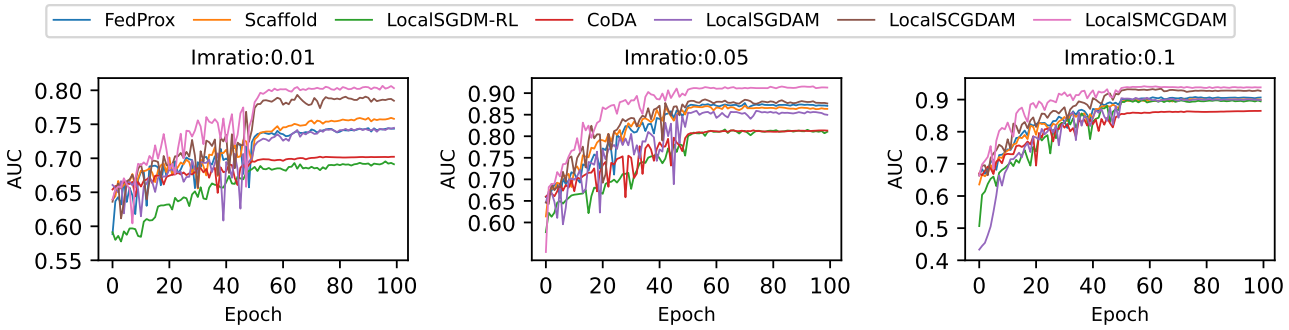
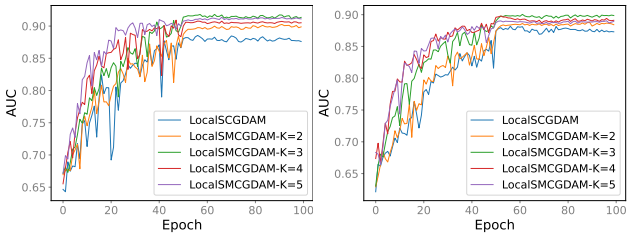


Figure 5. The test AUC score versus the number of epochs when using different imbalance ratios for CATvsDOG.

Experimental Results. In Figure 2, 3, and 4, we plot the test AUC score against the number of epochs when the communication period is set to 4, 8, and 16, respectively. Here, the imbalance ratio, denoted as imratio, is set to 0.05. In Figure 5, we plot the test AUC score against the number of epochs when using different imbalance ratios for the CATvsDOG dataset, where the communication period is 4.

It is evident that our algorithm LocalSMCGDAM consistently outperforms all baselines across all datasets under all scenarios. In particular, we have the following observations. (i) Compared with FedProx and Scaffold, our algorithm can showcase its effectiveness in handling the global imbalanced data distribution. (ii) Compared to LocalSCGDAM, which essentially serves as a simplified version with a one-step update for the inner-level function (i.e., the inner level $K = 1$), our LocalSMCGDAM designed for the multi-step update demonstrates significant improvement in optimizing the AUC loss. More experimental results on the AUC score of different algorithms can be found in Appendix A.



(a) CATvsDOG (b) CIFAR10

Figure 6. The test AUC score versus the number of epochs when using different number of levels K.

Additionally, we conduct experiments to demonstrate how the number of levels K affects the performance. In our loss function in Eq. (2), the inner-level function corresponds to training the classifier by minimizing the cross-entropy loss function, and the outer-level loss function corresponds to training the classifier by optimizing the AUC loss function. Then, intuitively, the inner-level function can be seen as the pretraining process, while the outer-level function can be regarded as the fine-tuning process. In fact, the AUC loss function is a minimax function, making it difficult to train the classifier from scratch. On the contrary, the cross-entropy loss function is convex, which makes it easier to train the classifier from scratch. Therefore, when using a multi-level inner function, we actually pretrain the classifier by minimizing the cross-entropy loss function for multiple steps. This typically can pretrain the classifier better than just using one step. Subsequently, fine-tuning the classifier with optimizing AUC loss can lead to improved performance. In Figure 6, we plot the test AUC score when using different number of levels K for CATvsDOG and CIFAR10, where the communication period is 4. It can be observed that the proposed multi-level algorithm LocalSMCGDAM always outperforms the baseline two-level algorithm LocalSCGDAM, where the number of inner update steps is $K = 1$ and the inner-level function estimator is based on moving average. This confirms the effectiveness of using multiple inner update steps. Meanwhile, it can be observed that a too large K degenerates the performance. This phenomenon can be explained from the perspective of pretraining. During the pretraining procedure, the cross-entropy loss function is used. As we know, the cross-entropy loss function does not perform well on imbalanced data. Therefore, when updating the classifier's weight by minimizing the cross-entropy loss for more steps, i.e., with a larger K, it makes the classifier's weight fit the cross-entropy loss too much,resulting in inferior performance. Therefore, K cannot be very large.

6. Conclusion

In this paper, we proposed a federated stochastic multi-level compositional minimax algorithm for deep AUC maximization named LocalSMCGDAM. Our theoretical analysis demonstrates that the proposed algorithm achieves level-independent convergence rates and linear speedup in the federated learning setting. This is the first time that these theoretical convergence results have been achieved for multi-level compositional minimax problems. Through comprehensive experimental validation, we present the effectiveness of our algorithm, showcasing its capability to address highly imbalanced classification tasks.

Acknowledgements

We thank anonymous reviewers for constructive comments. H. Gao was partially supported by Cisco Research Award.

Impact Statement

This paper introduces novel algorithms to enhance the performance of imbalanced classification tasks under a federated learning setting within the domain of Machine Learning. The primary objective is to contribute to the ongoing advancements in this field. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Balasubramanian, K., Ghadimi, S., and Nguyen, A. Stochastic multilevel composition optimization algorithms with level-independent convergence rates. *SIAM Journal on Optimization*, 32(2):519–544, 2022.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Chen, T., Sun, Y., and Yin, W. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021.

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

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

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

Elkan, C. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pp. 973–978. Lawrence Erlbaum Associates Ltd, 2001.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Gao, H. Decentralized multi-level compositional optimization algorithms with level-independent convergence rate. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4402–4410. PMLR, 02–04 May 2024.

Gao, H., Xu, A., and Huang, H. On the convergence of communication-efficient local sgd for federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7510–7518, 2021.

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

Ghadimi, S., Ruszczynski, A., and Wang, M. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM Journal on Optimization*, 30(1):960–979, 2020.

Guo, Z., Liu, M., Yuan, Z., Shen, L., Liu, W., and Yang, T. Communication-efficient distributed stochastic auc maximization with deep neural networks. In *International conference on machine learning*, pp. 3864–3874. PMLR, 2020.

Guo, Z., Jin, R., Luo, J., and Yang, T. Fedxl: Provable federated learning for deep x-risk optimization. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11934–11966. PMLR, 2023. URL https://proceedings.mlr.press/v202/guo23c.html.

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

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

Jamal, M. A., Brown, M., Yang, M.-H., Wang, L., and Gong, B. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7610–7619, 2020.

Jiang, W., Wang, B., Wang, Y., Zhang, L., and Yang, T. Optimal algorithms for stochastic multi-level compositional optimization. In *International Conference on Machine Learning*, pp. 10195–10216. PMLR, 2022.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

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

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

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

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

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

Shapiro, A., Dentcheva, D., and Ruszczynski, A. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

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

Tarzanagh, D. A., Li, M., Thrampoulidis, C., and Oymak, S. Fednest: Federated bilevel, minimax, and compositional optimization. In *International Conference on Machine Learning*, pp. 21146–21179. PMLR, 2022.

Tran-Dinh, Q., Pham, N. H., Phan, D. T., and Nguyen, L. M. A hybrid stochastic optimization framework for composite nonconvex optimization. *Mathematical Programming*, 191(2):1005–1071, 2022.

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

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

Wu, X., Sun, J., Hu, Z., Zhang, A., and Huang, H. Solving a class of non-convex minimax optimization in federated learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Yang, H., Liu, Z., Zhang, X., and Liu, J. Sagda: Achieving $\mathcal{O}(\epsilon^{-2})$ communication complexity in federated min-max learning. *arXiv preprint arXiv:2210.00611*, 2022.

Yang, S., Wang, M., and Fang, E. X. Multilevel stochastic gradient methods for nested composition optimization. *SIAM Journal on Optimization*, 29(1):616–659, 2019.

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

Yu, H., Wang, L., Wang, B., Liu, M., Yang, T., and Ji, S. Graphfm: Improving large-scale gnn training via feature momentum. In *International Conference on Machine Learning*, pp. 25684–25701. PMLR, 2022.

Yuan, H. and Hu, W. Stochastic recursive momentum method for non-convex compositional optimization. *arXiv preprint arXiv:2006.01688*, 2020.

Yuan, Z., Yan, Y., Sonka, M., and Yang, T. Robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. *arXiv preprint arXiv:2012.03173*, 8, 2020.

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

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

Yuan, Z., Yan, Y., Sonka, M., and Yang, T. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3040–3049, 2021c.

Zhang, J. and Xiao, L. Multilevel composite stochastic optimization via nested variance reduction. *SIAM Journal on Optimization*, 31(2):1131–1157, 2021.

Zhang, X., Zhang, Y., Yang, T., Souvenir, R., and Gao, H. Federated compositional deep auc maximization. *arXiv preprint arXiv:2304.10101*, 2023.

A. Appendix: More Experimental Settings and Results

A.1. Compositional AUC Loss Function

Below we provide the AUC loss function under the single-machine setting to explain Eq. (1). Specifically, according to (Yuan et al., 2021a), we can train the classifier by combining AUC loss and cross-entropy loss optimization as follows:

$$\min_{w,\tilde{w}_1,\tilde{w}_2} \max_{\tilde{w}_3} \mathbb{E}[\mathcal{L}_{AUC}(w - \rho\mathbb{E}[\nabla\mathcal{L}_{CE}(w;\xi)], \tilde{w}_1, \tilde{w}_2, \tilde{w}_3;\xi)] \,, \tag{16}$$

where $w \in \mathbb{R}^d$ is the classifier's weight, $\tilde{w}_1 \in \mathbb{R}$, $\tilde{w}_2 \in \mathbb{R}$, and $w_3 \in \mathbb{R}$ are the parameters to compute the AUC loss, ξ denotes the random sample, $\rho > 0$ is the step size, $\mathcal{L}_{AUC}$ denotes the AUC loss, and $\mathcal{L}_{CE}$ denotes the cross-entropy loss.

Then, by introducing $x = [w^T, \tilde{w}_1, \tilde{w}_2]^T$, $y = \tilde{w}_3$, and $\Delta(x,\xi) = [\nabla\mathcal{L}_{CE}(w;\xi)^T, 0, 0]^T$, we can denote $g(x;\xi) = x - \rho\Delta(x,\xi)$ and $f(g(x), y;\zeta) = \mathcal{L}_{AUC}(w - \rho\mathbb{E}[\nabla\mathcal{L}_{CE}(w;\xi)], \tilde{w}_1, \tilde{w}_2, \tilde{w}_3;\zeta)$, where ζ denotes the random samples for computing the AUC loss. Based on $g(x;\xi)$ and $f(g(x), y;\zeta)$, it is easy to obtain Eq. (1) under the federated learning setting.

A.2. More Experimental Settings and Results

More Experimental Settings. The details of different benchmark datasets are presented in Table 1. We partition each dataset into two groups based on its classes for the initial four datasets. The first half of the classes is considered as the positive class, while the second half is considered the negative class. Then we randomly drop samples from the positive class as shown to create highly imbalanced data, where the imbalance ratio is 0.05. For Melanoma, we utilize DenseNet121 (Huang et al., 2017), while ResNet20 (He et al., 2016) is employed for the remaining datasets. In each neural network, the dimensionality of the last layer is set to 1 for binary classification. For all experiments, the weight decay is set to $1e-4$. The details of the hyperparameters of different methods are provided in Table 2. Here, all algorithms employ similar learning rates to make a fair comparison. For the last three algorithms, the product of the learning rate and learning rate coefficient, i.e., $\gamma_x\eta$ and $\gamma_y\eta$, equals 0.099, close to the values used in the other algorithms. The learning rate is decayed at 50% and 75% of total training iterations by 10. Specifically, for our LocalSMCGDAM algorithms, the learning rate of each inner-level is decayed by 100 to avoid overfitting.

More Experimental Results. We report the best test AUC score for all methods in Table 3.

Table 1. Description of benchmark datasets. Here, #positive denotes the number of samples in the positive class, and #negative denotes the number of samples in the negative class.

Dataset	Training set		Testing set	
	#positive	#negative	#positive	#negative
CIFAR10	1,315	25,000	5,000	5,000
CIFAR100	1,315	25,000	5,000	5,000
STL10	131	2,500	4,000	4,000
CATvsDOG	527	10,016	2,516	2,888
Melanoma	868	25,670	117	6,881

Table 2. The hyperparameters of different methods.

Methods	Hyperparameters	Value
FedProx, Scaffold	learning rate	0.1
LocalSGDM-RL	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 ρ_x and ρ_y	3.3
LocalSCGDAM	learning rate for outer level η	0.3
	learning rate coefficient for outer level γ_x and γ_y	0.33
	momentum coefficient ρ_x and ρ_y for outer level	3.3
	learning rate for inner-level η'	0.1
	coefficient for inner-level α	3.0
LocalSMCGDAM (Ours)	learning rate for outermost-level η	0.3
	learning rate coefficient for outermost-level γ_x and γ_y	0.33
	momentum coefficient for outermost-level ρ_x and ρ_y	$\{0.2, 2.0, 3.3\}$
	learning rate for inner-levels η'	0.1
	coefficient α	3.0

Table 3. The comparison between the test AUC score of different methods on all datasets. Here, τ denotes the communication period.

Datasets	Period	Methods						
		Local SMCGDAM	Local SCGDAM	Local SGDAM	CoDA	Local SGDM-RL	Fed Prox	Scaffold
CATvsDOG	$\tau = 4$	**0.918**	0.885	0.859	0.814	0.816	0.877	0.872
	$\tau = 8$	**0.910**	0.886	0.861	0.811	0.818	0.853	0.855
	$\tau = 16$	**0.913**	0.884	0.854	0.813	0.792	0.822	0.851
CIFAR10	$\tau = 4$	**0.900**	0.882	0.853	0.791	0.793	0.829	0.845
	$\tau = 8$	**0.899**	0.874	0.860	0.782	0.803	0.816	0.837
	$\tau = 16$	**0.892**	0.871	0.848	0.778	0.753	0.793	0.836
CIFAR100	$\tau = 4$	**0.676**	0.657	0.656	0.652	0.631	0.654	0.650
	$\tau = 8$	**0.670**	0.663	0.658	0.650	0.606	0.650	0.659
	$\tau = 16$	**0.669**	0.660	0.651	0.645	0.607	0.647	0.654
STL10	$\tau = 4$	**0.794**	0.753	0.736	0.729	0.686	0.739	0.749
	$\tau = 8$	**0.778**	0.755	0.740	0.718	0.701	0.733	0.744
	$\tau = 16$	**0.764**	0.757	0.734	0.716	0.703	0.715	0.738
Melanoma	$\tau = 4$	**0.896**	0.878	0.804	0.828	0.748	0.881	0.876
	$\tau = 8$	**0.897**	0.870	0.840	0.840	0.741	0.887	0.887
	$\tau = 16$	**0.903**	0.878	0.785	0.833	0.760	0.875	0.884

B. Appendix: Theoretical Proof

Lemma B.1. *Given Assumption 4.1-4.3, we can know*

- $G^{(k)}(x)$ *is* C_g^k*-Lipschitz continuous for* $k \in \{1, \cdots, K-1\}$ *and* $G(x)$ *is* C_G*-Lipschitz continuous where* $C_G = C_g^K$*;*
- $\nabla G(x)$ *is* L_G*-Lipschitz continuous where* $L_G = \sum_{j=0}^{K-1} L_g C_g^{K-1+j}$*;*
- $\nabla \Phi(x)$ *is* L_Φ*-Lipschitz continuous where* $L_\Phi = C_G^2 L_f + C_G L_f L_{y^*} + C_f L_G$*;*
- $y^*(x)$ *is* L_{y^*}*-Lipschitz continuous where* $L_{y^*} = \frac{C_G L_f}{\mu}$.

Proof. For any $k \in \{1, \cdots, K\}$ and any $x_1, x_2 \in \mathbb{R}^{d_1}$, we have

$$\begin{aligned} &\|G^{(k)}(x_1) - G^{(k)}(x_2)\| \\ =& \|g^{(k)}(G^{(k-1)}(x_1)) - g^{(k)}(G^{(k-1)}(x_2))\| \\ \leq& C_g \|G^{(k-1)}(x_1) - G^{(k-1)}(x_2)\| \\ \leq& C_g^k \|x_1 - x_2\| , \end{aligned} \tag{17}$$

where the first inequality follows from Assumption 4.1, which completes the first property.

The second property can be proved as follows:

$$\begin{aligned} &\|\nabla G(x_1) - \nabla G(x_2)\| \\ =& \|\nabla g^{(1)}(x_1) \nabla g^{(2)}(G^{(1)}(x_1)) \nabla g^{(3)}(G^{(2)}(x_1)) \cdots \nabla g^{(K)}(G^{(K-1)}(x_1)) \\ &- \nabla g^{(1)}(x_2) \nabla g^{(2)}(G^{(1)}(x_2)) \nabla g^{(3)}(G^{(2)}(x_2)) \cdots \nabla g^{(K)}(G^{(K-1)}(x_2))\| \\ \leq& \|\nabla g^{(1)}(x_1) \nabla g^{(2)}(G^{(1)}(x_1)) \nabla g^{(3)}(G^{(2)}(x_1)) \cdots \nabla g^{(K)}(G^{(K-1)}(x_1)) \\ &\quad - \nabla g^{(1)}(x_2) \nabla g^{(2)}(G^{(1)}(x_1)) \nabla g^{(3)}(G^{(2)}(x_1)) \cdots \nabla g^{(K)}(G^{(K-1)}(x_1))\| \\ &+ \|\nabla g^{(1)}(x_2) \nabla g^{(2)}(G^{(1)}(x_1)) \nabla g^{(3)}(G^{(2)}(x_1)) \cdots \nabla g^{(K)}(G^{(K-1)}(x_1)) \\ &\quad - \nabla g^{(1)}(x_2) \nabla g^{(2)}(G^{(1)}(x_2)) \nabla g^{(3)}(G^{(2)}(x_1)) \cdots \nabla g^{(K)}(G^{(K-1)}(x_1))\| \\ &+ \|\nabla g^{(1)}(x_2) \nabla g^{(2)}(G^{(1)}(x_2)) \nabla g^{(3)}(G^{(2)}(x_1)) \cdots \nabla g^{(K)}(G^{(K-1)}(x_1)) \\ &\quad - \nabla g^{(1)}(x_2) \nabla g^{(2)}(G^{(1)}(x_2)) \nabla g^{(3)}(G^{(2)}(x_2)) \cdots \nabla g^{(K)}(G^{(K-1)}(x_1))\| \\ &+ \cdots + \|\nabla g^{(1)}(x_2) \nabla g^{(2)}(G^{(1)}(x_2)) \nabla g^{(3)}(G^{(2)}(x_2)) \cdots \nabla g^{(K)}(G^{(K-1)}(x_1)) \\ &\quad - \nabla g^{(1)}(x_2) \nabla g^{(2)}(G^{(1)}(x_2)) \nabla g^{(3)}(G^{(2)}(x_2)) \cdots \nabla g^{(K)}(G^{(K-1)}(x_2))\| \\ \leq& C_g^{K-1} L_g \|x_1 - x_2\| + C_g^{K-1} L_g \|G^{(1)}(x_1) - G^{(1)}(x_2)\| \\ &+ \cdots + C_g^{K-1} L_g \|G^{(K-1)}(x_1) - G^{(K-1)}(x_2)\| \\ \leq& C_g^{K-1} L_g \|x_1 - x_2\| + C_g^K L_g \|x_1 - x_2\| + \cdots + C_g^{K-1} L_g C_g^{K-1} \|x_1 - x_2\| \\ =& \sum_{j=0}^{K-1} L_g C_g^{K-1+j} \|x_1 - x_2\| , \end{aligned} \tag{18}$$

where the second inequality follows from Assumption 4.1.

The third property can be proved as follows:

$$\begin{aligned} &\|\nabla \Phi(x_1) - \nabla \Phi(x_2)\| \\ =& \|\nabla G(x_1) \nabla_1 f\left(G(x_1), y^*(x_1)\right) - \nabla G(x_2) \nabla_1 f\left(G(x_2), y^*(x_2)\right)\| \\ \leq& \|\nabla G(x_1) \nabla_1 f\left(G(x_1), y^*(x_1)\right) - \nabla G(x_1) \nabla_1 f\left(G(x_2), y^*(x_2)\right)\| \\ &+ \|\nabla G(x_1) \nabla_1 f\left(G(x_2), y^*(x_2)\right) - \nabla G(x_2) \nabla_1 f\left(G(x_2), y^*(x_2)\right)\| \end{aligned}$$

$$
\begin{aligned}
&\leq \|\nabla G(x_1)\|\|\nabla_1 f\left(G(x_1), y^*(x_1)\right) - \nabla_1 f\left(G(x_2), y^*(x_2)\right)\| \\
&\quad + \|\nabla G(x_1) - \nabla G(x_2)\|\|\nabla_1 f\left(G(x_2), y^*(x_2)\right)\| \\
&\leq C_G\|\nabla_1 f\left(G(x_1), y^*(x_1)\right) - \nabla_1 f\left(G(x_2), y^*(x_2)\right)\| + C_f\|\nabla G(x_1) - \nabla G(x_2)\| \\
&\leq C_G L_f\|G(x_1) - G(x_2)\| + C_G L_f\|y^*(x_1) - y^*(x_2)\| + C_f\|\nabla G(x_1) - \nabla G(x_2)\| \\
&\leq C_G^2 L_f\|x_1 - x_2\| + C_G L_f L_{y^*}\|x_1 - x_2\| + C_f L_G\|x_1 - x_2\| \\
&= (C_G^2 L_f + C_G L_f L_{y^*} + C_f L_G)\|x_1 - x_2\|\,, \qquad (19)
\end{aligned}
$$

where the fourth and fifth inequalities follow from Assumption 4.1.

Following (Zhang et al., 2023), the last property can be proved by showing $\|y^*(x_2) - y^*(x_1)\| \leq \frac{C_G L_f}{\mu}\|x_2 - x_1\|$.

□

Lemma B.2. *Given Assumption 4.1-4.3, we can know*

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^N p_{n,t+1} - \frac{1}{N}\sum_{n=1}^N \nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}), y_{n,t+1})\Big\|^2\Big] \\
&\leq (1-\rho_x\eta)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^N p_{n,t} - \frac{1}{N}\sum_{n=1}^N \nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}), y_{n,t})\Big\|^2\Big] \\
&\quad + \frac{4\eta\gamma_x^2}{\rho_x}\Big(C_g^{4K}L_f^2 + \sum_{k=0}^{K-1} C_g^{2(K-1+k)}C_f^2L_g^2\Big)\frac{K+1}{N}\sum_{n=1}^N\mathbb{E}\Big[\Big\|p_{n,t} - \bar{p}_t\Big\|^2\Big] \\
&\quad + \frac{4\eta\gamma_x^2}{\rho_x}(K+1)\Big(C_g^{4K}L_f^2 + \sum_{k=0}^{K-1} C_g^{2(K-1+k)}C_f^2L_g^2\Big)\mathbb{E}\Big[\Big\|\bar{p}_t\Big\|^2\Big] \\
&\quad + \frac{4\eta\gamma_y^2}{\rho_x}C_g^{2K}L_f^2\frac{K+1}{N}\sum_{n=1}^N\mathbb{E}\Big[\Big\|q_{n,t} - \bar{q}_t\Big\|^2\Big] + \frac{4\eta\gamma_y^2}{\rho_x}C_g^{2K}L_f^2(K+1)\mathbb{E}\Big[\Big\|\bar{q}_t\Big\|^2\Big] \\
&\quad + 4\rho_x\eta\frac{K}{N}\sum_{n=1}^N\sum_{k=1}^K A_k\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] + 8\rho_x\alpha^2\eta^5\frac{K}{N}\sum_{n=1}^N\sum_{k=1}^K\Big(\sum_{j=k+1}^K A_j(2C_g^2)^{j-k}\Big)\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] \\
&\quad + 8\rho_x\gamma_x^2\eta^3K\sum_{k=1}^K A_k(2C_g^2)^k\frac{1}{N}\sum_{n=1}^N\mathbb{E}\Big[\Big\|p_{n,t} - \bar{p}_t\Big\|^2\Big] + 8\rho_x\gamma_x^2\eta^3K\sum_{k=1}^K A_k(2C_g^2)^k\mathbb{E}\Big[\Big\|\bar{p}_t\Big\|^2\Big] \\
&\quad + 8\rho_x\alpha^2\eta^5\delta^2K\sum_{k=1}^K A_k\sum_{j=1}^{k-1}(2C_g^2)^j + \rho_x^2\eta^2(K(K+1)C_g^{2(K-1)}C_f^2 + (K+1)C_g^{2K})\frac{\sigma^2}{N}\,, \qquad (20)
\end{aligned}
$$

where $A_k = \left(C_g^{2(K-1)}C_f^2L_g^2\Big(\sum_{j=k}^{K-1}C_g^{j-k}\Big)^2 + C_g^{2K}L_f^2C_g^{2(K-k)}\right)$.

Proof.

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^N p_{n,t+1} - \frac{1}{N}\sum_{n=1}^N \nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}), y_{n,t+1})\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^N(1-\rho_x\eta)(p_{n,t} - \nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}), y_{n,t})) \\
&\quad + (1-\rho_x\eta)\frac{1}{N}\sum_{n=1}^N\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}), y_{n,t}) \\
&\quad - (1-\rho_x\eta)\frac{1}{N}\sum_{n=1}^N\nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}), y_{n,t+1}) \\
&\quad - \rho_x\eta\frac{1}{N}\sum_{n=1}^N\nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}), y_{n,t+1})
\end{aligned}
$$

$$
\begin{aligned}
&+ \rho_x\eta\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(h_{n,t+1}^{(0)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1})\\
&- \rho_x\eta\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(h_{n,t+1}^{(0)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1})\\
&+ \rho_x\eta\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(h_{n,t+1}^{(0)};\xi_{n,t+1}^{(1)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)};\xi_{n,t+1}^{(K-1)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)};\xi_{n,t+1}^{(K)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1};\zeta_{n,t+1})\Big\|^2\Big]\\
&= \mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}(1-\rho_x\eta)(p_{n,t}-\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}),y_{n,t}))\\
&+ (1-\rho_x\eta)\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}),y_{n,t})\\
&- (1-\rho_x\eta)\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1})\\
&- \rho_x\eta\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1})\\
&+ \rho_x\eta\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(h_{n,t+1}^{(0)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1})\Big\|^2\Big]\\
&+ \rho_x^2\eta^2\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(h_{n,t+1}^{(0)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1})\\
&- \frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(h_{n,t+1}^{(0)};\xi_{n,t+1}^{(1)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)};\xi_{n,t+1}^{(K-1)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)};\xi_{n,t+1}^{(K)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1};\zeta_{n,t+1})\Big\|^2\Big]\\
&\triangleq T_1+\rho_x^2\eta^2 T_2\,,
\end{aligned}
\tag{21}
$$

where the last step follows from that the sampling operations in different levels are independent.

Then, for T_1, we bound it as follows:

$$
\begin{aligned}
T_1 &\le (1-\rho_x\eta)^2(1+c)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}(p_{n,t}-\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}),y_{n,t}))\Big\|^2\Big]\\
&+ 2(1-\rho_x\eta)^2(1+1/c)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}),y_{n,t})\\
&\quad- \frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1})\Big\|^2\Big]\\
&+ 2\rho_x^2\eta^2(1+1/c)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1})\\
&\quad- \frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(h_{n,t+1}^{(0)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1})\Big\|^2\Big]\\
&\le (1-\rho_x\eta)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}(p_{n,t}-\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}),y_{n,t}))\Big\|^2\Big]\\
&+ \frac{2}{\rho_x\eta}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}),y_{n,t})\\
&\quad- \frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1})\Big\|^2\Big]\\
&+ 2\rho_x\eta\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1})
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(h_{n,t+1}^{(0)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\nabla_1 f_n(h_{n,t+1}^{(K)}, y_{n,t+1})\Big\|^2\Big] \\
&\triangleq (1-\rho_x\eta)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}(p_{n,t} - \nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}), y_{n,t}))\Big\|^2\Big] \\
&+ \frac{2}{\rho_x\eta}T_{1,1} + 2\rho_x\eta T_{1,2}\ , \qquad (22)
\end{aligned}
$$

where the last inequality follows from $c = \frac{\rho_x\eta}{1-\rho_x\eta}$. In the following, we bound $T_{1,1}$ and $T_{1,2}$, respectively.

$$
\begin{aligned}
T_{1,1} \leq\ & \frac{K+1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}), y_{n,t}) \\
&- \nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}), y_{n,t})\Big\|^2\Big] \\
&+ \frac{K+1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}), y_{n,t}) \\
&- \nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}), y_{n,t})\Big\|^2\Big] \\
&+ \cdots \\
&+ \frac{K+1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}), y_{n,t}) \\
&- \nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}), y_{n,t+1})\Big\|^2\Big] \\
\leq\ & \frac{K+1}{N}\sum_{n=1}^{N}C_g^{2(K-1)}C_f^2L_g^2\mathbb{E}\Big[\Big\|x_{n,t} - x_{n,t+1}\Big\|^2\Big] + \frac{K+1}{N}\sum_{n=1}^{N}C_g^{2(K-1)}C_f^2L_g^2\mathbb{E}\Big[\Big\|G_n^{(1)}(x_{n,t}) - G_n^{(1)}(x_{n,t+1})\Big\|^2\Big] \\
&+ \cdots + \frac{K+1}{N}\sum_{n=1}^{N}C_g^{2(K-1)}C_f^2L_g^2\mathbb{E}\Big[\Big\|G_n^{(K-1)}(x_{n,t}) - G_n^{(K-1)}(x_{n,t+1})\Big\|^2\Big] \\
&+ \frac{K+1}{N}\sum_{n=1}^{N}C_g^{2K}L_f^2\mathbb{E}\Big[\Big\|G_n^{(K)}(x_{n,t}) - G_n^{(K)}(x_{n,t+1})\|^2 + \frac{K+1}{N}\sum_{n=1}^{N}C_g^{2K}L_f^2\mathbb{E}\Big[\Big\|y_{n,t} - y_{n,t+1}\Big\|^2\Big] \\
\leq\ & \frac{K+1}{N}\sum_{n=1}^{N}C_g^{2(K-1)}C_f^2L_g^2\mathbb{E}\Big[\Big\|x_{n,t} - x_{n,t+1}\Big\|^2\Big] + \frac{K+1}{N}\sum_{n=1}^{N}C_g^{2(K-1)}C_f^2L_g^2C_g^2\mathbb{E}\Big[\Big\|x_{n,t} - x_{n,t+1}\Big\|^2\Big] \\
&+ \cdots + \frac{K+1}{N}\sum_{n=1}^{N}C_g^{2(K-1)}C_f^2L_g^2C_g^{2(K-1)}\mathbb{E}\Big[\Big\|x_{n,t} - x_{n,t+1}\Big\|^2\Big] \\
&+ \frac{K+1}{N}\sum_{n=1}^{N}C_g^{2K}L_f^2C_g^{2K}\mathbb{E}\Big[\Big\|x_{n,t} - x_{n,t+1}\Big\|^2\Big] + \frac{K+1}{N}\sum_{n=1}^{N}C_g^{2K}L_f^2\mathbb{E}\Big[\Big\|y_{n,t} - y_{n,t+1}\Big\|^2\Big] \\
=\ & \frac{K+1}{N}\sum_{n=1}^{N}\Big(C_g^{2K}L_f^2C_g^{2K} + \sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2\Big)\gamma_x^2\eta^2\mathbb{E}\Big[\Big\|p_{n,t}\Big\|^2\Big] + \frac{K+1}{N}\sum_{n=1}^{N}C_g^{2K}L_f^2\gamma_y^2\eta^2\mathbb{E}[|q_{n,t}\Big\|^2\Big] \\
\leq\ & 2\gamma_x^2\eta^2\Big(C_g^{2K}L_f^2C_g^{2K} + \sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2\Big)\frac{K+1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|p_{n,t} - \bar{p}_t|^2\Big] \\
&+ 2\gamma_x^2\eta^2(K+1)\Big(C_g^{2K}L_f^2C_g^{2K} + \sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2\Big)\mathbb{E}\Big[\Big\|\bar{p}_t\Big\|^2\Big] \\
&+ 2\gamma_y^2\eta^2C_g^{2K}L_f^2\frac{K+1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t} - \bar{q}_t\Big\|^2\Big] + 2\gamma_y^2\eta^2C_g^{2K}L_f^2(K+1)\mathbb{E}\Big[\Big\|\bar{q}_t\Big\|^2\Big]\ , \qquad (23)
\end{aligned}
$$

where the second inequality follows from Assumption 4.1, the third inequality follows from Lemma B.1. For $T_{1,2}$, due to $h_{n,t+1}^{(0)} = x_{n,t+1}$,

we have

$$T_{1,2} \leq \frac{1}{N}\sum_{n=1}^{N} C_g^2 \mathbb{E}\Big[\Big\|\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1}) \\ - \nabla g_n^{(2)}(h_{n,t+1}^{(1)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1})\Big\|^2\Big]. \tag{24}$$

Then, we have

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1}) \\
&\quad - \nabla g_n^{(2)}(h_{n,t+1}^{(1)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1})\Big\|\Big] \\
&\leq \mathbb{E}\Big[\Big\|\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1}) \\
&\quad - \nabla g_n^{(2)}(h_{n,t+1}^{(1)})\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1}) \\
&\quad + \nabla g_n^{(2)}(h_{n,t+1}^{(1)})\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1}) \\
&\quad + \cdots \\
&\quad + \nabla g_n^{(2)}(h_{n,t+1}^{(1)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1}) \\
&\quad - \nabla g_n^{(2)}(h_{n,t+1}^{(1)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1})\Big\|\Big] \\
&\leq \mathbb{E}\Big[\Big\|\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1}) \\
&\quad - \nabla g_n^{(2)}(h_{n,t+1}^{(1)})\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1})\Big\|\Big] \\
&\quad + \cdots \\
&\quad + \mathbb{E}\Big[\Big\|\nabla g_n^{(2)}(h_{n,t+1}^{(1)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1}) \\
&\quad - \nabla g_n^{(2)}(h_{n,t+1}^{(1)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1})\Big\|\Big] \\
&\leq C_g^{K-2}C_f\mathbb{E}\Big[\Big\|\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1})) - \nabla g_n^{(2)}(h_{n,t+1}^{(1)})\Big\|\Big] + C_g^{K-2}C_f\mathbb{E}\Big[\Big\|\nabla g_n^{(3)}(G_n^{(2)}(x_{n,t+1})) - \nabla g_n^{(3)}(h_{n,t+1}^{(2)})\Big\|\Big] \\
&\quad + \cdots + C_g^{K-2}C_f\mathbb{E}\Big[\Big\|\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1})) - \nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\Big\|\Big] \\
&\quad + C_g^{K-1}\mathbb{E}\Big[\Big\|\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1}) - \nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1})\Big\|\Big] \\
&\leq C_g^{K-2}C_f L_g\mathbb{E}\Big[\Big\|G_n^{(1)}(x_{n,t+1}) - h_{n,t+1}^{(1)}\|\Big] + C_g^{K-2}C_f L_g\mathbb{E}\Big[\Big\|G_n^{(2)}(x_{n,t+1}) - h_{n,t+1}^{(2)}\|\Big] \\
&\quad + \cdots + C_g^{K-2}C_f L_g\mathbb{E}\Big[\Big\|G_n^{(K-1)}(x_{n,t+1}) - h_{n,t+1}^{(K-1)}\|\Big] + C_g^{K-1}L_f\mathbb{E}\Big[\Big\|G_n^{(K)}(x_{n,t+1}) - h_{n,t+1}^{(K)}\|\Big] \\
&\leq C_g^{K-2}C_f L_g\sum_{k=1}^{K-1}\mathbb{E}\Big[\Big\|G_n^{(k)}(x_{n,t+1}) - h_{n,t+1}^{(k)}\|\Big] + C_g^{K-1}L_f\mathbb{E}\Big[\Big\|G_n^{(K)}(x_{n,t+1}) - h_{n,t+1}^{(K)}\|\Big] \\
&\leq C_g^{K-2}C_f L_g\sum_{k=1}^{K-1}\sum_{j=1}^{k} C_g^{(k-j)}\mathbb{E}\Big[\Big\|g_n^{(j)}(h_{n,t+1}^{(j-1)}) - h_{n,t+1}^{(j)}\|\Big] + C_g^{K-1}L_f\sum_{j=1}^{K} C_g^{K-j}\mathbb{E}\Big[\Big\|g_n^{(j)}(h_{n,t+1}^{(j-1)}) - h_{n,t+1}^{(j)}\|\Big] \\
&\leq C_g^{K-2}C_f L_g\sum_{k=1}^{K-1}\Big(\sum_{j=k}^{K-1} C_g^{j-k}\Big)\mathbb{E}\Big[\Big\|g_n^{(k)}(h_{n,t+1}^{(k-1)}) - h_{n,t+1}^{(k)}\|\Big] + C_g^{K-1}L_f\sum_{k=1}^{K} C_g^{K-k}\mathbb{E}\Big[\Big\|g_n^{(k)}(h_{n,t+1}^{(k-1)}) - h_{n,t+1}^{(k)}\|\Big],
\end{aligned} \tag{25}$$

where the second to last inequality follows from the following inequality.

For any $k \in \{1,\cdots,K\}$, we have

$$\begin{aligned}
\|G_n^{(k)}(x_{n,t}) - h_{n,t}^{(k)}\| &= \|g_n^{(k)}(G_n^{(k-1)}(x_{n,t})) - g_n^{(k)}(h_{n,t}^{(k-1)}) + g_n^{(k)}(h_{n,t}^{(k-1)}) - h_{n,t}^{(k)}\| \\
&\leq \|g_n^{(k)}(G_n^{(k-1)}(x_{n,t})) - g_n^{(k)}(h_{n,t}^{(k-1)})\| + \|g_n^{(k)}(h_{n,t}^{(k-1)}) - h_{n,t}^{(k)}\| \\
&\leq C_g\|G_n^{(k-1)}(x_{n,t}) - h_{n,t}^{(k-1)}\| + \|g_n^{(k)}(h_{n,t}^{(k-1)}) - h_{n,t}^{(k)}\| \\
&\leq \sum_{j=1}^{k} C_g^{k-j}\|g_n^{(j)}(h_{n,t}^{(j-1)}) - h_{n,t}^{(j)}\|,
\end{aligned} \tag{26}$$

where the second inequality follows from Assumption 4.1. Then, we have

$$
\begin{aligned}
T_{1,2} \leq & \frac{1}{N}\sum_{n=1}^{N} C_g^2 \mathbb{E}\Big[\Big\|\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1}) \\
& - \nabla g_n^{(2)}(h_{n,t+1}^{(1)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1})\Big\|^2\Big] \\
\leq & 2C_g^{2(K-1)}C_f^2 L_g^2 \frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Big(\sum_{j=k}^{K-1} C_g^{j-k}\Big)^2 \mathbb{E}\Big[\Big\|g_n^{(k)}(h_{n,t+1}^{(k-1)}) - h_{n,t+1}^{(k)}|^2\Big] \\
& + 2C_g^{2K} L_f^2 \frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} C_g^{2(K-k)}\mathbb{E}\Big[\Big\|g_n^{(k)}(h_{n,t+1}^{(k-1)}) - h_{n,t+1}^{(k)}\Big\|^2\Big]\,. \qquad (27)
\end{aligned}
$$

Then, we bound T_2 as follows:

$$
\begin{aligned}
T_2 = & \mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(h_{n,t+1}^{(0)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1}) \\
& - \frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(h_{n,t+1}^{(0)};\xi_{n,t+1}^{(1)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1}) \\
& + \cdots \\
& + \frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(h_{n,t+1}^{(0)};\xi_{n,t+1}^{(1)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)};\xi_{n,t+1}^{(K-1)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)};\xi_{n,t+1}^{(K)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1}) \\
& - \frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(h_{n,t+1}^{(0)};\xi_{n,t+1}^{(1)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)};\xi_{n,t+1}^{(K-1)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)};\xi_{n,t+1}^{(K)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1};\zeta_{n,t+1})\Big\|^2\Big] \\
\leq & (K+1)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(h_{n,t+1}^{(0)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1}) \\
& - \frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(h_{n,t+1}^{(0)};\xi_{n,t+1}^{(1)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1})\Big\|^2\Big] \\
& + \cdots \\
& + (K+1)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(h_{n,t+1}^{(0)};\xi_{n,t+1}^{(1)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)};\xi_{n,t+1}^{(K-1)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)};\xi_{n,t+1}^{(K)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1}) \\
& - \frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(h_{n,t+1}^{(0)};\xi_{n,t+1}^{(1)})\cdots\nabla g_n^{(K-1)}(h_{n,t+1}^{(K-2)};\xi_{n,t+1}^{(K-1)})\nabla g_n^{(K)}(h_{n,t+1}^{(K-1)};\xi_{n,t+1}^{(K)})\nabla_1 f_n(h_{n,t+1}^{(K)},y_{n,t+1};\zeta_{n,t+1})\Big\|^2\Big] \\
\leq & \frac{K(K+1)C_g^{2(K-1)}C_f^2}{N}\sigma^2 + \frac{(K+1)C_g^{2K}}{N}\sigma^2\,, \qquad (28)
\end{aligned}
$$

where the last inequality follows from that different workers have independent sampling operation in different levels.

Combine the result from T_1 and T_2, we have

$$
\begin{aligned}
& \mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N} p_{n,t+1} - \frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1})\Big\|^2\Big] \\
& \leq (1-\rho_x\eta)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}(p_{n,t} - \nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}),y_{n,t}))\Big\|^2\Big] \\
& \quad + \frac{2}{\rho_x\eta}T_{1,1} + 2\rho_x\eta T_{1,2} \\
& \leq (1-\rho_x\eta)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N} p_{n,t} - \frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}),y_{n,t})\Big\|^2\Big] \\
& \quad + \frac{4\eta\gamma_x^2}{\rho_x}\Big(C_g^{4K}L_f^2 + \sum_{k=0}^{K-1} C_g^{2(K-1+k)}C_f^2 L_g^2\Big)\frac{K+1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|p_{n,t} - \bar{p}_t\Big\|^2\Big]
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{4\eta\gamma_x^2}{\rho_x}(K+1)\Big(C_g^{4K}L_f^2+\sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2\Big)\mathbb{E}[\|\bar{p}_t\|^2]\\
&+\frac{4\eta\gamma_y^2}{\rho_x}C_g^{2K}L_f^2\frac{K+1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\|q_{n,t}-\bar{q}_t\|^2\Big]+\frac{4\eta\gamma_y^2}{\rho_x}C_g^{2K}L_f^2(K+1)\mathbb{E}\Big[\|\bar{q}_t\|^2\Big]\\
&+4\rho_x\eta C_g^{2(K-1)}C_f^2L_g^2\frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\Big(\sum_{j=k}^{K-1}C_g^{j-k}\Big)^2\mathbb{E}\Big[\|g_n^{(k)}(h_{n,t+1}^{(k-1)})-h_{n,t+1}^{(k)}\|^2\Big]\\
&+4\rho_x\eta C_g^{2K}L_f^2\frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}C_g^{2(K-k)}\mathbb{E}\Big[\|g_n^{(k)}(h_{n,t+1}^{(k-1)})-h_{n,t+1}^{(k)}\|^2\Big]+\rho_x^2\eta^2(K(K+1)C_g^{2(K-1)}C_f^2+(K+1)C_g^{2K})\frac{\sigma^2}{N}\\
\le\;&(1-\rho_x\eta)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}p_{n,t}-\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}),y_{n,t})\Big\|^2\Big]\\
&+\frac{4\eta\gamma_x^2}{\rho_x}\Big(C_g^{4K}L_f^2+\sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2\Big)\frac{K+1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\|p_{n,t}-\bar{p}_t\|^2\Big]\\
&+\frac{4\eta\gamma_x^2}{\rho_x}(K+1)\Big(C_g^{4K}L_f^2+\sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2\Big)\mathbb{E}\Big[\|\bar{p}_t\|^2\Big]\\
&+\frac{4\eta\gamma_y^2}{\rho_x}C_g^{2K}L_f^2\frac{K+1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\|q_{n,t}-\bar{q}_t\|^2\Big]+\frac{4\eta\gamma_y^2}{\rho_x}C_g^{2K}L_f^2(K+1)\mathbb{E}\Big[\|\bar{q}_t\|^2\Big]\\
&+4\rho_x\eta\frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}A_k\mathbb{E}\Big[\|g_n^{(k)}(h_{n,t+1}^{(k-1)})-h_{n,t+1}^{(k)}\|^2\Big]+\rho_x^2\eta^2(K(K+1)C_g^{2(K-1)}C_f^2+(K+1)C_g^{2K})\frac{\sigma^2}{N}\,, \qquad (29)
\end{aligned}
$$

where $A_k=\left(C_g^{2(K-1)}C_f^2L_g^2\Big(\sum_{j'=k}^{K-1}C_g^{j'-k}\Big)^2+C_g^{2K}L_f^2C_g^{2(K-k)}\right)$.

Then, from Lemma B.4, we can finally get

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}p_{n,t+1}-\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}),y_{n,t+1})\Big\|^2\Big]\\
&\le(1-\rho_x\eta)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}p_{n,t}-\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}),y_{n,t})\Big\|^2\Big]\\
&+\frac{4\eta\gamma_x^2}{\rho_x}\Big(C_g^{4K}L_f^2+\sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2\Big)\frac{K+1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\|p_{n,t}-\bar{p}_t\|^2\Big]\\
&+\frac{4\eta\gamma_x^2}{\rho_x}(K+1)\Big(C_g^{4K}L_f^2+\sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2\Big)\mathbb{E}\Big[\|\bar{p}_t\|^2\Big]\\
&+\frac{4\eta\gamma_y^2}{\rho_x}C_g^{2K}L_f^2\frac{K+1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\|q_{n,t}-\bar{q}_t\|^2\Big]+\frac{4\eta\gamma_y^2}{\rho_x}C_g^{2K}L_f^2(K+1)\mathbb{E}\Big[\|\bar{q}_t\|^2\Big]\\
&+4\rho_x\eta\frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}A_k\mathbb{E}\Big[\|h_{n,t}^{(k)}-g_n^{(k)}(h_{n,t}^{(k-1)})\|^2\Big]+8\rho_x\alpha^2\eta^5\frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\sum_{j=k+1}^{K}A_j(2C_g^2)^{j-k}\Big)\mathbb{E}\Big[\|h_{n,t}^{(k)}-g_n^{(k)}(h_{n,t}^{(k-1)})\|^2\Big]\\
&+8\rho_x\gamma_x^2\eta^3K\sum_{k=1}^{K}A_k(2C_g^2)^k\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\|p_{n,t}-\bar{p}_t\|^2\Big]+8\rho_x\gamma_x^2\eta^3K\sum_{k=1}^{K}A_k(2C_g^2)^k\mathbb{E}\Big[\|\bar{p}_t\|^2\Big]\\
&+8\rho_x\alpha^2\eta^5\delta^2K\sum_{k=1}^{K}A_k\sum_{j=1}^{k-1}(2C_g^2)^j+\rho_x^2\eta^2(K(K+1)C_g^{2(K-1)}C_f^2+(K+1)C_g^{2K})\frac{\sigma^2}{N}\,,
\end{aligned}
$$

where $A_k=\left(C_g^{2(K-1)}C_f^2L_g^2\Big(\sum_{j'=k}^{K-1}C_g^{j'-k}\Big)^2+C_g^{2K}L_f^2C_g^{2(K-k)}\right)$, the last step follows from Lemma B.4.

□

Lemma B.3. *Given Assumption 4.1-4.3, we can know*

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}q_{n,t+1}-\frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(G_n(x_{n,t+1}),y_{n,t+1})\Big\|^2\Big]\\
&\le(1-\rho_y\eta)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}q_{n,t}-\frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(G_n(x_{n,t}),y_{n,t})\Big\|^2\Big]\\
&\quad+\frac{4\eta\gamma_x^2L_f^2C_G^2}{\rho_y}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|p_{n,t}-\bar{p}_t\Big\|^2\Big]+\frac{4\eta\gamma_x^2L_f^2C_G^2}{\rho_y}\mathbb{E}\Big[\Big\|\bar{p}_t\Big\|^2\Big]\\
&\quad+\frac{4\eta\gamma_y^2L_f^2}{\rho_y}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t}-\bar{q}_t\Big\|^2\Big]+\frac{4\eta\gamma_y^2L_f^2}{\rho_y}\mathbb{E}\Big[\Big\|\bar{q}_t\Big\|^2\Big]+\rho_y^2\eta^2\frac{\sigma^2}{N}+4\alpha^2\rho_y\eta^5\delta^2L_f^2K\sum_{k=1}^{K}\sum_{j=1}^{k-1}C_g^{2(K-(k-j))}\\
&\quad+2\rho_y\eta L_f^2\frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}C_g^{2(K-k)}\mathbb{E}\Big[\Big\|h_{n,t}^{(k)}-g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big]\\
&\quad+4\rho_y\alpha^2\eta^5L_f^2\frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(2^{K+1}C_g^{2(K-k)}\Big)\mathbb{E}\Big[\Big\|h_{n,t}^{(k)}-g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big]\\
&\quad+4\rho_y\gamma_x^2\eta^3K^2C_G^2C_f^2L_f^2(2C_g^2)^K\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|p_{n,t}-\bar{p}_t\Big\|^2\Big]+4\rho_y\gamma_x^2\eta^3K^2C_G^2C_f^2L_f^2(2C_g^2)^K\mathbb{E}\Big[\Big\|\bar{p}_t\Big\|^2\Big]. \qquad (23)
\end{aligned}
$$

Proof.

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}q_{n,t+1}-\frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(G_n(x_{n,t+1}),y_{n,t+1})\Big\|^2\Big]\\
&=\mathbb{E}\Big[\Big\|(1-\rho_y\eta)\frac{1}{N}\sum_{n=1}^{N}(q_{n,t}-\nabla_2 f_n(G_n(x_{n,t}),y_{n,t}))\\
&\quad+(1-\rho_y\eta)\frac{1}{N}\sum_{n=1}^{N}\Big(\nabla_2 f_n(G_n(x_{n,t}),y_{n,t})-\nabla_2 f_n(G_n(x_{n,t+1}),y_{n,t+1})\Big)\\
&\quad+\rho_y\eta\frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(h_{n,t+1}^{(K)},y_{n,t+1};\zeta_{n,t+1})-\rho_y\eta\frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(h_{n,t+1}^{(K)},y_{n,t+1})\\
&\quad+\rho_y\eta\frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(h_{n,t+1}^{(K)},y_{n,t+1})-\rho_y\eta\frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(G(x_{n,t+1}),y_{n,t+1})\Big\|^2\Big]\\
&\le(1-\rho_y\eta)^2(1+c)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}q_{n,t}-\frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(G_n(x_{n,t}),y_{n,t})\Big\|^2\Big]\\
&\quad+2(1+1/c)(1-\rho_y\eta)^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\nabla_2 f_n(G_n(x_{n,t}),y_{n,t})-\nabla_2 f_n(G_n(x_{n,t+1}),y_{n,t+1})\Big\|^2\Big]\\
&\quad+2(1+1/c)\rho_y^2\eta^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\nabla_2 f_n(h_{n,t+1}^{(K)},y_{n,t+1})-\nabla_2 f_n(G_n(x_{n,t+1}),y_{n,t+1})\Big\|^2\Big]+\rho_y^2\eta^2\frac{\sigma^2}{N}\\
&\le(1-\rho_y\eta)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}q_{n,t}-\frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(G_n(x_{n,t}),y_{n,t})\Big\|^2\Big]\\
&\quad+\frac{2}{\rho_y\eta}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\nabla_2 f_n(G_n(x_{n,t}),y_{n,t})-\nabla_2 f_n(G_n(x_{n,t+1}),y_{n,t+1})\Big\|^2\Big]
\end{aligned}
$$

$$
\begin{aligned}
&+ 2\rho_y\eta\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\nabla_2 f_n(h_{n,t+1}^{(K)}, y_{n,t+1}) - \nabla_2 f_n(G_n(x_{n,t+1}), y_{n,t+1})\Big\|^2\Big] + \rho_y^2\eta^2\frac{\sigma^2}{N}\,,\\
&\leq (1-\rho_y\eta)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}q_{n,t} - \frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(G(x_{n,t}), y_{n,t})\Big\|^2\Big]\\
&\quad + \frac{2L_f^2C_G^2}{\rho_y\eta}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|x_{n,t} - x_{n,t+1}\Big\|^2\Big] + \frac{2L_f^2}{\rho_y\eta}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|y_{n,t} - y_{n,t+1}\Big\|^2\Big]\\
&\quad + 2\rho_y\eta L_f^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t+1}^{(K)} - G_n(x_{n,t+1})\Big\|^2\Big] + \rho_y^2\eta^2\frac{\sigma^2}{N}\\
&\leq (1-\rho_y\eta)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}q_{n,t} - \frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(G_n(x_{n,t}), y_{n,t})\Big\|^2\Big]\\
&\quad + \frac{4\eta\gamma_x^2L_f^2C_G^2}{\rho_y}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|p_{n,t} - \bar{p}_t\Big\|^2\Big] + \frac{4\eta\gamma_x^2L_f^2C_G^2}{\rho_y}\mathbb{E}\Big[\Big\|\bar{p}_t\Big\|^2\Big]\\
&\quad + \frac{4\eta\gamma_y^2L_f^2}{\rho_y}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t} - \bar{q}_t\Big\|^2\Big] + \frac{4\eta\gamma_y^2L_f^2}{\rho_y}\mathbb{E}\Big[\Big\|\bar{q}_t\Big\|^2\Big]\\
&\quad + 2\rho_y\eta L_f^2\frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}C_g^{2(K-k)}\mathbb{E}\Big[\Big\|g_n^{(k)}(h_{n,t+1}^{(k-1)}) - h_{n,t+1}^{(k)}\Big\|^2\Big] + \rho_y^2\eta^2\frac{\sigma^2}{N}\,,
\end{aligned}
\tag{24}
$$

where the third step follows from $c = \frac{\rho_y\eta}{1-\rho_y\eta}$, the last step follows from Eq. (26).

Then, based on Lemma B.4, we obtain

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}q_{n,t+1} - \frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(G_n(x_{n,t+1}), y_{n,t+1})\Big\|^2\Big]\\
&\leq (1-\rho_y\eta)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}q_{n,t} - \frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(G_n(x_{n,t}), y_{n,t})\Big\|^2\Big]\\
&\quad + \frac{4\eta\gamma_x^2L_f^2C_G^2}{\rho_y}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|p_{n,t} - \bar{p}_t\Big\|^2\Big] + \frac{4\eta\gamma_x^2L_f^2C_G^2}{\rho_y}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\bar{p}_t\Big\|^2\Big]\\
&\quad + \frac{4\eta\gamma_y^2L_f^2}{\rho_y}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t} - \bar{q}_t\Big\|^2\Big] + \frac{4\eta\gamma_y^2L_f^2}{\rho_y}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\bar{q}_t\Big\|^2\Big] + \rho_y^2\eta^2\frac{\sigma^2}{N} + 4\alpha^2\rho_y\eta^5\delta^2L_f^2K\sum_{k=1}^{K}\sum_{j=1}^{k-1}C_g^{2(K-(k-j))}\\
&\quad + 2\rho_y\eta L_f^2\frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}C_g^{2(K-k)}\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big]\\
&\quad + 4\rho_y\alpha^2\eta^5L_f^2\frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(2^{K+1}C_g^{2(K-k)}\Big)\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big]\\
&\quad + 4\rho_y\gamma_x^2\eta^3K^2C_G^2C_f^2L_f^2(2C_g^2)^K\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|p_{n,t} - \bar{p}_t\Big\|^2\Big] + 4\rho_y\gamma_x^2\eta^3K^2C_G^2C_f^2L_f^2(2C_g^2)^K\mathbb{E}\Big[\Big\|\bar{p}_t\Big\|^2\Big]\,.
\end{aligned}
\tag{25}
$$

□

Lemma B.4. *Given Assumption 4.1-4.3, for $k \in \{1, \cdots, K\}$, we can know*

$$
\mathbb{E}\Big[\Big\|h_{n,t+1}^{(k)} - g_n^{(k)}(h_{n,t+1}^{(k-1)})\Big\|^2\Big] \leq (1-\alpha\eta^2)\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] + 2C_g^2\mathbb{E}\Big[\Big\|h_{n,t+1}^{(k-1)} - h_{n,t}^{(k-1)}\Big\|^2\Big] + 2\alpha^2\eta^4\delta^2\,.
\tag{26}
$$

Additionally, for any $a_k > 0$ where $k \in \{1, \cdots, K\}$, we have

$$\begin{aligned}
&\sum_{k=1}^{K} a_k \mathbb{E}\Big[\Big\|h_{n,t+1}^{(k)} - g_n^{(k)}(h_{n,t+1}^{(k-1)})\Big\|^2\Big] \leq (1-\alpha\eta^2)\sum_{k=1}^{K} a_k \mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] \\
&+ 2\alpha^2\eta^4 \sum_{k=1}^{K}\Big(\sum_{j=k+1}^{K} a_j (2C_g^2)^{j-k}\Big)\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] \\
&+ 2\gamma_x^2\eta^2 \sum_{k=1}^{K} a_k (2C_g^2)^k \mathbb{E}\Big[\Big\|p_{n,t} - \bar{p}_t\Big\|^2\Big] + 2\gamma_x^2\eta^2 \sum_{k=1}^{K} a_k (2C_g^2)^k \mathbb{E}\Big[\Big\|\bar{p}_t\Big\|^2\Big] + 2\alpha^2\eta^4\delta^2 \sum_{k=1}^{K} a_k \sum_{j=1}^{k-1}(2C_g^2)^j \,. \qquad (27)
\end{aligned}$$

Proof. For $k \in \{1, \cdots, K\}$, we have

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|h_{n,t+1}^{(k)} - g_n^{(k)}(h_{n,t+1}^{(k-1)})\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|(1-\alpha\eta^2)(h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)}; \xi_{n,t+1}^{(k)})) + g_n^{(k)}(h_{n,t+1}^{(k-1)}; \xi_{n,t+1}^{(k)}) - g_n^{(k)}(h_{n,t+1}^{(k-1)})\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|(1-\alpha\eta^2)(h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})) \\
&\quad + (1-\alpha\eta^2)(g_n^{(k)}(h_{n,t+1}^{(k-1)}; \xi_{n,t+1}^{(k)}) - g_n^{(k)}(h_{n,t}^{(k-1)}; \xi_{n,t+1}^{(k)}) + g_n^{(k)}(h_{n,t}^{(k-1)}) - g_n^{(k)}(h_{n,t+1}^{(k-1)})) \\
&\quad + \alpha\eta^2 g_n^{(k)}(h_{n,t+1}^{(k-1)}; \xi_{n,t+1}^{(k)}) - \alpha\eta(g_n^{(k)}(h_{n,t+1}^{(k-1)}))\Big\|^2\Big] \\
&\leq (1-\alpha\eta^2)^2 \mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] \\
&\quad + 2(1-\alpha\eta^2)^2 \mathbb{E}\Big[\Big\|g_n^{(k)}(h_{n,t+1}^{(k-1)}; \xi_{n,t+1}^{(k)}) - g_n^{(k)}(h_{n,t}^{(k-1)}; \xi_{n,t+1}^{(k)}) + g_n^{(k)}(h_{n,t}^{(k-1)}) - g_n^{(k)}(h_{n,t+1}^{(k-1)})\Big\|^2\Big] \\
&\quad + 2\alpha^2\eta^4 \mathbb{E}\Big[\Big\|g_n^{(k)}(h_{n,t+1}^{(k-1)}; \xi_{n,t+1}^{(k)}) - g_n^{(k)}(h_{n,t+1}^{(k-1)})\Big\|^2\Big] \\
&\leq (1-\alpha\eta^2)^2 \mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] + 2(1-\alpha\eta^2)^2 \mathbb{E}\Big[\|g_n^{(k)}(h_{n,t+1}^{(k-1)}; \xi_{n,t+1}^{(k)}) - g_n^{(k)}(h_{n,t}^{(k-1)}; \xi_{n,t+1}^{(k)})\Big\|^2\Big] + 2\alpha^2\eta^4\delta^2 \\
&\leq (1-\alpha\eta^2)\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] + 2C_g^2 \mathbb{E}\Big[\Big\|h_{n,t+1}^{(k-1)} - h_{n,t}^{(k-1)}\Big\|^2\Big] + 2\alpha^2\eta^4\delta^2 \,, \qquad (28)
\end{aligned}$$

where the last step follows from $\alpha\eta^2 \leq 1$ and Assumptions 4.1-4.2.
Then, for any $a_k > 0$ where $k \in \{1, \cdots, K\}$, we have

$$a_k \mathbb{E}\Big[\Big\|h_{n,t+1}^{(k)} - g_n^{(k)}(h_{n,t+1}^{(k-1)})\Big\|^2\Big] \leq (1-\alpha\eta^2) a_k \mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] + 2a_k C_g^2 \mathbb{E}\Big[\Big\|h_{n,t+1}^{(k-1)} - h_{n,t}^{(k-1)}\Big\|^2\Big] + 2\alpha^2\eta^4\delta^2 a_k \,. \qquad (29)$$

By summing over k from 1 to K, we have

$$\begin{aligned}
&\sum_{k=1}^{K} a_k \mathbb{E}\Big[\Big\|h_{n,t+1}^{(k)} - g_n^{(k)}(h_{n,t+1}^{(k-1)})\Big\|^2\Big] \\
&\leq (1-\alpha\eta^2)\sum_{k=1}^{K} a_k \mathbb{E}\Big[\Big\| h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] + 2C_g^2 \sum_{k=1}^{K} a_k \mathbb{E}\Big[\Big\|h_{n,t+1}^{(k-1)} - h_{n,t}^{(k-1)}\Big\|^2\Big] + 2\alpha^2\eta^4\delta^2 \sum_{k=1}^{K} a_k \\
&= (1-\alpha\eta^2)\sum_{k=1}^{K} a_k \mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] + 2a_1 C_g^2 \mathbb{E}\Big[\Big\|x_{n,t+1} - x_{n,t}\Big\|^2\Big] \\
&\quad + 2\alpha^2\eta^4\delta^2 \sum_{k=1}^{K} a_k + 2C_g^2 \sum_{k=2}^{K} a_k \mathbb{E}\Big[\Big\|h_{n,t+1}^{(k-1)} - h_{n,t}^{(k-1)}\Big\|^2\Big] \,. \qquad (30)
\end{aligned}$$

From Lemma B.5, we can get

$$
\begin{aligned}
&\sum_{k=1}^{K} a_k \mathbb{E}\Big[\Big\|h_{n,t+1}^{(k)} - g_n^{(k)}(h_{n,t+1}^{(k-1)})\Big\|^2\Big] \\
&\leq (1-\alpha\eta^2)\sum_{k=1}^{K} a_k \mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] + 2a_1\gamma_x^2\eta^2 C_g^2 \mathbb{E}\Big[\Big\|p_{n,t}\Big\|^2\Big] \\
&\quad + 2\alpha^2\eta^4\delta^2 \sum_{k=1}^{K} a_k + 4\alpha^2\eta^4\delta^2 C_g^2 \sum_{k=2}^{K} a_k \sum_{j=1}^{k-1} (2C_g^2)^{j-1} \\
&\quad + 4\alpha^2\eta^4 C_g^2 \sum_{k=2}^{K} a_k \sum_{j=1}^{k-1} (2C_g^2)^{k-1-j} \mathbb{E}\Big[\Big\|h_{n,t}^{(j)} - g_n^{(j)}(h_{n,t}^{(j-1)})\Big\|^2\Big] + 2\gamma_x^2\eta^2 C_g^2 \sum_{k=2}^{K} a_k (2C_g^2)^{k-1} \mathbb{E}\Big[\Big\|p_{n,t}\Big\|^2\Big] \\
&= (1-\alpha\eta^2)\sum_{k=1}^{K} a_k \mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] + 2a_1\gamma_x^2\eta^2 C_g^2 \mathbb{E}\Big[\Big\|p_{n,t}\Big\|^2\Big] + 2\alpha^2\eta^4\delta^2 \sum_{k=1}^{K} a_k \sum_{j=1}^{k-1} (2C_g^2)^{j} \\
&\quad + 2\alpha^2\eta^4 \sum_{k=2}^{K} a_k \sum_{j=1}^{k-1} (2C_g^2)^{k-j} \mathbb{E}\Big[\Big\|h_{n,t}^{(j)} - g_n^{(j)}(h_{n,t}^{(j-1)})\Big\|^2\Big] + \gamma_x^2\eta^2 \sum_{k=2}^{K} a_k (2C_g^2)^{k} \mathbb{E}\Big[\Big\|p_{n,t}\Big\|^2\Big] \\
&\leq (1-\alpha\eta^2)\sum_{k=1}^{K} a_k \mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] + 2\alpha^2\eta^4 \sum_{k=1}^{K} \Big(\sum_{j=k+1}^{K} a_j (2C_g^2)^{j-k}\Big) \mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] \\
&\quad + \gamma_x^2\eta^2 \sum_{k=1}^{K} a_k (2C_g^2)^{k} \mathbb{E}\Big[\Big\|p_{n,t}\Big\|^2\Big] + 2\alpha^2\eta^4\delta^2 \sum_{k=1}^{K} a_k \sum_{j=1}^{k-1} (2C_g^2)^{j} \\
&\leq (1-\alpha\eta^2)\sum_{k=1}^{K} a_k \mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] + 2\alpha^2\eta^4 \sum_{k=1}^{K} \Big(\sum_{j=k+1}^{K} a_j (2C_g^2)^{j-k}\Big) \mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] \\
&\quad + 2\gamma_x^2\eta^2 \sum_{k=1}^{K} a_k (2C_g^2)^{k} \mathbb{E}\Big[\Big\|p_{n,t} - \bar{p}_t\Big\|^2\Big] + 2\gamma_x^2\eta^2 \sum_{k=1}^{K} a_k (2C_g^2)^{k} \mathbb{E}\Big[\Big\|\bar{p}_t\Big\|^2\Big] + 2\alpha^2\eta^4\delta^2 \sum_{k=1}^{K} a_k \sum_{j=1}^{k-1} (2C_g^2)^{j} \,. \qquad (31)
\end{aligned}
$$

□

Lemma B.5. *Given Assumption 4.1-4.3, for $k \in \{2, \cdots, K+1\}$, we can know*

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|h_{n,t+1}^{(k-1)} - h_{n,t}^{(k-1)}\Big\|^2\Big] \\
&\leq 2\alpha^2\eta^4 \sum_{j=1}^{k-1} (2C_g^2)^{k-1-j} \mathbb{E}\Big[\Big\|h_{n,t}^{(j)} - g_n^{(j)}(h_{n,t}^{(j-1)})\Big\|^2\Big] + (2C_g^2)^{k-1}\gamma_x^2\eta^2 \mathbb{E}\Big[\Big\|p_{n,t}\Big\|^2\Big] + 2\alpha^2\eta^4\delta^2 \sum_{j=1}^{k-1} (2C_g^2)^{j-1} \,. \qquad (32)
\end{aligned}
$$

Proof. For $k \in \{2, \cdots, K+1\}$, we have

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|h_{n,t+1}^{(k-1)} - h_{n,t}^{(k-1)}\Big\|^2\Big] = \mathbb{E}\Big[\Big\|(1-\alpha\eta^2)(h_{n,t}^{(k-1)} - g_n^{(k-1)}(h_{n,t}^{(k-2)}; \xi_{n,t+1}^{(k-1)})) + g_n^{(k-1)}(h_{n,t+1}^{(k-2)}; \xi_{n,t+1}^{(k-1)}) - h_{n,t}^{(k-1)}\Big\|^2\Big] \\
&\leq 2\mathbb{E}\Big[\Big\|g_n^{(k-1)}(h_{n,t+1}^{(k-2)}; \xi_{n,t+1}^{(k-1)}) - g_n^{(k-1)}(h_{n,t}^{(k-2)}; \xi_{n,t+1}^{(k-1)})\Big\|^2\Big] + 2\alpha^2\eta^4 \mathbb{E}\Big[\Big\|h_{n,t}^{(k-1)} - g_n^{(k-1)}(h_{n,t}^{(k-2)}; \xi_{n,t+1}^{(k-1)})\Big\|^2\Big] \\
&\leq 2C_g^2 \mathbb{E}\Big[\Big\|h_{n,t+1}^{(k-2)} - h_{n,t}^{(k-2)}\Big\|^2\Big] + 2\alpha^2\eta^4 \mathbb{E}\Big[\Big\|h_{n,t}^{(k-1)} - g_n^{(k-1)}(h_{n,t}^{(k-2)})\Big\|^2\Big] + 2\alpha^2\eta^4\delta^2 \\
&\leq (2C_g^2)^{k-1} \mathbb{E}\Big[\Big\|x_{n,t+1} - x_{n,t}\Big\|^2\Big] + 2\alpha^2\eta^4 \sum_{j=1}^{k-1} (2C_g^2)^{k-1-j} \mathbb{E}\Big[\Big\|h_{n,t}^{(j)} - g_n^{(j)}(h_{n,t}^{(j-1)})\Big\|^2\Big] + 2\alpha^2\eta^4\delta^2 \sum_{j=1}^{k-1} (2C_g^2)^{k-1-j} \\
&\leq 2\alpha^2\eta^4 \sum_{j=1}^{k-1} (2C_g^2)^{k-1-j} \mathbb{E}\Big[\Big\|h_{n,t}^{(j)} - g_n^{(j)}(h_{n,t}^{(j-1)})\Big\|^2\Big] + (2C_g^2)^{k-1}\gamma_x^2\eta^2 \mathbb{E}\Big[\Big\|p_{n,t}\Big\|^2\Big] + 2\alpha^2\eta^4\delta^2 \sum_{j=1}^{k-1} (2C_g^2)^{j-1} \,. \qquad (33)
\end{aligned}
$$

□

Lemma B.6. *Given Assumption 4.1-4.3 and $\eta \leq \frac{1}{2\gamma_x L_\Phi}$, we can know*

$$\begin{aligned}
\Phi\left(\bar{x}_{t+1}\right) \leq & \Phi(\bar{x}_t) - \frac{\gamma_x\eta}{2}\|\nabla\Phi(\bar{x}_t)\|^2 - \frac{\gamma_x\eta}{4}\|\bar{p}_t\|^2 + 2\gamma_x\eta C_G^2 L_f^2\|y^*(\bar{x}_t) - \bar{y}_t\|^2 \\
& + 2\gamma_x\eta(K+1)\Big(C_g^{4K}L_f^2 + \sum_{k=0}^{K-1} C_g^{2(K-1+k)}C_f^2L_g^2\Big)\frac{1}{N}\sum_{n=1}^N\|\bar{x}_t - x_{n,t}\|^2 + 2\gamma_x\eta(K+1)C_g^{2K}L_f^2\frac{1}{N}\sum_{n=1}^N\|\bar{y}_t - y_{n,t}\|^2 \\
& + 2\gamma_x\eta\|\frac{1}{N}\sum_{n=1}^N \nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}), y_{n,t}) - \frac{1}{N}\sum_{n=1}^N p_{n,t}\|^2 .
\end{aligned} \tag{34}$$

Proof.

$$\begin{aligned}
\Phi\left(\bar{x}_{t+1}\right) & \leq \Phi(\bar{x}_t) + \langle\nabla\Phi(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t\rangle + \frac{L_\Phi}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2 \\
& = \Phi(\bar{x}_t) - \gamma_x\eta\langle\nabla\Phi(\bar{x}_t), \bar{p}_t\rangle + \frac{\gamma_x^2\eta^2 L_\Phi}{2}\|\bar{p}_t\|^2 \\
& = \Phi(\bar{x}_t) - \frac{\gamma_x\eta}{2}\|\nabla\Phi(\bar{x}_t)\|^2 - (\frac{\gamma_x\eta}{2} - \frac{\gamma_x^2\eta^2 L_\Phi}{2})\|\bar{p}_t\|^2 + \frac{\gamma_x\eta}{2}\|\bar{p}_t - \nabla\Phi(\bar{x}_t)\|^2 \\
& \leq \Phi(\bar{x}_t) - \frac{\gamma_x\eta}{2}\|\nabla\Phi(\bar{x}_t)\|^2 - \frac{\gamma_x\eta}{4}\|\bar{p}_t\|^2 + \frac{\gamma_x\eta}{2}\|\bar{p}_t - \nabla\Phi(\bar{x}_t)\|^2 \\
& \leq \Phi(\bar{x}_t) - \frac{\gamma_x\eta}{2}\|\nabla\Phi(\bar{x}_t)\|^2 - \frac{\gamma_x\eta}{4}\|\bar{p}_t\|^2 + 2\gamma_x\eta\|\frac{1}{N}\sum_{n=1}^N\nabla_x f_n(G(\bar{x}_t), y^*(\bar{x}_t)) - \frac{1}{N}\sum_{n=1}^N\nabla_x f_n(G(\bar{x}_t), \bar{y}_t)\|^2 \\
& + 2\gamma_x\eta\|\frac{1}{N}\sum_{n=1}^N\nabla_x f_n(G(\bar{x}_t), \bar{y}_t) - \frac{1}{N}\sum_{n=1}^N \nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}), y_{n,t})\|^2 \\
& + 2\gamma_x\eta\|\frac{1}{N}\sum_{n=1}^N \nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}), y_{n,t}) - \frac{1}{N}\sum_{n=1}^N p_{n,t}\|^2 \\
& \leq \Phi(\bar{x}_t) - \frac{\gamma_x\eta}{2}\|\nabla\Phi(\bar{x}_t)\|^2 - \frac{\gamma_x\eta}{4}\|\bar{p}_t\|^2 + 2\gamma_x\eta C_G^2 L_f^2\|y^*(\bar{x}_t) - \bar{y}_t\|^2 \\
& + 2\gamma_x\eta(K+1)\Big(C_g^{4K}L_f^2 + \sum_{k=0}^{K-1} C_g^{2(K-1+k)}C_f^2L_g^2\Big)\frac{1}{N}\sum_{n=1}^N\|\bar{x}_t - x_{n,t}\|^2 + 2\gamma_x\eta(K+1)C_g^{2K}L_f^2\frac{1}{N}\sum_{n=1}^N\|\bar{y}_t - y_{n,t}\|^2 \\
& + 2\gamma_x\eta\|\frac{1}{N}\sum_{n=1}^N \nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}), y_{n,t}) - \frac{1}{N}\sum_{n=1}^N p_{n,t}\|^2 ,
\end{aligned} \tag{35}$$

where the second inequality follows from $\eta \leq \frac{1}{2\gamma_x L_\Phi}$, the last inequality follows from the following tow inequalities. First, we have

$$\begin{aligned}
& \|\frac{1}{N}\sum_{n=1}^N\nabla_x f_n(G(\bar{x}_t), y^*(\bar{x}_t)) - \frac{1}{N}\sum_{n=1}^N\nabla_x f_n(G(\bar{x}_t), \bar{y}_t)\|^2 \\
\leq & \frac{1}{N}\sum_{n=1}^N\|\nabla_x f_n(G^{(K)}(\bar{x}_t), y^*(\bar{x}_t)) - \nabla_x f_n(G(\bar{x}_t), \bar{y}_t)\|^2 \\
= & \frac{1}{N}\sum_{n=1}^N\|\nabla G^{(K)}(\bar{x}_t)\nabla_1 f_n(G^{(K)}(\bar{x}_t), y^*(\bar{x}_t)) - \nabla G^{(K)}(\bar{x}_t)\nabla_1 f_n(G^{(K)}(\bar{x}_t), \bar{y}_t)\|^2 \\
\leq & \frac{1}{N}\sum_{n=1}^N\|\nabla G^{(K)}(\bar{x}_t)\|^2\|\nabla_1 f_n(G^{(K)}(\bar{x}_t), y^*(\bar{x}_t)) - \nabla_1 f_n(G^{(K)}(\bar{x}_t), \bar{y}_t)\|^2 \\
\leq & C_G^2 L_f^2\|y^*(\bar{x}_t) - \bar{y}_t\|^2 .
\end{aligned} \tag{36}$$

Second, we have

$$\|\frac{1}{N}\sum_{n=1}^N\nabla_x f_n(G(\bar{x}_t), \bar{y}_t) - \frac{1}{N}\sum_{n=1}^N \nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}), y_{n,t})\|^2$$

$$
\begin{aligned}
&\leq \frac{1}{N}\sum_{n=1}^{N}\|\nabla g^{(1)}(\bar{x}_t)\nabla g^{(2)}(G^{(1)}(\bar{x}_t))\cdots\nabla g^{(K)}(G^{(K-1)}(\bar{x}_t))\nabla_1 f_n(G^{(K)}(\bar{x}_t),\bar{y}_t)\\
&\quad-\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}),y_{n,t})\|^2\\
&= \frac{1}{N}\sum_{n=1}^{N}\|\nabla g_n^{(1)}(\bar{x}_t)\nabla g_n^{(2)}(G_n^{(1)}(\bar{x}_t))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(\bar{x}_t))\nabla_1 f_n(G_n^{(K)}(\bar{x}_t),\bar{y}_t)\\
&\quad-\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}),y_{n,t})\|^2\\
&\leq \frac{1}{N}\sum_{n=1}^{N}\|\nabla g_n^{(1)}(\bar{x}_t)\nabla g_n^{(2)}(G_n^{(1)}(\bar{x}_t))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(\bar{x}_t))\nabla_1 f_n(G_n^{(K)}(\bar{x}_t),\bar{y}_t)\\
&\quad-\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(\bar{x}_t))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(\bar{x}_t))\nabla_1 f_n(G_n^{(K)}(\bar{x}_t),\bar{y}_t)\\
&\quad+\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(\bar{x}_t))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(\bar{x}_t))\nabla_1 f_n(G_n^{(K)}(\bar{x}_t),\bar{y}_t)\\
&\quad-\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(\bar{x}_t))\nabla_1 f_n(G_n^{(K)}(\bar{x}_t),\bar{y}_t)\\
&\quad\cdots\\
&\quad+\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(\bar{x}_t),\bar{y}_t)\\
&\quad-\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}),y_{n,t})\|^2\\
&\leq \frac{K+1}{N}\sum_{n=1}^{N}\|\nabla g_n^{(1)}(\bar{x}_t)\nabla g_n^{(2)}(G_n^{(1)}(\bar{x}_t))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(\bar{x}_t))\nabla_1 f_n(G_n^{(K)}(\bar{x}_t),\bar{y}_t)\\
&\quad-\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(\bar{x}_t))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(\bar{x}_t))\nabla_1 f_n(G_n^{(K)}(\bar{x}_t),\bar{y}_t)\|^2\\
&\quad+\frac{K+1}{N}\sum_{n=1}^{N}\|\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(\bar{x}_t))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(\bar{x}_t))\nabla_1 f_n(G_n^{(K)}(\bar{x}_t),\bar{y}_t)\\
&\quad-\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(\bar{x}_t))\nabla_1 f_n(G_n^{(K)}(\bar{x}_t),\bar{y}_t)\|^2\\
&\quad\cdots\\
&\quad+\frac{K+1}{N}\sum_{n=1}^{N}\|\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(\bar{x}_t),\bar{y}_t)\\
&\quad-\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}),y_{n,t})\|^2\\
&\leq \frac{K+1}{N}\sum_{n=1}^{N}C_g^{2(K-1)}C_f^2\|\nabla g_n^{(1)}(\bar{x}_t)-\nabla g_n^{(1)}(x_{n,t})\|^2+\frac{K+1}{N}\sum_{n=1}^{N}C_g^{2(K-1)}C_f^2\|\nabla g_n^{(2)}(G_n^{(1)}(\bar{x}_t))-\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\|^2\\
&\quad+\cdots+\frac{K+1}{N}\sum_{n=1}^{N}C_g^{2(K-1)}C_f^2\|\nabla g_n^{(K)}(G_n^{(K-1)}(\bar{x}_t))-\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\|^2\\
&\quad+\frac{K+1}{N}\sum_{n=1}^{N}C_g^{2K}\|\nabla_1 f_n(G_n^{(K)}(\bar{x}_t),\bar{y}_t)-\nabla_1 f_n(G_n^{(K)}(x_{n,t}),y_{n,t})\|^2\\
&\leq \frac{K+1}{N}\sum_{n=1}^{N}C_g^{2(K-1)}C_f^2L_g^2\|\bar{x}_t-x_{n,t}\|^2+\frac{K+1}{N}\sum_{n=1}^{N}C_g^{2(K-1)}C_f^2L_g^2C_g^2\|\bar{x}_t-x_{n,t}\|^2\\
&\quad+\cdots+\frac{K+1}{N}\sum_{n=1}^{N}C_g^{4(K-1)}C_f^2L_g^2\|\bar{x}_t-x_{n,t}\|^2+\frac{K+1}{N}\sum_{n=1}^{N}C_g^{4K}L_f^2\|\bar{x}_t-x_{n,t}\|^2+\frac{K+1}{N}\sum_{n=1}^{N}C_g^{2K}L_f^2\|\bar{y}_t-y_{n,t}\|^2\\
&= \frac{K+1}{N}\sum_{n=1}^{N}\Big(C_g^{4K}L_f^2+\sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2\Big)\|\bar{x}_t-x_{n,t}\|^2+\frac{K+1}{N}\sum_{n=1}^{N}C_g^{2K}L_f^2\|\bar{y}_t-y_{n,t}\|^2\,,
\end{aligned}
\tag{37}
$$

where the second step follows from the homogeneous data distribution setting.

□

Lemma B.7. *Given Assumption 4.1-4.3, we can know*

$$
\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|p_{n,t+1}-\bar{p}_{t+1}\Big\|^2\Big]\leq 6\tau^2\rho_x^2\eta^2C_G^2C_f^2\,,
\tag{38}
$$

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|x_{n,t+1}-\bar{x}_{t+1}\Big\|^2\Big] \leq 6\gamma_x^2\rho_x^2\tau^4\eta^4C_G^2C_f^2\,. \tag{39}$$

Proof.

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|p_{n,t+1}-\bar{p}_{t+1}\Big\|^2\Big]\\
&=\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|(1-\rho_x\eta)p_{n,t}+\rho_x\eta u_{n,t+1}-(1-\rho_x\eta)\bar{p}_t-\rho_x\eta\frac{1}{K}\sum_{n'=1}^{N}u_{n',t+1}\Big\|^2\Big]\\
&\leq(1-\rho_x\eta)^2(1+\frac{1}{\tau})\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|p_{n,t}-\bar{p}_t\Big\|^2\Big]+(1+\tau)\rho_x^2\eta^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|u_{n,t+1}-\frac{1}{N}\sum_{n'=1}^{N}u_{n',t+1}\Big\|^2\Big]\\
&\leq(1+\frac{1}{\tau})\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|p_{n,t}-\bar{p}_t\Big\|^2\Big]+2\tau\rho_x^2\eta^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|u_{n,t+1}-\frac{1}{N}\sum_{n'=1}^{N}u_{n',t+1}\Big\|^2\Big]\\
&\leq(1+\frac{1}{\tau})\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|p_{n,t}-\bar{p}_t\Big\|^2\Big]+2\tau\rho_x^2\eta^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|u_{n,t+1}\Big\|^2\Big]\\
&\leq(1+\frac{1}{\tau})\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|p_{n,t}-\bar{p}_t\Big\|^2\Big]+2\tau\rho_x^2\eta^2C_G^2C_f^2\\
&\leq 2\tau\rho_x^2\eta^2C_G^2C_f^2\sum_{t'=s_t\tau}^{t}(1+\frac{1}{\tau})^{t-t'}\\
&\leq 6\tau^2\rho_x^2\eta^2C_G^2C_f^2\,,
\end{aligned}\tag{40}$$

where $s_t=\lfloor(t+1)/\tau\rfloor$, the third to last step follows from $\|u_{n,t}\|\leq C_GC_f$, the last step follows from $(1+\frac{1}{\tau})^\tau<3$. Then, we have

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|x_{n,t+1}-\bar{x}_{t+1}\Big\|^2\Big]=\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|x_{n,s_t\tau}-\gamma_x\eta\sum_{t'=s_t\tau}^{t}p_{n,t'}-\bar{x}_{s_t\tau}+\gamma_x\eta\sum_{t'=s_t\tau}^{t}\bar{p}_{t'}\Big\|^2\Big]\\
&\leq\tau\gamma_x^2\eta^2\frac{1}{N}\sum_{n=1}^{N}\sum_{t'=s_tp}^{t}\mathbb{E}\Big[\Big\|p_{n,t'}-\bar{p}_{t'}\Big\|^2\Big]\leq 6\gamma_x^2\rho_x^2\tau^4\eta^4C_G^2C_f^2\,.
\end{aligned}\tag{41}$$

□

Lemma B.8. *Given Assumption 4.1-4.3, we can know*

$$\begin{aligned}
&\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|h_{n,t}^{(k)}-\bar{h}_t^{(k)}\Big\|^2\Big]\\
&\leq 24\alpha^2\tau^2\eta^4\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\sum_{j=k+1}^{K}(2C_g^2)^{j-k}\Big)\mathbb{E}\Big[\Big\|h_{n,t}^{(k)}-g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big]\\
&\quad+12\tau^2\gamma_x^2\eta^2C_G^2T\sum_{k=1}^{K}(2C_g^2)^k+24\alpha^2\eta^4\tau^2\delta^2T\sum_{k=1}^{K}\sum_{j=1}^{k-1}(2C_g^2)^j+192\tau^2\alpha^2\eta^4\delta^2TK\\
&\quad+24\gamma_x^2\eta^2\tau^2C_g^2C_G^2C_f^2T+1152\alpha^2\gamma_x^2\rho_x^2\tau^6\eta^8C_g^2C_G^2C_f^2T\,.
\end{aligned}\tag{42}$$

Proof. At first, we have

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[|h_{n,t+1}^{(k)}-\bar{h}_{t+1}^{(k)}\Big\|^2\Big]$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|(1-\alpha\eta^2)(h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)};\xi_{n,t+1}^{(k)})) + g_n^{(k)}(h_{n,t+1}^{(k-1)};\xi_{n,t+1}^{(k)}) \\
&\quad - \frac{1}{N}\sum_{n'=1}^{N}(1-\alpha\eta^2)(h_{n',t}^{(k)} - g_{n'}^{(k)}(h_{n',t}^{(k-1)};\xi_{n',t+1}^{(k)})) - \frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(k)}(h_{n',t+1}^{(k-1)};\xi_{n',t+1}^{(k)})\Big\|^2\Big] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|(1-\alpha\eta^2)(h_{n,t}^{(k)} - \frac{1}{N}\sum_{n'=1}^{N} h_{n',t}^{(k)}) - g_n^{(k)}(h_{n,t}^{(k-1)};\xi_{n,t+1}^{(k)}) + g_n^{(k)}(h_{n,t+1}^{(k-1)};\xi_{n,t+1}^{(k)}) \\
&\quad + \frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(k)}(h_{n',t}^{(k-1)};\xi_{n',t+1}^{(k)}) - \frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(k)}(h_{n',t+1}^{(k-1)};\xi_{n',t+1}^{(k)}) \\
&\quad + \alpha\eta^2 g_n^{(k)}(h_{n,t}^{(k-1)};\xi_{n,t+1}^{(k)}) - \alpha\eta^2\frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(k)}(h_{n',t}^{(k-1)};\xi_{n',t+1}^{(k)})\Big\|^2\Big] \\
&\le (1-\alpha\eta^2)^2(1+\frac{1}{\tau})\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - \frac{1}{N}\sum_{n'=1}^{N} h_{n',t}^{(k)}\Big\|^2\Big] + 2(1+\tau)\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\| - g_n^{(k)}(h_{n,t}^{(k-1)};\xi_{n,t+1}^{(k)}) + g_n^{(k)}(h_{n,t+1}^{(k-1)};\xi_{n,t+1}^{(k)}) \\
&\quad + \frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(k)}(h_{n',t}^{(k-1)};\xi_{n',t+1}^{(k)}) - \frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(k)}(h_{n',t+1}^{(k-1)};\xi_{n',t+1}^{(k)})\Big\|^2\Big] \\
&\quad + 2\alpha^2\eta^4(1+\tau)\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|g_n^{(k)}(h_{n,t}^{(k-1)};\xi_{n,t+1}^{(k)}) - \frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(k)}(h_{n',t}^{(k-1)};\xi_{n',t+1}^{(k)})\Big\|^2\Big] \\
&\le (1+\frac{1}{\tau})\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - \frac{1}{N}\sum_{n'=1}^{N} h_{n',t}^{(k)}\Big\|^2\Big] + 2(1+\tau)\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\| - g_n^{(k)}(h_{n,t}^{(k-1)};\xi_{n,t+1}^{(k)}) + g_n^{(k)}(h_{n,t+1}^{(k-1)};\xi_{n,t+1}^{(k)})\Big\|^2\Big] \\
&\quad + 2\alpha^2\eta^4(1+\tau)\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|g_n^{(k)}(h_{n,t}^{(k-1)};\xi_{n,t+1}^{(k)}) - \frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(k)}(h_{n',t}^{(k-1)};\xi_{n',t+1}^{(k)})\Big\|^2\Big] \\
&\le (1+\frac{1}{\tau})\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - \bar{h}_t^{(k)}\Big\|^2\Big] + 4\tau C_g^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t}^{(k-1)} - h_{n,t+1}^{(k-1)}\Big\|^2\Big] + 32\tau\alpha^2\eta^4\delta^2 \\
&\quad + 32\tau\alpha^2\eta^4 C_g^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t}^{(k-1)} - \bar{h}_t^{(k-1)}\Big\|^2\Big], \qquad (43)
\end{aligned}
$$

where the last step follows from $1+\tau \le 2\tau$ and the following inequality.

$$
\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|g_n^{(k)}(h_{n,t}^{(k-1)};\xi_{n,t+1}^{(k)}) - \frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(k)}(h_{n',t}^{(k-1)};\xi_{n',t+1}^{(k)})\Big\|^2\Big] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|g_n^{(k)}(h_{n,t}^{(k-1)};\xi_{n,t+1}^{(k)}) - g_n^{(k)}(h_{n,t}^{(k-1)}) + g_n^{(k)}(h_{n,t}^{(k-1)}) - g_n^{(k)}(\bar{h}_t^{(k-1)}) \\
&\quad + \frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(k)}(\bar{h}_t^{(k-1)}) - \frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(k)}(h_{n',t}^{(k-1)}) + \frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(k)}(h_{n',t}^{(k-1)}) - \frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(k)}(h_{n',t}^{(k-1)};\xi_{n',t+1}^{(k)})\Big\|^2\Big] \\
&\le 4\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|g_n^{(k)}(h_{n,t}^{(k-1)};\xi_{n,t+1}^{(k)}) - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] + 4\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|g_n^{(k)}(h_{n,t}^{(k-1)}) - g_n^{(k)}(\bar{h}_t^{(k-1)})\Big\|^2\Big] \\
&\quad + 4\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(k)}(\bar{h}_t^{(k-1)}) - \frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(k)}(h_{n',t}^{(k-1)})\Big\|^2\Big] \\
&\quad + 4\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(k)}(h_{n',t}^{(k-1)}) - \frac{1}{N}\sum_{n'=1}^{N} g_{n'}^{(k)}(h_{n',t}^{(k-1)};\xi_{n',t+1}^{(k)})\Big\|^2\Big] \\
&\le 8\delta^2 + 8C_g^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t}^{(k-1)} - \bar{h}_t^{(k-1)}\Big\|^2\Big]. \qquad (44)
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t}^{(k)}-\bar{h}_{t}^{(k)}\Big\|^2\Big]\\
&\leq (1+\frac{1}{\tau})\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t-1}^{(k)}-\bar{h}_{t-1}^{(k)}\Big\|^2\Big]+4\tau C_g^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t-1}^{(k-1)}-h_{n,t}^{(k-1)}\Big\|^2\Big]\\
&\quad+32\tau\alpha^2\eta^4\delta^2+32\tau\alpha^2\eta^4C_g^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t-1}^{(k-1)}-\bar{h}_{t-1}^{(k-1)}\Big\|^2\Big]\\
&\leq 4\tau C_g^2\sum_{t'=s_t\tau}^{t-1}(1+\frac{1}{\tau})^{t-1-t'}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t'}^{(k-1)}-h_{n,t'+1}^{(k-1)}\Big\|^2\Big]\\
&\quad+32\tau\alpha^2\eta^4\delta^2\sum_{t'=s_t\tau}^{t-1}(1+\frac{1}{\tau})^{t-1-t'}+32\tau\alpha^2\eta^4C_g^2\sum_{t'=s_t\tau}^{t-1}(1+\frac{1}{\tau})^{t-1-t'}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t'}^{(k-1)}-\bar{h}_{t'}^{(k-1)}\Big\|^2\Big]\\
&\leq 12\tau C_g^2\sum_{t'=s_t\tau}^{t-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t'}^{(k-1)}-h_{n,t'+1}^{(k-1)}\Big\|^2\Big]+96\tau^2\alpha^2\eta^4\delta^2+96\tau\alpha^2\eta^4C_g^2\sum_{t'=s_t\tau}^{t-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t'}^{(k-1)}-\bar{h}_{t'}^{(k-1)}\Big\|^2\Big]\\
&= 12\tau C_g^2\sum_{t'=s_t\tau+1}^{t}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t'-1}^{(k-1)}-h_{n,t'}^{(k-1)}\Big\|^2\Big]+96\tau^2\alpha^2\eta^4\delta^2+96\tau\alpha^2\eta^4C_g^2\sum_{t'=s_t\tau}^{t-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t'}^{(k-1)}-\bar{h}_{t'}^{(k-1)}\|^2\Big],
\end{aligned}
\tag{45}
$$

where the third inequality follows from $(1+\frac{1}{\tau})^{\tau}<3$.

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

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t}^{(k)}-\bar{h}_{t}^{(k)}\Big\|^2\Big]\\
&\leq 12\tau C_g^2\sum_{t=0}^{T-1}\sum_{t'=s_t\tau+1}^{t}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t'-1}^{(k-1)}-h_{n,t'}^{(k-1)}\Big\|^2\Big]+96\tau^2\alpha^2\eta^4\delta^2T\\
&\quad+96\tau\alpha^2\eta^4C_g^2\sum_{t=0}^{T-1}\sum_{t'=s_t\tau}^{t-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t'}^{(k-1)}-\bar{h}_{t'}^{(k-1)}\Big\|^2\Big]\\
&\leq 12\tau^2C_g^2\sum_{t=1}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t-1}^{(k-1)}-h_{n,t}^{(k-1)}\|^2+96\tau^2\alpha^2\eta^4\delta^2T+96\tau^2\alpha^2\eta^4C_g^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t}^{(k-1)}-\bar{h}_{t}^{(k-1)}\Big\|^2\Big]\\
&= 12\tau^2C_g^2\sum_{t=0}^{T-2}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t}^{(k-1)}-h_{n,t+1}^{(k-1)}\Big\|^2\Big]+96\tau^2\alpha^2\eta^4\delta^2T+96\tau^2\alpha^2\eta^4C_g^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t}^{(k-1)}-\bar{h}_{t}^{(k-1)}\Big\|^2\Big].
\end{aligned}
\tag{46}
$$

Then, by summing over k from 1 to K, we have

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|h_{n,t}^{(k)}-\bar{h}_{t}^{(k)}\Big\|^2\Big]\\
&\leq 12\tau^2C_g^2\sum_{t=0}^{T-2}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|h_{n,t}^{(k-1)}-h_{n,t+1}^{(k-1)}\Big\|^2\Big]+96\tau^2\alpha^2\eta^4\delta^2TK\\
&\quad+96\tau^2\alpha^2\eta^4C_g^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|h_{n,t}^{(k-1)}-\bar{h}_{t}^{(k-1)}\Big\|^2\Big]\\
&= 12\tau^2C_g^2\sum_{t=0}^{T-2}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=2}^{K}\mathbb{E}\Big[\Big\|h_{n,t}^{(k-1)}-h_{n,t+1}^{(k-1)}\Big\|^2\Big]+96\tau^2\alpha^2\eta^4\delta^2TK\\
&\quad+96\tau^2\alpha^2\eta^4C_g^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=2}^{K}\mathbb{E}\Big[\Big\|h_{n,t}^{(k-1)}-\bar{h}_{t}^{(k-1)}\Big\|^2\Big]\\
&\quad+12\tau^2C_g^2\sum_{t=0}^{T-2}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t}^{(0)}-h_{n,t+1}^{(0)}\Big\|^2\Big]+96\tau^2\alpha^2\eta^4C_g^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t}^{(0)}-\bar{h}_{t}^{(0)}\Big\|^2\Big]
\end{aligned}
$$

$$
\begin{aligned}
&\leq 96\tau^2\alpha^2\eta^4\delta^2TK + 96\tau^2\alpha^2\eta^4C_g^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - \bar{h}_t^{(k)}\Big\|^2\Big] + 12\tau^2C_g^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|x_{n,t} - x_{n,t+1}\Big\|^2\Big] \\
&\quad + 96\tau^2\alpha^2\eta^4C_g^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|x_{n,t} - \bar{x}_t\Big\|^2\Big] + 12\tau^2C_g^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=2}^{K}2\alpha^2\eta^4\sum_{j=1}^{k-1}(2C_g^2)^{k-1-j}\mathbb{E}\Big[\Big\|h_{n,t}^{(j)} - g_n^{(j)}(h_{n,t}^{(j-1)})\Big\|^2\Big] \\
&\quad + 12\tau^2C_g^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=2}^{K}(2C_g^2)^{k-1}\gamma_x^2\eta^2C_G^2C_f^2 + 12\tau^2C_g^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=2}^{K}2\alpha^2\eta^2\delta^2\sum_{j=1}^{k-1}(2C_g^2)^{j-1} \\
&\leq 96\tau^2\alpha^2\eta^4\delta^2TK + 96\tau^2\alpha^2\eta^4C_g^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - \bar{h}_t^{(k)}\Big\|^2\Big] + 12\gamma_x^2\eta^2\tau^2C_g^2C_G^2C_f^2T + 576\alpha^2\gamma_x^2\rho_x^2\tau^6\eta^8C_g^2C_G^2C_f^2T \\
&\quad + 12\alpha^2\tau^2\eta^4\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\sum_{j=k+1}^{K}(2C_g^2)^{j-k}\Big)\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] \\
&\quad + 6\tau^2\gamma_x^2\eta^2C_G^2C_f^2T\sum_{k=1}^{K}(2C_g^2)^k + 12\alpha^2\eta^4\tau^2\delta^2T\sum_{k=1}^{K}\sum_{j=1}^{k-1}(2C_g^2)^j\,, \qquad (47)
\end{aligned}
$$

where the second inequality follows from Lemma B.5 and $\|p_{n,t}\| \leq C_GC_f$, the last inequality follows from Lemma B.7.

Then, we have

$$
\begin{aligned}
&(1 - 96\tau^2\alpha^2\eta^4C_g^2)\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - \bar{h}_t^{(k)}\Big\|^2\Big] \\
&\leq 12\alpha^2\tau^2\eta^4\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\sum_{j=k+1}^{K}(2C_g^2)^{j-k}\Big)\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] \\
&\quad + 6\tau^2\gamma_x^2\eta^2C_G^2C_f^2T\sum_{k=1}^{K}(2C_g^2)^k + 12\alpha^2\eta^4\tau^2\delta^2T\sum_{k=1}^{K}\sum_{j=1}^{k-1}(2C_g^2)^j + 96\tau^2\alpha^2\eta^4\delta^2TK \\
&\quad + 12\gamma_x^2\eta^2\tau^2C_g^2C_G^2C_f^2T + 576\alpha^2\gamma_x^2\rho_x^2\tau^6\eta^8C_g^2C_G^2C_f^2T\,. \qquad (48)
\end{aligned}
$$

By setting $\eta \leq \frac{1}{4\sqrt{\tau\alpha C_g}}$, we have $1 - 96\tau^2\alpha^2\eta^4C_g^2 \geq \frac{1}{2}$. Then, we can obtain

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - \bar{h}_t^{(k)}\|^2\Big] \\
&\leq 24\alpha^2\tau^2\eta^4\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\sum_{j=k+1}^{K}(2C_g^2)^{j-k}\Big)\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] \\
&\quad + 12\tau^2\gamma_x^2\eta^2C_G^2C_f^2T\sum_{k=1}^{K}(2C_g^2)^k + 24\alpha^2\eta^4\tau^2\delta^2T\sum_{k=1}^{K}\sum_{j=1}^{k-1}(2C_g^2)^j + 192\tau^2\alpha^2\eta^4\delta^2TK \\
&\quad + 24\gamma_x^2\eta^2\tau^2C_g^2C_G^2C_f^2T + 1152\alpha^2\gamma_x^2\rho_x^2\tau^6\eta^8C_g^2C_G^2C_f^2T\,. \qquad (49)
\end{aligned}
$$

□

Lemma B.9. *Given Assumption 4.1-4.3, we can know*

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t} - \bar{q}_t\Big\|^2\Big] \\
&\leq 2304\alpha^2\tau^4\eta^6\rho_y^2L_f^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\sum_{j=k+1}^{K}(2C_g^2)^{j-k}\Big)\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] \\
&\quad + 1152\gamma_x^2\rho_y^2\tau^4\eta^4L_f^2C_G^2C_f^2T\sum_{k=1}^{K}(2C_g^2)^k + 2304\alpha^2\rho_y^2\tau^4\eta^6\delta^2L_f^2T\sum_{k=1}^{K}\sum_{j=1}^{k-1}(2C_g^2)^j
\end{aligned}
$$

$$+ 18432\alpha^2\rho_y^2\tau^4\eta^6\delta^2 L_f^2 TK + 2304\gamma_x^2\rho_y^2\tau^4\eta^4 L_f^2 C_g^2 C_G^2 C_f^2 T$$
$$+ 110592\alpha^2\gamma_x^2\rho_x^2\rho_y^2\tau^8\eta^{10} L_f^2 C_g^2 C_G^2 C_f^2 T + 76\tau^2\rho_y^2\eta^2\sigma^2 T\,, \tag{50}$$

$$\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|y_{n,t}-\bar{y}_t\Big\|^2\Big]$$
$$\leq 2304\alpha^2\gamma_y^2\rho_y^2\tau^6\eta^8 L_f^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\sum_{j=k+1}^{K}(2C_g^2)^{j-k}\Big)\mathbb{E}\Big[\Big\|h_{n,t}^{(k)}-g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big]$$
$$+ 1152\gamma_x^2\gamma_y^2\rho_y^2\tau^6\eta^6 L_f^2 C_G^2 C_f^2 T\sum_{k=1}^{K}(2C_g^2)^k + 2304\alpha^2\gamma_y^2\rho_y^2\tau^6\eta^8\delta^2 L_f^2 T\sum_{k=1}^{K}\sum_{j=1}^{k-1}(2C_g^2)^j$$
$$+ 18432\alpha^2\rho_y^2\gamma_y^2\tau^6\eta^8\delta^2 L_f^2 TK + 2304\gamma_x^2\gamma_y^2\rho_y^2\tau^6\eta^6 L_f^2 C_g^2 C_G^2 C_f^2 T$$
$$+ 110592\alpha^2\rho_x^2\gamma_x^2\rho_y^2\gamma_y^2\tau^{10}\eta^{12} L_f^2 C_g^2 C_G^2 C_f^2 T + 76\tau^4\eta^4\gamma_y^2\rho_y^2\sigma^2 T\,. \tag{51}$$

Proof.

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t+1}-\bar{q}_{t+1}\Big\|^2\Big]$$
$$= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|(1-\rho_y\eta)q_{n,t}+\rho_y\eta\nabla_2 f_n(h_{n,t+1}^{(K)},y_{n,t+1};\zeta_{n,t+1})$$
$$-(1-\rho_y\eta)\bar{q}_t-\rho_y\eta\frac{1}{N}\sum_{n'=1}^{N}\nabla_2 f_{n'}(h_{n',t+1}^{(K)},y_{n',t+1};\zeta_{n',t+1})\Big\|^2\Big]$$
$$\leq (1-\rho_y\eta)^2(1+\frac{1}{\tau})\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t}-\bar{q}_t\Big\|^2\Big]$$
$$+ (1+\tau)\rho_y^2\eta^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\nabla_2 f_n(h_{n,t+1}^{(K)},y_{n,t+1};\zeta_{n,t+1})-\frac{1}{N}\sum_{n'=1}^{N}\nabla_2 f_{n'}(h_{n',t+1}^{(K)},y_{n',t+1};\zeta_{n',t+1})\Big\|^2\Big]$$
$$\leq (1+\frac{1}{\tau})\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t}-\bar{q}_t\Big\|^2\Big]$$
$$+ 2\tau\rho_y^2\eta^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\nabla_2 f_n(h_{n,t+1}^{(K)},y_{n,t+1};\zeta_{n,t+1})-\frac{1}{N}\sum_{n'=1}^{N}\nabla_2 f_{n'}(h_{n',t+1}^{(K)},y_{n',t+1};\zeta_{n',t+1})\Big\|^2\Big]$$
$$\leq (1+\frac{1}{\tau})\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t}-\bar{q}_t\Big\|^2\Big] + 2\tau\rho_y^2\eta^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\nabla_2 f_n(h_{n,t+1}^{(K)},y_{n,t+1};\zeta_{n,t+1})-\nabla_2 f_n(h_{n,t+1}^{(K)},y_{n,t+1})$$
$$+ \nabla_2 f_n(h_{n,t+1}^{(K)},y_{n,t+1})-\nabla_2 f(\bar{h}_{t+1}^{(K)},\bar{y}_{t+1})+\nabla_2 f(\bar{h}_{t+1}^{(K)},\bar{y}_{t+1})-\frac{1}{N}\sum_{n'=1}^{N}\nabla_2 f_{n'}(h_{n',t+1}^{(K)},y_{n',t+1})$$
$$+ \frac{1}{N}\sum_{n'=1}^{N}\nabla_2 f_{n'}(h_{n',t+1}^{(K)},y_{n',t+1})-\frac{1}{N}\sum_{n'=1}^{N}\nabla_2 f_{n'}(h_{n',t+1}^{(K)},y_{n',t+1};\zeta_{n',t+1})\Big\|^2\Big]$$
$$\leq (1+\frac{1}{\tau})\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t}-\bar{q}_t\Big\|^2\Big] + 16\tau\rho_y^2\eta^2\sigma^2 + 8\tau\rho_y^2\eta^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\nabla_2 f_n(h_{n,t+1}^{(K)},y_{n,t+1})-\nabla_2 f_n(\bar{h}_{t+1}^{(K)},\bar{y}_{t+1})\Big\|^2\Big]$$
$$+ 8\tau\rho_y^2\eta^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\nabla_2 f(\bar{h}_{t+1}^{(K)},\bar{y}_{t+1})-\frac{1}{N}\sum_{n'=1}^{N}\nabla_2 f_{n'}(h_{n',t+1}^{(K)},y_{n',t+1})\Big\|^2\Big]$$
$$\leq (1+\frac{1}{\tau})\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t}-\bar{q}_t\Big\|^2\Big] + 16\tau\rho_y^2\eta^2\sigma^2 + 16\tau\rho_y^2\eta^2 L_f^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t+1}^{(K)}-\bar{h}_{t+1}^{(K)}\Big\|^2\Big]$$
$$+ 16\tau\rho_y^2\eta^2 L_f^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|y_{n,t+1}-\bar{y}_{t+1}\Big\|^2\Big]\,, \tag{52}$$

where the second to last step follows from the homogeneous data distribution setting. Due to

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|y_{n,t+1}-\bar{y}_{t+1}\Big\|^2\Big]\\
&=\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|y_{n,s_t\tau}+\gamma_y\eta\sum_{t'=s_t\tau}^{t}q_{n,t'}-\bar{y}_{s_t\tau}-\gamma_y\eta\sum_{t'=s_t\tau}^{t}\bar{q}_{t'}\Big\|^2\Big]\\
&\leq\tau\gamma_y^2\eta^2\frac{1}{N}\sum_{n=1}^{N}\sum_{t'=s_tp}^{t}\mathbb{E}\Big[\Big\|q_{n,t'}-\bar{q}_{t'}\Big\|^2\Big]\,,
\end{aligned}\tag{53}$$

we have

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t}-\bar{q}_t\Big\|^2\Big]\\
&\leq(1+\frac{1}{\tau})\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t-1}-\bar{q}_{t-1}\Big\|^2\Big]+16\tau\rho_y^2\eta^2\sigma^2+16\tau\rho_y^2\eta^2L_f^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t}^{(K)}-\bar{h}_t^{(K)}\Big\|^2\Big]\\
&\quad+16\tau\rho_y^2\eta^2L_f^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|y_{n,t}-\bar{y}_t\Big\|^2\Big]\\
&\leq(1+\frac{1}{\tau})\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t-1}-\bar{q}_{t-1}\Big\|^2\Big]+16\tau^2\rho_y^2\gamma_y^2\eta^4L_f^2\frac{1}{N}\sum_{n=1}^{N}\sum_{t'=s_tp}^{t-1}\mathbb{E}\Big[\Big\|q_{n,t'}-\bar{q}_{t'}\Big\|^2\Big]\\
&\quad+16\tau\rho_y^2\eta^2L_f^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t}^{(K)}-\bar{h}_t^{(K)}\Big\|^2\Big]+16\tau\rho_y^2\eta^2\sigma^2\\
&\leq16\tau^2\rho_y^2\gamma_y^2\eta^4L_f^2\sum_{t'=s_t\tau}^{t-1}(1+\frac{1}{\tau})^{t-1-t'}\frac{1}{N}\sum_{n=1}^{N}\sum_{t''=s_tp}^{t'}\mathbb{E}\Big[\Big\|q_{n,t''}-\bar{q}_{t''}\Big\|^2\Big]\\
&\quad+16\tau\rho_y^2\eta^2L_f^2\sum_{t'=s_t\tau}^{t-1}(1+\frac{1}{\tau})^{t-1-t'}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t'+1}^{(K)}-\bar{h}_{t'+1}^{(K)}\Big\|^2\Big]+16\tau\rho_y^2\eta^2\sigma^2\sum_{t'=s_t\tau}^{t-1}(1+\frac{1}{\tau})^{t-1-t'}\\
&\leq48\tau^3\rho_y^2\gamma_y^2\eta^4L_f^2\sum_{t'=s_t\tau}^{t-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t'}-\bar{q}_{t'}\Big\|^2\Big]\\
&\quad+48\tau\rho_y^2\eta^2L_f^2\sum_{t'=s_t\tau+1}^{t}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t'}^{(K)}-\bar{h}_{t'}^{(K)}\Big\|^2\Big]+38\tau^2\rho_y^2\eta^2\sigma^2\,,
\end{aligned}\tag{54}$$

where $s_t=\lfloor(t+1)/\tau\rfloor$, the last step follows from $(1+\frac{1}{\tau})^\tau<3$.

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

$$\begin{aligned}
&\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t}-\bar{q}_t\Big\|^2\Big]\\
&\leq48\tau^3\rho_y^2\gamma_y^2\eta^4L_f^2\sum_{t=0}^{T-1}\sum_{t'=s_t\tau}^{t-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t'}-\bar{q}_{t'}\Big\|^2\Big]\\
&\quad+48\tau\rho_y^2\eta^2L_f^2\sum_{t=0}^{T-1}\sum_{t'=s_t\tau+1}^{t}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t'}^{(K)}-\bar{h}_{t'}^{(K)}\Big\|^2\Big]+38\tau^2\rho_y^2\eta^2\sigma^2T\\
&\leq48\tau^4\rho_y^2\gamma_y^2\eta^4L_f^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t}-\bar{q}_t\Big\|^2\Big]\\
&\quad+48\tau^2\rho_y^2\eta^2L_f^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t}^{(K)}-\bar{h}_t^{(K)}\Big\|^2\Big]+38\tau^2\rho_y^2\eta^2\sigma^2T\,.
\end{aligned}\tag{55}$$

Then, by setting $\eta\leq\frac{1}{4\tau\sqrt{\rho_y\gamma_yL_f}}$, we have $1-48\tau^4\rho_y^2\gamma_y^2\eta^4L_f^2\geq\frac{1}{2}$. Finally, we can get

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t}-\bar{q}_t\Big\|^2\Big]\\
&\leq 96\tau^2\rho_y^2\eta^2L_f^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|h_{n,t}^{(K)}-\bar{h}_t^{(K)}\Big\|^2\Big]+76\tau^2\rho_y^2\eta^2\sigma^2T\\
&\leq 96\tau^2\rho_y^2\eta^2L_f^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}\Big[\Big\|h_{n,t}^{(k)}-\bar{h}_t^{(k)}\Big\|^2\Big]+76\tau^2\rho_y^2\eta^2\sigma^2T\\
&\leq 2304\alpha^2\tau^4\eta^6\rho_y^2L_f^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\sum_{j=k+1}^{K}(2C_g^2)^{j-k}\Big)\mathbb{E}\Big[\Big\|h_{n,t}^{(k)}-g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big]\\
&\quad+1152\gamma_x^2\rho_y^2\tau^4\eta^4L_f^2C_G^2C_f^2T\sum_{k=1}^{K}(2C_g^2)^k+2304\alpha^2\rho_y^2\tau^4\eta^6\delta^2L_f^2T\sum_{k=1}^{K}\sum_{j=1}^{k-1}(2C_g^2)^j\\
&\quad+18432\alpha^2\rho_y^2\tau^4\eta^6\delta^2L_f^2TK+2304\gamma_x^2\rho_y^2\tau^4\eta^4L_f^2C_g^2C_G^2C_f^2T\\
&\quad+110592\alpha^2\gamma_x^2\rho_x^2\rho_y^2\tau^8\eta^{10}L_f^2C_g^2C_G^2C_f^2T+76\tau^2\rho_y^2\eta^2\sigma^2T\,,
\end{aligned}
\tag{56}
$$

where the last step follows from Lemma B.8.

Additionally, we have

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|y_{n,t}-\bar{y}_t\Big\|^2\Big]\\
&\leq\tau\gamma_y^2\eta^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{t'=s_tp}^{t-1}\mathbb{E}\Big[\Big\|q_{n,t'}-\bar{q}_{t'}\Big\|^2\Big]\\
&\leq\tau^2\gamma_y^2\eta^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t}-\bar{q}_t\Big\|^2\Big]\\
&\leq 2304\alpha^2\gamma_y^2\rho_y^2\tau^6\eta^8L_f^2\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\sum_{j=k+1}^{K}(2C_g^2)^{j-k}\Big)\mathbb{E}\Big[\Big\|h_{n,t}^{(k)}-g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big]\\
&\quad+1152\gamma_x^2\gamma_y^2\rho_y^2\tau^6\eta^6L_f^2C_G^2C_f^2T\sum_{k=1}^{K}(2C_g^2)^k+2304\alpha^2\gamma_y^2\rho_y^2\tau^6\eta^8\delta^2L_f^2T\sum_{k=1}^{K}\sum_{j=1}^{k-1}(2C_g^2)^j\\
&\quad+18432\alpha^2\rho_y^2\gamma_y^2\tau^6\eta^8\delta^2L_f^2TK+2304\gamma_x^2\gamma_y^2\rho_y^2\tau^6\eta^6L_f^2C_g^2C_G^2C_f^2T\\
&\quad+110592\alpha^2\rho_x^2\gamma_x^2\rho_y^2\gamma_y^2\tau^{10}\eta^{12}L_f^2C_g^2C_G^2C_f^2T+76\tau^4\eta^4\gamma_y^2\rho_y^2\sigma^2T\,.
\end{aligned}
\tag{57}
$$

□

Lemma B.10. *Given Assumption 4.1-4.3 and $\gamma_y<\frac{1}{6L_f}$, we can know*

$$
\begin{aligned}
&\|\bar{y}_{t+1}-y^*(\bar{x}_{t+1})\|^2\\
&\leq(1-\frac{\eta\gamma_y\mu}{4})\|\bar{y}_t-y^*(\bar{x}_t)\|^2-\frac{3\eta\gamma_y^2}{4}\|\bar{q}_t\|^2+\frac{25\eta\gamma_x^2C_G^2L_f^2}{6\gamma_y\mu^3}\|\bar{p}_t\|^2\\
&\quad+\frac{25\eta\gamma_yL_f^2}{2\mu}\frac{1}{N}\sum_{n=1}^{N}\Big[\Big\|\bar{y}_t-y_{n,t}\Big\|^2\Big]+\frac{25\eta\gamma_yL_f^2C_G^2}{2\mu}\frac{1}{N}\sum_{n=1}^{N}\Big[\Big\|\bar{x}_t-x_{n,t}\Big\|^2\Big]\\
&\quad+\frac{25\eta\gamma_y}{2\mu}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\nabla_2f_n(G_n(x_{n,t}),y_{n,t})-\frac{1}{N}\sum_{n=1}^{N}q_{n,t}\Big\|^2\Big]\,.
\end{aligned}
\tag{58}
$$

Proof.

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

$$
\begin{aligned}
&\leq (1-\frac{\eta\gamma_y\mu}{4})\|\bar{y}_t - y^*(\bar{x}_t)\|^2 - \frac{3\eta\gamma_y^2}{4}\|\bar{q}_t\|^2 + \frac{25\eta\gamma_x^2 C_G^2 L_f^2}{6\gamma_y\mu^3}\|\bar{p}_t\|^2 + \frac{25\eta\gamma_y}{6\mu}\|\nabla_2 f(G(\bar{x}_t),\bar{y}_t) - \bar{q}_t\|^2 \\
&\leq (1-\frac{\eta\gamma_y\mu}{4})\|\bar{y}_t - y^*(\bar{x}_t)\|^2 - \frac{3\eta\gamma_y^2}{4}\|\bar{q}_t\|^2 + \frac{25\eta\gamma_x^2 C_G^2 L_f^2}{6\gamma_y\mu^3}\|\bar{p}_t\|^2 \\
&\quad + \frac{25\eta\gamma_y}{6\mu}\Big[\Big\|\frac{1}{N}\sum_{n=1}^N \nabla_2 f_n(G(\bar{x}_t),\bar{y}_t) - \frac{1}{N}\sum_{n=1}^N \nabla_2 f_n(G_n(\bar{x}_t), y_{n,t}) \\
&\qquad + \frac{1}{N}\sum_{n=1}^N \nabla_2 f_n(G_n(\bar{x}_t), y_{n,t}) - \frac{1}{N}\sum_{n=1}^N \nabla_2 f_n(G_n(x_{n,t}), y_{n,t}) \\
&\qquad + \frac{1}{N}\sum_{n=1}^N \nabla_2 f_n(G_n(x_{n,t}), y_{n,t}) - \frac{1}{N}\sum_{n=1}^N q_{n,t}\Big\|^2\Big] \\
&\leq (1-\frac{\eta\gamma_y\mu}{4})\|\bar{y}_t - y^*(\bar{x}_t)\|^2 - \frac{3\eta\gamma_y^2}{4}\|\bar{q}_t\|^2 + \frac{25\eta\gamma_x^2 C_G^2 L_f^2}{6\gamma_y\mu^3}\|\bar{p}_t\|^2 \\
&\quad + \frac{25\eta\gamma_y}{2\mu}\Big[\Big\|\frac{1}{N}\sum_{n=1}^N \nabla_2 f_n(G(\bar{x}_t),\bar{y}_t) - \frac{1}{N}\sum_{n=1}^N \nabla_2 f_n(G_n(\bar{x}_t), y_{n,t})\Big\|^2\Big] \\
&\quad + \frac{25\eta\gamma_y}{2\mu}\Big[\Big\|\frac{1}{N}\sum_{n=1}^N \nabla_2 f_n(G_n(\bar{x}_t), y_{n,t}) - \frac{1}{N}\sum_{n=1}^N \nabla_2 f_n(G_n(x_{n,t}), y_{n,t})\Big\|^2\Big] \\
&\quad + \frac{25\eta\gamma_y}{2\mu}\Big[\Big\|\frac{1}{N}\sum_{n=1}^N \nabla_2 f_n(G_n(x_{n,t}), y_{n,t}) - \frac{1}{N}\sum_{n=1}^N q_{n,t}\Big\|^2\Big] \\
&\leq (1-\frac{\eta\gamma_y\mu}{4})\|\bar{y}_t - y^*(\bar{x}_t)\|^2 - \frac{3\eta\gamma_y^2}{4}\|\bar{q}_t\|^2 + \frac{25\eta\gamma_x^2 C_G^2 L_f^2}{6\gamma_y\mu^3}\|\bar{p}_t\|^2 \\
&\quad + \frac{25\eta\gamma_y L_f^2}{2\mu}\frac{1}{N}\sum_{n=1}^N\Big[\Big\|\bar{y}_t - y_{n,t}\Big\|^2\Big] + \frac{25\eta\gamma_y L_f^2 C_G^2}{2\mu}\frac{1}{N}\sum_{n=1}^N\Big[\Big\|\bar{x}_t - x_{n,t}\Big\|^2\Big] \\
&\quad + \frac{25\eta\gamma_y}{2\mu}\Big[\Big\|\frac{1}{N}\sum_{n=1}^N \nabla_2 f_n(G_n(x_{n,t}), y_{n,t}) - \frac{1}{N}\sum_{n=1}^N q_{n,t}\Big\|^2\Big], \qquad (59)
\end{aligned}
$$

where the first step holds follows from Lemma 5 in (Gao et al., 2021), the second step follows from the homogeneous data distribution setting.

□

Now we are ready to establish the convergence rate. At first, we introduce a potential function as follows:

$$
\begin{aligned}
P_{t+1} &= \mathbb{E}[\Phi(\bar{x}_{t+1})] + w_0\mathbb{E}\Big[\Big\|\bar{y}_{t+1} - y^*(\bar{x}_{t+1})\|^2\Big] + \frac{1}{N}\sum_{n=1}^N\sum_{k=1}^K w_k\mathbb{E}\Big[\Big\|h_{n,t+1}^{(k)} - g_n^{(k)}(h_{n,t+1}^{(k-1)})\Big\|^2\Big] \\
&+ w_{K+1}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^N p_{n,t+1} - \frac{1}{N}\sum_{n=1}^N \nabla g_n^{(1)}(x_{n,t+1})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t+1})) \\
&\qquad \cdots \nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t+1}))\nabla_1 f_n(G_n^{(K)}(x_{n,t+1}), y_{n,t+1})\Big\|^2\Big] \\
&+ w_{K+2}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^N q_{n,t+1} - \frac{1}{N}\sum_{n=1}^N \nabla_2 f_n(G_n(x_{n,t+1}), y_{n,t+1})\Big\|^2\Big]. \qquad (60)
\end{aligned}
$$

Then, based on the aforementioned lemmas, we have

$$
P_{t+1} - P_t \leq -\frac{\gamma_x\eta}{2}\mathbb{E}\Big[\Big\|\nabla\Phi(\bar{x}_t)\Big\|^2\Big] - \frac{\gamma_x\eta}{4}\mathbb{E}[\|\bar{p}_t\|^2] + 2\gamma_x\eta C_G^2 L_f^2\mathbb{E}\|y^*(\bar{x}_t) - \bar{y}_t\|^2]
$$

$$
\begin{aligned}
&+ 2\gamma_x\eta(K+1)\Big(C_g^{4K}L_f^2 + \sum_{k=0}^{K-1} C_g^{2(K-1+k)}C_f^2L_g^2\Big)\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\bar{x}_t - x_{n,t}\Big\|^2\Big] + 2\gamma_x\eta(K+1)C_g^{2K}L_f^2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|\bar{y}_t - y_{n,t}\|^2]\\
&+ 2\gamma_x\eta\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}),y_{n,t}) - \frac{1}{N}\sum_{n=1}^{N}p_{n,t}\Big\|^2\Big]\\
&+ w_0\Big(-\frac{\eta\gamma_y\mu}{4}\mathbb{E}\Big[\Big\|\bar{y}_t - y^*(\bar{x}_t)\|^2\Big] - \frac{3\eta\gamma_y^2}{4}\mathbb{E}\Big[\Big\|\bar{q}_t\Big\|^2\Big] + \frac{25\eta\gamma_x^2C_G^2L_f^2}{6\gamma_y\mu^3}\mathbb{E}\Big[\Big\|\bar{p}_t\|^2\Big] + \frac{25\eta\gamma_yL_f^2}{2\mu}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\bar{y}_t - y_{n,t}\Big\|^2\Big]\\
&\quad + \frac{25\eta\gamma_yL_f^2C_G^2}{2\mu}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\bar{x}_t - x_{n,t}\Big\|^2\Big] + \frac{25\eta\gamma_y}{2\mu}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(G_n(x_{n,t}),y_{n,t}) - \frac{1}{N}\sum_{n=1}^{N}q_{n,t}\Big\|^2\Big]\Big)\\
&- \alpha\eta^2\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}w_k\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] + 2\alpha^2\eta^4\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\sum_{j=k+1}^{K}w_j(2C_g^2)^{j-k}\Big)\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big]\\
&+ 2\gamma_x^2\eta^2\sum_{k=1}^{K}w_k(2C_g^2)^k\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|p_{n,t} - \bar{p}_t\Big\|^2\Big] + 2\gamma_x^2\eta^2\sum_{k=1}^{K}w_k(2C_g^2)^k\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|\bar{p}_t\Big\|^2\Big] + 2\alpha^2\eta^4\delta^2\sum_{k=1}^{K}w_k\sum_{j=1}^{k-1}(2C_g^2)^j\\
&+ w_{K+1}\Big(-\rho_x\eta\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}p_{n,t} - \frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,t})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,t}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,t}))\nabla_1 f_n(G_n^{(K)}(x_{n,t}),y_{n,t})\Big\|^2\Big]\\
&\quad + \frac{4\eta\gamma_x^2}{\rho_x}\Big(C_g^{4K}L_f^2 + \sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2\Big)\frac{K+1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|p_{n,t} - \bar{p}_t\Big\|^2\Big]\\
&\quad + 8\rho_x\gamma_x^2\eta^3K\sum_{k=1}^{K}A_k(2C_g^2)^k\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|p_{n,t} - \bar{p}_{n,t}\Big\|^2\Big]\\
&\quad + \frac{4\eta\gamma_x^2}{\rho_x}(K+1)\Big(C_g^{4K}L_f^2 + \sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2\Big)\mathbb{E}\Big[\Big\|\bar{p}_t\Big\|^2\Big] + 8\rho_x\gamma_x^2\eta^3K\sum_{k=1}^{K}A_k(2C_g^2)^k\mathbb{E}\Big[\Big\|\bar{p}_{nt}\Big\|^2\Big]\\
&\quad + \frac{4\eta\gamma_y^2}{\rho_x}C_g^{2K}L_f^2\frac{K+1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t} - \bar{q}_t\Big\|^2\Big] + \frac{4\eta\gamma_y^2}{\rho_x}C_g^{2K}L_f^2(K+1)\mathbb{E}\Big[\Big\|\bar{q}_t\Big\|^2\Big]\\
&\quad + 8\rho_x\alpha^2\eta^5\frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(\sum_{j=k+1}^{K}A_j(2C_g^2)^{j-k}\Big)\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big]\\
&\quad + 8\rho_x\alpha^2\eta^5\delta^2K\sum_{k=1}^{K}A_k\sum_{j=1}^{k-1}(2C_g^2)^j + \rho_x^2\eta^2(K(K+1)C_g^{2(K-1)}C_f^2 + (K+1)C_g^{2K})\frac{\sigma^2}{N}\Big)\\
&+ w_{K+2}\Big(-\rho_y\eta\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}q_{n,t} - \frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(G_n(x_{n,t}),y_{n,t})\Big\|^2\Big] + \frac{4\eta\gamma_x^2L_f^2C_G^2}{\rho_y}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|p_{n,t} - \bar{p}_t\Big\|^2\Big] + \rho_y^2\eta^2\frac{\sigma^2}{N}\\
&\quad + 2\rho_y\eta L_f^2\frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}C_g^{2(K-k)}\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] + \frac{4\eta\gamma_y^2L_f^2}{\rho_y}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\Big\|q_{n,t} - \bar{q}_t\Big\|^2\Big] + \frac{4\eta\gamma_y^2L_f^2}{\rho_y}\mathbb{E}\Big[\Big\|\bar{q}_t\Big\|^2\Big]\\
&\quad + 4\rho_y\alpha^2\eta^5L_f^2\frac{K}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\Big(2^{K+1}C_g^{2(K-k)}\Big)\mathbb{E}\Big[\Big\|h_{n,t}^{(k)} - g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big] + \frac{4\eta\gamma_x^2L_f^2C_G^2}{\rho_y}\mathbb{E}\Big[\Big\|\bar{p}_t\Big\|^2\Big]\\
&\quad + 2\rho_y\gamma_x^2\eta^3K^2C_G^2C_f^2L_f^2(2C_g^2)^K + 4\alpha^2\rho_y\eta^5\delta^2L_f^2K\sum_{k=1}^{K}\sum_{j=1}^{k-1}C_g^{2(K-(k-j))}\Big), \qquad (61)
\end{aligned}
$$

where $A_k = \left(C_g^{2(K-1)}C_f^2L_g^2\Big(\sum_{j'=k}^{K-1}C_g^{j'-k}\Big)^2 + C_g^{2K}L_f^2C_g^{2(K-k)}\right)$.

From Lemma B.7, B.9, we have

$$
\begin{aligned}
&P_{t+1} - P_t \leq -\frac{\gamma_x\eta}{2}\mathbb{E}\Big[\Big\|\nabla\Phi(\bar{x}_t)\Big\|^2\Big] + (2\gamma_x\eta C_G^2L_f^2 - \frac{\eta\gamma_y\mu}{4}w_0)\mathbb{E}\|y^*(\bar{x}_t) - \bar{y}_t\|^2\\
&+ 6\gamma_x^2\rho_x^2\tau^4\eta^4C_G^2C_f^2\Big(\frac{25\eta\gamma_yL_f^2C_G^2}{2\mu}w_0 + 2\gamma_x\eta(K+1)\Big(C_g^{4K}L_f^2 + \sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2\Big)\Big)
\end{aligned}
$$

$$
\begin{aligned}
&+\Big(\frac{25\eta\gamma_y L_f^2}{2\mu}w_0+2\gamma_x\eta(K+1)C_g^{2K}L_f^2\Big)(1152\gamma_x^2\gamma_y^2\rho_y^2\tau^6\eta^6L_f^2C_G^2C_f^2\sum_{k=1}^K(2C_g^2)^k+2304\alpha^2\gamma_y^2\rho_y^2\tau^6\eta^8\delta^2L_f^2\sum_{k=1}^K\sum_{j=1}^{k-1}(2C_g^2)^j\\
&+18432\alpha^2\rho_y^2\gamma_y^2\tau^6\eta^8\delta^2L_f^2K+2304\gamma_x^2\gamma_y^2\rho_y^2\tau^6\eta^6L_f^2C_g^2C_G^2C_f^2+110592\alpha^2\rho_x^2\gamma_x^2\rho_y^2\gamma_y^2\tau^{10}\eta^{12}L_f^2C_g^2C_G^2C_f^2+76\tau^4\eta^4\gamma_y^2\rho_y^2\sigma^2)\\
&+\Big(\frac{4\eta\gamma_y^2}{\rho_x}C_g^{2K}L_f^2(K+1)w_{K+1}+\frac{4\eta\gamma_y^2L_f^2}{\rho_y}w_{K+2}-\frac{3\eta\gamma_y^2}{4}w_0\Big)\mathbb{E}[\|\bar q_t\|^2]\\
&+\Big(\frac{25\eta\gamma_x^2C_G^2L_f^2}{6\gamma_y\mu^3}w_0+2\gamma_x^2\eta^2\sum_{k=1}^K w_k(2C_g^2)^k+\frac{4\eta\gamma_x^2}{\rho_x}(K+1)\Big(C_g^{4K}L_f^2+\sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2\Big)w_{K+1}\\
&\quad+\frac{4\eta\gamma_x^2L_f^2C_G^2}{\rho_y}w_{K+2}-\frac{\gamma_x\eta}{4}\Big)\mathbb{E}[\|\bar p_t\|^2]\\
&+\Big(2\gamma_x\eta-\rho_x\eta w_{K+1}\Big)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^N p_{n,t}-\frac{1}{N}\sum_{n=1}^N\nabla_1 f_n(G_n(x_{n,t}),y_{n,t})\Big\|^2\Big]\\
&+\Big(\frac{25\eta\gamma_y}{2\mu}w_0-\rho_y\eta w_{K+2}\Big)\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^N\nabla_2 f_n(G_n(x_{n,t}),y_{n,t})-\frac{1}{N}\sum_{n=1}^N q_{n,t}\Big\|^2\Big]\Big)\\
&+\Bigg(4\rho_x\eta K\sum_{k=1}^K A_kw_{K+1}+8\rho_x\alpha^2\eta^5K\sum_{k=1}^K B_kw_{K+1}+2\alpha^2\eta^4\sum_{k=1}^K\Big(\sum_{j=k+1}^K w_j(2C_g^2)^{j-k}\Big)+2\rho_y\eta L_f^2K\sum_{k=1}^K C_g^{2(K-k)}w_{K+2}\\
&\quad+4\rho_y\alpha^2\eta^5L_f^2K\sum_{k=1}^K\Big(2^{K+1}C_g^{2(K-k)}\Big)w_{K+2}-\alpha\eta^2\sum_{k=1}^K w_k\Bigg)\frac{1}{N}\sum_{n=1}^N\mathbb{E}\Big[\Big\|h_{n,t}^{(k)}-g_n^{(k)}(h_{n,t}^{(k-1)})\Big\|^2\Big]\\
&+6\tau^2\rho_x^2\eta^2C_G^2C_f^2\Big(2\gamma_x^2\eta^2\sum_{k=1}^K w_k(2C_g^2)^k+\frac{4\eta\gamma_x^2}{\rho_x}\Big(C_g^{4K}L_f^2+\sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2\Big)(K+1)w_{K+1}+\frac{4\eta\gamma_x^2L_f^2C_G^2}{\rho_y}w_{K+2}\Big)\\
&+\Big(\frac{4\eta\gamma_y^2}{\rho_x}(K+1)C_g^{2K}L_f^2w_{K+1}+\frac{4\eta\gamma_y^2L_f^2}{\rho_y}w_{K+2}\Big)\Big(1152\gamma_x^2\rho_y^2\tau^4\eta^4L_f^2C_G^2C_f^2\sum_{k=1}^K(2C_g^2)^k+2304\alpha^2\rho_y^2\tau^4\eta^6\delta^2L_f^2\sum_{k=1}^K\sum_{j=1}^{k-1}(2C_g^2)^j\\
&+18432\alpha^2\rho_y^2\tau^4\eta^6\delta^2L_f^2K+2304\gamma_x^2\rho_y^2\tau^4\eta^4L_f^2C_g^2C_G^2C_f^2+110592\alpha^2\gamma_x^2\rho_x^2\rho_y^2\tau^8\eta^{10}L_f^2C_g^2C_G^2C_f^2+76\tau^2\rho_y^2\eta^2\sigma^2\Big)\\
&+2\alpha^2\eta^4\delta^2\sum_{k=1}^K w_k\sum_{j=1}^{k-1}(2C_g^2)^j+4\rho_x\gamma_x^2\eta^3C_G^2C_f^2K\sum_{k=1}^K A_k(2C_g^2)^kw_{K+1}\\
&+8\rho_x\alpha^2\eta^5\delta^2K\sum_{k=1}^K A_k\sum_{j=1}^{k-1}(2C_g^2)^jw_{K+1}+\rho_x^2\eta^2\Big(K(K+1)C_g^{2(K-1)}C_f^2+(K+1)C_g^{2K}\Big)\frac{\sigma^2}{N}w_{K+1}\\
&+w_{K+2}\rho_y^2\eta^2\frac{\sigma^2}{N}+2\rho_y\gamma_x^2\eta^3K^2C_G^2C_f^2L_f^2(2C_g^2)^Kw_{K+2}+4\alpha^2\rho_y\eta^5\delta^2L_f^2K\sum_{k=1}^K\sum_{j=1}^{k-1}C_g^{2(K-(k-j))}w_{K+2}\,,
\end{aligned}
\tag{62}
$$

where $A_k=\left(C_g^{2(K-1)}C_f^2L_g^2\Big(\sum_{j'=k}^{K-1}C_g^{j'-k}\Big)^2+C_g^{2K}L_f^2C_g^{2(K-k)}\right)$ and $B_k=\sum_{j=k+1}^K A_j(2C_g^2)^{j-k}$.

By setting $w_0=\frac{10\gamma_xC_G^2L_f^2}{\gamma_y\mu}$, we can get $2\gamma_x\eta C_G^2L_f^2-\frac{\eta\gamma_y\mu}{4}w_0\le-\frac{\gamma_x\eta C_G^2L_f^2}{2}$. By setting $w_{K+1}=\frac{2\gamma_x}{\rho_x}$, we can get $2\gamma_x\eta-\rho_x\eta w_{K+1}\le 0$. Moreover, by setting $w_{K+2}=\frac{125\gamma_xC_G^2L_f^2}{\rho_y\mu^2}$, we can get $\frac{25\eta\gamma_y}{2\mu}w_0-\rho_y\eta w_{K+2}\le 0$.

Then, we enforce

$$
\begin{aligned}
&4\rho_x\eta K\sum_{k=1}^K A_kw_{K+1}+8\rho_x\alpha^2\eta^5K\sum_{k=1}^K B_kw_{K+1}+2\alpha^2\eta^4\sum_{k=1}^K\Big(\sum_{j=k+1}^K w_j(2C_g^2)^{j-k}\Big)\\
&\quad+2\rho_y\eta L_f^2K\sum_{k=1}^K C_g^{2(K-k)}w_{K+2}+4\rho_y\alpha^2\eta^5L_f^2K\sum_{k=1}^K\Big(2^{K+1}C_g^{2(K-k)}\Big)w_{K+2}-\alpha\eta^2\sum_{k=1}^K w_k\le 0\,.
\end{aligned}
\tag{63}
$$

This is equivalent to enforce

$$
2\alpha^2\eta^4\Big(\sum_{j=k+1}^K w_j(2C_g^2)^{j-k}\Big)k-\alpha\eta^2w_k\le-\frac{1}{2}\alpha\eta^2w_k\,.
\tag{64}
$$

Then, we can get

$$\eta \leq \frac{1}{2}\sqrt{\frac{w_k}{\alpha\Big(\sum_{j=k+1}^{K} w_j (2C_g^2)^{j-k}\Big)}}\,. \tag{65}$$

We also have

$$\begin{aligned}
&8\eta\gamma_x K A_k + 16\alpha^2\eta^5\gamma_x K B_k + \frac{250\eta\gamma_x C_G^2 L_f^4}{\mu^2} K C_g^{2(K-k)} + \frac{500\alpha^2\eta^5\gamma_x C_G^2 L_f^4}{\mu^2} K\Big(2^{K+1} C_g^{2(K-k)}\Big) - \frac{1}{2}\alpha\eta^2 w_k \\
&\leq \alpha\eta^2\Big[\frac{1}{\alpha\eta} 8\gamma_x K A_k + 16\alpha\eta^2\gamma_x K B_k + \frac{1}{\alpha\eta}\frac{250\gamma_x C_G^2 L_f^4}{\mu^2} K C_g^{2(K-k)} + \frac{500\alpha\eta^2\gamma_x C_G^2 L_f^4}{\mu^2} K\Big(2^{K+1} C_g^{2(K-k)}\Big) - \frac{1}{2} w_k\Big]\,.
\end{aligned} \tag{66}$$

We enforce this upper to be non-positive so that

$$\begin{aligned}
&\frac{1}{\alpha\eta} 8\gamma_x K A_k + 16\alpha\eta^2\gamma_x K B_k + \frac{1}{\alpha\eta}\frac{250\gamma_x C_G^2 L_f^4}{\mu^2} K C_g^{2(K-k)} + \frac{500\alpha\eta^2\gamma_x C_G^2 L_f^4}{\mu^2} K\Big(2^{K+1} C_g^{2(K-k)}\Big) - \frac{1}{2} w_k \leq 0\,, \\
&w_k \geq \frac{1}{\alpha\eta} 16\gamma_x K A_k + 32\alpha\eta^2\gamma_x K B_k + \frac{1}{\alpha\eta}\frac{500\gamma_x C_G^2 L_f^4}{\mu^2} K C_g^{2(K-k)} + \frac{1000\alpha\eta^2\gamma_x C_G^2 L_f^4}{\mu^2} K\Big(2^{K+1} C_g^{2(K-k)}\Big) \\
&\quad = \frac{1}{\eta}\Big[\frac{1}{\alpha} 16\gamma_x K A_k + 32\alpha\eta^3\gamma_x K B_k + \frac{1}{\alpha}\frac{500\gamma_x C_G^2 L_f^4}{\mu^2} K C_g^{2(K-k)} + \frac{1000\alpha\eta^3\gamma_x C_G^2 L_f^4}{\mu^2} K\Big(2^{K+1} C_g^{2(K-k)}\Big)\Big]\,.
\end{aligned} \tag{67}$$

Due to $\alpha\eta^2 \leq 1$, $\eta \leq 1$, we can get

$$w_k \triangleq \frac{\gamma_x K}{\eta\mu^2}\tilde{w}_k = \frac{\gamma_x K}{\eta\mu^2}\Big[\frac{\mu^2}{\alpha} 16 A_k + 32 B_k\mu^2 + \frac{500 C_G^2 L_f^4}{\alpha} C_g^{2(K-k)} + 1000 C_G^2 L_f^4\Big(2^{K+1} C_g^{2(K-k)}\Big)\Big]\,. \tag{68}$$

Therefore, we can simplify the Eq. (65) as follows:

$$\eta \leq \frac{1}{2}\sqrt{\frac{\tilde{w}_k}{\alpha\Big(\sum_{j=k+1}^{K} \tilde{w}_j (2C_g^2)^{j-k}\Big)}}\,. \tag{69}$$

Then, we enforce

$$\begin{aligned}
&\frac{25\eta\gamma_x^2 C_G^2 L_f^2}{6\gamma_y\mu^3} w_0 + 2\gamma_x^2\eta^2 \sum_{k=1}^{K} w_k (2C_g^2)^k + \frac{4\eta\gamma_x^2}{\rho_x}(K+1)\Big(C_g^{4K} L_f^2 + \sum_{k=0}^{K-1} C_g^{2(K-1+k)} C_f^2 L_g^2\Big) w_{K+1} \\
&+ \frac{4\eta\gamma_x^2 L_f^2 C_G^2}{\rho_y} w_{K+2} - \frac{\gamma_x\eta}{4} \leq 0\,.
\end{aligned} \tag{70}$$

This is equivalent to enforce

$$\frac{125\gamma_x^2 C_G^4 L_f^4}{3\gamma_y^2\mu^4} + 2\gamma_x\eta \sum_{k=1}^{K} w_k (2C_g^2)^k + \frac{8\gamma_x^2}{\rho_x^2}(K+1)\Big(C_g^{4K} L_f^2 + \sum_{k=0}^{K-1} C_g^{2(K-1+k)} C_f^2 L_g^2\Big) + \frac{500\gamma_x^2 C_G^4 L_f^4}{\rho_y^2\mu^2} - \frac{1}{4} \leq 0\,. \tag{71}$$

Furthermore, we have the following inequalities

$$\begin{aligned}
\frac{125\gamma_x^2 C_G^4 L_f^4}{3\gamma_y^2\mu^4} - \frac{1}{2} &\leq -\frac{1}{4}\,, \\
\frac{2\gamma_x^2 K}{\mu^2} \sum_{k=1}^{K} \tilde{w}_k (2C_g^2)^k &\leq \frac{1}{12}\,, \\
\frac{8\gamma_x^2}{\rho_x^2}(K+1)(C_g^{4K} L_f^2 + \sum_{k=0}^{K-1} C_g^{2(K-1+k)} C_f^2 L_g^2) &\leq \frac{1}{12}\,,
\end{aligned}$$

$$\frac{500\gamma_x^2 L_f^4 C_G^4}{\rho_y^2\mu^2} \leq \frac{1}{12} \,. \tag{72}$$

By solving these inequalities, we can get

$$\gamma_x \leq \min\{\frac{\gamma_y\mu^2}{13C_G^2L_f^2}, \frac{\rho_y\mu}{78C_G^2L_f^2}, \frac{\rho_x}{\sqrt{96(K+1)(C_g^{4K}L_f^2+\sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2)}}, \frac{\mu}{\sqrt{24K\sum_{k=1}^{K}\tilde{w}_k(2C_g^2)^k}}\}. \tag{73}$$

Similarly,

$$\frac{4\eta\gamma_y^2}{\rho_x}C_g^{2K}L_f^2(K+1)w_{K+1} + \frac{4\eta\gamma_y^2L_f^2}{\rho_y}w_{K+2} - \frac{3\eta\gamma_y^2}{4}w_0 \leq 0 \,. \tag{74}$$

This is equivalent to enforce

$$\frac{8\gamma_y^2C_g^{2K}L_f^2(K+1)}{\rho_x^2} + \frac{500\gamma_y^2C_G^2L_f^4}{\rho_y^2\mu^2} - \frac{15\gamma_yC_G^2L_f^2}{2\mu} \leq 0 \,. \tag{75}$$

Furthermore, we have the following inequalities

$$\frac{500\gamma_yL_f^2}{\rho_y^2\mu^2} \leq \frac{15}{\mu} \,, \qquad\qquad \frac{8\gamma_y(K+1)}{\rho_x^2} \leq \frac{15}{\mu} \,, \tag{76}$$

By solving these inequalities, we can get

$$\gamma_y \leq \min\{\frac{3\rho_y^2\mu}{100L_f^2}, \frac{15\rho_x^2}{8\mu(K+1)}\}. \tag{77}$$

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

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}\Big[\Big\|\nabla\Phi(\bar{x}_t)\Big\|^2\Big] + C_G^2L_f^2\mathbb{E}\|y^*(\bar{x}_t)-\bar{y}_t\|^2) \\
&\leq \frac{2(P_0-P_T)}{\gamma_x\eta T} + 6\gamma_x^2\rho_x^2\tau^4\eta^4C_G^2C_f^2\Big(\frac{250C_G^4L_f^4}{\mu^2} + 4(K+1)\Big(C_g^{4K}L_f^2 + \sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2\Big)\Big) \\
&\quad + \Big(\frac{250C_G^2L_f^4}{\mu^2} + 4(K+1)C_g^{2K}L_f^2\Big)(1152\gamma_x^2\gamma_y^2\rho_y^2\tau^6\eta^6L_f^2C_G^2C_f^2\sum_{k=1}^{K}(2C_g^2)^k + 2304\alpha^2\gamma_y^2\rho_y^2\tau^6\eta^8\delta^2L_f^2\sum_{k=1}^{K}\sum_{j=1}^{k-1}(2C_g^2)^j \\
&\quad + 18432\alpha^2\rho_y^2\gamma_y^2\tau^6\eta^8\delta^2L_f^2K + 2304\gamma_x^2\gamma_y^2\rho_y^2\tau^6\eta^6L_f^2C_g^2C_G^2C_f^2 + 110592\alpha^2\rho_x^2\gamma_x^2\rho_y^2\gamma_y^2\tau^{10}\eta^{12}L_f^2C_g^2C_G^2C_f^2 + 76\tau^4\eta^4\gamma_y^2\rho_y^2\sigma^2) \\
&\quad + 6\tau^2\rho_x^2\eta^2C_G^2C_f^2\Big(\frac{4\gamma_x^2K}{\mu^2}\sum_{k=1}^{K}\tilde{w}_k(2C_g^2)^k + \frac{16\gamma_x^2}{\rho_x^2}\Big(C_g^{4K}L_f^2 + \sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2\Big)(K+1) + \frac{1000\gamma_x^2C_G^4L_f^4}{\rho_y^2\mu^2}\Big) \\
&\quad + \Big(\frac{16\gamma_y^2}{\rho_x^2}(K+1)C_g^{2K}L_f^2 + \frac{1000\gamma_y^2C_G^2L_f^4}{\rho_y^2\mu^2}\Big)\Big(1152\gamma_x^2\rho_y^2\tau^4\eta^4L_f^2C_G^2C_f^2\sum_{k=1}^{K}(2C_g^2)^k + 2304\alpha^2\rho_y^2\tau^4\eta^6\delta^2L_f^2\sum_{k=1}^{K}\sum_{j=1}^{k-1}(2C_g^2)^j \\
&\quad + 18432\alpha^2\rho_y^2\tau^4\eta^6\delta^2L_f^2K + 2304\gamma_x^2\rho_y^2\tau^4\eta^4L_f^2C_g^2C_G^2C_f^2 + 110592\alpha^2\gamma_x^2\rho_x^2\rho_y^2\tau^8\eta^{10}L_f^2C_g^2C_G^2C_f^2 + 76\tau^2\rho_y^2\eta^2\sigma^2\Big) \\
&\quad + 4\alpha^2\eta^2\delta^2\sum_{k=1}^{K}\frac{K}{\mu^2}\tilde{w}_k\sum_{j=1}^{k-1}(2C_g^2)^j + 16\gamma_x^2\eta^2C_G^2C_f^2K\sum_{k=1}^{K}A_k(2C_g^2)^k \\
&\quad + 32\alpha^2\eta^3\delta^2K\sum_{k=1}^{K}A_k\sum_{j=1}^{k-1}(2C_g^2)^j + 4\rho_x\eta\Big(K(K+1)C_g^{2(K-1)}C_f^2 + (K+1)C_g^{2K}\Big)\frac{\sigma^2}{N} \\
&\quad + \frac{250\rho_y\eta C_G^2L_f^2\sigma^2}{\mu^2N} + \frac{500\times 2^K\gamma_x^2\eta^2K^2C_f^2C_G^6L_f^4}{\mu^2} + \frac{1000\alpha^2\eta^4\delta^2C_G^2L_f^4K}{\mu^2}\sum_{k=1}^{K}\sum_{j=1}^{k-1}C_g^{2(K-(k-j))} \,.
\end{aligned} \tag{78}$$

The last step is to compute the value of P_0.

$$P_0 = \mathbb{E}[\Phi(\bar{x}_0)] + \frac{10\gamma_x C_G^2 L_f^2}{\gamma_y \mu}\mathbb{E}\Big[\Big\|\bar{y}_0 - y^*(\bar{x}_0)\|^2\Big] + \frac{\gamma_x K}{\eta\mu^2}\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\tilde{w}_k\mathbb{E}\Big[\Big\|h_{n,0}^{(k)} - g_n^{(k)}(h_{n,0}^{(k-1)})\Big\|^2\Big]$$
$$+ \frac{2\gamma_x}{\rho_x}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}p_{n,0} - \frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,0})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,0}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,0}))\nabla_1 f_n(G_n^{(K)}(x_{n,0}), y_{n,0})\Big\|^2\Big]$$
$$+ \frac{125\gamma_x C_G^2 L_f^2}{\rho_y \mu^2}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}q_{n,0} - \frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(G_n(x_{n,0}), y_{n,0})\Big\|^2\Big]\,. \tag{79}$$

For the initialization step, we select a mini-batch of samples denoted a M to initialize $h_{n,0}^{(k)}$ for $k = \{1, \cdots, K\}$. Then, we can get

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\tilde{w}_k\mathbb{E}\Big[\Big\|h_{n,0}^{(k)} - g_n^{(k)}(h_{n,0}^{(k-1)})\Big\|^2\Big] = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\tilde{w}_k\mathbb{E}\Big[\Big\|g_n^{(k)}(h_{n,0}^{(k-1)};\xi_{n,0}^{(k)}) - g_n^{(k)}(h_{n,0}^{(k-1)})\Big\|^2\Big] \leq K\frac{\delta^2}{M}\sum_{k=1}^{K}\tilde{w}_k\,. \tag{80}$$

Similarly, for the initialization of $p_{n,0}$, we have

$$\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}p_{n,0} - \frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,0})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,0}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,0}))\nabla_1 f_n(G_n^{(K)}(x_{n,0}), y_{n,0})\Big\|^2\Big]$$
$$= \mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,0};\xi_{n,0}^{(1)})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,0});\xi_{n,0}^{(1)})\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,0});\xi_{n,0}^{(K)})\nabla_1 f_n(G_n^{(K)}(x_{n,0}), y_{n,0};\zeta_{n,0})$$
$$- \frac{1}{N}\sum_{n=1}^{N}\nabla g_n^{(1)}(x_{n,0})\nabla g_n^{(2)}(G_n^{(1)}(x_{n,0}))\cdots\nabla g_n^{(K)}(G_n^{(K-1)}(x_{n,0}))\nabla_1 f_n(G_n^{(K)}(x_{n,0}), y_{n,0})\Big\|^2\Big]$$
$$\leq \frac{K(K+1)C_g^{2(K-1)}C_f^2}{N}\frac{\sigma^2}{M} + \frac{(K+1)C_g^{2K}}{N}\frac{\sigma^2}{M}\,. \tag{81}$$

And for the initialization of $q_{n,0}$, we also have

$$\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}q_{n,0} - \frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(G_n(x_{n,0}), y_{n,0})\Big\|^2\Big]$$
$$= \mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(G_n(x_{n,0}), y_{n,0};\zeta_{n,0}) - \frac{1}{N}\sum_{n=1}^{N}\nabla_2 f_n(G_n(x_{n,0}), y_{n,0})\Big\|^2\Big] \leq \frac{1}{N}\frac{\sigma^2}{M}\,. \tag{82}$$

Therefore, we have

$$P_0 = \mathbb{E}[\Phi(\bar{x}_0)] + \frac{10\gamma_x C_G^2 L_f^2}{\gamma_y\mu}\mathbb{E}[\|\bar{y}_0 - y^*(\bar{x}_0)\|^2] + \frac{\gamma_x K^2\delta^2}{\eta\mu^2}\sum_{k=1}^{K}\tilde{w}_k$$
$$+ \frac{2\gamma_x\sigma^2(K+1)C_g^{2(K-1)}}{\rho_x N}(KC_f^2 + C_g^2) + \frac{125\gamma_x C_G^2 L_f^2\sigma^2}{\rho_y\mu^2 N}\,. \tag{83}$$

At last, we successfully establish the convergence rate of our algorithm as follows:

$$\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|^2] + C_G^2 L_f^2\mathbb{E}\|y^*(\bar{x}_t) - \bar{y}_t\|^2])$$
$$\leq \frac{2(\Phi(x_0) - \Phi^*)}{\gamma_x\eta T} + \frac{20C_G^2 L_f^2}{\gamma_y\eta\mu T}\mathbb{E}[\|\bar{y}_0 - y^*(\bar{x}_0)\|^2]$$
$$+ 6\gamma_x^2\rho_x^2\tau^4\eta^4 C_G^2 C_f^2\Big(\frac{250C_G^4 L_f^4}{\mu^2} + 4(K+1)\Big(C_g^{4K}L_f^2 + \sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2 L_g^2\Big)\Big)$$
$$+ \Big(\frac{250C_G^2 L_f^4}{\mu^2} + 4(K+1)C_g^{2K}L_f^2\Big)(1152\gamma_x^2\gamma_y^2\rho_y^2\tau^6\eta^6 L_f^2 C_G^2 C_f^2\sum_{k=1}^{K}(2C_g^2)^k + 2304\alpha^2\gamma_y^2\rho_y^2\tau^6\eta^8\delta^2 L_f^2\sum_{k=1}^{K}\sum_{j=1}^{k-1}(2C_g^2)^j$$

$$
\begin{aligned}
&+ 18432\alpha^2\rho_y^2\gamma_y^2\tau^6\eta^8\delta^2L_f^2K + 2304\gamma_x^2\gamma_y^2\rho_y^2\tau^6\eta^6L_f^2C_g^2C_G^2C_f^2 + 110592\alpha^2\rho_x^2\gamma_x^2\rho_y^2\gamma_y^2\tau^{10}\eta^{12}L_f^2C_g^2C_G^2C_f^2 + 76\tau^4\eta^4\gamma_y^2\rho_y^2\sigma^2) \\
&+ 6\tau^2\rho_x^2\eta^2C_G^2C_f^2\Big(\frac{4\gamma_x^2K}{\mu^2}\sum_{k=1}^{K}\tilde{w}_k(2C_g^2)^k + \frac{16\gamma_x^2}{\rho_x^2}\Big(C_g^{4K}L_f^2 + \sum_{k=0}^{K-1}C_g^{2(K-1+k)}C_f^2L_g^2\Big)(K+1) + \frac{1000\gamma_x^2C_G^4L_f^4}{\rho_y^2\mu^2}\Big) \\
&+ \Big(\frac{16\gamma_y^2}{\rho_x^2}(K+1)C_g^{2K}L_f^2 + \frac{1000\gamma_y^2C_G^2L_f^4}{\rho_y^2\mu^2}\Big)\Big(1152\gamma_x^2\rho_y^2\tau^4\eta^4L_f^2C_G^2C_f^2\sum_{k=1}^{K}(2C_g^2)^k + 2304\alpha^2\rho_y^2\tau^4\eta^6\delta^2L_f^2\sum_{k=1}^{K}\sum_{j=1}^{k-1}(2C_g^2)^j \\
&+ 18432\alpha^2\rho_y^2\tau^4\eta^6\delta^2L_f^2K + 2304\gamma_x^2\rho_y^2\tau^4\eta^4L_f^2C_g^2C_G^2C_f^2 + 110592\alpha^2\gamma_x^2\rho_x^2\rho_y^2\tau^8\eta^{10}L_f^2C_g^2C_G^2C_f^2 + 76\tau^2\rho_y^2\eta^2\sigma^2\Big) \\
&+ 4\alpha^2\eta^2\delta^2\sum_{k=1}^{K}\frac{K}{\mu^2}\tilde{w}_k\sum_{j=1}^{k-1}(2C_g^2)^j + 16\gamma_x^2\eta^2C_G^2C_f^2K\sum_{k=1}^{K}A_k(2C_g^2)^k \\
&+ 32\alpha^2\eta^3\delta^2K\sum_{k=1}^{K}A_k\sum_{j=1}^{k-1}(2C_g^2)^j + 4\rho_x\eta\Big(K(K+1)C_g^{2(K-1)}C_f^2 + (K+1)C_g^{2K}\Big)\frac{\sigma^2}{N} \\
&+ \frac{250\rho_y\eta C_G^2L_f^2\sigma^2}{\mu^2N} + \frac{500\times 2^K\gamma_x^2\eta^2K^2C_f^2C_G^6L_f^4}{\mu^2} + \frac{1000\alpha^2\eta^4\delta^2C_G^2L_f^4K}{\mu^2}\sum_{k=1}^{K}\sum_{j=1}^{k-1}C_g^{2(K-(k-j))} \\
&+ \frac{2K^2\delta^2}{\mu^2\eta^2MT}\sum_{k=1}^{K}\tilde{w}_k + \frac{4\sigma^2(K+1)C_g^{2(K-1)}}{\rho_x\eta TNM}(KC_f^2 + C_g^2) + \frac{250C_G^2L_f^2\sigma^2}{\rho_y\mu^2NM\eta T}\,. \qquad (84)
\end{aligned}
$$

Since $\alpha, \rho_x, \rho_y, \gamma_x, \gamma_y$ are hyperparameters for the algorithm, they are independent of the number of iterations. Therefore, we can obtain

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|^2] + C_G^2L_f^2\mathbb{E}\|y^*(\bar{x}_t) - \bar{y}_t\|^2]) \\
&\leq \frac{2(\Phi(x_0)-\Phi^*)}{\gamma_x\eta T} + \frac{20C_G^2L_f^2}{\gamma_y\eta\mu T}\mathbb{E}[\|\bar{y}_0 - y^*(\bar{x}_0)\|^2] + O(K\tau^2\eta^2) + O(K\tau^4\eta^4) + O(K\tau^6\eta^6) \\
&\quad + O(K\tau^8\eta^8) + O(K\tau^{10}\eta^{10}) + O(\frac{K^2}{\eta T}) + O(\eta^2K^2) + O(\frac{\eta K^2}{N}) + O(\frac{K^2}{\eta^2MT})\,, \qquad (85)
\end{aligned}
$$

where Φ^* denotes the function value at the optimal solution, M is the mini-batch size in initialization step.