On the Convergence of Distributed Stochastic Bilevel Optimization Algorithms over a Network

Hongchang Gao
Temple University

Bin Gu
MBZUAI

My T. Thai
University of Florida

Abstract

Bilevel optimization has been applied to a wide variety of machine learning models and numerous stochastic bilevel optimization algorithms have been developed in recent years. However, most existing algorithms restrict their focus on the single-machine setting so that they are incapable of handling the distributed data. To address this issue, under the setting where all participants compose a network and perform peer-to-peer communication in this network, we developed two novel decentralized stochastic bilevel optimization algorithms based on the gradient tracking communication mechanism and two different gradient estimators. Additionally, we established their convergence rates for nonconvex-strongly-convex problems with novel theoretical analysis strategies. To our knowledge, this is the first work achieving these theoretical results. Finally, we applied our algorithms to practical machine learning models, and the experimental results confirmed the efficacy of our algorithms.

1 Introduction

Bilevel optimization is an important learning paradigm in machine learning. It consists of an upper-level optimization problem and a lower-level optimization problem, where the objective function of the upper-level optimization problem depends on the solution of the lower-level one. This kind of learning paradigm covers numerous machine learning models, such as hyperparameter optimization (Feurer and Hutter, 2019; Franceschi et al., 2017), meta-learning (Franceschi et al., 2018; Rajeswaran et al., 2019), neural architecture search (Liu et al., 2018), etc. Thus, it is of importance and necessity to develop efficient optimization algorithms to solve bilevel optimization problems.

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

In fact, the bilevel structure makes it difficult to compute the gradient of the outer-level optimization problem since it involves the computation of Hessian and Jacobian matrices. In the past few years, numerous optimization algorithms have been proposed to address this challenge. For instance, Ghadimi and Wang (2018); Hong et al. (2020); Ji et al. (2021); Chen et al. (2021) developed stochastic-gradient-based optimization algorithms, which are able to efficiently estimate Hessian and Jacobian matrices. Recently, to accelerate the convergence speed, Guo et al. (2021) developed a momentum-based optimization algorithm, and Yang et al. (2021); Khanduri et al. (2021b); Guo and Yang (2021) proposed the variance-reduced optimization algorithms. Both categories are able to improve the estimation of the full gradient. Thus, they can achieve faster convergence speed than the vanilla stochastic gradient based algorithms (Ghadimi and Wang, 2018; Hong et al., 2020; Ji et al., 2021; Chen et al., 2021) . However, all these bilevel optimization algorithms restrict their focus on the non-parallel setting. As a result, they are not applicable to the distributed setting.

In this work, we aim to develop decentralized bilevel optimization algorithms to solve the following bilevel distributed optimization problem:

$$\begin{aligned} &\min_{x\in\mathbb{R}^{d_x}} \frac{1}{K}\sum_{k=1}^{K} f^{(k)}(x, y^*(x)), \\ &s.t. \quad y^*(x) = \arg\min_{y\in\mathbb{R}^{d_y}} \frac{1}{K}\sum_{k=1}^{K} g^{(k)}(x,y)\ , \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}^{d_y}$ are the model parameters, $g^{(k)}(x,y) = \mathbb{E}_{\zeta\sim\mathcal{S}_g^{(k)}}[g^{(k)}(x,y;\zeta)]$ denotes the objective function of the lower-level subproblem in the k-th participant and $\mathcal{S}_g^{(k)}$ is the data distribution of the k-th participant for the low-level function, $f^{(k)}(x,y) = \mathbb{E}_{\xi\sim\mathcal{S}_f^{(k)}}[f^{(k)}(x,y;\xi)]$ is the objective function of the upper-level subproblem in the k-th participant and correspondingly $\mathcal{S}_f^{(k)}$ is the data distribution of the k-th participant for the upper-level function, K is the total number of the participants. From Eq. (1), it is easy to know that each participant possesses its own data, which will be used to learn the model parameters (x, y) via the collaboration among all participants. In this work, it is assumed that all participants compose a network where the participant performs peer-to-peer communication. Thus, it is a decentralized bilevel optimization problem.

Decentralized optimization has been extensively studied in recent years due to its great potential in real-world machine learning tasks, such as data analysis on Internet-of-Things (IoT) devices (Gao et al., 2023). To address the challenges under various settings, several decentralized optimization algorithms have been proposed. For example, Lian et al. (2017) developed the decentralized stochastic gradient descent (DSGD) algorithm based on the gossip communication mechanism and established its convergence rate for nonconvex problems. Sun et al. (2020); Xin et al. (2021); Zhang et al. (2021a); Zhan et al. (2022) proposed the decentralized stochastic variance-reduced gradient descent algorithms based on the gradient tracking communication mechanism, which improve the convergence rate of DSGD. Additionally, some efforts (Koloskova et al., 2019; Tang et al., 2019; Vogels et al., 2020; Li et al., 2019; Gao and Huang, 2020) were made to improve the communication complexity of decentralized optimization algorithms by compressing the communicated variables or skipping the communication round.

However, all aforementioned decentralized optimization algorithms restrict their focus on the *single-level* minimization problem, which are not applicable to the *bilevel* optimization problem. In particular, bilevel optimization involves the computation of Hessian and Jacobian matrices. If communicating these two matrices, it will incur a large communication complexity. Thus, it is not clear whether these matrices should be communicated like the gradient. Moreover, considering the interaction between the bilevel structure and the communication mechanism, how will the decentralized bilevel optimization algorithm converge? Particularly, the stochastic hypergradient regarding x of the outer-level objective function is a biased estimation of the full gradient. How does this biased estimator affect the consensus error? All these problems regarding the algorithmic design and theoretical analysis for decentralized bilevel optimization are still unexplored.

To address the aforementioned problems, we proposed two novel decentralized bilevel optimization algorithms. Specifically, on the algorithmic design side, we developed a momentum-based decentralized stochastic bilevel optimization (MDBO) algorithm, which takes advantage of the momentum to update model parameters, and a variance-reduction-based decentralized stochastic bilevel optimization (VRDBO) algorithm, which leverages the variance-reduced gradient estimator, STORM (Cutkosky and Orabona, 2019), to update model parameters. Both of them employ the gradient tracking communication mechanism. Importantly, in our two algorithms, only model parameters and gradient estimators are communicated among participants. In this way, the computation of Hessian and Jacobian matrices is restricted in each participant, avoiding large communication overhead. On the theoretical analysis side, we established the convergence rate of our two algorithms. Specifically, we investigated how the biased gradient estimator affects the consensus error in the presence of the momentum and variance-reduced gradient. With the help of well-designed potential functions, we show that MDBO achieves the $O(\frac{1}{\epsilon^2(1-\lambda)^2})$ convergence rate and VRDBO enjoys the $O(\frac{1}{K\epsilon^{3/2}(1-\lambda)^2})$ convergence rate to obtain the ϵ-accuracy solution under mild conditions. We further show that MDBO can also achieve linear speedup when employing stronger assumptions as existing works (Yang et al., 2022). In summary, our work have made the following contributions:

- We developed two novel decentralized bilevel optimization algorithms for solving Eq. (1), which demonstrated how to update model parameters locally and communicate them across different participants.
- We established the convergence rate of our proposed algorithms, which demonstrated how the bilevel structure, the gradient estimator, and the communication mechanism affect the convergence rate.
- We applied our algorithms to the practical machine learning task. The empirical results confirm the superiority of our algorithms.

2 Related Works

Bilevel optimization has been widely applied to numerous machine learning applications. For instance, in the hyperparameter optimization task, the upper-level problem optimizes the hyperparameter and the lower-level problem optimizes the machine learning model's parameter. In the meta-learning task, the upper-level problem learns the task-shared model parameters while the lower-level problem learns the task-specific model parameters (Ji et al., 2021). When optimizing these kinds of bilevel machine learning models, the challenge lies in the computation of the inverse Hessian matrix $(\nabla^2_{yy}g^{(k)})^{-1}$. To address this issue, Ghadimi and Wang (2018) developed a Hessian inverse approximation strategy, which employs stochastic samples to compute an approximation for $(\nabla^2_{yy}g^{(k)})^{-1}$. Meanwhile, it employs the double-loop mechanism, where y is updated for multiple times before updating x, to obtain a good approximation for $y^*(x)$. Ji et al. (2021) further employed a large batch size to improve the approximation for $y^*(x)$. On the contrary, Hong et al. (2020) developed a single-loop method, which employs different step sizes for the model parameters x and y such that each update y is a good approximation for the optimal solution $y^*(x)$. It is worth noting that these single-loop and double-loop algorithms employ stochastic gradients to update model parameters, which suffer from a large estimation variance. To address this problem, Guo et al. (2021) developed a single-loop algorithm, which utilizes the momentum to update model parameters. Khanduri et al. (2021a) proposed another single-loop algorithm, which leverages a variance-reduced gradient estimator to update the model parameter x. However, they fail to achieve a better theoretical convergence rate than the vanilla stochastic gradient based algorithms. Recently, Yang et al. (2021); Khanduri et al. (2021b); Guo and Yang (2021) resorted to more advanced variance-reduced gradient estimators to accelerate the convergence rate. Specifically, Yang et al. (2021); Khanduri et al. (2021b); Guo and Yang (2021) combined the STORM (Cutkosky and Orabona, 2019) gradient estimator and the single-loop mechanism so that they can achieve a better theoretical convergence rate than the stochastic-gradient-based algorithm and the momentum-based algorithm. Moreover, Yang et al. (2021) combined the SPIDER (Fang et al., 2018) gradient estimator and the double-loop mechanism, which actually can achieve the same theoretical convergence rate with that based on STORM. However, all these algorithms only investigate the non-parallel situation. Thus, their theoretical analysis does not hold anymore for the distributed scenario.

Decentralized optimization has also been applied to a wide variety of machine learning applications in recent years. Compared to the parameter-server setting, the decentralized communication is robust to the single-node failure since the participant conducts peer-to-peer communication. Recently, Lian et al. (2017) investigated the convergence rate of the standard decentralized stochastic gradient descent (SGD) algorithm for nonconvex problems. Yu et al. (2019) developed a decentralized stochastic gradient descent with momentum algorithm, which has the same theoretical convergence rate as (Lian et al., 2017). Pu and Nedić (2021); Lu et al. (2019) developed a decentralized SGD based on the gradient tracking communication mechanism. Later, some variance-reduced algorithms were proposed to accelerate the convergence rate. For instance, Sun et al. (2020) utilized the SPIDER (Fang et al., 2018) gradient estimator, Xin et al. (2021); Zhang et al. (2021a) employed the STORM (Cutkosky and Orabona, 2019) gradient estimator, and Zhan et al. (2022) resorted to the ZeroSARAH (Li et al., 2021) gradient estimator for improving the sample and communication complexities. Additionally, some works consider decentralized optimization problems that are beyond standard minimization problems, e.g., compositional optimization problems (Gao and Huang, 2021), minimax optimization problems (Tsaknakis et al., 2020; Xian et al., 2021; Zhang et al., 2021b; Gao, 2022b), constraint optimization problems (Wai et al., 2017; Mokhtari et al., 2018; Gao et al., 2021). However, all these decentralized optimization algorithms are not applicable to the decentralized bilevel optimization problem. On the one hand, they focus on the single-level problem. Thus, their theoretical analysis is incapable of handling the interaction between two levels of functions. On the other hand, those algorithms are based on the standard stochastic gradient, which is an unbiased estimator of the full gradient. On the contrary, the stochastic hypergradient is a biased estimator, which incurs new challenges when bounding the consensus error. Thus, it is necessary to develop new theoretical analysis strategies to investigate the convergence rate of decentralized bilevel optimization algorithms.

Recently, we are aware of two concurrent works (Chen et al., 2022; Yang et al., 2022) on decentralized bilevel optimizaiton. In particular, Chen et al. (2022) developed the decentralized optimization algorithms for Eq. (1) based on the *full* gradient and the vanilla *stochastic gradient*. On the contrary, our work leverages advanced gradient estimators, i.e., momentum and variance-reduced gradient, which makes our theoretical analysis more challenging. As for (Yang et al., 2022), it requires to communicate Hessian and Jacobian matrices, while our methods do not require that. Thus, our method is more efficient in communication. Moreover, Yang et al. (2022) employed the gossip communication mechanism, while our methods leverage the gradient tracking scheme and advanced gradient estimators. As a result, the theoretical analysis of our methods is more challenging. *It is worth noting that both of them (Chen et al., 2022; Yang et al., 2022) require much stronger assumption for the loss function. In particular, they require that the upper-level objective function is Lipschitz continuous with respect to x and the lower-level objective function is Lipschitz continuous with respect to y (See Assumption 2.1 in (Chen et al., 2022), Assumption 3.3 (iii) and 3.4 (iv) in (Yang et al., 2022)).* On the contrary, our theoretical analysis does not require these two strong assumptions, which incurs more challenges for theoretical analysis. All in all, our work is significantly different from these two concurrent works and the theoretical analysis is more challenging than them.

3 Preliminaries

Stochastic Hypergradient. Throughout this paper, we denote $F^{(k)}(x) = f^{(k)}(x, y^*(x))$ and $F(x) = \frac{1}{K}\sum_{k=1}^{K} F^{(k)}(x)$ and assume all participants have i.i.d. datesets. Then, according to Lemma 1 of (Gao, 2022a), we can compute the gradient of $F^{(k)}(x)$ as follows:

$$\begin{aligned} \nabla F^{(k)}(x) &= \nabla_x f^{(k)}(x, y^*(x)) \\ &- \nabla^2_{xy} g^{(k)}(x, y^*(x)) H_*^{-1} \nabla_y f^{(k)}(x, y^*(x)) \,, \end{aligned} \tag{2}$$

where $\nabla^2_{xy} g^{(k)}(x, y^*(x))$ is Jacobian matrix and $H_* = \nabla^2_{yy} g^{(k)}(x, y^*(x))$ is Hessian matrix. Note that $\nabla F^{(k)}(x)$ is also called *hypergradient*. Since $y^*(x)$ is typically not easy to obtain in each iteration, following (Ghadimi and

Wang, 2018), we can approximate it as follows:

$$\nabla F^{(k)}(x,y) = \nabla_x f^{(k)}(x,y) - \nabla^2_{xy} g^{(k)}(x,y) H^{-1} \nabla_y f^{(k)}(x,y) \;, \tag{3}$$

where $H = \nabla^2_{yy} g^{(k)}(x,y)$. Note that we just use $\nabla F^{(k)}(x,y)$ to approximate $\nabla F^{(k)}(x)$. It does not mean we introduce a function $F^{(k)}(x,y)$. Here, because the inverse of Hessian matrix is difficult to compute, following the Hessian inverse approximation strategy proposed in (Ghadimi and Wang, 2018), we can use the following stochastic hypergradient to approximate it:

$$\nabla \tilde{F}^{(k)}(x,y;\tilde{\xi}) = \nabla_x f^{(k)}(x,y;\xi) - \nabla^2_{xy} g^{(k)}(x,y;\zeta_0) \frac{J}{L_{g_y}} \tilde{H}_{\tilde{J}} \nabla_y f^{(k)}(x,y;\xi) \;, \tag{4}$$

where $\tilde{H}_{\tilde{J}} = \prod_{j=1}^{\tilde{J}} (I - \frac{1}{L_{g_y}} \nabla^2_{yy} g^{(k)}(x,y;\zeta_j))$, L_{g_y} is the Lipschitz-continuous constant, which is defined in Assumption 5, $\tilde{\xi} = \{\xi, \zeta_0, \zeta_1, \cdots, \zeta_J\}$ and $\tilde{J}$ are randomly selected from $\{0,1,2,\cdots,J\}$ where J is a positive integer. Note that we let $\prod_{j=1}^{\tilde{J}} (I - \frac{1}{L_{g_y}} \nabla^2_{yy} g^{(k)}(x,y;\zeta_j)) = I$ when $\tilde{J} = 0$. Moreover, we denote the expectation of the stochastic hypergradient as follows:

$$\nabla \tilde{F}^{(k)}(x,y) \triangleq \mathbb{E}[\nabla \tilde{F}^{(k)}(x,y;\tilde{\xi})] = \nabla_x f^{(k)}(x,y) - \nabla^2_{xy} g^{(k)}(x,y) \mathbb{E}\Big[\frac{J}{L_{g_y}} \tilde{H}_{\tilde{J}}\Big] \nabla_y f^{(k)}(x,y) \;, \tag{5}$$

Since $\nabla \tilde{F}^{(k)}(x,y) \neq \nabla F^{(k)}(x,y)$, the stochastic hypergradient $\nabla \tilde{F}^{(k)}(x,y;\tilde{\xi})$ is a biased estimator for $\nabla F^{(k)}(x,y)$. The detailed bias is shown in Lemma 3.

Notations. Throughout this paper, $x_t^{(k)}$ and $y_t^{(k)}$ denote the model parameters of the k-th participant in the t-th iteration. Moreover, we denote $\bar{x}_t = \frac{1}{K}\sum_{k=1}^K x_t^{(k)}$ and $\bar{y}_t = \frac{1}{K}\sum_{k=1}^K y_t^{(k)}$. Additionally, we denote

$$\begin{aligned} \Delta_t^{\tilde{F}_{\tilde{\xi}_t}} &= [\nabla \tilde{F}^{(1)}(x_t^{(1)}, y_t^{(1)}; \tilde{\xi}_t^{(1)}), \nabla \tilde{F}^{(2)}(x_t^{(2)}, y_t^{(2)}; \tilde{\xi}_t^{(2)}), \\ &\quad \cdots, \nabla \tilde{F}^{(k)}(x_t^{(K)}, y_t^{(K)}; \tilde{\xi}_t^{(K)})] \;, \\ \Delta_t^{g_{\zeta_t}} &= [\nabla_y g^{(k)}(x_t^{(1)}, y_t^{(1)}; \zeta_t^{(1)}), \nabla_y g^{(k)}(x_t^{(2)}, y_t^{(2)}; \zeta_t^{(2)}), \\ &\quad \cdots, \nabla_y g^{(k)}(x_t^{(K)}, y_t^{(K)}; \zeta_t^{(K)})] \;, \\ X_t &= [x_t^{(1)}, x_t^{(2)}, \cdots, x_t^{(K)}] \;, Y_t = [y_t^{(1)}, y_t^{(2)}, \cdots, y_t^{(K)}] \;. \end{aligned} \tag{6}$$

Furthermore, the adjacency matrix of the communication network is denoted by $W = [w_{ij}] \in \mathbb{R}_+^{K\times K}$, where $w_{ij} > 0$ indicates the i-th participant is connected with the j-th participant and otherwise $w_{ij} = 0$. The adjacency matrix satisfies the following assumption.

Assumption 1. *W satisfies $W^T = W$ and $W\mathbf{1} = \mathbf{1}$. Its eigenvalues satisfy $|\lambda_n| \leq \cdots \leq |\lambda_2| < |\lambda_1| = 1$.*

Then, the spectral gap of W can be represented by $1-\lambda$ where $\lambda \triangleq |\lambda_2|$ and $1-\lambda \in (0,1]$.

Algorithm 1 MDBO

Input: $x_0^{(k)} = x_0$, $y_0^{(k)} = y_0$, $\eta > 0$, $\alpha_1 > 0$, $\alpha_2 > 0$, $\beta_1 > 0$, $\beta_2 > 0$.

1: **for** $t = 0, \cdots, T-1$ **do**
2:   **if** $t == 0$ **then**
3:     $U_t = \Delta_t^{\tilde{F}_{\tilde{\xi}_t}}, V_t = \Delta_t^{g_{\zeta_t}}, Z_t^{\tilde{F}} = \Delta_t^{\tilde{F}_{\tilde{\xi}_t}}, Z_t^g = \Delta_t^{g_{\zeta_t}}$,
4:   **else**
5:     $U_t = (1-\alpha_1\eta)U_{t-1} + \alpha_1\eta\Delta_t^{\tilde{F}_{\tilde{\xi}_t}}$,
      $V_t = (1-\alpha_2\eta)V_{t-1} + \alpha_2\eta\Delta_t^{g_{\zeta_t}}$,
6:     $Z_t^{\tilde{F}} = Z_{t-1}^{\tilde{F}} W + U_t - U_{t-1}$,
      $Z_t^g = Z_{t-1}^g W + V_t - V_{t-1}$,
7:   **end if**
8:   $X_{t+1} = X_t - \eta X_t(I-W) - \beta_1\eta Z_t^{\tilde{F}}$,
    $Y_{t+1} = Y_t - \eta Y_t(I-W) - \beta_2\eta Z_t^g$,
9: **end for**

4 Decentralized Stochastic Bilevel Optimization Algorithms

Momentum-based Decentralized Stochastic Bilevel Optimization Algorithm. In Algorithm 1, we developed a momentum-based decentralized stochastic bilevel optimization (MDBO) algorithm. The main idea is to use the momentum to update the model parameters x and y at each participant and then perform communication. Specifically, the momentum is updated as follows:

$$\begin{aligned} U_t &= (1-\alpha_1\eta)U_{t-1} + \alpha_1\eta\Delta_t^{\tilde{F}_{\tilde{\xi}_t}} , \\ V_t &= (1-\alpha_2\eta)V_{t-1} + \alpha_2\eta\Delta_t^{g_{\zeta_t}} \;, \end{aligned} \tag{7}$$

where α_1, α_2, and η are positive, $\alpha_1\eta < 1$, $\alpha_2\eta < 1$, $U_t = [u_t^{(1)}, u_t^{(2)}, \cdots, u_t^{(K)}] \in \mathbb{R}^{d_x\times K}$ is the momentum of the stochastic hypergradient $\Delta_t^{\tilde{F}_{\tilde{\xi}_t}}$, and $V_t = [v_t^{(1)}, v_t^{(2)}, \cdots, v_t^{(K)}] \in \mathbb{R}^{d_y\times K}$ is the momentum of the stochastic gradient $\Delta_t^{g_{\zeta_t}}$. Here, $u_t^{(k)}$ and $v_t^{(k)}$ are the momentum in the k-th participant. Their updates are restricted in the corresponding participant. Then, MDBO employs the gradient tracking communication mechanism to exchange the momentum and model parameter across participants. In detail, $Z_t^{\tilde{F}} \in \mathbb{R}^{d_x\times K}$ and $Z_t^g \in \mathbb{R}^{d_y\times K}$ are the tracked momentum for U_t and V_t, respectively. In the first iteration, they are initialized as the stochastic gradient as shown in Line 3 of Algorithm 1. In other iterations, they are updated as follows:

$$\begin{aligned} Z_t^{\tilde{F}} &= Z_{t-1}^{\tilde{F}} W + U_t - U_{t-1}, \\ Z_t^g &= Z_{t-1}^g W + V_t - V_{t-1} \;, \end{aligned} \tag{8}$$

where $Z_{t-1}^{\tilde{F}} W$ denotes the communication operation. Based on the tracked momentum, the model parameters x and y are updated as follows:

$$\begin{aligned} X_{t+1} &= X_t - \eta X_t(I-W) - \beta_1\eta Z_t^{\tilde{F}} \;, \\ Y_{t+1} &= Y_t - \eta Y_t(I-W) - \beta_2\eta Z_t^g \;, \end{aligned} \tag{9}$$

where $\eta \in (0,1)$, β_1 and β_2 are positive, X_tW and Y_tW indicate the communication of model parameters across participants. In fact, by reformulating this updating rule, it is easy to know that X_{t+1} is the combination of the local model parameter X_t and the update $X_tW - \beta_1 Z_t^{\tilde{F}}$ that is based on the neighboring participants' information. In summary, the computation of stochastic gradients/hypergradients, Hessian matrix, and Jacobian matrix is restricted in each participant. Only the momentum and model parameters are communicated across participants.

Algorithm 2 VRDBO

Input: $x_0^{(k)} = x_0$, $y_0^{(k)} = y_0$, $\eta > 0$, $\alpha_1 > 0$, $\alpha_2 > 0$, $\beta_1 > 0$, $\beta_2 > 0$.

1: **for** $t = 0, \cdots, T-1$ **do**
2:   **if** $t == 0$ **then**
3:     With the mini-batch size B:
    $U_t = \Delta_t^{\tilde{F}_{\tilde{\xi}}}$, $V_t = \Delta_t^{g_\zeta}$, $Z_t^{\tilde{F}} = \Delta_t^{\tilde{F}_{\tilde{\xi}}}$, $Z_t^g = \Delta_t^{g_\zeta}$,
4:   **else**
5:     $U_t = (1-\alpha_1\eta^2)(U_{t-1} + \Delta_t^{\tilde{F}_{\tilde{\xi}_t}} - \Delta_{t-1}^{\tilde{F}_{\tilde{\xi}_t}}) + \alpha_1\eta^2\Delta_t^{\tilde{F}_{\tilde{\xi}_t}}$,
    $V_t = (1-\alpha_2\eta^2)(V_{t-1} + \Delta_t^{g_{\zeta_t}} - \Delta_{t-1}^{g_{\zeta_t}}) + \alpha_2\eta^2\Delta_t^{g_{\zeta_t}}$,
6:     $Z_t^{\tilde{F}} = Z_{t-1}^{\tilde{F}}W + U_t - U_{t-1}$,
    $Z_t^g = Z_{t-1}^g W + V_t - V_{t-1}$,
7:   **end if**
8:   $X_{t+1} = X_t - \eta X_t(I-W) - \beta_1\eta Z_t^{\tilde{F}}$,
  $Y_{t+1} = Y_t - \eta Y_t(I-W) - \beta_2\eta Z_t^g$,
9: **end for**

Variance-Reduction-based Decentralized Stochastic Bilevel Optimization Algorithm. Existing non-parallel algorithms have shown that the momentum-based approach does not achieve a better *theoretical* convergence rate even though it demonstrates better *empirical* convergence performance (Guo et al., 2021; Khanduri et al., 2021a). Thus, we further developed a new algorithm: variance-reduction-based decentralized stochastic bilevel optimization (VRDBO) algorithm, which takes advantage of the variance-reduced gradient estimator to accelerate the convergence rate. The details are shown in Algorithm 2. Specifically, VRDBO utilizes the STORM (Cutkosky and Orabona, 2019) gradient estimator to control the variance of stochastic gradients/hypergradients as follows:

$$U_t = (1-\alpha_1\eta^2)(U_{t-1} + \Delta_t^{\tilde{F}_{\tilde{\xi}_t}} - \Delta_{t-1}^{\tilde{F}_{\tilde{\xi}_t}}) + \alpha_1\eta^2\Delta_t^{\tilde{F}_{\tilde{\xi}_t}} , \tag{10}$$

where $\alpha > 0$, $\eta > 0$, and $\alpha\eta^2 \in (0,1)$. Note that $\Delta_{t-1}^{\tilde{F}_{\tilde{\xi}_t}}$ denotes the stochastic hypergradient which is computed based on the model parameters X_{t-1} and Y_{t-1} in the $(t-1)$-th iteration, as well as the selected samples in the t-th iteration. V_t is updated in the same way. Then, based on this variance-reduced gradient estimator, each participant leverages the gradient tracking communication mechanism to exchange the tracked gradients and model parameters to update local model parameters, which is shown in Lines 6 and 8 of Algorithm 2.

In summary, VRDBO utilizes a variance-reduced gradient estimator to control the variance of stochastic gradients. Thus, it can achieve a better convergence rate than MDBO, which will be shown in next section. Additionally, both MDBO and VRDBO do NOT require to communicate Hessian and Jacobian matrices. Only model parameters and tracked gradients are communicated. To the best of our knowledge, we are the first one developing the variance-reduced decentralized bilevel optimization algorithm.

5 Convergence Analysis

To investigate the convergence rate of our two algorithms, we first introduce two common assumptions for both algorithms and then introduce algorithm-specific assumptions.

Assumption 2. *For any fixed $x \in \mathbb{R}^{d_x}$ and $k \in \{1, 2, \cdots, K\}$, the lower-level function $g^{(k)}(x, y)$ is μ-strongly convex with respect to y.*

Assumption 3. *For any $k \in \{1, 2, \cdots, K\}$, the first and second order stochastic gradients of all loss functions have bounded variance σ^2 where $\sigma > 0$.*

5.1 Convergence Rate of Algorithm 1

Similar to the non-parallel algorithms (Ghadimi and Wang, 2018; Chen et al., 2021; Hong et al., 2020; Ji et al., 2021), our Algorithm 1 requires a weaker assumption regarding the smoothness of the loss function compared with Algorithm 2, which is shown as follows.

Assumption 4. *For any $k \in \{1, 2, \cdots, K\}$, $\nabla_x f^{(k)}(x, y)$ is Lipschitz continuous with the constant $L_{f_x} > 0$, $\nabla_y f^{(k)}(x, y)$ is Lipschitz continuous with the constant $L_{f_y} > 0$. Moreover, $\|\nabla_y f^{(k)}(x, y)\| \leq C_{f_y}$ with the constant $C_{f_y} > 0$ for $(x, y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$.*

Assumption 5. *For any $k \in \{1, 2, \cdots, K\}$, $\nabla_y g^{(k)}(x, y)$ is Lipschitz continuous with the constant $L_{g_y} > 0$, $\nabla^2_{xy} g^{(k)}(x, y)$ is Lipschitz continuous with the constant $L_{g_{xy}} > 0$, $\nabla^2_{yy} g^{(k)}(x, y)$ is Lipschitz continuous with the constant $L_{g_{yy}} > 0$. Moreover, $\|\nabla^2_{xy} g^{(k)}(x, y)\| \leq C_{g_{xy}}$ with the constant $C_{g_{xy}} > 0$ and $\mu\mathbf{1} \preceq \nabla^2_{yy} g^{(k)}(x, y; \zeta) \preceq L_{g_y}\mathbf{1}$ for $(x, y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$.*

Based on these assumptions, we are able to establish the convergence rate of Algorithm 1 as follows.

Theorem 1. *Given Assumptions 1-5, if $\alpha_1 > 0$, $\alpha_2 > 0$, $\eta < \min\{1, \frac{1}{2\beta_1 L_F^*}, \frac{1}{\alpha_1}, \frac{1}{\alpha_2}\}$, $\beta_1 \leq \min\{\beta_{1,a}, \beta_{1,b}, \beta_{1,c}\}$,* *and* $\beta_2 \leq \min\{\beta_{2,a}, \beta_{2,b}, \beta_{2,c}\}$*, where*

$$\beta_{1,a} = \frac{\beta_2\mu}{15L_yL_F} \, , \beta_{2,c} = \frac{1}{6L_{g_y}} \, ,$$
$$\beta_{1,b} = \frac{\mu}{4L_{g_y}\sqrt{((2+8/\alpha_1^2)L_{\tilde{F}}^2 + (100+200/\alpha_2^2)L_F^2)}} \, ,$$
$$\beta_{1,c} = \frac{\mu(1-\lambda)^2}{4L_{g_y}\sqrt{(6+18/\alpha_1^2)L_{\tilde{F}}^2 + (250+450/\alpha_2^2)L_F^2}} \, ,$$
$$\beta_{2,a} = \frac{9\mu L_F^2}{2((4+16/\alpha_1^2)L_{\tilde{F}}^2 + (200+400/\alpha_2^2)L_F^2)L_{g_y}^2} \, ,$$
$$\beta_{2,b} = \frac{5(1-\lambda)^2 L_F}{2L_{g_y}\sqrt{(12+36/\alpha_1^2)L_{\tilde{F}}^2 + (500+900/\alpha_2^2)L_F^2}} \, , \tag{11}$$

the convergence rate of Algorithm 1 is

$$\begin{aligned} &\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] + L_F^2\mathbb{E}[\|\bar{y}_t - y^*(\bar{x}_t)\|^2]) \\ &\leq \frac{2(F(x_0)-F(x_*))}{\eta\beta_1 T} + \frac{12L_F^2}{\beta_2\mu\eta T}\|\bar{y}_0 - y^*(\bar{x}_0)\|^2 \\ &\quad + \frac{6C_{g_{xy}}^2C_{f_y}^2}{\mu^2}(1-\frac{\mu}{L_{g_y}})^{2J} + \frac{250\alpha_2\eta L_F^2\sigma^2}{\mu^2} \\ &\quad + 9\alpha_1\eta\sigma_{\tilde{F}}^2 + \frac{10\sigma_{\tilde{F}}^2}{\alpha_1\eta T} + \frac{300L_F^2\sigma^2}{\alpha_2\mu^2\eta T} \, , \end{aligned} \tag{12}$$

where the definition of L_y*,* L_F^**,* L_F*,* $L_{\tilde{F}}$*,* $\sigma_{\tilde{F}}$ *is shown in Lemmas 1, 4, 5.*

Corollary 1. *Given the same condition with Theorem 1, by choosing* $T = O(\frac{1}{\epsilon^2(1-\lambda)^2})$*,* $\eta = O(\epsilon)$*,* $J = O(\log\frac{1}{\epsilon})$*,* $\beta_1 = O((1-\lambda)^2)$*,* $\beta_2 = O((1-\lambda)^2)$*,* $\alpha_1 = O(1)$*, and* $\alpha_2 = O(1)$*, Algorithm 1 can achieve the* ϵ*-accuracy solution:* $\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2 + L_F^2\|\bar{y}_t - y^*(\bar{x}_t)\|^2]) \leq O(\epsilon)$*. Then, the communication complexity is* $O(\frac{1}{\epsilon^2(1-\lambda)^2})$*, the gradient complexity and Jacobian-vector product complexity is* $O(\frac{1}{\epsilon^2(1-\lambda)^2})$*, and the Hessian-vector product complexity is* $\widetilde{O}(\frac{1}{\epsilon^2(1-\lambda)^2})$*.*

Discussion. The concurrent work (Chen et al., 2022) developed a decentralized stochastic bilevel optimization algorithm based on the vanilla stochastic gradient (See Algorithm 5 in (Chen et al., 2022)). Another concurrent work (Yang et al., 2022) developed a momentum-based decentralized stochastic bilevel optimization algorithm, which employs the gossip communication mechanism and requires to communicate stochastic gradients/hypergradients, Hessian matrix, and Jacobian matrix. Obviously, on the algorithmic design side, our Algorithm 1 is significantly different from those two works. Moreover, on the theoretical analysis side, those two works have much stronger assumptions. In particular, their theoretical analyses require that the second-order moments of (stochastic) hypergradient regarding x of the upper-level function and the (stochastic) gradient regarding y of the low-level function are upper bounded, which is shown in Assumption 6.

Assumption 6. *For any* $k \in \{1, 2, \cdots, K\}$*,* $(x, y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$*,* $\nabla_x f^{(k)}(x, y)$ *and* $\nabla_y g^{(k)}(x, y)$ *satisfies:*

$$\|\nabla_x f^{(k)}(x, y)\| \leq C_{f_x} \, , \, \|\nabla_y g^{(k)}(x, y)\| \leq C_{g_y} \, , \tag{13}$$

where the constant $C_{f_x} > 0$ *and* $C_{g_y} > 0$*.*

With this additional assumption, the convergence rate in (Yang et al., 2022) is able to achieve linear speedup with respect to the number of participants, while that in (Chen et al., 2022) fails to achieve linear speedup even with this strong assumption (See Lemma A.35 and Theorem 3.3 in (Chen et al., 2022)).

In the following, we show that our Algorithm 1 can also achieve the linear speedup effect as (Yang et al., 2022) with this additional assumption.

Theorem 2. *Given Assumptions 1-5 and* ***Assumption 6****, if* $\alpha_1 > 0$*,* $\alpha_2 > 0$*,* $\eta < \min\{1, \frac{1}{2\beta_1 L_F^*}, \frac{1}{\alpha_1}, \frac{1}{\alpha_2}\}$*,* $\beta_1 \leq \min\{\frac{\beta_2\mu}{15L_yL_F}, \frac{\mu}{4L_{g_y}\sqrt{6L_{\tilde{F}}^2/\alpha_1^2 + 100L_F^2/\alpha_2^2}}\}$*, and* $\beta_2 \leq \min\{\frac{1}{6L_{g_y}}, \frac{9\mu L_F^2}{2(12L_{\tilde{F}}^2/\alpha_1^2 + 200L_F^2/\alpha_2^2)L_{g_y}^2}\}$*, the convergence rate of Algorithm 1 is*

$$\begin{aligned} &\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] + L_F^2\mathbb{E}[\|\bar{y}_t - y^*(\bar{x}_t)\|^2]) \\ &\leq \frac{2(F(x_0)-F(x_*))}{\eta\beta_1 T} + \frac{12L_F^2}{\beta_2\mu\eta T}\|\bar{y}_0 - y^*(\bar{x}_0)\|^2 \\ &\quad + \frac{6C_{g_{xy}}^2C_{f_y}^2}{\mu^2}(1-\frac{\mu}{L_{g_y}})^{2J} + \frac{24\hat{C}_{\tilde{F}}^2}{\alpha_1\eta T} + \frac{400L_F^2\hat{C}_{g_y}^2}{\alpha_2\mu^2\eta T} \\ &\quad + \frac{48\alpha_1^2\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2L_{\tilde{F}}^2}{(1-\lambda)^4} + \frac{48\alpha_2^2\beta_2^2\eta^2\hat{C}_{g_y}^2L_{\tilde{F}}^2}{(1-\lambda)^4} \\ &\quad + \frac{800\alpha_1^2\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2L_{g_y}^2L_F^2}{\mu^2(1-\lambda)^4} + \frac{800\alpha_2^2\beta_2^2\eta^2\hat{C}_{g_y}^2L_F^2L_{g_y}^2}{\mu^2(1-\lambda)^4} \\ &\quad + \frac{432\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2L_{\tilde{F}}^2}{(1-\lambda)^4} + \frac{432\alpha_2^2\beta_2^2\eta^2\hat{C}_{g_y}^2L_{\tilde{F}}^2}{\alpha_1^2(1-\lambda)^4} \\ &\quad + \frac{7200\alpha_1^2\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2L_F^2L_{g_y}^2}{\alpha_2^2\mu^2(1-\lambda)^4} + \frac{7200\beta_2^2\eta^2\hat{C}_{g_y}^2L_F^2L_{g_y}^2}{\mu^2(1-\lambda)^4} \\ &\quad + \frac{6\alpha_1\eta\sigma_{\tilde{F}}^2}{K} + \frac{100\alpha_2\eta\sigma^2L_F^2}{\mu^2K} \, , \end{aligned} \tag{14}$$

where the definition of L_y*,* L_F^**,* L_F*,* $L_{\tilde{F}}$*,* $\sigma_{\tilde{F}}$*,* $\hat{C}_{\tilde{F}}$*,* $\hat{C}_{g_y}$ *is shown in Lemmas 1, 4, 5, 14.*

Corollary 2. *Given the same condition with Theorem 2, by choosing* $T = O(\frac{1}{K\epsilon^2})$*,* $\eta = O(K\epsilon)$*,* $J = O(\log\frac{1}{\epsilon})$*,* $\beta_1 = O(1)$*,* $\beta_2 = O(1)$*,* $\alpha_1 = O(1)$*,* $\alpha_2 = O(1)$*, Algorithm 1 can achieve the* ϵ*-accuracy solution. Then, the communication complexity is* $O(\frac{1}{K\epsilon^2})$*, the gradient complexity and*

Jacobian-vector product complexity is $O(\frac{1}{K\epsilon^2})$*, and the Hessian-vector product complexity is* $\tilde{O}(\frac{1}{K\epsilon^2})$*, which indicates the linear speedup regarding* K*.*

Remark 1. *Due to the additional assumption 6, it is much easier to bound the consensus error. For instance, to bound the consensus error* $\mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}}\|_F^2]$*, it is easy to get* $\frac{1}{K}\mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}}\|_F^2] \leq \frac{2\alpha_1^2\eta^2\hat{C}_{\tilde{F}}^2}{(1-\lambda)^2}$ *(See Lemma 15) since the second moment of the stochastic hypergradient is upper bounded. In fact, with this strong assumption, we can decouple different consensus errors to simplify the theoretical analysis. On the contrary, without Assumption 6, there exists inter-dependence between different consensus errors (See Lemmas 6, 8, 12), which makes it much more challenging to establish the convergence rate.*

5.2 Convergence Rate of Algorithm 2

Since Algorithm 2 employs the variance-reduced gradient estimator, we introduce the following mean-square Lipschitz smoothness assumption for the upper-level and lower-level objective functions, which is also used by existing variance-reduced bilevel optimization algorithms (Yang et al., 2021; Guo and Yang, 2021). Please note that all variance-reduced gradient descent algorithms (Cutkosky and Orabona, 2019; Fang et al., 2018) require the mean-square Lipschitz smoothness assumption to establish the convergence rate. In the following, we use z_i to denote (x_i, y_i) where $i \in \{1, 2\}$.

Assumption 7. *For any* $k \in \{1, 2, \cdots, K\}$, $\nabla_x f^{(k)}(x, y)$ *is Lipschitz continuous with the constant* $\ell_{f_x} > 0$, $\nabla_y f^{(k)}(x, y)$ *is Lipschitz continuous with the constant* $\ell_{f_y} > 0$*, i.e.,*

$$\mathbb{E}[\|\nabla_x f^{(k)}(z_1;\xi) - \nabla_x f^{(k)}(z_2;\xi)\|] \leq \ell_{f_x}\|z_1 - z_2\| ,$$
$$\mathbb{E}[\|\nabla_y f^{(k)}(z_1;\xi) - \nabla_y f^{(k)}(z_2;\xi)\|] \leq \ell_{f_y}\|z_1 - z_2\| , \tag{15}$$

hold for any $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$*. Moreover,* $\mathbb{E}[\|\nabla_y f^{(k)}(x, y; \xi)\|] \leq c_{f_y}$ *with the constant* $c_{f_y} > 0$ *for* $(x, y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$*.*

Assumption 8. *For any* $k \in \{1, 2, \cdots, K\}$, $\nabla_y g^{(k)}(x, y)$ *is Lipschitz continuous with the constant* $\ell_{g_y} > 0$, $\nabla^2_{xy} g^{(k)}(x, y)$ *is Lipschitz continuous with the constant* $\ell_{g_{xy}} > 0$, $\nabla^2_{yy} g^{(k)}(x, y)$ *is Lipschitz continuous with the constant* $\ell_{g_{yy}} > 0$*, i.e.,*

$$\mathbb{E}[\|\nabla_y g^{(k)}(z_1;\zeta) - \nabla_y g^{(k)}(z_2;\zeta)\|] \leq \ell_{g_y}\|z_1 - z_2\| ,$$
$$\mathbb{E}[\|\nabla^2_{xy} g^{(k)}(z_1;\zeta) - \nabla^2_{xy} g^{(k)}(z_2;\zeta)\|] \leq \ell_{g_{xy}}\|z_1 - z_2\| ,$$
$$\mathbb{E}[\|\nabla^2_{yy} g^{(k)}(z_1;\zeta) - \nabla^2_{yy} g^{(k)}(z_2;\zeta)\|] \leq \ell_{g_{yy}}\|z_1 - z_2\| , \tag{16}$$

hold for any $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$*. Moreover,* $\mathbb{E}[\|\nabla^2_{xy} g^{(k)}(x, y; \zeta)\|] \leq c_{g_{xy}}$ *with the constant* $c_{g_{xy}} > 0$ *and* $\mu\mathbf{1} \preceq \nabla^2_{yy} g^{(k)}(x, y; \zeta) \preceq \ell_{g_y}\mathbf{1}$ *for* $(x, y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$*.*

Theorem 3. *Given Assumptions 1-3, 7, 8, if* $\alpha_1 > 0$, $\alpha_2 > 0$, $\eta < \min\{1, \frac{1}{2\beta_1 L_F^*}, \frac{1}{\sqrt{\alpha_1}}, \frac{1}{\sqrt{\alpha_2}}\}$, $\beta_1 \leq \min\{\beta_{1,a}, \beta_{1,b}, \beta_{1,c}\}$ *and* $\beta_2 \leq \min\{\beta_{2,a}, \beta_{2,b}, \beta_{2,c}\}$*,*

$$\begin{aligned}
\beta_{1,a} &= \frac{\beta_2\mu}{15L_yL_F} , \beta_{2,c} = \frac{1}{6\ell_{g_y}} , \\
\beta_{1,b} &= \frac{\mu}{8\ell_{g_y}\sqrt{(3 + 3/(\alpha_1 K))L_{\tilde{F}}^2 + (3 + 50/(\alpha_2 K))L_F^2}} , \\
\beta_{1,c} &= \frac{\mu(1-\lambda)^2/\ell_{g_y}}{2\sqrt{(57 + 54/(\alpha_1 K))L_{\tilde{F}}^2 + (104 + 900/(\alpha_2 K))L_F^2}} , \\
\beta_{2,a} &= \frac{(1-\lambda)^2 L_F/\ell_{g_y}}{2\sqrt{(57 + 54/(\alpha_1 K))L_{\tilde{F}}^2 + (104 + 900/(\alpha_2 K))L_F^2}} , \\
\beta_{2,b} &= \frac{9\mu L_F^2}{8\ell_{g_y}^2((6 + 6/(\alpha_1 K))L_{\tilde{F}}^2 + (6 + 100/(\alpha_2 K))L_F^2)} ,
\end{aligned} \tag{17}$$

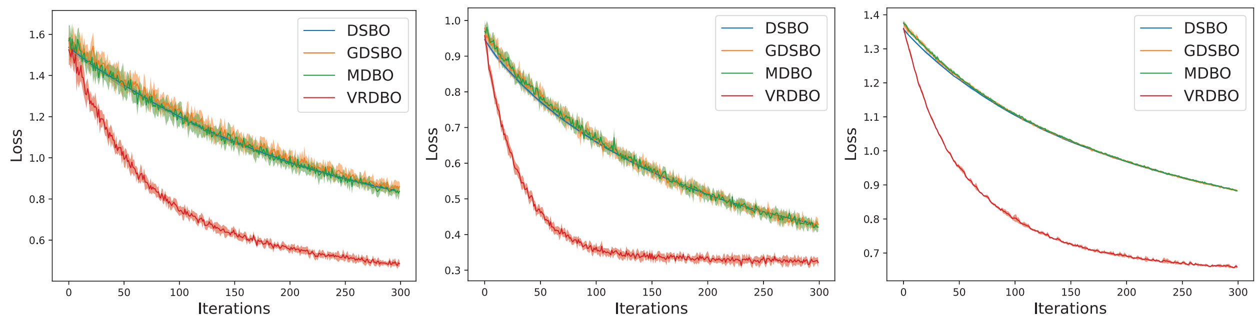
the convergence rate of Algorithm 2 is

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2 + L_F^2\|\bar{y}_t - y^*(\bar{x}_t)\|^2]) \\
&\leq \frac{2(F(x_0) - F(x_*))}{\beta_1\eta T} + \frac{12L_F^2}{\mu\beta_2\eta T}\|\bar{y}_0 - y^*(\bar{x}_0)\|^2 \\
&\quad + \frac{6C_{g_{xy}}^2 C_{f_y}^2}{\mu^2}(1 - \frac{\mu}{\ell_{g_y}})^{2J} + \frac{8\sigma_{\tilde{F}}^2}{\eta TB} + \frac{8L_F^2\sigma^2}{\eta TB\mu^2} \\
&\quad + \frac{6\sigma_{\tilde{F}}^2}{\alpha_1\eta^2 TBK} + \frac{100L_F^2\sigma^2}{\alpha_2\eta^2 TBK\mu^2} + 2\alpha_1^2\eta^3\sigma_{\tilde{F}}^2 \\
&\quad + \frac{2\alpha_2^2\eta^3\sigma^2 L_F^2}{\mu^2} + 8\alpha_1^2\eta^2\sigma_{\tilde{F}}^2 + \frac{8\alpha_2^2\eta^2\sigma^2 L_F^2}{\mu^2} \\
&\quad + \frac{12\alpha_1\eta^2\sigma_{\tilde{F}}^2}{K} + \frac{200\alpha_2\eta^2\sigma^2 L_F^2}{\mu^2 K} ,
\end{aligned} \tag{18}$$

where the definition of L_y, L_F^*, L_F, $L_{\tilde{F}}$, $\sigma_{\tilde{F}}$ *is shown in Lemmas 23, 26, 27.*

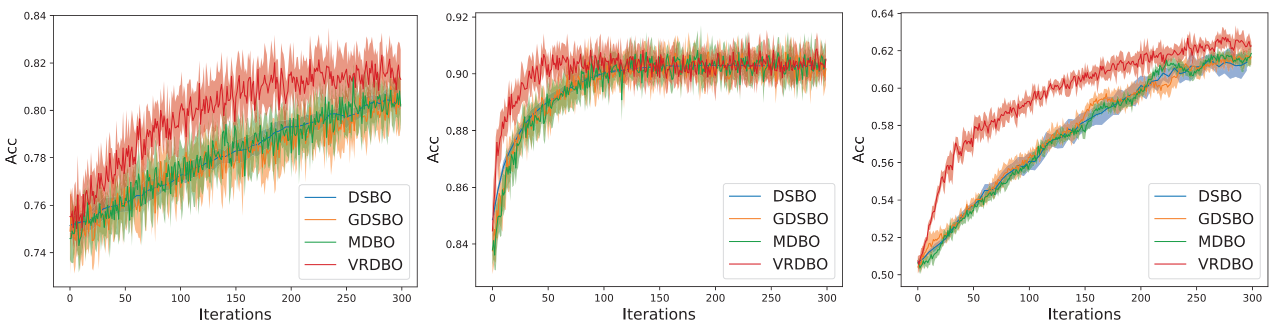
Corollary 3. *Given the same condition with Theorem 3, by choosing* $T = O(\frac{1}{K\epsilon^{3/2}(1-\lambda)^2})$, $\eta = O(K\epsilon^{1/2})$, $J = O(\log\frac{1}{\epsilon})$, $B = O(\frac{1}{\epsilon^{1/2}})$, $\beta_1 = O((1-\lambda)^2)$, $\beta_2 = O((1-\lambda)^2)$, $\alpha_1 = O(1/K)$*, and* $\alpha_2 = O(1/K)$*, Algorithm 2 can achieve the* ϵ*-accuracy solution. Then, the communication complexity is* $O(\frac{1}{K\epsilon^{3/2}(1-\lambda)^2})$*, which is better than* $O(\frac{1}{\epsilon^2(1-\lambda)^2})$ *of Algorithm 1. Additionally, the gradient complexity and Jacobian-vector product complexity of Algorithm 2 are* $O(\frac{1}{K\epsilon^{3/2}(1-\lambda)^2})$ *and the Hessian-vector product complexity is* $\tilde{O}(\frac{1}{K\epsilon^{3/2}(1-\lambda)^2})$*, indicating the linear speedup with respect to the number of participants* K*.*

Remark 2. *It is worth noting that we are able to set* $\alpha_i = O(1/K)$ *in Theorem 3. As a result,* β_i *and* α_i *are decoupled (See Eq. (17)). In this way, Algorithm 2 can achieve linear speedup with respect to the number of participants. In comparison,* α_i *has to be* $O(1)$ *in Theorem 1.*



(a) a9a (b) ijcnn1 (c) covtype

Figure 1: The upper-level training loss function value versus the update of variables.



(a) a9a (b) ijcnn1 (c) covtype

Figure 2: The prediction accuracy of validation set versus the update of variables.

6 Experiments

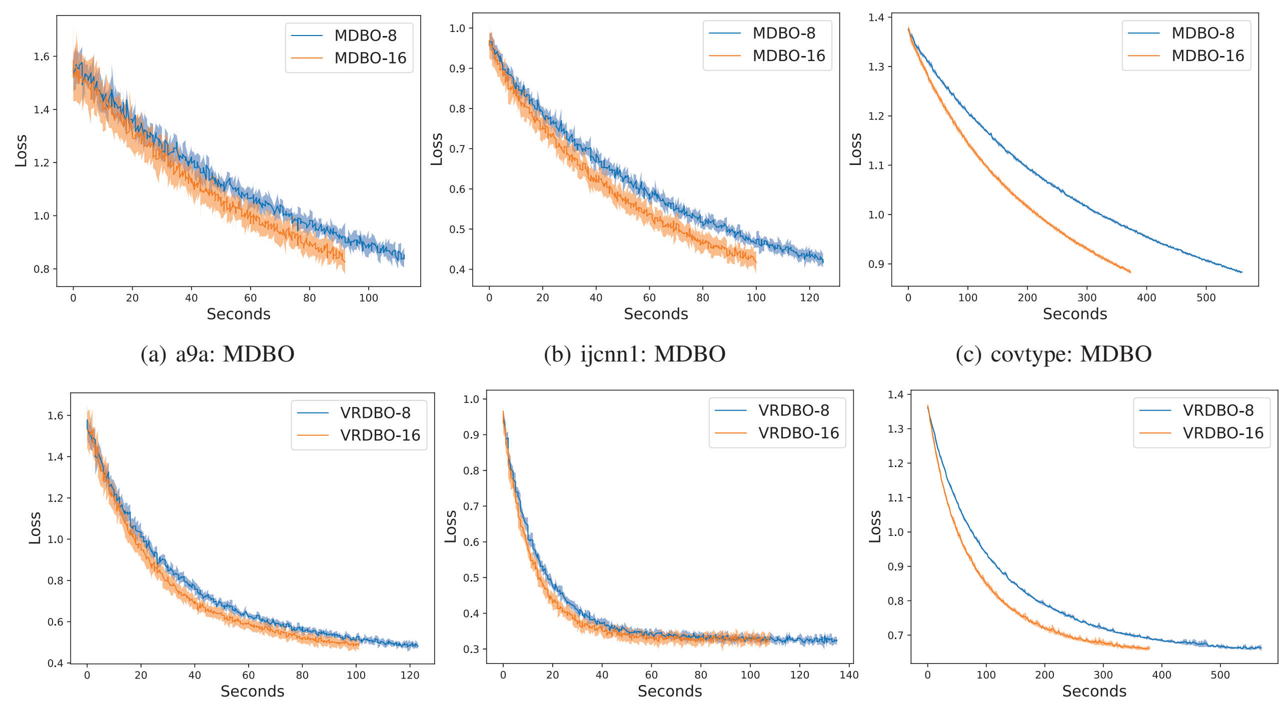
In this section, we conduct experiments to verify the performance of our proposed algorithms. In particular, we apply our algorithms to the hyperparameter optimization of the logistic regression model (Grazzi et al., 2020), which is defined as follows:

$$\begin{aligned} &\min_{x\in\mathbb{R}^d} \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_{val}^{(k)}}\sum_{i=1}^{n_{val}^{(k)}}\ell_{CE}(y^*(x)^T a_{val,i}^{(k)}, b_{val,i}^{(k)}) \\ &s.t.\ y^*(x) = \arg\min_{y\in\mathbb{R}^{d\times c}} \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_{tr}^{(k)}}\sum_{i=1}^{n_{tr}^{(k)}}\ell_{CE}(y^T a_{tr,i}^{(k)}, b_{tr,i}^{(k)}) \\ &\qquad + \frac{1}{cd}\sum_{p=1}^{c}\sum_{q=1}^{d}\exp(x_q)y_{pq}^2\ , \end{aligned} \tag{19}$$

where $(a_{val,i}^{(k)}, b_{val,i}^{(k)}) \in \mathbb{R}^d \times \mathbb{R}^c$ denotes the i-th validation sample's feature and label of the k-th participant, $(a_{tr,i}^{(k)}, b_{tr,i}^{(k)})$ represents the training sample, $n_{val}^{(k)}$ is the number of validation samples in the k-th participant, $n_{tr}^{(k)}$ is the number of training samples, ℓ_{CE} is the cross-entropy loss function, $x \in \mathbb{R}^d$ represents the hyperparameter, $y \in \mathbb{R}^{d\times c}$ denotes the model parameter.

In our experiments, we use three binary classification datasets [1]: a9a, ijcnn1, and covtype. In particular, a9a has 32,561 sample, ijcnn1 has 49,990 samples, and covtype has 581,012 samples. We randomly select 30% samples as the validation set and the remaining samples as the training set. Then, they are randomly and evenly put to each participant so that the data distribution is i.i.d. in all participants. To demonstrate the performance of our algorithms, we compare them with two baseline algorithms: DSBO (Chen et al., 2022) and GDSBO (Yang et al., 2022). Specifically, DSBO employs stochastic (hyper) gradient and gossip communication strategy, and GDSBO takes the momentum technique and gossip communication strategy. Note that They require to explicitly communicate Hessian or Jacobian matrices, which is prohibitive for practical applications. Thus, we implement a simplified version, where Hessian and Jacobian matrices are implicitly computed as (Yang et al., 2021) and only model parameters (and gradient estimators) are communicated via the gossip communication strategy. In our experiments, we use the same batch size for all algorithms. In particular, the batch size on each participant is $400/K$ where K is the total number of participants. When estimating the stochastic hypergradient, J

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/



(a) a9a: MDBO (b) ijcnn1: MDBO (c) covtype: MDBO

(d) a9a: VRDBO (e) ijcnn1: VRDBO (f) covtype: VRDBO

Figure 3: The upper-level loss function value versus the consumed time.

is set to 10 for all algorithms. Moreover, since the learning rate of DSBO, GDSBO, MDBO is in the order of $O(\epsilon)$ and that of VRDBO is $O(\epsilon^{1/2})$, we set the learning rate $\eta = 0.1$ for DSBO, GDSBO, MDBO and $\eta = 0.33$ for VRDBO. Additionally, we set $\beta_1 = \beta_2 = 1.0$ for both MDBO and VRDBO, $\alpha_1 = \alpha_2 = 1.0$ for MDBO, and $\alpha_1 = \alpha_2 = 5.0$ for VRDBO.

In Figure 1, we employ 8 workers and the network is a ring network. Here, we show the upper-level loss function value with respect to the update of variables. There are two observations. First, our MDBO has similar convergence performance as GDSBO based on both loss function value and prediction accuracy, which uses the same momentum technique as ours. However, they don't show significant improvement over DSBO. This is consistent with the theoretical convergence rate. Second, VRDBO converges much faster than MDBO and baseline methods, since it employs a variance-reduced gradient estimator. This observation confirms the correctness of our theoretical convergence rate and the effectiveness of our algorithm.

In Figure 2, we plot the prediction accuracy of validation set with respect to the update of variables. The experimental settings are the same with those of Figure 1. From Figure 2, we have two observations. First, our algorithms, MDBO and VRDBO, can achieve almost the same prediction accuracy with baseline algorithms, which confirms the correctness of our algorithms. Second, our VRDBO converges faster than MDBO in terms of the prediction accuracy, which is consistent with our theoretical convergence rates. In summary, Figure 1 and Figure 2 confirm the correctness and effectiveness of our two algorithms.

In Figure 3, we plot the upper-level loss function value with respect to the consumed time (seconds) for our MDBO and VRDBO algorithms. Here, to demonstrate the speedup effect, we use 8, 16 workers, respectively. The batch size of each worker is set to $400/K$ where K is the number of workers. Other experimental settings are the same with those of Figure 1. From Figure 3, we can find that using more workers is able to accelerate the practical convergence speed of our two algorithms.

7 Conclusions

In this paper, we studied how to facilitate bilevel optimization to the decentralized setting. In particular, we developed two decentralized bilevel optimization algorithms, which demonstrate how to update variables on each participant and communicate variables across participants. In addition, we established the convergence rate, demonstrating how the network topology, number of participants, and other hyperparameters affect the convergence rate. To our knowledge, this is the first work achieving these favorable results. Moreover, extensive experimental results confirm the correctness and effectiveness of our algorithms.

Acknowledgements

This material is partially based upon work supported by the National Science Foundation under grants CNS-1814614, CNS-2140477, and IIS-1908594.

References

T. Chen, Y. Sun, and W. Yin. Tighter analysis of alternating stochastic gradient method for stochastic nested problems. *arXiv preprint arXiv:2106.13781*, 2021.

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

A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

C. Fang, C. J. Li, Z. Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

M. Feurer and F. Hutter. Hyperparameter optimization. In *Automated machine learning*, pages 3–33. Springer, Cham, 2019.

L. Franceschi, M. Donini, P. Frasconi, and M. Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning*, pages 1165–1173. PMLR, 2017.

L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pages 1568–1577. PMLR, 2018.

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

H. Gao. Decentralized stochastic gradient descent ascent for finite-sum minimax problems. *arXiv preprint arXiv:2212.02724*, 2022b.

H. Gao and H. Huang. Periodic stochastic gradient descent with momentum for decentralized training. *arXiv preprint arXiv:2008.10435*, 2020.

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

H. Gao, H. Xu, and S. Vucetic. Sample efficient decentralized stochastic frank-wolfe methods for continuous dr-submodular maximization. In Z. Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3501–3507, 2021.

H. Gao, M. T. Thai, and J. Wu. When decentralized optimization meets federated learning. *IEEE Network*, 2023.

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

R. Grazzi, L. Franceschi, M. Pontil, and S. Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pages 3748–3758. PMLR, 2020.

Z. Guo and T. Yang. Randomized stochastic variance-reduced methods for stochastic bilevel optimization. *arXiv e-prints*, pages arXiv–2105, 2021.

Z. Guo, Y. Xu, W. Yin, R. Jin, and T. Yang. On stochastic moving-average estimators for non-convex optimization. *arXiv preprint arXiv:2104.14840*, 2021.

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

K. Ji, J. Yang, and Y. Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pages 4882–4892. PMLR, 2021.

P. Khanduri, S. Zeng, M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A momentum-assisted single-timescale stochastic approximation algorithm for bilevel optimization. *arXiv e-prints*, pages arXiv–2102, 2021a.

P. Khanduri, S. Zeng, M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in Neural Information Processing Systems*, 34, 2021b.

A. Koloskova, S. Stich, and M. Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International Conference on Machine Learning*, pages 3478–3487. PMLR, 2019.

X. Li, W. Yang, S. Wang, and Z. Zhang. Communication efficient decentralized training with multiple local updates. *arXiv preprint arXiv:1910.09126*, 5, 2019.

Z. Li, S. Hanzely, and P. Richtárik. Zerosarah: Efficient nonconvex finite-sum optimization with zero full gradient computation. *arXiv preprint arXiv:2103.01447*, 2021.

X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

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

S. Lu, X. Zhang, H. Sun, and M. Hong. Gnsd: A gradient-tracking based nonconvex stochastic algorithm for decentralized optimization. In *2019 IEEE Data Science Workshop (DSW)*, pages 315–321. IEEE, 2019.

A. Mokhtari, H. Hassani, and A. Karbasi. Decentralized submodular maximization: Bridging discrete and continuous settings. In *International Conference on Machine Learning*, pages 3616–3625. PMLR, 2018.

S. Pu and A. Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187(1):409–457, 2021.

A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

H. Sun, S. Lu, and M. Hong. Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking. In *International conference on machine learning*, pages 9217–9228. PMLR, 2020.

H. Tang, X. Lian, S. Qiu, L. Yuan, C. Zhang, T. Zhang, and J. Liu. Deepsqueeze: Decentralization meets error-compensated compression. *arXiv preprint arXiv:1907.07346*, 2019.

I. Tsaknakis, M. Hong, and S. Liu. Decentralized min-max optimization: Formulations, algorithms and applications in network poisoning attack. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5755–5759. IEEE, 2020.

T. Vogels, S. P. Karimireddy, and M. Jaggi. Powergossip: Practical low-rank communication compression in decentralized deep learning. *arXiv preprint arXiv:2008.01425*, 2020.

H.-T. Wai, J. Lafond, A. Scaglione, and E. Moulines. Decentralized frank–wolfe algorithm for convex and non-convex problems. *IEEE Transactions on Automatic Control*, 62(11):5522–5537, 2017.

W. Xian, F. Huang, Y. Zhang, and H. Huang. A faster decentralized algorithm for nonconvex minimax problems. *Advances in Neural Information Processing Systems*, 34: 25865–25877, 2021.

R. Xin, U. Khan, and S. Kar. A hybrid variance-reduced method for decentralized stochastic non-convex optimization. In *International Conference on Machine Learning*, pages 11459–11469. PMLR, 2021.

J. Yang, K. Ji, and Y. Liang. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

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

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

W. Zhan, G. Wu, and H. Gao. Efficient decentralized stochastic gradient descent method for nonconvex finite-sum optimization problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9006–9013, 2022.

X. Zhang, J. Liu, Z. Zhu, and E. S. Bentley. Low sample and communication complexities in decentralized learning: A triple hybrid approach. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2021a.

X. Zhang, Z. Liu, J. Liu, Z. Zhu, and S. Lu. Taming communication and sample complexities in decentralized policy evaluation for cooperative multi-agent reinforcement learning. In *NeurIPS*, pages 18825–18838, 2021b.

Supplementary Materials

A Proof

To investigate the convergence rate of our algorithms, other than the notations in Section 3, we introduce the following additional notations:

$$\begin{aligned}
&\bar{u}_t = \frac{1}{K}\sum_{k=1}^{K} u_t^{(k)}, \bar{v}_t = \frac{1}{K}\sum_{k=1}^{K} v_t^{(k)}, \bar{z}_t = \frac{1}{K}\sum_{k=1}^{K} z_t^{(k)}, \\
&\bar{X}_t = \frac{1}{K}X_t\mathbf{1}\mathbf{1}^T, \bar{Y}_t = \frac{1}{K}Y_t\mathbf{1}\mathbf{1}^T, \bar{U}_t = \frac{1}{K}U_t\mathbf{1}\mathbf{1}^T, \bar{V}_t = \frac{1}{K}V_t\mathbf{1}\mathbf{1}^T, \bar{Z}_t^{\tilde{F}} = \frac{1}{K}Z_t^{\tilde{F}}\mathbf{1}\mathbf{1}^T, \bar{Z}_t^{g} = \frac{1}{K}Z_t^{g}\mathbf{1}\mathbf{1}^T, \\
&\Delta_t^{\tilde{F}} = [\nabla \tilde{F}^{(1)}(x_t^{(1)}, y_t^{(1)}), \nabla \tilde{F}^{(2)}(x_t^{(2)}, y_t^{(2)}), \cdots, \nabla \tilde{F}^{(K)}(x_t^{(K)}, y_t^{(K)})], \\
&\Delta_t^{g} = [\nabla_y g^{(1)}(x_t^{(1)}, y_t^{(1)}), \nabla_y g^{(2)}(x_t^{(2)}, y_t^{(2)}), \cdots, \nabla_y g^{(K)}(x_t^{(K)}, y_t^{(K)})], \\
&\underline{\Delta}_t^{\tilde{F}} = [\nabla \tilde{F}^{(1)}(\bar{x}_t, \bar{y}_t), \nabla \tilde{F}^{(2)}(\bar{x}_t, \bar{y}_t), \cdots, \nabla \tilde{F}^{(K)}(\bar{x}_t, \bar{y}_t)], \\
&\underline{\Delta}_t^{g} = [\nabla_y g^{(1)}(\bar{x}_t, \bar{y}_t), \nabla_y g^{(2)}(\bar{x}_t, \bar{y}_t), \cdots, \nabla_y g^{(K)}(\bar{x}_t, \bar{y}_t)].
\end{aligned} \tag{20}$$

A.1 Proof Sketch

Proof Sketch of Theorem 1. Since there exists inter-dependence between different consensus errors, we proposed a novel potential function to establish the convergence rate of Algorithm 1, which is shown as follows:

$$\begin{aligned}
\mathcal{L}_{t+1} &= \mathbb{E}[F(x_{t+1})] + \frac{6\beta_1 L_F^2}{\beta_2\mu}\mathbb{E}[\|\bar{y}_{t+1} - y^*(\bar{x}_{t+1})\|^2] + w_2\frac{1}{K}\mathbb{E}[\|X_{t+1} - \bar{X}_{t+1}\|_F^2] + w_3\frac{1}{K}\mathbb{E}[\|Y_{t+1} - \bar{Y}_{t+1}\|_F^2] \\
&+ \frac{\beta_1(1-\lambda)}{2\alpha_1}\frac{1}{K}\mathbb{E}[\|Z_{t+1}^{\tilde{F}} - \bar{Z}_{t+1}^{\tilde{F}}\|_F^2] + \frac{25(1-\lambda)\beta_1 L_F^2}{\alpha_2\mu^2}\frac{1}{K}\mathbb{E}[\|Z_{t+1}^{g} - \bar{Z}_{t+1}^{g}\|_F^2] \\
&+ \frac{4\beta_1}{\alpha_1}\frac{1}{K}\mathbb{E}[\|\Delta_{t+1}^{\tilde{F}} - U_{t+1}\|_F^2] + \frac{100\beta_1 L_F^2}{\alpha_2\mu^2}\frac{1}{K}\mathbb{E}[\|\Delta_{t+1}^{g} - V_{t+1}\|_F^2],
\end{aligned} \tag{21}$$

where $w_2 = w_3 = \frac{2\beta_1 L_{g_y}^2((11+32/\alpha_1^2)L_{\tilde{F}}^2 + (450+800/\alpha_2^2)L_F^2)}{\mu^2(1-\lambda^2)}$. With this potential function, we established the upper bound for each item. Then, with appropriate hyperparameters as shown in Theorem 1, we are able to get

$$\begin{aligned}
\mathcal{L}_{t+1} - \mathcal{L}_t &\leq -\frac{\eta\beta_1}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] - \frac{\eta\beta_1 L_F^2}{2}\mathbb{E}[\|\bar{y}_t - y^*(\bar{x}_t)\|^2] \\
&+ \frac{9\alpha_1\beta_1\eta^2\sigma_{\tilde{F}}^2}{2} + \frac{125\beta_1\alpha_2\eta^2 L_F^2\sigma^2}{\mu^2} + \frac{3\eta\beta_1 C_{g_{xy}}^2 C_{f_y}^2}{\mu^2}(1 - \frac{\mu}{L_{g_y}})^{2J}.
\end{aligned} \tag{22}$$

With the help of this inequality, we can establish the convergence rate of Algorithm 1.

Proof Sketch of Theorem 2. With the additional Assumption 6, the consensus errors can be decoupled from each other. Then, we developed the following potential function for establishing the convergence rate of Algorithm 1.

$$\mathcal{L}_{t+1} = F(x_{t+1}) + \frac{6\beta_1 L_F^2}{\beta_2\mu}\mathbb{E}[\|\bar{y}_{t+1} - y^*(\bar{x}_{t+1})\|^2] + \frac{3\beta_1}{\alpha_1}\frac{1}{K}\mathbb{E}[\|\Delta_{t+1}^{\tilde{F}} - U_{t+1}\|_F^2] + \frac{50\beta_1 L_F^2}{\alpha_2\mu^2}\frac{1}{K}\mathbb{E}[\|\Delta_{t+1}^{g} - V_{t+1}\|_F^2]. \tag{23}$$

Similarly, we can know how the potential function evolves in each iteration and then establish the convergence rate.

Proof Sketch of Theorem 3. Establishing the convergence rate of Algorithm 2 is much more challenging due to the complicated variance-reduced gradient estimator. Directly employing the potential function in Eq. (21) cannot give us the desired result. To address this challenging problem, we developed a novel potential function, which is shown below.

$$\begin{aligned}\mathcal{L}_{t+1} &= \mathbb{E}[F(x_{t+1})] + \frac{6\beta_1 L_F^2}{\beta_2\mu}\mathbb{E}[\|\bar{y}_{t+1} - y^*(\bar{x}_{t+1})\|^2] \\ &\quad + w_2\frac{1}{K}\mathbb{E}[\|X_{t+1} - \bar{X}_{t+1}\|_F^2] + w_3\frac{1}{K}\mathbb{E}[\|Y_{t+1} - \bar{Y}_{t+1}\|_F^2] \\ &\quad + \beta_1(1-\lambda)\frac{1}{K}\mathbb{E}[\|Z_{t+1}^{\tilde{F}} - \bar{Z}_{t+1}^{\tilde{F}}\|_F^2] + \frac{\beta_1(1-\lambda)L_F^2}{\mu^2}\frac{1}{K}\mathbb{E}[\|Z_{t+1}^{g} - \bar{Z}_{t+1}^{g}\|_F^2] \\ &\quad + 2\beta_1\frac{1}{K}\mathbb{E}[\|\Delta_{t+1}^{\tilde{F}} - U_{t+1}\|_F^2] + \frac{2\beta_1 L_F^2}{\mu^2}\frac{1}{K}\mathbb{E}[\|\Delta_{t+1}^{g} - V_{t+1}\|_F^2] \\ &\quad + \frac{3\beta_1}{\alpha_1\eta}\mathbb{E}[\|(\Delta_{t+1}^{\tilde{F}} - U_{t+1})\frac{1}{K}\mathbf{1}\|^2] + \frac{50\beta_1 L_F^2}{\alpha_2\eta\mu^2}\mathbb{E}[\|(\Delta_{t+1}^{g} - V_{t+1})\frac{1}{K}\mathbf{1}\|^2] \ ,\end{aligned} \tag{24}$$

where $w_2 = w_3 = \frac{2\beta_1\ell_{g_y}^2((51+48/\alpha_1 K)L_{\tilde{F}}^2 + (98+800/\alpha_2 K)L_F^2)}{\mu^2(1-\lambda^2)}$. Compared with Eq. (21), there are two additional terms, which are critical to achieve linear speedup. In particular, with this novel design, we are able to set $\alpha_i = 1/K$. As a result, β_i and α_i are decoupled (See Eq. (17)). In this way, Algorithm 2 can achieve linear speedup with respect to the number of participants (See Corollary 3). In comparison, α_i has to be $O(1)$ in Theorem 1.

In summary, we proposed three novel potential functions for establishing the convergence rate in Theorems 1- 3, disclosing how hyperparameters affect convergence rates.

A.2 Proof of Theorem 1

A.2.1 Characterization of $F^{(k)}(x)$

Lemma 1. *Ghadimi and Wang (2018) Given Assumptions 2-5, the following inequalities hold.*

$$\begin{aligned}&\|\nabla F^{(k)}(x) - \nabla F^{(k)}(x,y)\| \le L_F\|y - y^*(x)\| \ , \\ &\|\nabla F^{(k)}(x_1) - \nabla F^{(k)}(x_2)\| \le L_F^*\|x_1 - x_2\| \ , \\ &\|y^*(x_1) - y^*(x_2)\| \le L_y\|x_1 - x_2\| \ ,\end{aligned} \tag{25}$$

where $L_F = L_{f_x} + \frac{L_{f_y}C_{g_{xy}}}{\mu} + \frac{C_{f_y}L_{g_{xy}}}{\mu} + \frac{L_{g_{yy}}C_{f_y}C_{g_{xy}}}{\mu^2}$ $L_F^* = L_F + \frac{L_F C_{g_{xy}}}{\mu}$, $L_y = \frac{C_{g_{xy}}}{\mu}$.

A.2.2 Characterization of $\nabla\tilde{F}^{(k)}(x,y)$

Lemma 2. *Ghadimi and Wang (2018) Given Assumptions 2-5, the following inequalities hold.*

$$\begin{aligned}&\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big] = \frac{1}{L_{g_y}}\sum_{j=0}^{J-1}\Big(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y)\Big)^j \ , \\ &\Big\|\Big(\nabla_{yy}^2 g^{(k)}(x,y)\Big)^{-1} - \mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\Big\| \le \frac{1}{\mu}(1 - \frac{\mu}{L_{g_y}})^J \ , \\ &\mathbb{E}\Big[\Big\|\Big(\nabla_{yy}^2 g^{(k)}(x,y)\Big)^{-1} - \frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big\|\Big] \le \frac{2}{\mu} \ , \\ &\mathbb{E}\Big[\Big\|\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big\|\Big] \le \frac{1}{\mu} \ .\end{aligned} \tag{26}$$

Lemma 3. *(Bias) Ghadimi and Wang (2018) Given Assumptions 2-5, the approximation error of* $\nabla\tilde{F}^{(k)}(x,y)$ *for* $\nabla F^{(k)}(x,y)$ *can be bounded as follows:*

$$\|\nabla F^{(k)}(x,y) - \nabla\tilde{F}^{(k)}(x,y)\| \le \frac{C_{g_{xy}}C_{f_y}}{\mu}(1 - \frac{\mu}{L_{g_y}})^J \ . \tag{27}$$

Lemma 4. *(Variance) Given Assumptions 2-5, the variance of the stochastic hypergradient can be bounded as follows:*

$$\mathbb{E}[\|\nabla \tilde{F}^{(k)}(x,y) - \nabla \tilde{F}^{(k)}(x,y;\tilde{\xi})\|^2] \leq \sigma_{\tilde{F}}^2 \,, \tag{28}$$

where $\sigma_{\tilde{F}}^2 = 4\sigma^2 + \frac{4C_{f_y}^2\sigma^2}{\mu^2} + \frac{4\sigma^2(\sigma^2+C_{g_{xy}}^2)}{\mu^2} + \frac{16(\sigma^2+C_{g_{xy}}^2)(\sigma^2+C_{f_y}^2)}{\mu^2}$.

Proof.

$$\begin{aligned}
&\mathbb{E}[\|\nabla \tilde{F}^{(k)}(x,y) - \nabla \tilde{F}^{(k)}(x,y;\tilde{\xi})\|^2] \\
&= \mathbb{E}\Big[\Big\|\nabla_x f^{(k)}(x,y) - \nabla_{xy}^2 g^{(k)}(x,y)\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\nabla_y f^{(k)}(x,y) \\
&\quad - \nabla_x f^{(k)}(x,y;\xi) + \nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\Big(\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big)\nabla_y f^{(k)}(x,y;\xi)\Big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|\nabla_x f^{(k)}(x,y) - \nabla_x f^{(k)}(x,y;\xi) \\
&\quad + \nabla_{xy}^2 g^{(k)}(x,y)\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\nabla_y f^{(k)}(x,y) \\
&\quad - \nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\nabla_y f^{(k)}(x,y) \\
&\quad + \nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\nabla_y f^{(k)}(x,y) \\
&\quad - \nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\nabla_y f^{(k)}(x,y;\xi) \\
&\quad + \nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\nabla_y f^{(k)}(x,y;\xi) \\
&\quad - \nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\Big(\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big)\nabla_y f^{(k)}(x,y;\xi)\Big\|^2\Big] \\
&\leq 4\sigma^2 + 4\mathbb{E}\Big[\Big\|\Big(\nabla_{xy}^2 g^{(k)}(x,y) - \nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\Big)\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\nabla_y f^{(k)}(x,y)\Big\|^2\Big] \\
&\quad + 4\mathbb{E}\Big[\Big\|\nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\Big(\nabla_y f^{(k)}(x,y) - \nabla_y f^{(k)}(x,y;\xi)\Big)\Big\|^2\Big] \\
&\quad + 4\mathbb{E}\Big[\Big\|\nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\Big(\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big] \\
&\qquad - \frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big)\nabla_y f^{(k)}(x,y;\xi)\Big\|^2\Big] \\
&\overset{(s_1)}{\leq} 4\sigma^2 + \frac{4C_{f_y}^2\sigma^2}{\mu^2} + \frac{4\sigma^2(\sigma^2+C_{g_{xy}}^2)}{\mu^2} + \frac{16(\sigma^2+C_{g_{xy}}^2)(\sigma^2+C_{f_y}^2)}{\mu^2} \,,
\end{aligned} \tag{29}$$

where (s_1) holds due to Lemma 2, Assumptions 4-5, and the following inequality.

$$\begin{aligned}
&\Big\|\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\Big\| = \Big\|\frac{1}{L_{g_y}}\sum_{j=0}^{J-1}\Big(I-\frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y)\Big)^j\Big\| \\
&\leq \frac{1}{L_{g_y}}\sum_{j=0}^{J-1}\Big\|\Big(I-\frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y)\Big)^j\Big\| \\
&\overset{(s_1)}{\leq} \frac{1}{L_{g_y}}\sum_{j=0}^{J-1}\Big(1-\frac{\mu}{L_{g_y}}\Big)^j \\
&\leq \frac{1}{\mu}\ ,
\end{aligned} \tag{30}$$

where (s_1) holds due to Assumption 2. □

Lemma 5. *(Smoothness) Given Assumptions 2-5, the approximated hypergradient $\nabla\tilde{F}^{(k)}(x,y)$ is $L_{\tilde{F}}$-Lipschitz continuous:*

$$\|\nabla\tilde{F}^{(k)}(x_1,y_1)-\nabla\tilde{F}^{(k)}(x_2,y_2)\|^2 \leq L_{\tilde{F}}^2(\|x_1-x_2\|^2+\|y_1-y_2\|^2), \tag{31}$$

where $L_{\tilde{F}}^2 = 4L_{f_x}^2 + 4C_{g_{xy}}^2 C_{f_y}^2\frac{J^2L_{g_{yy}}^2}{\mu^2L_{g_y}^2} + \frac{4C_{g_{xy}}^2L_{f_y}^2}{\mu^2} + \frac{L_{g_{xy}}^2C_{f_y}^2}{\mu^2}$, $(x_1,y_1)\in\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}$, $(x_2,y_2)\in\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}$.

Proof.

$$\begin{aligned}
&\|\nabla\tilde{F}^{(k)}(x_1,y_1)-\nabla\tilde{F}^{(k)}(x_2,y_2)\|^2 \\
&= \Big\|\nabla_x f^{(k)}(x_1,y_1) - \nabla_{xy}^2 g^{(k)}(x_1,y_1)\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x_1,y_1;\zeta_j))\Big]\nabla_y f^{(k)}(x_1,y_1) \\
&\quad - \nabla_x f^{(k)}(x_2,y_2) + \nabla_{xy}^2 g^{(k)}(x_2,y_2)\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x_2,y_2;\zeta_j))\Big]\nabla_y f^{(k)}(x_2,y_2)\Big\|^2 \\
&\leq \Big\|\nabla_x f^{(k)}(x_1,y_1) - \nabla_x f^{(k)}(x_2,y_2) \\
&\quad - \nabla_{xy}^2 g^{(k)}(x_1,y_1)\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x_1,y_1;\zeta_j))\Big]\nabla_y f^{(k)}(x_1,y_1) \\
&\quad + \nabla_{xy}^2 g^{(k)}(x_1,y_1)\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x_2,y_2;\zeta_j))\Big]\nabla_y f^{(k)}(x_1,y_1) \\
&\quad - \nabla_{xy}^2 g^{(k)}(x_1,y_1)\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x_2,y_2;\zeta_j))\Big]\nabla_y f^{(k)}(x_1,y_1) \\
&\quad + \nabla_{xy}^2 g^{(k)}(x_1,y_1)\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x_2,y_2;\zeta_j))\Big]\nabla_y f^{(k)}(x_2,y_2) \\
&\quad - \nabla_{xy}^2 g^{(k)}(x_1,y_1)\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x_2,y_2;\zeta_j))\Big]\nabla_y f^{(k)}(x_2,y_2) \\
&\quad + \nabla_{xy}^2 g^{(k)}(x_2,y_2)\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x_2,y_2;\zeta_j))\Big]\nabla_y f^{(k)}(x_2,y_2)\Big\|^2 \\
&\overset{(s_1)}{\leq} \Big(4L_{f_x}^2 + 4C_{g_{xy}}^2C_{f_y}^2\frac{J^2L_{g_{yy}}^2}{\mu^2L_{g_y}^2} + \frac{4C_{g_{xy}}^2L_{f_y}^2}{\mu^2} + \frac{L_{g_{xy}}^2C_{f_y}^2}{\mu^2}\Big)(\|x_1-x_2\|^2+\|y_1-y_2\|^2)\ ,
\end{aligned} \tag{32}$$

where (s_1) holds due to Assumptions 4-5 and the following inequality.

$$
\begin{aligned}
&\Big\|\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{L_{g_y}}\nabla^2_{yy}g^{(k)}(x_1,y_1;\zeta_j))\Big]-\mathbb{E}\Big[\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{L_{g_y}}\nabla^2_{yy}g^{(k)}(x_2,y_2;\zeta_j))\Big]\Big\| \\
&=\Big\|\frac{1}{L_{g_y}}\sum_{j=0}^{J-1}\Big(I-\frac{1}{L_{g_y}}\nabla^2_{yy}g^{(k)}(x_1,y_1)\Big)^j-\frac{1}{L_{g_y}}\sum_{j=0}^{J-1}\Big(I-\frac{1}{L_{g_y}}\nabla^2_{yy}g^{(k)}(x_2,y_2)\Big)^j\Big\| \\
&\leq\frac{1}{L_{g_y}}\sum_{j=1}^{J-1}\Big\|\Big(I-\frac{1}{L_{g_y}}\nabla^2_{yy}g^{(k)}(x_1,y_1)\Big)^j-\Big(I-\frac{1}{L_{g_y}}\nabla^2_{yy}g^{(k)}(x_2,y_2)\Big)^j\Big\| \\
&\overset{(s_1)}{=}\frac{1}{L_{g_y}}\sum_{j=1}^{J-1}\Bigg(\Big\|\Big(\frac{1}{L_{g_y}}\nabla^2_{yy}g^{(k)}(x_1,y_1)-\frac{1}{L_{g_y}}\nabla^2_{yy}g^{(k)}(x_2,y_2)\Big) \\
&\quad\times\Big(\sum_{i=0}^{j-1}\Big(I-\frac{1}{L_{g_y}}\nabla^2_{yy}g^{(k)}(x_1,y_1)\Big)^i\Big(I-\frac{1}{L_{g_y}}\nabla^2_{yy}g^{(k)}(x_2,y_2)\Big)^{j-1-i}\Big)\Big\|\Bigg) \\
&\leq\frac{1}{L_{g_y}}\sum_{j=1}^{J-1}\Bigg(\Big\|\frac{1}{L_{g_y}}\nabla^2_{yy}g^{(k)}(x_1,y_1)-\frac{1}{L_{g_y}}\nabla^2_{yy}g^{(k)}(x_2,y_2)\Big\| \\
&\quad\times\Big\|\sum_{i=0}^{j-1}\Big(I-\frac{1}{L_{g_y}}\nabla^2_{yy}g^{(k)}(x_2,y_2)\Big)^i\Big(I-\frac{1}{L_{g_y}}\nabla^2_{yy}g^{(k)}(x_2,y_2)\Big)^{j-1-i}\Big\|\Bigg) \\
&\leq\frac{1}{L_{g_y}}\sum_{j=1}^{J-1}\Big(\frac{L_{g_{yy}}}{L_{g_y}}\|(x_1,y_1)-(x_2,y_2)\|\times\sum_{i=0}^{j-1}(1-\frac{\mu}{L_{g_y}})^{j-1}\Big) \\
&\leq\frac{J}{L_{g_y}}\sum_{j=1}^{J-1}\Big(\frac{L_{g_{yy}}}{L_{g_y}}\|(x_1,y_1)-(x_2,y_2)\|\times(1-\frac{\mu}{L_{g_y}})^{j-1}\Big) \\
&\leq\frac{JL_{g_{yy}}}{\mu L_{g_y}}\|(x_1,y_1)-(x_2,y_2)\|\ ,
\end{aligned}
\tag{33}
$$

where (s_1) holds due to $a^n-b^n=(a-b)(\sum_{i=0}^{n-1}a^ib^{n-1-i})$.

□

A.2.3 Characterization of Gradient Estimators

Lemma 6. *Given Assumptions 1-5, the following inequality holds.*

$$
\begin{aligned}
&\frac{1}{K}\mathbb{E}[\|\Delta_t^{\tilde{F}}-U_t\|^2]\leq(1-\alpha_1\eta)\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^{\tilde{F}}-U_{t-1}\|^2]+\frac{L_{\tilde{F}}^2}{\alpha_1\eta}\frac{1}{K}\mathbb{E}[\|X_t-X_{t-1}\|^2] \\
&\quad+\frac{L_{\tilde{F}}^2}{\alpha_1\eta}\frac{1}{K}\mathbb{E}[\|Y_t-Y_{t-1}\|^2]+\alpha_1^2\eta^2\sigma_{\tilde{F}}^2\ .
\end{aligned}
\tag{34}
$$

Proof.

$$
\begin{aligned}
&\frac{1}{K}\mathbb{E}[\|\Delta_t^{\tilde{F}} - U_t\|^2] \\
&= \frac{1}{K}\mathbb{E}[\|\Delta_t^{\tilde{F}} - (1-\alpha_1\eta)U_{t-1} - \alpha_1\eta\Delta_t^{\tilde{F}_{\tilde{\xi}_t}}\|^2] \\
&= \frac{1}{K}\mathbb{E}[\|(1-\alpha_1\eta)(\Delta_{t-1}^{\tilde{F}} - U_{t-1}) + (1-\alpha_1\eta)(\Delta_t^{\tilde{F}} - \Delta_{t-1}^{\tilde{F}}) + \alpha_1\eta(\Delta_t^{\tilde{F}} - \Delta_t^{\tilde{F}_{\tilde{\xi}_t}})\|^2] \\
&\overset{(s_1)}{=} (1-\alpha_1\eta)^2\frac{1}{K}\mathbb{E}[\|(\Delta_{t-1}^{\tilde{F}} - U_{t-1}) + (\Delta_t^{\tilde{F}} - \Delta_{t-1}^{\tilde{F}})\|^2] + \alpha_1^2\eta^2\frac{1}{K}\mathbb{E}[\|\Delta_t^{\tilde{F}} - \Delta_t^{\tilde{F}_{\tilde{\xi}_t}}\|^2] \\
&\overset{(s_2)}{\leq} (1-\alpha_1\eta)\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^{\tilde{F}} - U_{t-1}\|^2] + \frac{1}{\alpha_1\eta}\frac{1}{K}\mathbb{E}[\|\Delta_t^{\tilde{F}} - \Delta_{t-1}^{\tilde{F}}\|^2] + \alpha_1^2\eta^2\sigma_{\tilde{F}}^2 \\
&\overset{(s_3)}{\leq} (1-\alpha_1\eta)\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^{\tilde{F}} - U_{t-1}\|^2] + \frac{L_{\tilde{F}}^2}{\alpha_1\eta}\frac{1}{K}\mathbb{E}[\|X_t - X_{t-1}\|^2] \\
&\quad + \frac{L_{\tilde{F}}^2}{\alpha_1\eta}\frac{1}{K}\mathbb{E}[\|Y_t - Y_{t-1}\|^2] + \alpha_1^2\eta^2\sigma_{\tilde{F}}^2 \ ,
\end{aligned} \tag{35}
$$

where (s_1) holds due to $\Delta_t^{\tilde{F}} = \mathbb{E}[\Delta_t^{\tilde{F}_{\tilde{\xi}_t}}]$, (s_2) holds due to Lemma 4 and Lemma 34 with $a = \frac{\alpha_1\eta}{1-\alpha_1\eta}$, (s_3) holds due to Lemma 5. □

Lemma 7. *Given Assumptions 1-5, the following inequality holds.*

$$
\begin{aligned}
&\frac{1}{K}\mathbb{E}[\|\Delta_t^{g} - V_t\|^2] \leq (1-\alpha_2\eta)\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^{g} - V_{t-1}\|^2] + \frac{L_{g_y}^2}{\alpha_2\eta}\frac{1}{K}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] \\
&\quad + \frac{L_{g_y}^2}{\alpha_2\eta}\frac{1}{K}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] + \alpha_2^2\eta^2\sigma^2 \ .
\end{aligned} \tag{36}
$$

Proof.

$$
\begin{aligned}
&\frac{1}{K}\mathbb{E}[\|\Delta_t^{g} - V_t\|^2] \\
&= \frac{1}{K}\mathbb{E}[\|\Delta_t^{g} - (1-\alpha_2\eta)V_{t-1} - \alpha_2\eta\Delta_t^{g_{\zeta_t}}\|^2] \\
&= \frac{1}{K}\mathbb{E}[\|(1-\alpha_2\eta)(\Delta_{t-1}^{g} - V_{t-1}) + (1-\alpha_2\eta)(\Delta_t^{g} - \Delta_{t-1}^{g}) + \alpha_2\eta(\Delta_t^{g} - \Delta_t^{g_{\zeta_t}})\|^2] \\
&\overset{(s_1)}{=} (1-\alpha_2\eta)^2\frac{1}{K}\mathbb{E}[\|(\Delta_{t-1}^{g} - V_{t-1}) + (\Delta_t^{g} - \Delta_{t-1}^{g})\|^2] + \alpha_2^2\eta^2\frac{1}{K}\mathbb{E}[\|\Delta_t^{g} - \Delta_t^{g_{\zeta_t}}\|^2] \\
&\overset{(s_2)}{\leq} (1-\alpha_2\eta)\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^{g} - V_{t-1}\|^2] + \frac{1}{\alpha_2\eta}\frac{1}{K}\mathbb{E}[\|\Delta_t^{g} - \Delta_{t-1}^{g}\|^2] + \alpha_2^2\eta^2\sigma^2 \\
&\overset{(s_3)}{\leq} (1-\alpha_2\eta)\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^{g} - V_{t-1}\|^2] + \frac{L_{g_y}^2}{\alpha_2\eta}\frac{1}{K}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] \\
&\quad + \frac{L_{g_y}^2}{\alpha_2\eta}\frac{1}{K}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] + \alpha_2^2\eta^2\sigma^2 \ ,
\end{aligned} \tag{37}
$$

where (s_1) holds due to $\Delta_t^{g} = \mathbb{E}[\Delta_t^{g_{\zeta_t}}]$, (s_2) holds due to Assumption 3 and Lemma 34 with $a = \frac{\alpha_2\eta}{1-\alpha_2\eta}$, (s_3) holds due to Assumption 5. □

A.2.4 Characterization of Consensus Errors

Lemma 8. *Given Assumptions 1-5, the following inequality holds.*

$$
\begin{aligned}
&\frac{1}{K}\mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}}\|_F^2] \leq \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{2\alpha_1^2\eta^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|U_{t-1} - \Delta_{t-1}^{\tilde{F}}\|_F^2] + \frac{\alpha_1^2\eta^2\sigma_{\tilde{F}}^2}{1-\lambda} \\
&\quad + \frac{2\alpha_1^2\eta^2L_{\tilde{F}}^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] + \frac{2\alpha_1^2\eta^2L_{\tilde{F}}^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] \ .
\end{aligned} \tag{38}
$$

Proof.

$$
\begin{aligned}
& \frac{1}{K}\mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}}\|_F^2] \\
& = \frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}}W + U_t - U_{t-1} - \bar{Z}_{t-1}^{\tilde{F}} - \bar{U}_t + \bar{U}_{t-1}\|_F^2] \\
& \stackrel{(s_1)}{\leq} \frac{1}{K}\lambda\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|U_t - U_{t-1} - \bar{U}_t + \bar{U}_{t-1}\|_F^2] \\
& \leq \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|U_t - U_{t-1}\|_F^2] \\
& = \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|(1-\alpha_1\eta)U_{t-1} + \alpha_1\eta\Delta_t^{\tilde{F}_{\tilde{\xi}_t}} - U_{t-1}\|_F^2] \\
& = \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\| - \alpha_1\eta U_{t-1} + \alpha_1\eta\Delta_t^{\tilde{F}} - \alpha_1\eta\Delta_t^{\tilde{F}} + \alpha_1\eta\Delta_t^{\tilde{F}_{\tilde{\xi}_t}}\|_F^2] \\
& \stackrel{(s_2)}{\leq} \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{\alpha_1^2\eta^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|U_{t-1} - \Delta_t^{\tilde{F}}\|_F^2] + \frac{\alpha_1^2\eta^2\sigma_{\tilde{F}}^2}{1-\lambda} \\
& \leq \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{2\alpha_1^2\eta^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|U_{t-1} - \Delta_{t-1}^{\tilde{F}}\|_F^2] + \frac{2\alpha_1^2\eta^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^{\tilde{F}} - \Delta_t^{\tilde{F}}\|_F^2] + \frac{\alpha_1^2\eta^2\sigma_{\tilde{F}}^2}{1-\lambda} \\
& \stackrel{(s_3)}{\leq} \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{2\alpha_1^2\eta^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|U_{t-1} - \Delta_{t-1}^{\tilde{F}}\|_F^2] + \frac{\alpha_1^2\eta^2\sigma_{\tilde{F}}^2}{1-\lambda} \\
& \quad + \frac{2\alpha_1^2\eta^2L_{\tilde{F}}^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] + \frac{2\alpha_1^2\eta^2L_{\tilde{F}}^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] \ ,
\end{aligned}
\tag{39}
$$

where (s_1) holds due to Lemma 34 with $a = \frac{1-\lambda}{\lambda}$, (s_2) holds due to Lemma 4, (s_3) holds due to Lemma 5.

□

Lemma 9. *Given Assumptions 1-5, the following inequality holds.*

$$
\begin{aligned}
\frac{1}{K}\mathbb{E}[\|Z_t^{g} - \bar{Z}_t^{g}\|_F^2] & \leq \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{g} - \bar{Z}_{t-1}^{g}\|_F^2] + \frac{2\alpha_2^2\eta^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|V_{t-1} - \Delta_{t-1}^{g}\|_F^2] + \frac{\alpha_2^2\eta^2\sigma^2}{1-\lambda} \\
& \quad + \frac{2\alpha_2^2\eta^2L_{g_y}^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] + \frac{2\alpha_2^2\eta^2L_{g_y}^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] \ .
\end{aligned}
\tag{40}
$$

Proof.

$$
\begin{aligned}
& \frac{1}{K}\mathbb{E}[\|Z_t^{g} - \bar{Z}_t^{g}\|_F^2] \\
& = \frac{1}{K}\mathbb{E}[\|Z_{t-1}^{g}W + V_t - V_{t-1} - \bar{Z}_{t-1}^{g} - \bar{V}_t + \bar{V}_{t-1}\|_F^2] \\
& \stackrel{(s_1)}{\leq} \frac{1}{K}\lambda\mathbb{E}[\|Z_{t-1}^{g} - \bar{Z}_{t-1}^{g}\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|V_t - V_{t-1} - \bar{V}_t + \bar{V}_{t-1}\|_F^2] \\
& \leq \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{g} - \bar{Z}_{t-1}^{g}\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|V_t - V_{t-1}\|_F^2] \\
& = \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{g} - \bar{Z}_{t-1}^{g}\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|(1-\alpha_2\eta)V_{t-1} + \alpha_2\eta\Delta_t^{g\zeta_t} - V_{t-1}\|_F^2] \\
& = \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{g} - \bar{Z}_{t-1}^{g}\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\| - \alpha_2\eta V_{t-1} + \alpha_2\eta\Delta_t^{g} - \alpha_2\eta\Delta_t^{g} + \alpha_2\eta\Delta_t^{g\zeta_t}\|_F^2] \\
& \stackrel{(s_2)}{\leq} \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{g} - \bar{Z}_{t-1}^{g}\|_F^2] + \frac{\alpha_2^2\eta^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|V_{t-1} - \Delta_t^{g}\|_F^2] + \frac{\alpha_2^2\eta^2\sigma^2}{1-\lambda} \\
& \leq \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{g} - \bar{Z}_{t-1}^{g}\|_F^2] + \frac{2\alpha_2^2\eta^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|V_{t-1} - \Delta_{t-1}^{g}\|_F^2] + \frac{2\alpha_2^2\eta^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^{g} - \Delta_t^{g}\|_F^2] + \frac{\alpha_2^2\eta^2\sigma^2}{1-\lambda} \\
& \stackrel{(s_3)}{\leq} \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{g} - \bar{Z}_{t-1}^{g}\|_F^2] + \frac{2\alpha_2^2\eta^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|V_{t-1} - \Delta_{t-1}^{g}\|_F^2] + \frac{\alpha_2^2\eta^2\sigma^2}{1-\lambda} \\
& \quad + \frac{2\alpha_2^2\eta^2L_{g_y}^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] + \frac{2\alpha_2^2\eta^2L_{g_y}^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] \ ,
\end{aligned}
\tag{41}
$$

where (s_1) holds due to Lemma 34 with $a = \frac{1-\lambda}{\lambda}$, (s_2) holds due to Assumption 3, (s_3) holds due to Assumption 5.

□

Lemma 10. *Given Assumptions 1-5, the following inequality holds.*

$$\mathbb{E}[\|X_{t+1} - \bar{X}_{t+1}\|_F^2] \leq \Big(1 - \frac{\eta(1-\lambda^2)}{2}\Big)\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + \frac{2\eta\beta_1^2}{1-\lambda^2}\mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}}\|_F^2] \ . \tag{42}$$

Proof.

$$\begin{aligned}
&\mathbb{E}[\|X_{t+1} - \bar{X}_{t+1}\|_F^2] \\
&\overset{(s_0)}{=} \mathbb{E}[\|X_t - \eta X_t(I-W) - \beta_1\eta Z_t^{\tilde{F}} - \bar{X}_t + \beta_1\eta\bar{Z}_t^{\tilde{F}}\|_F^2] \\
&= \mathbb{E}[\|(1-\eta)(X_t - \bar{X}_t) + \eta(X_tW - \bar{X}_t - \beta_1 Z_t^{\tilde{F}} + \beta_1\bar{Z}_t^{\tilde{F}})\|_F^2] \\
&\overset{(s_1)}{\leq} (1-\eta)\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + \eta\mathbb{E}[\|X_tW + \beta_1 Z_t^{\tilde{F}} - \bar{X}_t - \beta_1\bar{Z}_t^{\tilde{F}}\|_F^2] \\
&\overset{(s_2)}{\leq} (1-\eta)\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + \frac{\eta(1+\lambda^2)}{2\lambda^2}\mathbb{E}[\|X_tW - \bar{X}_t\|_F^2] + \frac{2\eta\beta_1^2}{1-\lambda^2}\mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}}\|_F^2] \\
&\overset{(s_3)}{\leq} \Big(1 - \frac{\eta(1-\lambda^2)}{2}\Big)\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + \frac{2\eta\beta_1^2}{1-\lambda^2}\mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}}\|_F^2] \ ,
\end{aligned} \tag{43}$$

where (s_0) holds due to $X_t(I-W)\mathbf{1} = 0$, (s_1) holds due to Lemma 34 with $a = \frac{\eta}{1-\eta}$, (s_2) holds due to Lemma 34 with $a = \frac{1-\lambda^2}{2\lambda^2}$, (s_3) holds due to $\|X_tW - \bar{X}_t\|_F^2 \leq \lambda^2\|X_t - \bar{X}_t\|_F^2$. □

Lemma 11. *Given Assumptions 1-5, the following inequality holds.*

$$\mathbb{E}[\|Y_{t+1} - \bar{Y}_{t+1}\|_F^2] \leq \Big(1 - \frac{\eta(1-\lambda^2)}{2}\Big)\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] + \frac{2\eta\beta_2^2}{1-\lambda^2}\mathbb{E}[\|Z_t^g - \bar{Z}_t^g\|_F^2] \ . \tag{44}$$

Proof. In terms of the definition of $x_{t+1}^{(k)}$, we have

$$\begin{aligned}
&\mathbb{E}[\|Y_{t+1} - \bar{Y}_{t+1}\|_F^2] \\
&\overset{(s_0)}{=} \mathbb{E}[\|Y_t - \eta Y_t(I-W) - \beta_2\eta Z_t^g - \bar{Y}_t + \beta_2\eta\bar{Z}_t^g\|_F^2] \\
&= \mathbb{E}[\|(1-\eta)(Y_t - \bar{Y}_t) + \eta(Y_tW - \bar{Y}_t - \beta_2 Z_t^g + \beta_2\bar{Z}_t^g)\|_F^2] \\
&\overset{(s_1)}{\leq} (1-\eta)\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] + \eta\mathbb{E}[\|Y_tW + \beta_2 Z_t^g - \bar{Y}_t - \beta_2\bar{Z}_t^g\|_F^2] \\
&\overset{(s_2)}{\leq} (1-\eta)\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] + \frac{\eta(1+\lambda^2)}{2\lambda^2}\mathbb{E}[\|Y_tW - \bar{Y}_t\|_F^2] + \frac{2\eta\beta_2^2}{1-\lambda^2}\mathbb{E}[\|Z_t^g - \bar{Z}_t^g\|_F^2] \\
&\overset{(s_3)}{\leq} \Big(1 - \frac{\eta(1-\lambda^2)}{2}\Big)\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] + \frac{2\eta\beta_2^2}{1-\lambda^2}\mathbb{E}[\|Z_t^g - \bar{Z}_t^g\|_F^2] \ ,
\end{aligned} \tag{45}$$

where (s_0) holds due to $Y_t(I-W)\mathbf{1} = 0$, (s_1) holds due to Lemma 34 with $a = \frac{\eta}{1-\eta}$, (s_2) holds due to Lemma 34 with $a = \frac{1-\lambda^2}{2\lambda^2}$, (s_3) holds due to $\|Y_tW - \bar{Y}_t\|_F^2 \leq \lambda^2\|Y_t - \bar{Y}_t\|_F^2$.

□

Lemma 12. *Given Assumptions 1-5, the following inequality holds.*

$$\mathbb{E}[\|X_{t+1} - X_t\|_F^2] \leq 8\eta^2\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + 4\eta^2\beta_1^2\mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}}\|_F^2] + 4\eta^2\beta_1^2\mathbb{E}[\|\bar{Z}_t^{\tilde{F}}\|_F^2] \ . \tag{46}$$

Proof.

$$
\begin{aligned}
& \mathbb{E}[\|X_{t+1} - X_t\|_F^2] \\
& = \mathbb{E}[\|X_t - \eta X_t(I-W) - \beta_1 \eta Z_t^{\tilde{F}} - X_t\|_F^2] \\
& = \mathbb{E}[\|\eta X_t(W-I) - \beta_1 \eta Z_t^{\tilde{F}}\|_F^2] \\
& \leq 2\eta^2 \mathbb{E}[\|X_t(W-I)\|_F^2] + 2\eta^2\beta_1^2 \mathbb{E}[\|Z_t^{\tilde{F}}\|_F^2] \\
& = 2\eta^2 \mathbb{E}[\|(X_t - \bar{X}_t)(W-I)\|_F^2] + 2\eta^2\beta_1^2 \mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}} + \bar{Z}_t^{\tilde{F}}\|_F^2] \\
& \overset{(s_1)}{\leq} 8\eta^2 \mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + 4\eta^2\beta_1^2 \mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}}\|_F^2] + 4\eta^2\beta_1^2 \mathbb{E}[\|\bar{Z}_t^{\tilde{F}}\|_F^2] \ ,
\end{aligned} \tag{47}
$$

where (s_1) holds due to $\|AB\|_F \leq \|A\|_2 \|B\|_F$ and $\|I - W\|_2 \leq 2$.

□

Lemma 13. *Given Assumptions 1-5, the following inequality holds.*

$$
\mathbb{E}[\|Y_{t+1} - Y_t\|_F^2] \leq 8\eta^2 \mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] + 4\eta^2\beta_2^2 \mathbb{E}[\|Z_t^{g} - \bar{Z}_t^{g}\|_F^2] + 4\eta^2\beta_2^2 \mathbb{E}[\|\bar{Z}_t^{g}\|_F^2] \ . \tag{48}
$$

Proof. In terms of the definition of $\tilde{x}_{t+1}^{(k)}$, we can get

$$
\begin{aligned}
& \mathbb{E}[\|Y_{t+1} - Y_t\|_F^2] \\
& = \mathbb{E}[\|Y_t - \eta Y_t(I-W) - \beta_2 \eta Z_t^{g} - Y_t\|_F^2] \\
& = \mathbb{E}[\|\eta Y_t(W-I) - \beta_2 \eta Z_t^{g}\|_F^2] \\
& \leq 2\eta^2 \mathbb{E}[\|Y_t(W-I)\|_F^2] + 2\eta^2\beta_2^2 \mathbb{E}[\|Z_t^{g}\|_F^2] \\
& = 2\eta^2 \mathbb{E}[\|(Y_t - \bar{Y}_t)(W-I)\|_F^2] + 2\eta^2\beta_2^2 \mathbb{E}[\|Z_t^{g} - \bar{Z}_t^{g} + \bar{Z}_t^{g}\|_F^2] \\
& \overset{(s_1)}{\leq} 8\eta^2 \mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] + 4\eta^2\beta_2^2 \mathbb{E}[\|Z_t^{g} - \bar{Z}_t^{g}\|_F^2] + 4\eta^2\beta_2^2 \mathbb{E}[\|\bar{Z}_t^{g}\|_F^2] \ ,
\end{aligned} \tag{49}
$$

where (s_1) holds due to $\|AB\|_F \leq \|A\|_2 \|B\|_F$ and $\|I - W\|_2 \leq 2$. □

A.2.5 Proof of Theorem 1

Proof. According to the smoothness of $F(x)$, we can get

$$
\begin{aligned}
& \mathbb{E}[F(\bar{x}_{t+1})] \leq \mathbb{E}[F(\bar{x}_t)] + \mathbb{E}[\langle \nabla F(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t \rangle] + \frac{L_F^*}{2} \mathbb{E}[\|\bar{x}_{t+1} - \bar{x}_t\|^2] \\
& = \mathbb{E}[F(\bar{x}_t)] - \eta\beta_1 \mathbb{E}[\langle \nabla F(\bar{x}_t), \bar{u}_t \rangle] + \frac{\eta^2\beta_1^2 L_F^*}{2} \mathbb{E}[\|\bar{u}_t\|^2] \\
& = \mathbb{E}[F(\bar{x}_t)] - \frac{\eta\beta_1}{2} \mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] + \Big(\frac{\eta^2\beta_1^2 L_F^*}{2} - \frac{\eta\beta_1}{2}\Big) \mathbb{E}[\|\bar{u}_t\|^2] + \frac{\eta\beta_1}{2} \mathbb{E}[\|\nabla F(\bar{x}_t) - \bar{u}_t\|^2] \\
& \overset{(s_1)}{\leq} \mathbb{E}[F(\bar{x}_t)] - \frac{\eta\beta_1}{2} \mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] - \frac{\eta\beta_1}{4} \mathbb{E}[\|\bar{u}_t\|^2] + \eta\beta_1 L_F^2 \mathbb{E}[\|\bar{y}_t - y^*(\bar{x}_t)\|^2] \\
& \quad + \frac{3\eta\beta_1 C_{g_{xy}}^2 C_{f_y}^2}{\mu^2} (1 - \frac{\mu}{L_{g_y}})^{2J} + \frac{3\eta\beta_1 L_{\tilde{F}}^2}{K} \mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + \frac{3\eta\beta_1 L_{\tilde{F}}^2}{K} \mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] \\
& \quad + 3\eta\beta_1 \frac{1}{K} \mathbb{E}[\|\Delta_t^{\tilde{F}} - U_t\|_F^2] \ ,
\end{aligned} \tag{50}
$$

where (s_1) holds due to $\eta \leq \frac{1}{2\beta_1 L_F^*}$ and the following inequality:

$$
\begin{aligned}
&\mathbb{E}[\|\nabla F(\bar{x}_t) - \bar{u}_t\|^2] \\
&\leq 2\mathbb{E}[\|\nabla F(\bar{x}_t) - \nabla F(\bar{x}_t, \bar{y}_t)\|^2] + 2\mathbb{E}[\|\nabla F(\bar{x}_t, \bar{y}_t) - \nabla \tilde{F}(\bar{x}_t, \bar{y}_t) + (\underline{\Delta}_t^{\tilde{F}} - \Delta_t^{\tilde{F}} + \Delta_t^{\tilde{F}} - U_t)\frac{1}{K}\mathbf{1}\|^2] \\
&\leq 2\mathbb{E}[\|\nabla F(\bar{x}_t) - \nabla F(\bar{x}_t, \bar{y}_t)\|^2] + 6\mathbb{E}[\|\nabla F(\bar{x}_t, \bar{y}_t) - \nabla \tilde{F}(\bar{x}_t, \bar{y}_t)\|^2] + 6\mathbb{E}\Big[\Big\|(\underline{\Delta}_t^{\tilde{F}} - \Delta_t^{\tilde{F}})\frac{1}{K}\mathbf{1}\Big\|^2\Big] \\
&\quad + 6\mathbb{E}\Big[\Big\|(\Delta_t^{\tilde{F}} - U_t)\frac{1}{K}\mathbf{1}\Big\|^2\Big] \\
&\overset{(s_1)}{\leq} 2L_F^2\mathbb{E}[\|\bar{y}_t - y^*(\bar{x}_t)\|^2] + \frac{6C_{g_{xy}}^2 C_{f_y}^2}{\mu^2}(1 - \frac{\mu}{L_{g_y}})^{2J} + \frac{6L_{\tilde{F}}^2}{K}\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + \frac{6L_{\tilde{F}}^2}{K}\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] \\
&\quad + 6\frac{1}{K}\mathbb{E}[\|\Delta_t^{\tilde{F}} - U_t\|_F^2] \ ,
\end{aligned}
\tag{51}
$$

where (s_1) holds due to Lemma 3 and Lemma 5. Additionally, we can bound $\mathbb{E}[\|\bar{y}_t - y^*(\bar{x}_t)\|^2]$ as follows:

$$
\begin{aligned}
&\mathbb{E}[\|\bar{y}_{t+1} - y^*(\bar{x}_{t+1})\|^2] \\
&\overset{(s_1)}{\leq} (1 - \frac{\beta_2\eta\mu}{4})\mathbb{E}[\|\bar{y}_t - y^*(\bar{x}_t)\|^2] - \frac{3\eta\beta_2^2}{4}\mathbb{E}[\|\bar{v}_t\|^2] + \frac{25\eta\beta_1^2 L_y^2}{6\beta_2\mu}\mathbb{E}[\|\bar{u}_t\|^2] + \frac{25\beta_2\eta}{6\mu}\mathbb{E}[\|(\underline{\Delta}_t^g - V_t)\frac{1}{K}\mathbf{1}\|^2] \\
&\overset{(s_2)}{\leq} (1 - \frac{\beta_2\eta\mu}{4})\mathbb{E}[\|\bar{y}_t - y^*(\bar{x}_t)\|^2] - \frac{3\eta\beta_2^2}{4}\mathbb{E}[\|\bar{v}_t\|^2] + \frac{25\eta\beta_1^2 L_y^2}{6\beta_2\mu}\mathbb{E}[\|\bar{u}_t\|^2] \\
&\quad + \frac{25\beta_2\eta L_{g_y}^2}{3\mu}\frac{1}{K}\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + \frac{25\beta_2\eta L_{g_y}^2}{3\mu}\frac{1}{K}\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] + \frac{25\beta_2\eta}{3\mu}\frac{1}{K}\mathbb{E}[\|\Delta_t^g - V_t\|_F^2] \ ,
\end{aligned}
\tag{52}
$$

where (s_1) holds due to Lemma 9 in Yang et al. (2021) with $\beta_2 < \frac{1}{6L_{g_y}}$ and $0 < \eta < 1$, (s_2) holds due to the following inequality.

$$
\begin{aligned}
&\mathbb{E}[\|(\underline{\Delta}_t^g - V_t)\frac{1}{K}\mathbf{1}\|^2] \\
&= \mathbb{E}[\|(\underline{\Delta}_t^g - \Delta_t^g + \Delta_t^g - V_t)\frac{1}{K}\mathbf{1}\|^2] \\
&\leq 2\mathbb{E}[\|(\underline{\Delta}_t^g - \Delta_t^g)\frac{1}{K}\mathbf{1}\|^2] + 2\mathbb{E}[\|(\Delta_t^g - V_t)\frac{1}{K}\mathbf{1}\|^2] \\
&\leq \frac{2L_{g_y}^2}{K}\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] + \frac{2L_{g_y}^2}{K}\mathbb{E}[\|Y_t - \bar{Y}_t\|_F^2] + 2\frac{1}{K}\mathbb{E}[\|\Delta_t^g - V_t\|_F^2] \ .
\end{aligned}
\tag{53}
$$

Then, to investigate the convergence rate of Algorithm 1, we introduce the following potential function:

$$
\begin{aligned}
\mathcal{L}_{t+1} &= \mathbb{E}[F(x_{t+1})] + \frac{6\beta_1 L_F^2}{\beta_2\mu}\mathbb{E}[\|\bar{y}_{t+1} - y^*(\bar{x}_{t+1})\|^2] \\
&\quad + \frac{2\beta_1 L_{g_y}^2((11 + 32/\alpha_1^2)L_{\tilde{F}}^2 + (450 + 800/\alpha_2^2)L_F^2)}{\mu^2(1-\lambda^2)}\frac{1}{K}\mathbb{E}[\|X_{t+1} - \bar{X}_{t+1}\|_F^2] \\
&\quad + \frac{2\beta_1 L_{g_y}^2((11 + 32/\alpha_1^2)L_{\tilde{F}}^2 + (450 + 800/\alpha_2^2)L_F^2)}{\mu^2(1-\lambda^2)}\frac{1}{K}\mathbb{E}[\|Y_{t+1} - \bar{Y}_{t+1}\|_F^2] \\
&\quad + \frac{\beta_1(1-\lambda)}{2\alpha_1}\frac{1}{K}\mathbb{E}[\|Z_{t+1}^{\tilde{F}} - \bar{Z}_{t+1}^{\tilde{F}}\|_F^2] + \frac{25(1-\lambda)\beta_1 L_F^2}{\alpha_2\mu^2}\frac{1}{K}\mathbb{E}[\|Z_{t+1}^g - \bar{Z}_{t+1}^g\|_F^2] \\
&\quad + \frac{4\beta_1}{\alpha_1}\frac{1}{K}\mathbb{E}[\|\Delta_{t+1}^{\tilde{F}} - U_{t+1}\|_F^2] + \frac{100\beta_1 L_F^2}{\alpha_2\mu^2}\frac{1}{K}\mathbb{E}[\|\Delta_{t+1}^g - V_{t+1}\|_F^2] \ .
\end{aligned}
\tag{54}
$$

Based on Lemmas 6, 7, 8, 9, 10, 11, 12, 13, we can get

$$
\begin{aligned}
&\mathcal{L}_{t+1}-\mathcal{L}_t \\
&\leq -\frac{\eta\beta_1}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] - \frac{\eta\beta_1 L_F^2}{2}\mathbb{E}[\|\bar{y}_t - y^*(\bar{x}_t)\|^2] + \frac{3\eta\beta_1 C_{g_{xy}}^2 C_{f_y}^2}{\mu^2}(1-\frac{\mu}{L_{g_y}})^{2J} \\
&\quad + (\frac{25\eta\beta_1^3 L_F^2 L_{g_y}^2}{\beta_2^2\mu^2} + 4\alpha_1\beta_1^3\eta^4 L_{\tilde{F}}^2 + \frac{200\alpha_2\beta_1^3\eta^4 L_{g_y}^2 L_F^2}{\mu^2} + \frac{16\eta\beta_1^3 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{400\eta\beta_1^3 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2} - \frac{\eta\beta_1}{4})\mathbb{E}[\|\bar{u}_t\|^2] \\
&\quad + (4\alpha_1\beta_1\beta_2^2\eta^4 L_{\tilde{F}}^2 + \frac{200\alpha_2\beta_1\beta_2^2\eta^4 L_{g_y}^2 L_F^2}{\mu^2} + \frac{16\eta\beta_1\beta_2^2 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{400\eta\beta_1\beta_2^2 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2} - \frac{9\eta\beta_1\beta_2 L_F^2}{2\mu})\mathbb{E}[\|\bar{v}_t\|^2] \\
&\quad + (\frac{4\eta\beta_1^3 L_{g_y}^2((11+32/\alpha_1^2)L_{\tilde{F}}^2 + (450+800/\alpha_2^2)L_F^2)}{\mu^2(1-\lambda^2)^2} - \frac{\beta_1(1-\lambda)^2}{2\alpha_1} + 4\alpha_1\beta_1^3\eta^4 L_{\tilde{F}}^2 \\
&\quad\quad + \frac{200\alpha_2\beta_1^3\eta^4 L_{g_y}^2 L_F^2}{\mu^2} + \frac{16\eta\beta_1^3 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{400\eta\beta_1^3 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2})\frac{1}{K}\mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}}\|_F^2] \\
&\quad + (\frac{4\eta\beta_1\beta_2^2 L_{g_y}^2((11+32/\alpha_1^2)L_{\tilde{F}}^2 + (450+800/\alpha_2^2)L_F^2)}{\mu^2(1-\lambda^2)^2} - \frac{25(1-\lambda)^2\beta_1 L_F^2}{\alpha_2\mu^2} + 4\alpha_1\beta_1\beta_2^2\eta^4 L_{\tilde{F}}^2 \\
&\quad\quad + \frac{200\alpha_2\beta_1\beta_2^2\eta^4 L_{g_y}^2 L_F^2}{\mu^2} + \frac{16\eta\beta_1\beta_2^2 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{400\eta\beta_1\beta_2^2 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2})\frac{1}{K}\mathbb{E}[\|Z_t^{g} - \bar{Z}_t^{g}\|_F^2] \\
&\quad + \frac{\alpha_1\beta_1\eta^2\sigma_{\tilde{F}}^2}{2} + \frac{25\beta_1\alpha_2\eta^2 L_F^2\sigma^2}{\mu^2} + 4\beta_1\alpha_1\eta^2\sigma_{\tilde{F}}^2 + \frac{100\beta_1\alpha_2\eta^2 L_F^2\sigma^2}{\mu^2} \ .
\end{aligned} \tag{55}
$$

By setting the coefficient of $\mathbb{E}[\|\bar{u}_t\|^2]$ to be non-positive, we can get

$$
\begin{aligned}
&\frac{25\eta\beta_1^3 L_F^2 L_{g_y}^2}{\beta_2^2\mu^2} + 4\alpha_1\beta_1^3\eta^4 L_{\tilde{F}}^2 + \frac{200\alpha_2\beta_1^3\eta^4 L_{g_y}^2 L_F^2}{\mu^2} + \frac{16\eta\beta_1^3 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{400\eta\beta_1^3 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2} - \frac{\eta\beta_1}{4} \leq 0 \ , \\
&\frac{25\beta_1^2 L_y^2 L_F^2}{\beta_2^2\mu^2} + 4\alpha_1\beta_1^2\eta^3 L_{\tilde{F}}^2 + \frac{200\alpha_2\beta_1^2\eta^3 L_{g_y}^2 L_F^2}{\mu^2} + \frac{16\beta_1^2 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{400\beta_1^2 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2} - \frac{1}{4} \leq 0 \ .
\end{aligned} \tag{56}
$$

By setting $\beta_1 \leq \frac{\beta_2\mu}{15L_yL_F}$, we can get $\frac{25\beta_1^2L_y^2L_F^2}{\beta_2^2\mu^2} - \frac{1}{4} \leq -\frac{1}{8}$. Because $\alpha_1\eta < 1, \alpha_2\eta < 1, \eta < 1, \frac{L_{g_y}^2}{\mu^2} > 1$, we can get

$$
\begin{aligned}
&\frac{25\beta_1^2 L_y^2 L_F^2}{\beta_2^2\mu^2} + 4\alpha_1\beta_1^2\eta^3 L_{\tilde{F}}^2 + \frac{200\alpha_2\beta_1^2\eta^3 L_{g_y}^2 L_F^2}{\mu^2} + \frac{16\beta_1^2 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{400\beta_1^2 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2} - \frac{1}{4} \\
&\leq \frac{25\beta_1^2 L_y^2 L_F^2}{\beta_2^2\mu^2} + \frac{4\beta_1^2 L_{\tilde{F}}^2 L_{g_y}^2}{\mu^2} + \frac{200\beta_1^2 L_{g_y}^2 L_F^2}{\mu^2} + \frac{16\beta_1^2 L_{\tilde{F}}^2}{\alpha_1^2}\frac{L_{g_y}^2}{\mu^2} + \frac{400\beta_1^2 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2} - \frac{1}{4} \\
&\leq \frac{4\beta_1^2 L_{\tilde{F}}^2 L_{g_y}^2}{\mu^2} + \frac{200\beta_1^2 L_{g_y}^2 L_F^2}{\mu^2} + \frac{16\beta_1^2 L_{\tilde{F}}^2}{\alpha_1^2}\frac{L_{g_y}^2}{\mu^2} + \frac{400\beta_1^2 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2} - \frac{1}{8} \ .
\end{aligned} \tag{57}
$$

By letting this upper bound non-positive, we can get

$$
\begin{aligned}
&\frac{4\beta_1^2 L_{\tilde{F}}^2 L_{g_y}^2}{\mu^2} + \frac{200\beta_1^2 L_{g_y}^2 L_F^2}{\mu^2} + \frac{16\beta_1^2 L_{\tilde{F}}^2}{\alpha_1^2}\frac{L_{g_y}^2}{\mu^2} + \frac{400\beta_1^2 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2} - \frac{1}{8} \leq 0 \ , \\
&\frac{(4+16/\alpha_1^2)L_{\tilde{F}}^2 L_{g_y}^2}{\mu^2}\beta_1^2 + \frac{(200+400/\alpha_2^2)L_F^2 L_{g_y}^2}{\mu^2}\beta_1^2 - \frac{1}{8} \leq 0 \ , \\
&\beta_1^2 \leq \frac{1}{8}\frac{\mu^2}{((4+16/\alpha_1^2)L_{\tilde{F}}^2 + (200+400/\alpha_2^2)L_F^2)L_{g_y}^2} \ , \\
&\beta_1 \leq \frac{\mu}{4L_{g_y}\sqrt{((2+8/\alpha_1^2)L_{\tilde{F}}^2 + (100+200/\alpha_2^2)L_F^2)}} \ .
\end{aligned} \tag{58}
$$

Therefore, by setting $\beta_1 \leq \min\left\{\frac{\beta_2\mu}{15L_yL_F}, \frac{\mu}{4L_{g_y}\sqrt{((2+8/\alpha_1^2)L_{\tilde{F}}^2+(100+200/\alpha_2^2)L_F^2)}}\right\}$, the coefficient of $\mathbb{E}[\|\bar{u}_t\|^2]$ is non-positive.

By setting the coefficient of $\mathbb{E}[\|\bar{v}_t\|^2]$ to be non-positive, we can get

$$\begin{aligned}
&4\alpha_1\beta_1\beta_2^2\eta^4 L_{\tilde{F}}^2 + \frac{200\alpha_2\beta_1\beta_2^2\eta^4 L_{g_y}^2 L_F^2}{\mu^2} + \frac{16\eta\beta_1\beta_2^2 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{400\eta\beta_1\beta_2^2 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2} - \frac{9\eta\beta_1\beta_2 L_F^2}{2\mu} \leq 0 \ , \\
&4\alpha_1\eta^3 L_{\tilde{F}}^2 + \frac{200\alpha_2\eta^3 L_{g_y}^2 L_F^2}{\mu^2} + \frac{16 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{400 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2} - \frac{9L_F^2}{2\beta_2\mu} \leq 0 \ .
\end{aligned} \tag{59}$$

Because $\alpha_1\eta < 1, \alpha_2\eta < 1, \eta < 1, \frac{L_{g_y}^2}{\mu^2} > 1$, we can get

$$\begin{aligned}
&4\alpha_1\eta^3 L_{\tilde{F}}^2 + \frac{200\alpha_2\eta^3 L_{g_y}^2 L_F^2}{\mu^2} + \frac{16 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{400 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2} - \frac{9L_F^2}{2\beta_2\mu} \\
&\leq \frac{4L_{\tilde{F}}^2 L_{g_y}^2}{\mu^2} + \frac{200 L_{g_y}^2 L_F^2}{\mu^2} + \frac{16 L_{\tilde{F}}^2 L_{g_y}^2}{\alpha_1^2\mu^2} + \frac{400 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2} - \frac{9L_F^2}{2\beta_2\mu} \\
&= \frac{(4 + 16/\alpha_1^2) L_{\tilde{F}}^2 L_{g_y}^2}{\mu^2} + \frac{(200 + 400/\alpha_2^2) L_{g_y}^2 L_F^2}{\mu^2} - \frac{9L_F^2}{2\beta_2\mu} \ .
\end{aligned} \tag{60}$$

By letting this upper bound non-positive, we can get

$$\begin{aligned}
&\frac{(4 + 16/\alpha_1^2) L_{\tilde{F}}^2 L_{g_y}^2}{\mu^2} + \frac{(200 + 400/\alpha_2^2) L_{g_y}^2 L_F^2}{\mu^2} - \frac{9L_F^2}{2\beta_2\mu} \leq 0 \ , \\
&\beta_2 \leq \frac{9\mu L_F^2}{2((4 + 16/\alpha_1^2) L_{\tilde{F}}^2 + (200 + 400/\alpha_2^2) L_F^2) L_{g_y}^2} \ .
\end{aligned} \tag{61}$$

By setting the coefficient of $\mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}}\|_F^2]$ to be non-positive, we can get

$$\begin{aligned}
&\frac{4\eta\beta_1^3 L_{g_y}^2((11 + 32/\alpha_1^2) L_{\tilde{F}}^2 + (450 + 800/\alpha_2^2) L_F^2)}{\mu^2(1-\lambda^2)^2} - \frac{\beta_1(1-\lambda)^2}{2\alpha_1} + 4\alpha_1\beta_1^3\eta^4 L_{\tilde{F}}^2 \\
&\quad + \frac{200\alpha_2\beta_1^3\eta^4 L_{g_y}^2 L_F^2}{\mu^2} + \frac{16\eta\beta_1^3 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{400\eta\beta_1^3 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2} \leq 0 \ , \\
&\beta_1^2 \frac{4L_{g_y}^2((11 + 32/\alpha_1^2) L_{\tilde{F}}^2 + (450 + 800/\alpha_2^2) L_F^2)}{\mu^2(1-\lambda^2)^2} - \frac{(1-\lambda)^2}{2\eta\alpha_1} + 4\beta_1^2\alpha_1\eta^3 L_{\tilde{F}}^2 \\
&\quad + \frac{200\alpha_2\eta^3 L_{g_y}^2 L_F^2}{\mu^2}\beta_1^2 + \frac{16\beta_1^2 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{400\beta_1^2 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2} \leq 0 \ .
\end{aligned} \tag{62}$$

Because $\alpha_1\eta < 1, \alpha_2\eta < 1, \eta < 1, \frac{L_{g_y}^2}{\mu^2} > 1, 1-\lambda^2 < 1$, we can get

$$\begin{aligned}
&\beta_1^2 \frac{4L_{g_y}^2((11 + 32/\alpha_1^2) L_{\tilde{F}}^2 + (450 + 800/\alpha_2^2) L_F^2)}{\mu^2(1-\lambda^2)^2} - \frac{(1-\lambda)^2}{2\eta\alpha_1} + 4\beta_1^2\alpha_1\eta^3 L_{\tilde{F}}^2 \\
&\quad + \frac{200\alpha_2\eta^3 L_{g_y}^2 L_F^2}{\mu^2}\beta_1^2 + \frac{16\beta_1^2 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{400\beta_1^2 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2} \\
&\leq \beta_1^2 \frac{4L_{g_y}^2((11 + 32/\alpha_1^2) L_{\tilde{F}}^2 + (450 + 800/\alpha_2^2) L_F^2)}{\mu^2(1-\lambda^2)^2} - \frac{(1-\lambda)^2}{2\eta\alpha_1} + \frac{4\beta_1^2 L_{\tilde{F}}^2 L_{g_y}^2}{\mu^2(1-\lambda^2)^2} \\
&\quad + \frac{200 L_{g_y}^2 L_F^2}{\mu^2(1-\lambda^2)^2}\beta_1^2 + \frac{16\beta_1^2 L_{\tilde{F}}^2 L_{g_y}^2}{\alpha_1^2\mu^2(1-\lambda^2)^2} + \frac{400\beta_1^2 L_{g_y}^2 L_F^2}{\alpha_2^2\mu^2(1-\lambda^2)^2} \ .
\end{aligned} \tag{63}$$

By letting this upper bound non-positive, we can get

$$\begin{aligned}
&\beta_1^2\frac{4L_{g_y}^2((11+32/\alpha_1^2)L_{\tilde{F}}^2+(450+800/\alpha_2^2)L_F^2)}{\mu^2(1-\lambda^2)^2}-\frac{(1-\lambda)^2}{2\eta\alpha_1}+\frac{4\beta_1^2L_{\tilde{F}}^2L_{g_y}^2}{\mu^2(1-\lambda^2)^2}\\
&\quad+\frac{200L_{g_y}^2L_F^2}{\mu^2(1-\lambda^2)^2}\beta_1^2+\frac{16\beta_1^2L_{\tilde{F}}^2L_{g_y}^2}{\alpha_1^2\mu^2(1-\lambda^2)^2}+\frac{400\beta_1^2L_{g_y}^2L_F^2}{\alpha_2^2\mu^2(1-\lambda^2)^2}\leq 0\ ,\\
&\beta_1^2\frac{4L_{g_y}^2((12+36/\alpha_1^2)L_{\tilde{F}}^2+(500+900/\alpha_2^2)L_F^2)}{\mu^2(1-\lambda^2)^2}\leq\frac{(1-\lambda)^2}{2\eta\alpha_1}\ ,\\
&\beta_1^2\leq\frac{\mu^2(1-\lambda)^4}{8L_{g_y}^2((12+36/\alpha_1^2)L_{\tilde{F}}^2+(500+900/\alpha_2^2)L_F^2)}\ ,\\
&\beta_1\leq\frac{\mu(1-\lambda)^2}{4L_{g_y}\sqrt{(6+18/\alpha_1^2)L_{\tilde{F}}^2+(250+450/\alpha_2^2)L_F^2}}\ ,
\end{aligned}\tag{64}$$

where the second to last step holds due to $\eta\alpha_1<1$.

By setting the coefficient of $\mathbb{E}[\|Z_t^g-\bar{Z}_t^g\|_F^2]$ to be non-positive, we can get

$$\begin{aligned}
&\frac{4\eta\beta_1\beta_2^2L_{g_y}^2((11+32/\alpha_1^2)L_{\tilde{F}}^2+(450+800/\alpha_2^2)L_F^2)}{\mu^2(1-\lambda^2)^2}-\frac{25(1-\lambda)^2\beta_1L_F^2}{\alpha_2\mu^2}+4\alpha_1\beta_1\beta_2^2\eta^4L_{\tilde{F}}^2\\
&\quad+\frac{200\alpha_2\beta_1\beta_2^2\eta^4L_{g_y}^2L_F^2}{\mu^2}+\frac{16\eta\beta_1\beta_2^2L_{\tilde{F}}^2}{\alpha_1^2}+\frac{400\eta\beta_1\beta_2^2L_{g_y}^2L_F^2}{\alpha_2^2\mu^2}\leq 0\ ,\\
&\beta_2^2\frac{4L_{g_y}^2((11+32/\alpha_1^2)L_{\tilde{F}}^2+(450+800/\alpha_2^2)L_F^2)}{\mu^2(1-\lambda^2)^2}-\frac{25(1-\lambda)^2L_F^2}{\eta\alpha_2\mu^2}\\
&\quad+4\alpha_1\beta_2^2\eta^3L_{\tilde{F}}^2+\frac{200\alpha_2\beta_2^2\eta^3L_{g_y}^2L_F^2}{\mu^2}+\frac{16\beta_2^2L_{\tilde{F}}^2}{\alpha_1^2}+\frac{400\beta_2^2L_{g_y}^2L_F^2}{\alpha_2^2\mu^2}\leq 0\ .
\end{aligned}\tag{65}$$

Because $\alpha_1\eta<1,\alpha_2\eta<1,\eta<1,\frac{L_{g_y}^2}{\mu^2}>1,1-\lambda^2<1$, we can get

$$\begin{aligned}
&\beta_2^2\frac{4L_{g_y}^2((11+32/\alpha_1^2)L_{\tilde{F}}^2+(450+800/\alpha_2^2)L_F^2)}{\mu^2(1-\lambda^2)^2}-\frac{25(1-\lambda)^2L_F^2}{\eta\alpha_2\mu^2}\\
&\quad+4\alpha_1\beta_2^2\eta^3L_{\tilde{F}}^2+\frac{200\alpha_2\beta_2^2\eta^3L_{g_y}^2L_F^2}{\mu^2}+\frac{16\beta_2^2L_{\tilde{F}}^2}{\alpha_1^2}+\frac{400\beta_2^2L_{g_y}^2L_F^2}{\alpha_2^2\mu^2}\\
&\leq\beta_2^2\frac{4L_{g_y}^2((11+32/\alpha_1^2)L_{\tilde{F}}^2+(450+800/\alpha_2^2)L_F^2)}{\mu^2(1-\lambda^2)^2}-\frac{25(1-\lambda)^2L_F^2}{\eta\alpha_2\mu^2}\\
&\quad+\frac{4\beta_2^2L_{\tilde{F}}^2L_{g_y}^2}{\mu^2(1-\lambda^2)^2}+\frac{200\beta_2^2L_{g_y}^2L_F^2}{\mu^2(1-\lambda^2)^2}+\frac{16\beta_2^2L_{\tilde{F}}^2L_{g_y}^2}{\alpha_1^2\mu^2(1-\lambda^2)^2}+\frac{400\beta_2^2L_{g_y}^2L_F^2}{\alpha_2^2\mu^2(1-\lambda^2)^2}\ .
\end{aligned}\tag{66}$$

By letting this upper bound non-positive, we can get

$$\begin{aligned}
&\beta_2^2\frac{4L_{g_y}^2((11+32/\alpha_1^2)L_{\tilde{F}}^2+(450+800/\alpha_2^2)L_F^2)}{\mu^2(1-\lambda^2)^2}-\frac{25(1-\lambda)^2L_F^2}{\eta\alpha_2\mu^2}\\
&\quad+\frac{4\beta_2^2L_{\tilde{F}}^2L_{g_y}^2}{\mu^2(1-\lambda^2)^2}+\frac{200\beta_2^2L_{g_y}^2L_F^2}{\mu^2(1-\lambda^2)^2}+\frac{16\beta_2^2L_{\tilde{F}}^2L_{g_y}^2}{\alpha_1^2\mu^2(1-\lambda^2)^2}+\frac{400\beta_2^2L_{g_y}^2L_F^2}{\alpha_2^2\mu^2(1-\lambda^2)^2}\leq 0\ ,\\
&\beta_2^2\frac{4L_{g_y}^2((12+36/\alpha_1^2)L_{\tilde{F}}^2+(500+900/\alpha_2^2)L_F^2)}{\mu^2(1-\lambda^2)^2}\leq\frac{25(1-\lambda)^2L_F^2}{\eta\alpha_2\mu^2}\ ,\\
&\beta_2^2\leq\frac{25(1-\lambda)^4L_F^2}{4L_{g_y}^2((12+36/\alpha_1^2)L_{\tilde{F}}^2+(500+900/\alpha_2^2)L_F^2)}\ ,\\
&\beta_2\leq\frac{5(1-\lambda)^2L_F}{2L_{g_y}\sqrt{(12+36/\alpha_1^2)L_{\tilde{F}}^2+(500+900/\alpha_2^2)L_F^2}}\ ,
\end{aligned}\tag{67}$$

where the second to last step holds due to $\eta\alpha_2 < 1$. As a result, by setting $\alpha_1\eta < 1, \alpha_2\eta < 1, \eta < 1$, and

$$\beta_1 \leq \min\Big\{\frac{\beta_2\mu}{15L_yL_F}, \frac{\mu}{4L_{g_y}\sqrt{((2+8/\alpha_1^2)L_{\tilde{F}}^2 + (100+200/\alpha_2^2)L_F^2)}}, \frac{\mu(1-\lambda)^2}{4L_{g_y}\sqrt{(6+18/\alpha_1^2)L_{\tilde{F}}^2 + (250+450/\alpha_2^2)L_F^2}}\Big\} ,$$
$$\beta_2 \leq \min\Big\{\frac{9\mu L_F^2}{2((4+16/\alpha_1^2)L_{\tilde{F}}^2 + (200+400/\alpha_2^2)L_F^2)L_{g_y}^2}, \frac{5(1-\lambda)^2 L_F}{2L_{g_y}\sqrt{(12+36/\alpha_1^2)L_{\tilde{F}}^2 + (500+900/\alpha_2^2)L_F^2}}\Big\} , \tag{68}$$

we can get

$$\begin{aligned} & \mathcal{L}_{t+1} - \mathcal{L}_t \\ & \leq -\frac{\eta\beta_1}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] - \frac{\eta\beta_1 L_F^2}{2}\mathbb{E}[\|\bar{y}_t - y^*(\bar{x}_t)\|^2] + \frac{3\eta\beta_1 C_{g_{xy}}^2 C_{f_y}^2}{\mu^2}(1-\frac{\mu}{L_{g_y}})^{2J} \\ & \quad + \frac{\alpha_1\beta_1\eta^2\sigma_{\tilde{F}}^2}{2} + \frac{25\beta_1\alpha_2\eta^2 L_F^2\sigma^2}{\mu^2} + 4\beta_1\alpha_1\eta^2\sigma_{\tilde{F}}^2 + \frac{100\beta_1\alpha_2\eta^2L_F^2\sigma^2}{\mu^2} . \end{aligned} \tag{69}$$

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

$$\begin{aligned} & \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] + L_F^2\mathbb{E}[\|\bar{y}_t - y^*(\bar{x}_t)\|^2] \\ & \leq \frac{2(\mathcal{L}_0 - \mathcal{L}_T)}{\eta\beta_1 T} + \frac{6C_{g_{xy}}^2C_{f_y}^2}{\mu^2}(1-\frac{\mu}{L_{g_y}})^{2J} + 9\alpha_1\eta\sigma_{\tilde{F}}^2 + \frac{250\alpha_2\eta L_F^2\sigma^2}{\mu^2} . \end{aligned} \tag{70}$$

As for $\mathcal{L}_0$, we have

$$\begin{aligned} & \frac{1}{K}\mathbb{E}[\|Z_0^{\tilde{F}} - \bar{Z}_0^{\tilde{F}}\|_F^2] \\ & = \frac{1}{K}\mathbb{E}[\|\Delta_0^{\tilde{F}_{\tilde{\xi}_0}} - \bar{\Delta}_0^{\tilde{F}_{\tilde{\xi}_0}}\|_F^2] \\ & = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla\tilde{F}^{(k)}(x_0, y_0; \tilde{\xi}_0^{(k)}) - \frac{1}{K}\sum_{k'=1}^{K}\nabla\tilde{F}^{(k')}(x_0, y_0; \tilde{\xi}_0^{(k')})\|_F^2] \\ & = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla\tilde{F}^{(k)}(x_0, y_0; \tilde{\xi}_0^{(k)}) - \nabla\tilde{F}^{(k)}(x_0, y_0) + \nabla\tilde{F}^{(k)}(x_0, y_0) \\ & \quad - \frac{1}{K}\sum_{k'=1}^{K}\nabla\tilde{F}^{(k')}(x_0, y_0) + \frac{1}{K}\sum_{k'=1}^{K}\nabla\tilde{F}^{(k')}(x_0, y_0) - \frac{1}{K}\sum_{k'=1}^{K}\nabla\tilde{F}^{(k')}(x_0, y_0; \tilde{\xi}_0^{(k')})\|_F^2] \\ & = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla\tilde{F}^{(k)}(x_0, y_0; \tilde{\xi}_0^{(k)}) - \nabla\tilde{F}^{(k)}(x_0, y_0) + \frac{1}{K}\sum_{k'=1}^{K}\nabla\tilde{F}^{(k')}(x_0, y_0) - \frac{1}{K}\sum_{k'=1}^{K}\nabla\tilde{F}^{(k')}(x_0, y_0; \tilde{\xi}_0^{(k')})\|_F^2] \\ & = 2\sigma_{\tilde{F}}^2 , \end{aligned} \tag{71}$$

and $\frac{1}{K}\mathbb{E}[\|Z_0^g - \bar{Z}_0^g\|_F^2] \leq 2\sigma^2$, $\frac{1}{K}\mathbb{E}[\|\Delta_0^{\tilde{F}} - U_0\|_F^2] = \frac{1}{K}\mathbb{E}[\|\Delta_0^{\tilde{F}} - \Delta_0^{\tilde{F}_{\tilde{\xi}_0}}\|_F^2] \leq \sigma_{\tilde{F}}^2$, $\frac{1}{K}\mathbb{E}[\|\Delta_0^g - V_0\|_F^2] \leq \sigma^2$. Then, we can get

$$\begin{aligned} \mathcal{L}_0 & = \mathbb{E}[F(x_0)] + \frac{6\beta_1 L_F^2}{\beta_2\mu}\mathbb{E}[\|\bar{y}_0 - y^*(\bar{x}_0)\|^2] \\ & \quad + \frac{\beta_1(1-\lambda)}{2\alpha_1}\frac{1}{K}\mathbb{E}[\|Z_0^{\tilde{F}} - \bar{Z}_0^{\tilde{F}}\|_F^2] + \frac{25(1-\lambda)\beta_1 L_F^2}{\alpha_2\mu^2}\frac{1}{K}\mathbb{E}[\|Z_0^g - \bar{Z}_0^g\|_F^2] \\ & \quad + \frac{4\beta_1}{\alpha_1}\frac{1}{K}\mathbb{E}[\|\Delta_0^{\tilde{F}} - U_0\|_F^2] + \frac{100\beta_1 L_F^2}{\alpha_2\mu^2}\frac{1}{K}\mathbb{E}[\|\Delta_0^g - V_0\|_F^2] \\ & \leq \mathbb{E}[F(x_0)] + \frac{6\beta_1 L_F^2}{\beta_2\mu}\mathbb{E}[\|\bar{y}_0 - y^*(\bar{x}_0)\|^2] + \frac{5\beta_1\sigma_{\tilde{F}}^2}{\alpha_1} + \frac{150\beta_1 L_F^2\sigma^2}{\alpha_2\mu^2} . \end{aligned} \tag{72}$$

Finally, we can get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2]+L_F^2\mathbb{E}[\|\bar{y}_t-y^*(\bar{x}_t)\|^2] \\
&\leq \frac{2(\mathcal{L}_0-\mathcal{L}_T)}{\eta\beta_1 T}+\frac{6C_{g_{xy}}^2C_{f_y}^2}{\mu^2}(1-\frac{\mu}{L_{g_y}})^{2J}+9\alpha_1\eta\sigma_{\tilde{F}}^2+\frac{250\alpha_2\eta L_F^2\sigma^2}{\mu^2} \\
&\leq \frac{2(F(x_0)-F(x_*))}{\eta\beta_1 T}+\frac{12L_F^2}{\beta_2\mu\eta T}\mathbb{E}[\|\bar{y}_0-y^*(\bar{x}_0)\|^2]+\frac{10\sigma_{\tilde{F}}^2}{\alpha_1\eta T}+\frac{300L_F^2\sigma^2}{\alpha_2\mu^2\eta T} \\
&\quad+\frac{6C_{g_{xy}}^2C_{f_y}^2}{\mu^2}(1-\frac{\mu}{L_{g_y}})^{2J}+9\alpha_1\eta\sigma_{\tilde{F}}^2+\frac{250\alpha_2\eta L_F^2\sigma^2}{\mu^2}\,.
\end{aligned} \tag{73}$$

□

A.2.6 Proof of Corollary 1

Proof. By setting $T=O(\frac{1}{\epsilon^2(1-\lambda)^2})$, $\eta=O(\epsilon)$, $J=O(\log\frac{1}{\epsilon})$, $\beta_1=O((1-\lambda)^2)$, $\beta_2=O((1-\lambda)^2)$, $\alpha_1=O(1)$, $\alpha_2=O(1)$, we can get

$$\begin{aligned}
&\frac{2(F(x_0)-F(x_*))}{\eta\beta_1 T}=O(\epsilon),\quad \frac{12L_F^2}{\beta_2\eta T\mu}\|\bar{y}_0-y^*(\bar{x}_0)\|^2=O(\epsilon),\quad \frac{12\sigma_{\tilde{F}}^2}{\alpha_1\eta T}=O(\epsilon),\quad \frac{400L_F^2\sigma^2}{\alpha_2\eta T\mu^2}=O(\epsilon),\\
&\frac{6C_{g_{xy}}^2C_{f_y}^2}{\mu^2}(1-\frac{\mu}{L_{g_y}})^{2J}=O(\epsilon),\quad 10\alpha_1\eta\sigma_{\tilde{F}}^2=O(\epsilon),\quad \frac{300\alpha_2\eta L_F^2\sigma^2}{\mu^2}=O(\epsilon).
\end{aligned} \tag{74}$$

Additionally, since the communication is conducted in each iteration, the communication complexity is equal to the number of iterations. Thus, it is $O(\frac{1}{\epsilon^2(1-\lambda)^2})$. When computing the stochastic gradient, the batch size is $O(1)$. Thus, the gradient complexity is $O(\frac{1}{\epsilon^2(1-\lambda)^2})$. When computing the stochastic hypergradient, the Jacobian-vector product is conducted for one time in each iteration and the Hessian-vector product is performed for J times in each iteration. Thus, the Jacobian-vector product complexity is $O(\frac{1}{\epsilon^2(1-\lambda)^2})$, and the Hessian-vector product complexity is $J\times T=O(\frac{1}{\epsilon^2(1-\lambda)^2}\log\frac{1}{\epsilon})=\tilde{O}(\frac{1}{\epsilon^2(1-\lambda)^2})$.

□

A.3 Proof of Theorem 2

To study the convergence of MDBO under Assumption 6, we employ a little different initialization condition, which is demonstrated in Algorithm 3.

Algorithm 3 MDBO

Input: $x_0^{(k)}=x_0, y_0^{(k)}=y_0, \eta>0, \alpha_1>0, \alpha_2>0, \beta_1>0, \beta_2>0$.

1: $U_{-1}=0$, $V_{-1}=0$, $Z_{-1}^{\tilde{F}}=0$, $Z_{-1}^{g}=0$,
2: **for** $t=0,\cdots,T-1$ **do**
3: $U_t=(1-\alpha_1\eta)U_{t-1}+\alpha_1\eta\Delta_t^{\tilde{F}_{\tilde{\xi}_t}}$,
$V_t=(1-\alpha_2\eta)V_{t-1}+\alpha_2\eta\Delta_t^{g_{\zeta_t}}$,
4: $Z_t^{\tilde{F}}=Z_{t-1}^{\tilde{F}}W+U_t-U_{t-1}$,
$Z_t^{g}=Z_{t-1}^{g}W+V_t-V_{t-1}$,
5: $X_{t+1}=X_t-\eta X_t(I-W)-\beta_1\eta Z_t^{\tilde{F}}$,
$Y_{t+1}=Y_t-\eta Y_t(I-W)-\beta_2\eta Z_t^{g}$,
6: **end for**

A.3.1 Characterization of Gradient Norm

Lemma 14. *Given Assumptions 1-6, the following inequality holds.*

$$
\begin{aligned}
&\hat{C}_{f_x}^2 \triangleq \mathbb{E}[\|\nabla_x f^{(k)}(x,y;\xi)\|^2] \leq \sigma^2 + C_{f_x}^2 \ , \\
&\hat{C}_{g_{xy}}^2 \triangleq \mathbb{E}[\|\nabla_{xy}^2 g^{(k)}(x,y;\zeta)\|^2] \leq \sigma^2 + C_{g_{xy}}^2 \ , \\
&\hat{C}_{f_y}^2 \triangleq \mathbb{E}[\|\nabla_y f^{(k)}(x,y;\xi)\|^2] \leq \sigma^2 + C_{f_y}^2 \ , \\
&\hat{C}_{\tilde{F}}^2 \triangleq \mathbb{E}[\|\nabla \tilde{F}^{(k)}(x,y;\tilde{\xi})\|^2] \leq 2\hat{C}_{f_x}^2 + \frac{2\hat{C}_{g_{xy}}^2\hat{C}_{f_y}^2}{\mu^2} \ , \\
&\hat{C}_{g_y}^2 \triangleq \mathbb{E}[\|\nabla_y g^{(k)}(x,y;\zeta)\|^2] \leq \sigma^2 + C_{g_y}^2 \ , \\
&\|U_t\|_F^2 \leq K\hat{C}_{\tilde{F}}^2, \quad \|V_t\|_F^2 \leq K\hat{C}_{g_y}^2 \ .
\end{aligned} \tag{75}
$$

Proof. Based on Assumptions 1-6, it is easy to get

$$
\begin{aligned}
&\hat{C}_{f_x}^2 \triangleq \mathbb{E}[\|\nabla_x f^{(k)}(x,y;\xi)\|^2] = \mathbb{E}[\|\nabla_x f^{(k)}(x,y;\xi) - \nabla_x f^{(k)}(x,y) + \nabla_x f^{(k)}(x,y)\|^2] \leq \sigma^2 + C_{f_x}^2 \ , \\
&\hat{C}_{g_{xy}}^2 \triangleq \mathbb{E}[\|\nabla_{xy}^2 g^{(k)}(x,y;\zeta)\|^2] = \mathbb{E}[\|\nabla_{xy}^2 g^{(k)}(x,y;\zeta) - \nabla_{xy}^2 g^{(k)}(x,y) + \nabla_{xy}^2 g^{(k)}(x,y)\|^2] \leq \sigma^2 + C_{g_{xy}}^2 \ , \\
&\hat{C}_{f_y}^2 \triangleq \mathbb{E}[\|\nabla_y f^{(k)}(x,y;\xi)\|^2] = \mathbb{E}[\|\nabla_y f^{(k)}(x,y;\xi) - \nabla_y f^{(k)}(x,y) + \nabla_y f^{(k)}(x,y)\|^2] \leq \sigma^2 + C_{f_y}^2 \ , \\
&\hat{C}_{g_y}^2 \triangleq \mathbb{E}[\|\nabla_y g^{(k)}(x,y;\zeta)\|^2] = \mathbb{E}[\|\nabla_y g^{(k)}(x,y;\zeta) - \nabla_y g^{(k)}(x,y) + \nabla_y g^{(k)}(x,y)\|^2] \leq \sigma^2 + C_{g_y}^2 \ .
\end{aligned} \tag{76}
$$

Then, we can get

$$
\begin{aligned}
&\hat{C}_{\tilde{F}}^2 \triangleq \mathbb{E}[\|\nabla \tilde{F}^{(k)}(x,y;\tilde{\xi})\|^2] \\
&= \mathbb{E}[\|\nabla_x f^{(k)}(x,y;\xi) - \nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\nabla_y f^{(k)}(x,y;\xi)\|^2] \\
&\leq 2\mathbb{E}[\|\nabla_x f^{(k)}(x,y;\xi)\|^2] + 2\mathbb{E}[\|\nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\nabla_y f^{(k)}(x,y;\xi)\|^2] \\
&\leq 2\mathbb{E}[\|\nabla_x f^{(k)}(x,y;\xi)\|^2] + 2\mathbb{E}\Big[\|\nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\|^2 \times \|\frac{J}{L_{g_y}}\prod_{j=1}^{\tilde{J}}(I - \frac{1}{L_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\|^2 \times \|\nabla_y f^{(k)}(x,y;\xi)\|^2\Big] \\
&\leq 2\hat{C}_{f_x}^2 + \frac{2\hat{C}_{g_{xy}}^2\hat{C}_{f_y}^2}{\mu^2} \ ,
\end{aligned} \tag{77}
$$

where we use Lemma 2 in the last step.

Then, we use induction approach to prove $\|U_t\|_F^2$. At first, when $t=0$, $\|U_0\|_F^2 = \|\alpha_1\eta\Delta_0^{\tilde{F}_{\tilde{\xi}_0}}\|_F^2 \leq K\hat{C}_{\tilde{F}}^2$ since $\alpha_1\eta \leq 1$. Then, assuming $\|U_t\|_F^2 \leq K\hat{C}_{\tilde{F}}^2$, we can get

$$
\begin{aligned}
\|U_{t+1}\|_F &= \|(1-\alpha_1\eta)U_t + \alpha_1\eta\Delta_{t+1}^{\tilde{F}_{\tilde{\xi}_{t+1}}}\|_F \\
&\leq (1-\alpha_1\eta)\|U_t\|_F + \alpha_1\eta\|\Delta_{t+1}^{\tilde{F}_{\tilde{\xi}_{t+1}}}\|_F \\
&\leq (1-\alpha_1\eta)\sqrt{K}\hat{C}_{\tilde{F}} + \alpha_1\eta\sqrt{K}\hat{C}_{\tilde{F}} \\
&= \sqrt{K}\hat{C}_{\tilde{F}} \ ,
\end{aligned} \tag{78}
$$

where the second step holds due to the convexity of Frobenius norm. This completes the proof for $\|U_t\|_F^2$. Similarly, we can prove the claim for $\|V_t\|_F^2$. □

A.3.2 Characterization of Consensus Errors

Lemma 15. *Given Assumptions 1-6, the following inequality holds.*

$$\frac{1}{K}\mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}}\|_F^2] \leq \frac{2\alpha_1^2\eta^2\hat{C}_{\tilde{F}}^2}{(1-\lambda)^2} \ . \tag{79}$$

Proof.

$$\begin{aligned}
&\frac{1}{K}\mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}}\|_F^2] \\
&= \frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}}W + U_t - U_{t-1} - \bar{Z}_{t-1}^{\tilde{F}} - \bar{U}_t + \bar{U}_{t-1}\|_F^2] \\
&\overset{(s_1)}{\leq} \frac{1}{K}\lambda\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|U_t - U_{t-1} - \bar{U}_t + \bar{U}_{t-1}\|_F^2] \\
&\leq \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|U_t - U_{t-1}\|_F^2] \\
&= \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|(1-\alpha_1\eta)U_{t-1} + \alpha_1\eta\Delta_t^{\tilde{F}_{\tilde{\xi}_t}} - U_{t-1}\|_F^2] \\
&= \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\| - \alpha_1\eta U_{t-1} + \alpha_1\eta\Delta_t^{\tilde{F}_{\tilde{\xi}_t}}\|_F^2] \\
&\overset{(s_2)}{\leq} \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{2\alpha_1^2\eta^2\hat{C}_{\tilde{F}}^2}{1-\lambda} \\
&\overset{(s_3)}{\leq} \frac{2\alpha_1^2\eta^2\hat{C}_{\tilde{F}}^2}{1-\lambda}\sum_{i=-1}^{t-1}\lambda^{t-1-i} \\
&\leq \frac{2\alpha_1^2\eta^2\hat{C}_{\tilde{F}}^2}{(1-\lambda)^2} \ ,
\end{aligned} \tag{80}$$

where (s_1) holds due to Lemma 34 with $a = \frac{1-\lambda}{\lambda}$, (s_2) holds due to Lemma 14, (s_3) holds due to recursive expansion and the initialisation conditions: $U_{-1} = 0, Z_{-1}^{\tilde{F}} = 0$.

□

Lemma 16. *Given Assumptions 1-5, the following inequality holds.*

$$\frac{1}{K}\mathbb{E}[\|Z_t^{g} - \bar{Z}_t^{g}\|_F^2] \leq \frac{2\alpha_2^2\eta^2\hat{C}_{g_y}^2}{(1-\lambda)^2} \ . \tag{81}$$

This lemma can be proved by following Lemma 15.

Lemma 17. *Given Assumptions 1-5, the following inequality holds.*

$$\frac{1}{K}\mathbb{E}[\|X_{t+1} - \bar{X}_{t+1}\|_F^2] \leq \frac{8\alpha_1^2\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2}{(1-\lambda)^4} \ . \tag{82}$$

Proof.

$$\begin{aligned}
&\frac{1}{K}\mathbb{E}[\|X_{t+1}-\bar{X}_{t+1}\|_F^2] \\
&\overset{(s_0)}{=} \frac{1}{K}\mathbb{E}[\|X_t-\eta X_t(I-W)-\beta_1\eta Z_t^{\tilde{F}}-\bar{X}_t+\beta_1\eta\bar{Z}_t^{\tilde{F}}\|_F^2] \\
&= \frac{1}{K}\mathbb{E}[\|(1-\eta)(X_t-\bar{X}_t)+\eta(X_tW-\bar{X}_t-\beta_1 Z_t^{\tilde{F}}+\beta_1\bar{Z}_t^{\tilde{F}})\|_F^2] \\
&\overset{(s_1)}{\leq} (1-\eta)\frac{1}{K}\mathbb{E}[\|X_t-\bar{X}_t\|_F^2]+\eta\frac{1}{K}\mathbb{E}[\|X_tW+\beta_1 Z_t^{\tilde{F}}-\bar{X}_t-\beta_1\bar{Z}_t^{\tilde{F}}\|_F^2] \\
&\overset{(s_2)}{\leq} (1-\eta)\frac{1}{K}\mathbb{E}[\|X_t-\bar{X}_t\|_F^2]+\frac{\eta(1+\lambda^2)}{2\lambda^2}\frac{1}{K}\mathbb{E}[\|X_tW-\bar{X}_t\|_F^2]+\frac{2\eta\beta_1^2}{1-\lambda^2}\frac{1}{K}\mathbb{E}[\|Z_t^{\tilde{F}}-\bar{Z}_t^{\tilde{F}}\|_F^2] \\
&\overset{(s_3)}{\leq} \Big(1-\frac{\eta(1-\lambda^2)}{2}\Big)\frac{1}{K}\mathbb{E}[\|X_t-\bar{X}_t\|_F^2]+\frac{2\eta\beta_1^2}{1-\lambda^2}\frac{1}{K}\mathbb{E}[\|Z_t^{\tilde{F}}-\bar{Z}_t^{\tilde{F}}\|_F^2] \\
&\overset{(s_4)}{\leq} \Big(1-\frac{\eta(1-\lambda^2)}{2}\Big)\frac{1}{K}\mathbb{E}[\|X_t-\bar{X}_t\|_F^2]+\frac{2\eta\beta_1^2}{1-\lambda^2}\frac{2\alpha_1^2\eta^2\hat{C}_{\tilde{F}}^2}{(1-\lambda)^2} \\
&\leq \frac{2\eta\beta_1^2}{1-\lambda^2}\frac{2\alpha_1^2\eta^2\hat{C}_{\tilde{F}}^2}{(1-\lambda)^2}\sum_{i=0}^{t}\Big(1-\frac{\eta(1-\lambda^2)}{2}\Big)^{t-i} \\
&\leq \frac{8\alpha_1^2\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2}{(1-\lambda)^4} \ ,
\end{aligned} \tag{83}$$

where (s_0) holds due to $X_t(I-W)\mathbf{1}=0$, (s_1) holds due to Lemma 34 with $a=\frac{\eta}{1-\eta}$, (s_2) holds due to Lemma 34 with $a=\frac{1-\lambda^2}{2\lambda^2}$, (s_3) holds due to $\|X_tW-\bar{X}_t\|_F^2\leq\lambda^2\|X_t-\bar{X}_t\|_F^2$, (s_4) holds due to Lemma 15. □

Lemma 18. *Given Assumptions 1-5, the following inequality holds.*

$$\frac{1}{K}\mathbb{E}[\|Y_{t+1}-\bar{Y}_{t+1}\|_F^2]\leq\frac{8\alpha_2^2\beta_2^2\eta^2\hat{C}_{g_y}^2}{(1-\lambda)^4} \ . \tag{84}$$

This lemma can be proved by following Lemma 17.

Lemma 19. *Given Assumptions 1-5, the following inequality holds.*

$$\frac{1}{K}\mathbb{E}[\|X_{t+1}-X_t\|_F^2]\leq\frac{72\alpha_1^2\beta_1^2\eta^4\hat{C}_{\tilde{F}}^2}{(1-\lambda)^4}+4\eta^2\beta_1^2\mathbb{E}[\|\bar{u}_t\|^2] \ . \tag{85}$$

Proof.

$$\begin{aligned}
&\frac{1}{K}\mathbb{E}[\|X_{t+1}-X_t\|_F^2] \\
&= \frac{1}{K}\mathbb{E}[\|X_t-\eta X_t(I-W)-\beta_1\eta Z_t^{\tilde{F}}-X_t\|_F^2] \\
&= \frac{1}{K}\mathbb{E}[\|\eta X_t(W-I)-\beta_1\eta Z_t^{\tilde{F}}\|_F^2] \\
&\leq 2\eta^2\frac{1}{K}\mathbb{E}[\|X_t(W-I)\|_F^2]+2\eta^2\beta_1^2\frac{1}{K}\mathbb{E}[\|Z_t^{\tilde{F}}\|_F^2] \\
&= 2\eta^2\frac{1}{K}\mathbb{E}[\|(X_t-\bar{X}_t)(W-I)\|_F^2]+2\eta^2\beta_1^2\frac{1}{K}\mathbb{E}[\|Z_t^{\tilde{F}}-\bar{Z}_t^{\tilde{F}}+\bar{Z}_t^{\tilde{F}}\|_F^2] \\
&\overset{(s_1)}{\leq} 8\eta^2\frac{1}{K}\mathbb{E}[\|X_t-\bar{X}_t\|_F^2]+4\eta^2\beta_1^2\frac{1}{K}\mathbb{E}[\|Z_t^{\tilde{F}}-\bar{Z}_t^{\tilde{F}}\|_F^2]+4\eta^2\beta_1^2\frac{1}{K}\mathbb{E}[\|\bar{Z}_t^{\tilde{F}}\|_F^2] \\
&\overset{(s_1)}{\leq} \frac{64\alpha_1^2\beta_1^2\eta^4\hat{C}_{\tilde{F}}^2}{(1-\lambda)^4}+\frac{8\alpha_1^2\beta_1^2\eta^4\hat{C}_{\tilde{F}}^2}{(1-\lambda)^2}+4\eta^2\beta_1^2\mathbb{E}[\|\bar{u}_t\|^2] \\
&\leq \frac{72\alpha_1^2\beta_1^2\eta^4\hat{C}_{\tilde{F}}^2}{(1-\lambda)^4}+4\eta^2\beta_1^2\mathbb{E}[\|\bar{u}_t\|^2] \ ,
\end{aligned} \tag{86}$$

where (s_1) holds due to $\|AB\|_F \leq \|A\|_2\|B\|_F$ and $\|I-W\|_2 \leq 2$, (s_1) holds due to Lemma 15 and Lemma 17.

□

Lemma 20. *Given Assumptions 1-5, the following inequality holds.*

$$\frac{1}{K}\mathbb{E}[\|Y_{t+1}-Y_t\|_F^2] \leq \frac{72\alpha_2^2\beta_2^2\eta^4\hat{C}_{g_y}^2}{(1-\lambda)^4} + 4\eta^2\beta_2^2\mathbb{E}[\|\bar{v}_t\|^2] \; . \tag{87}$$

A.3.3 Characterization of Gradient Estimators

Lemma 21. *Given Assumptions 1-5, the following inequality holds.*

$$\begin{aligned}\mathbb{E}[\|(\Delta_t^{\tilde{F}}-U_t)\frac{1}{K}\mathbf{1}\|^2] &\leq (1-\alpha_1\eta)\mathbb{E}[\|(\Delta_{t-1}^{\tilde{F}}-U_{t-1})\frac{1}{K}\mathbf{1}\|^2] + \frac{4\eta\beta_1^2L_{\tilde{F}}^2}{\alpha_1}\mathbb{E}[\|\bar{u}_{t-1}\|^2] + \frac{4\eta\beta_2^2L_{\tilde{F}}^2}{\alpha_1}\mathbb{E}[\|\bar{v}_{t-1}\|^2] \\ &\quad + \frac{72\alpha_1\beta_1^2\eta^3\hat{C}_{\tilde{F}}^2L_{\tilde{F}}^2}{(1-\lambda)^4} + \frac{72\alpha_2^2\beta_2^2\eta^3\hat{C}_{g_y}^2L_{\tilde{F}}^2}{\alpha_1(1-\lambda)^4} + \frac{\alpha_1^2\eta^2\sigma_{\tilde{F}}^2}{K} \; .\end{aligned} \tag{88}$$

Proof.

$$\begin{aligned}&\mathbb{E}[\|(\Delta_t^{\tilde{F}}-U_t)\frac{1}{K}\mathbf{1}\|^2] \\ &= \mathbb{E}[\|(\Delta_t^{\tilde{F}}-(1-\alpha_1\eta)U_{t-1}-\alpha_1\eta\Delta_t^{\tilde{F}_{\tilde{\xi}_t}})\frac{1}{K}\mathbf{1}\|^2] \\ &= \mathbb{E}[\|((1-\alpha_1\eta)(\Delta_{t-1}^{\tilde{F}}-U_{t-1})+(1-\alpha_1\eta)(\Delta_t^{\tilde{F}}-\Delta_{t-1}^{\tilde{F}})+\alpha_1\eta(\Delta_t^{\tilde{F}}-\Delta_t^{\tilde{F}_{\tilde{\xi}_t}}))\frac{1}{K}\mathbf{1}\|^2] \\ &\overset{(s_1)}{=} (1-\alpha_1\eta)^2\mathbb{E}[\|(\Delta_{t-1}^{\tilde{F}}-U_{t-1})\frac{1}{K}\mathbf{1}+(\Delta_t^{\tilde{F}}-\Delta_{t-1}^{\tilde{F}})\frac{1}{K}\mathbf{1}\|^2] + \alpha_1^2\eta^2\mathbb{E}[\|(\Delta_t^{\tilde{F}}-\Delta_t^{\tilde{F}_{\tilde{\xi}_t}})\frac{1}{K}\mathbf{1}\|^2] \\ &\overset{(s_2)}{\leq} (1-\alpha_1\eta)\mathbb{E}[\|(\Delta_{t-1}^{\tilde{F}}-U_{t-1})\frac{1}{K}\mathbf{1}\|^2] + \frac{1}{\alpha_1\eta}\frac{1}{K}\mathbb{E}[\|\Delta_t^{\tilde{F}}-\Delta_{t-1}^{\tilde{F}}\|_F^2] + \frac{\alpha_1^2\eta^2\sigma_{\tilde{F}}^2}{K} \\ &\overset{(s_3)}{\leq} (1-\alpha_1\eta)\mathbb{E}[\|(\Delta_{t-1}^{\tilde{F}}-U_{t-1})\frac{1}{K}\mathbf{1}\|^2] + \frac{L_{\tilde{F}}^2}{\alpha_1\eta}\frac{1}{K}\mathbb{E}[\|X_t-X_{t-1}\|_F^2] + \frac{L_{\tilde{F}}^2}{\alpha_1\eta}\frac{1}{K}\mathbb{E}[\|Y_t-Y_{t-1}\|_F^2] + \frac{\alpha_1^2\eta^2\sigma_{\tilde{F}}^2}{K} \\ &\overset{(s_4)}{\leq} (1-\alpha_1\eta)\mathbb{E}[\|(\Delta_{t-1}^{\tilde{F}}-U_{t-1})\frac{1}{K}\mathbf{1}\|^2] + \frac{L_{\tilde{F}}^2}{\alpha_1\eta}\frac{72\alpha_1^2\beta_1^2\eta^4\hat{C}_{\tilde{F}}^2}{(1-\lambda)^4} + \frac{L_{\tilde{F}}^2}{\alpha_1\eta}4\eta^2\beta_1^2\mathbb{E}[\|\bar{u}_{t-1}\|^2] \\ &\quad + \frac{L_{\tilde{F}}^2}{\alpha_1\eta}\frac{72\alpha_2^2\beta_2^2\eta^4\hat{C}_{g_y}^2}{(1-\lambda)^4} + \frac{L_{\tilde{F}}^2}{\alpha_1\eta}4\eta^2\beta_2^2\mathbb{E}[\|\bar{v}_{t-1}\|^2] + \frac{\alpha_1^2\eta^2\sigma_{\tilde{F}}^2}{K} \\ &\leq (1-\alpha_1\eta)\mathbb{E}[\|(\Delta_{t-1}^{\tilde{F}}-U_{t-1})\frac{1}{K}\mathbf{1}\|^2] + \frac{4\eta\beta_1^2L_{\tilde{F}}^2}{\alpha_1}\mathbb{E}[\|\bar{u}_{t-1}\|^2] + \frac{4\eta\beta_2^2L_{\tilde{F}}^2}{\alpha_1}\mathbb{E}[\|\bar{v}_{t-1}\|^2] \\ &\quad + \frac{72\alpha_1\beta_1^2\eta^3\hat{C}_{\tilde{F}}^2L_{\tilde{F}}^2}{(1-\lambda)^4} + \frac{72\alpha_2^2\beta_2^2\eta^3\hat{C}_{g_y}^2L_{\tilde{F}}^2}{\alpha_1(1-\lambda)^4} + \frac{\alpha_1^2\eta^2\sigma_{\tilde{F}}^2}{K} \; ,\end{aligned} \tag{89}$$

where (s_1) holds due to $\Delta_t^{\tilde{F}} = \mathbb{E}[\Delta_t^{\tilde{F}_{\tilde{\xi}_t}}]$, (s_2) holds due to Lemma 4 and Lemma 34 with $a = \frac{\alpha_1\eta}{1-\alpha_1\eta}$, (s_3) holds due to Lemma 5, (s_4) holds due to Lemma 19 and Lemma 20. □

Lemma 22. *Given Assumptions 1-5, the following inequality holds.*

$$\begin{aligned}\mathbb{E}[\|(\Delta_t^{g}-V_t)\frac{1}{K}\mathbf{1}\|^2] &\leq (1-\alpha_2\eta)\mathbb{E}[\|(\Delta_{t-1}^{g}-V_{t-1})\frac{1}{K}\mathbf{1}\|^2] + \frac{4\eta\beta_1^2L_{g_y}^2}{\alpha_2}\mathbb{E}[\|\bar{u}_{t-1}\|^2] + \frac{4\eta\beta_2^2L_{g_y}^2}{\alpha_2}\mathbb{E}[\|\bar{v}_{t-1}\|^2] \\ &\quad + \frac{72\alpha_1^2\beta_1^2\eta^3\hat{C}_{\tilde{F}}^2L_{g_y}^2}{\alpha_2(1-\lambda)^4} + \frac{72\alpha_2\beta_2^2\eta^3\hat{C}_{g_y}^2L_{g_y}^2}{(1-\lambda)^4} + \frac{\alpha_2^2\eta^2\sigma^2}{K} \; .\end{aligned} \tag{90}$$

Then, we introduce the following potential function:

$$\begin{aligned}\mathcal{L}_{t+1} &= F(x_{t+1}) + \frac{6\beta_1 L_F^2}{\beta_2\mu}\mathbb{E}[\|\bar{y}_{t+1} - y^*(\bar{x}_{t+1})\|^2] \\ &\quad + \frac{3\beta_1}{\alpha_1}\frac{1}{K}\mathbb{E}[\|\Delta_{t+1}^{\tilde{F}} - U_{t+1}\|_F^2] + \frac{50\beta_1 L_F^2}{\alpha_2\mu^2}\frac{1}{K}\mathbb{E}[\|\Delta_{t+1}^{g} - V_{t+1}\|_F^2]\ .\end{aligned} \tag{91}$$

Based on Lemmas 17, 18, 21, 22, we can get

$$\begin{aligned}
&\mathcal{L}_{t+1} - \mathcal{L}_t \\
&\leq -\frac{\eta\beta_1}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] - \frac{\eta\beta_1 L_F^2}{2}\mathbb{E}[\|\bar{y}_t - y^*(\bar{x}_t)\|^2] + \frac{3\eta\beta_1 C_{g_{xy}}^2 C_{f_y}^2}{\mu^2}(1-\frac{\mu}{L_{g_y}})^{2J} \\
&\quad + (\frac{25\eta\beta_1^2 L_{g_y}^2}{6\beta_2\mu}w_1 + w_6\frac{L_{\tilde{F}}^2}{\alpha_1\eta}4\eta^2\beta_1^2 + w_7\frac{L_{g_y}^2}{\alpha_2\eta}4\eta^2\beta_1^2 - \frac{\eta\beta_1}{4})\mathbb{E}[\|\bar{u}_t\|^2] \\
&\quad + (w_6\frac{L_{\tilde{F}}^2}{\alpha_1\eta}4\eta^2\beta_2^2 + w_7\frac{L_{g_y}^2}{\alpha_2\eta}4\eta^2\beta_2^2 - \frac{3\eta\beta_2^2}{4}w_1)\mathbb{E}[\|\bar{v}_t\|^2] \\
&\quad + 3\eta\beta_1 L_{\tilde{F}}^2\frac{8\alpha_1^2\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2}{(1-\lambda)^4} + 3\eta\beta_1 L_{\tilde{F}}^2\frac{8\alpha_2^2\beta_2^2\eta^2\hat{C}_{g_y}^2}{(1-\lambda)^4} + w_1\frac{25\beta_2\eta L_{g_y}^2}{3\mu}\frac{8\alpha_1^2\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2}{(1-\lambda)^4} + w_1\frac{25\beta_2\eta L_{g_y}^2}{3\mu}\frac{8\alpha_2^2\beta_2^2\eta^2\hat{C}_{g_y}^2}{(1-\lambda)^4} \\
&\quad + w_6\frac{L_{\tilde{F}}^2}{\alpha_1\eta}\frac{72\alpha_1^2\beta_1^2\eta^4\hat{C}_{\tilde{F}}^2}{(1-\lambda)^4} + w_6\frac{L_{\tilde{F}}^2}{\alpha_1\eta}\frac{72\alpha_2^2\beta_2^2\eta^4\hat{C}_{g_y}^2}{(1-\lambda)^4} + w_7\frac{L_{g_y}^2}{\alpha_2\eta}\frac{72\alpha_1^2\beta_1^2\eta^4\hat{C}_{\tilde{F}}^2}{(1-\lambda)^4} + w_7\frac{L_{g_y}^2}{\alpha_2\eta}\frac{72\alpha_2^2\beta_2^2\eta^4\hat{C}_{g_y}^2}{(1-\lambda)^4} \\
&\quad + w_6\frac{\alpha_1^2\eta^2\sigma_{\tilde{F}}^2}{K} + w_7\frac{\alpha_2^2\eta^2\sigma^2}{K} \\
&\leq -\frac{\eta\beta_1}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] - \frac{\eta\beta_1 L_F^2}{2}\mathbb{E}[\|\bar{y}_t - y^*(\bar{x}_t)\|^2] + \frac{3\eta\beta_1 C_{g_{xy}}^2 C_{f_y}^2}{\mu^2}(1-\frac{\mu}{L_{g_y}})^{2J} \\
&\quad + (\frac{25\eta\beta_1^2 L_{g_y}^2}{6\beta_2\mu}\frac{6\beta_1 L_F^2}{\beta_2\mu} + \frac{3\beta_1}{\alpha_1}\frac{L_{\tilde{F}}^2}{\alpha_1\eta}4\eta^2\beta_1^2 + \frac{150\beta_1 L_F^2}{3\alpha_2\mu^2}\frac{L_{g_y}^2}{\alpha_2\eta}4\eta^2\beta_1^2 - \frac{\eta\beta_1}{4})\mathbb{E}[\|\bar{u}_t\|^2] \\
&\quad + (\frac{3\beta_1}{\alpha_1}\frac{L_{\tilde{F}}^2}{\alpha_1\eta}4\eta^2\beta_2^2 + \frac{150\beta_1 L_F^2}{3\alpha_2\mu^2}\frac{L_{g_y}^2}{\alpha_2\eta}4\eta^2\beta_2^2 - \frac{3\eta\beta_2^2}{4}\frac{6\beta_1 L_F^2}{\beta_2\mu})\mathbb{E}[\|\bar{v}_t\|^2] \\
&\quad + 3\eta\beta_1 L_{\tilde{F}}^2\frac{8\alpha_1^2\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2}{(1-\lambda)^4} + 3\eta\beta_1 L_{\tilde{F}}^2\frac{8\alpha_2^2\beta_2^2\eta^2\hat{C}_{g_y}^2}{(1-\lambda)^4} + \frac{6\beta_1 L_F^2}{\beta_2\mu}\frac{25\beta_2\eta L_{g_y}^2}{3\mu}\frac{8\alpha_1^2\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2}{(1-\lambda)^4} + \frac{6\beta_1 L_F^2}{\beta_2\mu}\frac{25\beta_2\eta L_{g_y}^2}{3\mu}\frac{8\alpha_2^2\beta_2^2\eta^2\hat{C}_{g_y}^2}{(1-\lambda)^4} \\
&\quad + \frac{3\beta_1}{\alpha_1}\frac{L_{\tilde{F}}^2}{\alpha_1\eta}\frac{72\alpha_1^2\beta_1^2\eta^4\hat{C}_{\tilde{F}}^2}{(1-\lambda)^4} + \frac{3\beta_1}{\alpha_1}\frac{L_{\tilde{F}}^2}{\alpha_1\eta}\frac{72\alpha_2^2\beta_2^2\eta^4\hat{C}_{g_y}^2}{(1-\lambda)^4} + \frac{150\beta_1 L_F^2}{3\alpha_2\mu^2}\frac{L_{g_y}^2}{\alpha_2\eta}\frac{72\alpha_1^2\beta_1^2\eta^4\hat{C}_{\tilde{F}}^2}{(1-\lambda)^4} + \frac{150\beta_1 L_F^2}{3\alpha_2\mu^2}\frac{L_{g_y}^2}{\alpha_2\eta}\frac{72\alpha_2^2\beta_2^2\eta^4\hat{C}_{g_y}^2}{(1-\lambda)^4} \\
&\quad + \frac{3\beta_1}{\alpha_1}\frac{\alpha_1^2\eta^2\sigma_{\tilde{F}}^2}{K} + \frac{150\beta_1 L_F^2}{3\alpha_2\mu^2}\frac{\alpha_2^2\eta^2\sigma^2}{K} \\
&\leq -\frac{\eta\beta_1}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] - \frac{\eta\beta_1 L_F^2}{2}\mathbb{E}[\|\bar{y}_t - y^*(\bar{x}_t)\|^2] + \frac{3\eta\beta_1 C_{g_{xy}}^2 C_{f_y}^2}{\mu^2}(1-\frac{\mu}{L_{g_y}})^{2J} \\
&\quad + (\frac{25\eta\beta_1^3 L_F^2 L_y^2}{\beta_2^2\mu^2} + \frac{12\eta\beta_1^3 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{200\eta\beta_1^3 L_F^2 L_{g_y}^2}{\alpha_2^2\mu^2} - \frac{\eta\beta_1}{4})\mathbb{E}[\|\bar{u}_t\|^2] \\
&\quad + (\frac{12\eta\beta_1\beta_2^2 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{200\eta\beta_1\beta_2^2 L_F^2 L_{g_y}^2}{\alpha_2^2\mu^2} - \frac{9\eta\beta_1\beta_2 L_F^2}{2\mu})\mathbb{E}[\|\bar{v}_t\|^2] \\
&\quad + \frac{24\alpha_1^2\beta_1^3\eta^3\hat{C}_{\tilde{F}}^2 L_{\tilde{F}}^2}{(1-\lambda)^4} + \frac{24\alpha_2^2\beta_1\beta_2^2\eta^3\hat{C}_{g_y}^2 L_{\tilde{F}}^2}{(1-\lambda)^4} + \frac{400\alpha_1^2\beta_1^3\eta^3\hat{C}_{\tilde{F}}^2 L_{g_y}^2 L_F^2}{\mu^2(1-\lambda)^4} + \frac{400\alpha_2^2\beta_1\beta_2^2\eta^3\hat{C}_{g_y}^2 L_F^2 L_{g_y}^2}{\mu^2(1-\lambda)^4} \\
&\quad + \frac{216\beta_1^3\eta^3\hat{C}_{\tilde{F}}^2 L_{\tilde{F}}^2}{(1-\lambda)^4} + \frac{216\alpha_2^2\beta_1\beta_2^2\eta^3\hat{C}_{g_y}^2 L_{\tilde{F}}^2}{\alpha_1^2(1-\lambda)^4} + \frac{3600\alpha_1^2\beta_1^3\eta^3\hat{C}_{\tilde{F}}^2 L_F^2 L_{g_y}^2}{\alpha_2^2\mu^2(1-\lambda)^4} + \frac{3600\beta_1\beta_2^2\eta^3\hat{C}_{g_y}^2 L_F^2 L_{g_y}^2}{\mu^2(1-\lambda)^4} \\
&\quad + \frac{3\beta_1\alpha_1\eta^2\sigma_{\tilde{F}}^2}{K} + \frac{50\beta_1\alpha_2\eta^2\sigma^2 L_F^2}{\mu^2 K}\ .
\end{aligned} \tag{92}$$

By setting the coefficient of $\mathbb{E}[\|\bar{u}_t\|^2]$ to be non-positive, we can get

$$\begin{aligned} & \frac{25\eta\beta_1^3 L_F^2 L_y^2}{\beta_2^2\mu^2} + \frac{12\eta\beta_1^3 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{200\eta\beta_1^3 L_F^2 L_{g_y}^2}{\alpha_2^2\mu^2} - \frac{\eta\beta_1}{4} \leq 0 \ , \\ & \frac{25\beta_1^2 L_F^2 L_y^2}{\beta_2^2\mu^2} + \frac{12\beta_1^2 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{200\beta_1^2 L_F^2 L_{g_y}^2}{\alpha_2^2\mu^2} - \frac{1}{4} \leq 0 \ . \end{aligned} \tag{93}$$

By setting $\beta_1 \leq \frac{\beta_2\mu}{15L_F L_y}$, we can get $\frac{25\beta_1^2 L_F^2 L_y^2}{\beta_2^2\mu^2} - \frac{1}{4} \leq -\frac{1}{8}$. Then, we have

$$\begin{aligned} & \frac{25\beta_1^2 L_F^2 L_y^2}{\beta_2^2\mu^2} + \frac{12\beta_1^2 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{200\beta_1^2 L_F^2 L_{g_y}^2}{\alpha_2^2\mu^2} - \frac{1}{4} \\ & \leq \frac{12\beta_1^2 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{200\beta_1^2 L_F^2 L_{g_y}^2}{\alpha_2^2\mu^2} - \frac{1}{8} \\ & \leq \frac{12\beta_1^2 L_{\tilde{F}}^2 L_{g_y}^2}{\alpha_1^2\mu^2} + \frac{200\beta_1^2 L_F^2 L_{g_y}^2}{\alpha_2^2\mu^2} - \frac{1}{8} \ , \end{aligned} \tag{94}$$

where the last step holds due to $L_{g_y}/\mu > 1$. By letting this upper bound non-positive, we can get

$$\begin{aligned} & \frac{12\beta_1^2 L_{\tilde{F}}^2 L_{g_y}^2}{\alpha_1^2\mu^2} + \frac{200\beta_1^2 L_F^2 L_{g_y}^2}{\alpha_2^2\mu^2} - \frac{1}{8} \leq 0 \ , \\ & \frac{(12L_{\tilde{F}}^2/\alpha_1^2 + 200L_F^2/\alpha_2^2)\beta_1^2 L_{g_y}^2}{\mu^2} \leq \frac{1}{8} \ , \\ & \beta_1 \leq \frac{\mu}{4L_{g_y}\sqrt{6L_{\tilde{F}}^2/\alpha_1^2 + 100L_F^2/\alpha_2^2}} \ . \end{aligned} \tag{95}$$

By setting the coefficient of $\mathbb{E}[\|\bar{v}_t\|^2]$ to be non-positive, we can get

$$\begin{aligned} & \frac{12\eta\beta_1\beta_2^2 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{200\eta\beta_1\beta_2^2 L_F^2 L_{g_y}^2}{\alpha_2^2\mu^2} - \frac{9\eta\beta_1\beta_2 L_F^2}{2\mu} \leq 0 \ , \\ & \frac{12\beta_2 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{200\beta_2 L_F^2 L_{g_y}^2}{\alpha_2^2\mu^2} - \frac{9L_F^2}{2\mu} \leq 0 \ , \end{aligned} \tag{96}$$

Because $L_{g_y}/\mu > 1$, we can get

$$\begin{aligned} & \frac{12\beta_2 L_{\tilde{F}}^2}{\alpha_1^2} + \frac{200\beta_2 L_F^2 L_{g_y}^2}{\alpha_2^2\mu^2} - \frac{9L_F^2}{2\mu} \\ & \leq \frac{12\beta_2 L_{\tilde{F}}^2 L_{g_y}^2}{\alpha_1^2\mu^2} + \frac{200\beta_2 L_F^2 L_{g_y}^2}{\alpha_2^2\mu^2} - \frac{9L_F^2}{2\mu} \\ & = \frac{(12L_{\tilde{F}}^2/\alpha_1^2 + 200L_F^2/\alpha_2^2)L_{g_y}^2}{\mu^2}\beta_2 - \frac{9L_F^2}{2\mu} \ . \end{aligned} \tag{97}$$

By letting this upper bound non-positive, we can get

$$\begin{aligned} & \frac{(12L_{\tilde{F}}^2/\alpha_1^2 + 200L_F^2/\alpha_2^2)L_{g_y}^2}{\mu^2}\beta_2 \leq \frac{9L_F^2}{2\mu} \ , \\ & \beta_2 \leq \frac{9\mu L_F^2}{2(12L_{\tilde{F}}^2/\alpha_1^2 + 200L_F^2/\alpha_2^2)L_{g_y}^2} \ . \end{aligned} \tag{98}$$

As a result, by setting

$$\begin{aligned} & \beta_1 \leq \min\left\{\frac{\beta_2\mu}{15L_F L_y}, \frac{\mu}{4L_{g_y}\sqrt{6L_{\tilde{F}}^2/\alpha_1^2 + 100L_F^2/\alpha_2^2}}\right\}, \\ & \beta_2 \leq \frac{9\mu L_F^2}{2(12L_{\tilde{F}}^2/\alpha_1^2 + 200L_F^2/\alpha_2^2)L_{g_y}^2} \ , \end{aligned} \tag{99}$$

we can get

$$
\begin{aligned}
&\mathcal{L}_{t+1}-\mathcal{L}_t \\
&\leq -\frac{\eta\beta_1}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2]-\frac{\eta\beta_1 L_F^2}{2}\mathbb{E}[\|\bar{y}_t-y^*(\bar{x}_t)\|^2]+\frac{3\eta\beta_1 C_{g_{xy}}^2 C_{f_y}^2}{\mu^2}(1-\frac{\mu}{L_{g_y}})^{2J} \\
&\quad+\frac{24\alpha_1^2\beta_1^3\eta^3\hat{C}_{\tilde{F}}^2 L_{\tilde{F}}^2}{(1-\lambda)^4}+\frac{24\alpha_2^2\beta_1\beta_2^2\eta^3\hat{C}_{g_y}^2 L_{\tilde{F}}^2}{(1-\lambda)^4}+\frac{400\alpha_1^2\beta_1^3\eta^3\hat{C}_{\tilde{F}}^2 L_{g_y}^2 L_F^2}{\mu^2(1-\lambda)^4}+\frac{400\alpha_2^2\beta_1\beta_2^2\eta^3\hat{C}_{g_y}^2 L_F^2 L_{g_y}^2}{\mu^2(1-\lambda)^4} \\
&\quad+\frac{216\beta_1^3\eta^3\hat{C}_{\tilde{F}}^2 L_{\tilde{F}}^2}{(1-\lambda)^4}+\frac{216\alpha_2^2\beta_1\beta_2^2\eta^3\hat{C}_{g_y}^2 L_{\tilde{F}}^2}{\alpha_1^2(1-\lambda)^4}+\frac{3600\alpha_1^2\beta_1^3\eta^3\hat{C}_{\tilde{F}}^2 L_F^2 L_{g_y}^2}{\alpha_2^2\mu^2(1-\lambda)^4}+\frac{3600\beta_1\beta_2^2\eta^3\hat{C}_{g_y}^2 L_F^2 L_{g_y}^2}{\mu^2(1-\lambda)^4} \\
&\quad+\frac{3\beta_1\alpha_1\eta^2\sigma_{\tilde{F}}^2}{K}+\frac{50\beta_1\alpha_2\eta^2\sigma^2 L_F^2}{\mu^2 K}\ .
\end{aligned} \tag{100}
$$

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

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2]+L_F^2\mathbb{E}[\|\bar{y}_t-y^*(\bar{x}_t)\|^2] \\
&\leq \frac{2(\mathcal{L}_0-\mathcal{L}_T)}{\eta\beta_1 T}+\frac{6C_{g_{xy}}^2 C_{f_y}^2}{\mu^2}(1-\frac{\mu}{L_{g_y}})^{2J} \\
&\quad+\frac{48\alpha_1^2\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2 L_{\tilde{F}}^2}{(1-\lambda)^4}+\frac{48\alpha_2^2\beta_2^2\eta^2\hat{C}_{g_y}^2 L_{\tilde{F}}^2}{(1-\lambda)^4}+\frac{800\alpha_1^2\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2 L_{g_y}^2 L_F^2}{\mu^2(1-\lambda)^4}+\frac{800\alpha_2^2\beta_2^2\eta^2\hat{C}_{g_y}^2 L_F^2 L_{g_y}^2}{\mu^2(1-\lambda)^4} \\
&\quad+\frac{432\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2 L_{\tilde{F}}^2}{(1-\lambda)^4}+\frac{432\alpha_2^2\beta_2^2\eta^2\hat{C}_{g_y}^2 L_{\tilde{F}}^2}{\alpha_1^2(1-\lambda)^4}+\frac{7200\alpha_1^2\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2 L_F^2 L_{g_y}^2}{\alpha_2^2\mu^2(1-\lambda)^4}+\frac{7200\beta_2^2\eta^2\hat{C}_{g_y}^2 L_F^2 L_{g_y}^2}{\mu^2(1-\lambda)^4} \\
&\quad+\frac{6\alpha_1\eta\sigma_{\tilde{F}}^2}{K}+\frac{100\alpha_2\eta\sigma^2 L_F^2}{\mu^2 K}\ .
\end{aligned} \tag{101}
$$

Due to $\mathbb{E}[\|(\Delta_0^{\tilde{F}}-U_0)\frac{1}{K}\mathbf{1}\|^2] \leq \frac{2}{K^2}\mathbb{E}[\|\sum_{k=1}^{K}\nabla\tilde{F}^{(k)}(x_0^{(k)},y_0^{(k)})\|^2]+\frac{2\alpha_1^2\eta^2}{K^2}\mathbb{E}[\|\sum_{k=1}^{K}\nabla\tilde{F}^{(k)}(x_0^{(k)},y_0^{(k)};\tilde{\xi}_0^{(k)})\|^2] \leq 4\hat{C}_{\tilde{F}}^2$ and $\mathbb{E}[\|(\Delta_0^{g}-V_0)\frac{1}{K}\mathbf{1}\|^2] \leq 4\hat{C}_{g_y}^2$, we can get

$$
\begin{aligned}
\mathcal{L}_0 &= \mathbb{E}[F(x_0)]+\frac{6\beta_1 L_F^2}{\beta_2\mu}\mathbb{E}[\|\bar{y}_0-y^*(\bar{x}_0)\|^2]+\frac{3\beta_1}{\alpha_1}\mathbb{E}[\|(\Delta_0^{\tilde{F}}-U_0)\frac{1}{K}\mathbf{1}\|^2]+\frac{50\beta_1 L_F^2}{\alpha_2\mu^2}\mathbb{E}[\|(\Delta_0^{g}-V_0)\frac{1}{K}\mathbf{1}\|^2] \\
&\leq \mathbb{E}[F(x_0)]+\frac{6\beta_1 L_F^2}{\beta_2\mu}\mathbb{E}[\|\bar{y}_0-y^*(\bar{x}_0)\|^2]+\frac{12\beta_1\hat{C}_{\tilde{F}}^2}{\alpha_1}+\frac{200\beta_1 L_F^2\hat{C}_{g_y}^2}{\alpha_2\mu^2}\ .
\end{aligned} \tag{102}
$$

Finally, we can get

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2]+L_F^2\mathbb{E}[\|\bar{y}_t-y^*(\bar{x}_t)\|^2]) \\
&\leq \frac{2(F(x_0)-F(x_*))}{\eta\beta_1 T}+\frac{12L_F^2}{\beta_2\mu\eta T}\mathbb{E}[\|\bar{y}_0-y^*(\bar{x}_0)\|^2]+\frac{6C_{g_{xy}}^2 C_{f_y}^2}{\mu^2}(1-\frac{\mu}{L_{g_y}})^{2J} \\
&\quad+\frac{48\alpha_1^2\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2 L_{\tilde{F}}^2}{(1-\lambda)^4}+\frac{48\alpha_2^2\beta_2^2\eta^2\hat{C}_{g_y}^2 L_{\tilde{F}}^2}{(1-\lambda)^4}+\frac{800\alpha_1^2\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2 L_{g_y}^2 L_F^2}{\mu^2(1-\lambda)^4}+\frac{800\alpha_2^2\beta_2^2\eta^2\hat{C}_{g_y}^2 L_F^2 L_{g_y}^2}{\mu^2(1-\lambda)^4} \\
&\quad+\frac{432\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2 L_{\tilde{F}}^2}{(1-\lambda)^4}+\frac{432\alpha_2^2\beta_2^2\eta^2\hat{C}_{g_y}^2 L_{\tilde{F}}^2}{\alpha_1^2(1-\lambda)^4}+\frac{7200\alpha_1^2\beta_1^2\eta^2\hat{C}_{\tilde{F}}^2 L_F^2 L_{g_y}^2}{\alpha_2^2\mu^2(1-\lambda)^4}+\frac{7200\beta_2^2\eta^2\hat{C}_{g_y}^2 L_F^2 L_{g_y}^2}{\mu^2(1-\lambda)^4} \\
&\quad+\frac{6\alpha_1\eta\sigma_{\tilde{F}}^2}{K}+\frac{100\alpha_2\eta\sigma^2 L_F^2}{\mu^2 K}+\frac{24\hat{C}_{\tilde{F}}^2}{\alpha_1\eta T}+\frac{400L_F^2\hat{C}_{g_y}^2}{\alpha_2\mu^2\eta T}\ .
\end{aligned} \tag{103}
$$

A.4 Proof of Theorem 3

A.4.1 Characterization of $F^{(k)}(x)$

Lemma 23. *Given Assumptions 2, 3, 7, 8, the following inequalities hold.*

$$\begin{aligned} &\|\nabla F^{(k)}(x) - \nabla F^{(k)}(x,y)\| \leq L_F \|y - y^*(x)\| \,, \\ &\|\nabla F^{(k)}(x_1) - \nabla F^{(k)}(x_2)\| \leq L_F^* \|x_1 - x_2\| \,, \\ &\|y^*(x_1) - y^*(x_2)\| \leq L_y \|x_1 - x_2\| \,, \end{aligned} \tag{104}$$

where $L_F = \ell_{f_x} + \frac{\ell_{f_y} c_{g_{xy}}}{\mu} + \frac{c_{f_y} \ell_{g_{xy}}}{\mu} + \frac{\ell_{g_{yy}} c_{f_y} c_{g_{xy}}}{\mu^2}$ $L_F^* = L_F + \frac{L_F c_{g_{xy}}}{\mu}$, $L_y = \frac{c_{g_{xy}}}{\mu}$.

These inequalities can also be proved by following Lemma 2.2 in Ghadimi and Wang (2018) when given the mean-square Lipschitz smoothness assumption, since $\|\mathbb{E}[a]\|^2 \leq \mathbb{E}[\|a\|^2]$ holds for any random variable a.

A.4.2 Characterization of $\nabla \tilde{F}^{(k)}(x,y)$

Lemma 24. *Given Assumptions 2, 3, 7, 8, the following inequalities hold.*

$$\begin{aligned} &\mathbb{E}\Big[\frac{J}{\ell_{g_y}} \prod_{j=1}^{\tilde{J}} (I - \frac{1}{\ell_{g_y}} \nabla^2_{yy} g^{(k)}(x,y;\zeta_j))\Big] = \frac{1}{\ell_{g_y}} \sum_{j=0}^{J-1} \Big(I - \frac{1}{\ell_{g_y}} \nabla^2_{yy} g^{(k)}(x,y)\Big)^j \,, \\ &\Big\| \Big(\nabla^2_{yy} g^{(k)}(x,y)\Big)^{-1} - \mathbb{E}\Big[\frac{J}{\ell_{g_y}} \prod_{j=1}^{\tilde{J}} (I - \frac{1}{\ell_{g_y}} \nabla^2_{yy} g^{(k)}(x,y;\zeta_j))\Big] \Big\| \leq \frac{1}{\mu} (1 - \frac{\mu}{\ell_{g_y}})^J \,, \\ &\mathbb{E}\Big[\Big\| \Big(\nabla^2_{yy} g^{(k)}(x,y)\Big)^{-1} - \frac{J}{\ell_{g_y}} \prod_{j=1}^{\tilde{J}} (I - \frac{1}{\ell_{g_y}} \nabla^2_{yy} g^{(k)}(x,y;\zeta_j)) \Big\|\Big] \leq \frac{2}{\mu} \,, \\ &\mathbb{E}\Big[\Big\| \frac{J}{\ell_{g_y}} \prod_{j=1}^{\tilde{J}} (I - \frac{1}{\ell_{g_y}} \nabla^2_{yy} g^{(k)}(x,y;\zeta_j)) \Big\|\Big] \leq \frac{1}{\mu} \,. \end{aligned} \tag{105}$$

These inequalities can be proved by following Lemma 3.2 in Ghadimi and Wang (2018) when given the mean-square Lipschitz smoothness assumption since $\|\mathbb{E}[a]\|^2 \leq \mathbb{E}[\|a\|^2]$ holds for any random variable a.

Lemma 25. *(Bias) Given Assumptions 2, 3, 7, 8, the approximation error of $\nabla \tilde{F}^{(k)}(x,y)$ for $\nabla F^{(k)}(x,y)$ can be bounded as follows:*

$$\|\nabla F^{(k)}(x,y) - \nabla \tilde{F}^{(k)}(x,y)\| \leq \frac{c_{g_{xy}} c_{f_y}}{\mu} (1 - \frac{\mu}{\ell_{g_y}})^J \,. \tag{106}$$

These inequalities can be proved by following Lemma 3.2 in Ghadimi and Wang (2018) when given the mean-square Lipschitz smoothness assumption since $\|\mathbb{E}[a]\|^2 \leq \mathbb{E}[\|a\|^2]$ holds for any random variable a.

Lemma 26. *(Variance) Given Assumptions 2, 3, 7, 8, the variance of the stochastic hypergradient can be bounded as follows:*

$$\mathbb{E}[\|\nabla\tilde{F}^{(k)}(x,y)-\nabla\tilde{F}^{(k)}(x,y;\tilde{\xi})\|^2]\leq\sigma_{\tilde{F}}^2\ , \tag{107}$$

where $\sigma_{\tilde{F}}^2=4\sigma^2+\frac{4c_{f_y}^2\sigma^2}{\mu^2}+\frac{4c_{g_{xy}}^2\sigma^2}{\mu^2}+\frac{16c_{g_{xy}}^2c_{f_y}^2}{\mu^2}$.

Proof.

$$\begin{aligned}
&\mathbb{E}[\|\nabla\tilde{F}^{(k)}(x,y)-\nabla\tilde{F}^{(k)}(x,y;\tilde{\xi})\|^2]\\
&=\mathbb{E}\Big[\Big\|\nabla_x f^{(k)}(x,y)-\nabla_{xy}^2 g^{(k)}(x,y)\mathbb{E}\Big[\frac{J}{\ell_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{\ell_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\nabla_y f^{(k)}(x,y)\\
&\quad-\nabla_x f^{(k)}(x,y;\xi)+\nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\Big(\frac{J}{\ell_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{\ell_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big)\nabla_y f^{(k)}(x,y;\xi)\Big\|^2\Big]\\
&=\mathbb{E}\Big[\Big\|\nabla_x f^{(k)}(x,y)-\nabla_x f^{(k)}(x,y;\xi)\\
&\quad+\nabla_{xy}^2 g^{(k)}(x,y)\mathbb{E}\Big[\frac{J}{\ell_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{\ell_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\nabla_y f^{(k)}(x,y)\\
&\quad-\nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\mathbb{E}\Big[\frac{J}{\ell_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{\ell_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\nabla_y f^{(k)}(x,y)\\
&\quad+\nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\mathbb{E}\Big[\frac{J}{\ell_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{\ell_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\nabla_y f^{(k)}(x,y)\\
&\quad-\nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\mathbb{E}\Big[\frac{J}{\ell_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{\ell_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\nabla_y f^{(k)}(x,y;\xi)\\
&\quad+\nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\mathbb{E}\Big[\frac{J}{\ell_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{\ell_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\nabla_y f^{(k)}(x,y;\xi)\\
&\quad-\nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\Big(\frac{J}{\ell_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{\ell_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big)\nabla_y f^{(k)}(x,y;\xi)\Big\|^2\Big]\\
&\leq 4\sigma^2+4\mathbb{E}\Big[\Big\|\Big(\nabla_{xy}^2 g^{(k)}(x,y)-\nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\Big)\mathbb{E}\Big[\frac{J}{\ell_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{\ell_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\nabla_y f^{(k)}(x,y)\Big\|^2\Big]\\
&\quad+4\mathbb{E}\Big[\Big\|\nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\mathbb{E}\Big[\frac{J}{\ell_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{\ell_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\Big(\nabla_y f^{(k)}(x,y)-\nabla_y f^{(k)}(x,y;\xi)\Big)\Big\|^2\Big]\\
&\quad+4\mathbb{E}\Big[\Big\|\nabla_{xy}^2 g^{(k)}(x,y;\zeta_0)\Big(\mathbb{E}\Big[\frac{J}{\ell_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{\ell_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big]\\
&\qquad-\frac{J}{\ell_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{\ell_{g_y}}\nabla_{yy}^2 g^{(k)}(x,y;\zeta_j))\Big)\nabla_y f^{(k)}(x,y;\xi)\Big\|^2\Big]\\
&\overset{(s_1)}{\leq} 4\sigma^2+\frac{4c_{f_y}^2\sigma^2}{\mu^2}+\frac{4c_{g_{xy}}^2\sigma^2}{\mu^2}+\frac{16c_{g_{xy}}^2c_{f_y}^2}{\mu^2}\ ,
\end{aligned} \tag{108}$$

where (s_1) holds due to Lemma 24, Assumptions 3, 7, 8, and the following inequality.

$$
\begin{aligned}
&\Big\|\mathbb{E}\Big[\frac{J}{\ell_{g_y}}\prod_{j=1}^{\tilde{J}}(I-\frac{1}{\ell_{g_y}}\nabla^2_{yy}g^{(k)}(x,y;\zeta_j))\Big]\Big\| = \Big\|\frac{1}{\ell_{g_y}}\sum_{j=0}^{J-1}\Big(I-\frac{1}{\ell_{g_y}}\nabla^2_{yy}g^{(k)}(x,y)\Big)^j\Big\| \\
&\leq \frac{1}{\ell_{g_y}}\sum_{j=0}^{J-1}\Big\|\Big(I-\frac{1}{\ell_{g_y}}\nabla^2_{yy}g^{(k)}(x,y)\Big)^j\Big\| \\
&\overset{(s_1)}{\leq} \frac{1}{\ell_{g_y}}\sum_{j=0}^{J-1}\Big(1-\frac{\mu}{\ell_{g_y}}\Big)^j \\
&\leq \frac{1}{\mu}\ ,
\end{aligned}
\tag{109}
$$

where (s_1) holds due to Assumption 2. □

Lemma 27. *(Smoothness) Khanduri et al. (2021b) Given Assumptions 2, 3, 7, 8, the approximated hypergradient $\nabla\tilde{F}^{(k)}(x,y;\tilde{xi})$ is $L_{\tilde{F}}$-Lipschitz continuous, i.e.,*

$$
\mathbb{E}[\|\nabla\tilde{F}^{(k)}(x_1,y_1;\tilde{\xi})-\nabla\tilde{F}^{(k)}(x_2,y_2;\tilde{\xi})\|^2] \leq L^2_{\tilde{F}}(\|x_1-x_2\|^2+\|y_1-y_2\|^2),
\tag{110}
$$

where $L^2_{\tilde{F}} = 2\ell^2_{f_x} + \frac{6(c^2_{g_{xy}}\ell^2_{f_y}+c^2_{f_y}\ell^2_{g_{xy}})J}{2\mu\ell_{g_y}-\mu^2} + \frac{6c^2_{g_{xy}}c^2_{f_y}\ell^2_{g_y}J^3}{(2\mu\ell_{g_y}-\mu^2)(\ell_{g_y}-\mu)^2}$, *for any* (x_1,y_1), $(x_2,y_2)\in\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}$.

A.4.3 Characterization of Gradient Estimators

Lemma 28. *Given Assumptions 1, 2, 3, 7, 8, the following inequality holds.*

$$
\begin{aligned}
&\mathbb{E}[\|(\Delta^{\tilde{F}}_t-U_t)\frac{1}{K}\mathbf{1}\|^2] \leq (1-\alpha_1\eta^2)\mathbb{E}[\|(\Delta^{\tilde{F}}_{t-1}-U_{t-1})\frac{1}{K}\mathbf{1}\|^2] + 2L^2_{\tilde{F}}\frac{1}{K^2}\mathbb{E}[\|X_t-X_{t-1}\|^2_F] \\
&+2L^2_{\tilde{F}}\frac{1}{K^2}\mathbb{E}[\|Y_t-Y_{t-1}\|^2_F] + \frac{2\alpha_1^2\eta^4\sigma^2_{\tilde{F}}}{K}\ .
\end{aligned}
\tag{111}
$$

Proof.

$$
\begin{aligned}
&\mathbb{E}[\|(\Delta^{\tilde{F}}_t-U_t)\frac{1}{K}\mathbf{1}\|^2] \\
&= \mathbb{E}[\|(\Delta^{\tilde{F}}_t-(1-\alpha_1\eta^2)(U_{t-1}-\Delta^{\tilde{F}_{\tilde{\xi}_t}}_{t-1})-\Delta^{\tilde{F}_{\tilde{\xi}_t}}_t)\frac{1}{K}\mathbf{1}\|^2] \\
&= \mathbb{E}[\|((1-\alpha_1\eta^2)(\Delta^{\tilde{F}}_{t-1}-U_{t-1})+(1-\alpha_1\eta^2)(\Delta^{\tilde{F}_{\tilde{\xi}_t}}_{t-1}-\Delta^{\tilde{F}}_{t-1})-(\Delta^{\tilde{F}_{\tilde{\xi}_t}}_t-\Delta^{\tilde{F}}_t))\frac{1}{K}\mathbf{1}\|^2_F] \\
&\overset{(s_1)}{=} (1-\alpha_1\eta^2)^2\mathbb{E}[\|(\Delta^{\tilde{F}}_{t-1}-U_{t-1})\frac{1}{K}\mathbf{1}\|^2] + \mathbb{E}[\|(1-\alpha_1\eta^2)(\Delta^{\tilde{F}_{\tilde{\xi}_t}}_{t-1}-\Delta^{\tilde{F}}_{t-1}-\Delta^{\tilde{F}_{\tilde{\xi}_t}}_t+\Delta^{\tilde{F}}_t)\frac{1}{K}\mathbf{1} \\
&\quad -\alpha_1\eta^2(\Delta^{\tilde{F}_{\tilde{\xi}_t}}_t-\Delta^{\tilde{F}}_t)\frac{1}{K}\mathbf{1}\|^2] \\
&\leq (1-\alpha_1\eta^2)^2\mathbb{E}[\|(\Delta^{\tilde{F}}_{t-1}-U_{t-1})\frac{1}{K}\mathbf{1}\|^2] + 2(1-\alpha_1\eta^2)^2\frac{1}{K^2}\mathbb{E}[\|\Delta^{\tilde{F}}_{t-1}-\Delta^{\tilde{F}_{\tilde{\xi}_t}}_{t-1}-\Delta^{\tilde{F}}_t+\Delta^{\tilde{F}_{\tilde{\xi}_t}}_t\|^2_F] \\
&\quad +2\alpha_1^2\eta^4\frac{1}{K^2}\mathbb{E}[\|\Delta^{\tilde{F}}_t-\Delta^{\tilde{F}_{\tilde{\xi}_t}}_t\|^2_F] \\
&\overset{(s_2)}{\leq} (1-\alpha_1\eta^2)^2\mathbb{E}[\|(\Delta^{\tilde{F}}_{t-1}-U_{t-1})\frac{1}{K}\mathbf{1}\|^2] + 2(1-\alpha_1\eta^2)^2\frac{1}{K^2}\mathbb{E}[\|\Delta^{\tilde{F}_{\tilde{\xi}_t}}_t-\Delta^{\tilde{F}_{\tilde{\xi}_t}}_{t-1}\|^2_F] + \frac{2\alpha_1^2\eta^4\sigma^2_{\tilde{F}}}{K} \\
&\overset{(s_3)}{\leq} (1-\alpha_1\eta^2)\mathbb{E}[\|(\Delta^{\tilde{F}}_{t-1}-U_{t-1})\frac{1}{K}\mathbf{1}\|^2] + 2(1-\alpha_1\eta^2)^2L^2_{\tilde{F}}\frac{1}{K^2}\mathbb{E}[\|X_t-X_{t-1}\|^2_F] \\
&\quad +2(1-\alpha_1\eta^2)^2L^2_{\tilde{F}}\frac{1}{K^2}\mathbb{E}[\|Y_t-Y_{t-1}\|^2_F] + \frac{2\alpha_1^2\eta^4\sigma^2_{\tilde{F}}}{K} \\
&\leq (1-\alpha_1\eta^2)\mathbb{E}[\|(\Delta^{\tilde{F}}_{t-1}-U_{t-1})\frac{1}{K}\mathbf{1}\|^2] + 2L^2_{\tilde{F}}\frac{1}{K^2}\mathbb{E}[\|X_t-X_{t-1}\|^2_F] \\
&\quad +2L^2_{\tilde{F}}\frac{1}{K^2}\mathbb{E}[\|Y_t-Y_{t-1}\|^2_F] + \frac{2\alpha_1^2\eta^4\sigma^2_{\tilde{F}}}{K}\ ,
\end{aligned}
\tag{112}
$$

where (s_1) holds due to $\Delta_t^{\tilde{F}} = \mathbb{E}[\Delta_t^{\tilde{F}_{\tilde{\xi}_t}}]$ and $\Delta_{t-1}^{\tilde{F}} = \mathbb{E}[\Delta_{t-1}^{\tilde{F}_{\tilde{\xi}_t}}]$, (s_2) holds due to Lemma 26, (s_3) holds due to Lemma 27. □

Lemma 29. *Given Assumptions 1, 2, 3, 7, 8, the following inequality holds.*

$$\begin{aligned}&\mathbb{E}[\|(\Delta_t^g - V_t)\frac{1}{K}\mathbf{1}\|^2] \leq (1-\alpha_2\eta^2)\mathbb{E}[\|(\Delta_{t-1}^g - V_{t-1})\frac{1}{K}\mathbf{1}\|^2] + 2\ell_{g_y}^2\frac{1}{K^2}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] \\ &\quad + 2\ell_{g_y}^2\frac{1}{K^2}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] + \frac{2\alpha_2^2\eta^4\sigma^2}{K} \ .\end{aligned} \tag{113}$$

Proof.

$$\begin{aligned}&\mathbb{E}[\|(\Delta_t^g - V_t)\frac{1}{K}\mathbf{1}\|^2] \\ &= \mathbb{E}[\|(\Delta_t^g - (1-\alpha_2\eta^2)(V_{t-1} - \Delta_{t-1}^{g_{\zeta_t}}) - \Delta_t^{g_{\zeta_t}})\frac{1}{K}\mathbf{1}\|^2] \\ &= \mathbb{E}[\|(1-\alpha_2\eta^2)(\Delta_{t-1}^g - V_{t-1})\frac{1}{K}\mathbf{1} + (1-\alpha_2\eta^2)(\Delta_{t-1}^{g_{\zeta_t}} - \Delta_{t-1}^g) - (\Delta_t^{g_{\zeta_t}} - \Delta_t^g)\frac{1}{K}\mathbf{1}\|^2] \\ &= \mathbb{E}[\|(1-\alpha_2\eta^2)(\Delta_{t-1}^g - V_{t-1})\frac{1}{K}\mathbf{1} + (1-\alpha_2\eta^2)(\Delta_{t-1}^{g_{\zeta_t}} - \Delta_{t-1}^g - \Delta_t^{g_{\zeta_t}} + \Delta_t^g)\frac{1}{K}\mathbf{1} - \alpha_2\eta^2(\Delta_t^{g_{\zeta_t}} - \Delta_t^g)\frac{1}{K}\mathbf{1}\|^2] \\ &\leq (1-\alpha_2\eta^2)^2\mathbb{E}[\|(\Delta_{t-1}^g - V_{t-1})\frac{1}{K}\mathbf{1}\|^2] \\ &\quad + 2(1-\alpha_2\eta^2)^2\frac{1}{K^2}\mathbb{E}[\|\Delta_{t-1}^{g_{\zeta_t}} - \Delta_{t-1}^g - \Delta_t^{g_{\zeta_t}} + \Delta_t^g\|_F^2] + 2\alpha_2^2\eta^4\frac{1}{K^2}\mathbb{E}[\|\Delta_t^g - \Delta_t^{g_{\zeta_t}}\|_F^2] \\ &\leq (1-\alpha_2\eta^2)^2\mathbb{E}[\|(\Delta_{t-1}^g - V_{t-1})\frac{1}{K}\mathbf{1}\|^2] + 2(1-\alpha_2\eta^2)^2\frac{1}{K^2}\mathbb{E}[\|\Delta_{t-1}^{g_{\zeta_t}} - \Delta_t^{g_{\zeta_t}}\|_F^2] + 2\alpha_2^2\eta^4\frac{1}{K^2}\mathbb{E}[\|\Delta_t^g - \Delta_t^{g_{\zeta_t}}\|_F^2] \\ &\overset{(s_1)}{\leq} (1-\alpha_2\eta^2)\mathbb{E}[\|(\Delta_{t-1}^g - V_{t-1})\frac{1}{K}\mathbf{1}\|^2] + 2\ell_{g_y}^2\frac{1}{K^2}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] \\ &\quad + 2\ell_{g_y}^2\frac{1}{K^2}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] + \frac{2\alpha_2^2\eta^4\sigma^2}{K} \ ,\end{aligned} \tag{114}$$

where (s_1) holds due to Assumption 8 and Assumption 3. □

Lemma 30. *Given Assumptions 1, 2, 3, 7, 8, the following inequality holds.*

$$\begin{aligned}&\frac{1}{K}\mathbb{E}[\|\Delta_t^{\tilde{F}} - U_t\|_F^2] \leq (1-\alpha_1\eta^2)\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^{\tilde{F}} - U_{t-1}\|_F^2] + 2(1-\alpha_1\eta^2)^2L_{\tilde{F}}^2\frac{1}{K}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] \\ &+ 2(1-\alpha_1\eta^2)^2L_{\tilde{F}}^2\frac{1}{K}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] + 2\alpha_1^2\eta^4\sigma_{\tilde{F}}^2 \ .\end{aligned} \tag{115}$$

Proof.

$$\begin{aligned}
&\frac{1}{K}\mathbb{E}[\|\Delta_t^{\tilde{F}} - U_t\|_F^2] \\
&= \frac{1}{K}\mathbb{E}[\|\Delta_t^{\tilde{F}} - (1-\alpha_1\eta^2)(U_{t-1} - \Delta_{t-1}^{\tilde{F}_{\tilde{\xi}_t}}) - \Delta_t^{\tilde{F}_{\tilde{\xi}_t}}\|_F^2] \\
&= \frac{1}{K}\mathbb{E}[\|(1-\alpha_1\eta^2)(\Delta_{t-1}^{\tilde{F}} - U_{t-1}) + (1-\alpha_1\eta^2)(\Delta_{t-1}^{\tilde{F}_{\tilde{\xi}_t}} - \Delta_{t-1}^{\tilde{F}}) - (\Delta_t^{\tilde{F}_{\tilde{\xi}_t}} - \Delta_t^{\tilde{F}})\|_F^2] \\
&\overset{(s_1)}{=} (1-\alpha_1\eta^2)^2\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^{\tilde{F}} - U_{t-1}\|_F^2] + \frac{1}{K}\mathbb{E}[\|(1-\alpha_1\eta^2)(\Delta_{t-1}^{\tilde{F}_{\tilde{\xi}_t}} - \Delta_{t-1}^{\tilde{F}} - \Delta_t^{\tilde{F}_{\tilde{\xi}_t}} + \Delta_t^{\tilde{F}}) \\
&\quad - \alpha_1\eta^2(\Delta_t^{\tilde{F}_{\tilde{\xi}_t}} - \Delta_t^{\tilde{F}})\|_F^2] \\
&\leq (1-\alpha_1\eta^2)^2\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^{\tilde{F}} - U_{t-1}\|_F^2] + 2(1-\alpha_1\eta^2)^2\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^{\tilde{F}} - \Delta_{t-1}^{\tilde{F}_{\tilde{\xi}_t}} - \Delta_t^{\tilde{F}} + \Delta_t^{\tilde{F}_{\tilde{\xi}_t}}\|_F^2] \\
&\quad + 2\alpha_1^2\eta^4\frac{1}{K}\mathbb{E}[\|\Delta_t^{\tilde{F}} - \Delta_t^{\tilde{F}_{\tilde{\xi}_t}}\|_F^2] \\
&\overset{s_2}{\leq} (1-\alpha_1\eta^2)^2\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^{\tilde{F}} - U_{t-1}\|_F^2] + 2(1-\alpha_1\eta^2)^2\frac{1}{K}\mathbb{E}[\|\Delta_t^{\tilde{F}_{\tilde{\xi}_t}} - \Delta_{t-1}^{\tilde{F}_{\tilde{\xi}_t}}\|_F^2] + 2\alpha_1^2\eta^4\sigma_{\tilde{F}}^2 \\
&\overset{(s_3)}{\leq} (1-\alpha_1\eta^2)\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^{\tilde{F}} - U_{t-1}\|_F^2] + 2(1-\alpha_1\eta^2)^2L_{\tilde{F}}^2\frac{1}{K}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] \\
&\quad + 2(1-\alpha_1\eta^2)^2L_{\tilde{F}}^2\frac{1}{K}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] + 2\alpha_1^2\eta^4\sigma_{\tilde{F}}^2 ,
\end{aligned} \tag{116}$$

where (s_1) holds due to $\Delta_t^{\tilde{F}} = \mathbb{E}[\Delta_t^{\tilde{F}_{\tilde{\xi}_t}}]$ and $\Delta_{t-1}^{\tilde{F}} = \mathbb{E}[\Delta_{t-1}^{\tilde{F}_{\tilde{\xi}_t}}]$, (s_2) holds due to Lemma 26, (s_3) holds due to Lemma 27. □

Lemma 31. *Given Assumptions 1, 2, 3, 7, 8, the following inequality holds.*

$$\begin{aligned}
&\frac{1}{K}\mathbb{E}[\|\Delta_t^g - V_t\|_F^2] \leq (1-\alpha_2\eta^2)\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^g - V_{t-1}\|_F^2] + 2(1-\alpha_2\eta^2)^2\ell_{g_y}^2\frac{1}{K}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] \\
&\quad + 2(1-\alpha_2\eta^2)^2\ell_{g_y}^2\frac{1}{K}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] + 2\alpha_2^2\eta^4\sigma^2 .
\end{aligned} \tag{117}$$

Proof.

$$\begin{aligned}
&\frac{1}{K}\mathbb{E}[\|\Delta_t^g - V_t\|_F^2] \\
&= \frac{1}{K}\mathbb{E}[\|\Delta_t^g - (1-\alpha_2\eta^2)(V_{t-1} - \Delta_{t-1}^{g_{\zeta_t}}) - \Delta_t^{g_{\zeta_t}}\|_F^2] \\
&= \frac{1}{K}\mathbb{E}[\|(1-\alpha_2\eta^2)(\Delta_{t-1}^g - V_{t-1}) + (1-\alpha_2\eta^2)(\Delta_{t-1}^{g_{\zeta_t}} - \Delta_{t-1}^g) - (\Delta_t^{g_{\zeta_t}} - \Delta_t^g)\|_F^2] \\
&= \frac{1}{K}\mathbb{E}[\|(1-\alpha_2\eta^2)(\Delta_{t-1}^g - V_{t-1}) + (1-\alpha_2\eta^2)(\Delta_{t-1}^{g_{\zeta_t}} - \Delta_{t-1}^g - \Delta_t^{g_{\zeta_t}} + \Delta_t^g) - \alpha_2\eta^2(\Delta_t^{g_{\zeta_t}} - \Delta_t^g)\|_F^2] \\
&\leq (1-\alpha_2\eta^2)^2\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^g - V_{t-1}\|_F^2] \\
&\quad + 2(1-\alpha_2\eta^2)^2\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^{g_{\zeta_t}} - \Delta_{t-1}^g - \Delta_t^{g_{\zeta_t}} + \Delta_t^g\|_F^2] + 2\alpha_2^2\eta^4\frac{1}{K}\mathbb{E}[\|\Delta_t^g - \Delta_t^{g_{\zeta_t}}\|_F^2] \\
&\leq (1-\alpha_2\eta^2)^2\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^g - V_{t-1}\|_F^2] + 2(1-\alpha_2\eta^2)^2\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^{g_{\zeta_t}} - \Delta_t^{g_{\zeta_t}}\|_F^2] + 2\alpha_2^2\eta^4\frac{1}{K}\mathbb{E}[\|\Delta_t^g - \Delta_t^{g_{\zeta_t}}\|_F^2] \\
&\overset{(s_1)}{\leq} (1-\alpha_2\eta^2)\frac{1}{K}\mathbb{E}[\|\Delta_{t-1}^g - V_{t-1}\|_F^2] + 2(1-\alpha_2\eta^2)^2\ell_{g_y}^2\frac{1}{K}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] \\
&\quad + 2(1-\alpha_2\eta^2)^2\ell_{g_y}^2\frac{1}{K}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] + 2\alpha_2^2\eta^4\sigma^2 ,
\end{aligned} \tag{118}$$

where (s_1) holds due to Assumption 8and Assumption 3. □

A.4.4 Characterization of Consensus Errors

Lemma 32. *Given Assumptions 1, 2, 3, 7, 8, the following inequality holds.*

$$\begin{aligned} &\frac{1}{K}\mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}}\|_F^2] \leq \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{2\ell_{\tilde{F}}^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] + \frac{2\ell_{\tilde{F}}^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] \\ &\quad + \frac{2\alpha_1^2\eta^4}{1-\lambda}\frac{1}{K}\mathbb{E}[\|U_{t-1} - \Delta_{t-1}^{\tilde{F}}\|_F^2] + \frac{\alpha_1^2\eta^4\sigma_{\tilde{F}}^2}{1-\lambda} \ . \end{aligned} \tag{119}$$

Proof.

$$\begin{aligned} &\frac{1}{K}\mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}}\|_F^2] \\ &= \frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}}W + U_t - U_{t-1} - \bar{Z}_{t-1}^{\tilde{F}} - \bar{U}_t + \bar{U}_{t-1}\|_F^2] \\ &\overset{(s_1)}{\leq} \frac{1}{K}\lambda\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|U_t - U_{t-1} - \bar{U}_t + \bar{U}_{t-1}\|_F^2] \\ &\leq \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|U_t - U_{t-1}\|_F^2] \\ &= \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|(1-\alpha_1\eta^2)(U_{t-1} - \Delta_{t-1}^{\tilde{F}_{\tilde{\xi}_t}}) + \Delta_t^{\tilde{F}_{\tilde{\xi}_t}} - U_{t-1}\|_F^2] \\ &= \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|\Delta_t^{\tilde{F}_{\tilde{\xi}_t}} - \Delta_{t-1}^{\tilde{F}_{\tilde{\xi}_t}} - \alpha_1\eta^2(U_{t-1} - \Delta_{t-1}^{\tilde{F}}) - \alpha_1\eta^2(\Delta_{t-1}^{\tilde{F}} - \Delta_{t-1}^{\tilde{F}_{\tilde{\xi}_t}})\|_F^2] \\ &\overset{(s_1)}{\leq} \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|\Delta_t^{\tilde{F}_{\tilde{\xi}_t}} - \Delta_{t-1}^{\tilde{F}_{\tilde{\xi}_t}}\|_F^2] + \frac{2\alpha_1^2\eta^4}{1-\lambda}\frac{1}{K}\mathbb{E}[\|U_{t-1} - \Delta_{t-1}^{\tilde{F}}\|_F^2] + \frac{\alpha_1^2\eta^4\sigma_{\tilde{F}}^2}{1-\lambda} \\ &\overset{(s_2)}{\leq} \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{\tilde{F}} - \bar{Z}_{t-1}^{\tilde{F}}\|_F^2] + \frac{2L_{\tilde{F}}^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] + \frac{2L_{\tilde{F}}^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] \\ &\quad + \frac{2\alpha_1^2\eta^4}{1-\lambda}\frac{1}{K}\mathbb{E}[\|U_{t-1} - \Delta_{t-1}^{\tilde{F}}\|_F^2] + \frac{\alpha_1^2\eta^4\sigma_{\tilde{F}}^2}{1-\lambda} \ , \end{aligned} \tag{120}$$

where (s_1) holds due to Lemma 34 with $a = \frac{1-\lambda}{\lambda}$, (s_2) holds due to $\mathbb{E}[\Delta_{t-1}^{\tilde{F}_{\tilde{\xi}_t}}] = \Delta_{t-1}^{\tilde{F}}$ and Lemma 26, (s_3) holds due to Lemma 27. □

Lemma 33. *Given Assumptions 1, 2, 3, 7, 8, the following inequality holds.*

$$\begin{aligned} &\frac{1}{K}\mathbb{E}[\|Z_t^{g} - \bar{Z}_t^{g}\|_F^2] \leq \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^{g} - \bar{Z}_{t-1}^{g}\|_F^2] + \frac{2\ell_{g_y}^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] + \frac{2\ell_{g_y}^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] \\ &\quad + \frac{2\alpha_2^2\eta^4}{1-\lambda}\frac{1}{K}\mathbb{E}[\|V_{t-1} - \Delta_{t-1}^{g}\|_F^2] + \frac{\alpha_2^2\eta^4\sigma^2}{1-\lambda} \ . \end{aligned} \tag{121}$$

Proof.

$$\begin{aligned}
&\frac{1}{K}\mathbb{E}[\|Z_t^g - \bar{Z}_t^g\|_F^2] \\
&= \frac{1}{K}\mathbb{E}[\|Z_{t-1}^g W + V_t - V_{t-1} - \bar{Z}_{t-1}^g - \bar{V}_t + \bar{V}_{t-1}\|_F^2] \\
&\overset{(s_1)}{\leq} \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^g - \bar{Z}_{t-1}^g\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|V_t - V_{t-1} - \bar{V}_t + \bar{V}_{t-1}\|_F^2] \\
&\leq \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^g - \bar{Z}_{t-1}^g\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|V_t - V_{t-1}\|_F^2] \\
&= \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^g - \bar{Z}_{t-1}^g\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|(1-\alpha_2\eta^2)(V_{t-1} - \Delta_{t-1}^{g_{\zeta_t}}) + \Delta_t^{g_{\zeta_t}} - V_{t-1}\|_F^2] \\
&= \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^g - \bar{Z}_{t-1}^g\|_F^2] + \frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|\Delta_t^{g_{\zeta_t}} - \Delta_{t-1}^{g_{\zeta_t}} - \alpha_2\eta^2(V_{t-1} - \Delta_{t-1}^g) - \alpha_2\eta^2(\Delta_{t-1}^g - \Delta_{t-1}^{g_{\zeta_t}})\|_F^2] \\
&\overset{(s_2)}{\leq} \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^g - \bar{Z}_{t-1}^g\|_F^2] + \frac{2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|\Delta_t^{g_{\zeta_t}} - \Delta_{t-1}^{g_{\zeta_t}}\|_F^2] + \frac{2\alpha_2^2\eta^4}{1-\lambda}\frac{1}{K}\mathbb{E}[\|V_{t-1} - \Delta_{t-1}^g\|_F^2] + \frac{\alpha_2^2\eta^4\sigma^2}{1-\lambda} \\
&\overset{(s_3)}{\leq} \lambda\frac{1}{K}\mathbb{E}[\|Z_{t-1}^g - \bar{Z}_{t-1}^g\|_F^2] + \frac{2\ell_{g_y}^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|X_t - X_{t-1}\|_F^2] + \frac{2\ell_{g_y}^2}{1-\lambda}\frac{1}{K}\mathbb{E}[\|Y_t - Y_{t-1}\|_F^2] \\
&\quad + \frac{2\alpha_2^2\eta^4}{1-\lambda}\frac{1}{K}\mathbb{E}[\|V_{t-1} - \Delta_{t-1}^g\|_F^2] + \frac{\alpha_2^2\eta^4\sigma^2}{1-\lambda} \ .
\end{aligned} \tag{122}$$

where (s_1) holds due to Lemma 34 with $a = \frac{1-\lambda}{\lambda}$, (s_2) holds due to $\mathbb{E}[\Delta_{t-1}^{g_{\zeta_t}}] = \Delta_{t-1}^g$ and Assumption 3, (s_3) holds due to Assumption 8.

□

Note that Lemmas 10, 11, 12, 13 still hold. Then, based on these lemmas, we begin to prove Theorem 3.

A.4.5 Proof of Theorem 3

Proof. At first, we introduce the following potential function for investigating the convergence rate of Algorithm 2.

$$\begin{aligned}
\mathcal{L}_{t+1} &= \mathbb{E}[F(x_{t+1})] + \frac{6\beta_1 L_F^2}{\beta_2\mu}\mathbb{E}[\|\bar{y}_{t+1} - y^*(\bar{x}_{t+1})\|^2] \\
&\quad + \frac{2\beta_1\ell_{g_y}^2((51+48/\alpha_1 K)L_{\tilde{F}}^2 + (98+800/\alpha_2 K)L_F^2)}{\mu^2(1-\lambda^2)}\frac{1}{K}\mathbb{E}[\|X_{t+1} - \bar{X}_{t+1}\|_F^2] \\
&\quad + \frac{2\beta_1\ell_{g_y}^2((51+48/\alpha_1 K)L_{\tilde{F}}^2 + (98+800/\alpha_2 K)L_F^2)}{\mu^2(1-\lambda^2)}\frac{1}{K}\mathbb{E}[\|Y_{t+1} - \bar{Y}_{t+1}\|_F^2] \\
&\quad + \beta_1(1-\lambda)\frac{1}{K}\mathbb{E}[\|Z_{t+1}^{\tilde{F}} - \bar{Z}_{t+1}^{\tilde{F}}\|_F^2] + \frac{\beta_1(1-\lambda)L_F^2}{\mu^2}\frac{1}{K}\mathbb{E}[\|Z_{t+1}^g - \bar{Z}_{t+1}^g\|_F^2] \\
&\quad + 2\beta_1\frac{1}{K}\mathbb{E}[\|\Delta_{t+1}^{\tilde{F}} - U_{t+1}\|_F^2] + \frac{2\beta_1 L_F^2}{\mu^2}\frac{1}{K}\mathbb{E}[\|\Delta_{t+1}^g - V_{t+1}\|_F^2] \\
&\quad + \frac{3\beta_1}{\alpha_1\eta}\mathbb{E}[\|(\Delta_{t+1}^{\tilde{F}} - U_{t+1})\frac{1}{K}\mathbf{1}\|^2] + \frac{50\beta_1 L_F^2}{\alpha_2\eta\mu^2}\mathbb{E}[\|(\Delta_{t+1}^g - V_{t+1})\frac{1}{K}\mathbf{1}\|^2] \ .
\end{aligned} \tag{123}$$

According to the aforementioned lemmas, Eq. (50), and Eq. (52) where L_{g_y} is replaced with ℓ_{g_y}, we can get

$$\begin{aligned}
&\mathcal{L}_{t+1} - \mathcal{L}_t \\
&\leq -\frac{\eta\beta_1}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] - \frac{\eta\beta_1 L_F^2}{2}\mathbb{E}[\|\bar{y}_t - y^*(\bar{x}_t)\|^2] + \frac{3\eta\beta_1 C_{g_{xy}}^2 C_{f_y}^2}{\mu^2}(1-\frac{\mu}{L_{g_y}})^{2J} \\
&\quad + (\frac{6\beta_1 L_F^2}{\beta_2\mu}\frac{25\eta\beta_1^2 L_y^2}{6\beta_2\mu} - \frac{\eta\beta_1}{4} + 4\eta^2\beta_1^2\tilde{w})\mathbb{E}[\|\bar{u}_t\|^2] + (4\eta^2\beta_2^2\tilde{w} - \frac{6\beta_1 L_F^2}{\beta_2\mu}\frac{3\eta\beta_2^2}{4})\mathbb{E}[\|\bar{v}_t\|^2] \\
&\quad + (\frac{2\eta\beta_1^2}{1-\lambda^2}\frac{2\beta_1\ell_{g_y}^2((51+48/\alpha_1 K)L_{\tilde{F}}^2 + (98+800/\alpha_2 K)L_F^2)}{\mu^2(1-\lambda^2)} - \beta_1(1-\lambda)^2 + 4\eta^2\beta_1^2\tilde{w})\frac{1}{K}\mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}}\|_F^2] \\
&\quad + (\frac{2\eta\beta_2^2}{1-\lambda^2}\frac{2\beta_1\ell_{g_y}^2((51+48/\alpha_1 K)L_{\tilde{F}}^2 + (98+800/\alpha_2 K)L_F^2)}{\mu^2(1-\lambda^2)} - \frac{\beta_1(1-\lambda)^2 L_F^2}{\mu^2} + 4\eta^2\beta_2^2\tilde{w})\frac{1}{K}\mathbb{E}[\|Z_t^{g} - \bar{Z}_t^{g}\|_F^2] \\
&\quad + \beta_1\alpha_1^2\eta^4\sigma_{\tilde{F}}^2 + \frac{\beta_1\alpha_2^2\eta^4\sigma^2 L_F^2}{\mu^2} + 4\beta_1\alpha_1^2\eta^3\sigma_{\tilde{F}}^2 + \frac{4\beta_1\alpha_2^2\eta^3\sigma^2 L_F^2}{\mu^2} + \frac{6\beta_1\alpha_1\eta^3\sigma_{\tilde{F}}^2}{K} + \frac{100\beta_1\alpha_2\eta^3\sigma^2 L_F^2}{\mu^2 K} \ ,
\end{aligned} \tag{124}$$

where $\tilde{w} = 2\beta_1 L_{\tilde{F}}^2 + \frac{2\beta_1 L_F^2\ell_{g_y}^2}{\mu^2} + \frac{4\beta_1 L_{\tilde{F}}^2}{\eta} + \frac{4\beta_1 L_F^2\ell_{g_y}^2}{\eta\mu^2} + \frac{\beta_1(6/\alpha_1 K)L_{\tilde{F}}^2}{\eta} + \frac{(100/\alpha_2 K)\beta_1 L_F^2\ell_{g_y}^2}{\eta\mu^2}$.

By setting the coefficient of $\mathbb{E}[\|Z_t^{\tilde{F}} - \bar{Z}_t^{\tilde{F}}\|_F^2]$ to be non-positive, we can get

$$\begin{aligned}
&\frac{2\beta_1^2}{1-\lambda^2}\frac{2\ell_{g_y}^2((51+48/\alpha_1 K)L_{\tilde{F}}^2 + (98+800/\alpha_2 K)L_F^2)}{\mu^2(1-\lambda^2)} - (1-\lambda)^2 \\
&\quad + 4\eta\beta_1^2\Big(\frac{2L_{\tilde{F}}^2\ell_{g_y}^2}{\mu^2} + \frac{2L_F^2\ell_{g_y}^2}{\mu^2} + \frac{4L_{\tilde{F}}^2\ell_{g_y}^2}{\eta\mu^2} + \frac{4L_F^2\ell_{g_y}^2}{\eta\mu^2} + \frac{(6/\alpha_1 K)L_{\tilde{F}}^2\ell_{g_y}^2}{\eta\mu^2} + \frac{(100/\alpha_2 K)L_F^2\ell_{g_y}^2}{\eta\mu^2}\Big) \leq 0 \ .
\end{aligned} \tag{125}$$

Due to $\eta < 1$ and $\lambda < 1$, we can get

$$\begin{aligned}
&\frac{2\beta_1^2}{1-\lambda^2}\frac{2\ell_{g_y}^2((51+48/\alpha_1 K)L_{\tilde{F}}^2 + (98+800/\alpha_2 K)L_F^2)}{\mu^2(1-\lambda^2)} - (1-\lambda)^2 \\
&\quad + 4\eta\beta_1^2\Big(\frac{2L_{\tilde{F}}^2\ell_{g_y}^2}{\mu^2} + \frac{2L_F^2\ell_{g_y}^2}{\mu^2} + \frac{4L_{\tilde{F}}^2\ell_{g_y}^2}{\eta\mu^2} + \frac{4L_F^2\ell_{g_y}^2}{\eta\mu^2} + \frac{(6/\alpha_1 K)L_{\tilde{F}}^2\ell_{g_y}^2}{\eta\mu^2} + \frac{(100/\alpha_2 K)L_F^2\ell_{g_y}^2}{\eta\mu^2}\Big) \\
&\leq \frac{2\beta_1^2}{1-\lambda^2}\frac{2\ell_{g_y}^2((51+48/\alpha_1 K)L_{\tilde{F}}^2 + (98+800/\alpha_2 K)L_F^2)}{\mu^2(1-\lambda^2)} - (1-\lambda)^2 \\
&\quad + 4\beta_1^2\Big(\frac{2L_{\tilde{F}}^2\ell_{g_y}^2}{\mu^2} + \frac{2L_F^2\ell_{g_y}^2}{\mu^2} + \frac{4L_{\tilde{F}}^2\ell_{g_y}^2}{\mu^2} + \frac{4L_F^2\ell_{g_y}^2}{\mu^2} + \frac{(6/\alpha_1 K)L_{\tilde{F}}^2\ell_{g_y}^2}{\mu^2} + \frac{(100/\alpha_2 K)L_F^2\ell_{g_y}^2}{\mu^2}\Big) \\
&\leq \frac{4\beta_1^2\ell_{g_y}^2((51+48/\alpha_1 K)L_{\tilde{F}}^2 + (98+800/\alpha_2 K)L_F^2)}{\mu^2(1-\lambda^2)^2} - (1-\lambda)^2 + 4\beta_1^2\ell_{g_y}^2\frac{(6+6/\alpha_1 K)L_{\tilde{F}}^2 + (6+100/\alpha_2 K)L_F^2}{\mu^2} \\
&\leq 4\beta_1^2\ell_{g_y}^2\frac{(57+54/\alpha_1 K)L_{\tilde{F}}^2 + (104+900/\alpha_2 K)L_F^2}{\mu^2(1-\lambda^2)^2} - (1-\lambda)^2 \ .
\end{aligned} \tag{126}$$

By letting this upper bound non-positive, we can get

$$\begin{aligned}
&4\beta_1^2\ell_{g_y}^2\frac{(57+54/\alpha_1 K)L_{\tilde{F}}^2 + (104+900/\alpha_2 K)L_F^2}{\mu^2(1-\lambda^2)^2} \leq (1-\lambda)^2 \\
&\beta_1 \leq \frac{\mu(1-\lambda)^2}{2\ell_{g_y}\sqrt{(57+54/\alpha_1 K)L_{\tilde{F}}^2 + (104+900/\alpha_2 K)L_F^2}} \ .
\end{aligned} \tag{127}$$

By setting the coefficient of $\mathbb{E}[\|Z_t^{g} - \bar{Z}_t^{g}\|_F^2]$ to be non-positive, we can get

$$\begin{aligned}
&\frac{2\beta_2^2}{1-\lambda^2}\frac{2\ell_{g_y}^2((51+48/\alpha_1 K)L_{\tilde{F}}^2 + (98+800/\alpha_2 K)L_F^2)}{\mu^2(1-\lambda^2)} - \frac{(1-\lambda)^2 L_F^2}{\mu^2} \\
&\quad + 4\eta\beta_2^2\Big(\frac{2L_{\tilde{F}}^2\ell_{g_y}^2}{\mu^2} + \frac{2L_F^2\ell_{g_y}^2}{\mu^2} + \frac{4L_{\tilde{F}}^2\ell_{g_y}^2}{\eta\mu^2} + \frac{4L_F^2\ell_{g_y}^2}{\eta\mu^2} + \frac{(6/\alpha_1 K)L_{\tilde{F}}^2\ell_{g_y}^2}{\eta\mu^2} + \frac{(100/\alpha_2 K)L_F^2\ell_{g_y}^2}{\eta\mu^2}\Big) \leq 0 \ .
\end{aligned} \tag{128}$$

Due to $\eta < 1$, $\lambda < 1$, and $\ell_{g_y}/\mu > 1$, we can get

$$
\begin{aligned}
&\frac{2\beta_2^2}{1-\lambda^2}\frac{2\ell_{g_y}^2((51+48/\alpha_1 K)L_{\tilde{F}}^2+(98+800/\alpha_2 K)L_F^2)}{\mu^2(1-\lambda^2)} - \frac{(1-\lambda)^2 L_F^2}{\mu^2} \\
&\quad + 4\eta\beta_2^2\Big(\frac{2L_{\tilde{F}}^2\ell_{g_y}^2}{\mu^2} + \frac{2L_F^2\ell_{g_y}^2}{\mu^2} + \frac{4L_{\tilde{F}}^2\ell_{g_y}^2}{\eta\mu^2} + \frac{4L_F^2\ell_{g_y}^2}{\eta\mu^2} + \frac{(6/\alpha_1 K)L_{\tilde{F}}^2\ell_{g_y}^2}{\eta\mu^2} + \frac{(100/\alpha_2 K)L_F^2\ell_{g_y}^2}{\eta\mu^2}\Big) \\
&\leq \frac{4\beta_2^2\ell_{g_y}^2((51+48/\alpha_1 K)L_{\tilde{F}}^2+(98+800/\alpha_2 K)L_F^2)}{\mu^2(1-\lambda^2)^2} - \frac{(1-\lambda)^2 L_F^2}{\mu^2} \\
&\quad + 4\beta_2^2\ell_{g_y}^2\frac{(6+6/\alpha_1 K)L_{\tilde{F}}^2+(6+100/\alpha_2 K)L_F^2}{\mu^2} \\
&\leq 4\beta_2^2\ell_{g_y}^2\frac{(57+54/\alpha_1 K)L_{\tilde{F}}^2+(104+900/\alpha_2 K)L_F^2}{\mu^2(1-\lambda^2)^2} - \frac{(1-\lambda)^2 L_F^2}{\mu^2} \,.
\end{aligned} \tag{129}
$$

By letting this upper bound non-positive, we can get

$$
\beta_2 \leq \frac{(1-\lambda)^2 L_F}{2\ell_{g_y}\sqrt{(57+54/\alpha_1 K)L_{\tilde{F}}^2+(104+900/\alpha_2 K)L_F^2}} \,. \tag{130}
$$

By setting the coefficient of $\mathbb{E}[\|\bar{u}_t\|^2]$ to be non-positive, we can get

$$
\frac{6L_F^2}{\beta_2\mu}\frac{25\beta_1^2 L_y^2}{6\beta_2\mu} - \frac{1}{4} + 4\eta\beta_1^2\Big(\frac{2L_{\tilde{F}}^2\ell_{g_y}^2}{\mu^2} + \frac{2L_F^2\ell_{g_y}^2}{\mu^2} + \frac{4L_{\tilde{F}}^2\ell_{g_y}^2}{\eta\mu^2} + \frac{4L_F^2\ell_{g_y}^2}{\eta\mu^2} + \frac{(6/\alpha_1 K)L_{\tilde{F}}^2\ell_{g_y}^2}{\eta\mu^2} + \frac{(100/\alpha_2 K)L_F^2\ell_{g_y}^2}{\eta\mu^2}\Big) \leq 0 \,. \tag{131}
$$

Due to $\eta_1 < 1$ and $\ell_{g_y}/\mu > 1$, we can get

$$
\begin{aligned}
&\frac{6L_F^2}{\beta_2\mu}\frac{25\beta_1^2 L_y^2}{6\beta_2\mu} - \frac{1}{4} + 4\eta\beta_1^2\Big(\frac{2L_{\tilde{F}}^2\ell_{g_y}^2}{\mu^2} + \frac{2L_F^2\ell_{g_y}^2}{\mu^2} + \frac{4L_{\tilde{F}}^2\ell_{g_y}^2}{\eta\mu^2} + \frac{4L_F^2\ell_{g_y}^2}{\eta\mu^2} + \frac{(6/\alpha_1 K)L_{\tilde{F}}^2\ell_{g_y}^2}{\eta\mu^2} + \frac{(100/\alpha_2 K)L_F^2\ell_{g_y}^2}{\eta\mu^2}\Big) \\
&\leq \frac{6L_F^2}{\beta_2\mu}\frac{25\beta_1^2 L_y^2}{6\beta_2\mu} - \frac{1}{4} + 4\beta_1^2\ell_{g_y}^2\frac{(6+6/\alpha_1 K)L_{\tilde{F}}^2+(6+100/\alpha_2 K)L_F^2}{\mu^2} \,.
\end{aligned} \tag{132}
$$

By setting $\beta_1 \leq \frac{\beta_2\mu}{15L_F L_y}$, we can get $\frac{6L_F^2}{\beta_2\mu}\frac{25\beta_1^2 L_y^2}{6\beta_2\mu} - \frac{1}{4} \leq -\frac{1}{8}$. Then, we have

$$
\begin{aligned}
&\frac{6L_F^2}{\beta_2\mu}\frac{25\beta_1^2 L_y^2}{6\beta_2\mu} - \frac{1}{4} + 4\beta_1^2\ell_{g_y}^2\frac{(6+6/\alpha_1 K)L_{\tilde{F}}^2+(6+100/\alpha_2 K)L_F^2}{\mu^2} \\
&\leq 4\beta_1^2\ell_{g_y}^2\frac{(6+6/\alpha_1 K)L_{\tilde{F}}^2+(6+100/\alpha_2 K)L_F^2}{\mu^2} - \frac{1}{8} \,.
\end{aligned} \tag{133}
$$

By letting this upper bound non-positive, we can get

$$
\beta_1 \leq \frac{\mu}{8\ell_{g_y}\sqrt{(3+3/\alpha_1 K)L_{\tilde{F}}^2+(3+50/\alpha_2 K)L_F^2}} \,. \tag{134}
$$

By setting the coefficient of $\mathbb{E}[\|\bar{v}_t\|^2]$ to be non-positive, we can get

$$
4\eta\beta_2^2\beta_1\Big(\frac{2L_{\tilde{F}}^2\ell_{g_y}^2}{\mu^2} + \frac{2L_F^2\ell_{g_y}^2}{\mu^2} + \frac{4L_{\tilde{F}}^2\ell_{g_y}^2}{\eta\mu^2} + \frac{4L_F^2\ell_{g_y}^2}{\eta\mu^2} + \frac{(6/\alpha_1 K)L_{\tilde{F}}^2\ell_{g_y}^2}{\eta\mu^2} + \frac{(100/\alpha_2 K)L_F^2\ell_{g_y}^2}{\eta\mu^2}\Big) - \frac{6\beta_1 L_F^2}{\beta_2\mu}\frac{3\beta_2^2}{4} \leq 0 \,. \tag{135}
$$

Due to $\eta < 1$ and $\ell_{g_y}/\mu > 1$, we can get

$$
\begin{aligned}
&4\eta\beta_2^2\beta_1\Big(\frac{2L_{\tilde{F}}^2\ell_{g_y}^2}{\mu^2} + \frac{2L_F^2\ell_{g_y}^2}{\mu^2} + \frac{4L_{\tilde{F}}^2\ell_{g_y}^2}{\eta\mu^2} + \frac{4L_F^2\ell_{g_y}^2}{\eta\mu^2} + \frac{(6/\alpha_1 K)L_{\tilde{F}}^2\ell_{g_y}^2}{\eta\mu^2} + \frac{(100/\alpha_2 K)L_F^2\ell_{g_y}^2}{\eta\mu^2}\Big) - \frac{6\beta_1 L_F^2}{\beta_2\mu}\frac{3\beta_2^2}{4} \\
&\leq 4\beta_2^2\beta_1\ell_{g_y}^2\frac{(6+6/\alpha_1 K)L_{\tilde{F}}^2+(6+100/\alpha_2 K)L_F^2}{\mu^2} - \frac{6\beta_1 L_F^2}{\beta_2\mu}\frac{3\beta_2^2}{4} \,.
\end{aligned} \tag{136}
$$

By letting this upper bound non-positive, we can get

$$\beta_2 \leq \frac{9\mu L_F^2}{8\ell_{g_y}^2((6+6/\alpha_1 K)L_{\tilde{F}}^2 + (6+100/\alpha_2 K)L_F^2)} \ . \tag{137}$$

Therefore, by combining all these conditions, we can get

$$\begin{aligned}
&\beta_1 \leq \min\Big\{\frac{\beta_2\mu}{15L_F L_y}, \frac{\mu}{8\ell_{g_y}\sqrt{(3+3/\alpha_1 K)L_{\tilde{F}}^2 + (3+50/\alpha_2 K)L_F^2}}, \frac{\mu(1-\lambda)^2}{2\ell_{g_y}\sqrt{(57+54/\alpha_1 K)L_{\tilde{F}}^2 + (104+900/\alpha_2 K)L_F^2}}\Big\} , \\
&\beta_2 \leq \min\Big\{\frac{(1-\lambda)^2 L_F}{2\ell_{g_y}\sqrt{(57+54/\alpha_1 K)L_{\tilde{F}}^2 + (104+900/\alpha_2 K)L_F^2}}, \frac{9\mu L_F^2}{8\ell_{g_y}^2((6+6/\alpha_1 K)L_{\tilde{F}}^2 + (6+100/\alpha_2 K)L_F^2)}\Big\} \ .
\end{aligned} \tag{138}$$

As a result, we can get

$$\begin{aligned}
&\mathcal{L}_{t+1} - \mathcal{L}_t \\
&\leq -\frac{\eta\beta_1}{2}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2] - \frac{\eta\beta_1 L_F^2}{2}\mathbb{E}[\|\bar{y}_t - y^*(\bar{x}_t)\|^2] + \frac{3\eta\beta_1 C_{g_{xy}}^2 C_{f_y}^2}{\mu^2}(1-\frac{\mu}{L_{g_y}})^{2J} \\
&\quad + \beta_1\alpha_1^2\eta^4\sigma_{\tilde{F}}^2 + \frac{\beta_1\alpha_2^2\eta^4\sigma^2 L_F^2}{\mu^2} + 4\beta_1\alpha_1^2\eta^3\sigma_{\tilde{F}}^2 + \frac{4\beta_1\alpha_2^2\eta^3\sigma^2 L_F^2}{\mu^2} + \frac{6\beta_1\alpha_1\eta^3\sigma_{\tilde{F}}^2}{K} + \frac{100\beta_1\alpha_2\eta^3\sigma^2 L_F^2}{\mu^2 K} \ .
\end{aligned} \tag{139}$$

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

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2 + L_F^2\|\bar{y}_t - y^*(\bar{x}_t)\|^2] \\
&\leq \frac{2(\mathcal{L}_0 - \mathcal{L}_T)}{\eta\beta_1 T} + \frac{6C_{g_{xy}}^2 C_{f_y}^2}{\mu^2}(1-\frac{\mu}{\ell_{g_y}})^{2J} \\
&\quad + 2\alpha_1^2\eta^3\sigma_{\tilde{F}}^2 + \frac{2\alpha_2^2\eta^3\sigma^2 L_F^2}{\mu^2} + 8\alpha_1^2\eta^2\sigma_{\tilde{F}}^2 + \frac{8\alpha_2^2\eta^2\sigma^2 L_F^2}{\mu^2} + \frac{12\alpha_1\eta^2\sigma_{\tilde{F}}^2}{K} + \frac{200\alpha_2\eta^2\sigma^2 L_F^2}{\mu^2 K} \ .
\end{aligned} \tag{140}$$

In terms of the initialisation, we can get

$$\begin{aligned}
\mathcal{L}_0 &= \mathbb{E}[F(x_0)] + \frac{6\beta_1 L_F^2}{\beta_2\mu}\mathbb{E}[\|\bar{y}_0 - y^*(\bar{x}_0)\|^2] \\
&\quad + \beta_1(1-\lambda)\frac{1}{K}\mathbb{E}[\|Z_0^{\tilde{F}} - \bar{Z}_0^{\tilde{F}}\|_F^2] + \frac{\beta_1(1-\lambda)L_F^2}{\mu^2}\frac{1}{K}\mathbb{E}[\|Z_0^g - \bar{Z}_0^g\|_F^2] \\
&\quad + 2\beta_1\frac{1}{K}\mathbb{E}[\|\Delta_0^{\tilde{F}} - U_0\|_F^2] + \frac{2\beta_1 L_F^2}{\mu^2}\frac{1}{K}\mathbb{E}[\|\Delta_0^g - V_0\|_F^2] \\
&\quad + \frac{3\beta_1}{\alpha_1\eta}\mathbb{E}[\|(\Delta_0^{\tilde{F}} - U_0)\frac{1}{K}\mathbf{1}\|^2] + \frac{50\beta_1 L_F^2}{\alpha_2\eta\mu^2}\mathbb{E}[\|(\Delta_0^g - V_0)\frac{1}{K}\mathbf{1}\|^2] \ .
\end{aligned} \tag{141}$$

As for $\frac{1}{K}\mathbb{E}[\|Z_0^{\tilde{F}}-\bar{Z}_0^{\tilde{F}}\|_F^2]$, we have

$$
\begin{aligned}
&\frac{1}{K}\mathbb{E}[\|Z_0^{\tilde{F}}-\bar{Z}_0^{\tilde{F}}\|_F^2] \\
&= \frac{1}{K}\mathbb{E}[\|\Delta_0^{\tilde{F}_{\tilde{\xi}_0}}-\bar{\Delta}_0^{\tilde{F}_{\tilde{\xi}_0}}\|_F^2] \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla\tilde{F}^{(k)}(x_0,y_0;\tilde{\xi}_0^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}\nabla\tilde{F}^{(k')}(x_0,y_0;\tilde{\xi}_0^{(k')})\|_F^2] \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla\tilde{F}^{(k)}(x_0,y_0;\tilde{\xi}_0^{(k)})-\nabla\tilde{F}^{(k)}(x_0,y_0)+\nabla\tilde{F}^{(k)}(x_0,y_0) \\
&\quad -\frac{1}{K}\sum_{k'=1}^{K}\nabla\tilde{F}^{(k')}(x_0,y_0)+\frac{1}{K}\sum_{k'=1}^{K}\nabla\tilde{F}^{(k')}(x_0,y_0)-\frac{1}{K}\sum_{k'=1}^{K}\nabla\tilde{F}^{(k')}(x_0,y_0;\tilde{\xi}_0^{(k')})\|_F^2] \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla\tilde{F}^{(k)}(x_0,y_0;\tilde{\xi}_0^{(k)})-\nabla\tilde{F}^{(k)}(x_0,y_0)+\frac{1}{K}\sum_{k'=1}^{K}\nabla\tilde{F}^{(k')}(x_0,y_0)-\frac{1}{K}\sum_{k'=1}^{K}\nabla\tilde{F}^{(k')}(x_0,y_0;\tilde{\xi}_0^{(k')})\|_F^2] \\
&\leq 2\sigma_{\tilde{F}}^2 \ .
\end{aligned}
\tag{142}
$$

Similarly, we can get $\frac{1}{K}\mathbb{E}[\|Z_0^{g}-\bar{Z}_0^{g}\|_F^2]\leq 2\sigma^2$, $\frac{1}{K}\mathbb{E}[\|\Delta_0^{\tilde{F}}-U_0\|_F^2]=\frac{1}{K}\mathbb{E}[\|\Delta_0^{\tilde{F}}-\Delta_0^{\tilde{F}_{\tilde{\xi}_0}}\|_F^2]\leq\sigma_{\tilde{F}}^2$, $\frac{1}{K}\mathbb{E}[\|\Delta_0^{g}-V_0\|_F^2]\leq\sigma^2$, $\mathbb{E}[\|(\Delta_0^{\tilde{F}}-U_0)\frac{1}{K}\mathbf{1}\|^2]=\mathbb{E}[\|(\Delta_0^{\tilde{F}}-\Delta_0^{\tilde{F}_{\tilde{\xi}_0}})\frac{1}{K}\mathbf{1}\|^2]\leq\frac{\sigma_{\tilde{F}}^2}{K}$, $\mathbb{E}[\|(\Delta_0^{g}-V_0)\frac{1}{K}\mathbf{1}\|^2]\leq\frac{\sigma^2}{K}$. When the mini-batch size is set to B, we can get

$$
\begin{aligned}
\mathcal{L}_0 &\leq \mathbb{E}[F(x_0)]+\frac{6\beta_1L_F^2}{\beta_2\mu}\mathbb{E}[\|\bar{y}_0-y^*(\bar{x}_0)\|^2] \\
&\quad+\frac{2\beta_1\sigma_{\tilde{F}}^2}{B}+\frac{2\beta_1L_F^2\sigma^2}{B\mu^2}+\frac{2\beta_1\sigma_{\tilde{F}}^2}{B}+\frac{2\beta_1L_F^2\sigma^2}{B\mu^2}+\frac{3\beta_1}{\alpha_1\eta}\frac{\sigma_{\tilde{F}}^2}{BK}+\frac{50\beta_1L_F^2}{\alpha_2\eta\mu^2}\frac{\sigma^2}{BK} \\
&= \mathbb{E}[F(x_0)]+\frac{6\beta_1L_F^2}{\beta_2\mu}\mathbb{E}[\|\bar{y}_0-y^*(\bar{x}_0)\|^2]+\frac{4\beta_1\sigma_{\tilde{F}}^2}{B}+\frac{4\beta_1L_F^2\sigma^2}{B\mu^2}+\frac{3\beta_1}{\alpha_1\eta}\frac{\sigma_{\tilde{F}}^2}{BK}+\frac{50\beta_1L_F^2}{\alpha_2\eta\mu^2}\frac{\sigma^2}{BK} \ .
\end{aligned}
\tag{143}
$$

Then, we can get

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}[\|\nabla F(\bar{x}_t)\|^2+L_F^2\|\bar{y}_t-y^*(\bar{x}_t)\|^2]) \\
&\leq \frac{2(F(x_0)-F(x_*))}{\eta\beta_1T}+\frac{12L_F^2}{\eta\beta_2T\mu}\mathbb{E}[\|\bar{y}_0-y^*(\bar{x}_0)\|^2]+\frac{6C_{g_{xy}}^2C_{f_y}^2}{\mu^2}(1-\frac{\mu}{\ell_{g_y}})^{2J} \\
&\quad+\frac{8\sigma_{\tilde{F}}^2}{\eta TB}+\frac{8L_F^2\sigma^2}{\eta TB\mu^2}+\frac{6\sigma_{\tilde{F}}^2}{\alpha_1\eta^2TBK}+\frac{100L_F^2\sigma^2}{\alpha_2\eta^2TBK\mu^2} \\
&\quad+2\alpha_1^2\eta^3\sigma_{\tilde{F}}^2+\frac{2\alpha_2^2\eta^3\sigma^2L_F^2}{\mu^2}+8\alpha_1^2\eta^2\sigma_{\tilde{F}}^2+\frac{8\alpha_2^2\eta^2\sigma^2L_F^2}{\mu^2}+\frac{12\alpha_1\eta^2\sigma_{\tilde{F}}^2}{K}+\frac{200\alpha_2\eta^2\sigma^2L_F^2}{\mu^2K} \ ,
\end{aligned}
\tag{144}
$$

which completes the proof.

□

A.4.6 Proof of Corollary 3

Corollary 3 can be proved by following the proof of Corollary 1.

A.5 Additional Lemmas

Lemma 34. *For any matrices $X\in\mathbb{R}^{m\times n}$, $Y\in\mathbb{R}^{m\times n}$, the following inequality hold*

$$
\|X+Y\|_F^2\leq(1+a)\|X\|_F^2+(1+\frac{1}{a})\|Y\|_F^2 \ , \tag{145}
$$

for any $a>0$.