Nonconvex Decentralized Stochastic Bilevel Optimization under Heavy-Tailed Noise

Xinwen Zhang[*]
Temple University
Philadelphia, PA, USA
ellenz@temple.edu

Yihan Zhang[*]
Temple University
Philadelphia, PA, USA
yihan.zhang0002@temple.edu

Heng Liang
Temple University
Philadelphia, PA, USA
heng.liang@temple.edu

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

Abstract

Existing decentralized stochastic optimization methods assume the lower-level loss function is strongly convex and the stochastic gradient noise has finite variance. These strong assumptions typically are not satisfied in real-world machine learning models. For example, learning on language data typically leads to heavy-tailed gradient. To address these limitations, we develop a novel decentralized stochastic bilevel optimization algorithm for the nonconvex bilevel optimization problem under heavy-tailed noise. Specifically, we develop a normalized stochastic variance-reduced bilevel gradient descent algorithm, which does not rely on any clipping operation. Moreover, we establish its convergence rate by innovatively bounding interdependent gradient sequences under heavy-tailed noise for nonconvex decentralized bilevel optimization problems. As far as we know, this is the first decentralized bilevel optimization algorithm with rigorous theoretical guarantees under heavy-tailed noise. The extensive experimental results confirm the effectiveness of our algorithm in handling heavy-tailed noise.

1 Introduction

Stochastic bilevel optimization consists of two levels of optimization subproblems, where the upper-level subproblem depends on the optimal solution of the lower-level subproblem. It has received a surge of attention in recent years because it lays the optimization foundation for a series of machine learning models, such as model-agnostic meta-learning [10], hyperparameter optimization [11, 42], imbalanced data classification [58], reinforcement learning [47, 27], large language models [45, 28], etc. To facilitate stochastic bilevel optimization for distributed machine learning models, where data are distributed across different workers, a series of decentralized stochastic bilevel optimization algorithms have been developed in recent years. Specifically, in a decentralized setting, each device computes stochastic gradients based on its local training data to update the variables of both the upper-level and lower-level subproblems, and then communicates these updates with neighboring workers in a peer-to-peer manner.

Compared to traditional single-level optimization problems, a unique challenge in decentralized stochastic bilevel optimization lies in computing the stochastic hypergradient, that is, the stochastic

[*]Equal contributions.
[†]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

gradient of the upper-level loss function with respect to its variable. This challenge is caused by the unique characteristic of bilevel optimization: the upper-level subproblem relies on the optimal solution of the lower-level subproblem, and therefore the local hypergradient requires the global Hessian inverse matrix. To address this challenge, three categories of decentralized stochastic bilevel optimization algorithms [57, 13, 5, 6, 66, 22, 69, 37, 34, 35, 53, 43, 64] have been developed. The first category, such as [57], uses the Neumann series expansion approach to approximate the Hessian inverse on each device and then communicates it between workers, suffering from high communication costs. The second category, such as [66, 69], estimates the Hessian-inverse-vector product by solving an auxiliary quadratic optimization problem with gradient descent on each device and then communicates this estimator, which helps reduce communication costs. However, both the first and second categories incur significant computational overhead due to the need to compute second-order Hessian information. The third category, such as [53], addresses this challenge by reformulating the decentralized stochastic bilevel problem as a single-level optimization problem and then solving it with only first-order gradients. By avoiding the computation of second-order gradients, this category significantly reduces computational overhead.

However, existing decentralized stochastic bilevel optimization algorithms suffer from significant limitations. First, these algorithms require the loss function of the lower-level subproblem to be strongly convex. This strong assumption is not satisfied by most practical machine learning models, such as deep neural networks, which are inherently nonconvex. To address this issue, several recent works in the single-machine setting have investigated nonconvex bilevel optimization algorithms [29, 55, 46, 23, 4, 3]. Second, they assume the stochastic noise in the gradient has finite variance. However, existing studies [48, 60] have demonstrated that this bounded variance assumption does not hold for the commonly used deep neural networks. In practice, the stochastic noise often follows a heavy-tailed distribution. Moreover, [26, 21, 52] show that learning on language data typically leads to heavy-tailed gradients because different words have different frequencies. Hence, in these practical scenarios, existing algorithmic designs and theoretical foundations for decentralized bilevel optimization no longer hold. It is therefore necessary to develop new decentralized stochastic bilevel optimization algorithms that can accommodate a broader range of machine learning models and provide solid theoretical guarantees. To this end, the goal of this paper is to develop an efficient decentralized stochastic bilevel optimization algorithm for *nonconvex* bilevel problems under *heavy-tailed noise*, with rigorous theoretical guarantees. Since the first-order methods in the aforementioned third category offer low computational overhead and communication costs, this paper focuses on the first-order method.

For standard single-level optimization problems in the single-machine setting, a commonly used approach to handling heavy-tailed noise is Clipped SGD [60], which mitigates the effect of heavy-tailed noise by clipping the norm of the stochastic gradient below a predefined threshold. Nevertheless, tuning the clipping threshold can be challenging. Recently, several works [36, 18, 50] have shown that the gradient normalization technique is sufficient to guarantee the convergence of stochastic gradient descent in-expectation under heavy-tailed noise without assuming bounded gradients. For instance, [18] proves that the batched normalized SGD (batched-NSGD) can converge for a smooth nonconvex minimization problem under heavy-tailed noise. However, it requires a large batch size. [50] achieves a similar conclusion for NSGD with momentum using a stronger assumption: the individual Lipschitz smoothness. [36] established the in-expectation convergence rate of the batched normalized stochastic gradient descent with momentum (batched-NSGDM) algorithm under heavy-tailed noise by innovatively bounding the accumulated noise from an online learning perspective.

Since the aforementioned approaches focus solely on single-level optimization in a single-machine setting, they are not applicable to decentralized stochastic bilevel optimization problems. In practice, this setting presents several unique challenges, outlined as follows.

1. In bilevel optimization, **multiple gradients interact with one another**. Each of these gradients is affected by the heavy-tailed noise, which in turn impacts convergence. Therefore, it is challenging to control all of them and establish a convergence rate under heavy-tailed noise.
2. In the decentralized setting, **the consensus error with respect to gradients is also affected by heavy-tailed noise**. It remains unclear how to design algorithms and analyses that effectively control this noise to ensure convergence.
3. The aforementioned first-order method for bilevel optimization requires advanced gradient estimators, such as the variance-reduced gradient, to avoid the quite slow convergence rate

under the finite variance assumption, as shown in [23]. However, **no existing work for bilevel optimization problems in both single-machine and distributed settings has demonstrated that the advanced gradient estimator can ensure convergence under heavy-tailed noise**.

In summary, it is challenging to achieve a fast convergence rate for the first-order gradient-based decentralized bilevel optimization algorithm under heavy-tailed noise. To address these unique challenges, we develop a novel decentralized normalized stochastic gradient with variance reduction algorithm to solve Eq. (1). Rather than using gradient clipping, our algorithm **requires only normalized first-order gradients**, making it more efficient and effective in handling heavy-tailed noise, which is lacking in existing second-order-based methods. Importantly, our algorithm demonstrates how gradient normalization should be applied in decentralized bilevel optimization. **To the best of our knowledge, this is the first algorithm capable of handling heavy-tailed noise in bilevel optimization without using gradient clipping**. We further establish the convergence rate of the developed algorithm under heavy-tailed noise. Specifically, to address challenges arising from the interaction between gradients of different variables, we explicitly characterize their interdependence by innovatively handling the optimization subproblems associated with each variable. In addition, we provide a novel analysis of the consensus errors related to these gradients, which are also influenced by heavy-tailed noise. **To the best of our knowledge, this is the first work to bound interdependent gradient sequences under heavy-tailed noise in bilevel optimization**. Finally, the established convergence rate clearly illustrates how the properties of a decentralized system influence overall convergence, and extensive experimental results validate the effectiveness of the proposed algorithm in handling heavy-tailed noise.

2 Related Work

2.1 Decentralized Stochastic Bilevel Optimization

Decentralized stochastic bilevel optimization enables the decentralized optimization framework for bilevel optimization problems. Due to the two-level characteristics of this problem, there are some unique challenges for computation and communication compared to the decentralization of traditional single-level optimization problems. Specifically, the hypergradient on each worker relies on the global Jacobian matrix and the inverse of the global Hessian matrix. Directly communicating or computing them on each worker can result in a large communication and computation overhead, such as [57, 5] in the aforementioned first category, which communicate Jacobian or Hessian matrix in each iteration. To avoid this issue, [66] developed the first single-loop decentralized algorithm, which computes and communicates the Hessian-inverse-vector product to reduce both computation and communication overhead. This approach has also been applied to the full gradient method [9], stochastic gradient [69], and the momentum-based method [22]. However, these methods require computing the second-order Jacobian and Hessian matrix, which can incur large memory and computation overhead for high-dimensional problems. To avoid computing second-order gradients, in the single-machine setting, [46, 24, 23, 4, 39, 38] propose converting the bilevel optimization problem into a single-level optimization problem via the penalty approach and then only the first-order gradient is needed to solve it, which can save computation overhead significantly. Based on this reformulation, [53] developed a decentralized first-order method, which only requires the standard stochastic gradient. Therefore, its practical computational time is much smaller than that of second-order gradient-based methods. However, [53] still suffers from some limitations. On the one hand, it can only handle the strongly-convex lower-level loss function, which is also a limitation of all aforementioned decentralized methods [57, 13, 5, 6, 66, 22, 69, 37, 34, 35, 53]. On the other hand, [53] suffers from a quite slow convergence rate, $O(1/T^{1/7})$, where T is the number of iterations, while the first-order method [23] in the single-machine setting can achieve a convergence rate of $O(1/T^{1/5})$. Recently, [68] achieved this convergence rate in the decentralized setting. Finally, it is worth noting that all existing decentralized bilevel optimization methods assume that the stochastic noise in the gradient has finite variance. Therefore, these algorithms cannot handle heavy-tailed noise.

2.2 Stochastic Optimization under Heavy-Tailed Noise

Some recent works [60] have shown that the finite variance assumption is too restrictive for modern machine learning models. In practice, commonly used deep neural networks, such as image classification models [49, 2] and attention-based models [60, 1], have stochastic gradients

whose noise follows a heavy-tailed distribution. This observation has sparked the recent interest [60, 7, 33, 40, 31, 36, 18, 50, 15, 19] in the study of stochastic optimization under heavy-tailed noise. For example, [60] established the in-expectation convergence rate of Clipped SGD for strongly convex and nonconvex loss functions. As discussed earlier, Clipped SGD requires a clipping threshold, which introduces more difficulties for tuning the optimizer. Therefore, some recent efforts [36, 18, 50] have been made to get rid of the clipping operation, while keeping the normalization operation. For example, [50] established the in-expectation convergence rate of normalized SGD with momentum based on a strong assumption of the individual Lipschitz smoothness. [18] also achieved this result for normalized SGD without using this strong assumption in the cost of a large batch size. However, establishing the convergence rate of normalized SGD with momentum under mild conditions is not trivial. Recently, [36] developed an innovative approach from the online learning perspective and successfully addressed this issue, establishing the in-expectation convergence rate of normalized SGD without relying on the strong smoothness assumption or a large batch size. However, it remains unclear whether the approach in [36] can be applied to the normalized SGD with variance reduction.

In the distributed setting, the heavy-tailed noise has been less studied, although [16] has shown that noise in the decentralized setting tends to have heavier tails than in the centralized setting. Moreover, existing distributed methods for handling heavy-tailed noise [44, 56, 25] still rely on the gradient clipping technique. Recently, [62, 61] handle heavy-tailed noise without gradient clipping, but they are limited to federated learning. The recent work [51, 59] studied gradient normalization for decentralized nonconvex minimization problems. However, it remains unclear whether it works for decentralized bilevel optimization. Furthermore, to the best of our knowledge, gradient normalization without clipping has not yet been explored for decentralized bilevel optimization or decentralized minimax optimization under heavy-tailed noise. The only recent concurrent work on bilevel optimization under heavy-tailed noise in the single-machine setting is [32]. However, it relies on a restrictive assumption that the lower-level problem is strongly convex. Moreover, it still relies on gradient clipping. In addition, the double-loop structure incurs substantial computational overhead due to the inner-loop iterations. Thus, it is important to fill this gap.

3 Problem Setup

3.1 Problem Definition

In this paper, we assume that there are K workers, indexed by $k \in \{1, 2, \cdots, K\}$, which form a communication graph and perform peer-to-peer communication within it. These workers collaboratively optimize a nonconvex decentralized stochastic bilevel optimization problem, defined as:

$$\min_{x\in\mathbb{R}^{d_1}, y\in y^*(x)} \frac{1}{K}\sum_{k=1}^{K} f^{(k)}(x,y) \quad s.t. \quad y^*(x) = \arg\min_{y\in\mathbb{R}^{d_2}} \frac{1}{K}\sum_{k=1}^{K} g^{(k)}(x,y)\,. \tag{1}$$

In Eq. (1), $f(x,y) = \frac{1}{K}\sum_{k=1}^{K} f^{(k)}(x,y)$ is the global upper-level loss function, where $f^{(k)}(x,y) = \mathbb{E}[f^{(k)}(x,y;\xi^{(k)})]$ is the local one on the k-th worker and $\xi^{(k)}$ denotes random samples on that worker. Additionally, $g(x,y) = \frac{1}{K}\sum_{k=1}^{K} g^{(k)}(x,y)$ is the global lower-level loss function, where $g^{(k)}(x,y) = \mathbb{E}[g^{(k)}(x,y;\zeta^{(k)})]$ is the lower-level one on the k-th worker and $\zeta^{(k)}$ represents the corresponding random samples. Unlike existing decentralized bilevel optimization methods [57, 13, 5, 6, 66, 22, 69, 37, 34, 35, 53], which assume that $g(x,y)$ is strongly convex with respect to y, we assume that $g(x,y)$ is a nonconvex loss function with respect to y, but satisfies the Polyak–Łojasiewicz (PL) condition with respect to y for any given x, which is strictly weaker than strong convexity and thus a more practical assumption.

3.2 Minimax Reformulation

Because $g(x,y)$ is nonconvex with respect to y, the second-order method, which relies on the Hessian inverse with respect to y of $g(x,y)$, is not applicable to Eq. (1). Hence, we employ the first-order method to solve it. Specifically, [23] shows that the lower-level subproblem in Eq. (1) can be converted into a constraint: $g(x,y) \le \min_{z\in\mathbb{R}^{d_2}} g(x,z)$, and then it can be converted into a minimax

optimization problem based on the penalty method, which is defined as follows:

$$\min_{x\in\mathbb{R}^{d_1},y\in\mathbb{R}^{d_2}} \max_{z\in\mathbb{R}^{d_2}} \frac{1}{K}\sum_{k=1}^{K} f^{(k)}(x,y) + \frac{1}{\delta}\left(\frac{1}{K}\sum_{k=1}^{K} g^{(k)}(x,y) - \frac{1}{K}\sum_{k=1}^{K} g^{(k)}(x,z)\right), \quad (2)$$

where $\delta > 0$ denotes the penalty parameter. With this reformulation, we only need to compute the first-order gradient with respect to x, y, and z to update them.

To solve Eq. (2) and measure its approximation for Eq. (1), we introduce the following functions:

$$\Phi(x) = \min_{y\in y^*(x)} \frac{1}{K}\sum_{k=1}^{K} f^{(k)}(x,y), \quad \Phi_\delta(x) = \min_{y\in\mathbb{R}^{d_2}} \max_{z\in\mathbb{R}^{d_2}} \frac{1}{\delta}\frac{1}{K}\sum_{k=1}^{K}(h_\delta^{(k)}(x,y) - g^{(k)}(x,z)), \quad (3)$$

where $h_\delta^{(k)}(x,y) = \delta f^{(k)}(x,y) + g^{(k)}(x,y)$ and $h_\delta(x,y) = \frac{1}{K}\sum_{k=1}^{K} h_\delta^{(k)}(x,y)$. [4] shows that $\Phi_\delta(x)$ can approximate $\Phi(x)$ well, including both their loss functions and gradients, by controlling the penalty parameter δ, which is shown in Appendix C.1. Importantly, $\min_{x\in\mathbb{R}^{d_1}} \Phi_\delta(x)$ is tractable compared to $\min_{x\in\mathbb{R}^{d_1}} \Phi(x)$. With the minimax reformulation, in the single-machine setting, [23] shows that the convergence rate when using first-order stochastic gradients is $O(1/T^{1/7})$ and can be improved to $O(1/T^{1/5})$ when using first-order stochastic variance-reduced gradients. Note that this reformulation for nonconvex bilevel optimization cannot achieve the $O(1/T^{1/3})$ convergence rate as the single-level method when using variance-reduced gradients. In fact, it is still an open problem to achieve that convergence rate. **The purpose of this paper is not to bridge this gap. Instead, our goal is to design a decentralized algorithm to solve Eq. (2) under heavy-tailed noise and then formally show how its solution solves Eq. (1).** Note that there are currently no decentralized minimax optimization methods capable of handling heavy-tailed noise without gradient clipping. Moreover, due to the penalty term, establishing the convergence rate is significantly more challenging than in existing single-level or standard minimax methods. Therefore, *solving Eq. (2) as a means to solve Eq. (1) under heavy-tailed noise requires new algorithm design and convergence analysis.*

3.3 Assumptions

To solve Eq. (1), we introduce some commonly used assumptions, which have been used in existing nonconvex bilevel optimization methods, such as [23, 4].

Assumption 3.1. Let $\theta = (x,y) \in \mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$, then the upper-level function $f^{(k)}(\theta)$ and lower-level function $g^{(k)}(\theta)$ on the k-th worker, and the penalty function $h_\delta(\theta)$ satisfy the following conditions:

1. For any θ_1 and θ_2, $\mathbb{E}[\|\nabla f^{(k)}(\theta_1;\xi) - \nabla f^{(k)}(\theta_2;\xi)\|^2] \le L_f^2\|\theta_1-\theta_2\|^2$ where the constant $L_f > 0$; $\|\nabla_2 f^{(k)}(\theta)\| \le C_f$ where the constant $C_f > 0$; $\mathbb{E}[\|\nabla^2 f^{(k)}(\theta_1;\xi) - \nabla^2 f^{(k)}(\theta_2;\xi)\|^2] \le \ell_f^2\|\theta_1-\theta_2\|^2$ where the constant $\ell_f > 0$.
2. For any θ_1 and θ_2, $\mathbb{E}[\|\nabla g^{(k)}(\theta_1;\zeta) - \nabla g^{(k)}(\theta_2;\zeta)\|^2] \le L_g^2\|\theta_1-\theta_2\|^2$ where the constant $L_g > 0$; $\mathbb{E}[\|\nabla^2 g^{(k)}(\theta_1;\zeta) - \nabla^2 g^{(k)}(\theta_2;\zeta)\|^2] \le \ell_g^2\|\theta_1-\theta_2\|^2$ where the constant $\ell_g > 0$.
3. $g(x,y)$ satisfies the μ-PL with respect to y where the constant $\mu > 0$; $h_\delta(x,y)$ satisfies the μ-PL with respect to y.

In particular, the third assumption, namely the PL condition on the lower-level function and the penalty function, is commonly adopted in [23, 4]. Moreover, recent studies have shown that over-parameterized neural networks can induce loss functions that satisfy the PL condition [30]. The Hessian-Lipschitz condition is used only in the analysis to ensure the smoothness of $\Phi(x)$ (Lemma C.1), while our algorithm is fully first-order and never computes Hessians.

Assumption 3.2. (**Heavy-tailed noise**) All stochastic first-order and second-order gradients are the unbiased estimators for the corresponding deterministic gradients. Moreover, there exist $s \in (1,2]$ and $\sigma > 0$ such that $\mathbb{E}[\|\nabla f^{(k)}(\theta;\xi) - \nabla f^{(k)}(\theta)\|^s] \le \sigma^s$ and $\mathbb{E}[\|\nabla g^{(k)}(\theta;\zeta) - \nabla g^{(k)}(\theta)\|^s] \le \sigma^s$ for any $\theta = (x,y) \in \mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$.

Assumption 3.3. For the adjacency matrix $E = [e_{ij}] \in \mathbb{R}_+^{K\times K}$ of the communication graph, $e_{ij} > 0$ indicates that the i-th worker and the j-th worker are connected. Otherwise, $e_{ij} = 0$. In addition, $\mathcal{N}_k = \{j | e_{kj} > 0\}$ denotes the neighboring workers of the k-th worker. Moreover, it satisfies the following conditions:

1. $E^T = E$, $E\mathbf{1} = \mathbf{1}$, $\mathbf{1}^T E = \mathbf{1}^T$, where $\mathbf{1} \in \mathbb{R}^K$ is the vector of all ones.
2. Its eigenvalues can be ordered by magnitude as: $|\lambda_K| \leq |\lambda_{K-1}| \leq \cdots \leq |\lambda_2| < |\lambda_1| = 1$.

By denoting $\lambda = |\lambda_2|$, the spectral gap is $1 - \lambda$.

Notations. In this paper, we define $\ell = \max\{L_f, L_g, \ell_f, \ell_g\}$, denote the condition number by $\kappa = \ell/\mu$, and represent the gradient with respect to the i-th variable with ∇_i.

Algorithm 1 D-NSVRGDA

Input: $\eta_x > 0, \eta_y > 0, \eta_z > 0, \gamma_x > 0, \gamma_y > 0, \gamma_z > 0$.
$\mathbb{I}_{t>0} = 1$ when $t > 0$. Otherwise, $\mathbb{I}_{t>0} = 0$. The batch size is B_0 when $t = 0$. Otherwise, it is $O(1)$.
Initialization on the k-th worker: $x_0^{(k)} = x_0$, $y_0^{(k)} = y_0$, $z_0^{(k)} = z_0$,

1: **for** $t = 0, \cdots, T-1$, the k-th worker **do**
2: Variance-reduced gradient estimators:
$$u_{1,t}^{(k)} = (1-\gamma_x)(u_{1,t-1}^{(k)} - \nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}; \xi_t^{(k)}))\mathbb{I}_{t>0} + \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_t^{(k)}),$$
$$u_{2,t}^{(k)} = (1-\gamma_x)(u_{2,t-1}^{(k)} - \nabla_1 g^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}; \zeta_t^{(k)}))\mathbb{I}_{t>0} + \nabla_1 g^{(k)}(x_t^{(k)}, y_t^{(k)}; \zeta_t^{(k)}),$$
$$u_{3,t}^{(k)} = (1-\gamma_x)(u_{3,t-1}^{(k)} - \nabla_1 g^{(k)}(x_{t-1}^{(k)}, z_{t-1}^{(k)}; \zeta_t^{(k)}))\mathbb{I}_{t>0} + \nabla_1 g^{(k)}(x_t^{(k)}, z_t^{(k)}; \zeta_t^{(k)}),$$
$$v_{1,t}^{(k)} = (1-\gamma_y)(v_{1,t-1}^{(k)} - \nabla_2 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}; \xi_t^{(k)}))\mathbb{I}_{t>0} + \nabla_2 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_t^{(k)}),$$
$$v_{2,t}^{(k)} = (1-\gamma_y)(v_{2,t-1}^{(k)} - \nabla_2 g^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}; \zeta_t^{(k)}))\mathbb{I}_{t>0} + \nabla_2 g^{(k)}(x_t^{(k)}, y_t^{(k)}; \zeta_t^{(k)}),$$
$$w_{1,t}^{(k)} = (1-\gamma_z)(w_{1,t-1}^{(k)} - \nabla_2 g^{(k)}(x_{t-1}^{(k)}, z_{t-1}^{(k)}; \zeta_t^{(k)}))\mathbb{I}_{t>0} + \nabla_2 g^{(k)}(x_t^{(k)}, z_t^{(k)}; \zeta_t^{(k)}),$$
3: Combine gradient estimators together for each variable:
$$u_t^{(k)} = u_{1,t}^{(k)} + \tfrac{1}{\delta}(u_{2,t}^{(k)} - u_{3,t}^{(k)}), \quad v_t^{(k)} = v_{1,t}^{(k)} + \tfrac{1}{\delta} v_{2,t}^{(k)}, \quad w_t^{(k)} = \tfrac{1}{\delta} w_{1,t}^{(k)},$$
4: Gradient tracking:
$$\tilde{p}_t^{(k)} = (p_{t-1}^{(k)} - u_{t-1}^{(k)})\mathbb{I}_{t>0} + u_t^{(k)}, \quad p_t^{(k)} = \textstyle\sum_{j\in\mathcal{N}_k} e_{kj}\tilde{p}_t^{(j)},$$
$$\tilde{q}_t^{(k)} = (q_{t-1}^{(k)} - v_{t-1}^{(k)})\mathbb{I}_{t>0} + v_t^{(k)}, \quad q_t^{(k)} = \textstyle\sum_{j\in\mathcal{N}_k} e_{kj}\tilde{q}_t^{(j)},$$
$$\tilde{r}_t^{(k)} = (r_{t-1}^{(k)} - w_{t-1}^{(k)})\mathbb{I}_{t>0} + w_t^{(k)}, \quad r_t^{(k)} = \textstyle\sum_{j\in\mathcal{N}_k} e_{kj}\tilde{r}_t^{(j)},$$
5: Updating:
$$\tilde{x}_{t+1}^{(k)} = x_t^{(k)} - \eta_x \tfrac{p_t^{(k)}}{\|p_t^{(k)}\|}, \quad x_{t+1}^{(k)} = \textstyle\sum_{j\in\mathcal{N}_k} e_{kj}\tilde{x}_{t+1}^{(j)} \ ,$$
$$\tilde{y}_{t+1}^{(k)} = y_t^{(k)} - \eta_y \tfrac{q_t^{(k)}}{\|q_t^{(k)}\|}, \quad y_{t+1}^{(k)} = \textstyle\sum_{j\in\mathcal{N}_k} e_{kj}\tilde{y}_{t+1}^{(j)} \ ,$$
$$\tilde{z}_{t+1}^{(k)} = z_t^{(k)} - \eta_z \tfrac{r_t^{(k)}}{\|r_t^{(k)}\|}, \quad z_{t+1}^{(k)} = \textstyle\sum_{j\in\mathcal{N}_k} e_{kj}\tilde{z}_{t+1}^{(j)} \ ,$$
6: **end for**

4 Decentralized Normalized Stochastic Gradient Descent Ascent with Variance Reduction Algorithm

4.1 Algorithm Design

To solve the reformulated Eq. (2), we develope a novel decentralized normalized stochastic gradient descent ascent with variance reduction (D-NSVRGDA) algorithm, which is presented in Algorithm 1. Specifically, we use the normalized variance-reduced gradient to update three variables: x, y, and z. More specifically, in Step 3 of Algorithm 1, we compute the variance-reduced gradient estimator for three variables as follows:

$$u_t^{(k)} = u_{1,t}^{(k)} + \frac{1}{\delta}(u_{2,t}^{(k)} - u_{3,t}^{(k)}) \ , \quad v_t^{(k)} = v_{1,t}^{(k)} + \frac{1}{\delta} v_{2,t}^{(k)} \ , \quad w_t^{(k)} = \frac{1}{\delta} w_{1,t}^{(k)} \ , \tag{4}$$

In Eq. (4), $u_{1,t}^{(k)}$, $u_{2,t}^{(k)}$, and $u_{3,t}^{(k)}$ are the variance-reduced gradient estimators for $\nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)})$, $\nabla_1 g^{(k)}(x_t^{(k)}, y_t^{(k)})$, and $\nabla_1 g^{(k)}(x_t^{(k)}, z_t^{(k)})$, respectively. Similarly, $v_{1,t}^{(k)}$ and $v_{2,t}^{(k)}$ estimate $\nabla_2 f^{(k)}(x_t^{(k)}, y_t^{(k)})$ and $\nabla_2 g^{(k)}(x_t^{(k)}, y_t^{(k)})$, respectively, while $w_{1,t}^{(k)}$ is used to estimate $\nabla_2 g^{(k)}(x_t^{(k)}, z_t^{(k)})$. All these gradient estimators are computed using the STORM method [8], as described in Step 2 of Algorithm 1, where $\gamma_x \in (0,1)$, $\gamma_y \in (0,1)$, and $\gamma_z \in (0,1)$ are three hyperparameters.

Then, our algorithm uses the gradient tracking approach to communicate these gradient estimators, which is shown in Step 4. For example, $p_t^{(k)} = \sum_{j\in\mathcal{N}_k} e_{kj}\tilde{p}_t^{(j)}$ represents the aggregation of gradient estimators $\tilde{p}_t^{(j)}$ from the neighboring workers $\mathcal{N}_k$ of the k-th worker. Finally, in Step 5, our algorithm uses the normalized gradient estimator to update the variables. For example, the k-th worker uses the normalized gradient estimator to update its local variable x and communicates the updated variable $\tilde{x}_{t+1}^{(k)}$ as follows:

$$\tilde{x}_{t+1}^{(k)} = x_t^{(k)} - \eta_x \frac{p_t^{(k)}}{\|p_t^{(k)}\|}\,, \quad x_{t+1}^{(k)} = \sum_{j\in\mathcal{N}_k} e_{kj}\tilde{x}_{t+1}^{(j)}\,, \tag{5}$$

where $\eta_x > 0$ denotes the learning rate for variable x, $\frac{p_t^{(k)}}{\|p_t^{(k)}\|}$ denotes the normalized gradient estimator, and the second equation represents the aggregation of updated variables $\tilde{x}_{t+1}^{(j)}$ from the neighboring workers $\mathcal{N}_k$ of the k-th worker. The other two variables are updated in the same way.

In Algorithm 1, we use only the normalized gradient estimator to update variables, without employing gradient clipping. To the best of our knowledge, **this is the first decentralized algorithm for bilevel optimization under heavy-tailed noise that does not rely on gradient clipping**. Furthermore, we believe that our algorithm can also be applied to standard minimax optimization under heavy-tailed noise, for which a decentralized algorithm without gradient clipping is also lacking.

4.2 Convergence Rate

Based on Assumptions 3.1-3.3, we establish the theoretical convergence rate of Algorithm 1 in the following theorem.

Theorem 4.1. *Given Assumptions 3.1-3.3 and initial points satisfying* $\mathbb{E}[h_\delta(x_0,y_0) - \min_y h_\delta(x_0,y)] \le \delta$ *and* $\mathbb{E}[g(x_0,z_0) - \min_z g(x_0,z)] \le \delta$*, by denoting* $a = \frac{s}{s-1} \ge 2$ *and setting*

$$\begin{aligned}
&\eta_x = O\left(\frac{(1-\lambda)^2 K\epsilon^{a+2}}{(\kappa^3\ell)^{a/2+2}\kappa^{a/2-1}\sigma^{a/2}}\right)\,, \quad \eta_y = \eta_x\frac{4(\delta L_f + L_g)}{\mu}\,, \quad \eta_z = \eta_x\frac{4L_g}{\mu}\,,\\
&\gamma_x = \gamma_y = \gamma_z = \gamma = O\left(\frac{K\epsilon^{2a}}{(\kappa^3\ell)^a\kappa^a\sigma^a}\right)\,, \quad \delta = O\left(\frac{\epsilon}{\kappa^3\ell}\right)\,,\\
&T = 4\max\left\{1, \frac{(\kappa^3\ell)^{a/2+2}\kappa^{a/2-1}\sigma^{a/2}}{(1-\lambda)^2K\epsilon^{a+3}}, M^{a/(a+1)}\right\}\,, \quad B_0 = \max\left\{1, \left(\frac{4M}{T}\right)^a\right\},
\end{aligned} \tag{6}$$

where $M = \frac{\kappa\sigma}{\delta\epsilon}\left(\frac{1}{\gamma} + \frac{\lambda\sqrt{K}}{1-\lambda}\right)$*,* B_0 *is the batch size at* $t=0$*, and* $T \ge 4M^{a/(a+1)}$ *ensures* $B_0 \le T$*, then our Algorithm 1 achieves the* ϵ*-accuracy solution:* $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|] \le O(\epsilon)$*, where* $\epsilon > 0$ *is a sufficiently small constant value.*

Remark 4.2. The convergence rate in Theorem 4.1 can recover the finite-variance setting. Specifically, when $s=2$, the number of iterations of our algorithm is $O\left(\frac{\kappa^9}{(1-\lambda)^2K\epsilon^5}\right)$ when ignoring other factors, which is the same as the convergence rate of [68] under the finite-variance assumption. In particular, the number of iterations is scaled by $1/K$, indicating that our algorithm can achieve linear speedup with respect to the number of workers K. The number of iterations is scaled by $1/(1-\lambda)^2$. This dependence on the spectral gap is consistent with that of classical decentralized minimax optimization [54, 63] and decentralized first-order bilevel optimization under the finite-variance assumption [68].

Remark 4.3. The initial points required in Theorem 4.1 are a warm-start condition similar to that in [53]. They can be obtained by gradient descent on the μ-PL problems $h_\delta(x_0,\cdot)$ and $g(x_0,\cdot)$, which converges at a linear rate. This cost is dominated by T and thus does not degenerate the overall complexity.

4.3 Proof Sketch

Establishing the convergence rate for Algorithm 1 is significantly more challenging than for existing methods [36, 18] that address single-level problems in a single-machine setting. The main difficulties

arise from: 1) **the interaction between gradients with respect to three variables due to the bilevel structure**, and 2) **the consensus error introduced by the decentralized setting**. On the other hand, both challenges are compounded by heavy-tailed noise, which makes the analysis more difficult than that in existing decentralized bilevel optimization methods that rely on the finite variance assumption. In Appendix B, we provide a proof sketch to demonstrate how these challenges are addressed. In Appendix C, we provide the detailed proof of Theorem 4.1.

5 Experiments

In our experiments, we evaluate our algorithm on two machine learning applications: hyperparameter optimization and model pruning. Due to space constraints, we present the results on two synthetic datasets and three real-world datasets related to hyperparameter optimization here. **Additional experimental results for both hyperparameter optimization and model pruning are provided in Appendix A**.

5.1 Nonconvex Hyperparameter Optimization

To validate the performance of D-NSVRGDA, we consider a nonconvex hyperparameter optimization problem, with the corresponding loss function defined in Eq. (7). Specifically, in the lower-level optimization subproblem, we optimize the weights of a two-layer fully connected neural network. Although this is a nonconvex optimization problem, existing work has shown that it can satisfy the Polyak-Łojasiewicz (PL) condition in the over-parameterized regime [30]. In the upper-level optimization subproblem, we optimize the hyperparameters that are used to regularize the neural network weights. Formally, it is defined as below:

$$\begin{aligned}&\min_{x=\{x_1,x_2\}} \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}^{(k)}(y^*(x);\mathcal{D}_{vl}^{(k)})\\ &\quad s.t.\ y^*(x)=\arg\min_{y=\{y_1,y_2\}} \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}^{(k)}(y;\mathcal{D}_{tr}^{(k)})+\mathcal{R}_1(x)+\mathcal{R}_2(x)\,,\end{aligned} \tag{7}$$

where $y_1 = [y_{1,pq}] \in \mathbb{R}^{m_1\times m_2}$ is the weight of the first layer, $y_2 = [y_{2,pq}] \in \mathbb{R}^{m_2\times m_3}$ is the weight of the second layer, $x_1 = [x_{1,q}] \in \mathbb{R}^{m_2}$ and $x_2 = [x_{2,q}] \in \mathbb{R}^{m_3}$ are hyperparameters for the regularization term: $\mathcal{R}_1(x) = \frac{1}{m_2}\sum_{q=1}^{m_2}\exp(x_{1,q})\frac{1}{m_1}\sum_{p=1}^{d_1} y_{1,pq}^2$ and $\mathcal{R}_2(x) = \frac{1}{m_3}\sum_{q=1}^{m_3}\exp(x_{2,q})\frac{1}{m_2}\sum_{p=1}^{m_2} y_{2,pq}^2$. In our experiments, m_1 is set to the number of input features, m_2 is set to 20, and m_3 is set to 1 for binary classification. We use ReLU as the activation function for the hidden layer and logistic loss for binary classification.

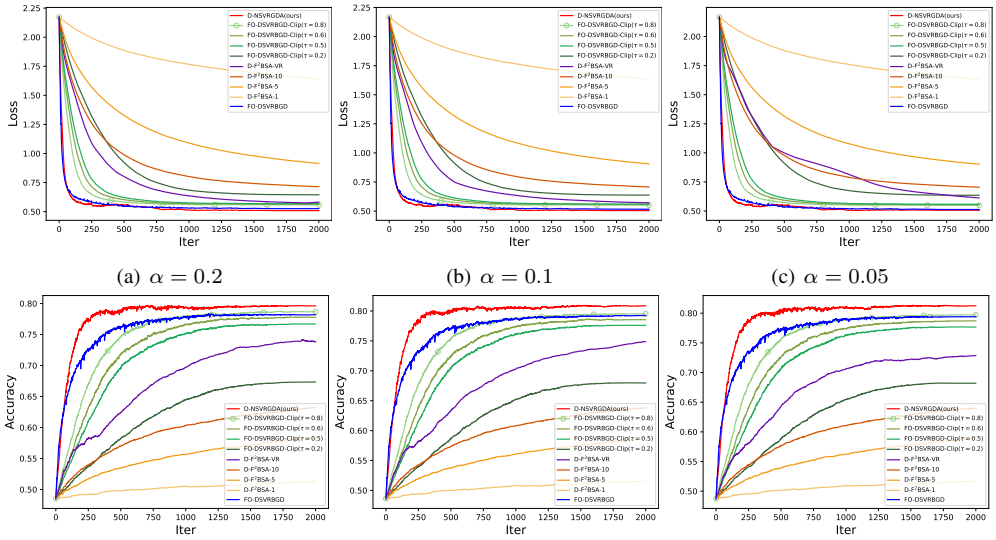
5.1.1 Synthetic Dataset I

We use a synthetic dataset to allow full control over the heavy-tailed feature distribution. For each sample i, the features are drawn as $x_{ij} \overset{\text{i.i.d.}}{\sim} t_\nu$, $j = 1,\dots,d$, where t_ν is the Student-t distribution with $\nu = 1.5$ degrees of freedom, and $d = 100$. The label is generated by

$$y_i^{(0)} = \operatorname{sgn}\left(x_i^\top w + \alpha\,\xi_i\right), \qquad y_i = b_i\, y_i^{(0)}\,,$$

where $w \in \mathbb{R}^{100}$ is drawn from a standard Gaussian distribution and normalized to unit norm, $\xi_i \sim \text{Cauchy}(0,1)$ perturbs the latent classification score, $\alpha \geq 0$ controls the magnitude of this perturbation, and $b_i \in \{-1,+1\}$ is an independent label flip with $\Pr(b_i = -1) = p_{\text{flip}} = 0.1$. Since $1 < \nu < 2$, the features have finite first moments but infinite second moments. The independent label flips prevent the logistic loss from suppressing large-feature observations, yielding stochastic gradients with finite means but infinite variances with respect to the first-layer weights. We generate $10{,}000$ training samples, and use the same approach to generate validation and testing sets with the same number of samples.

Most existing decentralized bilevel optimization algorithms require a strongly convex lower-level loss function. Although a recent work, FO-DSVRBGD [68], considers *nonconvex* lower-level problems, it does not address heavy-tailed noise or employ gradient normalization. Therefore, we use FO-DSVRBGD as a baseline to evaluate the effectiveness of gradient normalization in handling heavy-tailed noise. In addition, we replace gradient normalization with gradient clipping to obtain



(a) $\alpha = 0.2$ (b) $\alpha = 0.1$ (c) $\alpha = 0.05$

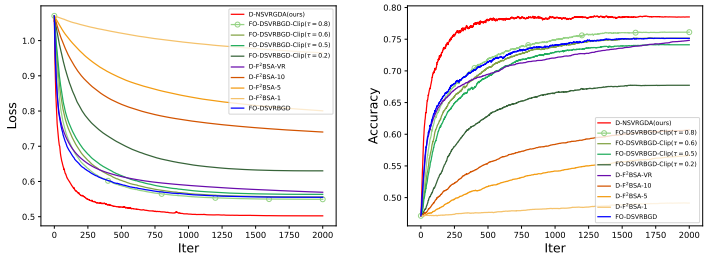
(d) $\alpha = 0.2$ (e) $\alpha = 0.1$ (f) $\alpha = 0.05$

Figure 1: The upper-level loss function value and test accuracy on Synthetic Dataset I with different magnitudes α.

another baseline, denoted as D-SVRGDA-Clip. We then compare the performance of D-NSVRGDA with that of FO-DSVRBGD and D-SVRGDA-Clip. For all three algorithms, we use the same hyperparameter settings, with the momentum coefficient set to 0.9 and the penalty parameter set to 0.3. The learning rate is tuned separately for each algorithm to achieve the best performance. As for D-SVRGDA-Clip, we evaluate different clipping threshold to fully demonstrate its performance. Additionally, there are eight workers, which are connected into a LINE graph. The batch size on each worker is set to 32.

Moreover, we compare our algorithm with methods developed for nonconvex bilevel problems in the single-machine setting under the bounded-variance assumption: F^{2}BSA [23]. We consider both single-loop and double-loop variants of F^{2}BSA under the decentralized setting. For the double-loop approach, the inner-loop iterations are set to one, five, and ten, which we denote as D-F^{2}BSA-1, D-F^{2}BSA-5, and D-F^{2}BSA-10, respectively. For the single-loop approach, which also uses the STORM variance reduction technique, we denote it by D-F^{2}BSA-VR. The learning rates of these baselines are tuned according to Corollaries 5.2 and 5.5 of [23].

Figure 1 shows the upper-level loss function value and the test accuracy of all methods on three datasets generate with $\alpha = \{0.2, 0.1, 0.05\}$. Both the loss function value and the test accuracy in Figure 1 confirm the effectiveness of our algorithm D-NSVRGDA in accommodating different levels of heavy-tailed noise compared to FO-DSVRBGD. In addition, we can find that D-SVRGDA-Clip is heavily affected by the clipping threshold τ. Therefore, D-SVRGDA-Clip is much more difficult to tune than our method. Furthermore, the double-loop approach (D-F^{2}BSA) depends heavily on the number of inner-loop iterations; increasing it may improve performance but incurs substantial computational overhead. Although the single-loop approach with variance reduction (D-F^{2}BSA-VR) performs better than the other baselines, it remains inferior to our method.



(a) Loss (b) Accuracy

Figure 2: The upper-level loss function value and test accuracy on Synthetic Dataset II.

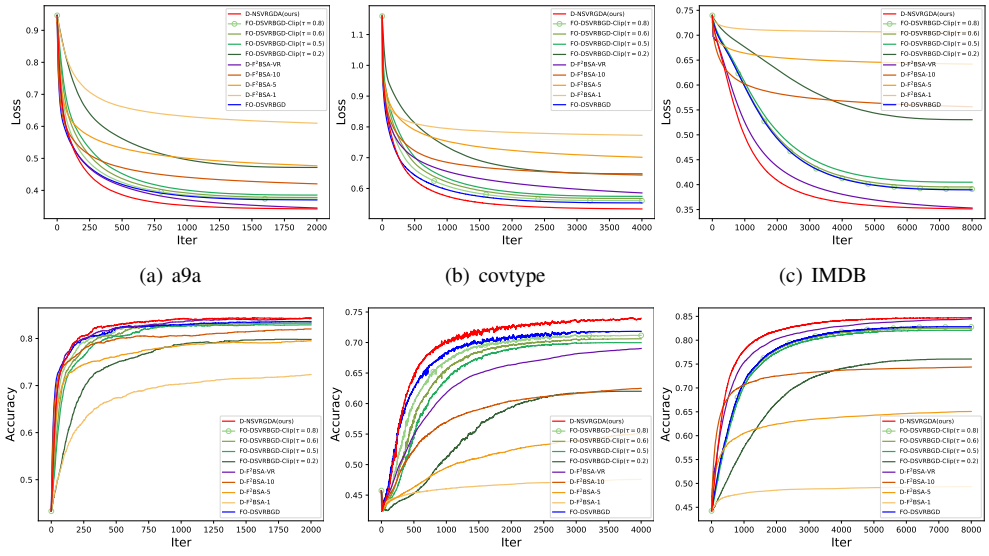
5.1.2 Synthetic Dataset II

In this experiment, we introduce a new synthetic dataset to simulate heavy-tailed noise in language data. Specifically, in natural language, some words appear much more frequently than others, and word frequencies follow a heavy-tailed distribution. To simulate this phenomenon, we split features into the common and rare features. Specifically, following [25], we assume 10% of features are the common ones, X_{common}, which are drawn from a Bernoulli distribution with the probability being 0.9, and 90% are the rare ones, X_{rare}, which are drawn from a Bernoulli distribution with probability 0.1. Then, the generated samples are represented by $X = [X_{\text{common}}, X_{\text{rare}}]$. The labels and all other settings follow Synthetic Dataset I, where α is 0.1. Figure 2 shows the upper-level loss function value and the test accuracy of all methods. Both the loss function value and the test accuracy in Figure 2 further confirm the effectiveness of our algorithm D-NSVRGDA in accommodating heavy-tailed noise compared to other baselines.

5.1.3 Real-world Datasets

In this experiment, we evaluate the performance of D-NSVRGDA on three real-world datasets: a9a, covtype, and IMDB, all of which are available from LIBSVM [3]. The experimental settings, including the communication graph, the batch size, the learning rate, and the penalty parameter, are the same as those in the first two experiments.

Figure 3 shows the upper-level loss function value and the test accuracy of D-NSVRGDA and other baselines on three real-world datasets. Similar to the first two experiments, both the loss function value and the test accuracy in Figure 3 further confirm the effectiveness of our algorithm D-NSVRGDA. In particular, IMDB is a text dataset whose features naturally follow a heavy-tailed distribution, and our algorithm demonstrates significant improvement over the baselines.



(a) a9a (b) covtype (c) IMDB

(d) a9a (e) covtype (f) IMDB

Figure 3: The upper-level loss function value and test accuracy on real-world datasets.

6 Conclusion

Heavy-tailed noise is common in practical machine learning models, yet it has not been studied in the context of decentralized bilevel optimization. To bridge this gap, our paper developed the first decentralized bilevel optimization algorithm to handle heavy-tailed noise using gradient normalization without gradient clipping. Moreover, our paper provided a theoretical convergence rate for our algorithm under heavy-tailed noise. To the best of our knowledge, this is the first theoretical result for nonconvex decentralized bilevel optimization under heavy-tailed noise. Finally, the extensive experiments validate the effectiveness of the proposed algorithm in handling heavy-tailed noise.

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

Acknowledgements

We thank the anonymous reviewers for their constructive comments. X. Zhang, Y. Zhang, H. Liang, and H. Gao were partially supported by U.S. NSF CAREER 2339545 and NSF IIS 2416607.

References

[1] K. Ahn, X. Cheng, M. Song, C. Yun, A. Jadbabaie, and S. Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *The Twelfth International Conference on Learning Representations*, 2024.

[2] B. Battash, L. Wolf, and O. Lindenbaum. Revisiting the noise model of stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 4780–4788. PMLR, 2024.

[3] H. Chen, J. Li, and A. M.-C. So. Set smoothness unlocks clarke hyper-stationarity in bilevel optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[4] L. Chen, J. Xu, and J. Zhang. On finding small hyper-gradients in bilevel optimization: Hardness results and improved analysis. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 947–980. PMLR, 2024.

[5] X. Chen, M. Huang, and S. Ma. Decentralized bilevel optimization. *Optimization Letters*, 19(7):1249–1313, 2025.

[6] X. Chen, M. Huang, S. Ma, and K. Balasubramanian. Decentralized stochastic bilevel optimization with improved per-iteration complexity. In *International conference on machine learning*, pages 4641–4671. PMLR, 2023.

[7] A. Cutkosky and H. Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. *Advances in Neural Information Processing Systems*, 34:4883–4895, 2021.

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

[9] Y. Dong, S. Ma, J. Yang, and C. Yin. A single-loop algorithm for decentralized bilevel optimization. *Mathematics of Operations Research*, 2025.

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

[11] 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.

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

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

[14] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

[15] E. Gorbunov, A. Sadiev, M. Danilova, S. Horváth, G. Gidel, P. Dvurechensky, A. Gasnikov, and P. Richtárik. High-probability convergence for composite and distributed stochastic minimization and variational inequalities with heavy-tailed noise. In *International Conference on Machine Learning*, pages 15951–16070. PMLR, 2024.

[16] M. Gürbüzbalaban, Y. Hu, U. Şimşekli, K. Yuan, and L. Zhu. Heavy-tail phenomenon in decentralized sgd. *IISE Transactions*, pages 1–15, 2024.

[17] F. Huang and S. Chen. Near-optimal decentralized momentum method for nonconvex-pl minimax problems. *arXiv preprint arXiv:2304.10902*, 2023.

[18] F. Hübler, I. Fatkhullin, and N. He. From gradient clipping to normalization for heavy tailed sgd. In *International Conference on Artificial Intelligence and Statistics*, pages 2413–2421. PMLR, 2025.

[19] W. Jiang, H. Bai, and Z. Li. Last-iterate convergence of clipped sgd with linear decay. *arXiv preprint arXiv:2610.18555*, 2026.

[20] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.

[21] J. Kim, E. Nichani, D. Wu, A. Bietti, and J. D. Lee. Sharp capacity scaling of spectral optimizers in learning associative memory. *arXiv preprint arXiv:2603.26554*, 2026.

[22] B. Kong, S. Zhu, S. Lu, X. Huang, and K. Yuan. Decentralized bilevel optimization: A perspective from transient iteration complexity. *Journal of Machine Learning Research*, 26(240):1–64, 2025.

[23] J. Kwon, D. Kwon, S. Wright, and R. Nowak. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. In *The Twelfth International Conference on Learning Representations*, 2024.

[24] J. Kwon, D. Kwon, S. Wright, and R. D. Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pages 18083–18113. PMLR, 2023.

[25] S. H. Lee, M. Zaheer, and T. Li. Efficient distributed optimization under heavy-tailed noise. In *International Conference on Machine Learning*, pages 33833–33882. PMLR, 2025.

[26] B. Li, K. Wang, H. Zhong, P. Lu, and L. Wang. Muon in associative memory learning: Training dynamics and scaling laws. *arXiv preprint arXiv:2602.05725*, 2026.

[27] C. Li, S. Zeng, Z. Liao, J. Li, D. Kang, A. Garcia, and M. Hong. Learning reward and policy jointly from demonstration and preference improves alignment. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*, 2024.

[28] J. Li, S. Zeng, H. T. Wai, C. Li, A. Garcia, and M. Hong. Getting more juice out of the SFT data: Reward learning from human demonstration improves SFT for LLM alignment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[29] B. Liu, M. Ye, S. Wright, P. Stone, and Q. Liu. Bome! bilevel optimization made easy: A simple first-order approach. *Advances in neural information processing systems*, 35:17248–17262, 2022.

[30] C. Liu, L. Zhu, and M. Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

[31] L. Liu, Y. Wang, and L. Zhang. High-probability bound for non-smooth non-convex stochastic optimization with heavy tails. In *Forty-first International Conference on Machine Learning*, 2024.

[32] Z. Liu and L. Luo. Stochastic bilevel optimization with heavy-tailed noise. *arXiv preprint arXiv:2509.14952*, 2025.

[33] Z. Liu, J. Zhang, and Z. Zhou. Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2266–2290. PMLR, 2023.

[34] Z. Liu, X. Zhang, P. Khanduri, S. Lu, and J. Liu. Interact: achieving low sample and communication complexities in decentralized bilevel learning over networks. In *Proceedings of the Twenty-Third International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 61–70, 2022.

[35] Z. Liu, X. Zhang, P. Khanduri, S. Lu, and J. Liu. Prometheus: taming sample and communication complexities in constrained decentralized stochastic bilevel learning. In *International Conference on Machine Learning*, pages 22420–22453. PMLR, 2023.

[36] Z. Liu and Z. Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *The Thirteenth International Conference on Learning Representations*, 2025.

[37] S. Lu, X. Cui, M. S. Squillante, B. Kingsbury, and L. Horesh. Decentralized bilevel optimization for personalized client learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5543–5547. IEEE, 2022.

[38] Z. Lu and S. Mei. First-order penalty methods for bilevel optimization. *SIAM Journal on Optimization*, 34(2):1937–1969, 2024.

[39] Z. Lu and S. Mei. Solving bilevel optimization via sequential minimax optimization. *Mathematics of Operations Research*, 2026.

[40] T. D. Nguyen, T. H. Nguyen, A. Ene, and H. Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. *Advances in Neural Information Processing Systems*, 36:24191–24222, 2023.

[41] M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[42] F. Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pages 737–746. PMLR, 2016.

[43] Z. Qin, Z. Liu, S. Lu, Y. Liang, and J. Liu. DUET: Decentralized bilevel optimization without lower-level strong convexity. In *The Thirteenth International Conference on Learning Representations*, 2025.

[44] A. Sadiev, M. Danilova, E. Gorbunov, S. Horváth, G. Gidel, P. Dvurechensky, A. Gasnikov, and P. Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *International Conference on Machine Learning*, pages 29563–29648. PMLR, 2023.

[45] H. Shen, P.-Y. Chen, P. Das, and T. Chen. SEAL: Safety-enhanced aligned LLM fine-tuning via bilevel data selection. In *The Thirteenth International Conference on Learning Representations*, 2025.

[46] H. Shen and T. Chen. On penalty-based bilevel gradient descent method. In *International Conference on Machine Learning*, pages 30992–31015. PMLR, 2023.

[47] H. Shen, Z. Yang, and T. Chen. Principled penalty-based methods for bilevel reinforcement learning and rlhf. *Journal of Machine Learning Research*, 26(114):1–49, 2025.

[48] U. Şimşekli, M. Gürbüzbalaban, T. H. Nguyen, G. Richard, and L. Sagun. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*, 2019.

[49] U. Simsekli, L. Sagun, and M. Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

[50] T. Sun, X. Liu, and K. Yuan. Gradient normalization provably benefits nonconvex sgd under heavy-tailed noise. *arXiv preprint arXiv:2410.16561*, 2024.

[51] M. Wang, Z. Liu, and L. Luo. Near-optimal decentralized stochastic nonconvex optimization with heavy-tailed noise. *arXiv preprint arXiv:2601.11435*, 2026.

[52] S. Wang, F. Zhang, J. Li, C. Du, C. Du, T. Pang, Z. Yang, M. Hong, and V. Tan. Muon outperforms adam in tail-end associative memory learning. In *International Conference on Learning Representations*, volume 2026, pages 37382–37419, 2026.

[53] X. Wang, X. Chen, S. Ma, and T. Zhang. Fully first-order methods for decentralizedbilevel optimization. *IEEE Transactions on Signal Processing*, 73:4734–4747, 2025.

[54] 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.

[55] Q. Xiao, S. Lu, and T. Chen. A generalized alternating method for bilevel optimization under the polyak-łojasiewicz condition. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 63847–63873, 2023.

[56] H. Yang, P. Qiu, and J. Liu. Taming fat-tailed ("heavier-tailed" with potentially infinite variance) noise in federated learning. *Advances in Neural Information Processing Systems*, 35:17017–17029, 2022.

[57] S. Yang, X. Zhang, and M. Wang. Decentralized gossip-based stochastic bilevel optimization over communication networks. *Advances in neural information processing systems*, 35:238–252, 2022.

[58] T. Yang. Algorithmic foundations of empirical x-risk minimization. *arXiv preprint arXiv:2206.00439*, 2022.

[59] S. Yu, D. Jakovetic, and S. Kar. Decentralized nonconvex optimization under heavy-tailed noise: Normalization and optimal convergence. In *International Conference on Learning Representations*, volume 2026, pages 8319–8332, 2026.

[60] J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. Reddi, S. Kumar, and S. Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

[61] X. Zhang and H. Gao. Federated stochastic minimax optimization under heavy-tailed noises. *arXiv preprint arXiv:2511.04456*, 2025.

[62] X. Zhang and H. Gao. On provable benefits of muon in federated learning. *arXiv preprint arXiv:2510.03866*, 2025.

[63] X. Zhang, G. Mancino-Ball, N. S. Aybat, and Y. Xu. Jointly improving the sample and communication complexities in decentralized stochastic minimax optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20865–20873, 2024.

[64] X. Zhang, Y. Zhang, H. Gao, and H. Huang. Distributed stochastic k-level optimization over networks. In *Forty-third International Conference on Machine Learning*, 2026.

[65] Y. Zhang, X. Shi, M. Qiu, Y. Wang, and H. Gao. On the convergence of decentralized stochastic minimax optimization algorithm with compressed communication. In *Forty-third International Conference on Machine Learning*, 2026.

[66] Y. Zhang, M. T. Thai, J. Wu, and H. Gao. On the communication complexity of decentralized stochastic bilevel optimization. *Machine Learning*, 115(5):101, 2026.

[67] Y. Zhang, Y. Yao, P. Ram, P. Zhao, T. Chen, M. Hong, Y. Wang, and S. Liu. Advancing model pruning via bi-level optimization. *Advances in Neural Information Processing Systems*, 35:18309–18326, 2022.

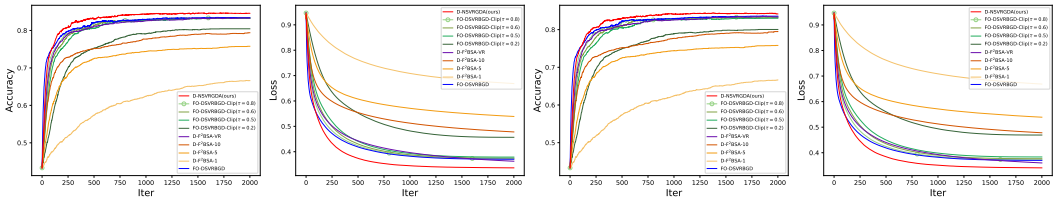
[68] Y. Zhang, X. Zhang, M. T. Thai, J. Wu, and H. Gao. Convergence analysis of decentralized hessian-/jacobian-free algorithm for nonconvex stochastic bilevel optimization. In *Forty-third International Conference on Machine Learning*, 2026.

[69] S. Zhu, B. Kong, S. Lu, X. Huang, and K. Yuan. Sparkle: a unified single-loop primal-dual framework for decentralized bilevel optimization. *Advances in Neural Information Processing Systems*, 37:62912–62987, 2024.

A More Experiments

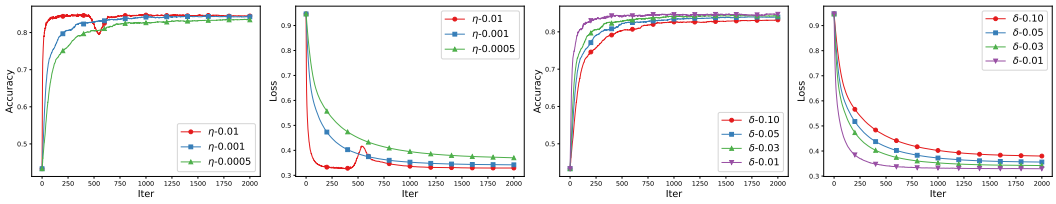
A.1 Additional Results on Nonconvex Hyperparameter Optimization

We provide further experiments on the a9a dataset under different communication graphs (Figure 4) and different hyperparameter settings (Figure 5). Figure 4 shows that our algorithm consistently outperforms the baselines across different graph topologies. Figure 5 demonstrates that the convergence rate improves with a larger learning rate and a smaller penalty parameter δ.



(a) Random Graph (b) Random Graph (c) Torus Graph (d) Torus Graph

Figure 4: The upper-level loss function value and test accuracy under different graphs.



(a) Different η (b) Different η (c) Different δ (d) Different δ

Figure 5: The upper-level loss function value and test accuracy under different hyperparameters.

To evaluate the robustness, we conduct three independent runs on the a9a dataset and report the mean and standard deviation of the test accuracy and test loss in Table 1. Our method consistently achieves the highest test accuracy and the lowest test loss.

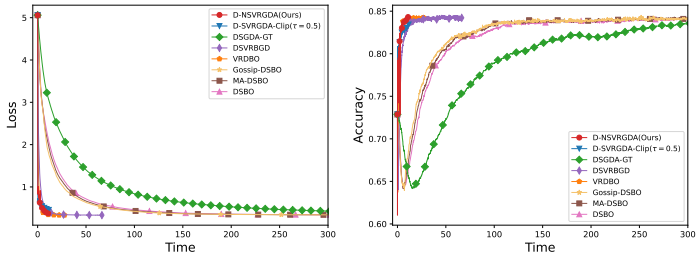
Table 1: Test accuracy and test loss (mean $\pm$ std over three independent runs).

Method	Test accuracy	Test loss
D-NSVRGDA (Ours)	$\mathbf{0.8397 \pm 0.0043}$	$\mathbf{0.3485 \pm 0.0064}$
D-SVRGDA-Clip ($\tau = 0.8$)	0.8320 ± 0.0024	0.3753 ± 0.0036
D-SVRGDA-Clip ($\tau = 0.6$)	0.8309 ± 0.0028	0.3808 ± 0.0064
D-SVRGDA-Clip ($\tau = 0.5$)	0.8281 ± 0.0020	0.3888 ± 0.0096
D-SVRGDA-Clip ($\tau = 0.2$)	0.7838 ± 0.0211	0.4849 ± 0.0453
D-F^{2}BSA-VR	0.8376 ± 0.0056	0.3623 ± 0.0154
D-F^{2}BSA-10	0.7937 ± 0.0230	0.4601 ± 0.0356
D-F^{2}BSA-5	0.7662 ± 0.0253	0.5213 ± 0.0449
D-F^{2}BSA-1	0.6802 ± 0.0811	0.6612 ± 0.1068
FO-DSVRBGD	0.8337 ± 0.0029	0.3721 ± 0.0017

A.2 Hyperparameter Optimization on Nonconvex-Strongly-Convex Bilevel Optimization Problem

To provide a more comprehensive comparison with existing decentralized bilevel baselines, we further conduct experiments on a hyperparameter optimization task under a nonconvex–strongly-convex bilevel formulation. Specifically, we focus on the hyperparameter optimization task, where the classifier is a logistic regression model. Then, the lower-level optimization problem is to learn the weight of the logistic regression model, and the upper-level optimization problem is to learn the coefficient of the regularization term like Eq. (7). Since strong convexity implies the PL condition, our method can be directly applied in this setting. In addition to our variant with gradient clipping, we compare against the following representative baselines: DSBO [5], MA-DSBO [6], Gossip-DSBO [57], VRDBO [13], DSVRBGD [66], and DSGDA-GT [53]. Note that all of these methods

rely on second-order information for updates, except DSGDA-GT, which is fully first-order. In our experiments, we set the learning rate of these baseline methods according to their theoretical results in the original paper.

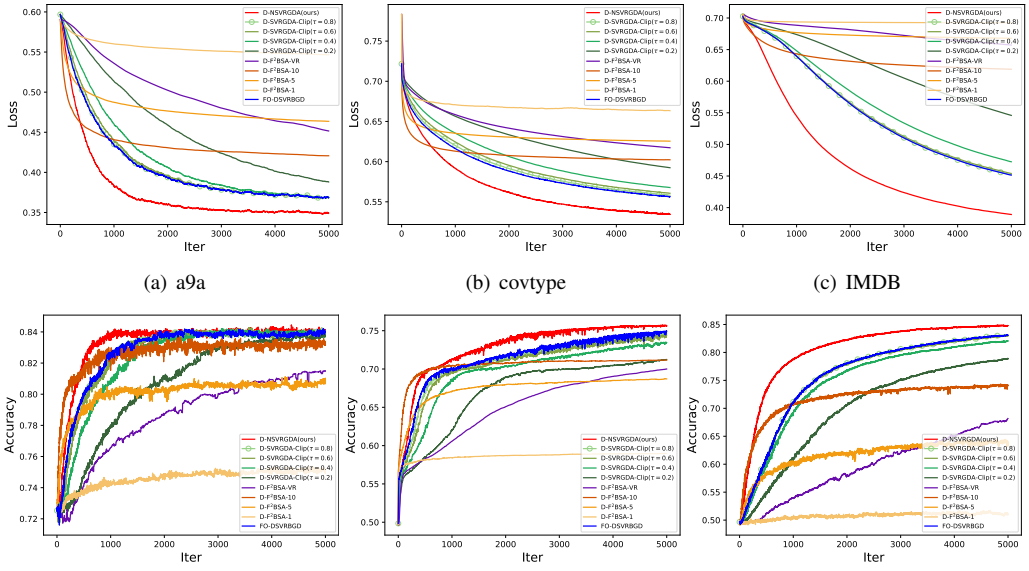


(a) Loss (b) Test Accuracy

Figure 6: The upper-level loss function value and test accuracy for the nonconvex-strongly-convex bilevel problem with respect to the time consumed (seconds) on a9a dataset.

From Figure 6, we can clearly observe that our algorithm, which relies only on first-order variance-reduced gradient updates, requires substantially less time to converge compared with methods that depend on second-order Jacobians or Hessians. Although DSGDA-GT also uses fully first-order information, its high complexity of $O(\epsilon^{-7})$ leads to significantly slower convergence, offering limited practical advantage.

A.3 Model Pruning for MLP



(a) a9a (b) covtype (c) IMDB

(d) a9a (e) covtype (f) IMDB

Figure 7: The upper-level loss function value and test accuracy on real-world datasets for MLP pruning task.

In this experiment, we verify the performance of our algorithm on the model pruning task. Following [67], model pruning can be formulated as a bilevel optimization problem. Formally, in the decentralized setting, its loss function is defined as follows:

$$\min_x \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}^{(k)}(x \odot y^*(x)) \qquad s.t. \quad y^*(x) = \arg\min_y \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}^{(k)}(x \odot y)\,, \tag{8}$$

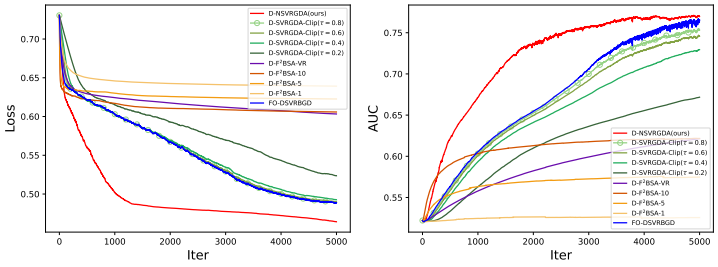
where $y \in \mathbb{R}^d$ denotes the parameter of a deep neural network, and $x \in \{0,1\}^d$ is a binary mask, where 0 indicates pruning the corresponding neuron. Since [30] shows that optimizing an overparameterized deep neural network satisfies the PL condition, the model pruning problem satisfies

the nonconvex-PL assumption when pruning a deep neural network. In this experiment, we use the same neural network architecture as in the first three experiments and keep the other experimental settings unchanged. For the pruning rate, we prune 80% of the neurons.

Figure 7 shows the upper-level loss value and test accuracy of D-NSVRGDA and other baselines on the model pruning task defined in Eq. (8). We also evaluate D-SVRGDA-Clip under different clipping threshold values. From the figure, we observe that our algorithm, D-NSVRGDA, consistently outperforms the baseline methods in terms of both the loss value and test accuracy. This further confirms the effectiveness of our algorithm in handling heavy-tailed noise in new applications.

A.4 Model Pruning for RNN

In this section, we add an additional experiment to further verify the performance of our algorithm on more complicated applications. Specifically, we consider the model pruning task in Eq. (8) for the text classification application using a recurrent neural network, as the language data typically incurs the heavy-tailed noise. In detail, we use Sentiment140 dataset [14] and use a two-layer recurrent neural network as the classifier where the embedding size is 300 and the number of hidden neurons is 128. Then, in the lower-level optimization problem, we learn the weight of the recurrent neural network in the lower-level optimization problem, while learning the pruning mask in the upper-level optimization problem. In this experiment, we use the same experimental settings for all methods as the last experiment regarding MLP pruning.



(a) Loss (b) Test AUC

Figure 8: The upper-level loss function value and test AUC score in the RNN pruning task.

Figure 8 shows the upper-level loss function value and test AUC (area-under-the-curve) score of D-NSVRGDA and all baselines on the RNN pruning task. Similar to the MLP pruning task, our algorithm, D-NSVRGDA, consistently outperforms all baseline methods in terms of both the loss value and test accuracy for the RNN pruning task. This further confirms the effectiveness of our algorithm in handling heavy-tailed noise in large-scale real-world applications.

A.5 Model Pruning for Transformer

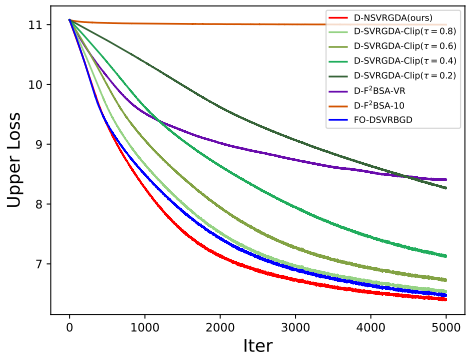


Figure 9: The upper-level loss function value for the Transformer pruning task on the EnViMT dataset.

To further evaluate the effectiveness of our algorithm on larger models, we consider the model pruning task in Eq. (8) for machine translation using a Transformer model on the EnViMT dataset. Both

the encoder and decoder contain two blocks, with 128 hidden features and 4 attention heads in each block. We adopt the same decentralized optimization protocol as in the previous pruning experiments, using 8 nodes with a ring topology and a batch size of 16. Differently, we apply unstructured weight pruning to the feed-forward layers of the Transformer, retaining 80% of the weights.

Figure 9 shows the upper-level loss function values of D-NSVRGDA and the baseline methods. From the figure, we observe that D-NSVRGDA consistently achieves lower upper-level loss values than all baseline methods throughout the training process. These results further demonstrate the effectiveness of our algorithm on more complex neural network architectures.

B Proof Sketch

Establishing the convergence rate for Algorithm 1 is significantly more challenging than for existing methods [36, 18] that address single-level problems in a single-machine setting. The main difficulties arise from: 1) **the interaction between gradients with respect to three variables due to the bilevel structure**, and 2) **the consensus error introduced by the decentralized setting**. On the other hand, both challenges are compounded by heavy-tailed noise, which makes the analysis more difficult than that in existing decentralized bilevel optimization methods that rely on the finite variance assumption.

B.1 Novelty of Our Convergence Analysis

Here, we highlight the novelty of our convergence analysis in handling the unique challenges caused by the heavy-tailed noise for the decentralized bilevel optimization. Specifically, bounding $\mathbb{E}[||\nabla\Phi(\bar{x}_t)||]$ is quite challenging in the presence of heavy-tailed noise for nonconvex bilevel optimization. **The reason is that its upper bound relies on the optimization errors regarding y and z, and the update of y and z relies on normalized gradient estimators to handle heavy-tailed noise.**

B.1.1 Novelty over methods with bounded variance

When there does not exist heavy-tailed noise, the commonly used approach [23] for handling the optimization errors regarding y and z is to bound $\|\bar{y}_t - y^*_\delta(\bar{x}_t)\|^2$ and $\|\bar{z}_t - y^*(\bar{x}_t)\|^2$. However, **this approach does NOT work for the normalized gradient estimator**. The reason is that **it requires the standard stochastic gradient estimator without normalization and requires strong convexity**. Specifically, the second to last step in Lemma C.1 of [23] holds only for the original gradient and strong convexity. As a result, Lemma C.1 of [23] cannot handle the normalized gradient estimator in our algorithm. For example, if using Lemma C.1 of [23] to bound $\|\bar{z}_{t+1} - y^*(\bar{x}_{t+1})\|^2$, it is incapable of handling $\|\bar{z}_t - \eta_z \frac{1}{K}\sum_{k=1}^K \frac{r_t^{(k)}}{\|r_t^{(k)}\|} - y^*(\bar{x}_{t+1})\|^2$ **in the presence of the normalized gradient and the absence of strong convexity**. Specifically, if we bound it via

$$\begin{aligned}
&\left\|\bar{z}_t - \eta_z \frac{1}{K}\sum_{k=1}^K \frac{r_t^{(k)}}{\|r_t^{(k)}\|} - y^*(\bar{x}_{t+1})\right\|^2 \\
&\leq 2\left\|\bar{z}_t - \eta_z \frac{1}{K}\sum_{k=1}^K r_t^{(k)} - y^*(\bar{x}_{t+1})\right\|^2 + 2\eta_z^2 \frac{1}{K}\sum_{k=1}^K \left\| r_t^{(k)} - \frac{r_t^{(k)}}{\|r_t^{(k)}\|}\right\|^2 . \qquad (9)
\end{aligned}$$

It is unclear how to bound the last term, as **it could introduce a constant term and thus only guarantee convergence to a neighborhood of the stationary point**. Therefore, directly bounding $\|\bar{z}_{t+1} - y^*(\bar{x}_{t+1})\|^2$ is not feasible. To the best of our knowledge, **we are not aware of any existing works that can bound it in the presence of a normalized gradient**.

In our proof, we proposed a novel approach to handle the **normalized gradient estimator** when bounding the optimization errors regarding y and z. Generally speaking, **we bound optimization errors regarding y and z from the perspective of function values, instead of variables**. Specifically, as shown in Lemma B.1, we proposed bounding $\mathbb{E}[||\nabla_2 h_\delta(\bar{x}_t, \bar{y}_t)||]$ and $\mathbb{E}[||\nabla_2 g(\bar{x}_t, \bar{z}_t)||]$, instead of $||\bar{y}_t - y^*_\delta(\bar{x}_t)||$ and $||\bar{z}_t - y^*(\bar{x}_t)||$, to **effectively handle the normalized gradient estimator**. For example, to bound $\mathbb{E}[||\nabla_2 h_\delta(\bar{x}_t, \bar{y}_t)||]$, we study the evolvement of the function values: $h_\delta(\bar{x}_{t+1}, \bar{y}_{t+1})$ and $h^*_\delta(\bar{x}_{t+1})$. In particular, by upper bounding $h_\delta(\bar{x}_{t+1}, \bar{y}_{t+1}) - h_\delta(\bar{x}_t, \bar{y}_t)$ and

$h_\delta^*(\bar{x}_t) - h_\delta^*(\bar{x}_{t+1})$, where the normalized gradient estimator is much easier to handle and the strong convexity is NOT required, we can obtain the upper bound of $\mathbb{E}[||\nabla_2 h_\delta(\bar{x}_t, \bar{y}_t)||]$ in Lemma C.8. Similarly, we bound $\mathbb{E}[||\nabla_2 g(\bar{x}_t, \bar{z}_t)||]$ in Lemma C.7. Note that, with this new strategy, we obtain the inner product between the original gradient and normalized gradient, e.g., $\mathbb{E}\left[\left\langle \bar{r}_t, \frac{1}{K}\sum_{k=1}^K \frac{r_t^{(k)}}{\|r_t^{(k)}\|}\right\rangle\right]$ in Lemma C.7 instead of the difference like $\frac{1}{K}\sum_{k=1}^K \left\| r_t^{(k)} - \frac{r_t^{(k)}}{\|r_t^{(k)}\|}\right\|^2$. This kind of inner product is much easier to bound because $\left\|\frac{r_t}{\|r_t\|}\right\| = 1$.

As such, we can successfully address the challenge about the optimization error with respect to y and z **in the presence of the normalized gradient and the absence of strong convexity**. Note that **bounding $\mathbb{E}[||\nabla_2 h_\delta(\bar{x}_t, \bar{y}_t)||]$ and $\mathbb{E}[||\nabla_2 g(\bar{x}_t, \bar{z}_t)||]$ is significantly different from the setting where the gradient noise has finite variance**. In particular, when handling T_1 and T_2 in Lemma C.8 and Lemma C.7, we should handle the normalized gradient estimator rather than the standard gradient estimator. Therefore, it is totally different from the setting where the gradient noise has finite variance.

B.1.2 Novelty over methods for single-level optimization

The single-level optimization method for heavy-tailed noise, such as [36], **cannot handle the interaction among three variables in our bilevel optimization problems**. For example, when bounding $\mathbb{E}[||\nabla_2 h_\delta(\bar{x}_t, \bar{y}_t)||]$ in Lemma C.8, we need to handle the interaction between x and y. The existing single-level approaches [36] are NOT capable of handling this interaction.

In our proof, we develop a novel approach to handle the interaction between two variables when bounding $\mathbb{E}[||\nabla_2 h_\delta(\bar{x}_t, \bar{y}_t)||]$ and $\mathbb{E}[||\nabla_2 g(\bar{x}_t, \bar{z}_t)||]$. For example, we use three steps to address this interaction when bounding $\mathbb{E}[||\nabla_2 h_\delta(\bar{x}_t, \bar{y}_t)||]$ in Lemma C.8, which is shown below:

- First, we figure out how the update of y affects the evolvement of the function value $h_\delta(\bar{x}_{t+1}, \bar{y}_{t+1})$, i.e., studying $h_\delta(\bar{x}_{t+1}, \bar{y}_{t+1}) - h_\delta(\bar{x}_{t+1}, \bar{y}_t)$.
- Second, we study how the update of x affects the evolvement of the function value $h_\delta(\bar{x}_{t+1}, \bar{y}_t)$, i.e., bounding $h_\delta(\bar{x}_{t+1}, \bar{y}_t) - h_\delta(\bar{x}_t, \bar{y}_t)$.
- Third, we investigate how the update of x affects the evolvement of $h_\delta^*(\bar{x}_{t+1})$, i.e., bounding $h_\delta^*(\bar{x}_t) - h_\delta^*(\bar{x}_{t+1})$.

Finally, by combining these three upper bounds to obtain the upper bound of $h_\delta(\bar{x}_{t+1}, \bar{y}_{t+1}) - h_\delta^*(\bar{x}_{t+1})$, this can provide the upper bound of the optimization error regarding y, i.e., bounding $\mathbb{E}[||\nabla_2 h_\delta(\bar{x}_t, \bar{y}_t)||]$. With such a novel approach, we can successfully address the challenge caused by the interaction between three variables.

In summary, our proof is novel and has addressed unique challenges caused by the heavy-tailed noise for nonconvex decentralized bilevel optimization. To the best of our knowledge, this is the first paper proposing this technique to handle the heavy-tailed noise for nonconvex bilevel optimization.

B.2 Solution for the First Challenge

First Step: Given that the gradients with respect to three variables interact with each other, we first disclose how they interact with each other in Lemma B.1.

Lemma B.1. *Given Assumption 3.1, we can obtain*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|] \le \frac{\mathbb{E}[\Phi(\bar{x}_0)] - \mathbb{E}[\Phi(\bar{x}_T)]}{\eta_x T} + 2\frac{1}{T}\sum_{t=0}^{T-1}\underbrace{\mathbb{E}[\|\nabla\Phi(\bar{x}_t) - \nabla\Phi_\delta(\bar{x}_t)\|]}_{\textit{Approximation Error caused by the minimax reformulation}}$$
$$+ \frac{2(\delta L_f + L_g)}{\mu}\frac{1}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\underbrace{\mathbb{E}[||\nabla_2 h_\delta(\bar{x}_t, \bar{y}_t)||]}_{\textit{Gradient regarding } y} + \frac{2L_g}{\mu}\frac{1}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\underbrace{\mathbb{E}[||\nabla_2 g(\bar{x}_t, \bar{z}_t)||]}_{\textit{Gradient regarding } z} + \frac{\eta_x L_\Phi}{2}$$
$$+ \textbf{\textit{Gradient Errors}} + \textbf{\textit{Consensus Errors}}\ . \tag{10}$$

Here, **Gradient Errors** include $2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_1 f^{(k)}(x_t^{(k)},y_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}u_{1,t}^{(k)}\|]$, $2\frac{1}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_1 g^{(k)}(x_t^{(k)},y_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}u_{2,t}^{(k)}\|]$, and $2\frac{1}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_1 g^{(k)}(x_t^{(k)},z_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}u_{3,t}^{(k)}\|]$. **Consensus Errors** include: $2(L_f+\frac{2L_g}{\delta})\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_t^{(k)}-\bar{x}_t\|]$, $2(L_f+\frac{L_g}{\delta})\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_t^{(k)}-\bar{y}_t\|]$, $2\frac{L_g}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|z_t^{(k)}-\bar{z}_t\|]$, and $\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t-p_t^{(k)}\|]$, where the first three terms are the consensus error with respect to variables, while the last is about the gradient.

Lemma B.1 discloses that the gradient $\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|]$ regarding x is influenced by $\mathbb{E}[\|\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)\|]$ regarding y and $\mathbb{E}[\|\nabla_2 g(\bar{x}_t,\bar{z}_t)\|]$ regarding z. Meanwhile, Lemma B.1 reveals that the gradient $\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|]$ is also affected by the consensus errors regarding both variables and gradients. After revealing this explicit interaction, the remainder of the proof boils down to bounding each factor.

Second Step: After revealing the explicit interaction between three gradients, our next step is to bound $\mathbb{E}[\|\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)\|]$ with respect to y and $\mathbb{E}[\|\nabla_2 g(\bar{x}_t,\bar{z}_t)\|]$ regarding z, so that $\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|]$ can be bounded. However, **this is challenging because $\mathbb{E}[\|\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)\|]$ is affected by the update of two variables simultaneously (the same applies to $\mathbb{E}[\|\nabla_2 g(\bar{x}_t,\bar{z}_t)\|]$), and is thus affected by two normalized variance-reduced gradients**. In our proof, we innovatively handle those normalized variance-reduced gradients and establish the following lemma.

Lemma B.2. *Given Assumption 3.1 and $\eta_x \le \eta_y\frac{\mu}{2(\delta L_f+L_g)}$, we can obtain that*

$$\frac{1}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)\|] \le \frac{2(\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_0,\bar{y}_0)-h_\delta^*(\bar{x}_0)]-\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_T,\bar{y}_T)-h_\delta^*(\bar{x}_T)])}{\eta_y T} + \frac{1}{\delta}2\eta_x(\delta L_f+L_g)+\frac{1}{\delta}\eta_y(\delta L_f+L_g)+\frac{1}{\delta}\frac{\eta_x^2(\delta L_f+L_g)}{\eta_y}+\frac{1}{\delta}\frac{\eta_x^2 L_{h_\delta^*}}{\eta_y} + \textbf{\textit{Gradient Errors}} + \textbf{\textit{Consensus Errors}}\ . \tag{11}$$

Here, $h_\delta^*(x) = h_\delta(x,y^*(x))$ where $y^*(x) = \arg\min_y h_\delta(x,y)$. **Gradient Errors** include: $4\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_2 f^{(k)}(x_t^{(k)},y_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}v_{1,t}^{(k)}\|]$ and $4\frac{1}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_2 g^{(k)}(x_t^{(k)},y_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}v_{2,t}^{(k)}\|]$. **Consensus Errors** include: $4\left(L_f+\frac{L_g}{\delta}\right)\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t-x_t^{(k)}\|]$, $4\left(L_f+\frac{L_g}{\delta}\right)\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t-y_t^{(k)}\|]$, and $2\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{q}_t-q_t^{(k)}\|]$.

Lemma B.2 shows that $\mathbb{E}[\|\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)\|]$ is only affected by **Gradient Errors**, **Consensus Errors**, and some other terms that are not explicitly related to $\mathbb{E}[\|\nabla_2 g(\bar{x}_t,\bar{z}_t)\|]$ and $\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|]$. Therefore, we only need to provide the upper bound of **Gradient Errors** and **Consensus Errors** in order to bound $\mathbb{E}[\|\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)\|]$. Similarly, we can bound $\mathbb{E}[\|\nabla_2 g(\bar{x}_t,\bar{z}_t)\|]$ as Lemma B.2, which is deferred to Lemma C.7 in Appendix C.2 due to space limitation.

Summarization. First, it is worth noting that our proof is fundamentally different from existing decentralized bilevel optimization [57, 13, 5, 6, 66, 22, 69, 37, 34, 35, 53] or decentralized minimax optimization [54, 12, 63, 17, 65] that rely on the finite-variance assumption. For example, the upper bound for $\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|]$ in those methods has a term with regard to σ^2, which could be infinity under heavy-tailed noise. On the contrary, our upper bound does not have this kind of terms. In fact, this is the first work showing how to handle the normalized variance-reduced gradient and heavy-tailed noise for decentralized bilevel optimization. Second, from Lemmas B.1, B.2, C.7, we can observe that they all are affected by **Gradient Errors** and **Consensus Errors**. Then, we need to bound them under heavy-tailed noise.

B.3 Solution for the Second Challenge

First Step. Since the consensus error regarding gradients involves the gradient estimator, e.g., $u_{1,t}^{(k)}$, it can be influenced by both **stochastic noises** and **gradient errors**. For example, Eq. (91) in Appendix C.5 shows that the consensus error regarding the gradient, $\mathbb{E}[\|p_t^{(k)}-\bar{p}_t\|]$, is influenced by stochastic noises, e.g., $\mathbb{E}[\|\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})-\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)};\xi_j^{(k)})\|]$, and gradient errors, e.g., $\mathbb{E}[\|u_{1,j-1}^{(k)}-\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})\|]$. Then, our fist step for this challenge is to establish the

upper bound for **Gradient Errors**. For example, in Lemma B.3, we establish the upper bound for the Gradient Error, $\mathbb{E}[\|u_{1,t}^{(k)} - \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)})\|]$.

Lemma B.3. *Given Assumptions 3.1-3.3, we can obtain*

$$\sum_{k=1}^{K} \mathbb{E}[\|u_{1,t}^{(k)} - \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)})\|] \leq (1-\gamma_x)^t \frac{2\sqrt{2}\sigma K}{B_0^{1-1/s}} + \frac{8(\eta_x+\eta_y)L_f}{(1-\lambda)\sqrt{\gamma_x}} K + 2\sqrt{2}\gamma_x^{1-1/s}\sigma K \,. \tag{12}$$

Note that bounding gradient errors requires addressing the communication step. Lemma B.3 demonstrates the influence of the spectral gap $1-\lambda$ on this bound, which differs from the single-machine setting. Similarly, we established other Gradient Errors in Lemmas C.17- C.21 in Appendix C.4.

Second Step. The second step is to bound the consensus error regarding gradients in terms of **Gradient Errors**. For example, in Lemma B.4, we provide the upper bound for $\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_t^{(k)} - \bar{p}_t\|]$, demonstrating how the heavy-tailed noise (σ), hyperparameters (η_x, η_y, η_z, γ_x), penalty parameter (δ), and spectral gap ($1-\lambda$) affect this upper bound.

Lemma B.4. *Given Assumptions 3.1-3.3, we can obtain*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_t^{(k)} - \bar{p}_t\|] \leq \frac{2\lambda}{(1-\lambda)T}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(x_0^{(k)}, y_0^{(k)})\|] \\
&+ \frac{2\lambda}{(1-\lambda)T}\frac{1}{\delta}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_0^{(k)}, y_0^{(k)})\|] + \frac{2\lambda}{(1-\lambda)T}\frac{1}{\delta}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_0^{(k)}, z_0^{(k)})\|] \\
&+ \frac{\lambda}{(1-\lambda)T}\frac{4\sqrt{2}\sqrt{K}}{B_0^{1-1/s}}\left(1+\frac{2}{\delta}\right)\sigma + \frac{\gamma_x\lambda\sqrt{K}\sigma}{(1-\lambda)^{3/2}}\left(1+\frac{2}{\delta}\right) + \frac{4\eta_x\lambda\sqrt{K}}{(1-\lambda)^{5/2}}\left(L_f + \frac{2L_g}{\delta}\right) \\
&+ \frac{4\eta_y\lambda\sqrt{K}}{(1-\lambda)^{5/2}}\left(L_f + \frac{L_g}{\delta}\right) + \frac{4\eta_z\lambda\sqrt{K}}{(1-\lambda)^{5/2}}\frac{L_g}{\delta} + \frac{\lambda\sqrt{K}}{T(1-\lambda)^{3/2}}\frac{2\sqrt{2}\sigma}{B_0^{1-1/s}}\left(1+\frac{2}{\delta}\right) \\
&+ \frac{2\sqrt{2}\gamma_x^{2-1/s}\lambda\sqrt{K}}{(1-\lambda)^{3/2}}\sigma\left(1+\frac{2}{\delta}\right) + \frac{8\eta_x\sqrt{\gamma_x}\lambda\sqrt{K}}{(1-\lambda)^{5/2}}\left(L_f + \frac{2L_g}{\delta}\right) + \frac{8\eta_y\sqrt{\gamma_x}\lambda\sqrt{K}}{(1-\lambda)^{5/2}}\left(L_f + \frac{2L_g}{\delta}\right) .
\end{aligned} \tag{13}$$

Similarly, we established the upper bounds for other consensus errors regarding gradients in Lemmas C.30, C.31 in Appendix C.5.

After obtaining the upper bounds for the gradients, $\mathbb{E}[\|\nabla_2 h_\delta(\bar{x}_t, \bar{y}_t)\|]$ and $\mathbb{E}[\|\nabla_2 g(\bar{x}_t, \bar{z}_t)\|]$, the upper bounds for **Consensus Errors**, and the upper bounds for **Gradient Errors**, we plug them into Lemma B.1, we can finally obtain the convergence rate of Algorithm 1 in Theorem 4.1.

C Main Proof

This section is organized as follows:

1. **Appendix C.1: Supporting Terminologies and Lemmas**
2. **Appendix C.2: Characterizing Interdependence between Gradients**
3. **Appendix C.3: Bounding Consecutive Updates**
4. **Appendix C.4: Bounding Gradient Errors**
5. **Appendix C.5: Bounding Consensus Errors**
6. **Appendix C.6: Proof of Theorem 4.1**

The proof of Theorem 4.1 follows the structure presented in Section B. Specifically, we first characterize the interdependence between different gradients in Appendix C.2 and then bound Gradient Errors in Appendix C.4 and Consensus Errors in Appendix C.5. Based on them, we prove Theorem 4.1 in Appendix C.6.

C.1 Supporting Terminologies and Lemmas

We define the following terminologies for convergence analysis:

$$
\begin{aligned}
&X_t = [x_t^{(1)}, x_t^{(2)}, \cdots, x_t^{(K)}]\,, \quad Y_t = [y_t^{(1)}, y_t^{(2)}, \cdots, y_t^{(K)}]\,, \quad Z_t = [z_t^{(1)}, z_t^{(2)}, \cdots, z_t^{(K)}]\,,\\
&P_t = [p_t^{(1)}, p_t^{(2)}, \cdots, p_t^{(K)}]\,, \quad Q_t = [q_t^{(1)}, q_t^{(2)}, \cdots, q_t^{(K)}]\,, \quad R_t = [r_t^{(1)}, r_t^{(2)}, \cdots, r_t^{(K)}]\,,\\
&U_t = [u_t^{(1)}, u_t^{(2)}, \cdots, u_t^{(K)}]\,, \quad V_t = [v_t^{(1)}, v_t^{(2)}, \cdots, v_t^{(K)}]\,, \quad W_t = [w_t^{(1)}, w_t^{(2)}, \cdots, w_t^{(K)}]\,,\\
&\hat{P}_t = [\frac{p_t^{(1)}}{\|p_t^{(1)}\|}, \frac{p_t^{(2)}}{\|p_t^{(2)}\|}, \cdots, \frac{p_t^{(K)}}{\|p_t^{(K)}\|}]\,, \quad \hat{Q}_t = [\frac{q_t^{(1)}}{\|q_t^{(1)}\|}, \frac{q_t^{(2)}}{\|q_t^{(2)}\|}, \cdots, \frac{q_t^{(K)}}{\|q_t^{(K)}\|}]\,,\\
&\hat{R}_t = [\frac{r_t^{(1)}}{\|r_t^{(1)}\|}, \frac{r_t^{(2)}}{\|r_t^{(2)}\|}, \cdots, \frac{r_t^{(K)}}{\|r_t^{(K)}\|}]\,,\\
&\bar{X}_t = X_t\frac{\mathbf{1}\mathbf{1}^T}{K}\,, \quad \bar{Y}_t = Y_t\frac{\mathbf{1}\mathbf{1}^T}{K}\,, \quad \bar{Z}_t = Z_t\frac{\mathbf{1}\mathbf{1}^T}{K}\,, \quad \bar{P}_t = P_t\frac{\mathbf{1}\mathbf{1}^T}{K}\,, \quad \bar{Q}_t = Q_t\frac{\mathbf{1}\mathbf{1}^T}{K}\,, \quad \bar{R}_t = R_t\frac{\mathbf{1}\mathbf{1}^T}{K}\,,\\
&\bar{\hat{P}}_t = \hat{P}_t\frac{\mathbf{1}\mathbf{1}^T}{K}\,, \quad \bar{\hat{Q}}_t = \hat{Q}_t\frac{\mathbf{1}\mathbf{1}^T}{K}\,, \quad \bar{\hat{R}}_t = \hat{R}_t\frac{\mathbf{1}\mathbf{1}^T}{K}\,, \quad \bar{U}_t = U_t\frac{\mathbf{1}\mathbf{1}^T}{K}\,, \quad \bar{V}_t = V_t\frac{\mathbf{1}\mathbf{1}^T}{K}\,, \quad \bar{W}_t = W_t\frac{\mathbf{1}\mathbf{1}^T}{K}\,.
\end{aligned}
\tag{14}
$$

Lemma C.1. *[4] Given Assumptions 3.1, then $\Phi(x)$ is L_Φ-smooth, where the constant $L_\Phi = O(\ell\kappa^3)$.*

Lemma C.2. *[4] Given Assumptions 3.1, then $Y^*(x)$ is continuous, i.e., for any $x_1, x_2 \in \mathbb{R}^{d_1}$, the following inequality holds:*

$$
\textit{Dist}(Y^*(x_1), Y^*(x_2)) \leq C_{y^*}\|x_1 - x_2\|\,, \tag{15}
$$

where $C_{y^*} = \frac{L_g}{\mu} = O(\kappa)$, Dist$(\cdot,\cdot)$ denotes the distance between two sets.

Lemma C.3. *(Appendix A of [20]) Given Assumptions 3.1, the following inequality holds:*

$$
\|y^*(x) - z\|^2 \leq \frac{1}{\mu^2}\|\nabla_2 g(x,z)\|^2\,, \qquad \|y_\delta^*(x) - y\|^2 \leq \frac{1}{\mu^2}\|\nabla_2 h_\delta(x,y)\|^2\,. \tag{16}
$$

Lemma C.4. *Given Assumptions 3.1, then $\nabla g^*(x)$ is continuous and $\nabla h_\delta^*(x)$ is also continuous, i.e., for any $x_1, x_2 \in \mathbb{R}^{d_1}$, the following inequalities hold:*

$$
\|\nabla g^*(x_1) - \nabla g^*(x_2)\| \leq L_{g^*}\|x_1 - x_2\|\,, \quad \|\nabla h_\delta^*(x_1) - \nabla h_\delta^*(x_2)\| \leq L_{h_\delta^*}\|x_1 - x_2\|\,, \tag{17}
$$

where $L_{g^*} = L_g(1 + \frac{L_g}{\mu}) = O(\ell\kappa)$ and $L_{h_\delta^*} = (\delta L_f + L_g)(1 + \frac{\delta L_f + L_g}{\mu}) = O(\ell\kappa)$.

This lemma be easily proved by following Lemma A.5 in [41].

Lemma C.5. *[36] Given random vectors v_t that satisfies $\mathbb{E}[v_t|\mathcal{F}_{t-1}] = 0$, where $\mathcal{F}_{t-1}$ is a natural filtration and $t \in \mathbb{N}$, then the following inequality holds:*

$$
\mathbb{E}\left[\left\|\sum_{t=1}^{T} v_t\right\|\right] \leq 2\sqrt{2}\mathbb{E}\left[\left(\sum_{t=1}^{T}\|v_t\|^s\right)^{\frac{1}{s}}\right]\,, \tag{18}
$$

where $T \in \mathbb{N}$ and $s \in (1, 2]$.

C.2 Characterizing Interdependence between Gradients

Lemma C.6. *Given Assumption 3.1, we obtain*

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|] \leq\ & \frac{\mathbb{E}[\Phi(\bar{x}_0)] - \mathbb{E}[\Phi(\bar{x}_T)]}{\eta_x T} + 2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\bar{x}_t) - \nabla\Phi_\delta(\bar{x}_t)\|]\\
&+ \frac{2(\delta L_f + L_g)}{\mu}\frac{1}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_2 h_\delta(\bar{x}_t, \bar{y}_t)\|] + \frac{2L_g}{\mu}\frac{1}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_2 g(\bar{x}_t, \bar{z}_t)\|]\\
&+ 2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}u_{1,t}^{(k)}\|]
\end{aligned}
$$

$$+2\frac{1}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_1 g^{(k)}(x_t^{(k)},y_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}u_{2,t}^{(k)}\|]$$
$$+2\frac{1}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_1 g^{(k)}(x_t^{(k)},z_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}u_{3,t}^{(k)}\|]$$
$$+2(L_f+\frac{2L_g}{\delta})\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_t^{(k)}-\bar{x}_t\|]+2(L_f+\frac{L_g}{\delta})\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_t^{(k)}-\bar{y}_t\|]$$
$$+2\frac{L_g}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|z_t^{(k)}-\bar{z}_t\|]+\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t-p_t^{(k)}\|]+\frac{\eta_x L_\Phi}{2}\,. \tag{19}$$

Proof. Due to the smoothness of $\Phi(x)$, we obtain

$$\begin{aligned}
&\mathbb{E}[\Phi(\bar{x}_{t+1})]\le\mathbb{E}[\Phi(\bar{x}_t)]+\mathbb{E}[\langle\nabla\Phi(\bar{x}_t),\bar{x}_{t+1}-\bar{x}_t\rangle]+\frac{L_\Phi}{2}\mathbb{E}[\|\bar{x}_{t+1}-\bar{x}_t\|^2]\\
&=\mathbb{E}[\Phi(\bar{x}_t)]-\eta_x\mathbb{E}[\langle\nabla\Phi(\bar{x}_t),\frac{1}{K}\sum_{k=1}^{K}\frac{p_t^{(k)}}{\|p_t^{(k)}\|}\rangle]+\frac{\eta_x^2L_\Phi}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\frac{p_t^{(k)}}{\|p_t^{(k)}\|}\|^2]\\
&\overset{(a)}{\le}\mathbb{E}[\Phi(\bar{x}_t)]\underbrace{-\eta_x\mathbb{E}[\langle\nabla\Phi(\bar{x}_t)-\bar{p}_t,\frac{1}{K}\sum_{k=1}^{K}\frac{p_t^{(k)}}{\|p_t^{(k)}\|}\rangle]}_{T_1}\underbrace{-\eta_x\mathbb{E}[\langle\bar{p}_t,\frac{1}{K}\sum_{k=1}^{K}\frac{p_t^{(k)}}{\|p_t^{(k)}\|}\rangle]}_{T_2}+\frac{\eta_x^2L_\Phi}{2}\,,
\end{aligned}\tag{20}$$

where (a) holds due to $\|\frac{p_t^{(k)}}{\|p_t^{(k)}\|}\|=1$.

For T_1, we bound it as follows:

$$T_1\le\eta_x\mathbb{E}[\|\nabla\Phi(\bar{x}_t)-\bar{p}_t\|\|\frac{1}{K}\sum_{k=1}^{K}\frac{p_t^{(k)}}{\|p_t^{(k)}\|}\|]\le\eta_x\mathbb{E}[\|\nabla\Phi(\bar{x}_t)-\bar{p}_t\|]\,. \tag{21}$$

For T_2, we bound it as follows:

$$\begin{aligned}
T_2&=-\eta_x\mathbb{E}[\langle\bar{p}_t,\frac{1}{K}\sum_{k=1}^{K}\frac{p_t^{(k)}}{\|p_t^{(k)}\|}-\frac{\bar{p}_t}{\|\bar{p}_t\|}\rangle]-\eta_x\mathbb{E}[\langle\bar{p}_t,\frac{\bar{p}_t}{\|\bar{p}_t\|}\rangle]\\
&\le\eta_x\mathbb{E}[\|\bar{p}_t\|\|\frac{1}{K}\sum_{k=1}^{K}\frac{p_t^{(k)}}{\|p_t^{(k)}\|}-\frac{\bar{p}_t}{\|\bar{p}_t\|}\|]-\eta_x\mathbb{E}[\|\bar{p}_t\|]\\
&=\eta_x\mathbb{E}[\|\bar{p}_t\|\|\frac{1}{K}\sum_{k=1}^{K}\frac{p_t^{(k)}}{\|p_t^{(k)}\|}-\frac{1}{K}\sum_{k=1}^{K}\frac{p_t^{(k)}}{\|\bar{p}_t\|}\|]-\eta_x\mathbb{E}[\|\bar{p}_t\|]\\
&\le\eta_x\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t\|\|p_t^{(k)}\|\|\frac{1}{\|p_t^{(k)}\|}-\frac{1}{\|\bar{p}_t\|}\|]-\eta_x\mathbb{E}[\|\bar{p}_t\|]\\
&\le\eta_x\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t-p_t^{(k)}\|]-\eta_x\mathbb{E}[\|\bar{p}_t\|]\\
&\overset{(a)}{\le}\eta_x\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t-p_t^{(k)}\|]-\eta_x\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|]+\eta_x\mathbb{E}[\|\nabla\Phi(\bar{x}_t)-\bar{p}_t\|]\,,
\end{aligned}\tag{22}$$

where (a) holds due to the following inequality:

$$\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|]\le\mathbb{E}[\|\nabla\Phi(\bar{x}_t)-\bar{p}_t\|]+\mathbb{E}[\|\bar{p}_t\|]\,. \tag{23}$$

Therefore, we obtain

$$\mathbb{E}[\Phi(\bar{x}_{t+1})]\le\mathbb{E}[\Phi(\bar{x}_t)]-\eta_x\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|]+\eta_x\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t-p_t^{(k)}\|]+\frac{\eta_x^2L_\Phi}{2}$$

$$+ 2\eta_x \mathbb{E}[\|\nabla\Phi(\bar{x}_t) - \bar{p}_t\|] \,. \tag{24}$$

For $\mathbb{E}[\|\nabla\Phi(\bar{x}_t) - \bar{p}_t\|]$, we bound it as follows:

$$\begin{aligned}
&\mathbb{E}[\|\nabla\Phi(\bar{x}_t) - \bar{p}_t\|] \\
&\leq \mathbb{E}[\|\nabla\Phi(\bar{x}_t) - \nabla\Phi_\delta(\bar{x}_t)\|] + \mathbb{E}[\|\nabla\Phi_\delta(\bar{x}_t) - \nabla_x\Phi_\delta(\bar{x}_t, \bar{y}_t, \bar{z}_t)\|] + \mathbb{E}[\|\nabla_x\Phi_\delta(\bar{x}_t, \bar{y}_t, \bar{z}_t) - \bar{p}_t\|] \\
&\overset{(a)}{\leq} \mathbb{E}[\|\nabla\Phi(\bar{x}_t) - \nabla\Phi_\delta(\bar{x}_t)\|] + (L_f + \frac{L_g}{\delta})\mathbb{E}[\|y_\delta^*(\bar{x}_t) - \bar{y}_t\|] + \frac{L_g}{\delta}\mathbb{E}[\|y^*(\bar{x}_t) - \bar{z}_t\|] \\
&\quad + \mathbb{E}[\|\nabla_x\Phi_\delta(\bar{x}_t, \bar{y}_t, \bar{z}_t) - \bar{p}_t\|] \\
&\overset{(b)}{\leq} \mathbb{E}[\|\nabla\Phi(\bar{x}_t) - \nabla\Phi_\delta(\bar{x}_t)\|] + \frac{1}{\mu}(L_f + \frac{L_g}{\delta})\mathbb{E}[\|\nabla_2 h_\delta(\bar{x}_t, \bar{y}_t)\|] + \frac{1}{\mu}\frac{L_g}{\delta}\mathbb{E}[\|\nabla_2 g(\bar{x}_t, \bar{z}_t)\|] \\
&\quad + \mathbb{E}[\|\nabla_1 f(\bar{x}_t, \bar{y}_t) - \bar{u}_{1,t}\|] + \frac{1}{\delta}\mathbb{E}[\|\nabla_1 g(\bar{x}_t, \bar{y}_t) - \bar{u}_{2,t}\|] + \frac{1}{\delta}\mathbb{E}[\|\nabla_1 g(\bar{x}_t, \bar{z}_t) - \bar{u}_{3,t}\|] \\
&\leq \mathbb{E}[\|\nabla\Phi(\bar{x}_t) - \nabla\Phi_\delta(\bar{x}_t)\|] + \frac{1}{\mu}(L_f + \frac{L_g}{\delta})\mathbb{E}[\|\nabla_2 h_\delta(\bar{x}_t, \bar{y}_t)\|] + \frac{1}{\mu}\frac{L_g}{\delta}\mathbb{E}[\|\nabla_2 g(\bar{x}_t, \bar{z}_t)\|] \\
&\quad + \mathbb{E}[\|\nabla_1 f(\bar{x}_t, \bar{y}_t) - \frac{1}{K}\sum_{k=1}^K \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)})\|] + \mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K u_{1,t}^{(k)}\|] \\
&\quad + \frac{1}{\delta}\mathbb{E}[\|\nabla_1 g(\bar{x}_t, \bar{y}_t) - \frac{1}{K}\sum_{k=1}^K \nabla_1 g^{(k)}(x_t^{(k)}, y_t^{(k)})\|] + \frac{1}{\delta}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_1 g^{(k)}(x_t^{(k)}, y_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K u_{2,t}^{(k)}\|] \\
&\quad + \frac{1}{\delta}\mathbb{E}[\|\nabla_1 g(\bar{x}_t, \bar{z}_t) - \frac{1}{K}\sum_{k=1}^K \nabla_1 g^{(k)}(x_t^{(k)}, z_t^{(k)})\|] + \frac{1}{\delta}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_1 g^{(k)}(x_t^{(k)}, z_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K u_{3,t}^{(k)}\|] \\
&\overset{(c)}{\leq} \mathbb{E}[\|\nabla\Phi(\bar{x}_t) - \nabla\Phi_\delta(\bar{x}_t)\|] + \frac{1}{\mu}(L_f + \frac{L_g}{\delta})\mathbb{E}[\|\nabla_2 h_\delta(\bar{x}_t, \bar{y}_t)\|] + \frac{1}{\mu}\frac{L_g}{\delta}\mathbb{E}[\|\nabla_2 g(\bar{x}_t, \bar{z}_t)\|] \\
&\quad + \mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K u_{1,t}^{(k)}\|] \\
&\quad + \frac{1}{\delta}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_1 g^{(k)}(x_t^{(k)}, y_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K u_{2,t}^{(k)}\|] + \frac{1}{\delta}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_1 g^{(k)}(x_t^{(k)}, z_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K u_{3,t}^{(k)}\|] \\
&\quad + (L_f + \frac{2L_g}{\delta})\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|x_t^{(k)} - \bar{x}_t\|] + (L_f + \frac{L_g}{\delta})\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|y_t^{(k)} - \bar{y}_t\|] + \frac{L_g}{\delta}\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|z_t^{(k)} - \bar{z}_t\|] \,,
\end{aligned}$$

where (a) holds due to Assumption 3.1, (b) holds due to Lemma C.3, and (c) holds due to Assumption 3.1.

By combining the above two inequalities, we complete the proof.

□

Lemma C.7. *Given Assumption 3.1 and $\eta_x \leq \frac{\mu}{2L_g}\eta_z$, we obtain*

$$\begin{aligned}
\frac{1}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_2 g(\bar{x}_t, \bar{z}_t)\|] &\leq \frac{2(\frac{1}{\delta}\mathbb{E}[g(\bar{x}_0, \bar{z}_0) - g^*(\bar{x}_0)] - \frac{1}{\delta}\mathbb{E}[g(\bar{x}_T, \bar{z}_T) - g^*(\bar{x}_T)])}{\eta_z T} \\
&\quad + 4\frac{1}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_2 g^{(k)}(x_t^{(k)}, z_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K w_{1,t}^{(k)}\|] \\
&\quad + 4\frac{L_g}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|x_t^{(k)} - \bar{x}_t\|] + 4\frac{L_g}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|z_t^{(k)} - \bar{z}_t\|] \\
&\quad + 2\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{r}_t - r_t^{(k)}\|] + \frac{1}{\delta}2\eta_x L_g + \frac{1}{\delta}\eta_z L_g + \frac{1}{\delta}\frac{\eta_x^2 L_g}{\eta_z} + \frac{1}{\delta}\frac{\eta_x^2 L_{g^*}}{\eta_z} \,.
\end{aligned} \tag{25}$$

Proof. Due to the smoothness of g, we obtain

$$\begin{aligned}
&\frac{1}{\delta}\mathbb{E}[g(\bar{x}_{t+1},\bar{z}_{t+1})] \le \frac{1}{\delta}\mathbb{E}[g(\bar{x}_{t+1},\bar{z}_t)] + \frac{1}{\delta}\mathbb{E}[\langle \nabla_2 g(\bar{x}_{t+1},\bar{z}_t), \bar{z}_{t+1}-\bar{z}_t\rangle] + \frac{1}{\delta}\frac{L_g}{2}\mathbb{E}[\|\bar{z}_{t+1}-\bar{z}_t\|^2] \\
&= \frac{1}{\delta}\mathbb{E}[g(\bar{x}_{t+1},\bar{z}_t)] - \eta_z\mathbb{E}[\langle \frac{1}{\delta}\nabla_2 g(\bar{x}_{t+1},\bar{z}_t), \frac{1}{K}\sum_{k=1}^K \frac{r_t^{(k)}}{\|r_t^{(k)}\|}\rangle] + \frac{1}{\delta}\frac{\eta_z^2 L_g}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \frac{r_t^{(k)}}{\|r_t^{(k)}\|}\|^2] \\
&\overset{(a)}{\le} \frac{1}{\delta}\mathbb{E}[g(\bar{x}_{t+1},\bar{z}_t)] + \frac{1}{\delta}\frac{\eta_z^2 L_g}{2} \\
&\underbrace{-\eta_z\mathbb{E}[\langle \frac{1}{\delta}\nabla_2 g(\bar{x}_{t+1},\bar{z}_t) - \bar{r}_t, \frac{1}{K}\sum_{k=1}^K \frac{r_t^{(k)}}{\|r_t^{(k)}\|}\rangle]}_{T_1} \underbrace{-\eta_z\mathbb{E}[\langle \bar{r}_t, \frac{1}{K}\sum_{k=1}^K \frac{r_t^{(k)}}{\|r_t^{(k)}\|}\rangle]}_{T_2},
\end{aligned} \tag{26}$$

where (a) holds due to $\|\frac{r_t^{(k)}}{\|r_t^{(k)}\|}\| = 1$.

Similar to the proof of Lemma C.6, for T_1, we obtain

$$\begin{aligned}
T_1 &\le \eta_z\mathbb{E}[\|\frac{1}{\delta}\nabla_2 g(\bar{x}_{t+1},\bar{z}_t) - \bar{r}_t\|] \\
&\le \eta_z\mathbb{E}[\|\frac{1}{\delta}\nabla_2 g(\bar{x}_{t+1},\bar{z}_t) - \frac{1}{\delta}\nabla_2 g(\bar{x}_t,\bar{z}_t)\|] + \eta_z\mathbb{E}[\|\frac{1}{\delta}\nabla_2 g(\bar{x}_t,\bar{z}_t) - \bar{r}_t\|] \\
&\le \eta_z\frac{L_g}{\delta}\mathbb{E}[\|\bar{x}_{t+1}-\bar{x}_t\|] + \eta_z\mathbb{E}[\|\frac{1}{\delta}\nabla_2 g(\bar{x}_t,\bar{z}_t) - \bar{r}_t\|] \\
&= \eta_x\eta_z\frac{L_g}{\delta}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \frac{p_t^{(k)}}{\|p_t^{(k)}\|}\|] + \eta_z\mathbb{E}[\|\frac{1}{\delta}\nabla_2 g(\bar{x}_t,\bar{z}_t) - \bar{r}_t\|] \\
&\le \eta_x\eta_z\frac{L_g}{\delta} + \eta_z\mathbb{E}[\|\frac{1}{\delta}\nabla_2 g(\bar{x}_t,\bar{z}_t) - \bar{r}_t\|] .
\end{aligned} \tag{27}$$

In addition, similar to the proof of Lemma C.6, for T_2, we obtain

$$\begin{aligned}
T_2 &\le \eta_z\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{r}_t - r_t^{(k)}\|] - \eta_z\mathbb{E}[\|\bar{r}_t\|] \\
&\le \eta_z\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{r}_t - r_t^{(k)}\|] - \eta_z\mathbb{E}[\|\frac{1}{\delta}\nabla_2 g(\bar{x}_t,\bar{z}_t)\|] + \eta_z\mathbb{E}[\|\frac{1}{\delta}\nabla_2 g(\bar{x}_t,\bar{z}_t) - \bar{r}_t\|] .
\end{aligned} \tag{28}$$

Then, we obtain

$$\begin{aligned}
&\frac{1}{\delta}\mathbb{E}[g(\bar{x}_{t+1},\bar{z}_{t+1})] \le \frac{1}{\delta}\mathbb{E}[g(\bar{x}_{t+1},\bar{z}_t)] - \eta_z\mathbb{E}[\|\frac{1}{\delta}\nabla_2 g(\bar{x}_t,\bar{z}_t)\|] + \eta_z\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{r}_t - r_t^{(k)}\|] \\
&\quad + 2\eta_z\mathbb{E}[\|\frac{1}{\delta}\nabla_2 g(\bar{x}_t,\bar{z}_t) - \bar{r}_t\|] + \eta_x\eta_z\frac{L_g}{\delta} + \frac{1}{\delta}\frac{\eta_z^2 L_g}{2} .
\end{aligned} \tag{29}$$

For $\mathbb{E}[\|\frac{1}{\delta}\nabla_2 g(\bar{x}_t,\bar{z}_t) - \bar{r}_t\|]$, we bound it as follows:

$$\begin{aligned}
&\mathbb{E}[\|\frac{1}{\delta}\nabla_2 g(\bar{x}_t,\bar{z}_t) - \bar{r}_t\|] \\
&\le \mathbb{E}[\|\frac{1}{\delta}\nabla_2 g(\bar{x}_t,\bar{z}_t) - \frac{1}{\delta}\frac{1}{K}\sum_{k=1}^K \nabla_2 g^{(k)}(x_t^{(k)},z_t^{(k)})\|] + \mathbb{E}[\|\frac{1}{\delta}\frac{1}{K}\sum_{k=1}^K \nabla_2 g^{(k)}(x_t^{(k)},z_t^{(k)}) - \frac{1}{\delta}\frac{1}{K}\sum_{k=1}^K w_{1,t}^{(k)}\|] \\
&\overset{(a)}{\le} \frac{L_g}{\delta}\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|x_t^{(k)} - \bar{x}_t\|] + \frac{L_g}{\delta}\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|z_t^{(k)} - \bar{z}_t\|] + \frac{1}{\delta}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_2 g^{(k)}(x_t^{(k)},z_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K w_{1,t}^{(k)}\|] ,
\end{aligned} \tag{30}$$

where (a) holds due to Assumption 3.1.

By combining the above two inequalities, we obtain

$$\begin{aligned}
\frac{1}{\delta}\mathbb{E}[g(\bar{x}_{t+1},\bar{z}_{t+1})] \leq \frac{1}{\delta}\mathbb{E}[g(\bar{x}_{t+1},\bar{z}_t)] - \eta_z\mathbb{E}[\|\frac{1}{\delta}\nabla_2 g(\bar{x}_t,\bar{z}_t)\|] + \eta_z\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{r}_t - r_t^{(k)}\|] \\
+ 2\eta_z\frac{L_g}{\delta}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_t^{(k)} - \bar{x}_t\|] + 2\eta_z\frac{L_g}{\delta}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|z_t^{(k)} - \bar{z}_t\|] \\
+ 2\eta_z\frac{1}{\delta}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_2 g^{(k)}(x_t^{(k)},z_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}w_{1,t}^{(k)}\|] + \eta_x\eta_z\frac{L_g}{\delta} + \frac{1}{\delta}\frac{\eta_z^2 L_g}{2} \ . \qquad (31)
\end{aligned}$$

Moreover, due to the smoothness of g, we obtain

$$\begin{aligned}
&\frac{1}{\delta}\mathbb{E}[g(\bar{x}_{t+1},\bar{z}_t)] \leq \frac{1}{\delta}\mathbb{E}[g(\bar{x}_t,\bar{z}_t)] + \frac{1}{\delta}\mathbb{E}[\langle\nabla_1 g(\bar{x}_t,\bar{z}_t),\bar{x}_{t+1}-\bar{x}_t\rangle] + \frac{1}{\delta}\frac{L_g}{2}\mathbb{E}[\|\bar{x}_{t+1}-\bar{x}_t\|^2] \\
&= \frac{1}{\delta}\mathbb{E}[g(\bar{x}_t,\bar{z}_t)] + \frac{1}{\delta}\mathbb{E}[\langle\nabla_1 g(\bar{x}_t,\bar{z}_t) - \nabla_x g(\bar{x}_t,y^*(\bar{x}_t)),\bar{x}_{t+1}-\bar{x}_t\rangle] \\
&\quad + \frac{1}{\delta}\mathbb{E}[\langle\nabla_x g(\bar{x}_t,y^*(\bar{x}_t)),\bar{x}_{t+1}-\bar{x}_t\rangle] + \frac{1}{\delta}\frac{L_g}{2}\mathbb{E}[\|\bar{x}_{t+1}-\bar{x}_t\|^2] \\
&= \frac{1}{\delta}\mathbb{E}[g(\bar{x}_t,\bar{z}_t)] - \eta_x\mathbb{E}[\langle\frac{1}{\delta}(\nabla_1 g(\bar{x}_t,\bar{z}_t) - \nabla_x g(\bar{x}_t,y^*(\bar{x}_t))),\frac{1}{K}\sum_{k=1}^{K}\frac{p_t^{(k)}}{\|p_t^{(k)}\|}\rangle] \\
&\quad + \frac{1}{\delta}\mathbb{E}[\langle\nabla_x g(\bar{x}_t,y^*(\bar{x}_t)),\bar{x}_{t+1}-\bar{x}_t\rangle] + \frac{1}{\delta}\frac{\eta_x^2 L_g}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\frac{p_t^{(k)}}{\|p_t^{(k)}\|}\|^2] \\
&\overset{(a)}{\leq} \frac{1}{\delta}\mathbb{E}[g(\bar{x}_t,\bar{z}_t)] + \eta_x\frac{1}{\delta}\mathbb{E}[\|\nabla_1 g(\bar{x}_t,\bar{z}_t) - \nabla_1 g(\bar{x}_t,y^*(\bar{x}_t))\|] \\
&\quad + \frac{1}{\delta}\mathbb{E}[\langle\nabla_x g(\bar{x}_t,y^*(\bar{x}_t)),\bar{x}_{t+1}-\bar{x}_t\rangle] + \frac{1}{\delta}\frac{\eta_x^2 L_g}{2} \\
&\overset{(b)}{\leq} \frac{1}{\delta}\mathbb{E}[g(\bar{x}_t,\bar{z}_t)] + \eta_x\frac{L_g}{\delta}\mathbb{E}[\|\bar{z}_t - y^*(\bar{x}_t)\|] + \frac{1}{\delta}\mathbb{E}[\langle\nabla_x g(\bar{x}_t,y^*(\bar{x}_t)),\bar{x}_{t+1}-\bar{x}_t\rangle] + \frac{1}{\delta}\frac{\eta_x^2 L_g}{2} \\
&\overset{(c)}{\leq} \frac{1}{\delta}\mathbb{E}[g(\bar{x}_t,\bar{z}_t)] + \eta_x\frac{L_g}{\mu\delta}\mathbb{E}[\|\nabla_2 g(\bar{x}_t,\bar{z}_t)\|] + \frac{1}{\delta}\mathbb{E}[\langle\nabla_x g(\bar{x}_t,y^*(\bar{x}_t)),\bar{x}_{t+1}-\bar{x}_t\rangle] + \frac{1}{\delta}\frac{\eta_x^2 L_g}{2} \ , \qquad (32)
\end{aligned}$$

where (a) holds due to $\nabla_x g(\bar{x}_t,y^*(\bar{x}_t)) = \nabla_1 g(\bar{x}_t,y^*(\bar{x}_t)) + \nabla y^*(\bar{x}_t)\nabla_2 g(\bar{x}_t,y^*(\bar{x}_t)) = \nabla_1 g(\bar{x}_t,y^*(\bar{x}_t))$ and $\|\frac{p_t^{(k)}}{\|p_t^{(k)}\|}\| = 1$, (b) holds due to Assumption 3.1, and (c) holds due to Lemma C.3.

Furthermore, due to the smoothness of $g^*(x)$ as shown in Lemma C.4, we obtain

$$\begin{aligned}
&\frac{1}{\delta}g^*(\bar{x}_{t+1}) \geq \frac{1}{\delta}g^*(\bar{x}_t) + \frac{1}{\delta}\langle\nabla g^*(\bar{x}_t),\bar{x}_{t+1}-\bar{x}_t\rangle - \frac{1}{\delta}\frac{L_{g^*}}{2}\|\bar{x}_{t+1}-\bar{x}_t\|^2 \\
&= \frac{1}{\delta}g^*(\bar{x}_t) + \frac{1}{\delta}\langle\nabla_x g(\bar{x}_t,y^*(\bar{x}_t)),\bar{x}_{t+1}-\bar{x}_t\rangle - \frac{1}{\delta}\frac{\eta_x^2 L_{g^*}}{2}\|\frac{1}{K}\sum_{k=1}^{K}\frac{p_t^{(k)}}{\|p_t^{(k)}\|}\|^2 \\
&\geq \frac{1}{\delta}g^*(\bar{x}_t) + \frac{1}{\delta}\langle\nabla_x g(\bar{x}_t,y^*(\bar{x}_t)),\bar{x}_{t+1}-\bar{x}_t\rangle - \frac{1}{\delta}\frac{\eta_x^2 L_{g^*}}{2} \ . \qquad (33)
\end{aligned}$$

Then, we obtain

$$\frac{1}{\delta}g^*(\bar{x}_t) - \frac{1}{\delta}g^*(\bar{x}_{t+1}) \leq -\frac{1}{\delta}\langle\nabla_x g(\bar{x}_t,y^*(\bar{x}_t)),\bar{x}_{t+1}-\bar{x}_t\rangle + \frac{1}{\delta}\frac{\eta_x^2 L_{g^*}}{2} \ . \qquad (34)$$

Finally, we obtain

$$\frac{1}{\delta}\mathbb{E}[g(\bar{x}_{t+1},\bar{z}_{t+1})] - \frac{1}{\delta}\mathbb{E}[g^*(\bar{x}_{t+1})]$$

$$
\begin{aligned}
&= \frac{1}{\delta}\mathbb{E}[g(\bar{x}_{t+1},\bar{z}_{t+1})] - \frac{1}{\delta}\mathbb{E}[g(\bar{x}_{t+1},\bar{z}_t)] + \frac{1}{\delta}\mathbb{E}[g(\bar{x}_{t+1},\bar{z}_t)] - \frac{1}{\delta}\mathbb{E}[g(\bar{x}_t,\bar{z}_t)] \\
&\quad + \frac{1}{\delta}\mathbb{E}[g(\bar{x}_t,\bar{z}_t)] - \frac{1}{\delta}\mathbb{E}[g^*(\bar{x}_t)] + \frac{1}{\delta}\mathbb{E}[g^*(\bar{x}_t)] - \frac{1}{\delta}\mathbb{E}[g^*(\bar{x}_{t+1})] \\
&\leq \frac{1}{\delta}\mathbb{E}[g(\bar{x}_t,\bar{z}_t)] - \frac{1}{\delta}\mathbb{E}[g^*(\bar{x}_t)] - \eta_z\mathbb{E}[\|\frac{1}{\delta}\nabla_2 g(\bar{x}_t,\bar{z}_t)\|] + \eta_z\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{r}_t - r_t^{(k)}\|] \\
&\quad + 2\eta_z\frac{L_g}{\delta}\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|x_t^{(k)} - \bar{x}_t\|] + 2\eta_z\frac{L_g}{\delta}\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|z_t^{(k)} - \bar{z}_t\|] \\
&\quad + 2\eta_z\frac{1}{\delta}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K\nabla_2 g^{(k)}(x_t^{(k)},z_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K w_{1,t}^{(k)}\|] + \eta_x\eta_z\frac{L_g}{\delta} + \frac{1}{\delta}\frac{\eta_z^2 L_g}{2} \\
&\quad + \eta_x\frac{L_g}{\mu\delta}\mathbb{E}[\|\nabla_2 g(\bar{x}_t,\bar{z}_t)\|] + \frac{1}{\delta}\mathbb{E}[\langle\nabla_x g(\bar{x}_t,y^*(\bar{x}_t)),\bar{x}_{t+1}-\bar{x}_t\rangle] + \frac{1}{\delta}\frac{\eta_x^2 L_g}{2} \\
&\quad - \frac{1}{\delta}\mathbb{E}[\langle\nabla_x g(\bar{x}_t,y^*(\bar{x}_t)),\bar{x}_{t+1}-\bar{x}_t\rangle] + \frac{1}{\delta}\frac{\eta_x^2 L_{g^*}}{2} \\
&= \frac{1}{\delta}\mathbb{E}[g(\bar{x}_t,\bar{z}_t)] - \frac{1}{\delta}\mathbb{E}[g^*(\bar{x}_t)] + \left(\eta_x\frac{L_g}{\mu} - \eta_z\right)\frac{1}{\delta}\mathbb{E}[\|\nabla_2 g(\bar{x}_t,\bar{z}_t)\|] + \eta_z\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{r}_t - r_t^{(k)}\|] \\
&\quad + 2\eta_z\frac{L_g}{\delta}\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|x_t^{(k)} - \bar{x}_t\|] + 2\eta_z\frac{L_g}{\delta}\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|z_t^{(k)} - \bar{z}_t\|] + \frac{1}{\delta}\eta_x\eta_z L_g + \frac{1}{\delta}\frac{\eta_z^2 L_g}{2} \\
&\quad + 2\eta_z\frac{1}{\delta}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K\nabla_2 g^{(k)}(x_t^{(k)},z_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K w_{1,t}^{(k)}\|] + \frac{1}{\delta}\frac{\eta_x^2 L_g}{2} + \frac{1}{\delta}\frac{\eta_x^2 L_{g^*}}{2}\,. \qquad (35)
\end{aligned}
$$

By setting $\eta_x \leq \frac{\mu}{2L_g}\eta_z$, we complete the proof.

□

Lemma C.8. *Given Assumption 3.1 and $\eta_x \leq \eta_y\frac{\mu}{2L_{h_\delta}}$, where $L_{h_\delta} = \delta L_f + L_g$, we obtain that*

$$
\begin{aligned}
&\frac{1}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)\|] \leq \frac{2(\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_0,\bar{y}_0) - h_\delta^*(\bar{x}_0)] - \frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_T,\bar{y}_T) - h_\delta^*(\bar{x}_T)])}{\eta_y T} \\
&+ 4\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K\nabla_2 f^{(k)}(x_t^{(k)},y_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K v_{1,t}^{(k)}\|] \\
&+ 4\frac{1}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K\nabla_2 g^{(k)}(x_t^{(k)},y_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K v_{2,t}^{(k)}\|] \\
&+ 4\left(L_f + \frac{L_g}{\delta}\right)\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|] + 4\left(L_f + \frac{L_g}{\delta}\right)\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|] \\
&+ 2\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{q}_t - q_t^{(k)}\|] + \frac{1}{\delta}2\eta_x L_{h_\delta} + \frac{1}{\delta}\eta_y L_{h_\delta} + \frac{1}{\delta}\frac{\eta_x^2 L_{h_\delta}}{\eta_y} + \frac{1}{\delta}\frac{\eta_x^2 L_{h_\delta^*}}{\eta_y}\,. \qquad (36)
\end{aligned}
$$

Proof. Given Assumptions 3.1, it is easy to know that $h_\delta(x,y) = \delta f(x,y) + g(x,y)$ is L_{h_δ}-smooth with $L_{h_\delta} = \delta L_f + L_g$.

Then, based on its smoothness, we obtain

$$
\begin{aligned}
&\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_{t+1},\bar{y}_{t+1})] \leq \frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_{t+1},\bar{y}_t)] + \frac{1}{\delta}\mathbb{E}[\langle\nabla_2 h_\delta(\bar{x}_{t+1},\bar{y}_t),\bar{y}_{t+1}-\bar{y}_t\rangle] + \frac{1}{\delta}\frac{L_{h_\delta}}{2}\mathbb{E}[\|\bar{y}_{t+1}-\bar{y}_t\|^2] \\
&= \frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_{t+1},\bar{y}_t)] - \eta_y\mathbb{E}[\langle\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_{t+1},\bar{y}_t),\frac{1}{K}\sum_{k=1}^K\frac{q_t^{(k)}}{\|q_t^{(k)}\|}\rangle] + \frac{1}{\delta}\frac{\eta_y^2 L_{h_\delta}}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K\frac{q_t^{(k)}}{\|q_t^{(k)}\|}\|^2]
\end{aligned}
$$

$$\overset{(a)}{\leq} \frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_{t+1},\bar{y}_t)] + \frac{1}{\delta}\frac{\eta_y^2 L_{h_\delta}}{2}$$
$$\underbrace{-\eta_y\mathbb{E}[\langle\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_{t+1},\bar{y}_t)-\bar{q}_t, \frac{1}{K}\sum_{k=1}^K\frac{q_t^{(k)}}{\|q_t^{(k)}\|}\rangle]}_{T_1}\underbrace{-\eta_y\mathbb{E}[\langle\bar{q}_t, \frac{1}{K}\sum_{k=1}^K\frac{q_t^{(k)}}{\|q_t^{(k)}\|}\rangle]}_{T_2}, \tag{37}$$

where (a) holds due to $\|\frac{q_t^{(k)}}{\|q_t^{(k)}\|}\| = 1$.

For T_1, we bound it as follows:

$$\begin{aligned}
T_1 &\leq \eta_y\mathbb{E}[\|\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_{t+1},\bar{y}_t)-\bar{q}_t\|\|\frac{1}{K}\sum_{k=1}^K\frac{q_t^{(k)}}{\|q_t^{(k)}\|}\|] \\
&\leq \eta_y\mathbb{E}[\|\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_{t+1},\bar{y}_t)-\bar{q}_t\|] \\
&\leq \eta_y\mathbb{E}[\|\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_{t+1},\bar{y}_t)-\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)\|] + \eta_y\mathbb{E}[\|\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)-\bar{q}_t\|] \\
&\overset{(a)}{\leq} \eta_y L_{h_\delta}\frac{1}{\delta}\mathbb{E}[\|\bar{x}_{t+1}-\bar{x}_t\|] + \eta_y\mathbb{E}[\|\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)-\bar{q}_t\|] \\
&\overset{(b)}{=} \frac{1}{\delta}\eta_x\eta_y L_{h_\delta} + \eta_y\mathbb{E}[\|\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)-\bar{q}_t\|],
\end{aligned} \tag{38}$$

where (a) holds due to Assumption 3.1, and (b) holds due to $\|\frac{p_t^{(k)}}{\|p_t^{(k)}\|}\| = 1$.

Similar to the proof of Lemma C.6, for T_2, we bound it as follows:

$$\begin{aligned}
T_2 &\leq \eta_y\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{q}_t - q_t^{(k)}\|] - \eta_y\mathbb{E}[\|\bar{q}_t\|] \\
&\leq \eta_y\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{q}_t - q_t^{(k)}\|] - \eta_y\mathbb{E}[\|\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)\|] + \eta_y\mathbb{E}[\|\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)-\bar{q}_t\|].
\end{aligned} \tag{39}$$

Then, we obtain

$$\begin{aligned}
\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_{t+1},\bar{y}_{t+1})] &\leq \frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_{t+1},\bar{y}_t)] - \eta_y\mathbb{E}[\|\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)\|] + \frac{1}{\delta}\frac{\eta_y^2 L_{h_\delta}}{2} + \frac{1}{\delta}\eta_x\eta_y L_{h_\delta} \\
&\quad + \eta_y\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{q}_t - q_t^{(k)}\|] + 2\eta_y\mathbb{E}[\|\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)-\bar{q}_t\|].
\end{aligned} \tag{40}$$

For $\mathbb{E}[\|\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)-\bar{q}_t\|]$, we bound it as follows:

$$\begin{aligned}
&\mathbb{E}[\|\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)-\bar{q}_t\|] = \mathbb{E}[\|\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)-\bar{v}_t\|] \\
&\leq \mathbb{E}[\|\nabla_2 f(\bar{x}_t,\bar{y}_t)-\bar{v}_{1,t}\|] + \mathbb{E}[\|\frac{1}{\delta}\nabla_2 g(\bar{x}_t,\bar{y}_t)-\frac{1}{\delta}\bar{v}_{2,t}\|] \\
&\leq \mathbb{E}[\|\nabla_2 f(\bar{x}_t,\bar{y}_t)-\frac{1}{K}\sum_{k=1}^K\nabla_2 f^{(k)}(x_t^{(k)},y_t^{(k)})\|] + \mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K\nabla_2 f^{(k)}(x_t^{(k)},y_t^{(k)})-\frac{1}{K}\sum_{k=1}^K v_{1,t}^{(k)}\|] \\
&\quad + \frac{1}{\delta}\mathbb{E}[\|\nabla_2 g(\bar{x}_t,\bar{y}_t)-\frac{1}{K}\sum_{k=1}^K\nabla_2 g^{(k)}(x_t^{(k)},y_t^{(k)})\|] + \frac{1}{\delta}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K\nabla_2 g^{(k)}(x_t^{(k)},y_t^{(k)})-\frac{1}{K}\sum_{k=1}^K v_{2,t}^{(k)}\|] \\
&\overset{(a)}{\leq} \left(L_f+\frac{L_g}{\delta}\right)\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|] + \left(L_f+\frac{L_g}{\delta}\right)\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|]
\end{aligned}$$

$$+\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_2 f^{(k)}(x_t^{(k)},y_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}v_{1,t}^{(k)}\|]+\frac{1}{\delta}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_2 g^{(k)}(x_t^{(k)},y_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}v_{2,t}^{(k)}\|]\,, \tag{41}$$

where (a) holds due to Assumption 3.1.

By combining the above two inequalities, we obtain

$$\begin{aligned}
&\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_{t+1},\bar{y}_{t+1})]\le\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_{t+1},\bar{y}_t)]-\eta_y\mathbb{E}[\|\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)\|]+\frac{1}{\delta}\frac{\eta_y^2 L_{h_\delta}}{2}+\frac{1}{\delta}\eta_x\eta_y L_{h_\delta}\\
&+\eta_y\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{q}_t-q_t^{(k)}\|]+2\eta_y\left(L_f+\frac{L_g}{\delta}\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t-x_t^{(k)}\|]\\
&+2\eta_y\left(L_f+\frac{L_g}{\delta}\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t-y_t^{(k)}\|]+2\eta_y\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_2 f^{(k)}(x_t^{(k)},y_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}v_{1,t}^{(k)}\|]\\
&+2\eta_y\frac{1}{\delta}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_2 g^{(k)}(x_t^{(k)},y_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}v_{2,t}^{(k)}\|]\,.
\end{aligned}\tag{42}$$

In addition, due to the smoothness of $h_\delta(x,y)$, we further obtain

$$\begin{aligned}
&\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_{t+1},\bar{y}_t)]\le\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_t,\bar{y}_t)]+\frac{1}{\delta}\mathbb{E}[\langle\nabla_1 h_\delta(\bar{x}_t,\bar{y}_t),\bar{x}_{t+1}-\bar{x}_t\rangle]+\frac{1}{\delta}\frac{L_{h_\delta}}{2}\mathbb{E}[\|\bar{x}_{t+1}-\bar{x}_t\|^2]\\
&=\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_t,\bar{y}_t)]+\frac{1}{\delta}\frac{L_{h_\delta}}{2}\mathbb{E}[\|\bar{x}_{t+1}-\bar{x}_t\|^2]\\
&\quad+\mathbb{E}[\langle\frac{1}{\delta}\nabla_1 h_\delta(\bar{x}_t,\bar{y}_t)-\frac{1}{\delta}\nabla h_\delta^*(\bar{x}_t),\bar{x}_{t+1}-\bar{x}_t\rangle]+\mathbb{E}[\langle\frac{1}{\delta}\nabla h_\delta^*(\bar{x}_t),\bar{x}_{t+1}-\bar{x}_t\rangle]\\
&=\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_t,\bar{y}_t)]+\frac{1}{\delta}\frac{\eta_x^2 L_{h_\delta}}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\frac{p_t^{(k)}}{\|p_t^{(k)}\|}\|^2]+\mathbb{E}[\langle\frac{1}{\delta}\nabla_x h_\delta(\bar{x}_t,y_\delta^*(\bar{x}_t)),\bar{x}_{t+1}-\bar{x}_t\rangle\\
&\quad-\eta_x\mathbb{E}[\langle\frac{1}{\delta}\nabla_1 h_\delta(\bar{x}_t,\bar{y}_t)-\frac{1}{\delta}\nabla_x h_\delta(\bar{x}_t,y_\delta^*(\bar{x}_t)),\frac{1}{K}\sum_{k=1}^{K}\frac{p_t^{(k)}}{\|p_t^{(k)}\|}\rangle]]\\
&\overset{(a)}{\le}\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_t,\bar{y}_t)]+\eta_x\frac{1}{\delta}\mathbb{E}[\|\nabla_1 h_\delta(\bar{x}_t,\bar{y}_t)-\nabla_1 h_\delta(\bar{x}_t,y_\delta^*(\bar{x}_t))\|]\\
&\quad+\mathbb{E}[\langle\frac{1}{\delta}\nabla_x h_\delta(\bar{x}_t,y_\delta^*(\bar{x}_t)),\bar{x}_{t+1}-\bar{x}_t\rangle]+\frac{1}{\delta}\frac{\eta_x^2 L_{h_\delta}}{2}\\
&\overset{(b)}{\le}\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_t,\bar{y}_t)]+\eta_x\frac{L_{h_\delta}}{\delta}\mathbb{E}[\|\bar{y}_t-y_\delta^*(\bar{x}_t)\|]+\mathbb{E}[\langle\frac{1}{\delta}\nabla_x h_\delta(\bar{x}_t,y_\delta^*(\bar{x}_t)),\bar{x}_{t+1}-\bar{x}_t\rangle]+\frac{1}{\delta}\frac{\eta_x^2 L_{h_\delta}}{2}\\
&\overset{(c)}{\le}\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_t,\bar{y}_t)]+\eta_x\frac{1}{\delta}\frac{L_{h_\delta}}{\mu}\mathbb{E}[\|\nabla_2 h_\delta(\bar{x}_t,\bar{y}_t)\|]+\mathbb{E}[\langle\frac{1}{\delta}\nabla_x h_\delta(\bar{x}_t,y_\delta^*(\bar{x}_t)),\bar{x}_{t+1}-\bar{x}_t\rangle]+\frac{1}{\delta}\frac{\eta_x^2 L_{h_\delta}}{2}\,,
\end{aligned}\tag{43}$$

where (a) holds due to $\|\frac{p_t^{(k)}}{\|p_t^{(k)}\|}\|=1$ and $\nabla h_\delta^*(\bar{x}_t)=\nabla_x h_\delta(\bar{x}_t,y_\delta^*(\bar{x}_t))=\nabla_1 h_\delta(\bar{x}_t,y_\delta^*(\bar{x}_t))+\nabla y_\delta^*(\bar{x}_t)\nabla_2 h_\delta(\bar{x}_t,y_\delta^*(\bar{x}_t))=\nabla_1 h_\delta(\bar{x}_t,y_\delta^*(\bar{x}_t))$, (b) holds due to Assumption 3.1, and (c) holds due to Lemma C.3.

Furthermore, due to the smoothness of $h_\delta^*(x_t)$ as shown in Lemma C.4, we obtain

$$\begin{aligned}
&\frac{1}{\delta}h_\delta^*(\bar{x}_{t+1})\ge\frac{1}{\delta}h_\delta^*(\bar{x}_t)+\frac{1}{\delta}\langle\nabla h_\delta^*(\bar{x}_t),\bar{x}_{t+1}-\bar{x}_t\rangle-\frac{1}{\delta}\frac{L_{h_\delta^*}}{2}\|\bar{x}_{t+1}-\bar{x}_t\|^2\\
&=\frac{1}{\delta}h_\delta^*(\bar{x}_t)+\frac{1}{\delta}\langle\nabla_x h_\delta(\bar{x}_t,y_\delta^*(\bar{x}_t)),\bar{x}_{t+1}-\bar{x}_t\rangle-\frac{1}{\delta}\frac{\eta_x^2 L_{h_\delta^*}}{2}\|\frac{1}{K}\sum_{k=1}^{K}\frac{p_t^{(k)}}{\|p_t^{(k)}\|}\|^2\\
&=\frac{1}{\delta}h_\delta^*(\bar{x}_t)+\frac{1}{\delta}\langle\nabla_x h_\delta(\bar{x}_t,y_\delta^*(\bar{x}_t)),\bar{x}_{t+1}-\bar{x}_t\rangle-\frac{1}{\delta}\frac{\eta_x^2 L_{h_\delta^*}}{2}\,.
\end{aligned}\tag{44}$$

Then, we obtain

$$\frac{1}{\delta}h^*_\delta(\bar{x}_t) - \frac{1}{\delta}h^*_\delta(\bar{x}_{t+1}) \leq -\frac{1}{\delta}\langle \nabla_x h_\delta(\bar{x}_t, y^*_\delta(\bar{x}_t)), \bar{x}_{t+1} - \bar{x}_t\rangle + \frac{1}{\delta}\frac{\eta_x^2 L_{h^*_\delta}}{2} \,. \tag{45}$$

Finally, we obtain

$$\begin{aligned}
&\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_{t+1}, \bar{y}_{t+1})] - \frac{1}{\delta}\mathbb{E}[h^*_\delta(\bar{x}_{t+1})] \\
&= \frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_{t+1}, \bar{y}_{t+1})] - \frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_{t+1}, \bar{y}_t)] + \frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_{t+1}, \bar{y}_t)] - \frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_t, \bar{y}_t)] \\
&\quad + \frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_t, \bar{y}_t)] - \frac{1}{\delta}\mathbb{E}[h^*_\delta(\bar{x}_t)] + \frac{1}{\delta}\mathbb{E}[h^*_\delta(\bar{x}_t)] - \frac{1}{\delta}\mathbb{E}[h^*_\delta(\bar{x}_{t+1})] \\
&\leq \frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_t, \bar{y}_t)] - \frac{1}{\delta}\mathbb{E}[h^*_\delta(\bar{x}_t)] - \eta_y\mathbb{E}[\|\frac{1}{\delta}\nabla_2 h_\delta(\bar{x}_t, \bar{y}_t)\|] + \frac{1}{\delta}\frac{\eta_y^2 L_{h_\delta}}{2} + \frac{1}{\delta}\eta_x\eta_y L_{h_\delta} \\
&\quad + \eta_y\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{q}_t - q_t^{(k)}\|] + 2\eta_y\left(L_f + \frac{L_g}{\delta}\right)\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|] \\
&\quad + 2\eta_y\left(L_f + \frac{L_g}{\delta}\right)\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|] + 2\eta_y\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_2 f^{(k)}(x_t^{(k)}, y_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K v_{1,t}^{(k)}\|] \\
&\quad + 2\eta_y\frac{1}{\delta}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_2 g^{(k)}(x_t^{(k)}, y_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K v_{2,t}^{(k)}\|] \\
&\quad + \eta_x\frac{1}{\delta}\frac{L_{h_\delta}}{\mu}\mathbb{E}[\|\nabla_2 h_\delta(\bar{x}_t, \bar{y}_t)\|] + \mathbb{E}[\langle\frac{1}{\delta}\nabla_x h_\delta(\bar{x}_t, y^*_\delta(\bar{x}_t)), \bar{x}_{t+1} - \bar{x}_t\rangle] + \frac{1}{\delta}\frac{\eta_x^2 L_{h_\delta}}{2} \\
&\quad - \frac{1}{\delta}\mathbb{E}[\langle\nabla_x h_\delta(\bar{x}_t, y^*_\delta(\bar{x}_t)), \bar{x}_{t+1} - \bar{x}_t\rangle] + \frac{1}{\delta}\frac{\eta_x^2 L_{h^*_\delta}}{2} \\
&= \frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_t, \bar{y}_t)] - \frac{1}{\delta}\mathbb{E}[h^*_\delta(\bar{x}_t)] + \left(\eta_x\frac{L_{h_\delta}}{\mu} - \eta_y\right)\frac{1}{\delta}\mathbb{E}[\|\nabla_2 h_\delta(\bar{x}_t, \bar{y}_t)\|] + \eta_y\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{q}_t - q_t^{(k)}\|] \\
&\quad + 2\eta_y\left(L_f + \frac{L_g}{\delta}\right)\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|] + 2\eta_y\left(L_f + \frac{L_g}{\delta}\right)\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|] \\
&\quad + 2\eta_y\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_2 f^{(k)}(x_t^{(k)}, y_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K v_{1,t}^{(k)}\|] + \frac{1}{\delta}\frac{\eta_x^2 L_{h_\delta}}{2} + \frac{1}{\delta}\frac{\eta_x^2 L_{h^*_\delta}}{2} \\
&\quad + 2\eta_y\frac{1}{\delta}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_2 g^{(k)}(x_t^{(k)}, y_t^{(k)}) - \frac{1}{K}\sum_{k=1}^K v_{2,t}^{(k)}\|] + \frac{1}{\delta}\frac{\eta_y^2 L_{h_\delta}}{2} + \frac{1}{\delta}\eta_x\eta_y L_{h_\delta} \,.
\end{aligned} \tag{46}$$

By setting $\eta_x \leq \eta_y\frac{\mu}{2L_{h_\delta}}$, we complete the proof.

□

C.3 Bounding Consecutive Updates

Lemma C.9. *Given Assumptions 3.1-3.3, we obtain*

$$\begin{aligned}
&\sum_{k=1}^K \|x_{t+1}^{(k)} - x_t^{(k)}\| \leq \frac{4\eta_x}{1-\lambda}K \,; \quad \sum_{k=1}^K \|y_{t+1}^{(k)} - y_t^{(k)}\| \leq \frac{4\eta_y}{1-\lambda}K \,; \\
&\sum_{k=1}^K \|z_{t+1}^{(k)} - z_t^{(k)}\| \leq \frac{4\eta_z}{1-\lambda}K \,.
\end{aligned} \tag{47}$$

Proof.

$$\|X_{t+1} - X_t\|_F^2 = \|(X_t - \eta_x\hat{P}_t)E - X_t\|_F^2$$

$$
\begin{aligned}
&\leq 2\|X_t E - X_t\|_F^2 + 2\eta_x^2 \|\hat{P}_t E\|_F^2 \\
&= 2\|(X_t - \bar{X}_t)(E - I)\|_F^2 + 2\eta_x^2 \|\hat{P}_t E\|_F^2 \\
&\leq 2\|X_t - \bar{X}_t\|_F^2 \|E - I\|_2^2 + 2\eta_x^2 \|\hat{P}_t\|_F^2 \|E\|_2^2 \\
&\overset{(a)}{\leq} 8\|X_t - \bar{X}_t\|_F^2 + 2\eta_x^2 \|\hat{P}_t\|_F^2 \\
&\leq 8\|X_t - \bar{X}_t\|_F^2 + 2\eta_x^2 \sum_{k=1}^{K} \left\| \frac{p_t^{(k)}}{\|p_t^{(k)}\|} \right\|^2 \\
&\leq 8\|X_t - \bar{X}_t\|_F^2 + 2\eta_x^2 K \\
&\overset{(b)}{\leq} \frac{8\eta_x^2 \lambda^2}{(1-\lambda)^2} K + 2\eta_x^2 K \overset{(c)}{\leq} \frac{10\eta_x^2}{(1-\lambda)^2} K \,,
\end{aligned} \tag{48}
$$

where (a) holds due to $\|E - I\|_2 \leq 2$ and $\|E\|_2 \leq 1$, (b) holds due to Lemma C.28, and (c) holds due to $\lambda < 1$.

Then, we obtain

$$
\begin{aligned}
&\sum_{k=1}^{K} \|x_{t+1}^{(k)} - x_t^{(k)}\| = \sqrt{\left(\sum_{k=1}^{K} \|x_{t+1}^{(k)} - x_t^{(k)}\| \right)^2} \\
&\leq \sqrt{K \sum_{k=1}^{K} \|x_{t+1}^{(k)} - x_t^{(k)}\|^2} = \sqrt{K} \sqrt{\|X_{t+1} - X_t\|_F^2} \leq \frac{4\eta_x}{1-\lambda} K \,.
\end{aligned} \tag{49}
$$

The other two inequalities can be proved in a same approach. □

Lemma C.10. *Given Assumptions 3.1-3.3, for $t > 0$, we obtain*

$$
\begin{aligned}
\sum_{k=1}^{K} \mathbb{E}[\|u_{1,t}^{(k)} - u_{1,t-1}^{(k)}\|] &\leq \gamma_x \sum_{k=1}^{K} \mathbb{E}[\|\nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}) - \nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}; \xi_t^{(k)})\|] \\
&+ \gamma_x \sum_{k=1}^{K} \mathbb{E}[\|u_{1,t-1}^{(k)} - \nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)})\|] + \frac{4\eta_x L_f}{1-\lambda} K + \frac{4\eta_y L_f}{1-\lambda} K \,.
\end{aligned} \tag{50}
$$

Proof.

$$
\begin{aligned}
&\sum_{k=1}^{K} \|u_{1,t}^{(k)} - u_{1,t-1}^{(k)}\| \\
&= \sum_{k=1}^{K} \|(1-\gamma_x)(u_{1,t-1}^{(k)} - \nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}; \xi_t^{(k)})) + \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_t^{(k)}) - u_{1,t-1}^{(k)}\| \\
&\leq \sum_{k=1}^{K} \|\nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_t^{(k)}) - \nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}; \xi_t^{(k)})\| + \gamma_x \sum_{k=1}^{K} \|u_{1,t-1}^{(k)} - \nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)})\| \\
&\quad + \gamma_x \sum_{k=1}^{K} \|\nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}) - \nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}; \xi_t^{(k)})\| \\
&\leq L_f \sum_{k=1}^{K} \|x_t^{(k)} - x_{t-1}^{(k)}\| + L_f \sum_{k=1}^{K} \|y_t^{(k)} - y_{t-1}^{(k)}\| + \gamma_x \sum_{k=1}^{K} \|u_{1,t-1}^{(k)} - \nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)})\| \\
&\quad + \gamma_x \sum_{k=1}^{K} \|\nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}) - \nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}; \xi_t^{(k)})\| \\
&\overset{(a)}{\leq} \frac{4\eta_x L_f}{1-\lambda} K + \frac{4\eta_y L_f}{1-\lambda} K + \gamma_x \sum_{k=1}^{K} \|u_{1,t-1}^{(k)} - \nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)})\|
\end{aligned}
$$

$$+ \gamma_x \sum_{k=1}^{K} \|\nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}) - \nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}; \xi_t^{(k)})\| \,, \tag{51}$$

where (a) holds due to Lemma C.9.

□

Lemma C.11. *Given Assumptions 3.1-3.3, for $t > 0$, we obtain*

$$\sum_{k=1}^{K} \mathbb{E}[\|u_{2,t}^{(k)} - u_{2,t-1}^{(k)}\|] \leq \gamma_x \sum_{k=1}^{K} \mathbb{E}[\|\nabla_1 g^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}) - \nabla_1 g^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}; \zeta_t^{(k)})\|] + \gamma_x \sum_{k=1}^{K} \mathbb{E}[\|u_{2,t-1}^{(k)} - \nabla_1 g^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)})\|] + \frac{4\eta_x L_g}{1-\lambda} K + \frac{4\eta_y L_g}{1-\lambda} K \,. \tag{52}$$

Lemma C.12. *Given Assumptions 3.1-3.3, for $t > 0$, we obtain*

$$\sum_{k=1}^{K} \mathbb{E}[\|u_{3,t}^{(k)} - u_{3,t-1}^{(k)}\|] \leq \gamma_x \sum_{k=1}^{K} \mathbb{E}[\|\nabla_1 g^{(k)}(x_{t-1}^{(k)}, z_{t-1}^{(k)}) - \nabla_1 g^{(k)}(x_{t-1}^{(k)}, z_{t-1}^{(k)}; \zeta_t^{(k)})\|] + \gamma_x \sum_{k=1}^{K} \mathbb{E}[\|u_{3,t-1}^{(k)} - \nabla_1 g^{(k)}(x_{t-1}^{(k)}, z_{t-1}^{(k)})\|] + \frac{4\eta_x L_g}{1-\lambda} K + \frac{4\eta_z L_g}{1-\lambda} K \,. \tag{53}$$

Lemma C.13. *Given Assumptions 3.1-3.3, for $t > 0$, we obtain*

$$\sum_{k=1}^{K} \mathbb{E}[\|v_{1,t}^{(k)} - v_{1,t-1}^{(k)}\|] \leq \gamma_y \sum_{k=1}^{K} \mathbb{E}[\|\nabla_2 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}) - \nabla_2 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}; \xi_t^{(k)})\|] + \gamma_y \sum_{k=1}^{K} \mathbb{E}[\|v_{1,t-1}^{(k)} - \nabla_2 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)})\|] + \frac{4\eta_x L_f}{1-\lambda} K + \frac{4\eta_y L_f}{1-\lambda} K \,. \tag{54}$$

Lemma C.14. *Given Assumptions 3.1-3.3, for $t > 0$, we obtain*

$$\sum_{k=1}^{K} \mathbb{E}[\|v_{2,t}^{(k)} - v_{2,t-1}^{(k)}\|] \leq \gamma_y \sum_{k=1}^{K} \mathbb{E}[\|\nabla_2 g^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}) - \nabla_2 g^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}; \zeta_t^{(k)})\|] + \gamma_y \sum_{k=1}^{K} \mathbb{E}[\|v_{2,t-1}^{(k)} - \nabla_2 g^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)})\|] + \frac{4\eta_x L_g}{1-\lambda} K + \frac{4\eta_y L_g}{1-\lambda} K \,. \tag{55}$$

Lemma C.15. *Given Assumptions 3.1-3.3, for $t > 0$, we obtain*

$$\sum_{k=1}^{K} \mathbb{E}[\|w_{1,t}^{(k)} - w_{1,t-1}^{(k)}\|] \leq \gamma_z \sum_{k=1}^{K} \mathbb{E}[\|\nabla_2 g^{(k)}(x_{t-1}^{(k)}, z_{t-1}^{(k)}) - \nabla_2 g^{(k)}(x_{t-1}^{(k)}, z_{t-1}^{(k)}; \zeta_t^{(k)})\|] + \gamma_z \sum_{k=1}^{K} \mathbb{E}[\|w_{1,t-1}^{(k)} - \nabla_2 g^{(k)}(x_{t-1}^{(k)}, z_{t-1}^{(k)})\|] + \frac{4\eta_x L_g}{1-\lambda} K + \frac{4\eta_z L_g}{1-\lambda} K \,. \tag{56}$$

Lemmas C.11 - C.15 can be easily proved by following Lemma C.10.

C.4 Bounding Gradient Errors

Lemma C.16. *Given Assumptions 3.1-3.3, we obtain*

$$\sum_{k=1}^{K} \mathbb{E}[\|u_{1,t}^{(k)} - \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)})\|] \leq (1-\gamma_x)^t \frac{2\sqrt{2}\sigma K}{B_0^{1-1/s}} + \frac{8(\eta_x + \eta_y) L_f}{(1-\lambda)\sqrt{\gamma_x}} K + 2\sqrt{2}\gamma_x^{1-1/s} \sigma K \,. \tag{57}$$

Proof. When $t > 0$, based on Algorithm 1, we obtain

$$
\begin{aligned}
&u_{1,t}^{(k)} - \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}) \\
&= (1-\gamma_x)(u_{1,t-1}^{(k)} - \nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)})) + (1-\gamma_x)\Big(\nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_t^{(k)}) \\
&\qquad\qquad - \nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}; \xi_t^{(k)}) - \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}) + \nabla_1 f^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)})\Big) \\
&\quad + \gamma_x(\nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_t^{(k)}) - \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)})) \\
&= (1-\gamma_x)^t(u_{1,0}^{(k)} - \nabla_1 f^{(k)}(x_0^{(k)}, y_0^{(k)})) \\
&\quad + \sum_{j=1}^{t}(1-\gamma_x)^{t-j+1}\Big(\nabla_1 f^{(k)}(x_j^{(k)}, y_j^{(k)}; \xi_j^{(k)}) - \nabla_1 f^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)}; \xi_j^{(k)}) \\
&\qquad\qquad + \nabla_1 f^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)}) - \nabla_1 f^{(k)}(x_j^{(k)}, y_j^{(k)})\Big) \\
&\quad + \sum_{j=1}^{t}\gamma_x(1-\gamma_x)^{t-j}(\nabla_1 f^{(k)}(x_j^{(k)}, y_j^{(k)}; \xi_j^{(k)}) - \nabla_1 f^{(k)}(x_j^{(k)}, y_j^{(k)}))\,.
\end{aligned} \tag{58}
$$

Then, we obtain

$$
\begin{aligned}
\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t}^{(k)} - \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)})\|] &\le (1-\gamma_x)^t \sum_{k=1}^{K}\mathbb{E}[\|u_{1,0}^{(k)} - \nabla_1 f^{(k)}(x_0^{(k)}, y_0^{(k)})\|] \\
&\quad + \sum_{k=1}^{K}\mathbb{E}[\|\sum_{j=1}^{t}(1-\gamma_x)^{t-j+1}\Big(\nabla_1 f^{(k)}(x_j^{(k)}, y_j^{(k)}; \xi_j^{(k)}) - \nabla_1 f^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)}; \xi_j^{(k)}) \\
&\qquad\qquad + \nabla_1 f^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)}) - \nabla_1 f^{(k)}(x_j^{(k)}, y_j^{(k)})\Big)\|] \\
&\quad + \sum_{k=1}^{K}\mathbb{E}[\|\gamma_x\sum_{j=1}^{t}(1-\gamma_x)^{t-j}(\nabla_1 f^{(k)}(x_j^{(k)}, y_j^{(k)}; \xi_j^{(k)}) - \nabla_1 f^{(k)}(x_j^{(k)}, y_j^{(k)}))\|]\,.
\end{aligned} \tag{59}
$$

For the first term on the right-hand side of Eq. (59), we bound it as follows:

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}[\|u_{1,0}^{(k)} - \nabla_1 f^{(k)}(x_0^{(k)}, y_0^{(k)})\|] \\
&= \sum_{k=1}^{K}\mathbb{E}[\|\frac{1}{B_0}\sum_{b=1}^{B_0}\nabla_1 f^{(k)}(x_0^{(k)}, y_0^{(k)}; \xi_{0,b}^{(k)}) - \nabla_1 f^{(k)}(x_0^{(k)}, y_0^{(k)})\|] \\
&= \sum_{k=1}^{K}\frac{1}{B_0}\mathbb{E}[\|\sum_{b=1}^{B_0}(\nabla_1 f^{(k)}(x_0^{(k)}, y_0^{(k)}; \xi_{0,b}^{(k)}) - \nabla_1 f^{(k)}(x_0^{(k)}, y_0^{(k)}))\|] \\
&\overset{(a)}{\le} \sum_{k=1}^{K}\frac{2\sqrt{2}}{B_0}\mathbb{E}[(\sum_{b=1}^{B_0}\|\nabla_1 f^{(k)}(x_0^{(k)}, y_0^{(k)}; \xi_0^{(k)}) - \nabla_1 f^{(k)}(x_0^{(k)}, y_0^{(k)})\|^s)^{\frac{1}{s}}] \\
&\overset{(b)}{\le} \sum_{k=1}^{K}\frac{2\sqrt{2}}{B_0}(\sum_{b=1}^{B_0}\mathbb{E}[\|\nabla_1 f^{(k)}(x_0^{(k)}, y_0^{(k)}; \xi_0^{(k)}) - \nabla_1 f^{(k)}(x_0^{(k)}, y_0^{(k)})\|^s])^{\frac{1}{s}} \\
&\overset{(c)}{\le} \frac{2\sqrt{2}K}{B_0^{1-1/s}}\sigma\,,
\end{aligned} \tag{60}
$$

where B_0 represents the batch size in the initial iteration, (a) holds due to Lemma C.5, (b) holds due to Jensen's inequality, and (c) holds due to Assumption 3.2.

Similar to the proof of Theorem 2 in [50], to bound the second term on the right-hand side of Eq. (59), we first bound the following one:

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}[\|\sum_{j=1}^{t}(1-\gamma_x)^{t-j+1}\Big(\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)};\xi_j^{(k)})\\
&\qquad+\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})-\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)})\Big)\|^2]\\
&=\sum_{k=1}^{K}\sum_{j=1}^{t}(1-\gamma_x)^{2(t-j+1)}\mathbb{E}[\|\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)};\xi_j^{(k)})\\
&\qquad+\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})-\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)})\|^2]\\
&\le 2\sum_{k=1}^{K}\sum_{j=1}^{t}(1-\gamma_x)^{2(t-j+1)}\mathbb{E}[\|\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)};\xi_j^{(k)})\|^2]\\
&\quad+2\sum_{k=1}^{K}\sum_{j=1}^{t}(1-\gamma_x)^{2(t-j+1)}\mathbb{E}[\|\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)})-\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})\|^2]\\
&\le 4\sum_{k=1}^{K}\sum_{j=1}^{t}(1-\gamma_x)^{2(t-j+1)}L_f^2(\mathbb{E}[\|x_j^{(k)}-x_{j-1}^{(k)}\|^2]+\mathbb{E}[\|y_j^{(k)}-y_{j-1}^{(k)}\|^2])\\
&=4\sum_{j=1}^{t}(1-\gamma_x)^{2(t-j+1)}L_f^2(\mathbb{E}[\|X_j-X_{j-1}\|_F^2]+\mathbb{E}[\|Y_j-Y_{j-1}\|_F^2])\\
&\overset{(a)}{\le} 4\sum_{j=1}^{t}(1-\gamma_x)^{2(t-j+1)}L_f^2\left(\frac{10\eta_x^2}{(1-\lambda)^2}K+\frac{10\eta_y^2}{(1-\lambda)^2}K\right)\\
&\le \frac{1}{1-(1-\gamma_x)^2}L_f^2\left(\frac{40\eta_x^2}{(1-\lambda)^2}K+\frac{40\eta_y^2}{(1-\lambda)^2}K\right)\\
&\overset{(b)}{\le} \frac{40\eta_x^2}{(1-\lambda)^2}\frac{L_f^2}{\gamma_x}K+\frac{40\eta_y^2}{(1-\lambda)^2}\frac{L_f^2}{\gamma_x}K\,,
\end{aligned}
\tag{61}
$$

where (a) holds due to Eq. (48), and (b) holds due to $\gamma_x<1$.

Then, similar to Eq. (49), we bound the second term on the right-hand side of Eq. (59) as follows:

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}[\|\sum_{j=1}^{t}(1-\gamma_x)^{t-j+1}\Big(\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)};\xi_j^{(k)})\\
&\qquad+\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})-\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)})\Big)\|]\\
&\le\sqrt{K}\sqrt{\frac{40\eta_x^2}{(1-\lambda)^2}\frac{L_f^2}{\gamma_x}K+\frac{40\eta_y^2}{(1-\lambda)^2}\frac{L_f^2}{\gamma_x}K}\\
&\le\frac{8\eta_x L_f}{(1-\lambda)\sqrt{\gamma_x}}K+\frac{8\eta_y L_f}{(1-\lambda)\sqrt{\gamma_x}}K\,.
\end{aligned}
\tag{62}
$$

For the third term on the right-hand side of Eq. (59), we bound it as follows:

$$
\begin{aligned}
&\mathbb{E}[\|\sum_{j=1}^{t}\gamma_x(1-\gamma_x)^{t-j}(\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)}))\|]\\
&\overset{(a)}{\le} 2\sqrt{2}\mathbb{E}\left[\left(\sum_{j=1}^{t}\|\gamma_x(1-\gamma_x)^{t-j}(\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)}))\|^s\right)^{1/s}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= 2\sqrt{2}\mathbb{E}\left[\left(\sum_{j=1}^{t}\gamma_x^s(1-\gamma_x)^{s(t-j)}\|\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)})\|^s\right)^{1/s}\right]\\
&\overset{(b)}{\leq} 2\sqrt{2}\left(\mathbb{E}\left[\sum_{j=1}^{t}\gamma_x^s(1-\gamma_x)^{s(t-j)}\|\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)})\|^s\right]\right)^{1/s}\\
&\overset{(c)}{\leq} 2\sqrt{2}\left(\sum_{j=1}^{t}\gamma_x^s(1-\gamma_x)^{s(t-j)}\right)^{1/s}\sigma\,, \qquad (63)
\end{aligned}
$$

where (a) holds due to Lemma C.5, (b) holds due to Jensen's inequality, and (c) holds due to Assumption 3.2.

Finally, when $t>0$, from

$$
\left(\sum_{j=1}^{t}(1-\gamma_x)^{s(t-j)}\right)^{1/s}\leq\left(\frac{1}{1-(1-\gamma_x)^s}\right)^{1/s}\leq\left(\frac{1}{1-(1-\gamma_x)}\right)^{1/s}\leq\gamma_x^{-1/s}\,, \qquad (64)
$$

we obtain

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t}^{(k)}-\nabla_1 f^{(k)}(x_t^{(k)},y_t^{(k)})\|]\\
&\leq(1-\gamma_x)^t\frac{2\sqrt{2}K}{B_0^{1-1/s}}\sigma+\frac{8\eta_x L_f}{(1-\lambda)\sqrt{\gamma_x}}K+\frac{8\eta_y L_f}{(1-\lambda)\sqrt{\gamma_x}}K+2\sqrt{2}\gamma_x^{1-1/s}\sigma K\,. \qquad (65)
\end{aligned}
$$

Based on Eq. (60), it is easy to know that this upper bound also holds when $t=0$. □

Lemma C.17. *Given Assumptions 3.1-3.3, we obtain*

$$
\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t}^{(k)}-\nabla_1 g^{(k)}(x_t^{(k)},y_t^{(k)})\|]\leq(1-\gamma_x)^t\frac{2\sqrt{2}\sigma K}{B_0^{1-1/s}}+\frac{8(\eta_x+\eta_y)L_g}{(1-\lambda)\sqrt{\gamma_x}}K+2\sqrt{2}\gamma_x^{1-1/s}\sigma K\,. \qquad (66)
$$

Lemma C.18. *Given Assumptions 3.1-3.3, we obtain*

$$
\sum_{k=1}^{K}\mathbb{E}[\|u_{3,t}^{(k)}-\nabla_1 g^{(k)}(x_t^{(k)},z_t^{(k)})\|]\leq(1-\gamma_x)^t\frac{2\sqrt{2}\sigma K}{B_0^{1-1/s}}x+\frac{8(\eta_x+\eta_z)L_g}{(1-\lambda)\sqrt{\gamma_x}}K+2\sqrt{2}\gamma_x^{1-1/s}\sigma K\,. \qquad (67)
$$

Lemma C.19. *Given Assumptions 3.1-3.3, we obtain*

$$
\sum_{k=1}^{K}\mathbb{E}[\|v_{1,t}^{(k)}-\nabla_2 f^{(k)}(x_t^{(k)},y_t^{(k)})\|]\leq(1-\gamma_y)^t\frac{2\sqrt{2}\sigma K}{B_0^{1-1/s}}+\frac{8(\eta_x+\eta_y)L_f}{(1-\lambda)\sqrt{\gamma_y}}K+2\sqrt{2}\gamma_y^{1-1/s}\sigma K\,. \qquad (68)
$$

Lemma C.20. *Given Assumptions 3.1-3.3, we obtain*

$$
\sum_{k=1}^{K}\mathbb{E}[\|v_{2,t}^{(k)}-\nabla_2 g^{(k)}(x_t^{(k)},y_t^{(k)})\|]\leq(1-\gamma_y)^t\frac{2\sqrt{2}\sigma K}{B_0^{1-1/s}}+\frac{8(\eta_x+\eta_y)L_g}{(1-\lambda)\sqrt{\gamma_y}}K+2\sqrt{2}\gamma_y^{1-1/s}\sigma K\,. \qquad (69)
$$

Lemma C.21. *Given Assumptions 3.1-3.3, we obtain*

$$
\sum_{k=1}^{K}\mathbb{E}[\|w_{1,t}^{(k)}-\nabla_2 g^{(k)}(x_t^{(k)},z_t^{(k)})\|]\leq(1-\gamma_z)^t\frac{2\sqrt{2}\sigma K}{B_0^{1-1/s}}+\frac{8(\eta_x+\eta_z)L_g}{(1-\lambda)\sqrt{\gamma_z}}K+2\sqrt{2}\gamma_z^{1-1/s}\sigma K\,. \qquad (70)
$$

Lemmas C.17 - C.21 can be easily proved by following Lemma C.16.

Lemma C.22. *Given Assumptions 3.1-3.3, we obtain*

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}u_{1,t}^{(k)}-\frac{1}{K}\sum_{k=1}^{K}\nabla_1 f^{(k)}(x_t^{(k)},y_t^{(k)})\|]\\ \leq&\frac{1}{\gamma_x T}\frac{2\sqrt{2}\sigma}{B_0^{1-1/s}}+\frac{8(\eta_x+\eta_y)L_f}{(1-\lambda)\sqrt{\gamma_x}}\frac{1}{\sqrt{K}}+\frac{2\sqrt{2}\gamma_x^{1-1/s}\sigma}{K^{1-1/s}}\,.\end{aligned}\tag{71}$$

Proof. When $t>0$, same as the proof of Lemma C.16, we obtain

$$\begin{aligned}&\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}(u_{1,t}^{(k)}-\nabla_1 f^{(k)}(x_t^{(k)},y_t^{(k)}))\|]\\ \leq&(1-\gamma_x)^t\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}(u_{1,0}^{(k)}-\nabla_1 f^{(k)}(x_0^{(k)},y_0^{(k)}))\|]\\ &+\mathbb{E}[\|\sum_{j=1}^{t}(1-\gamma_x)^{t-j+1}\frac{1}{K}\sum_{k=1}^{K}\Big(\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)};\xi_j^{(k)})\\ &\quad+\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})-\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)})\Big)\|]\\ &+\mathbb{E}[\|\gamma_x\sum_{j=1}^{t}(1-\gamma_x)^{t-j}\frac{1}{K}\sum_{k=1}^{K}(\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)}))\|]\,.\end{aligned}\tag{72}$$

Then, for the first term on the right-hand side of Eq. (72), we obtain

$$\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}(u_{1,0}^{(k)}-\nabla_1 f^{(k)}(x_0^{(k)},y_0^{(k)}))\|]\leq\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,0}^{(k)}-\nabla_1 f^{(k)}(x_0^{(k)},y_0^{(k)})\|]\overset{(a)}{\leq}\frac{2\sqrt{2}}{B_0^{1-1/s}}\sigma\,,$$

where (a) holds due to Eq. (60).

Similar to the proof of Theorem 2 in [50], to bound the second term on the right-hand side of Eq. (72), we first bound the following one:

$$\begin{aligned}&\mathbb{E}[\|\sum_{j=1}^{t}(1-\gamma_x)^{t-j+1}\frac{1}{K}\sum_{k=1}^{K}(\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)};\xi_j^{(k)})\\ &\quad+\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})-\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)}))\|^2]\\ =&\frac{1}{K^2}\sum_{k=1}^{K}\sum_{j=1}^{t}(1-\gamma_x)^{2(t-j+1)}\mathbb{E}[\|\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)};\xi_j^{(k)})\\ &\quad+\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})-\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)})\|^2]\\ \leq&2\frac{1}{K^2}\sum_{k=1}^{K}\sum_{j=1}^{t}(1-\gamma_x)^{2(t-j+1)}\mathbb{E}[\|\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)};\xi_j^{(k)})\|^2]\\ &+2\frac{1}{K^2}\sum_{k=1}^{K}\sum_{j=1}^{t}(1-\gamma_x)^{2(t-j+1)}\mathbb{E}[\|\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)})-\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})\|^2]\\ \leq&4\frac{1}{K^2}\sum_{k=1}^{K}\sum_{j=1}^{t}(1-\gamma_x)^{2(t-j+1)}L_f^2(\mathbb{E}[\|x_j^{(k)}-x_{j-1}^{(k)}\|^2]+\mathbb{E}[\|y_j^{(k)}-y_{j-1}^{(k)}\|^2])\\ =&4\frac{1}{K^2}\sum_{j=1}^{t}(1-\gamma_x)^{2(t-j+1)}L_f^2(\mathbb{E}[\|X_j-X_{j-1}\|_F^2]+\mathbb{E}[\|Y_j-Y_{j-1}\|_F^2])\end{aligned}$$

$$\overset{(a)}{\leq} 4\frac{1}{K^2}\sum_{j=1}^{t}(1-\gamma_x)^{2(t-j+1)}L_f^2\left(\frac{10\eta_x^2}{(1-\lambda)^2}K+\frac{10\eta_y^2}{(1-\lambda)^2}K\right)$$
$$\leq \frac{4}{1-(1-\gamma_x)^2}L_f^2\left(\frac{10\eta_x^2}{(1-\lambda)^2}\frac{1}{K}+\frac{10\eta_y^2}{(1-\lambda)^2}\frac{1}{K}\right)$$
$$\overset{(b)}{\leq} \frac{40\eta_x^2}{(1-\lambda)^2}\frac{L_f^2}{\gamma_x}\frac{1}{K}+\frac{40\eta_y^2}{(1-\lambda)^2}\frac{L_f^2}{\gamma_x}\frac{1}{K}\,, \tag{73}$$

where (a) holds due to Eq. (48), and (b) holds due to $\gamma_x < 1$.

Then, we bound the second term on the right-hand side of Eq. (59) as follows:

$$\mathbb{E}[\|\sum_{j=1}^{t}(1-\gamma_x)^{t-j+1}\frac{1}{K}\sum_{k=1}^{K}\Big(\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)};\xi_j^{(k)})$$
$$+\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})-\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)})\Big)\|]$$
$$\leq \frac{8\eta_x L_f}{(1-\lambda)\sqrt{\gamma_x}}\frac{1}{\sqrt{K}}+\frac{8\eta_y L_f}{(1-\lambda)\sqrt{\gamma_x}}\frac{1}{\sqrt{K}}\,. \tag{74}$$

For the third term on the right-hand side of Eq. (72), we bound it as follows:

$$\mathbb{E}[\|\sum_{j=1}^{t}\gamma_x(1-\gamma_x)^{t-j}\frac{1}{K}\sum_{k=1}^{K}(\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)}))\|]$$
$$=\frac{1}{K}\mathbb{E}[\|\sum_{j=1}^{t}\gamma_x(1-\gamma_x)^{t-j}\sum_{k=1}^{K}(\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)}))\|]$$
$$\overset{(a)}{\leq} \frac{2\sqrt{2}}{K}\mathbb{E}\left[\left(\sum_{k=1}^{K}\sum_{j=1}^{t}\|\gamma_x(1-\gamma_x)^{t-j}(\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)}))\|^s\right)^{1/s}\right]$$
$$=\frac{2\sqrt{2}}{K}\mathbb{E}\left[\left(\sum_{k=1}^{K}\sum_{j=1}^{t}\gamma_x^s(1-\gamma_x)^{s(t-j)}\|\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)})\|^s\right)^{1/s}\right]$$
$$\overset{(b)}{\leq} \frac{2\sqrt{2}}{K}\left(\mathbb{E}\left[\sum_{k=1}^{K}\sum_{j=1}^{t}\gamma_x^s(1-\gamma_x)^{s(t-j)}\|\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)};\xi_j^{(k)})-\nabla_1 f^{(k)}(x_j^{(k)},y_j^{(k)})\|^s\right]\right)^{1/s}$$
$$\overset{(c)}{\leq} \frac{2\sqrt{2}}{K}\left(\sum_{k=1}^{K}\sum_{j=1}^{t}\gamma_x^s(1-\gamma_x)^{s(t-j)}\right)^{1/s}\sigma\,, \tag{75}$$

where (a) holds due to Lemma C.5, (b) holds due to Jensen's inequality, and (c) holds due to Assumption 3.2.

Finally, when $t > 0$, from

$$\left(\sum_{k=1}^{K}\sum_{j=1}^{t}\gamma_x^s(1-\gamma_x)^{s(t-j)}\right)^{1/s}\leq\left(\frac{K\gamma_x^s}{1-(1-\gamma_x)^s}\right)^{1/s}\leq\left(\frac{K\gamma_x^s}{1-(1-\gamma_x)}\right)^{1/s}\leq\frac{\gamma_x^{1-1/s}}{K^{-1/s}}\,,$$

we obtain

$$\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}u_{1,t}^{(k)}-\frac{1}{K}\sum_{k=1}^{K}\nabla_1 f^{(k)}(x_t^{(k)},y_t^{(k)})\|]$$
$$\leq(1-\gamma_x)^t\frac{2\sqrt{2}}{B_0^{1-1/s}}\sigma+\frac{8\eta_x L_f}{(1-\lambda)\sqrt{\gamma_x}}\frac{1}{\sqrt{K}}+\frac{8\eta_y L_f}{(1-\lambda)\sqrt{\gamma_x}}\frac{1}{\sqrt{K}}+\frac{2\sqrt{2}}{K^{1-1/s}}\gamma_x^{1-1/s}\sigma\,. \tag{76}$$

Similarly, it is easy to know that this upper bound also holds when $t = 0$. Then, we obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}u_{1,t}^{(k)}-\frac{1}{K}\sum_{k=1}^{K}\nabla_1 f^{(k)}(x_t^{(k)},y_t^{(k)})\|] \\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}(1-\gamma_x)^t\frac{2\sqrt{2}}{B_0^{1-1/s}}\sigma+\frac{8\eta_x L_f}{(1-\lambda)\sqrt{\gamma_x}}\frac{1}{\sqrt{K}}+\frac{8\eta_y L_f}{(1-\lambda)\sqrt{\gamma_x}}\frac{1}{\sqrt{K}}+\frac{2\sqrt{2}}{K^{1-1/s}}\gamma_x^{1-1/s}\sigma \\
&\leq \frac{1}{\gamma_x T}\frac{2\sqrt{2}}{B_0^{1-1/s}}\sigma+\frac{8\eta_x L_f}{(1-\lambda)\sqrt{\gamma_x}}\frac{1}{\sqrt{K}}+\frac{8\eta_y L_f}{(1-\lambda)\sqrt{\gamma_x}}\frac{1}{\sqrt{K}}+\frac{2\sqrt{2}}{K^{1-1/s}}\gamma_x^{1-1/s}\sigma\,. \qquad (77)
\end{aligned}$$

□

Lemma C.23. *Given Assumptions 3.1-3.3, we obtain*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}u_{2,t}^{(k)}-\frac{1}{K}\sum_{k=1}^{K}\nabla_1 g^{(k)}(x_t^{(k)},y_t^{(k)})\|] \\
&\leq \frac{1}{\gamma_x T}\frac{2\sqrt{2}\sigma}{B_0^{1-1/s}}+\frac{8(\eta_x+\eta_y)L_g}{(1-\lambda)\sqrt{\gamma_x}}\frac{1}{\sqrt{K}}+\frac{2\sqrt{2}\gamma_x^{1-1/s}\sigma}{K^{1-1/s}}\,. \qquad (78)
\end{aligned}$$

Lemma C.24. *Given Assumptions 3.1-3.3, we obtain*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}u_{3,t}^{(k)}-\frac{1}{K}\sum_{k=1}^{K}\nabla_1 g^{(k)}(x_t^{(k)},z_t^{(k)})\|] \\
&\leq \frac{1}{\gamma_x T}\frac{2\sqrt{2}\sigma}{B_0^{1-1/s}}+\frac{8(\eta_x+\eta_z)L_g}{(1-\lambda)\sqrt{\gamma_x}}\frac{1}{\sqrt{K}}+\frac{2\sqrt{2}\gamma_x^{1-1/s}\sigma}{K^{1-1/s}}\,. \qquad (79)
\end{aligned}$$

Lemma C.25. *Given Assumptions 3.1-3.3, we obtain*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}v_{1,t}^{(k)}-\frac{1}{K}\sum_{k=1}^{K}\nabla_2 f^{(k)}(x_t^{(k)},y_t^{(k)})\|] \\
&\leq \frac{1}{\gamma_y T}\frac{2\sqrt{2}}{B_0^{1-1/s}}\sigma+\frac{8(\eta_x+\eta_y)L_f}{(1-\lambda)\sqrt{\gamma_y}}\frac{1}{\sqrt{K}}+\frac{2\sqrt{2}\gamma_y^{1-1/s}\sigma}{K^{1-1/s}}\,. \qquad (80)
\end{aligned}$$

Lemma C.26. *Given Assumptions 3.1-3.3, we obtain*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}v_{2,t}^{(k)}-\frac{1}{K}\sum_{k=1}^{K}\nabla_2 g^{(k)}(x_t^{(k)},y_t^{(k)})\|] \\
&\leq \frac{1}{\gamma_y T}\frac{2\sqrt{2}}{B_0^{1-1/s}}\sigma+\frac{8(\eta_x+\eta_y)L_g}{(1-\lambda)\sqrt{\gamma_y}}\frac{1}{\sqrt{K}}+\frac{2\sqrt{2}\gamma_y^{1-1/s}\sigma}{K^{1-1/s}}\,. \qquad (81)
\end{aligned}$$

Lemma C.27. *Given Assumptions 3.1-3.3, we obtain*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}w_{1,t}^{(k)}-\frac{1}{K}\sum_{k=1}^{K}\nabla_2 g^{(k)}(x_t^{(k)},z_t^{(k)})\|] \\
&\leq \frac{1}{\gamma_z T}\frac{2\sqrt{2}\sigma}{B_0^{1-1/s}}+\frac{8(\eta_x+\eta_z)L_g}{(1-\lambda)\sqrt{\gamma_z}}\frac{1}{\sqrt{K}}+\frac{2\sqrt{2}\gamma_z^{1-1/s}\sigma}{K^{1-1/s}}\,. \qquad (82)
\end{aligned}$$

C.5 Bounding Consensus Errors

Lemma C.28. *Given Assumptions 3.1-3.3, we obtain*

$$\frac{1}{K}\sum_{k=1}^{K}\|x_t^{(k)}-\bar{x}_t\|\leq\frac{\eta_x\lambda}{1-\lambda}\,;\quad \frac{1}{K}\sum_{k=1}^{K}\|y_t^{(k)}-\bar{y}_t\|\leq\frac{\eta_y\lambda}{1-\lambda}\,;\quad \frac{1}{K}\sum_{k=1}^{K}\|z_t^{(k)}-\bar{z}_t\|\leq\frac{\eta_z\lambda}{1-\lambda}\,. \qquad (83)$$

Proof.

$$
\begin{aligned}
&\frac{1}{K}\|X_t - \bar{X}_t\|_F^2 \\
&= \frac{1}{K}\|(X_{t-1} - \eta_x \hat{P}_{t-1})E - (\bar{X}_{t-1} - \eta_x \bar{\hat{P}}_{t-1})\|_F^2 \\
&\overset{(a)}{=} \frac{1}{K}\|((X_{t-1} - \eta_x \hat{P}_{t-1}) - (\bar{X}_{t-1} - \eta_x \bar{\hat{P}}_{t-1}))(E - \frac{\mathbf{1}\mathbf{1}^T}{K})\|_F^2 \\
&\leq \frac{1}{K}\|(X_{t-1} - \eta_x \hat{P}_{t-1}) - (\bar{X}_{t-1} - \eta_x \bar{\hat{P}}_{t-1})\|_F^2 \|E - \frac{\mathbf{1}\mathbf{1}^T}{K}\|_2^2 \\
&\overset{(b)}{\leq} \lambda^2 \frac{1}{K}\|(X_{t-1} - \eta_x \hat{P}_{t-1}) - (\bar{X}_{t-1} - \eta_x \bar{\hat{P}}_{t-1})\|_F^2 \\
&\leq \lambda^2(1 + 1/a)\frac{1}{K}\|X_{t-1} - \bar{X}_{t-1}\|_F^2 + \eta_x^2\lambda^2(1+a)\frac{1}{K}\|\hat{P}_{t-1} - \bar{\hat{P}}_{t-1}\|_F^2 \\
&\overset{(c)}{\leq} \lambda \frac{1}{K}\|X_{t-1} - \bar{X}_{t-1}\|_F^2 + \eta_x^2 \frac{\lambda^2}{1-\lambda}\frac{1}{K}\|\hat{P}_{t-1} - \bar{\hat{P}}_{t-1}\|_F^2 \\
&\leq \lambda \frac{1}{K}\|X_{t-1} - \bar{X}_{t-1}\|_F^2 + \eta_x^2 \frac{\lambda^2}{1-\lambda}\frac{1}{K}\|\hat{P}_{t-1}\|_F^2 \\
&= \lambda \frac{1}{K}\|X_{t-1} - \bar{X}_{t-1}\|_F^2 + \eta_x^2 \frac{\lambda^2}{1-\lambda}\frac{1}{K}\sum_{k=1}^K \|\frac{p_{t-1}^{(k)}}{\|p_{t-1}^{(k)}\|}\|^2 \\
&= \lambda \frac{1}{K}\|X_{t-1} - \bar{X}_{t-1}\|_F^2 + \eta_x^2 \frac{\lambda^2}{1-\lambda} \\
&\leq \frac{\eta_x^2\lambda^2}{(1-\lambda)^2}\,,
\end{aligned} \tag{84}
$$

where (a) holds because E is a doubly stochastic matrix, (b) holds due to Assumption 3.3, (c) holds due to $a = \frac{\lambda}{1-\lambda}$.

Then, we obtain

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^K \|x_t^{(k)} - \bar{x}_t\| = \sqrt{\frac{1}{K^2}\left(\sum_{k=1}^K \|x_t^{(k)} - \bar{x}_t\|\right)^2} \\
&\leq \sqrt{\frac{1}{K^2}K\sum_{k=1}^K \|x_t^{(k)} - \bar{x}_t\|^2} \leq \sqrt{\frac{1}{K}\|X_t - \bar{X}_t\|_F^2} \leq \frac{\eta_x\lambda}{1-\lambda}\,.
\end{aligned} \tag{85}
$$

The other two inequalities can be proved in a same approach. □

Lemma C.29. *Given Assumptions 3.1-3.3, we obtain*

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|p_t^{(k)} - \bar{p}_t\|] \\
&\leq \frac{2\lambda}{(1-\lambda)T}\frac{1}{\sqrt{K}}\sum_{k=1}^K \mathbb{E}[\|\nabla_1 f^{(k)}(x_0^{(k)}, y_0^{(k)})\|] + \frac{1}{\delta}\frac{2\lambda}{(1-\lambda)T}\frac{1}{\sqrt{K}}\sum_{k=1}^K \mathbb{E}[\|\nabla_1 g^{(k)}(x_0^{(k)}, y_0^{(k)})\|] \\
&\quad + \frac{1}{\delta}\frac{2\lambda}{(1-\lambda)T}\frac{1}{\sqrt{K}}\sum_{k=1}^K \mathbb{E}[\|\nabla_1 g^{(k)}(x_0^{(k)}, z_0^{(k)})\|] + \frac{\lambda\sqrt{K}}{(1-\lambda)T}\frac{6\sqrt{2}}{B_0^{1-1/s}}\left(1 + \frac{2}{\delta}\right)\sigma \\
&\quad + \frac{\gamma_x\lambda\sqrt{K}\sigma}{1-\lambda}\left(1 + \frac{2}{\delta}\right) + \frac{4(1+2\sqrt{\gamma_x})\eta_x\lambda\sqrt{K}}{(1-\lambda)^2}\left(L_f + \frac{2L_g}{\delta}\right) \\
&\quad + \frac{4(1+2\sqrt{\gamma_x})\eta_y\lambda\sqrt{K}}{(1-\lambda)^2}\left(L_f + \frac{L_g}{\delta}\right) + \frac{4(1+2\sqrt{\gamma_x})\lambda\eta_z\sqrt{K}}{(1-\lambda)^2}\frac{L_g}{\delta}
\end{aligned}
$$

$$+\frac{2\sqrt{2}\gamma_x^{2-1/s}\lambda\sqrt{K}}{1-\lambda}\sigma\left(1+\frac{2}{\delta}\right)\ . \tag{86}$$

Proof. When $t>0$, similar to the proof of Lemma C.28, we obtain

$$\begin{aligned}
&\|P_t-\bar{P}_t\|_F\\
&=\|(P_{t-1}-U_{t-1}+U_t)E-(\bar{P}_{t-1}-\bar{U}_{t-1}+\bar{U}_t)\|_F\\
&\overset{(a)}{\leq}\lambda\|(P_{t-1}-U_{t-1}+U_t)-(\bar{P}_{t-1}-\bar{U}_{t-1}+\bar{U}_t)\|_F\\
&\leq\lambda\|P_{t-1}-\bar{P}_{t-1}\|_F+\lambda\|U_t-U_{t-1}-(\bar{U}_t-\bar{U}_{t-1})\|_F\\
&\leq\lambda\|P_{t-1}-\bar{P}_{t-1}\|_F+\lambda\|U_t-U_{t-1}\|_F\\
&\leq\lambda^t\|P_0-\bar{P}_0\|_F+\sum_{j=1}^{t}\lambda^{t-j+1}\|U_j-U_{j-1}\|_F\\
&\overset{(b)}{\leq}\lambda^{t+1}\|U_0-\bar{U}_0\|_F+\sum_{j=1}^{t}\lambda^{t-j+1}\|U_j-U_{j-1}\|_F\ ,
\end{aligned}\tag{87}$$

where (a) holds due to Assumption 3.3, (b) holds due to the following inequality.

$$\begin{aligned}
&\|P_0-\bar{P}_0\|_F=\|U_0E-\bar{U}_0\|_F=\|(U_0-\bar{U}_0)(E-\frac{\mathbf{1}\mathbf{1}^T}{K})\|_F\\
&\leq\|U_0-\bar{U}_0\|_F\|E-\frac{\mathbf{1}\mathbf{1}^T}{K}\|_2\leq\lambda\|U_0-\bar{U}_0\|_F\ .
\end{aligned}\tag{88}$$

Then, we obtain

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_t^{(k)}-\bar{p}_t\|]=\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|p_t^{(k)}-\bar{p}_t\|\right]\\
&\overset{(a)}{\leq}\frac{1}{\sqrt{K}}\mathbb{E}[\|P_t-\bar{P}_t\|_F]\\
&\leq\lambda^{t+1}\frac{1}{\sqrt{K}}\mathbb{E}[\|U_0-\bar{U}_0\|_F]+\frac{1}{\sqrt{K}}\sum_{j=1}^{t}\lambda^{t-j+1}\mathbb{E}[\|U_j-U_{j-1}\|_F]\\
&\overset{(b)}{\leq}\lambda^{t+1}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|u_0^{(k)}-\bar{u}_0\|]+\frac{1}{\sqrt{K}}\sum_{j=1}^{t}\lambda^{t-j+1}\sum_{k=1}^{K}\mathbb{E}[\|u_j^{(k)}-u_{j-1}^{(k)}\|]\\
&\leq\lambda^{t+1}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|u_0^{(k)}-\bar{u}_0\|]+\frac{1}{\sqrt{K}}\sum_{j=1}^{t}\lambda^{t-j+1}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,j}^{(k)}-u_{1,j-1}^{(k)}\|]\\
&\quad+\frac{1}{\sqrt{K}}\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,j}^{(k)}-u_{2,j-1}^{(k)}\|]\\
&\quad+\frac{1}{\sqrt{K}}\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,j}^{(k)}-u_{3,j-1}^{(k)}\|]\ ,
\end{aligned}\tag{89}$$

where (a) holds due to $\frac{1}{K}\sum_{k=1}^{K}a_k\leq\sqrt{\frac{1}{K}\sum_{k=1}^{K}a_k^2}$ as $0\leq\frac{1}{K}\sum_{k=1}^{K}(a_k-\bar{a})^2=\frac{1}{K}\sum_{k=1}^{K}a_k^2-\bar{a}^2$ for any $a_k\geq 0$ and $\bar{a}=\frac{1}{K}\sum_{k=1}^{K}a_k$, (b) holds due to $\|[a_1,\cdots,a_K]\|_F=\sqrt{\sum_{k=1}^{K}\|a_k\|^2}\leq\sum_{k=1}^{K}\|a_k\|$ for any vector a_i.

Similarly, when $t=0$, we have

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_0^{(k)}-\bar{p}_0\|]\leq\frac{1}{\sqrt{K}}\mathbb{E}[\|P_0-\bar{P}_0\|_F]\leq\frac{\lambda}{\sqrt{K}}\mathbb{E}[\|U_0-\bar{U}_0\|_F]\leq\frac{\lambda}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|u_0^{(k)}-\bar{u}_0\|]\ . \tag{90}$$

Then, due to Lemmas C.10 - C.12, we obtain

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_t^{(k)}-\bar{p}_t\|] &\le \lambda^{t+1}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|u_0^{(k)}-\bar{u}_0\|] \\
&+\gamma_x\frac{1}{\sqrt{K}}\sum_{j=1}^{t}\lambda^{t-j+1}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})-\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)};\xi_j^{(k)})\|] \\
&+\gamma_x\frac{1}{\sqrt{K}}\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})-\nabla_1 g^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)};\zeta_j^{(k)})\|] \\
&+\gamma_x\frac{1}{\sqrt{K}}\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_{j-1}^{(k)},z_{j-1}^{(k)})-\nabla_1 g^{(k)}(x_{j-1}^{(k)},z_{j-1}^{(k)};\zeta_j^{(k)})\|] \\
&+\gamma_x\frac{1}{\sqrt{K}}\sum_{j=1}^{t}\lambda^{t-j+1}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,j-1}^{(k)}-\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})\|] \\
&+\gamma_x\frac{1}{\sqrt{K}}\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,j-1}^{(k)}-\nabla_1 g^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})\|] \\
&+\gamma_x\frac{1}{\sqrt{K}}\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,j-1}^{(k)}-\nabla_1 g^{(k)}(x_{j-1}^{(k)},z_{j-1}^{(k)})\|] \\
&+\frac{4\eta_x\sqrt{K}}{1-\lambda}\left(L_f+\frac{2L_g}{\delta}\right)\sum_{j=1}^{t}\lambda^{t-j+1} \\
&+\frac{4\eta_y\sqrt{K}}{1-\lambda}\left(L_f+\frac{L_g}{\delta}\right)\sum_{j=1}^{t}\lambda^{t-j+1}+\frac{4\eta_z\sqrt{K}}{1-\lambda}\frac{L_g}{\delta}\sum_{j=1}^{t}\lambda^{t-j+1}\,. \qquad (91)
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_t^{(k)}-\bar{p}_t\|]=\frac{1}{T}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_0^{(k)}-\bar{p}_0\|]+\frac{1}{T}\sum_{t=1}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_t^{(k)}-\bar{p}_t\|] \\
\le\ &\frac{1}{T}\sum_{t=0}^{T-1}\lambda^{t+1}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|u_0^{(k)}-\bar{u}_0\|] \\
&+\gamma_x\frac{1}{\sqrt{K}}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})-\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)};\xi_j^{(k)})\|] \\
&+\gamma_x\frac{1}{\sqrt{K}}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})-\nabla_1 g^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)};\zeta_j^{(k)})\|] \\
&+\gamma_x\frac{1}{\sqrt{K}}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_{j-1}^{(k)},z_{j-1}^{(k)})-\nabla_1 g^{(k)}(x_{j-1}^{(k)},z_{j-1}^{(k)};\zeta_j^{(k)})\|] \\
&+\gamma_x\frac{1}{\sqrt{K}}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,j-1}^{(k)}-\nabla_1 f^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})\|] \\
&+\gamma_x\frac{1}{\sqrt{K}}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,j-1}^{(k)}-\nabla_1 g^{(k)}(x_{j-1}^{(k)},y_{j-1}^{(k)})\|] \\
&+\gamma_x\frac{1}{\sqrt{K}}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,j-1}^{(k)}-\nabla_1 g^{(k)}(x_{j-1}^{(k)},z_{j-1}^{(k)})\|]
\end{aligned}$$

$$+\frac{4\eta_x\sqrt{K}}{1-\lambda}\left(L_f+\frac{2L_g}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}$$
$$+\frac{4\eta_y\sqrt{K}}{1-\lambda}\left(L_f+\frac{L_g}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}+\frac{4\eta_z\sqrt{K}}{1-\lambda}\frac{L_g}{\delta}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}\,. \tag{92}$$

Note that $\nabla_1 f^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)}; \xi_j^{(k)})$ is computed with the samples in the $j \geq 1$-th iteration, where the batch size is 1, then for any $j \in \{1, \cdots, t\}$, we obtain

$$\begin{aligned}
&\mathbb{E}[\|\nabla_1 f^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)}) - \nabla_1 f^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)}; \xi_j^{(k)})\|] \\
&= \mathbb{E}[(\|\nabla_1 f^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)}) - \nabla_1 f^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)}; \xi_j^{(k)})\|^s)^{1/s}] \\
&\overset{(a)}{\leq} (\mathbb{E}[\|\nabla_1 f^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)}) - \nabla_1 f^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)}; \xi_j^{(k)})\|^s])^{1/s} \overset{(b)}{\leq} \sigma\,,
\end{aligned} \tag{93}$$

where (a) holds due to Jensen's inequality, and (b) holds due to Assumption 3.2.

Similarly, we obtain

$$\begin{aligned}
&\mathbb{E}[\|\nabla_1 g^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)}) - \nabla_1 g^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)}; \zeta_j^{(k)})\|] \leq \sigma\,, \\
&\mathbb{E}[\|\nabla_1 g^{(k)}(x_{j-1}^{(k)}, z_{j-1}^{(k)}) - \nabla_1 g^{(k)}(x_{j-1}^{(k)}, z_{j-1}^{(k)}; \zeta_j^{(k)})\|] \leq \sigma\,.
\end{aligned} \tag{94}$$

Then, we obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_t^{(k)} - \bar{p}_t\|] \\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}\lambda^{t+1}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|u_0^{(k)} - \bar{u}_0\|] + \gamma_x\sqrt{K}\sigma\left(1+\frac{2}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} \\
&\quad + \frac{4\eta_x\sqrt{K}}{1-\lambda}\left(L_f+\frac{2L_g}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} \\
&\quad + \frac{4\eta_y\sqrt{K}}{1-\lambda}\left(L_f+\frac{L_g}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} + \frac{4\eta_z\sqrt{K}}{1-\lambda}\frac{L_g}{\delta}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} \\
&\quad + \gamma_x\frac{1}{\sqrt{K}}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,j-1}^{(k)} - \nabla_1 f^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)})\|] \\
&\quad + \gamma_x\frac{1}{\sqrt{K}}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,j-1}^{(k)} - \nabla_1 g^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)})\|] \\
&\quad + \gamma_x\frac{1}{\sqrt{K}}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,j-1}^{(k)} - \nabla_1 g^{(k)}(x_{j-1}^{(k)}, z_{j-1}^{(k)})\|]\,.
\end{aligned} \tag{95}$$

Then, based on Lemmas C.16, C.17, C.18, we obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_t^{(k)} - \bar{p}_t\|] \\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}\lambda^{t+1}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|u_0^{(k)} - \bar{u}_0\|] + \gamma_x\sqrt{K}\sigma\left(1+\frac{2}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} \\
&\quad + \frac{4\eta_x\sqrt{K}}{1-\lambda}\left(L_f+\frac{2L_g}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{4\eta_y\sqrt{K}}{1-\lambda}\left(L_f + \frac{L_g}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} + \frac{4\eta_z\sqrt{K}}{1-\lambda}\frac{L_g}{\delta}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} \\
&+ \gamma_x\sqrt{K}\frac{2\sqrt{2}\sigma}{B_0^{1-1/s}}\left(1+\frac{2}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}(1-\gamma_x)^{j-1} \\
&+ \frac{8\eta_x\sqrt{\gamma_x}\sqrt{K}}{(1-\lambda)}\left(L_f + \frac{2L_g}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} + \frac{8\eta_y\sqrt{\gamma_x}\sqrt{K}}{(1-\lambda)}\left(L_f + \frac{L_g}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} \\
&+ \frac{8\eta_z\sqrt{\gamma_x}\sqrt{K}}{(1-\lambda)}\frac{L_g}{\delta}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} + 2\sqrt{2}\sqrt{K}\gamma_x^{2-1/s}\sigma\left(1+\frac{2}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}\,. \quad (96)
\end{aligned}
$$

Because

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\lambda^{t+1} \le \frac{\lambda}{(1-\lambda)T}\,, \\
&\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} = \frac{1}{T}\sum_{t=1}^{T-1}\sum_{r=1}^{t}\lambda^{r} = \frac{\lambda}{T(1-\lambda)}\sum_{t=1}^{T-1}(1-\lambda^t) \le \frac{T-1}{T}\frac{\lambda}{1-\lambda} \le \frac{\lambda}{1-\lambda}\,, \\
&\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}(1-\gamma_x)^{j-1} = \frac{1}{T}\sum_{j=1}^{T-1}(1-\gamma_x)^{j-1}\sum_{t=j}^{T-1}\lambda^{t-j+1} \\
&\quad \le \frac{\lambda}{T(1-\lambda)}\sum_{j=1}^{T-1}(1-\gamma_x)^{j-1} \le \frac{\lambda}{T\gamma_x(1-\lambda)}\,,
\end{aligned}
$$

we obtain

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_t^{(k)} - \bar{p}_t\|] \\
&\le \frac{\lambda}{(1-\lambda)T}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|u_0^{(k)} - \bar{u}_0\|] + \frac{\gamma_x\lambda\sqrt{K}\sigma}{1-\lambda}\left(1+\frac{2}{\delta}\right) \\
&\quad + \frac{4\eta_x\lambda\sqrt{K}}{(1-\lambda)^2}\left(L_f + \frac{2L_g}{\delta}\right) + \frac{4\eta_y\lambda\sqrt{K}}{(1-\lambda)^2}\left(L_f + \frac{L_g}{\delta}\right) + \frac{4\lambda\eta_z\sqrt{K}}{(1-\lambda)^2}\frac{L_g}{\delta} \\
&\quad + \frac{\lambda\sqrt{K}}{T(1-\lambda)}\frac{2\sqrt{2}\sigma}{B_0^{1-1/s}}\left(1+\frac{2}{\delta}\right) + \frac{2\sqrt{2}\gamma_x^{2-1/s}\lambda\sqrt{K}}{1-\lambda}\sigma\left(1+\frac{2}{\delta}\right) \\
&\quad + \frac{8\eta_x\sqrt{\gamma_x}\lambda\sqrt{K}}{(1-\lambda)^2}\left(L_f + \frac{2L_g}{\delta}\right) + \frac{8\eta_y\sqrt{\gamma_x}\lambda\sqrt{K}}{(1-\lambda)^2}\left(L_f + \frac{L_g}{\delta}\right) + \frac{8\eta_z\sqrt{\gamma_x}\lambda\sqrt{K}}{(1-\lambda)^2}\frac{L_g}{\delta}\,. \quad (97)
\end{aligned}
$$

For $\sum_{k=1}^{K}\mathbb{E}[\|u_0^{(k)} - \bar{u}_0\|]$, we bound it as follows:

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}[\|u_0^{(k)} - \bar{u}_0\|] = \sum_{k=1}^{K}\mathbb{E}[\|u_{1,0}^{(k)} + \frac{1}{\delta}u_{2,0}^{(k)} + \frac{1}{\delta}u_{3,0}^{(k)} - \frac{1}{K}\sum_{j=1}^{K}(u_{1,0}^{(j)} + \frac{1}{\delta}u_{2,0}^{(j)} + \frac{1}{\delta}u_{3,0}^{(j)})\|] \\
&\le \sum_{k=1}^{K}\mathbb{E}[\|u_{1,0}^{(k)} - \frac{1}{K}\sum_{j=1}^{K}u_{1,0}^{(j)}\|] + \sum_{k=1}^{K}\mathbb{E}[\|\frac{1}{\delta}u_{2,0}^{(k)} - \frac{1}{K}\sum_{j=1}^{K}\frac{1}{\delta}u_{2,0}^{(j)}\|] + \sum_{k=1}^{K}\mathbb{E}[\|\frac{1}{\delta}u_{3,0}^{(k)} - \frac{1}{K}\sum_{j=1}^{K}\frac{1}{\delta}u_{3,0}^{(j)}\|] \\
&\le \sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(x_0^{(k)}, y_0^{(k)}; \xi_0^{(k)}) - \frac{1}{K}\sum_{j=1}^{K}\nabla_1 f^{(j)}(x_0^{(j)}, y_0^{(j)}; \xi_0^{(j)})\|] \\
&\quad + \frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_0^{(k)}, y_0^{(k)}; \zeta_0^{(k)}) - \frac{1}{K}\sum_{j=1}^{K}\nabla_1 g^{(j)}(x_0^{(j)}, y_0^{(j)}; \zeta_0^{(j)})\|]
\end{aligned}
$$

$$+\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_0^{(k)},z_0^{(k)};\zeta_0^{(k)})-\frac{1}{K}\sum_{j=1}^{K}\nabla_1 g^{(j)}(x_0^{(j)},z_0^{(j)};\zeta_0^{(j)})\|]\,. \tag{98}$$

For the first term on the last step of Eq. (98), we bound it as follows:

$$\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(x_0^{(k)},y_0^{(k)};\xi_0^{(k)})-\frac{1}{K}\sum_{j=1}^{K}\nabla_1 f^{(j)}(x_0^{(j)},y_0^{(j)};\xi_0^{(j)})\|]\\
&\le\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(x_0^{(k)},y_0^{(k)};\xi_0^{(k)})-\nabla_1 f^{(k)}(x_0^{(k)},y_0^{(k)})\|]\\
&\quad+\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(x_0^{(k)},y_0^{(k)})-\frac{1}{K}\sum_{j=1}^{K}\nabla_1 f^{(j)}(x_0^{(j)},y_0^{(j)})\|]\\
&\quad+\sum_{k=1}^{K}\mathbb{E}[\|\frac{1}{K}\sum_{j=1}^{K}\nabla_1 f^{(j)}(x_0^{(j)},y_0^{(j)})-\frac{1}{K}\sum_{j=1}^{K}\nabla_1 f^{(j)}(x_0^{(j)},y_0^{(j)};\xi_0^{(j)})\|]\\
&\le 2\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(x_0^{(k)},y_0^{(k)};\xi_0^{(k)})-\nabla_1 f^{(k)}(x_0^{(k)},y_0^{(k)})\|]+2\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(x_0^{(k)},y_0^{(k)})\|]\\
&\overset{(a)}{\le} 2\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(x_0^{(k)},y_0^{(k)})\|]+\frac{4\sqrt{2}K}{B_0^{1-1/s}}\sigma\,,
\end{aligned} \tag{99}$$

where (a) holds due to Eq. (60)

Similarly, we bound the second term on the last step of Eq. (98) as follows:

$$\begin{aligned}
&\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_0^{(k)},y_0^{(k)};\zeta_0^{(k)})-\frac{1}{K}\sum_{j=1}^{K}\nabla_1 g^{(j)}(x_0^{(j)},y_0^{(j)};\zeta_0^{(j)})\|]\\
&\le\frac{2}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_0^{(k)},y_0^{(k)})\|]+\frac{4\sqrt{2}K}{B_0^{1-1/s}}\frac{1}{\delta}\sigma\,.
\end{aligned} \tag{100}$$

By combining them together, we obtain

$$\begin{aligned}
\sum_{k=1}^{K}\mathbb{E}[\|u_0^{(k)}-\bar{u}_0\|]&\le 2\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(x_0^{(k)},y_0^{(k)})\|]+\frac{2}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_0^{(k)},y_0^{(k)})\|]\\
&\quad+\frac{2}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_0^{(k)},z_0^{(k)})\|]+\frac{4\sqrt{2}K}{B_0^{1-1/s}}\left(1+\frac{2}{\delta}\right)\sigma\,.
\end{aligned} \tag{101}$$

Finally, by substituting the above inequality into Eq. (97), the proof is complete.

□

Lemma C.30. *Given Assumptions 3.1-3.3, we obtain*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|q_t^{(k)}-\bar{q}_t\|]\\
&\le\frac{2\lambda}{(1-\lambda)T}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_2 f^{(k)}(x_0^{(k)},y_0^{(k)})\|]+\frac{2\lambda}{(1-\lambda)T}\frac{1}{\delta}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_2 g^{(k)}(x_0^{(k)},y_0^{(k)})\|]\\
&\quad+\frac{\lambda}{(1-\lambda)T}\frac{6\sqrt{2}\sqrt{K}}{B_0^{1-1/s}}\left(1+\frac{1}{\delta}\right)\sigma+\frac{\gamma_y\lambda\sqrt{K}\sigma}{1-\lambda}\left(1+\frac{1}{\delta}\right)\\
&\quad+\frac{4(1+2\sqrt{\gamma_y})\eta_x\lambda\sqrt{K}}{(1-\lambda)^2}\left(L_f+\frac{L_g}{\delta}\right)+\frac{4(1+2\sqrt{\gamma_y})\eta_y\lambda\sqrt{K}}{(1-\lambda)^2}\left(L_f+\frac{L_g}{\delta}\right)
\end{aligned}$$

$$+ \frac{2\sqrt{2}\gamma_y^{2-1/s}\lambda\sqrt{K}}{1-\lambda}\sigma\left(1+\frac{1}{\delta}\right) . \tag{102}$$

Proof. Following the proof of Lemma C.29, for any $t \geq 0$, we obtain

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|q_t^{(k)} - \bar{q}_t\|] \\
&\leq \lambda^{t+1}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|v_0^{(k)} - \bar{v}_0\|] + \sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|v_j^{(k)} - v_{j-1}^{(k)}\|] \\
&\leq \lambda^{t+1}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|v_0^{(k)} - \bar{v}_0\|] + \frac{1}{\sqrt{K}}\sum_{j=1}^{t}\lambda^{t-j+1}\sum_{k=1}^{K}\mathbb{E}[\|v_{1,j}^{(k)} - v_{1,j-1}^{(k)}\|] \\
&\quad + \frac{1}{\sqrt{K}}\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|v_{2,j}^{(k)} - v_{2,j-1}^{(k)}\|] .
\end{aligned} \tag{103}$$

Based on Lemma C.13 and Lemma C.14, we obtain

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|q_t^{(k)} - \bar{q}_t\|] &\leq \lambda^{t+1}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|v_0^{(k)} - \bar{v}_0\|] \\
&+ \gamma_y\frac{1}{\sqrt{K}}\sum_{j=1}^{t}\lambda^{t-j+1}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_2 f^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)}) - \nabla_2 f^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)}; \xi_j^{(k)})\|] \\
&+ \gamma_y\frac{1}{\sqrt{K}}\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_2 g^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)}) - \nabla_2 g^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)}; \zeta_j^{(k)})\|] \\
&+ \gamma_y\frac{1}{\sqrt{K}}\sum_{j=1}^{t}\lambda^{t-j+1}\sum_{k=1}^{K}\mathbb{E}[\|v_{1,j-1}^{(k)} - \nabla_2 f^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)})\|] \\
&+ \gamma_y\frac{1}{\sqrt{K}}\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|v_{2,j-1}^{(k)} - \nabla_2 g^{(k)}(x_{j-1}^{(k)}, y_{j-1}^{(k)})\|] \\
&+ \frac{4\eta_x\sqrt{K}}{1-\lambda}\left(L_f + \frac{L_g}{\delta}\right)\sum_{j=1}^{t}\lambda^{t-j+1} + \frac{4\eta_y\sqrt{K}}{1-\lambda}\left(L_f + \frac{L_g}{\delta}\right)\sum_{j=1}^{t}\lambda^{t-j+1} .
\end{aligned} \tag{104}$$

Then, we obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|q_t^{(k)} - \bar{q}_t\|] \\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}\lambda^{t+1}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|v_0^{(k)} - \bar{v}_0\|] + \gamma_y\sqrt{K}\sigma\left(1+\frac{1}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} \\
&\quad + \frac{4\eta_x\sqrt{K}}{1-\lambda}\left(L_f + \frac{L_g}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} + \frac{4\eta_y\sqrt{K}}{1-\lambda}\left(L_f + \frac{L_g}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} \\
&\quad + \gamma_y\sqrt{K}\frac{2\sqrt{2}\sigma}{B_0^{1-1/s}}\left(1+\frac{1}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}(1-\gamma_y)^{j-1} \\
&\quad + \frac{8\eta_x\sqrt{\gamma_y}\sqrt{K}}{(1-\lambda)}\left(L_f + \frac{L_g}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} + \frac{8\eta_y\sqrt{\gamma_y}\sqrt{K}}{(1-\lambda)}\left(L_f + \frac{L_g}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} \\
&\quad + 2\sqrt{2}\sqrt{K}\gamma_y^{2-1/s}\sigma\left(1+\frac{1}{\delta}\right)\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\lambda}{(1-\lambda)T}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|v_0^{(k)}-\bar{v}_0\|]+\frac{\gamma_y\lambda\sqrt{K}\sigma}{1-\lambda}\left(1+\frac{1}{\delta}\right)\\
&+\frac{4\eta_x\lambda\sqrt{K}}{(1-\lambda)^2}\left(L_f+\frac{L_g}{\delta}\right)+\frac{4\eta_y\lambda\sqrt{K}}{(1-\lambda)^2}\left(L_f+\frac{L_g}{\delta}\right)\\
&+\frac{\lambda\sqrt{K}}{T(1-\lambda)}\frac{2\sqrt{2}\sigma}{B_0^{1-1/s}}\left(1+\frac{1}{\delta}\right)+\frac{2\sqrt{2}\gamma_y^{2-1/s}\lambda\sqrt{K}}{1-\lambda}\sigma\left(1+\frac{1}{\delta}\right)\\
&+\frac{8\eta_x\sqrt{\gamma_y}\lambda\sqrt{K}}{(1-\lambda)^2}\left(L_f+\frac{L_g}{\delta}\right)+\frac{8\eta_y\sqrt{\gamma_y}\lambda\sqrt{K}}{(1-\lambda)^2}\left(L_f+\frac{L_g}{\delta}\right). \qquad (105)
\end{aligned}
$$

For $\sum_{k=1}^{K}\mathbb{E}[\|v_0^{(k)}-\bar{v}_0\|]$, we bound it as follows:

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}[\|v_0^{(k)}-\bar{v}_0\|]=\sum_{k=1}^{K}\mathbb{E}[\|v_{1,0}^{(k)}+\frac{1}{\delta}v_{2,0}^{(k)}-\frac{1}{K}\sum_{j=1}^{K}(v_{1,0}^{(j)}+\frac{1}{\delta}v_{2,0}^{(j)})\|]\\
&\leq\sum_{k=1}^{K}\mathbb{E}[\|v_{1,0}^{(k)}-\frac{1}{K}\sum_{j=1}^{K}v_{1,0}^{(j)}\|]+\sum_{k=1}^{K}\mathbb{E}[\|\frac{1}{\delta}v_{2,0}^{(k)}-\frac{1}{K}\sum_{j=1}^{K}\frac{1}{\delta}v_{2,0}^{(j)}\|]\\
&\leq\sum_{k=1}^{K}\mathbb{E}[\|\nabla_2 f^{(k)}(x_0^{(k)},y_0^{(k)};\xi_0^{(k)})-\frac{1}{K}\sum_{j=1}^{K}\nabla_2 f^{(j)}(x_0^{(j)},y_0^{(j)};\xi_0^{(j)})\|]\\
&\quad+\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_2 g^{(k)}(x_0^{(k)},y_0^{(k)};\zeta_0^{(k)})-\frac{1}{K}\sum_{j=1}^{K}\nabla_2 g^{(j)}(x_0^{(j)},y_0^{(j)};\zeta_0^{(j)})\|]. \qquad (106)
\end{aligned}
$$

For the first term on the last step of Eq. (106), we bound it as follows:

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}[\|\nabla_2 f^{(k)}(x_0^{(k)},y_0^{(k)};\xi_0^{(k)})-\frac{1}{K}\sum_{j=1}^{K}\nabla_2 f^{(j)}(x_0^{(j)},y_0^{(j)};\xi_0^{(j)})\|]\\
&\leq 2\sum_{k=1}^{K}\mathbb{E}[\|\nabla_2 f^{(k)}(x_0^{(k)},y_0^{(k)})\|]+\frac{4\sqrt{2}K}{B_0^{1-1/s}}\sigma. \qquad (107)
\end{aligned}
$$

Similarly, we bound the second term on the last step of Eq. (106) as follows:

$$
\begin{aligned}
&\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_2 g^{(k)}(x_0^{(k)},y_0^{(k)};\zeta_0^{(k)})-\frac{1}{K}\sum_{j=1}^{K}\nabla_2 g^{(j)}(x_0^{(j)},y_0^{(j)};\zeta_0^{(j)})\|]\\
&\leq\frac{2}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_2 g^{(k)}(x_0^{(k)},y_0^{(k)})\|]+\frac{4\sqrt{2}K}{B_0^{1-1/s}}\frac{1}{\delta}\sigma. \qquad (108)
\end{aligned}
$$

By combining them together, we obtain

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}[\|v_0^{(k)}-\bar{v}_0\|]\leq 2\sum_{k=1}^{K}\mathbb{E}[\|\nabla_2 f^{(k)}(x_0^{(k)},y_0^{(k)})\|]\\
&+\frac{2}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_2 g^{(k)}(x_0^{(k)},y_0^{(k)})\|]+\frac{4\sqrt{2}K}{B_0^{1-1/s}}\left(1+\frac{1}{\delta}\right)\sigma. \qquad (109)
\end{aligned}
$$

Finally, by substituting the above inequality into Eq. (105), the proof is complete.

□

Lemma C.31. *Given Assumptions 3.1-3.3, we obtain*

$$
\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|r_t^{(k)}-\bar{r}_t\|]
$$

$$\leq \frac{2\lambda}{(1-\lambda)T}\frac{1}{\delta}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_2 g^{(k)}(x_0^{(k)}, z_0^{(k)})\|] + \frac{\lambda\sqrt{K}}{T(1-\lambda)}\frac{6\sqrt{2}}{B_0^{1-1/s}}\frac{1}{\delta}\sigma + \frac{\gamma_z\lambda\sqrt{K}}{1-\lambda}\frac{1}{\delta}\sigma$$
$$+ \frac{4(1+2\sqrt{\gamma_z})\eta_x\lambda\sqrt{K}}{(1-\lambda)^2}\frac{L_g}{\delta} + \frac{4(1+2\sqrt{\gamma_z})\eta_z\lambda\sqrt{K}}{(1-\lambda)^2}\frac{L_g}{\delta} + \frac{2\sqrt{2}\gamma_z^{2-1/s}\lambda\sqrt{K}}{1-\lambda}\frac{1}{\delta}\sigma\,. \tag{110}$$

Proof. Following the proof of Lemma C.29, we obtain

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|r_t^{(k)} - \bar{r}_t\|]$$
$$\leq \lambda^{1+t}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|w_0^{(k)} - \bar{w}_0\|] + \sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|w_j^{(k)} - w_{j-1}^{(k)}\|]$$
$$\leq \lambda^{1+t}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|w_0^{(k)} - \bar{w}_0\|] + \sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|w_{1,j}^{(k)} - w_{1,j-1}^{(k)}\|]\,.$$

Based on Lemma C.15, we obtain

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|r_t^{(k)} - \bar{r}_t\|] \leq \lambda^{1+t}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|w_0^{(k)} - \bar{w}_0\|]$$
$$+ \gamma_z\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_2 g^{(k)}(x_{j-1}^{(k)}, z_{j-1}^{(k)}) - \nabla_2 g^{(k)}(x_{j-1}^{(k)}, z_{j-1}^{(k)}; \zeta_j^{(k)})\|]$$
$$+ \gamma_z\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|w_{1,j-1}^{(k)} - \nabla_2 g^{(k)}(x_{j-1}^{(k)}, z_{j-1}^{(k)})\|]$$
$$+ \frac{4\eta_x\sqrt{K}}{1-\lambda}\frac{L_g}{\delta}\sum_{j=1}^{t}\lambda^{t-j+1} + \frac{4\eta_z\sqrt{K}}{1-\lambda}\frac{L_g}{\delta}\sum_{j=1}^{t}\lambda^{t-j+1}\,. \tag{111}$$

Then, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|r_t^{(k)} - \bar{r}_t\|] \leq \frac{1}{T}\sum_{t=0}^{T-1}\lambda^{1+t}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|w_0^{(k)} - \bar{w}_0\|]$$
$$+ \gamma_z\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_2 g^{(k)}(x_{j-1}^{(k)}, z_{j-1}^{(k)}) - \nabla_2 g^{(k)}(x_{j-1}^{(k)}, z_{j-1}^{(k)}; \zeta_j^{(k)})\|]$$
$$+ \gamma_z\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}\frac{1}{\delta}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|w_{1,j-1}^{(k)} - \nabla_2 g^{(k)}(x_{j-1}^{(k)}, z_{j-1}^{(k)})\|]$$
$$+ \frac{4\eta_x\sqrt{K}}{1-\lambda}\frac{L_g}{\delta}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} + \frac{4\eta_z\sqrt{K}}{1-\lambda}\frac{L_g}{\delta}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}$$
$$\leq \frac{1}{T}\sum_{t=0}^{T-1}\lambda^{1+t}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|w_0^{(k)} - \bar{w}_0\|] + \gamma_z\sqrt{K}\sigma\frac{1}{\delta}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}$$
$$+ \gamma_z\sqrt{K}\frac{2\sqrt{2}\sigma}{B_0^{1-1/s}}\frac{1}{\delta}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}(1-\gamma_z)^{j-1} + 2\sqrt{2}\gamma_z^{2-1/s}\sqrt{K}\frac{1}{\delta}\sigma\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}$$
$$+ \frac{8\eta_x\sqrt{\gamma_z}\sqrt{K}}{(1-\lambda)}\frac{L_g}{\delta}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} + \sqrt{K}\frac{8\eta_z\sqrt{\gamma_z}}{(1-\lambda)}\frac{L_g}{\delta}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}$$
$$+ \frac{4\eta_x\sqrt{K}}{1-\lambda}\frac{L_g}{\delta}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1} + \frac{4\eta_z\sqrt{K}}{1-\lambda}\frac{L_g}{\delta}\frac{1}{T}\sum_{t=1}^{T-1}\sum_{j=1}^{t}\lambda^{t-j+1}$$

$$
\begin{aligned}
&\leq \frac{\lambda}{(1-\lambda)T}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{E}[\|w_0^{(k)}-\bar{w}_0\|]+\frac{\gamma_z\lambda\sqrt{K}}{1-\lambda}\frac{1}{\delta}\sigma \\
&+\frac{4\eta_x\lambda\sqrt{K}}{(1-\lambda)^2}\frac{L_g}{\delta}+\frac{4\eta_z\lambda\sqrt{K}}{(1-\lambda)^2}\frac{L_g}{\delta} \\
&+\frac{\lambda\sqrt{K}}{T(1-\lambda)}\frac{2\sqrt{2}\sigma}{B_0^{1-1/s}}\frac{1}{\delta}+\frac{2\sqrt{2}\gamma_z^{2-1/s}\lambda\sqrt{K}}{1-\lambda}\frac{1}{\delta}\sigma \\
&+\frac{8\eta_x\sqrt{\gamma_z}\lambda\sqrt{K}}{(1-\lambda)^2}\frac{L_g}{\delta}+\frac{8\eta_z\sqrt{\gamma_z}\lambda\sqrt{K}}{(1-\lambda)^2}\frac{L_g}{\delta}\,. \qquad (112)
\end{aligned}
$$

For $\sum_{k=1}^{K}\mathbb{E}[\|w_0^{(k)}-\bar{w}_0\|]$, we bound it as follows:

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}[\|w_0^{(k)}-\bar{w}_0\|]=\sum_{k=1}^{K}\mathbb{E}[\|\frac{1}{\delta}w_{1,0}^{(k)}-\frac{1}{K}\sum_{j=1}^{K}\frac{1}{\delta}w_{1,0}^{(j)}\|] \\
&=\frac{1}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_2 g^{(k)}(x_0^{(k)},z_0^{(k)};\zeta_0^{(k)})-\frac{1}{K}\sum_{j=1}^{K}\nabla_2 g^{(j)}(x_0^{(j)},z_0^{(j)};\zeta_0^{(j)})\|] \\
&\leq\frac{2}{\delta}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_2 g^{(k)}(x_0^{(k)},z_0^{(k)})\|]+\frac{4\sqrt{2}K}{B_0^{1-1/s}}\frac{1}{\delta}\sigma\,. \qquad (113)
\end{aligned}
$$

Finally, by substituting the above inequality into Eq. (112), the proof is complete.

□

C.6 Proof of Theorem 4.1

Proof. By plugging the inequalities in Lemmas C.8, C.7 into Lemma C.6, we obtain

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|] &\leq \frac{\mathbb{E}[\Phi(\bar{x}_0)-\Phi(\bar{x}_T)]}{\eta_x T}+2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\bar{x}_t)-\nabla\Phi_\delta(\bar{x}_t)\|] \\
&+\frac{4(\delta L_f+L_g)}{\mu}\frac{(\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_0,\bar{y}_0)-h_\delta^*(\bar{x}_0)]-\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_T,\bar{y}_T)-h_\delta^*(\bar{x}_T)])}{\eta_y T} \\
&+\frac{4L_g}{\mu}\frac{(\frac{1}{\delta}\mathbb{E}[g(\bar{x}_0,\bar{z}_0)-g^*(\bar{x}_0)]-\frac{1}{\delta}\mathbb{E}[g(\bar{x}_T,\bar{z}_T)-g^*(\bar{x}_T)])}{\eta_z T} \\
&+2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_1 f^{(k)}(x_t^{(k)},y_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}u_{1,t}^{(k)}\|] \\
&+\frac{2}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_1 g^{(k)}(x_t^{(k)},y_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}u_{2,t}^{(k)}\|] \\
&+\frac{2}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_1 g^{(k)}(x_t^{(k)},z_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}u_{3,t}^{(k)}\|] \\
&+\frac{8L_g}{\mu}\frac{1}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_2 g^{(k)}(x_t^{(k)},z_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}w_{1,t}^{(k)}\|] \\
&+\frac{8(\delta L_f+L_g)}{\mu}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_2 f^{(k)}(x_t^{(k)},y_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}v_{1,t}^{(k)}\|] \\
&+\frac{8(\delta L_f+L_g)}{\mu}\frac{1}{\delta}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_2 g^{(k)}(x_t^{(k)},y_t^{(k)})-\frac{1}{K}\sum_{k=1}^{K}v_{2,t}^{(k)}\|]
\end{aligned}
$$

$$
\begin{aligned}
&+\left(2(L_f+\frac{2L_g}{\delta})+\frac{8(\delta L_f+L_g)}{\mu}\left(L_f+\frac{L_g}{\delta}\right)+\frac{8L_g^2}{\mu}\frac{1}{\delta}\right)\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t-x_t^{(k)}\|]\\
&+\left(2(L_f+\frac{L_g}{\delta})+\frac{8(\delta L_f+L_g)}{\mu}\left(L_f+\frac{L_g}{\delta}\right)\right)\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t-y_t^{(k)}\|]\\
&+\left(2\frac{L_g}{\delta}+\frac{8L_g^2}{\mu}\frac{1}{\delta}\right)\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|z_t^{(k)}-\bar{z}_t\|]+\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t-p_t^{(k)}\|]\\
&+\frac{4(\delta L_f+L_g)}{\mu}\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{q}_t-q_t^{(k)}\|]+\frac{4L_g}{\mu}\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{r}_t-r_t^{(k)}\|]\\
&+\frac{\eta_x L_\Phi}{2}+\frac{1}{\delta}\frac{4\eta_x(\delta L_f+L_g)^2}{\mu}+\frac{1}{\delta}\frac{2\eta_y(\delta L_f+L_g)^2}{\mu}+\frac{1}{\delta}\frac{\eta_x^2}{\eta_y}\frac{2(\delta L_f+L_g)^2}{\mu}\\
&+\frac{1}{\delta}\frac{\eta_x^2}{\eta_y}\frac{2L_{h_\delta^*}(\delta L_f+L_g)}{\mu}+\frac{1}{\delta}\frac{4\eta_x L_g^2}{\mu}+\frac{1}{\delta}\frac{2\eta_z L_g^2}{\mu}+\frac{1}{\delta}\frac{\eta_x^2}{\eta_z}\frac{2L_g^2}{\mu}+\frac{1}{\delta}\frac{\eta_x^2}{\eta_z}\frac{2L_{g^*}L_g}{\mu}\,. \qquad (114)
\end{aligned}
$$

By plugging Lemmas C.22, C.23, C.24, C.25, C.26, C.27, C.28, C.29, C.30, C.31 into the above inequality, we have

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|] \le{}& \frac{\mathbb{E}[\Phi(\bar{x}_0)-\Phi(\bar{x}_T)]}{\eta_x T}+2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\bar{x}_t)-\nabla\Phi_\delta(\bar{x}_t)\|]\\
&+\frac{4(\delta L_f+L_g)}{\mu}\frac{(\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_0,\bar{y}_0)-h_\delta^*(\bar{x}_0)]-\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_T,\bar{y}_T)-h_\delta^*(\bar{x}_T)])}{\eta_y T}\\
&+\frac{4L_g}{\mu}\frac{(\frac{1}{\delta}\mathbb{E}[g(\bar{x}_0,\bar{z}_0)-g^*(\bar{x}_0)]-\frac{1}{\delta}\mathbb{E}[g(\bar{x}_T,\bar{z}_T)-g^*(\bar{x}_T)])}{\eta_z T}\\
&+\frac{2\lambda}{(1-\lambda)T}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\left(\mathbb{E}[\|\nabla_1 f^{(k)}(x_0,y_0)\|]+\frac{1}{\delta}\mathbb{E}[\|\nabla_1 g^{(k)}(x_0,y_0)\|]+\frac{1}{\delta}\mathbb{E}[\|\nabla_1 g^{(k)}(x_0,z_0)\|]\right)\\
&+\frac{4(\delta L_f+L_g)}{\mu}\frac{2\lambda}{(1-\lambda)T}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\left(\mathbb{E}[\|\nabla_2 f^{(k)}(x_0,y_0)\|]+\frac{1}{\delta}\mathbb{E}[\|\nabla_2 g^{(k)}(x_0,y_0)\|]\right)\\
&+\frac{4L_g}{\mu}\frac{2\lambda}{(1-\lambda)T}\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{1}{\delta}\mathbb{E}[\|\nabla_2 g^{(k)}(x_0,z_0)\|]\\
&+\left(\frac{4\sqrt{2}}{\gamma_x}\left(1+\frac{2}{\delta}\right)+\frac{2\sqrt{2}}{\gamma_y}\frac{8(\delta L_f+L_g)}{\mu}\left(1+\frac{1}{\delta}\right)+\frac{2\sqrt{2}}{\gamma_z}\frac{8L_g}{\mu}\frac{1}{\delta}\right)\frac{1}{TB_0^{1-1/s}}\sigma\\
&+\left(\left(1+\frac{2}{\delta}\right)+\frac{4(\delta L_f+L_g)}{\mu}\left(1+\frac{1}{\delta}\right)+\frac{4L_g}{\mu}\frac{1}{\delta}\right)\frac{6\sqrt{2}\lambda\sqrt{K}}{(1-\lambda)TB_0^{1-1/s}}\sigma\\
&+\left(\gamma_x^{1-1/s}\left(1+\frac{2}{\delta}\right)+\gamma_y^{1-1/s}\frac{4(\delta L_f+L_g)}{\mu}\left(1+\frac{1}{\delta}\right)+\gamma_z^{1-1/s}\frac{4L_g}{\mu}\frac{1}{\delta}\right)\frac{4\sqrt{2}}{K^{1-1/s}}\sigma\\
&+\left(\gamma_x^{2-1/s}\left(1+\frac{2}{\delta}\right)+\gamma_y^{2-1/s}\frac{4(\delta L_f+L_g)}{\mu}\left(1+\frac{1}{\delta}\right)+\gamma_z^{2-1/s}\frac{4L_g}{\mu}\frac{1}{\delta}\right)\frac{2\sqrt{2}\lambda\sqrt{K}}{1-\lambda}\sigma\\
&+\left(\gamma_x\left(1+\frac{2}{\delta}\right)+\gamma_y\frac{4(\delta L_f+L_g)}{\mu}\left(1+\frac{1}{\delta}\right)+\gamma_z\frac{4L_g}{\mu}\frac{1}{\delta}\right)\frac{\lambda\sqrt{K}}{1-\lambda}\sigma\\
&+\left(\eta_x\left(L_f+\frac{2L_g}{\delta}\right)+\eta_y\left(L_f+\frac{L_g}{\delta}\right)+\eta_z\frac{L_g}{\delta}\right)\frac{16}{(1-\lambda)\sqrt{\gamma_x K}}\\
&+(\eta_x+\eta_y)\frac{8(\delta L_f+L_g)}{\mu}\left(L_f+\frac{L_g}{\delta}\right)\frac{8}{(1-\lambda)\sqrt{\gamma_y K}}+(\eta_x+\eta_z)\frac{8L_g}{\mu}\frac{L_g}{\delta}\frac{8}{(1-\lambda)\sqrt{\gamma_z K}}
\end{aligned}
$$

$$
\begin{aligned}
&+\left(2(L_f+\frac{2L_g}{\delta})+\frac{8(\delta L_f+L_g)}{\mu}\left(L_f+\frac{L_g}{\delta}\right)+\frac{8L_g^2}{\mu}\frac{1}{\delta}\right)\frac{\eta_x\lambda}{1-\lambda}\\
&+\left(2(L_f+\frac{L_g}{\delta})+\frac{8(\delta L_f+L_g)}{\mu}\left(L_f+\frac{L_g}{\delta}\right)\right)\frac{\eta_y\lambda}{1-\lambda}+\left(2\frac{L_g}{\delta}+\frac{8L_g^2}{\mu}\frac{1}{\delta}\right)\frac{\eta_z\lambda}{1-\lambda}\\
&+\left(\eta_x\left(L_f+\frac{2L_g}{\delta}\right)+\eta_y\left(L_f+\frac{L_g}{\delta}\right)+\eta_z\frac{L_g}{\delta}\right)\frac{4(1+2\sqrt{\gamma_x})\lambda\sqrt{K}}{(1-\lambda)^2}\\
&+(\eta_x+\eta_y)\frac{4(\delta L_f+L_g)}{\mu}\left(L_f+\frac{L_g}{\delta}\right)\frac{4(1+2\sqrt{\gamma_y})\lambda\sqrt{K}}{(1-\lambda)^2}+(\eta_x+\eta_z)\frac{4L_g}{\mu}\frac{L_g}{\delta}\frac{4(1+2\sqrt{\gamma_z})\lambda\sqrt{K}}{(1-\lambda)^2}\\
&+\eta_x\left(\frac{L_\Phi}{2}+\frac{1}{\delta}\frac{4(\delta L_f+L_g)^2}{\mu}+\frac{1}{\delta}\frac{4L_g^2}{\mu}\right)+\eta_y\frac{1}{\delta}\frac{2(\delta L_f+L_g)^2}{\mu}+\eta_z\frac{1}{\delta}\frac{2L_g^2}{\mu}\\
&+\frac{\eta_x^2}{\eta_y}\frac{1}{\delta}\left(\frac{2(\delta L_f+L_g)^2}{\mu}+\frac{2L_{h_\delta^*}(\delta L_f+L_g)}{\mu}\right)+\frac{\eta_x^2}{\eta_z}\frac{1}{\delta}\left(\frac{2L_g^2}{\mu}+\frac{2L_{g^*}L_g}{\mu}\right). \qquad (115)
\end{aligned}
$$

Because $\kappa>1$, $1-\lambda<1$, $s\in(1,2]$, $L_\Phi=O(\ell\kappa^3)$, $L_{h_\delta^*}=O(\ell\kappa)$, and $L_{g^*}=O(\ell\kappa)$, by setting $\gamma_x=\gamma_y=\gamma_z=\gamma<1$, it can be simplified to the following inequality:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|]\le\frac{\mathbb{E}[\Phi(\bar{x}_0)-\Phi(\bar{x}_T)]}{\eta_xT}+2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\bar{x}_t)-\nabla\Phi_\delta(\bar{x}_t)\|]\\
&+\frac{4(\delta L_f+L_g)}{\mu}\frac{(\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_0,\bar{y}_0)-h_\delta^*(\bar{x}_0)]-\frac{1}{\delta}\mathbb{E}[h_\delta(\bar{x}_T,\bar{y}_T)-h_\delta^*(\bar{x}_T)])}{\eta_yT}\\
&+\frac{4L_g}{\mu}\frac{(\frac{1}{\delta}\mathbb{E}[g(\bar{x}_0,\bar{z}_0)-g^*(\bar{x}_0)]-\frac{1}{\delta}\mathbb{E}[g(\bar{x}_T,\bar{z}_T)-g^*(\bar{x}_T)])}{\eta_zT}\\
&+O\left(\frac{\lambda}{(1-\lambda)T}\right)\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\left(\mathbb{E}[\|\nabla_1f^{(k)}(x_0,y_0)\|]+\frac{1}{\delta}\mathbb{E}[\|\nabla_1g^{(k)}(x_0,y_0)\|]+\frac{1}{\delta}\mathbb{E}[\|\nabla_1g^{(k)}(x_0,z_0)\|]\right)\\
&+O\left(\frac{\lambda\kappa}{(1-\lambda)T}\right)\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\left(\mathbb{E}[\|\nabla_2f^{(k)}(x_0,y_0)\|]+\frac{1}{\delta}\mathbb{E}[\|\nabla_2g^{(k)}(x_0,y_0)\|]\right)\\
&+O\left(\frac{\lambda\kappa}{(1-\lambda)T}\right)\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\frac{1}{\delta}\mathbb{E}[\|\nabla_2g^{(k)}(x_0,z_0)\|]\\
&+O\left(\frac{\kappa}{\delta}\frac{1}{\gamma TB_0^{1-1/s}}\sigma\right)+O\left(\frac{\kappa}{\delta}\frac{\lambda\sqrt{K}}{(1-\lambda)TB_0^{1-1/s}}\sigma\right)\\
&+O\left(\frac{\kappa}{\delta}\frac{\gamma^{1-1/s}}{K^{1-1/s}}\sigma\right)+O\left(\frac{\kappa}{\delta}\frac{\gamma^{2-1/s}\lambda\sqrt{K}}{1-\lambda}\sigma\right)+O\left(\frac{\kappa}{\delta}\frac{\gamma\lambda\sqrt{K}}{1-\lambda}\sigma\right)\\
&+O\left(\frac{\kappa\ell}{\delta}\frac{\eta_x\lambda}{1-\lambda}\right)+O\left(\frac{\kappa\ell}{\delta}\frac{\eta_y\lambda}{1-\lambda}\right)+O\left(\frac{\kappa\ell}{\delta}\frac{\eta_z\lambda}{1-\lambda}\right)\\
&+O\left(\frac{\kappa\ell}{\delta}\frac{\eta_x}{(1-\lambda)\sqrt{\gamma K}}\right)+O\left(\frac{\kappa\ell}{\delta}\frac{\eta_y}{(1-\lambda)\sqrt{\gamma K}}\right)+O\left(\frac{\kappa\ell}{\delta}\frac{\eta_z}{(1-\lambda)\sqrt{\gamma K}}\right)\\
&+O\left(\frac{\kappa\ell}{\delta}\frac{\eta_x\lambda\sqrt{K}}{(1-\lambda)^2}\right)+O\left(\frac{\kappa\ell}{\delta}\frac{\eta_y\lambda\sqrt{K}}{(1-\lambda)^2}\right)+O\left(\frac{\kappa\ell}{\delta}\frac{\eta_z\lambda\sqrt{K}}{(1-\lambda)^2}\right)\\
&+O\left(\eta_x\kappa^3\ell\right)+O\left(\eta_x\frac{\kappa\ell}{\delta}\right)+O\left(\eta_y\frac{\kappa\ell}{\delta}\right)+O\left(\eta_z\frac{\kappa\ell}{\delta}\right)+O\left(\frac{\eta_x^2}{\eta_y}\frac{\kappa^2\ell}{\delta}\right)+O\left(\frac{\eta_x^2}{\eta_z}\frac{\kappa^2\ell}{\delta}\right). \qquad (116)
\end{aligned}
$$

By denoting $a = \frac{s}{s-1} \geq 2$ and setting

$$\eta_x = O\left(\frac{(1-\lambda)^2 K\epsilon^{a+2}}{(\kappa^3\ell)^{a/2+2}\kappa^{a/2-1}\sigma^{a/2}}\right), \quad \eta_y = \eta_x \frac{4(\delta L_f + L_g)}{\mu}, \quad \eta_z = \eta_x \frac{4L_g}{\mu},$$
$$\gamma_x = \gamma_y = \gamma_z = \gamma = O\left(\frac{K\epsilon^{2a}}{(\kappa^3\ell)^a \kappa^a \sigma^a}\right), \quad \delta = O\left(\frac{\epsilon}{\kappa^3\ell}\right),$$
$$T = 4\max\left\{1, \frac{(\kappa^3\ell)^{a/2+2}\kappa^{a/2-1}\sigma^{a/2}}{(1-\lambda)^2 K\epsilon^{a+3}}, M^{a/(a+1)}\right\}, \quad B_0 = \max\left\{1, \left(\frac{4M}{T}\right)^a\right\}, \tag{117}$$

where $M = \frac{\kappa\sigma}{\delta\epsilon}\left(\frac{1}{\gamma} + \frac{\lambda\sqrt{K}}{1-\lambda}\right)$ and $T \geq 4M^{a/(a+1)}$ ensures $B_0 \leq T$, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|] &\leq O(\epsilon) + \frac{1}{\delta}\frac{\mathbb{E}[(h_\delta(\bar{x}_0,\bar{y}_0) - h_\delta^*(\bar{x}_0)) - (h_\delta(\bar{x}_T,\bar{y}_T) - h_\delta^*(\bar{x}_T))]}{\eta_x T} \\
&\quad + \frac{1}{\delta}\frac{\mathbb{E}[(g(\bar{x}_0,\bar{z}_0) - g^*(\bar{x}_0)) - (g(\bar{x}_T,\bar{z}_T) - g^*(\bar{x}_T))]}{\eta_x T} \\
&\leq O(\epsilon) + \frac{1}{\delta}\frac{\mathbb{E}[h_\delta(\bar{x}_0,\bar{y}_0) - h_\delta^*(\bar{x}_0)]}{\eta_x T} + \frac{1}{\delta}\frac{\mathbb{E}[g(\bar{x}_0,\bar{z}_0) - g^*(\bar{x}_0)]}{\eta_x T}.
\end{aligned} \tag{118}$$

As for the last two terms of the upper bound in Eq. (118), both are affected by $\frac{1}{\delta}$. To avoid the degeneration of the convergence rate, we can provide good initial points (x_0, y_0) and (x_0, z_0) such that $\mathbb{E}[h_\delta(x_0,y_0) - h_\delta(x_0, y_\delta^*(x_0))] \leq \delta$ and $\mathbb{E}[g(x_0,z_0) - g(x_0,y^*(x_0))] \leq \delta$ to mitigate the adverse affect from $\frac{1}{\delta}$. Since both $h_\delta(x,y)$ and $g(x,z)$ satisfy the μ-PL condition with respect to the second variable, we can use a full gradient descent method to obtain such solutions, which has a linear convergence rate and therefore does not affect the other terms in Eq. (118). As a result, we can obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\bar{x}_t)\|] \leq O(\epsilon). \tag{119}$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: They reflect our paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 3.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: In Section 3, 4, and Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In the section of experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: In the section of experiments, but the code is not opened in the review phase.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: In the section of experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In the section of experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In the section of experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: In every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In the section of experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.