SMOOTHED GRADIENT METHOD FOR NONCONVEX FEDERATED STOCHASTIC BILEVEL OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, federated stochastic bilevel optimization has attracted increasing attention due to its wide range of applications in machine learning. To reduce the computational overhead associated with second-order Hessian and Jacobian matrices, several first-order methods have been proposed. However, existing methods typically impose restrictive assumptions on the lower-level function, suffer from a strong dependence on the condition number in their convergence rates, and require different learning-rate scales for variables across the upper- and lower-level problems, limiting their practical applicability and complicating hyperparameter tuning. To address these challenges, we propose a stochastic doubly smoothed gradient method for nonconvex federated stochastic bilevel optimization problems, which decouples the learning rates of upper- and lower-level variables and does not require a strongly-convex lower-level loss function. We establish rigorous theoretical guarantees for the proposed algorithm, demonstrating an improved convergence rate of $O(\kappa^{15/2}/\epsilon^5)$ and a communication complexity of $O(\kappa^4/\epsilon^3)$, where κ denotes the condition number and ϵ represents the solution accuracy. Notably, these bounds exhibit significantly better dependence on the condition number κ than those of existing methods. Extensive experiments validate the effectiveness of our algorithm.

1 INTRODUCTION

Federated bilevel optimization has emerged as an important paradigm for large-scale machine learning, with applications spanning model-agnostic meta-learning (Finn et al., 2017; Rajeswaran et al., 2019), hyperparameter optimization (Feurer & Hutter, 2019), neural architecture search (Liu et al., 2018; Zhang et al., 2021), and reinforcement learning (Li et al., 2024; Shen et al., 2025). In many modern scenarios, data are distributed across clients due to privacy and communication constraints, motivating the following federated stochastic bilevel optimization problem:

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

where K denotes the number of clients, $f^{(k)}(x, y^*(x)) = \mathbb{E}_\xi[f^{(k)}(x, y^*(x); \xi^{(k)})]$ is the upper-level objective, and $g^{(k)}(x, y) = \mathbb{E}_\xi[g^{(k)}(x, y; \xi^{(k)})]$ is the lower-level objective on the k-th client.

Solving Eq. (1) in federated learning setting is challenging because each client's upper-level objective depends on the global lower-level solution. Classical federated bilevel methods (Huang, 2022; Gao, 2022; Tarzanagh et al., 2022; Huang et al., 2023; Li et al., 2023) typically rely on implicit differentiation with costly Jacobian- and Hessian-related computations, resulting in substantial computational and communication overhead. Recent works (Yang et al., 2025; Zhang et al., 2025) alleviate this issue through fully first-order reformulations that avoid explicit second-order computations. A representative strategy (Kwon et al., 2023; Shen & Chen, 2023; Lu & Mei, 2026; 2024; Liu et al., 2024) uses a penalty-based formulation to transform the bilevel problem into an unconstrained minimax problem solvable using only first-order stochastic gradients.

Despite this progress, existing first-order federated bilevel methods have several limitations. First, they require different learning-rate scales for the upper- and lower-level variables, making hyperparameter tuning difficult and potentially destabilizing the coupled optimization dynamics. Second,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of our algorithm with existing stochastic first-order methods for bilevel optimization. NC-SC refers to the setting where the upper-level problem is nonconvex and the lower-level problem is strongly convex, while NC-PL refers to the setting where the lower-level problem satisfies the PL condition. ϵ measures the accuracy for stationary point, κ denotes the condition number; "·" indicates that the cited work does not report an explicit κ-dependence in closed form.

Algorithms	Assumption	Convergence Rate	Communication Complexity	Unified LR Scale
Single-machine Setting				
F^{2}BA(Kwon et al., 2023)	NC-SC	$O(\cdot/\epsilon^7), O(\cdot/\epsilon^5)$	–	✗
Prox-F^{2}BA(Kwon et al., 2024)	NC-PL	$O(\cdot/\epsilon^7), O(\cdot/\epsilon^5)$	–	✗
F^{2}BA(Chen et al., 2024)	NC-PL	$O(\kappa^{12}/\epsilon^6)$	–	✗
Federated Learning Setting				
MemFBO (Yang et al., 2025)	NC-SC	$O(\cdot/\epsilon^7)$	$O(\cdot/\epsilon^3)$	✗
FedSVRBGD-FO (Zhang et al., 2025)	NC-SC	$O(\kappa^{10}/\epsilon^5)$	$O(\kappa^9/\epsilon^4)$	✗
FedDS-BVRFO(Corollary 4.9)	NC-PL	$O(\kappa^{15/2}/\epsilon^5)$	$O(\kappa^4/\epsilon^3)$	✓

their theoretical guarantees generally rely on strong convexity of the lower-level objective, excluding a broad class of nonconvex problems. Third, their convergence and communication complexities have unfavorable dependence on the condition number κ. These limitations motivate more efficient and robust first-order algorithms for federated bilevel optimization.

To address these challenges, we revisit stochastic smoothing techniques (Zhang et al., 2020; Yang et al., 2022; Zheng et al., 2023), which can stabilize optimization, enable unified learning-rate scaling, and improve condition-number dependence in stochastic minimax optimization. Recent works extend smoothing to federated stochastic minimax optimization (Shen et al., 2024), stochastic multi-level compositional minimax problems (Zhang & Gao, 2025), and constrained bilevel problems (Lu, 2023; 2025). However, stochastic smoothing in federated bilevel optimization, particularly doubly smoothed schemes, remains largely unexplored. This raises the following question:

Can stochastic smoothing be effectively incorporated into federated bilevel optimization to simultaneously improve theoretical guarantees and practical performance?

However, incorporating smoothing into federated bilevel optimization is nontrivial. Unlike standard minimax problems, federated bilevel optimization involves tightly coupled upper- and lower-level variables together with a global lower-level problem. Moreover, the penalty-based first-order reformulation introduces additional interactions between the penalty and smoothing parameters, making the appropriate smoothing strategy unclear. Stochastic gradient estimation and periodic communication further complicate the algorithm design. Theoretically, establishing convergence under the relaxed PL condition while improving condition-number dependence poses additional challenges.

To address these challenges, we propose a stochastic *doubly smoothed* framework for federated bilevel optimization. Our method integrates smoothing into a penalty-based first-order reformulation for *both the upper- and lower-level variables*, yielding a fully first-order algorithm for solving Eq. (1). This formulation enables unified learning-rate scaling without explicit Jacobian or Hessian computations. Moreover, our analysis allows lower-level objectives satisfying the PL condition with improved condition-number dependence. Our main contributions are summarized as follows:

- **Algorithmic contribution:** We propose a stochastic *doubly smoothed* framework for federated bilevel optimization and develop a fully first-order algorithm, termed **FedDS-BVRFO**. By smoothing both the upper- and lower-level updates, our method enables the same order of learning rates with respect to the condition number, simplifying hyperparameter tuning. To the best of our knowledge, this is the first stochastic doubly smoothed method for federated bilevel optimization with unified learning-rate scaling.
- **Theoretical contribution:** We establish convergence guarantees under the PL condition on the lower-level objective, relaxing the strong-convexity assumption of existing federated bilevel methods. Our method achieves a sample complexity of $O(\kappa^{15/2}/(K\epsilon^5))$ and communication complexity of $O(\kappa^4/\epsilon^3)$, improving the condition-number dependence over existing fully first-order federated bilevel methods while preserving linear speedup in the number of clients.

Finally, extensive experimental results validate the effectiveness of the proposed algorithm.

2 Related Work

2.1 Federated and Stochastic Bilevel Optimization

Federated stochastic bilevel optimization has recently attracted growing attention due to its applications in privacy-preserving meta-learning and hyperparameter optimization (Chen et al., 2018; Jiang et al., 2019; Fallah et al., 2020; Yang et al., 2023a). Compared with classical single-machine bilevel optimization (Ghadimi & Wang, 2018; Ji et al., 2021; Chen et al., 2021), the federated setting introduces additional challenges due to the coupling between local upper-level objectives and a global lower-level problem. A primary line of work is based on approximate implicit differentiation techniques, such as Neumann-series-based estimators, leading to representative methods including FedNEST and its variants (Tarzanagh et al., 2022; Huang et al., 2023; Yang et al., 2023b). These approaches are further improved by incorporating momentum or variance-reduction strategies (Gao, 2022; Li et al., 2023) to enhance communication efficiency and training stability. More recently, fully first-order methods have been proposed via penalty-based Lagrangian reformulations (Yang et al., 2025; Zhang et al., 2025), which enable gradient-based updates without explicit second-order computations. In addition, several works have investigated nonconvex lower-level problems, primarily in the single-machine setting (Shen & Chen, 2023; Kwon et al., 2024; Chen et al., 2024; Liu et al., 2024; Lu, 2023; 2025). These studies relax strong convexity assumptions on the lower-level problem, but their extension to federated settings remains largely unexplored.

2.2 Smoothed Techniques

Stochastic smoothing has been widely studied for stabilizing optimization dynamics and improving convergence behavior. Zhang et al. (2020) first introduced it into stochastic minimax optimization to mitigate gradient oscillations and ensure convergence to stationary solutions. Under the PL condition, Yang et al. (2022) further showed that smoothing improves the condition-number dependence from $O(\kappa^3)$ to $O(\kappa^2)$. Subsequent works extend smoothing to more complex settings. In particular, Zheng et al. (2023) proposed doubly smoothed schemes for nonconvex–nonconcave problems, while Shen et al. (2024) incorporated smoothing into federated stochastic minimax optimization, achieving an $O(\kappa^2)$ dependence. More recently, Zhang & Gao (2025) extended smoothing to stochastic multi-level compositional minimax problems, achieving an improved $O(\kappa^{3/2})$ dependence when combined with variance reduction. Beyond minimax optimization, Lu (2023; 2025) incorporated smoothing into Lagrangian methods for constrained bilevel optimization. Note that our reformulation differs from those in recent works (Lu, 2023; 2025). Specifically, Lu (2023; 2025) apply the smoothing technique only for the primal variable, while our method apply it to both the primal and dual variables in the reformulated minimax problem. Despite these advances, the role of stochastic smoothing in federated bilevel optimization, particularly the use of doubly smoothed schemes, remains largely unexplored.

3 Method

3.1 Doubly Smoothed Bilevel Objective

Following the fully first-order reformulation in Yang et al. (2025); Zhang et al. (2025), we introduce a penalty parameter $\rho > 0$ to transform the federated bilevel problem in Eq. (1) into the following unconstrained minimax formulation:

$$\min_{x,y} \max_{z} \underbrace{\frac{1}{K}\sum_{k=1}^{K} f_\rho^{(k)}(x,y,z)}_{f_\rho(x,y,z)}, \quad f_\rho^{(k)}(x,y,z) = f^{(k)}(x,y) + \frac{1}{\rho}(g^{(k)}(x,y) - g^{(k)}(x,z)). \quad (2)$$

At optimality, the maximizer with respect to $z \in \mathbb{R}^{d_y}$ coincides with the lower-level solution $y^*(x)$. Importantly, the gradients with respect to x, y, and z involve only first-order information, thereby avoiding explicit Jacobian and Hessian computations and enabling more efficient optimization in federated learning settings.

Motivated by smoothing techniques in classical stochastic minimax optimization (Zhang et al., 2020; Yang et al., 2022; Zheng et al., 2023), we introduce auxiliary proximal variables $p \in \mathbb{R}^{d_x}$ and $q \in \mathbb{R}^{d_y}$ for the primal updates x and y, respectively, and an auxiliary proximal variables $r \in \mathbb{R}^{d_y}$

for the dual maximizer z. We then construct the following doubly smoothed objective:

$$F(x,y,z;p,q,r) = \rho\frac{1}{K}\sum_{k=1}^{K} f_\rho^{(k)}(x,y,z) + \underbrace{\frac{\gamma_1}{2}(\|x-p\|^2+\|y-q\|^2)}_{\textbf{primal smoothing on } (x,y)} - \underbrace{\frac{\gamma_2}{2}\|z-r\|^2}_{\textbf{dual smoothing on } z}, \quad (3)$$

where $\gamma_1 > 0$ and $\gamma_2 > 0$ are the primal and dual smoothing parameters.

The term $\rho f_\rho^{(k)}(x,y,z)$ is L_ρ-smooth with respect to x and y, where $L_\rho \triangleq \rho L_f + 2L_g$. By choosing $\gamma_1 = O(L_\rho)$ with $\gamma_1 > L_\rho$, the function $F(x,y,z;p,q,r)$ is $(\gamma_1 - L_\rho)$-strongly convex and $(\gamma_1 + L_\rho)$-smooth in x and y. In particular, the strong convexity and smoothness constants are given by $\mu_F = (\gamma_1 - L_\rho)$ and $L_F = (\gamma_1 + L_\rho)$, respectively. Similarly, by smoothing the dual variable z with $\gamma_2 > L_\rho$, $F(x,y,z;p,q,r)$ is $(\gamma_2 - L_\rho)$-strongly concave and $(\gamma_2 + L_\rho)$-smooth with respect to z.

By rearranging the above expression, the doubly smoothed objective can be written in the following decomposed form:

$$F(x,y,z;p,q,r) = \frac{1}{K}\sum_{k=1}^{K}\left(\rho f^{(k)}(x,y) + g_{\gamma_1}^{(k)}(x,y;p,q) - g_{\gamma_2}^{(k)}(x,z;r)\right), \quad (4)$$

where the smoothed lower-level component is defined as

$$\begin{aligned} g_{\gamma_1}^{(k)}(x,y;p,q) &= g^{(k)}(x,y) + \frac{\gamma_1}{2}(\|x-p\|^2+\|y-q\|^2), \\ g_{\gamma_2}^{(k)}(x,z;r) &= g^{(k)}(x,z) + \frac{\gamma_2}{2}\|z-r\|^2. \end{aligned} \quad (5)$$

Here, $g_{\gamma_1}^{(k)}$ smooths the primal minimization updates (x,y), whereas $g_{\gamma_2}^{(k)}$ smooths the dual maximization update z. This simultaneous primal-dual smoothing is the key feature of our doubly smoothed formulation, and yields a well-conditioned strongly-convex-strongly-concave surrogate that supports unified learning-rate scaling.

Algorithm 1 FedDS-BVRFO

Input: $\alpha_x > 0, \alpha_y > 0, \alpha_z > 0, \alpha_p > 0, \alpha_q > 0, \alpha_r > 0, \beta_x > 0, \beta_y > 0, \beta_z > 0, \eta > 0$.
Initialization on each worker $k \in \{1,\cdots,K\}$: $x_0^{(k)} = x_0$, $y_0^{(k)} = y_0$, $z_0^{(k)} = z_0$,
for $t = 0,\cdots,T-1$ **do**
  Update variables x, y, and z via Eq. (10)
  Update variables p, q, and r via Eq. (11)
  Update gradients u, v, and w via Eq. (9)
  if $\mathrm{mod}(t+1,\tau) == 0$ **then**
    Communication via Eq. (12)
  end if
end for

3.2 Proposed Algorithm

To solve Eq. (4), we propose a new algorithm termed Federated doubly smoothed Bilevel Variance-Reduced First-Order (FedDS-BVRFO), summarized in Algorithm 1. The full algorithmic description is provided in the Appendix B.

For the smoothed objective in Eq. (4), the partial gradients are given by

$$\begin{aligned} \nabla_x F(x_t,y_t,z_t;p_t,q_t,r_t) &= \rho\nabla_1 f(x_t,y_t) + (\nabla_1 g_{\gamma_1}(x_t,y_t;p_t,q_t) - \nabla_1 g_{\gamma_2}(x_t,z_t;r_t)), \\ \nabla_y F(x_t,y_t,z_t;p_t,q_t,r_t) &= \rho\nabla_2 f(x_t,y_t) + \nabla_2 g_{\gamma_1}(x_t,y_t;p_t,q_t), \\ \nabla_z F(x_t,y_t,z_t;p_t,q_t,r_t) &= -\nabla_2 g_{\gamma_2}(x_t,z_t;r_t), \end{aligned} \quad (6)$$

where

$$\nabla_1 g_{\gamma_1}(x_t,y_t;p_t,q_t) = \nabla_1 g(x_t,y_t) + \gamma_1(x_t - p_t),$$

$$\nabla_2 g_{\gamma_1}(x_t, y_t; p_t, q_t) = \nabla_2 g(x_t, y_t) + \gamma_1(y_t - q_t) ,$$
$$\nabla_1 g_{\gamma_2}(x_t, z_t; r_t) = \nabla_1 g(x_t, z_t) ,$$
$$\nabla_2 g_{\gamma_2}(x_t, z_t; r_t) = \nabla_2 g(x_t, z_t) + \gamma_2(z_t - r_t) . \tag{7}$$

To construct variance-reduced estimators, we decompose the gradient into multiple components and maintain separate momentum variables. Specifically, $u_{1,t}$, $u_{2,t}$, and $u_{3,t}$ track the three components of the x-gradient, corresponding to $\nabla_1 f(x_t, y_t)$, $\nabla_1 g_\gamma(x_t, y_t, p_t, q_t)$ and $\nabla_1 g(x_t, z_t)$, respectively. Similarly, $v_{1,t}$ and $v_{2,t}$ track the two components of the y-gradient, while w_t tracks the z-gradient.

On each worker k, we adopt STORM-style variance-reduced updates (Cutkosky & Orabona, 2019) for each gradient component, including $u_{1,t+1}^{(k)}$, $u_{2,t+1}^{(k)}$, $u_{3,t+1}^{(k)}$, $v_{1,t+1}^{(k)}$, $v_{2,t+1}^{(k)}$, $w_{t+1}^{(k)}$. For illustration, we present the update for $u_{1,t+1}^{(k)}$ as follows:

$$u_{1,t+1}^{(k)} = (1 - \beta_x \eta^2)(u_{1,t}^{(k)} - \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_{t+1}^{(k)})) + \nabla_1 f^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; \xi_{t+1}^{(k)}) . \tag{8}$$

Thus, the final variance-reduced momentum estimators for the x-, y-, and z-updates are constructed as:

$$u_{t+1}^{(k)} = \rho u_{1,t+1}^{(k)} + (u_{2,t+1}^{(k)} - u_{3,t+1}^{(k)}) , \quad v_{t+1}^{(k)} = \rho v_{1,t+1}^{(k)} + v_{2,t+1}^{(k)} , \quad w_{t+1}^{(k)} = w_{t+1}^{(k)} . \tag{9}$$

After obtaining these estimators, the local model variables are updated as

$$x_{t+1}^{(k)} = x_t^{(k)} - \alpha_x \eta u_t^{(k)} , \quad y_{t+1}^{(k)} = y_t^{(k)} - \alpha_y \eta v_t^{(k)} , \quad z_{t+1}^{(k)} = z_t^{(k)} - \alpha_z \eta w_t^{(k)} . \tag{10}$$

In addition, the auxiliary smoothing variables $p_{t+1}^{(k)}$, $q_{t+1}^{(k)}$, and $r_{t+1}^{(k)}$ are updated as:

$$p_{t+1}^{(k)} = p_t^{(k)} + \alpha_p \eta (x_{t+1}^{(k)} - p_t^{(k)}) , \quad q_{t+1}^{(k)} = q_t^{(k)} + \alpha_q \eta (y_{t+1}^{(k)} - q_t^{(k)}) ,$$
$$r_{t+1}^{(k)} = r_t^{(k)} + \alpha_r \eta (z_{t+1}^{(k)} - r_t^{(k)}) . \tag{11}$$

To improve communication efficiency, we adopt a periodic synchronization strategy. Specifically, every τ iterations, each local variable is replaced by the average across all participating workers. This includes the model variables $x_{t+1}^{(k)}$, $y_{t+1}^{(k)}$, $z_{t+1}^{(k)}$; the auxiliary variables $p_{t+1}^{(k)}$, $q_{t+1}^{(k)}$, and $r_{t+1}^{(k)}$; and the variance-reduced momentum estimators $u_{t+1}^{(k)}$, $v_{t+1}^{(k)}$, $w_{t+1}^{(k)}$. Specifically, for any $S \in \{x, y, z, p, q, r, u, v, w\}$, the synchronization step is given by

$$S_{t+1}^{(k)} = \frac{1}{K} \sum_{k'=1}^{K} S_{t+1}^{(k')} . \tag{12}$$

4 Theoretical Analysis

4.1 Assumptions

We adopt the following standard assumptions commonly used in federated stochastic bilevel optimization (Kwon et al., 2023; 2024; Chen et al., 2024; Zhang et al., 2025; Yang et al., 2025).

Assumption 4.1. *For $k \in \{1, \cdots, K\}$, $f^{(k)}(\cdot, \cdot)$ satisfies the mean-squared smoothness with the constant $L_f > 0$ and its second-order gradient satisfies the Lipschitz continuity with the constant $\ell_f > 0$. In addition, the partial gradient with respect to the second argument is bounded as $\|\nabla_2 f^{(k)}(\cdot, \cdot)\| \leq C_f$, where $C_f > 0$ is a constant.*

Assumption 4.2. *For $k \in \{1, \cdots, K\}$, $g^{(k)}(\cdot, \cdot)$ satisfies the mean-squared smoothness with the constant $L_g > 0$, and its second-order gradient satisfies the Lipschitz continuity with the constant $\ell_g > 0$.*

Assumption 4.3. *For $k \in \{1, \cdots, K\}$, $g^{(k)}(\cdot, \cdot)$ satisfies the μ-PL condition with respect to the second argument with the constant $\mu > 0$. In addition, $h_\rho(\cdot, \cdot) = \rho f(\cdot, \cdot) + g(\cdot, \cdot)$ also satisfies the μ-PL condition with respect to the second argument.*

Note that most federated bilevel optimization methods in the literature (Tarzanagh et al., 2022; Huang et al., 2023; Gao, 2022; Li et al., 2023; Zhang et al., 2025; Yang et al., 2025) rely on the strong convexity of $g^{(k)}(\cdot, \cdot)$ with respect to the second argument. In contrast, our federated analysis assumes only a μ-PL condition on the lower-level objective, which has recently been studied in single-machine bilevel optimization. Since the PL condition is strictly weaker than strong convexity, our framework applies to a broader range of practical scenarios.

Assumption 4.4. *For $k \in \{1, \cdots, K\}$, the stochastic gradients of $f^{(k)}(\cdot,\cdot)$ and $g^{(k)}(\cdot,\cdot)$ have variances upper bounded by σ^2 where $\sigma > 0$ is a constant.*

Assumption 4.5. *The data distribution between workers is heterogeneous and the upper-level and lower-level gradients satisfy the following conditions:*

$$\frac{1}{K}\sum_{k=1}^{K}\|\nabla f^{(k)}(x,y) - \frac{1}{K}\sum_{k'=1}^{K}\nabla f^{(k')}(x,y)\|^2 \le \delta^2\,,$$
$$\frac{1}{K}\sum_{k=1}^{K}\|\nabla g^{(k)}(x,y) - \frac{1}{K}\sum_{k'=1}^{K}\nabla g^{(k')}(x,y)\|^2 \le \delta^2\,, \tag{13}$$

where $\delta > 0$ is a constant, $x \in \mathbb{R}^{d_x}$, and $y \in \mathbb{R}^{d_y}$.

Definition 4.6. *Given Assumptions 4.1-4.3, we define $\ell = \max\{C_f, L_f, \ell_f, L_g, \ell_g\}$ and the condition number $\kappa = \ell/\mu$.*

Definition 4.7. *(Yang et al., 2022) (α,θ) is an (ϵ_1,ϵ_2)-stationary point of $f(\cdot,\cdot)$, if $\|\nabla_\alpha f(\alpha,\theta)\| \le \epsilon_1$ and $\|\nabla_\theta f(\alpha,\theta)\| \le \epsilon_2$. α is an ϵ-stationary point of Φ if $\|\nabla\Phi(\alpha)\| \le \epsilon$, where $\Phi(\alpha) = f(\alpha,\theta^*(\alpha))$ and $\theta^*(\alpha) = \arg\max_{\theta\in\mathbb{R}^{d_2}} f(\alpha,\theta)$.*

4.2 Convergence Rate

Based on Assumptions 4.1-4.5, we establish the convergence rate of our algorithm below:

Theorem 4.8. *Given Assumptions 4.1-4.5, when $\beta_x > 0$, $\beta_y > 0$, $\beta_z > 0$, $\gamma_1 = O(L_\rho)$, $\gamma_2 = O(L_\rho)$, the condition about α_x in Eq. (56), and*

$$\alpha_y = \alpha_x\,,\quad \alpha_z = \alpha_x \underbrace{\frac{(\gamma_1 - L_\rho)^2}{72L_g^2}}_{c_{\alpha_z}=O(1)}\,,\quad \alpha_p = \alpha_q = \alpha_x \underbrace{\frac{(\gamma_1 - L_\rho)^4(\gamma_2 - L_\rho)^2\mu}{72\times 768\gamma_1 L_g^2(5L_g^2\mu + 16\gamma_2^2(\gamma_1 - L_\rho))}}_{c_{\alpha_{pq}}=O(1/\kappa)}\,,$$
$$\alpha_r = \alpha_x \underbrace{\frac{(\gamma_1 - L_\rho)^4(\gamma_2 - L_\rho)^4}{72\times 24\times 640\gamma_1\gamma_2^2 L_g^4}}_{c_{\alpha_r}=O(1)}\,,$$

Algorithm 1 has the following convergence upper bound:

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}[\|\nabla_x f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2] + \mathbb{E}[\|\nabla_y f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2] + \kappa\mathbb{E}[\|\nabla_z f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2]\right)$$
$$\le O\Big(\frac{\kappa}{\alpha_x\eta\rho T}\Big) + O\Big(\frac{\kappa}{\beta\eta^2\rho^2 BKT}\Big) + O\Big(\frac{\kappa\beta\eta^2}{\rho^2 K}\Big) + O\Big(\frac{\kappa\tau^2\beta\eta^4\alpha_x^2}{\rho^2 K}\Big) + O\Big(\frac{\kappa\tau^4\alpha_x^2\beta^2\eta^6}{\rho^2}\Big)\,. \tag{14}$$

Note that the convergence rate is established with respect to the original federated bilevel objective in Eq. (2), rather than the smoothed objective in Eq. (4).

Corollary 4.9. *Given Assumptions 4.1-4.5, by setting*

$$\rho = O\left(\frac{\epsilon}{\kappa^3}\right)\,,\quad \alpha_x = O(1)\,,\quad \beta_x = O\left(\frac{1}{K}\right)\,,\quad \beta_y = O\left(\frac{1}{K}\right)\,,\quad \beta_z = O\left(\frac{1}{K}\right)\,,$$
$$\eta = O\left(\frac{K\epsilon^2}{\kappa^{7/2}}\right)\,,\quad \tau = O\left(\frac{\kappa^{7/2}}{K\epsilon^2}\right)\,,\quad B = O\left(\frac{\kappa^{13/2}}{K\epsilon^3}\right)\,,\quad T = O\left(\frac{\kappa^{15/2}}{K\epsilon^5}\right)\,. \tag{15}$$

we obtain an $(\epsilon, \epsilon/\sqrt{\kappa})$-stationary point, in the sense that

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(\rho^2\mathbb{E}[\|\nabla_x f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2] + \rho^2\mathbb{E}[\|\nabla_y f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2] + \kappa\rho^2\mathbb{E}[\|\nabla_z f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2]\right) \le O(\epsilon^2)\,.$$

Note that B denotes only the initial batch size used in the first iteration; throughout the subsequent training process, the algorithm operates with an $O(1)$ batch size.

Remark 4.10. ***(Stationarity translation.)*** *Following Yang et al. (2022), we need to translate the $(\epsilon, \epsilon/\sqrt{\kappa})$-stationary solution to the ϵ-stationary solution. The complete translation procedure and analysis are provided in Appendix G. Importantly, after this translation, the iteration complexity to find the ϵ-stationary solution is still $T = O\left(\frac{\kappa^{15/2}}{K\epsilon^5}\right)$.*

Remark 4.11. ***(Better convergence rate under milder assumption.)*** *The iteration complexity $T = O\left(\frac{\kappa^{15/2}}{K\epsilon^5}\right)$ indicates that the proposed method enjoys linear speedup with respect to the number of clients K. Compared with existing first-order federated bilevel methods, our result improves upon the $O(\frac{1}{K\epsilon^7})$ complexity established in Yang et al. (2025) and matches the optimal $O(\frac{1}{K\epsilon^5})$ dependence reported in Zhang et al. (2025). In terms of condition-number dependence κ, our method achieves a complexity of $O(\kappa^{15/2})$, which significantly improves over the $O(\kappa^{10})$ bounds in Zhang et al. (2025), and $O(\kappa^{12})$ bounds in Chen et al. (2024), respectively. Moreover, this result is established under a milder assumption in federated bilevel optimization, requiring only a PL condition on the lower-level function rather than assuming strong convexity as in Yang et al. (2025); Zhang et al. (2025).*

Remark 4.12. ***(Better communication complexity.)*** *Regarding communication efficiency, Algorithm 1 requires $T/\tau = O\left(\kappa^4/\epsilon^3\right)$ communication rounds. This significantly improves upon the $O\left(\kappa^9/\epsilon^4\right)$ communication complexity reported in Zhang et al. (2025), in terms of both the condition-number dependence κ and accuracy ϵ.*

Remark 4.13. ***(Unified learning rate scale.)*** *Existing first-order federated bilevel methods (Yang et al., 2025; Zhang et al., 2025) typically require multi-scale learning rates for different variables in the upper- and lower-level problems, with the step sizes tightly coupled by the condition number κ. For instance, in Zhang et al. (2025), the prescribed learning-rate ratios between x and y, as well as between x and z, are on the order of $O(1/\kappa^2)$. By contrast, the proposed smoothed formulation enables the updates of x, y, and z to operate under the same order of step sizes with respect to κ. This unified scaling substantially simplifies hyperparameter tuning and improves practical robustness.*

The proof sketch and full proof are provided in Appendices C – F.

5 Experiments

5.1 Nonconvex Hyperparameter Optimization

We consider a nonconvex hyperparameter optimization problem with objective defined in Eq. (16). In this bilevel formulation, the lower-level problem trains a two-layer fully connected neural network, while the upper-level problem tunes the hyperparameters. Although the lower-level objective is nonconvex, prior work shows that overparameterized neural networks can satisfy the PL condition (Liu et al., 2022). Specifically, the lower-level problem optimizes network parameters $y = \{y_1, y_2\}$ using training data across clients, whereas the upper-level problem updates hyperparameters $x = \{x_1, x_2\}$ based on validation loss. The resulting bilevel formulation is shown as follows:

$$\begin{aligned} \min_{x=\{x_1,x_2\}} \quad & \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}^{(k)}(y^*(x); \mathcal{D}_{val}^{(k)}) \\ s.t. \quad & y^*(x) = \arg\min_{y=\{y_1,y_2\}} \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}^{(k)}(y; \mathcal{D}_{tr}^{(k)}) + \mathcal{R}_1(x) + \mathcal{R}_2(x)\,. \end{aligned} \tag{16}$$

Here, $y_1 = [y_{1,pq}] \in \mathbb{R}^{d_1 \times d_2}$ and $y_2 = [y_{2,pq}] \in \mathbb{R}^{d_2 \times d_3}$ denote the weight matrices of the first and second network layers. The hyperparameters $x_1 = [x_{1,q}] \in \mathbb{R}^{d_2}$ and $x_2 = [x_{2,q}] \in \mathbb{R}^{d_3}$ control the regularization terms: $\mathcal{R}_1(x) = \frac{1}{d_2}\sum_{q=1}^{d_2} \exp(x_{1,q}) \frac{1}{d_1}\sum_{p=1}^{d_1} y_{1,pq}^2$ and $\mathcal{R}_2(x) = \frac{1}{d_3}\sum_{q=1}^{d_3} \exp(x_{2,q}) \frac{1}{d_2}\sum_{p=1}^{d_2} y_{2,pq}^2$. In our experiments, d_1 is set to the number of input features, $d_2 = 20$, and $d_3 = 1$ for binary classification.

Since the lower-level objective in Eq. (16) is nonconvex, existing federated bilevel methods do not have convergence guarantees and cannot directly apply to this setting. In particular, both second-order methods (Tarzanagh et al., 2022; Huang et al., 2023; Gao, 2022; Li et al., 2023) and fully first-order methods (Zhang et al., 2025; Yang et al., 2025) rely critically on the strong convexity assumption for the lower-level problem. To enable a fair and meaningful comparison, we therefore benchmark against algorithms specifically designed for nonconvex bilevel optimization, most notably F^{2}BSA (Kwon et al., 2024), adapted to the federated environment. We evaluate both the

double-loop and single-loop variants. For the double-loop version, we set the number of inner iterations to 1, 5, and 10, denoted as **Fed-F^{2}BSA-1**, **Fed-F^{2}BSA-5**, and **Fed-F^{2}BSA-10**, respectively. We also include the variance-reduced single-loop variant, denoted as **Fed-F^{2}BSA-VR**, which incorporates a STORM-type estimator. All baseline learning rates follow the guidelines in Corollaries 5.2 and 5.5 of Kwon et al. (2024), which require different learning-rate scales across variables.

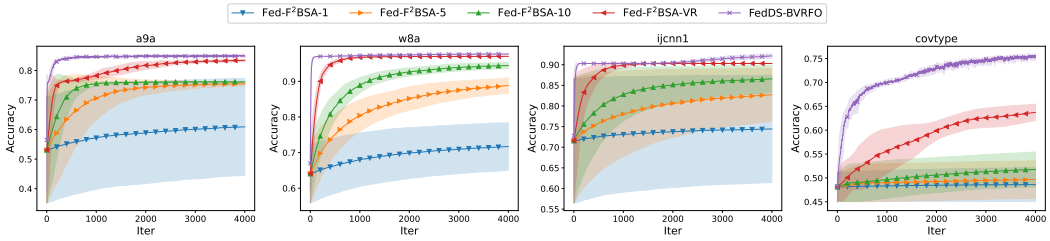


Figure 1: The test accuracy versus the training epochs, $\tau = 4$, nonconvex.

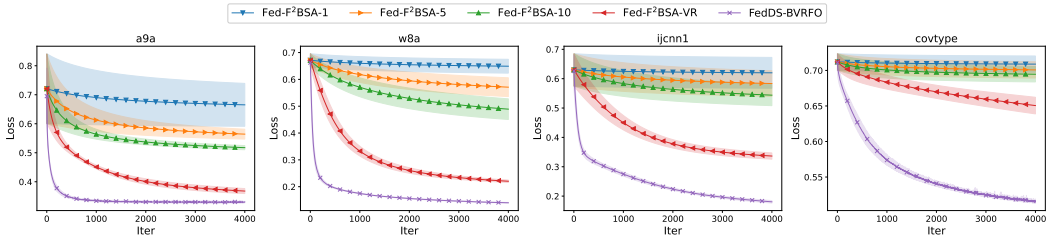
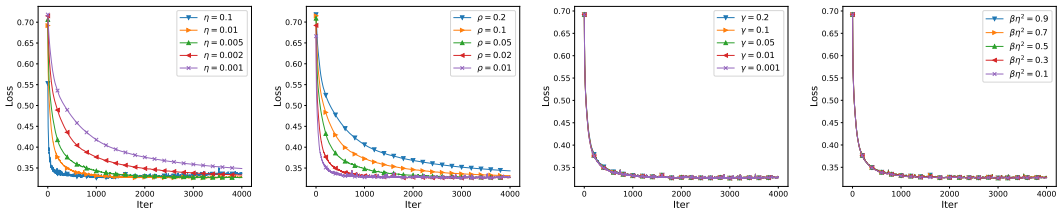


Figure 2: The upper-level loss function value versus the training epochs, $\tau = 4$, nonconvex.

We evaluate our method on four standard LIBSVM datasets: a9a, w8a, ijcnn1, and covtype[1]. For each dataset, we randomly reserve 10% of the samples for testing, and split the remaining data into training and validation sets with a $70\%/30\%$ ratio. To simulate the federated setting, both training and validation data are uniformly distributed across eight workers. Each worker uses a mini-batch size of 10. For our method, we set $\epsilon = 0.1$, $\rho = \epsilon/5$. The learning rates are chosen according to Corollary 4.9, using a unified scale for both upper- and lower-level variables. For all methods, we conduct experiments over five random seeds and report the mean and standard deviation.

When the communication period is set to $\tau = 4$, the results in Figure 1 and Figure 2 show that the proposed method consistently achieves higher test accuracy and lower upper-level objective values than all baselines, indicating faster and more stable convergence. Moreover, its smaller standard deviations across five random seeds indicate reduced sensitivity to stochastic variation. We also observe that the double-loop variants of Fed-F^{2}BSA are sensitive to the number of inner iterations: increasing the inner-loop length improves performance but incurs higher computational cost. Although the variance-reduced single-loop variant Fed-F^{2}BSA-VR alleviates this issue, it remains inferior to our method. These results highlight the benefit of the proposed method in achieving efficient and stable convergence without additional inner-loop updates. Additional results for $\tau = 16$ are provided in Appendix A.



(a) Different η (b) Different ρ (c) Different γ (d) Different $\beta\eta^2$

Figure 3: The upper-level loss function value versus the training epochs on a9a, $\tau = 16$, nonconvex.

We further study the sensitivity of our method to several hyperparameters, including the learning rate η, penalty parameter ρ, smoothing parameter $\gamma := \gamma_1 = \gamma_2$, and the momentum coefficient

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

$\beta\eta^2$, where $\beta := \beta_x = \beta_y = \beta_z$, as shown in Figure 3. Figure 3(a) indicates that a larger η accelerates convergence but may cause instability, while Figure 3(b) shows that a smaller ρ (i.e., a larger penalty $1/\rho$) improves convergence speed. Moreover, Figures 3(c) and 3(d) indicate that our method is relatively robust to γ and β, suggesting that the proposed smoothing and momentum parameters do not require extensive tuning in practice.

Finally, additional comparisons with federated bilevel baselines on nonconvex–strongly-convex tasks are reported in Appendix A.

5.2 Model Pruning

To provide additional empirical evidence, we evaluate our method on a structured parameter reduction problem, namely model pruning (Zhang et al., 2022). This task aims to identify a sparse subnetwork by selectively disabling redundant weights via a binary selection mask. In this framework, the sparsity pattern and the parameter retraining process are coupled through a bilevel formulation: the lower-level problem optimizes the network weights under a fixed selection pattern, while the upper-level problem determines which parameters should be retained and which should be pruned. The corresponding formulation is given by

$$\min_x \underbrace{\frac{1}{K}\sum_{k=1}^{K}\ell^{(k)}(x \odot y^*(x))}_{\text{Pruning task}} \quad s.t. \quad \underbrace{y^*(x) = \arg\min_y \frac{1}{K}\sum_{k=1}^{K}\ell^{(k)}(x \odot y)}_{\text{Sparsity-fixed model retraining}}, \tag{17}$$

where $y \in \mathbb{R}^d$ represents the network weights and $x \in \{0,1\}^d$ encodes the selection pattern, with zero entries indicating removed connections.

To solve Eq. (17), we instantiate this setup using a ResNet20 architecture trained on CIFAR10 and CIFAR100. Due to the overparameterized nature of deep neural networks relative to the complexity of these datasets, the lower-level objective satisfies the PL condition under standard assumptions (Liu et al., 2022). Our algorithm is then applied to jointly determine the model parameters and the sparsity structure, with a target sparsity level of 80% (i.e., retaining 20% of the weights).

All experiments are conducted under the same setup as in the previous section, except that we consider a heterogeneous data distribution across clients ($\delta > 0$). Specifically, the data are partitioned according to a Dirichlet distribution (Hsu et al., 2019) with $Dir(0.5)$, resulting in a non-IID setting. The testing accuracy and upper-level loss over iterations are reported in Figure 4. As shown in the figure, our method consistently achieves higher accuracy and lower objective values at a faster rate compared to competing approaches, demonstrating improved optimization efficiency in this setting.

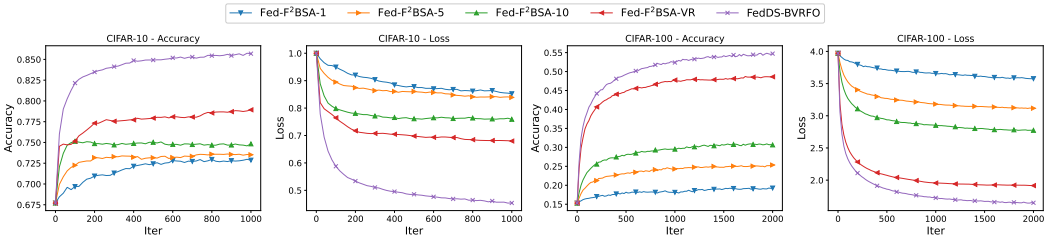


Figure 4: Testing accuracy and upper-level loss for model pruning, $\tau = 4$, nonconvex.

6 Conclusion

In this paper, we proposed a stochastic doubly smoothed framework for federated bilevel optimization, leading to a fully first-order algorithm with unified learning-rate scaling. By integrating smoothing into a Lagrangian reformulation, the proposed method effectively stabilizes the coupled dynamics across variables and avoids expensive second-order computations. We further established convergence guarantees under relaxed nonconvex lower-level conditions and derived improved sample and communication complexities with favorable dependence on the condition number. Extensive experiments on benchmark tasks, including hyperparameter optimization and model pruning, demonstrate the efficiency and robustness of our approach compared to existing methods.

AI USE STATEMENT

In this work, we did not use generative AI tools for tasks requiring disclosure, including developing theoretical models or conceptual frameworks, formulating mathematical claims, assisting with proofs, designing research methodology or experiments, implementing methods, or interpreting experimental results. These research tasks were conducted independently by the authors. Other tasks requiring disclosure, such as generating synthetic datasets, cleaning or reformatting datasets, and translation, were not applicable to this work. We used generative AI tools only for editing and polishing the manuscript to improve readability and check grammar. The tools were not used to generate or modify the technical contributions, theoretical results, experimental results, or scientific claims of the paper. We carefully reviewed all AI-assisted edits to ensure that they preserved the intended technical meaning and did not alter any results or claims. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

For the proposed algorithm, a detailed description and complete pseudocode are provided in Section 3 and the Appendix B. For the theoretical results, all assumptions are stated in Section 4.1, and complete proofs of the theoretical claims are provided in the Appendices C – F. For the experiments, Section 5 and Appendix A provide the implementation details, including the datasets, experimental settings, and hyperparameter choices. The source code will be made publicly available upon acceptance.

REFERENCES

Fei Chen, Mi Luo, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. Federated meta-learning with fast convergence and efficient communication. *arXiv preprint arXiv:1802.07876*, 2018.

Lesi Chen, Jing Xu, and Jingzhao Zhang. On finding small hyper-gradients in bilevel optimization: Hardness results and improved analysis. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 947–980. PMLR, 2024.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Tighter analysis of alternating stochastic gradient method for stochastic nested problems. *arXiv preprint arXiv:2106.13781*, 2021.

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

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

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in neural information processing systems*, 33:3557–3568, 2020.

Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated machine learning*, pp. 3–33. Springer, Cham, 2019.

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

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

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

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Feihu Huang. Fast adaptive federated bilevel optimization. *arXiv preprint arXiv:2211.01122*, 2022.

Minhui Huang, Dewei Zhang, and Kaiyi Ji. Achieving linear speedup in non-iid federated bilevel learning. In *International conference on machine learning*, pp. 14039–14059. PMLR, 2023.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pp. 4882–4892. PMLR, 2021.

Yihan Jiang, Jakub Konečnỳ, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pp. 18083–18113. PMLR, 2023.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. In *The Twelfth International Conference on Learning Representations*, 2024.

Jiaxiang Li, Siliang Zeng, Hoi-To Wai, Chenliang Li, Alfredo Garcia, and Mingyi Hong. Getting more juice out of the sft data: Reward learning from human demonstration improves sft for llm alignment. *Advances in Neural Information Processing Systems*, 37:124292–124318, 2024.

Junyi Li, Feihu Huang, and Heng Huang. Communication-efficient federated bilevel optimization with global and local lower level problems. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pp. 6083–6093. PMLR, 2020.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. In *International Conference on Learning Representations*, 2018.

Risheng Liu, Zhu Liu, Wei Yao, Shangzhi Zeng, and Jin Zhang. Moreau envelope for nonconvex bi-level optimization: A single-loop and hessian-free solution strategy. In *Forty-first International Conference on Machine Learning*, 2024.

Songtao Lu. Slm: A smoothed first-order lagrangian method for structured constrained nonconvex optimization. *Advances in Neural Information Processing Systems*, 36:80414–80454, 2023.

Songtao Lu. TSP: A two-sided smoothed primal-dual method for nonconvex bilevel optimization. In *Forty-second International Conference on Machine Learning*, 2025.

Zhaosong Lu and Sanyou Mei. First-order penalty methods for bilevel optimization. *SIAM Journal on Optimization*, 34(2):1937–1969, 2024.

Zhaosong Lu and Sanyou Mei. Solving bilevel optimization via sequential minimax optimization. *Mathematics of Operations Research*, 2026.

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

Han Shen and Tianyi Chen. On penalty-based bilevel gradient descent method. In *International Conference on Machine Learning*, pp. 30992–31015. PMLR, 2023.

Han Shen, Zhuoran Yang, and Tianyi Chen. Principled penalty-based methods for bilevel reinforcement learning and rlhf. *Journal of Machine Learning Research*, 26(114):1–49, 2025.

Wei Shen, Minhui Huang, Jiawei Zhang, and Cong Shen. Stochastic smoothed gradient descent ascent for federated minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3988–3996. PMLR, 2024.

Davoud Ataee Tarzanagh, Mingchen Li, Christos Thrampoulidis, and Samet Oymak. Fednest: Federated bilevel, minimax, and compositional optimization. In *International Conference on Machine Learning*, pp. 21146–21179. PMLR, 2022.

Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In *International Conference on Artificial Intelligence and Statistics*, pp. 5485–5517. PMLR, 2022.

Lei Yang, Jiaming Huang, Wanyu Lin, and Jiannong Cao. Personalized federated learning on non-iid data via group-based meta-learning. *ACM Transactions on Knowledge Discovery from Data*, 17(4):1–20, 2023a.

Yifan Yang, Peiyao Xiao, and Kaiyi Ji. Simfbo: Towards simple, flexible and communication-efficient federated bilevel learning. *Advances in Neural Information Processing Systems*, 36: 33027–33040, 2023b.

Yifan Yang, Peiyao Xiao, Shiqian Ma, and Kaiyi Ji. First-order federated bilevel learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22029–22037, 2025.

Jiawei Zhang, Peijun Xiao, Ruoyu Sun, and Zhiquan Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in neural information processing systems*, 33:7377–7389, 2020.

Miao Zhang, Steven W Su, Shirui Pan, Xiaojun Chang, Ehsan M Abbasnejad, and Reza Haffari. idarts: Differentiable architecture search with stochastic implicit gradients. In *International Conference on Machine Learning*, pp. 12557–12566. PMLR, 2021.

Xinwen Zhang and Hongchang Gao. On the convergence of stochastic smoothed multi-level compositional gradient descent ascent. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yihan Zhang, Rohit Dhaipule, Chiu C Tan, Haibin Ling, and Hongchang Gao. Federated stochastic bilevel optimization with fully first-order gradients. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 7047–7055, 2025.

Yihua Zhang, Yuguang Yao, Parikshit Ram, Pu Zhao, Tianlong Chen, Mingyi Hong, Yanzhi Wang, and Sijia Liu. Advancing model pruning via bi-level optimization. *Advances in Neural Information Processing Systems*, 35:18309–18326, 2022.

Taoli Zheng, Linglingzhi Zhu, Anthony Man-Cho So, José Blanchet, and Jiajin Li. Universal gradient descent ascent method for nonconvex-nonconcave minimax optimization. *Advances in Neural Information Processing Systems*, 36:54075–54110, 2023.

A Additional Experiments

A.1 Nonconvex Hyperparameter Optimization

We further evaluate all methods under a larger communication interval ($\tau = 16$), as shown in Figure 5 and Figure 6. A similar performance pattern is observed, where our method continues to achieve superior accuracy and lower upper-level loss, demonstrating its robustness and efficiency under reduced communication frequency.

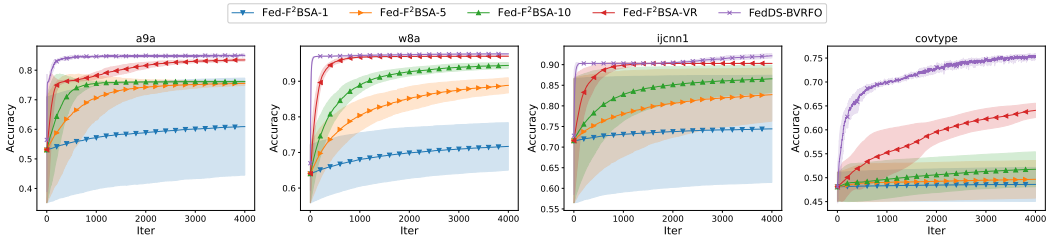


Figure 5: The test accuracy versus the training epochs, $\tau = 16$, nonconvex.

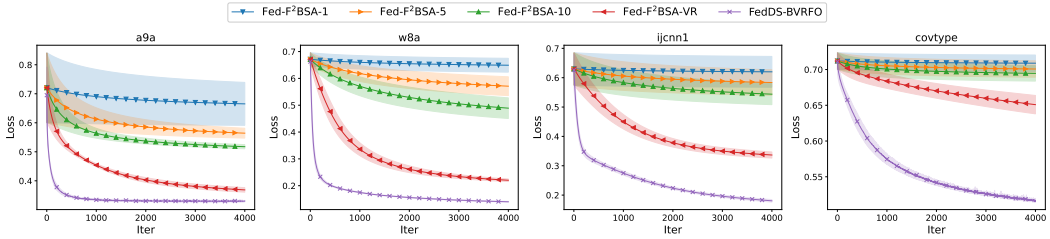


Figure 6: The upper-level loss function value versus the training epochs, $\tau = 16$, nonconvex.

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

To provide a more comprehensive comparison with existing federated stochastic bilevel baselines, we conduct additional experiments on a hyperparameter optimization task under a nonconvex–strongly-convex bilevel formulation. Specifically,

$$\min_{x\in\mathbb{R}^d} \frac{1}{K}\sum_{k=1}^{K} \frac{1}{m^{(n)}} \hat{\mathcal{L}}^{(k)}(y^*(x)^T; \mathcal{D}_{val}^{(k)}) \tag{18}$$

$$s.t. \quad y^*(x) = \arg\min_{y\in\mathbb{R}^{d\times c}} \frac{1}{K}\sum_{k=1}^{K} \hat{\mathcal{L}}^{(k)}(y; \mathcal{D}_{tr}^{(k)}) + \hat{\mathcal{R}}(x)\ ,$$

where the lower-level problem learns the parameter $y \in \mathbb{R}^{d\times c}$ of a logistic regression model using the training data, while the upper-level problem optimizes the hyperparameter $x \in \mathbb{R}^d$ with regularization term $\hat{\mathcal{R}}(x) = \frac{1}{d}\sum_{q=1}^{d} \exp(x_q) \frac{1}{c}\sum_{p=1}^{c} y_{pq}^2$ based on the validation data.

We compare Algorithm 1 with four state-of-the-art second-order methods: FedNEST (Tarzanagh et al., 2022), FedMBO (Huang et al., 2023), LocalBSGVRM (Gao, 2022), and FedBiOAcc (Li et al., 2023), as well as two first-order methods: MemFBO (Yang et al., 2025) and FedSVRBGD-FO (Zhang et al., 2025). For a fair comparison, we fix the target test accuracy for each dataset and set hyperparameters according to their theoretical guidelines. We report the upper-level objective value required to reach this accuracy versus running time in Figure 7. The results show that first-order methods converge faster than second-order methods, as they avoid the computation of Jacobian and Hessian matrices. Compared with MemFBO and FedSVRBGD-FO, our method achieves comparable or better convergence behavior. These results indicate that the proposed smoothing approach effectively enhances convergence speed in nonconvex–strongly-convex bilevel settings.

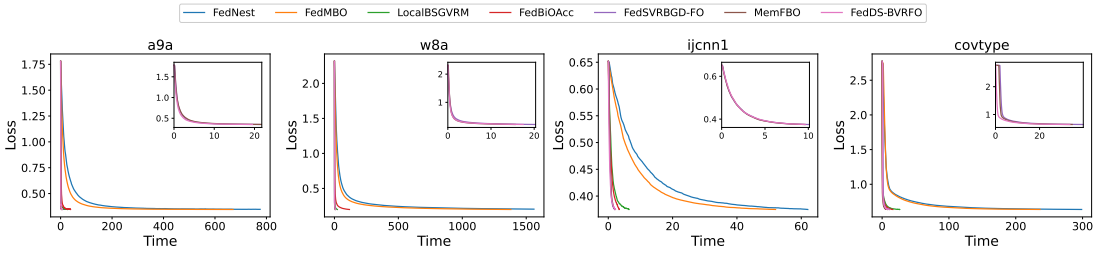


Figure 7: The upper-level loss function value versus the running time, $\tau = 4$, strongly-convex.

B Complete Algorithm

The full algorithm is presented below.

Algorithm 2 FedDS-BVRFO

Input: $\alpha_x > 0, \alpha_y > 0, \alpha_z > 0, \alpha_p > 0, \alpha_q > 0, \alpha_r > 0, \beta_x > 0, \beta_y > 0, \beta_z > 0, \eta > 0$.

Initialization on each worker $k \in \{1, \cdots, K\}$: $x_0^{(k)} = x_0$, $y_0^{(k)} = y_0$, $z_0^{(k)} = z_0$,

1: **for** $t = 0, \cdots, T-1$ **do**

2: Update variables x, y, and z:

$$x_{t+1}^{(k)} = x_t^{(k)} - \alpha_x \eta u_t^{(k)}, \qquad y_{t+1}^{(k)} = y_t^{(k)} - \alpha_y \eta v_t^{(k)}, \qquad z_{t+1}^{(k)} = z_t^{(k)} - \alpha_z \eta w_t^{(k)},$$

3: Update variables p, q, and r:

$$p_{t+1}^{(k)} = p_t^{(k)} + \alpha_p \eta (x_{t+1}^{(k)} - p_t^{(k)}), \qquad q_{t+1}^{(k)} = q_t^{(k)} + \alpha_q \eta (y_{t+1}^{(k)} - q_t^{(k)}), \qquad r_{t+1}^{(k)} = r_t^{(k)} + \alpha_r \eta (z_{t+1}^{(k)} - r_t^{(k)}),$$

4: Update gradient u:

$$u_{1,t+1}^{(k)} = (1 - \beta_x \eta^2)(u_{1,t}^{(k)} - \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_{t+1}^{(k)})) + \nabla_1 f^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; \xi_{t+1}^{(k)}),$$
$$u_{2,t+1}^{(k)} = (1 - \beta_x \eta^2)(u_{2,t}^{(k)} - \nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}, p_t^{(k)}, q_t^{(k)}; \zeta_{t+1}^{(k)})) + \nabla_1 g_{\gamma_1}^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}, p_{t+1}^{(k)}, q_{t+1}^{(k)}; \zeta_{t+1}^{(k)}),$$
$$u_{3,t+1}^{(k)} = (1 - \beta_x \eta^2)(u_{3,t}^{(k)} - \nabla_1 g_{\gamma_2}^{(k)}(x_t^{(k)}, z_t^{(k)}, r_t^{(k)}; \zeta_{t+1}^{(k)})) + \nabla_1 g_{\gamma_2}^{(k)}(x_{t+1}^{(k)}, z_{t+1}^{(k)}, r_{t+1}^{(k)}; \zeta_{t+1}^{(k)}),$$
$$u_{t+1}^{(k)} = \rho u_{1,t+1}^{(k)} + (u_{2,t+1}^{(k)} - u_{3,t+1}^{(k)}),$$

5: Update gradient v:

$$v_{1,t+1}^{(k)} = (1 - \beta_y \eta^2)(v_{1,t}^{(k)} - \nabla_2 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_{t+1}^{(k)})) + \nabla_2 f^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; \xi_{t+1}^{(k)}),$$
$$v_{2,t+1}^{(k)} = (1 - \beta_y \eta^2)(v_{2,t}^{(k)} - \nabla_2 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}, p_t^{(k)}, q_t^{(k)}; \zeta_{t+1}^{(k)})) + \nabla_2 g_{\gamma_1}^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}, p_{t+1}^{(k)}, q_{t+1}^{(k)}; \zeta_{t+1}^{(k)}),$$
$$v_{t+1}^{(k)} = \rho v_{1,t+1}^{(k)} + v_{2,t+1}^{(k)},$$

6: Update gradient w:

$$w_{t+1}^{(k)} = (1 - \beta_z \eta^2)(w_t^{(k)} - \nabla_2 g_{\gamma_2}^{(k)}(x_t^{(k)}, z_t^{(k)}, r_t^{(k)}; \zeta_{t+1}^{(k)})) + \nabla_2 g_{\gamma_2}^{(k)}(x_{t+1}^{(k)}, z_{t+1}^{(k)}, r_{t+1}^{(k)}; \zeta_{t+1}^{(k)}),$$

7: **if** $\mathrm{mod}(t+1, \tau) == 0$ **then**

8: Communication:

$$x_{t+1}^{(k)} = \frac{1}{K} \sum_{k'=1}^{K} x_{t+1}^{(k')}, \quad y_{t+1}^{(k)} = \frac{1}{K} \sum_{k'=1}^{K} y_{t+1}^{(k')}, \quad z_{t+1}^{(k)} = \frac{1}{K} \sum_{k'=1}^{K} z_{t+1}^{(k')},$$
$$p_{t+1}^{(k)} = \frac{1}{K} \sum_{k'=1}^{K} p_{t+1}^{(k')}, \quad q_{t+1}^{(k)} = \frac{1}{K} \sum_{k'=1}^{K} q_{t+1}^{(k')}, \quad r_{t+1}^{(k)} = \frac{1}{K} \sum_{k'=1}^{K} r_{t+1}^{(k')},$$
$$u_{t+1}^{(k)} = \frac{1}{K} \sum_{k'=1}^{K} u_{t+1}^{(k')}, \quad v_{t+1}^{(k)} = \frac{1}{K} \sum_{k'=1}^{K} v_{t+1}^{(k')}, \quad w_{t+1}^{(k)} = \frac{1}{K} \sum_{k'=1}^{K} w_{t+1}^{(k')},$$

9: **end if**

10: **end for**

C PROOF SKETCH

To facilitate the convergence analysis, we introduce the following auxiliary quantities, following the constructions in Zhang et al. (2020); Yang et al. (2022); Zheng et al. (2023):

$$
\begin{aligned}
&F_d(z;p,q,r) = \min_{x,y} F(x,y,z;p,q,r) = F(x^*(z;p,q,r), y^*(z;p,q,r), z;p,q,r)\ ,\\
&(x^*(z;p,q,r), y^*(z;p,q,r)) = \arg\min_{x,y} F(x,y,z;p,q,r)\\
&F_p(x,y;p,q,r) = \max_{z} F(x,y,z;p,q,r) = F(x,y,z^*(x,y;p,q,r);p,q,r)\ ,\\
&z^*(x,y;p,q,r) = \arg\max_{z} F(x,y,z;p,q,r)\ ,\\
&H(p,q,r) = \max_{z} F_d(z;p,q,r) = \min_{x,y}\max_{z} F(x,y,z;p,q,r)\ ,\\
&z^*(p,q,r) = \arg\max_{z} F_d(z;p,q,r)\ ,\\
&x^*(p,q,r) = x^*(z^*(p,q,r);p,q,r)\ , \qquad y^*(p,q,r) = y^*(z^*(p,q,r);p,q,r)\ ,\\
&H_z(p,q) = \max_{r} H(p,q,r)\ , \qquad r^*(p,q) = \arg\max_{r} H(p,q,r)\ ,\\
&z^+ = z_t + \alpha_z\eta\nabla_z F_d(z_t;p_t,q_t,r_t)\ , \qquad r^+ = \bar{r}_t + \alpha_r\eta(z^*(\bar{p}_t,\bar{q}_t,\bar{r}_t) - \bar{r}_t)\ .
\end{aligned}
\tag{19}
$$

We next provide a proof sketch for the convergence result. The analysis consists of two main components: (i) the smoothing technique (Appendix D), and (ii) the federated analysis (Appendix E).

Smoothing Analysis. We begin by establishing several auxiliary results that characterize the properties of the optimal solution mappings. Specifically, Lemmas D.1–D.4 provide continuity and regularity properties of the associated optimal functions.

Building on these results, we derive descent properties for the smoothed objective components in Lemmas D.5–D.7. These results enable us to introduce the following smoothing potential function (Lemma D.8):

$$
\Psi_t = \underbrace{F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)}_{\text{Lemma D.5}} - 2\underbrace{F_d(\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)}_{\text{Lemma D.6}} + 2\underbrace{H(\bar{p}_t,\bar{q}_t)}_{\text{Lemma D.7}}\ .
$$

This potential function captures the joint progress of all variables and serves as the key tool for the convergence analysis. Based on this construction, we show that the step sizes of x, y, and z can be chosen at the same order regarding κ:

$$
\alpha_y = \alpha_x\ , \quad \alpha_z = \alpha_x\frac{(\gamma_1 - L_\rho)^2}{72L_g^2}\ , \tag{20}
$$

which highlights the stabilizing effect of the smoothing technique.

Federated Analysis. We begin by establishing two useful lemmas, Lemmas E.1–E.2, which characterize the evolution of local variables and their consensus errors. We then analyze the property of controlling gradient estimation errors and consensus errors.

First, we bound the six gradient estimation errors using Lemmas E.3 and E.4. In particular, Lemma E.4 establishes three novel bounds associated with the smoothed gradients of g_{γ_1} and g_{γ_2}, which are essential for handling the doubly-smoothed terms.

Next, we control the corresponding consensus errors in Lemmas E.5 and E.6. Among them, Lemma E.6 captures the additional error terms introduced by doubly-smoothed. Finally, these bounds are aggregated in Lemma E.7, which provides an overall consensus error bound across all six components.

Final Proof. We construct a novel Lyapunov function $\mathcal{P}_t$ in Appendix F:

$$
\mathcal{P}_t = \Psi_t + c_1\rho^2\mathcal{G}_{u,1,t} + c_2\mathcal{G}_{u,2,t} + c_3\mathcal{G}_{u,3,t} + c_4\rho^2\mathcal{G}_{v,1,t} + c_5\mathcal{G}_{v,2,t} + c_6\mathcal{G}_{w,t}\ ,
$$

where we denote

$$
\mathcal{G}_{u,1,t} = \mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}) - \bar{u}_{1,t}\|^2]\ ,
$$

$$\mathcal{G}_{u,2,t} = \mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)}) - \bar{u}_{2,t}\|^2] ,$$

$$\mathcal{G}_{u,3,t} = \mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_1 g_{\gamma_2}^{(k)}(x_t^{(k)}, z_t^{(k)}; r_t^{(k)}) - \bar{u}_{3,t}\|^2] ,$$

$$\mathcal{G}_{v,1,t} = \mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_2 f^{(k)}(x_t^{(k)}, y_t^{(k)}) - \bar{v}_{1,t}\|^2] ,$$

$$\mathcal{G}_{v,2,t} = \mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_2 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)}) - \bar{v}_{2,t}\|^2] ,$$

$$\mathcal{G}_{w,t} = \mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_2 g_{\gamma_2}^{(k)}(x_t^{(k)}, z_t^{(k)}; r_t^{(k)}) - \bar{w}_t\|^2] . \tag{21}$$

This Lyapunov function integrates the smoothing potential Ψ_t from Appendix D with the six gradient error terms from Appendix E. By carefully choosing the coefficients $\{c_j\}_{j=1}^6$ and properly selecting the hyperparameter α_x, we establish a descent property for $\mathcal{P}_t$, which leads to the desired convergence result.

D SMOOTHED LEMMAS

Lemma D.1. *Given Assumptions 4.1-4.4, the following inequality holds:*

$$(i).\ \|x^*(z;p,q,r) - x^*(z;p',q',r)\|^2 + \|y^*(z;p,q,r) - y^*(z;p',q',r)\|^2 \le \omega_1^2(\|p-p'\|^2 + \|q-q'\|^2) ,$$
$$(ii).\ \|x^*(p,q,r) - x^*(p',q',r)\|^2 + \|y^*(p,q,r) - y^*(p',q',r)\|^2 \le \omega_1^2(\|p-p'\|^2 + \|q-q'\|^2) ,$$
$$(iii).\ \|x^*(z;p,q,r) - x^*(z';p,q,r)\|^2 + \|y^*(z;p,q,r) - y^*(z';p,q,r)\|^2 \le \omega_2^2\|z-z'\|^2 . \tag{22}$$

where $\omega_1 = \frac{\gamma_1}{(\gamma_1 - L_\rho)}, \omega_2 = \frac{L_g}{(\gamma_1 - L_\rho)}$.

Proof. **Dependence of** (x^*, y^*) **on** (p, q)**.** To prove (i), on the one hand, we split

$$F(x^*(z;p,q,r), y^*(z;p,q,r), z; p', q', r) - F(x^*(z;p',q',r), y^*(z;p',q',r), z; p', q', r)$$

into three terms

$$T_1 := F(x^*(z;p,q,r), y^*(z;p,q,r), z; p', q', r) - F(x^*(z;p,q,r), y^*(z;p,q,r), z; p, q, r) ,$$
$$T_2 := F(x^*(z;p,q,r), y^*(z;p,q,r), z; p, q, r) - F(x^*(z;p',q',r), y^*(z;p',q',r), z; p, q, r) ,$$
$$T_3 := F(x^*(z;p',q',r), y^*(z;p',q',r), z; p, q, r) - F(x^*(z;p',q',r), y^*(z;p',q',r), z; p', q', r) .$$

For T_2, due to the $(\gamma_1 - L_\rho)$-strong convexity of F with respect to x and y, we obtain

$$T_2 \le -\frac{(\gamma_1 - L_\rho)}{2}\|x^*(z;p,q,r) - x^*(z;p',q',r)\|^2 - \frac{(\gamma_1 - L_\rho)}{2}\|y^*(z;p,q,r) - y^*(z;p',q',r)\|^2 .$$

For T_1 and T_3 , we derive

$$\begin{aligned} T_1 &\le \frac{\gamma_1}{2}\|x^*(z;p,q,r) - p'\|^2 + \frac{\gamma_1}{2}\|y^*(z;p,q,r) - q'\|^2 \\ &\quad - \frac{\gamma_1}{2}\|x^*(z;p,q,r) - p\|^2 - \frac{\gamma_1}{2}\|y^*(z;p,q,r) - q\|^2 \\ &\le \frac{\gamma_1}{2}(2\langle x^*(z;p,q,r), p - p'\rangle + \|p'\|^2 - \|p\|^2) + \frac{\gamma_1}{2}(2\langle y^*(z;p,q,r), q - q'\rangle + \|q'\|^2 - \|q\|^2) . \\ T_3 &\le \frac{\gamma_1}{2}\|x^*(z;p',q',r) - p\|^2 + \frac{\gamma_1}{2}\|y^*(z;p',q',r) - q\|^2 \\ &\quad - \frac{\gamma_1}{2}\|x^*(z;p',q',r) - p'\|^2 - \frac{\gamma_1}{2}\|y^*(z;p',q',r) - q'\|^2 \\ &\le \frac{\gamma_1}{2}(2\langle x^*(z;p',q',r), p' - p\rangle + \|p\|^2 - \|p'\|^2) + \frac{\gamma_1}{2}(2\langle y^*(z;p',q',r), q' - q\rangle + \|q\|^2 - \|q'\|^2) . \end{aligned}$$

Combining T_1 and T_3 yields

$$T_1 + T_3 \leq \gamma_1 \langle x^*(z;p,q,r) - x^*(z;p',q',r), p - p' \rangle + \gamma_1 \langle y^*(z;p,q,r) - y^*(z;p',q',r), q - q' \rangle .$$

Thus, combining $T_1 + T_2 + T_3$ yields

$$\begin{aligned} &F(x^*(z;p,q,r), y^*(z;p,q,r), z; p',q',r) - F(x^*(z;p',q',r), y^*(z;p',q',r), z; p',q',r) \\ \leq &-\frac{(\gamma_1 - L_\rho)}{2}\|x^*(z;p,q,r) - x^*(z;p',q',r)\|^2 - \frac{(\gamma_1 - L_\rho)}{2}\|y^*(z;p,q,r) - y^*(z;p',q',r)\|^2 \\ &+ \gamma_1 \langle x^*(z;p,q,r) - x^*(z;p',q',r), p - p' \rangle + \gamma_1 \langle y^*(z;p,q,r) - y^*(z;p',q',r), q - q' \rangle . \end{aligned}$$

On the other hand, due to the $(\gamma_1 - L_\rho)$-strong convexity of F with respect to x and y, we have

$$\begin{aligned} &F(x^*(z;p,q,r), y^*(z;p,q,r), z; p',q',r) - F(x^*(z;p',q',r), y^*(z;p',q',r), z; p',q',r) \\ \geq &\frac{(\gamma_1 - L_\rho)}{2}\|x^*(z;p,q,r) - x^*(z;p',q',r)\|^2 + \frac{(\gamma_1 - L_\rho)}{2}\|y^*(z;p,q,r) - y^*(z;p',q',r)\|^2 . \end{aligned}$$

By combining the above two inequalities, we have

$$\begin{aligned} &(\gamma_1 - L_\rho)\|x^*(z;p,q,r) - x^*(z;p',q',r)\|^2 + (\gamma_1 - L_\rho)\|y^*(z;p,q,r) - y^*(z;p',q',r)\|^2 \\ \leq &\gamma_1 \langle x^*(z;p,q,r) - x^*(z;p',q',r), p - p' \rangle + \gamma_1 \langle y^*(z;p,q) - y^*(z;p',q'), q - q' \rangle . \end{aligned}$$

By denoting $a = [x^*(z;p,q,r) - x^*(z;p',q',r), y^*(z;p,q,r) - y^*(z;p',q',r)]^T$ and $b = [p - p', q - q']^T$, the above inequality can be represented as follows:

$$(\gamma_1 - L_\rho)\|a\|^2 \leq \gamma_1 \langle a, b \rangle \leq \gamma_1 \|a\|\|b\| .$$

It indicates

$$(\gamma_1 - L_\rho)\|a\| \leq \gamma_1 \|b\| .$$

As a result, we have

$$\|x^*(z;p,q,r) - x^*(z;p',q',r)\|^2 + \|y^*(z;p,q,r) - y^*(z;p',q',r)\|^2 \leq \omega_1^2(\|p - p'\|^2 + \|q - q'\|^2) .$$

where $\omega_1 = \frac{\gamma_1}{(\gamma_1 - L_\rho)}$. The second inequalities (ii) in Eq. (22) can be proved in a similar manner.

Dependence of (x^*, y^*) on z. To prove (iii), due to $(\gamma_1 - L_\rho)$-strong convexity of F with respect to x and y, we obtain

$$\begin{aligned} &(\gamma_1 - L_\rho)\|(x^*(z';p,q,r), y^*(z';p,q,r)) - (x^*(z;p,q,r), y^*(z;p,q,r))\| \\ \leq &\|\nabla_{x,y} F(x^*(z';p,q,r), y^*(z';p,q,r), z; p,q,r)\| \\ = &\|\nabla_{x,y} F(x^*(z';p,q,r), y^*(z';p,q,r), z; p,q,r) - \nabla_{x,y} F(x^*(z';p,q,r), y^*(z';p,q,r), z'; p,q,r)\| , \end{aligned}$$

where the last equality uses $\nabla_{x,y} F(x^*(z';p,q,r), y^*(z';p,q,r), z'; p,q,r) = 0$. Recalling $\nabla_x F(x,y,z;p,q,r) = \rho \nabla_1 f(x,y) + \nabla_1 g(x,y) - \nabla_1 g(x,z) + \gamma_1(x - p)$ and $\nabla_y F(x,y,z;p,q,r) = \rho \nabla_2 f(x,y) + \nabla_2 g(x,y) + \gamma_1(y - q)$, only the x-component of $\nabla_{x,y} F$ depends on z. Hence

$$\begin{aligned} &\|\nabla_{x,y} F(x^*(z';p,q,r), y^*(z';p,q,r), z; p,q,r) - \nabla_{x,y} F(x^*(z';p,q,r), y^*(z';p,q,r), z'; p,q,r)\| \\ = &\|\nabla_1 g(x^*(z'), z') - \nabla_1 g(x^*(z'), z)\| \leq L_g \|z - z'\| . \end{aligned}$$

Combining the two displays, we obtain

$$\|x^*(z;p,q,r) - x^*(z';p,q,r)\|^2 + \|y^*(z;p,q,r) - y^*(z';p,q,r)\|^2 \leq \omega_2^2 \|z - z'\|^2 ,$$

where $\omega_2 = \frac{L_g}{(\gamma_1 - L_\rho)}$. □

Lemma D.2. *Given Assumptions 4.1-4.4, the following inequalities hold:*

$$\begin{aligned} &(i).\ \|z^*(x,y;p,q,r) - z^*(x,y;p,q,r')\|^2 \leq \omega_3^2 \|r - r'\|^2 , \\ &(ii).\ \|z^*(p,q,r) - z^*(p,q,r')\|^2 \leq \omega_3^2 \|r - r'\|^2 , \\ &(iii).\ \|z^*(x,y;p,q,r) - z^*(x',y;p,q,r)\|^2 \leq \omega_4^2 \|x - x'\|^2 , \\ &(iv).\ \|z^*(p,q,r) - z^*(p',q',r)\|^2 \leq \omega_5^2 \big(\|p - p'\|^2 + \|q - q'\|^2\big) , \end{aligned} \tag{23}$$

where $\omega_3 = \frac{\gamma_2}{(\gamma_2 - L_\rho)}$, $\omega_4 = \frac{L_g}{(\gamma_2 - L_\rho)}$, *and* $\omega_5 = \frac{\gamma_1 \omega_2}{(\gamma_2 - L_\rho)} + \frac{\omega_1}{\omega_2}$.

Proof. **Dependence of z^* on r.** We split

$$F(x,y,z^*(x,y;p,q,r);p,q,r') - F(x,y,z^*(x,y;p,q,r');p,q,r')$$

into three terms

$$\begin{aligned}
S_1 &:= F(x,y,z^*(x,y;p,q,r);p,q,r') - F(x,y,z^*(x,y;p,q,r);p,q,r) \,, \\
S_2 &:= F(x,y,z^*(x,y;p,q,r);p,q,r) - F(x,y,z^*(x,y;p,q,r');p,q,r) \,, \\
S_3 &:= F(x,y,z^*(x,y;p,q,r');p,q,r) - F(x,y,z^*(x,y;p,q,r');p,q,r') \,.
\end{aligned}$$

Since (x,y,p,q) are held fixed, only the term $-\frac{\gamma_2}{2}\|z-r\|^2$ in F contributes to S_1 and S_3; a direct expansion gives

$$S_1 + S_3 = -\gamma_2 \langle z^* - z^{*\prime},\, r - r' \rangle \,.$$

For S_2, since z^* maximizes $F(x,y,\cdot;p,q,r)$ and F is $(\gamma_2 - L_\rho)$-strongly concave in z, we have $S_2 \geq \frac{\gamma_2 - L_\rho}{2}\|z^* - z^{*\prime}\|^2$.

On the other hand, since $z^{*\prime}$ maximizes $F(x,y,\cdot;p,q,r')$, strong concavity also gives $F(x,y,z^*;p,q,r') \leq F(x,y,z^{*\prime};p,q,r') - \frac{\gamma_2 - L_\rho}{2}\|z^* - z^{*\prime}\|^2$, i.e., $S_1 + S_2 + S_3 \leq -\frac{\gamma_2 - L_\rho}{2}\|z^* - z^{*\prime}\|^2$. Combining the two bounds on $S_1 + S_2 + S_3$,

$$\frac{\gamma_2 - L_\rho}{2}\|z^* - z^{*\prime}\|^2 - \gamma_2\langle z^* - z^{*\prime}, r - r'\rangle \leq S_1 + S_2 + S_3 \leq -\frac{\gamma_2 - L_\rho}{2}\|z^* - z^{*\prime}\|^2 \,,$$

which yields $(\gamma_2 - L_\rho)\|z^* - z^{*\prime}\|^2 \leq \gamma_2\langle z^* - z^{*\prime}, r - r'\rangle \leq \gamma_2\|z^* - z^{*\prime}\|\|r - r'\|$, and hence

$$\|z^*(x,y;p,q,r) - z^*(x,y;p,q,r')\|^2 \leq \omega_3^2\|r - r'\|^2 \,,$$

where $\omega_3 = \frac{\gamma_2}{\gamma_2 - L_\rho}$. The second inequalities (ii) in Eq. (23) can be proved in a similar manner.

Dependence of z^* on x. We first note that $z^*(x,y;p,q,r)$ does not depend on y at all. By the $(\gamma_2 - L_\rho)$-strong concavity of F in z,

$$\begin{aligned}
&(\gamma_2 - L_\rho)\|z^*(x,y;p,q,r) - z^*(x',y;p,q,r)\| \\
&\leq \|\nabla_z F(x,y,z^*(x',y;p,q,r);p,q,r)\| \\
&= \|\nabla_z F(x,y,z^*(x',y;p,q,r);p,q,r) - \nabla_z F(x',y,z^*(x',y;p,q,r);p,q,r)\| \,, \\
&\leq L_g\|x - x'\| \,,
\end{aligned}$$

where the equality uses $\nabla_z F(x',y,z^*(x',y;p,q,r);p,q,r) = 0$. Thus. we obtain

$$\|z^*(x,y;p,q,r) - z^*(x',y;p,q,r)\|^2 \leq \omega_4^2\|x - x'\|^2 \,,$$

where $\omega_4 = \frac{L_g}{\gamma_2 - L_\rho}$.

Dependence of z^* on p,q. By strong concavity of F_d, we obtain

$$\begin{aligned}
F_d(z^*(p,q,r);p,q,r) - F_d(z^*(p',q',r);p,q,r) &\geq \frac{\gamma_2 - L_\rho}{2}\|z^*(p,q,r) - z^*(p',q',r)\|^2 \,, \\
F_d(z^*(p,q,r);p',q',r) - F_d(z^*(p',q',r);p',q',r) &\leq -\frac{\gamma_2 - L_\rho}{2}\|z^*(p,q,r) - z^*(p',q',r)\|^2 \,.
\end{aligned}$$

Combining these two inequalities derives

$$\begin{aligned}
&(\gamma_2 - L_\rho)\|z^*(p,q,r) - z^*(p',q',r)\|^2 \\
&\leq \underbrace{\big[F_d(z^*(p,q,r);p,q,r) - F_d(z^*(p,q,r);p',q',r)\big]}_{=:U_1} \\
&\quad + \underbrace{\big[F_d(z^*(p',q',r);p',q',r) - F_d(z^*(p',q',r);p,q,r)\big]}_{=:U_2} \,.
\end{aligned}$$

For U_1, we obtain

$$U_1 \leq \gamma_1\langle x^*(z^*(p,q,r);p',q',r), p - p'\rangle + \gamma_1\langle y^*(z^*(p,q,r);p',q',r), q - q'\rangle$$

$$+\frac{\gamma_1}{2}(\|p'\|^2-\|p\|^2+\|q'\|^2-\|q\|^2)\,.$$

Symmetrically,

$$\begin{aligned}U_2 \le{}& \gamma_1\langle x^*(z^*(p',q',r);p,q,r),p'-p\rangle+\gamma_1\langle y^*(z^*(p',q',r);p,q,r),q'-q\rangle\\ &+\frac{\gamma_1}{2}(\|p\|^2-\|p'\|^2+\|q\|^2-\|q'\|^2)\,.\end{aligned}$$

Adding U_1 and U_2, the quadratic terms cancel, leaving

$$\begin{aligned}U_1+U_2 \le{}& \gamma_1\langle x^*(z^*(p,q,r);p',q',r)-x^*(z^*(p',q',r);p,q,r),p-p'\rangle\\ &+\gamma_1\langle y^*(z^*(p,q,r);p',q',r)-y^*(z^*(p',q',r);p,q,r),q-q'\rangle\,.\end{aligned}$$

Denoting

$$\begin{aligned}a :=&[x^*(z^*(p,q,r);p',q',r)-x^*(z^*(p',q',r);p,q,r),\\ &\ y^*(z^*(p,q,r);p',q',r)-y^*(z^*(p',q',r);p,q,r)]^T\,,\\ b :=&[p-p',q-q']^T\,,\end{aligned}$$

the display above gives

$$\begin{aligned}&(\gamma_2-L_\rho)\|z^*(p,q,r)-z^*(p',q',r)\|^2\le\gamma_1\langle a,b\rangle\le\gamma_1\|a\|\|b\|\\ \le{}&\gamma_1(\omega_1\|b\|+\omega_2\|z^*(p,q,r)-z^*(p',q',r)\|)\|b\|\,,\end{aligned}$$

where the last step from Lemma D.1.

Setting $\xi := \frac{\|z^*(p,q,r)-z^*(p',q',r)\|}{\|b\|}$, we have $\xi^2 \le \frac{\gamma_1\omega_2}{(\gamma_2-L_\rho)}\xi+\frac{\gamma_1\omega_1}{(\gamma_2-L_\rho)}$. Following Zheng et al. (2023), using AM-GM inequality, we obtain

$$\|z^*(p,q,r)-z^*(p',q',r)\|^2\le\omega_5^2\big(\|p-p'\|^2+\|q-q'\|^2\big)\,,$$

where $\omega_5=\frac{\gamma_1\omega_2}{(\gamma_2-L_\rho)}+\frac{\omega_1}{\omega_2}$. □

Lemma D.3. *Given Assumptions 4.1-4.4, by defining* $\hat{r} := r^*(\bar{p}_{t+1},\bar{q}_{t+1})$, $r^+ = \bar{r}_t+\alpha_r\eta(z^*(\bar{p}_t,\bar{q}_t,\bar{r}_t)-\bar{r}_t)$, $z^+=\bar{z}_t+\alpha_z\eta\nabla_z F_d(\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)$, *and* $C_r=\frac{4\gamma_2^2}{\mu(\gamma_1-L_\rho)}\big(\omega_3^2+\frac{(1-\alpha_r\eta)^2}{\alpha_r^2\eta^2}\big)$, *the following inequality holds:*

$$\begin{aligned}&\|x^*(\bar{p}_{t+1},\bar{q}_{t+1},r^+)-x^*(\bar{p}_{t+1},\bar{q}_{t+1},\hat{r})\|^2+\|y^*(\bar{p}_{t+1},\bar{q}_{t+1},r^+)-y^*(\bar{p}_{t+1},\bar{q}_{t+1},\hat{r})\|^2\\ \le{}&\frac{2\omega_5^2\gamma_2^2}{\mu(\gamma_1-L_\rho)}\big(\|\bar{p}_{t+1}-\bar{p}_t\|^2+\|\bar{q}_{t+1}-\bar{q}_t\|^2\big)+2C_r\|\bar{r}_{t+1}-\bar{r}_t\|^2\\ &+4C_r\alpha_r^2\eta^2\Big(\alpha_z^2\eta^2+\frac{2}{(\gamma_2-L_\rho)^2}\Big)\|\bar{w}_t\|^2+\frac{16C_r\alpha_r^2\eta^2L_g^2}{(\gamma_1-L_\rho)^2(\gamma_2-L_\rho)^2}\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2\\ &+\frac{16C_r\alpha_r^2\eta^2}{(\gamma_2-L_\rho)^2}\|\nabla_z F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)-\bar{w}_t\|^2\,.\end{aligned}\tag{24}$$

Proof. By the (γ_1-L_ρ)-strong convexity of F_p in (x,y), denoting $x_+^* := x^*(\bar{p}_{t+1},\bar{q}_{t+1},r^+)$, $y_+^* := y^*(\bar{p}_{t+1},\bar{q}_{t+1},r^+)$ we have

$$\begin{aligned}&\frac{\gamma_1-L_\rho}{2}\Big(\|x_+^*-x^*(\bar{p}_{t+1},\bar{q}_{t+1},\hat{r})\|^2+\|y_+^*-y^*(\bar{p}_{t+1},\bar{q}_{t+1},\hat{r})\|^2\Big)\\ \le{}&\max_r F_p(x_+^*,y_+^*;\bar{p}_{t+1},\bar{q}_{t+1},r)-F_p(x_+^*,y_+^*;\bar{p}_{t+1},\bar{q}_{t+1},\hat{r})\\ \le{}&\max_r F_p(x_+^*,y_+^*;\bar{p}_{t+1},\bar{q}_{t+1},r)-F_p(x_+^*,y_+^*;\bar{p}_{t+1},\bar{q}_{t+1},r^+)\\ \le{}&\frac{1}{2\mu}\|\nabla_r F_p(x_+^*,y_+^*;\bar{p}_{t+1},\bar{q}_{t+1},r^+)\|^2\\ ={}&\frac{\gamma_2^2}{2\mu}\|z^*(\bar{p}_{t+1},\bar{q}_{t+1},r^+)-r^+\|^2\,,\end{aligned}$$

where the second-to-last step holds due to the PL-property.

Then, since $r^+ = \bar{r}_t + \alpha_r\eta(z^*(\bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{r}_t)$, by Young's inequality, we obtain

$$\begin{aligned}
&\|z^*(\bar{p}_{t+1}, \bar{q}_{t+1}, r^+) - r^+\|^2 \\
&\leq 2\|z^*(\bar{p}_{t+1}, \bar{q}_{t+1}, r^+) - z^*(\bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 + 2(1-\alpha_r\eta)^2\|z^*(\bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{r}_t\|^2 \\
&\leq 4\|z^*(\bar{p}_{t+1}, \bar{q}_{t+1}, r^+) - z^*(\bar{p}_t, \bar{q}_t, r^+)\|^2 + 4\|z^*(\bar{p}_t, \bar{q}_t, r^+) - z^*(\bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 \\
&\quad + 2(1-\alpha_r\eta)^2\|z^*(\bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{r}_t\|^2 \\
&\leq 4\omega_5^2\big(\|\bar{p}_{t+1} - \bar{p}_t\|^2 + \|\bar{q}_{t+1} - \bar{q}_t\|^2\big) + 4\omega_3^2\|r^+ - \bar{r}_t\|^2 + \frac{2(1-\alpha_r\eta)^2}{\alpha_r^2\eta^2}\|r^+ - \bar{r}_t\|^2 \\
&\leq 4\omega_5^2\big(\|\bar{p}_{t+1} - \bar{p}_t\|^2 + \|\bar{q}_{t+1} - \bar{q}_t\|^2\big) + 4\big(\omega_3^2 + \frac{(1-\alpha_r\eta)^2}{\alpha_r^2\eta^2}\big)\|r^+ - \bar{r}_t\|^2 .
\end{aligned}$$

Finally, we derive

$$\begin{aligned}
&\|r^+ - \bar{r}_t\|^2 \leq 2\|\bar{r}_{t+1} - \bar{r}_t\|^2 + 2\|r^+ - \bar{r}_{t+1}\|^2 \\
&= 2\|\bar{r}_{t+1} - \bar{r}_t\|^2 + 2\|\bar{r}_t + \alpha_r\eta(z^*(\bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{r}_t) - \bar{r}_t - \alpha_r\eta(\bar{z}_{t+1} - \bar{r}_t)\|^2 \\
&= 2\|\bar{r}_{t+1} - \bar{r}_t\|^2 + 2\alpha_r^2\eta^2\|\bar{z}_{t+1} - z^*(\bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 \\
&\leq 2\|\bar{r}_{t+1} - \bar{r}_t\|^2 + 4\alpha_r^2\eta^2\|\bar{z}_{t+1} - \bar{z}_t\|^2 + 4\alpha_r^2\eta^2\|\bar{z}_t - z^*(\bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 \\
&\leq 2\|\bar{r}_{t+1} - \bar{r}_t\|^2 + 4\alpha_r^2\eta^2\alpha_z^2\eta^2\|\bar{w}_t\|^2 + \frac{4\alpha_r^2}{\alpha_z^2(\gamma_2 - L_\rho)^2}\|z^+ - \bar{z}_t\|^2 ,
\end{aligned}$$

where the last step holds due to the the $(\gamma_2 - L_\rho)$-strong concavity of $F_d(\cdot; \bar{p}_t, \bar{q}_t, \bar{r}_t)$, and the notation of $z^+ = \bar{z}_t + \alpha_z\eta\nabla_z F_d(\bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)$:

$$(\gamma_2 - L_\rho)\|\bar{z}_t - z^*(\bar{p}_t, \bar{q}_t, \bar{r}_t)\| \leq \|\nabla_z F_d(\bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\| = \frac{1}{\alpha_z\eta}\|z^+ - \bar{z}_t\| .$$

Similarly, we derive

$$\begin{aligned}
&\|z^+ - \bar{z}_t\|^2 \leq 2\|z^+ - \bar{z}_{t+1}\|^2 + 2\|\bar{z}_{t+1} - \bar{z}_t\|^2 \\
&= 2\alpha_z^2\eta^2\|\nabla_z F_d(\bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{w}_t\|^2 + 2\alpha_z^2\eta^2\|\bar{w}_t\|^2 \\
&\leq 4\alpha_z^2\eta^2\|\nabla_z F_d(\bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 \\
&\quad + 4\alpha_z^2\eta^2\|\nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{w}_t\|^2 + 2\alpha_z^2\eta^2\|\bar{w}_t\|^2 \\
&\leq 4\alpha_z^2\eta^2 L_g^2\|x^*(\bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{x}_t\|^2 + 4\alpha_z^2\eta^2\|\nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{w}_t\|^2 \\
&\quad + 2\alpha_z^2\eta^2\|\bar{w}_t\|^2 \\
&\leq \frac{4\alpha_z^2\eta^2 L_g^2}{(\gamma_1 - L_\rho)^2}\|\nabla_x F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 + 4\alpha_z^2\eta^2\|\nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{w}_t\|^2 \\
&\quad + 2\alpha_z^2\eta^2\|\bar{w}_t\|^2 ,
\end{aligned} \tag{25}$$

where the last step holds from the $(\gamma_1 - L_\rho)$-strongly convex of $F(\cdot, \cdot, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)$:

$$\|x^*(\bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{x}_t\|^2 \leq \frac{1}{(\gamma_1 - L_\rho)^2}\|\nabla_x F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 .$$

Thus, we have

$$\begin{aligned}
&\|r^+ - \bar{r}_t\|^2 \leq 2\|\bar{r}_{t+1} - \bar{r}_t\|^2 + 4\alpha_r^2\eta^2\Big(\alpha_z^2\eta^2 + \frac{2}{(\gamma_2 - L_\rho)^2}\Big)\|\bar{w}_t\|^2 \\
&+ \frac{16\alpha_r^2\eta^2 L_g^2}{(\gamma_1 - L_\rho)^2(\gamma_2 - L_\rho)^2}\|\nabla_x F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 \\
&+ \frac{16\alpha_r^2\eta^2}{(\gamma_2 - L_\rho)^2}\|\nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{w}_t\|^2 .
\end{aligned} \tag{26}$$

Combining the above and dividing by $\frac{\gamma_1 - L_\rho}{2}$ completes the proof. □

Lemma D.4. *Given Assumptions 4.1-4.4, by defining $\hat{r} := r^*(\bar{p}_{t+1}, \bar{q}_{t+1})$, and $z^+ = z + \alpha_z\eta\nabla_z F(x^*(z;p,q,r), y^*(z;p,q,r), z; p, q, r)$, the following inequality holds:*

$$\begin{aligned}
&\|x^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t) - x^*(\bar{p}_{t+1},\bar{q}_{t+1},r^+)\|^2 \\
&+ \|y^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t) - y^*(\bar{p}_{t+1},\bar{q}_{t+1},r^+)\|^2 \\
&\le \Big(10\omega_2^2\alpha_z^2\eta^2\big(1+2\omega_3^2\alpha_r^2\eta^2\big) + \frac{10\omega_2^2\big(1+4\omega_3^2\alpha_r^2\eta^2\big)}{(\gamma_2-L_\rho)^2}\Big)\|\bar{w}_t\|^2 \\
&\quad + \Big(\frac{30\omega_2^2\alpha_z^2\eta^2L_g^2}{(\gamma_1-L_\rho)^2} + \frac{20\omega_2^2L_g^2\big(1+4\omega_3^2\alpha_r^2\eta^2\big)}{(\gamma_1-L_\rho)^2(\gamma_2-L_\rho)^2}\Big)\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2 \\
&\quad + \Big(30\omega_2^2\alpha_z^2\eta^2 + \frac{20\omega_2^2\big(1+4\omega_3^2\alpha_r^2\eta^2\big)}{(\gamma_2-L_\rho)^2}\Big)\|\nabla_z F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t) - \bar{w}_t\|^2 \\
&\quad + 5\omega_2^2\omega_5^2(\|\bar{p}_{t+1}-\bar{p}_t\|^2 + \|\bar{q}_{t+1}-\bar{q}_t\|^2) + 10\omega_2^2\omega_3^2\|\bar{r}_{t+1}-\bar{r}_t\|^2 \,. \qquad (27)
\end{aligned}$$

Proof. At first, from Lemmas D.1 and D.2, we decompose

$$\begin{aligned}
&\|x^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t) - x^*(\bar{p}_{t+1},\bar{q}_{t+1},r^+)\|^2 \\
&+ \|y^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t) - y^*(\bar{p}_{t+1},\bar{q}_{t+1},r^+)\|^2 \\
&\le \omega_2^2\|\bar{z}_{t+1} - z^*(\bar{p}_{t+1},\bar{q}_{t+1},r^+)\|^2 \\
&\le 5\omega_2^2\|\bar{z}_{t+1}-z^+\|^2 + 5\omega_2^2\|z^+-\bar{z}_t\|^2 + 5\omega_2^2\|\bar{z}_t - z^*(\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2 \\
&\quad + 5\omega_2^2\|z^*(\bar{p}_t,\bar{q}_t,\bar{r}_t) - z^*(\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t)\|^2 + 5\omega_2^2\|z^*(\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t) - z^*(\bar{p}_{t+1},\bar{q}_{t+1},r^+)\|^2 \\
&\le 5\omega_2^2\|\bar{z}_{t+1}-z^+\|^2 + 5\omega_2^2\Big(1+\frac{1}{\alpha_z^2\eta^2(\gamma_2-L_\rho)^2}\Big)\|z^+-\bar{z}_t\|^2 \\
&\quad + 5\omega_2^2\omega_5^2(\|\bar{p}_{t+1}-\bar{p}_t\|^2 + \|\bar{q}_{t+1}-\bar{q}_t\|^2) + 5\omega_2^2\omega_3^2\|r^+-\bar{r}_t\|^2 \,,
\end{aligned}$$

where the last step holds due to the $(\gamma_2 - L_\rho)$-strong concavity of $F_d(\cdot;\bar{p}_t,\bar{q}_t,\bar{r}_t)$ and $z^+ = \bar{z}_t + \alpha_z\eta\nabla_z F_d(\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)$,

$$(\gamma_2-L_\rho)\|\bar{z}_t - z^*(\bar{p}_t,\bar{q}_t,\bar{r}_t)\| \le \|\nabla_z F_d(\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\| = \frac{1}{\alpha_z\eta}\|z^+-\bar{z}_t\| \,.$$

In addition, following Eq. (25), we derive

$$\begin{aligned}
\|\bar{z}_{t+1}-z^+\|^2 \le{}& \frac{2\alpha_z^2\eta^2L_g^2}{(\gamma_1-L_\rho)^2}\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2 \\
&+ 2\alpha_z^2\eta^2\|\nabla_z F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t) - \bar{w}_t\|^2 \,.
\end{aligned}$$

Combining above inequalities with Eqs. (25) and (26) completes the proof. □

Lemma D.5. *Given Assumptions 4.1-4.4, by setting $\eta \le \min\{\frac{1}{\alpha_p}, \frac{1}{\alpha_q}, \frac{1}{2\alpha_x L_{F,1}}\}$, and and defining $L_{F,1} := \gamma_1 + L_\rho$, $L_{F,2} := \gamma_2 + L_\rho$, the following inequality holds:*

$$\begin{aligned}
&F(\bar{x}_{t+1},\bar{y}_{t+1},\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_{t+1}) - F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t) \le -\frac{\gamma_1}{2\alpha_p\eta}\|\bar{p}_{t+1}-\bar{p}_t\|^2 \\
&- \frac{\gamma_1}{2\alpha_q\eta}\|\bar{q}_{t+1}-\bar{q}_t\|^2 + \frac{\gamma_2(1-(1-\alpha_r\eta)^2)}{2\alpha_r^2\eta^2}\|\bar{r}_{t+1}-\bar{r}_t\|^2 - \frac{\alpha_x\eta}{4}\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2 \\
&- \frac{\alpha_y\eta}{2}\|\nabla_y F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2 + \frac{\alpha_z\eta}{2}\|\nabla_z F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2 \qquad (28) \\
&+ \alpha_x\eta\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t) - \bar{u}_t\|^2 + \frac{\alpha_y\eta}{2}\|\nabla_y F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t) - \bar{v}_t\|^2 \\
&+ \Big(4\alpha_x^2\alpha_z\eta^3L_g^2 - \frac{\alpha_x\eta}{4}\Big)\|\bar{u}_t\|^2 + \Big(\frac{3\alpha_x\alpha_y^2\eta^3L_\rho^2}{2} - \frac{\alpha_y\eta}{4}\Big)\|\bar{v}_t\|^2 + \Big(\frac{\alpha_z^2\eta^2L_{F,2}}{2} + \frac{3\alpha_z\eta}{4}\Big)\|\bar{w}_t\|^2 \,.
\end{aligned}$$

Proof. At first, we decompose $F(\bar{x}_{t+1},\bar{y}_{t+1},\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_{t+1}) - F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)$ as follows:

$$F(\bar{x}_{t+1},\bar{y}_{t+1},\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_{t+1}) - F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)$$

$$
\begin{aligned}
&= \underbrace{F(\bar{x}_{t+1}, \bar{y}_{t+1}, \bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_{t+1}) - F(\bar{x}_{t+1}, \bar{y}_{t+1}, \bar{z}_{t+1}; \bar{p}_t, \bar{q}_t, \bar{r}_t)}_{T_0} \\
&\quad + \underbrace{F(\bar{x}_{t+1}, \bar{y}_{t+1}, \bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - F(\bar{x}_{t+1}, \bar{y}_{t+1}, \bar{z}_{t+1}; \bar{p}_t, \bar{q}_t, \bar{r}_t)}_{T_1} \\
&\quad + \underbrace{F(\bar{x}_{t+1}, \bar{y}_{t+1}, \bar{z}_{t+1}; \bar{p}_t, \bar{q}_t, \bar{r}_t) - F(\bar{x}_{t+1}, \bar{y}_{t+1}, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)}_{T_2} \\
&\quad + \underbrace{F(\bar{x}_{t+1}, \bar{y}_{t+1}, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - F(\bar{x}_t, \bar{y}_{t+1}, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)}_{T_3} \\
&\quad + \underbrace{F(\bar{x}_t, \bar{y}_{t+1}, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)}_{T_4} .
\end{aligned}
$$

Based on this decomposition, we bound T_0, T_1, T_2, T_3, and T_4, respectively.

Regarding T_0, it can be bounded as follows:

$$
T_0 = -\frac{\gamma_2}{2}\|\bar{z}_{t+1} - \bar{r}_{t+1}\|^2 + \frac{\gamma_2}{2}\|\bar{z}_{t+1} - \bar{r}_t\|^2 = \frac{\gamma_2(1-(1-\alpha_r\eta)^2)}{2\alpha_r^2\eta^2}\|\bar{r}_{t+1} - \bar{r}_t\|^2 ,
$$

Regarding T_1, it can be bounded as follows:

$$
\begin{aligned}
T_1 &= \frac{\gamma_1}{2}\|\bar{x}_{t+1} - \bar{p}_{t+1}\|^2 + \frac{\gamma_1}{2}\|\bar{y}_{t+1} - \bar{q}_{t+1}\|^2 - \frac{\gamma_1}{2}\|\bar{x}_{t+1} - \bar{p}_t\|^2 - \frac{\gamma_1}{2}\|\bar{y}_{t+1} - \bar{q}_t\|^2 \\
&= \frac{\gamma_1}{2}\left(\|\bar{x}_{t+1} - \bar{p}_{t+1}\|^2 - \|\bar{x}_{t+1} - \bar{p}_t\|^2\right) + \frac{\gamma_1}{2}\left(\|\bar{y}_{t+1} - \bar{q}_{t+1}\|^2 - \|\bar{y}_{t+1} - \bar{q}_t\|^2\right) \\
&= \frac{\gamma_1(1-\alpha_p\eta)^2 - 1)}{2}\|\bar{x}_{t+1} - \bar{p}_t\|^2 + \frac{\gamma_1(1-\alpha_q\eta)^2 - 1)}{2}\|\bar{y}_{t+1} - \bar{q}_t\|^2 \\
&= \frac{\gamma_1(1-\alpha_p\eta)^2 - 1)}{2\alpha_p^2\eta^2}\|\bar{p}_{t+1} - \bar{p}_t\|^2 + \frac{\gamma_1(1-\alpha_q\eta)^2 - 1)}{2\alpha_q^2\eta^2}\|\bar{q}_{t+1} - \bar{q}_t\|^2 \\
&\leq -\frac{\gamma_1}{2\alpha_p\eta}\|\bar{p}_{t+1} - \bar{p}_t\|^2 - \frac{\gamma_1}{2\alpha_q\eta}\|\bar{q}_{t+1} - \bar{q}_t\|^2 ,
\end{aligned}
$$

where the last step follows from $\eta < \frac{1}{\alpha_p}$ and $\eta < \frac{1}{\alpha_q}$.

Regarding T_2, it can be bounded as follows:

$$
\begin{aligned}
T_2 &\leq \langle \nabla_z F(\bar{x}_{t+1}, \bar{y}_{t+1}, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t), \bar{z}_{t+1} - \bar{z}_t \rangle + \frac{L_{F,2}}{2}\|\bar{z}_{t+1} - \bar{z}_t\|^2 \\
&= \alpha_z\eta\langle \nabla_z F(\bar{x}_{t+1}, \bar{y}_{t+1}, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t), \bar{w}_t \rangle + \frac{\alpha_z^2\eta^2 L_{F,2}}{2}\|\bar{w}_t\|^2 \\
&= \alpha_z\eta\langle \nabla_z F(\bar{x}_{t+1}, \bar{y}_{t+1}, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t), \bar{w}_t \rangle \\
&\quad + \alpha_z\eta\langle \nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t), \bar{w}_t \rangle + \frac{\alpha_z^2\eta^2 L_{F,2}}{2}\|\bar{w}_t\|^2 \\
&\leq 4\alpha_z\eta\|\nabla_z F(\bar{x}_{t+1}, \bar{y}_{t+1}, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 + \frac{\alpha_z\eta}{4}\|\bar{w}_t\|^2 \\
&\quad + \frac{\alpha_z\eta}{2}\|\nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 + \frac{\alpha_z\eta}{2}\|\bar{w}_t\|^2 + \frac{\alpha_z^2\eta^2 L_{F,2}}{2}\|\bar{w}_t\|^2 \\
&\leq 4\alpha_x^2\alpha_z\eta^3 L_g^2\|\bar{u}_t\|^2 + \frac{\alpha_z\eta}{2}\|\nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 + \left(\frac{\alpha_z^2\eta^2 L_{F,2}}{2} + \frac{3\alpha_z\eta}{4}\right)\|\bar{w}_t\|^2 ,
\end{aligned}
$$

where the last step holds due to

$$
\begin{aligned}
&\|\nabla_z F(\bar{x}_{t+1}, \bar{y}_{t+1}, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 \\
&= \|\nabla_2 g(\bar{x}_{t+1}, \bar{z}_t) - \nabla_2 g(\bar{x}_t, \bar{z}_t)\|^2 \leq L_g^2\|\bar{x}_{t+1} - \bar{x}_t\|^2 \leq \alpha_x^2\eta^2 L_g^2\|\bar{u}_t\|^2 .
\end{aligned}
$$

Regarding T_3, it can be bounded as follows:

$$
T_3 \leq \langle \nabla_x F(\bar{x}_t, \bar{y}_{t+1}, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t), \bar{x}_{t+1} - \bar{x}_t \rangle + \frac{L_{F,1}}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2
$$

$$
\begin{aligned}
&= -\alpha_x\eta\langle\nabla_x F(\bar{x}_t,\bar{y}_{t+1},\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t),\bar{u}_t\rangle + \frac{\alpha_x^2\eta^2 L_{F,1}}{2}\|\bar{u}_t\|^2 \\
&= \frac{\alpha_x\eta}{2}\|\nabla_x F(\bar{x}_t,\bar{y}_{t+1},\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)-\bar{u}_t\|^2 - \frac{\alpha_x\eta}{2}\|\nabla_x F(\bar{x}_t,\bar{y}_{t+1},\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2 \\
&\quad - \frac{\alpha_x\eta}{2}\|\bar{u}_t\|^2 + \frac{\alpha_x^2\eta^2 L_{F,1}}{2}\|\bar{u}_t\|^2 \\
&\le -\frac{\alpha_x\eta}{4}\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2 + \frac{\alpha_x\eta}{2}\|\nabla_x F(\bar{x}_t,\bar{y}_{t+1},\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)-\bar{u}_t\|^2 - \frac{\alpha_x\eta}{4}\|\bar{u}_t\|^2 \\
&\quad + \frac{\alpha_x\eta}{2}\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)-\nabla_x F(\bar{x}_t,\bar{y}_{t+1},\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2 \\
&\le -\frac{\alpha_x\eta}{4}\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2 + \alpha_x\eta\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)-\bar{u}_t\|^2 - \frac{\alpha_x\eta}{4}\|\bar{u}_t\|^2 \\
&\quad + \frac{3\alpha_x\eta}{2}\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)-\nabla_x F(\bar{x}_t,\bar{y}_{t+1},\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2 \\
&\le -\frac{\alpha_x\eta}{4}\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2 + \alpha_x\eta\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)-\bar{u}_t\|^2 - \frac{\alpha_x\eta}{4}\|\bar{u}_t\|^2 \\
&\quad + \frac{3\alpha_x\alpha_y^2\eta^3 L_\rho^2}{2}\|\bar{v}_t\|^2 ,
\end{aligned}
$$

where the fourth step is due to $-\|\nabla_x F(\bar{x}_t,\bar{y}_{t+1},\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2 \le -\frac{1}{2}\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2 + \|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)-\nabla_x F(\bar{x}_t,\bar{y}_{t+1},\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2$ and $\eta \le \frac{1}{2\alpha_x L_{F,1}}$.

Regarding T_4, it can be bounded as follows:

$$
\begin{aligned}
T_4 &\le \langle\nabla_y F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t),\bar{y}_{t+1}-\bar{y}_t\rangle + \frac{L_{F,1}}{2}\|\bar{y}_{t+1}-\bar{y}_t\|^2 \\
&= -\alpha_y\eta\langle\nabla_y F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t),\bar{v}_t\rangle + \frac{\alpha_y^2\eta^2 L_{F,1}}{2}\|\bar{v}_t\|^2 \\
&= \frac{\alpha_y\eta}{2}\|\nabla_y F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)-\bar{v}_t\|^2 - \frac{\alpha_y\eta}{2}\|\nabla_y F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2 \\
&\quad - \frac{\alpha_y\eta}{2}\|\bar{v}_t\|^2 + \frac{\alpha_y^2\eta^2 L_{F,1}}{2}\|\bar{v}_t\|^2 \\
&\le \frac{\alpha_y\eta}{2}\|\nabla_y F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)-\bar{v}_t\|^2 - \frac{\alpha_y\eta}{2}\|\nabla_y F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2 - \frac{\alpha_y\eta}{4}\|\bar{v}_t\|^2 ,
\end{aligned}
$$

where the last step is due to $\eta \le \frac{1}{2\alpha_y L_{F,1}}$.

By combining T_0, T_1, T_2, T_3, and T_4, we complete the proof.

□

Lemma D.6. *Given Assumptions 4.1-4.4, by defining*

$$F_d(\bar{z};\bar{p},\bar{q},\bar{r}) = F(x^*(\bar{z};\bar{p},\bar{q},\bar{r}),y^*(\bar{z};\bar{p},\bar{q},\bar{r}),\bar{z};\bar{p},\bar{q},\bar{r}) ,$$

$F_d(\bar{z};\bar{p},\bar{q},\bar{r})$ is L_d-smooth, where $L_d = L_\rho(1+\omega_2)+\gamma_2$, and the following inequality holds:

$$
\begin{aligned}
&F_d(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_{t+1}) - F_d(\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t) \ge \alpha_z\eta\langle\nabla_z F_d(\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t),\bar{w}_t\rangle - \frac{\alpha_z^2\eta^2 L_d}{2}\|w_t\|^2 \\
&+ \frac{\gamma_1}{2}\langle\bar{p}_{t+1}-\bar{p}_t,\bar{p}_{t+1}+\bar{p}_t-2x^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t)\rangle \qquad (29) \\
&+ \frac{\gamma_1}{2}\langle\bar{q}_{t+1}-\bar{q}_t,\bar{q}_{t+1}+\bar{q}_t-2y^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t)\rangle + \frac{\gamma_2(1-(1-\alpha_r\eta)^2)}{2\alpha_r^2\eta^2}\|\bar{r}_{t+1}-\bar{r}_t\|^2 .
\end{aligned}
$$

Proof. At first, we decompose $F_d(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_{t+1}) - F_d(\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)$ as follows:

$$
\begin{aligned}
&F_d(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_{t+1}) - F_d(\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t) = \underbrace{F_d(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_{t+1}) - F_d(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t)}_{T_0} \\
&+ \underbrace{F_d(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t) - F_d(\bar{z}_{t+1};\bar{p}_t,\bar{q}_t,\bar{r}_t)}_{T_1} + \underbrace{F_d(\bar{z}_{t+1};\bar{p}_t,\bar{q}_t,\bar{r}_t) - F_d(\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)}_{T_2} .
\end{aligned}
$$

In the following, we bound T_0, T_1 and T_2, respectively.

Regarding T_0, we derive

$$T_0 = -\frac{\gamma_2}{2}\|\bar{z}_{t+1} - \bar{r}_{t+1}\|^2 + \frac{\gamma_2}{2}\|\bar{z}_{t+1} - \bar{r}_t\|^2 = \frac{\gamma_2(1-(1-\alpha_r\eta)^2)}{2\alpha_r^2\eta^2}\|\bar{r}_{t+1} - \bar{r}_t\|^2 .$$

Regarding T_1, we derive:

$$\begin{aligned}
T_1 &= F(x^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t), y^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t), \bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t) \\
&\quad - F(x^*(\bar{z}_{t+1};\bar{p}_t,\bar{q}_t,\bar{r}_t), y^*(\bar{z}_{t+1};\bar{p}_t,\bar{q}_t,\bar{r}_t), \bar{z}_{t+1};\bar{p}_t,\bar{q}_t,\bar{r}_t) \\
&\geq F(x^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t), y^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t), \bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t) \\
&\quad - F(x^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t), y^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t), \bar{z}_{t+1};\bar{p}_t,\bar{q}_t,\bar{r}_t) \\
&\geq \frac{\gamma_1}{2}\left(\|x^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t) - \bar{p}_{t+1}\|^2 - \|x^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t) - \bar{p}_t\|^2\right) \\
&\quad + \frac{\gamma_1}{2}\left(\|y^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t) - \bar{q}_{t+1}\|^2 - \|y^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t) - \bar{q}_t\|^2\right) \\
&= \frac{\gamma_1}{2}\langle \bar{p}_{t+1} - \bar{p}_t, \bar{p}_{t+1} + \bar{p}_t - 2x^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t)\rangle \\
&\quad + \frac{\gamma_1}{2}\langle \bar{q}_{t+1} - \bar{q}_t, \bar{q}_{t+1} + \bar{q}_t - 2y^*(\bar{z}_{t+1};\bar{p}_{t+1},\bar{q}_{t+1},\bar{r}_t)\rangle .
\end{aligned}$$

Regarding T_2, it can be bounded as follows:

$$\begin{aligned}
T_2 &\geq \langle \nabla_z F_d(\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t), \bar{z}_{t+1} - \bar{z}_t\rangle - \frac{L_d}{2}\|\bar{z}_{t+1} - \bar{z}_t\|^2 \\
&= \alpha_z\eta\langle \nabla_z F_d(\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t), \bar{w}_t\rangle - \frac{\alpha_z^2\eta^2 L_d}{2}\|w_t\|^2 .
\end{aligned}$$

By combining T_0, T_1, T_2, we complete the proof. □

Lemma D.7. *Given Assumptions 4.1-4.4, by defining*

$$\begin{aligned}
H(p,q,r) &= \min_{x,y}\max_z F(x,y,z;p,q,r) = \min_{x,y} F_p(x,y;p,q,r) , \\
H_z(p,q) &= \max_r H(p,q,r) , \quad r^*(p,q) = \arg\max_r H(p,q,r) ,
\end{aligned}$$

and denoting $\hat{r} := r^*(\bar{p}_{t+1},\bar{q}_{t+1})$, the following inequality holds:

$$\begin{aligned}
H_z(\bar{p}_{t+1},\bar{q}_{t+1}) - H_z(\bar{p}_t,\bar{q}_t) &\leq \frac{\gamma_1}{2}\langle \bar{p}_{t+1} - \bar{p}_t, \bar{p}_{t+1} + \bar{p}_t - 2x^*(\bar{p}_t,\bar{q}_t,\hat{r})\rangle \\
&\quad + \frac{\gamma_1}{2}\langle \bar{q}_{t+1} - \bar{q}_t, \bar{q}_{t+1} + \bar{q}_t - 2y^*(\bar{p}_t,\bar{q}_t,\hat{r})\rangle . \qquad (30)
\end{aligned}$$

Proof. Since $H_z(\bar{p}_t,\bar{q}_t) = H(\bar{p}_t,\bar{q}_t,r^*(\bar{p}_t,\bar{q}_t))$, denoting $\hat{r} := r^*(\bar{p}_{t+1},\bar{q}_{t+1})$, we derive

$$\begin{aligned}
&H_z(\bar{p}_{t+1},\bar{q}_{t+1}) - H_z(\bar{p}_t,\bar{q}_t) = H(\bar{p}_{t+1},\bar{q}_{t+1},\hat{r}) - H(\bar{p}_t,\bar{q}_t,r^*(\bar{p}_t,\bar{q}_t)) \\
&\leq H(\bar{p}_{t+1},\bar{q}_{t+1},\hat{r}) - H(\bar{p}_t,\bar{q}_t,\hat{r}) \\
&= F_p(x^*(\bar{p}_{t+1},\bar{q}_{t+1},\hat{r}), y^*(\bar{p}_{t+1},\bar{q}_{t+1},\hat{r});\bar{p}_{t+1},\bar{q}_{t+1},\hat{r}) - F_p(x^*(\bar{p}_t,\bar{q}_t,\hat{r}), y^*(\bar{p}_t,\bar{q}_t,\hat{r});\bar{p}_t,\bar{q}_t,\hat{r}) \\
&\leq F_p(x^*(\bar{p}_t,\bar{q}_t,\hat{r}), y^*(\bar{p}_t,\bar{q}_t,\hat{r});\bar{p}_{t+1},\bar{q}_{t+1},\hat{r}) - F_p(x^*(\bar{p}_t,\bar{q}_t,\hat{r}), y^*(\bar{p}_t,\bar{q}_t,\hat{r});\bar{p}_t,\bar{q}_t,\hat{r}) .
\end{aligned}$$

Hence, only the (p,q)-smoothing terms differ:

$$\begin{aligned}
H_z(\bar{p}_{t+1},\bar{q}_{t+1}) - H_z(\bar{p}_t,\bar{q}_t) &\leq \frac{\gamma_1}{2}\left(\|x^*(\bar{p}_t,\bar{q}_t,\hat{r}) - \bar{p}_{t+1}\|^2 - \|x^*(\bar{p}_t,\bar{q}_t,\hat{r}) - \bar{p}_t\|^2\right) \\
&\quad + \frac{\gamma_1}{2}\left(\|y^*(\bar{p}_t,\bar{q}_t,\hat{r}) - \bar{q}_{t+1}\|^2 - \|y^*(\bar{p}_t,\bar{q}_t,\hat{r}) - \bar{q}_t\|^2\right) \\
&= \frac{\gamma_1}{2}\langle \bar{p}_{t+1} - \bar{p}_t, \bar{p}_{t+1} + \bar{p}_t - 2x^*(\bar{p}_t,\bar{q}_t,\hat{r})\rangle + \frac{\gamma_1}{2}\langle \bar{q}_{t+1} - \bar{q}_t, \bar{q}_{t+1} + \bar{q}_t - 2y^*(\bar{p}_t,\bar{q}_t,\hat{r})\rangle .
\end{aligned}$$

□

Lemma D.8. *Given Assumptions 4.1-4.4, by defining*

$$\Psi_{t+1} = F(\bar{x}_{t+1}, \bar{y}_{t+1}, \bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_{t+1}) - 2F_d(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_{t+1}) + 2H_z(\bar{p}_{t+1}, \bar{q}_{t+1}) \,, \tag{31}$$

and setting $\alpha_{pq} = \alpha_p = \alpha_q$,

$$\alpha_z = \alpha_x \underbrace{\frac{(\gamma_1 - L_\rho)^2}{72L_g^2}}_{c_{\alpha_z}=O(1)}, \quad \alpha_{pq} = \alpha_x \underbrace{\frac{(\gamma_1 - L_\rho)^4(\gamma_2 - L_\rho)^2\mu}{72 \times 768\gamma_1 L_g^2(5L_g^2\mu + 16\gamma_2^2(\gamma_1 - L_\rho))}}_{c_{\alpha_{pq}}=O(1/\kappa)},$$
$$\alpha_r = \alpha_x \underbrace{\frac{(\gamma_1 - L_\rho)^4(\gamma_2 - L_\rho)^4}{72 \times 24 \times 640\gamma_1\gamma_2^2 L_g^4}}_{c_{\alpha_r}=O(1)}, \quad \alpha_x \le \min\Big\{\frac{9}{80\gamma_1}, \frac{\sqrt{\mu(\gamma_1 - L_\rho)^3}(\gamma_2 - L_\rho)^2}{24 \times 16\sqrt{6\gamma_1 c_{\alpha_{pq}}} L_g c_{\alpha_r}\gamma_2^2}\Big\},$$

and denoting $C_r = \frac{4\gamma_2^2}{\mu(\gamma_1 - L_\rho)}\big(\omega_3^2 + \frac{(1-\alpha_r\eta)^2}{\alpha_r^2\eta^2}\big)$, $C_\rho = 16\gamma_1\alpha_{pq}\eta\big(5\omega_2^2 + 20\omega_2^2\omega_3^2\alpha_r^2\eta^2 + 4C_r\alpha_r^2\eta^2\big)$, *the following inequality holds:*

$$\begin{aligned}
\Psi_{t+1} - \Psi_t \le & -\frac{\alpha_x\eta}{8}\|\nabla_x F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 - \frac{\alpha_y\eta}{2}\|\nabla_y F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 \\
& - \frac{\alpha_z\eta}{2}\|\nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 + \Big(16\gamma_1\alpha_{pq}\eta\big(5\omega_2^2\omega_3^2 + C_r\big) - \frac{\gamma_2}{2\alpha_r\eta}\Big)\|\bar{r}_{t+1} - \bar{r}_t\|^2 \\
& + \Big(1 + \omega_1^2 + 8\alpha_{pq}\eta\omega_5^2\big(5\omega_2^2 + \frac{2\gamma_2^2}{\mu(\gamma_1 - L_\rho)}\big) - \frac{1}{4\alpha_{pq}\eta}\Big)\gamma_1\big(\|\bar{p}_{t+1} - \bar{p}_t\|^2 + \|\bar{q}_{t+1} - \bar{q}_t\|^2\big) \\
& + \alpha_x\eta\|\nabla_x F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{u}_t\|^2 + \frac{\alpha_y\eta}{2}\|\nabla_y F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{v}_t\|^2 \\
& + \frac{3\alpha_z\eta}{2}\|\nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{w}_t\|^2 + \Big(4\alpha_x^2\alpha_z\eta^3 L_g^2 - \frac{\alpha_x\eta}{4}\Big)\|\bar{u}_t\|^2 \\
& + \Big(\frac{3\alpha_x\alpha_y^2\eta^3 L_\rho^2}{2} - \frac{\alpha_y\eta}{4}\Big)\|\bar{v}_t\|^2 + \Big(\frac{\alpha_z^2\eta^2}{2}(L_{F,2} + 2L_d + 2C_\rho) - \frac{\alpha_z\eta}{16}\Big)\|\bar{w}_t\|^2 \,.
\end{aligned} \tag{32}$$

Proof. Firstly, based on Lemmas D.5, D.6, and D.7, we have

$$\begin{aligned}
\Psi_{t+1} - \Psi_t \le & -\frac{\gamma_1}{2\alpha_p\eta}\|\bar{p}_{t+1} - \bar{p}_t\|^2 - \frac{\gamma_1}{2\alpha_q\eta}\|\bar{q}_{t+1} - \bar{q}_t\|^2 - \frac{\gamma_2}{2\alpha_r\eta}\|\bar{r}_{t+1} - \bar{r}_t\|^2 \\
& - \frac{\alpha_x\eta}{4}\|\nabla_x F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 - \frac{\alpha_y\eta}{2}\|\nabla_y F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 \\
& + \frac{\alpha_z\eta}{2}\|\nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 + \alpha_x\eta\|\nabla_x F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{u}_t\|^2 \\
& + \frac{\alpha_y\eta}{2}\|\nabla_y F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{v}_t\|^2 + \Big(4\alpha_x^2\alpha_z\eta^3 L_g^2 - \frac{\alpha_x\eta}{4}\Big)\|\bar{u}_t\|^2 \\
& + \Big(\frac{3\alpha_x\alpha_y^2\eta^3 L_\rho^2}{2} - \frac{\alpha_y\eta}{4}\Big)\|\bar{v}_t\|^2 + \Big(\frac{\alpha_z^2\eta^2 L_{F,2}}{2} + \frac{3\alpha_z\eta}{4} + \alpha_z^2\eta^2 L_d\Big)\|\bar{w}_t\|^2 \\
& - 2\alpha_z\eta\langle\nabla_z F_d(\bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t), \bar{w}_t\rangle \\
& + 2\gamma_1\langle\bar{p}_{t+1} - \bar{p}_t,\ x^*(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - x^*(\bar{p}_t, \bar{q}_t, \hat{r})\rangle \\
& + 2\gamma_1\langle\bar{q}_{t+1} - \bar{q}_t,\ y^*(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - y^*(\bar{p}_t, \bar{q}_t, \hat{r})\rangle \,.
\end{aligned} \tag{33}$$

In particular, we obtain

$$\begin{aligned}
& - 2\langle\nabla_z F_d(\bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t), \bar{w}_t\rangle \\
= & -2\langle\nabla_z F(x^*(\bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t), y^*(\bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t), \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t), \bar{w}_t\rangle \\
= & -2\langle\nabla_z F(x^*(\bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t), y^*(\bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t), \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t), \bar{w}_t\rangle \\
& - 2\langle\nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t), \bar{w}_t\rangle \\
\le & \, 8\|\nabla_z F(x^*(\bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t), y^*(\bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t), \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 \\
& + \frac{1}{8}\|\bar{w}_t\|^2 + \|\nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{w}_t\|^2 - \|\nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 - \|\bar{w}_t\|^2
\end{aligned}$$

$$\leq \frac{8L_g^2}{(\gamma_1 - L_\rho)^2}\|\nabla_x F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t)\|^2 + \|\nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t) - \bar{w}_t\|^2$$
$$- \|\nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t)\|^2 - \frac{7}{8}\|\bar{w}_t\|^2 . \tag{34}$$

In addition, we derive

$$\begin{aligned}
&2\gamma_1\langle \bar{p}_{t+1} - \bar{p}_t, x^*(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - x^*(\bar{p}_t, \bar{q}_t, \hat{r})\rangle \\
&= 2\gamma_1\langle \bar{p}_{t+1} - \bar{p}_t, x^*(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - x^*(\bar{p}_{t+1}, \bar{q}_{t+1}, \hat{r})\rangle \\
&\quad + 2\gamma_1\langle \bar{p}_{t+1} - \bar{p}_t, x^*(\bar{p}_{t+1}, \bar{q}_{t+1}, \hat{r}) - x^*(\bar{p}_t, \bar{q}_t, \hat{r})\rangle \\
&\leq \frac{\gamma_1}{4\alpha_p\eta}\|\bar{p}_{t+1} - \bar{p}_t\|^2 + 4\gamma_1\alpha_p\eta\|x^*(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - x^*(\bar{p}_{t+1}, \bar{q}_{t+1}, \hat{r})\|^2 \\
&\quad + \gamma_1\|\bar{p}_{t+1} - \bar{p}_t\|^2 + \gamma_1\|x^*(\bar{p}_{t+1}, \bar{q}_{t+1}, \hat{r}) - x^*(\bar{p}_t, \bar{q}_t, \hat{r})\|^2 .
\end{aligned} \tag{35}$$

Similarly, we derive

$$\begin{aligned}
&2\gamma_1\langle \bar{q}_{t+1} - \bar{q}_t, y^*(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - y^*(\bar{p}_t, \bar{q}_t, \hat{r})\rangle \\
&\leq \frac{\gamma_1}{4\alpha_q\eta}\|\bar{q}_{t+1} - \bar{q}_t\|^2 + 4\gamma_1\alpha_q\eta\|y^*(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - y^*(\bar{p}_{t+1}, \bar{q}_{t+1}, \hat{r})\|^2 \\
&\quad + \gamma_1\|\bar{q}_{t+1} - \bar{q}_t\|^2 + \gamma_1\|y^*(\bar{p}_{t+1}, \bar{q}_{t+1}, \hat{r}) - y^*(\bar{p}_t, \bar{q}_t, \hat{r})\|^2 .
\end{aligned}$$

Combining these two inequalities, and applying Lemma D.1 on the last two terms, we obtain

$$\begin{aligned}
&2\gamma_1\langle \bar{p}_{t+1} - \bar{p}_t, x^*(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - x^*(\bar{p}_t, \bar{q}_t, \hat{r})\rangle \\
&+ 2\gamma_1\langle \bar{q}_{t+1} - \bar{q}_t, y^*(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - y^*(\bar{p}_t, \bar{q}_t, \hat{r})\rangle \\
&\leq \frac{\gamma_1}{4\alpha_p\eta}\|\bar{p}_{t+1} - \bar{p}_t\|^2 + 4\gamma_1\alpha_p\eta\|x^*(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - x^*(\bar{p}_{t+1}, \bar{q}_{t+1}, \hat{r})\|^2 \\
&\quad + \frac{\gamma_1}{4\alpha_q\eta}\|\bar{q}_{t+1} - \bar{q}_t\|^2 + 4\gamma_1\alpha_q\eta\|y^*(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - y^*(\bar{p}_{t+1}, \bar{q}_{t+1}, \hat{r})\|^2 \\
&\quad + \gamma_1\|\bar{p}_{t+1} - \bar{p}_t\|^2 + \gamma_1\omega_1^2\|\bar{p}_{t+1} - \bar{p}_t\|^2 + \gamma_1\|\bar{q}_{t+1} - \bar{q}_t\|^2 + \gamma_1\omega_1^2\|\bar{q}_{t+1} - \bar{q}_t\|^2 \\
&= \Big(\frac{1}{4\alpha_p\eta} + 1 + \omega_1^2\Big)\gamma_1\|\bar{p}_{t+1} - \bar{p}_t\|^2 + \Big(\frac{1}{4\alpha_q\eta} + 1 + \omega_1^2\Big)\gamma_1\|\bar{q}_{t+1} - \bar{q}_t\|^2 \\
&\quad + 4\gamma_1\alpha_p\eta\|x^*(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - x^*(\bar{p}_{t+1}, \bar{q}_{t+1}, \hat{r})\|^2 \\
&\quad + 4\gamma_1\alpha_q\eta\|y^*(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - y^*(\bar{p}_{t+1}, \bar{q}_{t+1}, \hat{r})\|^2 .
\end{aligned} \tag{36}$$

Applying Lemma D.3 and Lemma D.4, we obtain

$$\begin{aligned}
&\|x^*(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - x^*(\bar{p}_{t+1}, \bar{q}_{t+1}, \hat{r})\|^2 + \|y^*(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - y^*(\bar{p}_{t+1}, \bar{q}_{t+1}, \hat{r})\|^2 \\
&\leq 2\|x^*(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - x^*(\bar{p}_{t+1}, \bar{q}_{t+1}, r^+)\|^2 + 2\|y^*(\bar{z}_{t+1}; \bar{p}_{t+1}, \bar{q}_{t+1}, \bar{r}_t) - y^*(\bar{p}_{t+1}, \bar{q}_{t+1}, r^+)\|^2 \\
&\quad + 2\|x^*(\bar{p}_{t+1}, \bar{q}_{t+1}, r^+) - x^*(\bar{p}_{t+1}, \bar{q}_{t+1}, \hat{r})\|^2 + 2\|y^*(\bar{p}_{t+1}, \bar{q}_{t+1}, r^+) - y^*(\bar{p}_{t+1}, \bar{q}_{t+1}, \hat{r})\|^2 \\
&\leq \Big[20\omega_2^2\alpha_z^2\eta^2\big(1 + 2\omega_3^2\alpha_r^2\eta^2\big) + \frac{20\omega_2^2\big(1 + 4\omega_3^2\alpha_r^2\eta^2\big)}{(\gamma_2 - L_\rho)^2} + 8C_r\alpha_r^2\eta^2\Big(\alpha_z^2\eta^2 + \frac{2}{(\gamma_2 - L_\rho)^2}\Big)\Big]\|\bar{w}_t\|^2 \\
&\quad + \Big[\frac{60\omega_2^2\alpha_z^2\eta^2L_g^2}{(\gamma_1 - L_\rho)^2} + \frac{40\omega_2^2L_g^2\big(1 + 4\omega_3^2\alpha_r^2\eta^2\big) + 32C_r\alpha_r^2\eta^2L_g^2}{(\gamma_1 - L_\rho)^2(\gamma_2 - L_\rho)^2}\Big]\|\nabla_x F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2 \\
&\quad + \Big[60\omega_2^2\alpha_z^2\eta^2 + \frac{40\omega_2^2\big(1 + 4\omega_3^2\alpha_r^2\eta^2\big) + 32C_r\alpha_r^2\eta^2}{(\gamma_2 - L_\rho)^2}\Big]\|\nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{w}_t\|^2 \\
&\quad + \Big(10\omega_2^2\omega_5^2 + \frac{4\omega_5^2\gamma_2^2}{\mu(\gamma_1 - L_\rho)}\Big)(\|\bar{p}_{t+1} - \bar{p}_t\|^2 + \|\bar{q}_{t+1} - \bar{q}_t\|^2) + \Big(20\omega_2^2\omega_3^2 + 4C_r\Big)\|\bar{r}_{t+1} - \bar{r}_t\|^2 .
\end{aligned}$$

Combining Eqs. (35), (36), (37), and denoting $\alpha_{pq} = \alpha_p = \alpha_q$, $C_\rho = 16\gamma_1\alpha_{pq}\eta\big(5\omega_2^2 + 20\omega_2^2\omega_3^2\alpha_r^2\eta^2 + 4C_r\alpha_r^2\eta^2\big)$, we obtain

$$\Psi_{t+1} - \Psi_t \leq \Big(1 + \omega_1^2 + 8\alpha_{pq}\eta\omega_5^2\big(5\omega_2^2 + \frac{2\gamma_2^2}{\mu(\gamma_1 - L_\rho)}\big) - \frac{1}{4\alpha_{pq}\eta}\Big)\gamma_1\big(\|\bar{p}_{t+1} - \bar{p}_t\|^2 + \|\bar{q}_{t+1} - \bar{q}_t\|^2\big)$$

$$
\begin{aligned}
&+\Big(16\gamma_1\alpha_{pq}\eta\big(5\omega_2^2\omega_3^2+C_r\big)-\frac{\gamma_2}{2\alpha_r\eta}\Big)\|\bar{r}_{t+1}-\bar{r}_t\|^2\\
&+\Big[\frac{L_g^2}{(\gamma_1-L_\rho)^2}\Big(8\alpha_z\eta+240\gamma_1\alpha_{pq}\eta\alpha_z^2\eta^2\omega_2^2+\frac{2C_\rho}{(\gamma_2-L_\rho)^2}\Big)-\frac{\alpha_x\eta}{4}\Big]\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2\\
&-\frac{\alpha_y\eta}{2}\|\nabla_y F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2-\frac{\alpha_z\eta}{2}\|\nabla_z F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2\\
&+\alpha_x\eta\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)-\bar{u}_t\|^2+\frac{\alpha_y\eta}{2}\|\nabla_y F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)-\bar{v}_t\|^2\\
&+\Big(\alpha_z\eta+240\gamma_1\alpha_{pq}\eta\alpha_z^2\eta^2\omega_2^2+\frac{2C_\rho}{(\gamma_2-L_\rho)^2}\Big)\|\nabla_z F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)-\bar{w}_t\|^2\\
&+\Big(4\alpha_x^2\alpha_z\eta^3L_g^2-\frac{\alpha_x\eta}{4}\Big)\|\bar{u}_t\|^2+\Big(\frac{3\alpha_x\alpha_y^2\eta^3L_\rho^2}{2}-\frac{\alpha_y\eta}{4}\Big)\|\bar{v}_t\|^2\\
&+\Big(\frac{\alpha_z^2\eta^2}{2}(L_{F,2}+2L_d+2C_\rho)+\frac{C_\rho}{(\gamma_2-L_\rho)^2}-\frac{\alpha_z\eta}{8}\Big)\|\bar{w}_t\|^2\,. \qquad (37)
\end{aligned}
$$

For the coefficient of $\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2$, we enforce

$$
\frac{L_g^2}{(\gamma_1-L_\rho)^2}\Big(8\alpha_z\eta+240\gamma_1\alpha_{pq}\eta\alpha_z^2\eta^2\omega_2^2+\frac{2C_\rho}{(\gamma_2-L_\rho)^2}\Big)-\frac{\alpha_x\eta}{4}\le-\frac{\alpha_x\eta}{8}\,.
$$

In particular, from $C_r=\frac{4\gamma_2^2}{\mu(\gamma_1-L_\rho)}\big(\omega_3^2+\frac{(1-\alpha_r\eta)^2}{\alpha_r^2\eta^2}\big)$, and $C_\rho=16\gamma_1\alpha_{pq}\eta\big(5\omega_2^2+20\omega_2^2\omega_3^2\alpha_r^2\eta^2+4C_r\alpha_r^2\eta^2\big)$, we enforce

$$
\begin{aligned}
&\frac{8\alpha_z\eta L_g^2}{(\gamma_1-L_\rho)^2}\le\frac{\alpha_x\eta}{9}\,,\quad \frac{240\gamma_1\alpha_{pq}\eta\alpha_z^2\eta^2\omega_2^2L_g^2}{(\gamma_1-L_\rho)^2}\le\frac{\alpha_x\eta}{24\times8}\,,\\
&\frac{32\gamma_1\alpha_{pq}\eta L_g^2}{(\gamma_1-L_\rho)^2}\Big(\frac{5\omega_2^2}{(\gamma_2-L_\rho)^2}+\frac{4(1-\alpha_r\eta)^2}{(\gamma_2-L_\rho)^2}\frac{4\gamma_2^2}{\mu(\gamma_1-L_\rho)}\Big)\le\frac{\alpha_x\eta}{72\times24}\,,\\
&\frac{32\gamma_1\alpha_{pq}\eta L_g^2}{(\gamma_1-L_\rho)^2}\frac{20\omega_2^2\omega_3^2\alpha_r^2\eta^2}{(\gamma_2-L_\rho)^2}\le\frac{\alpha_x\eta}{72\times24}\,,\\
&\frac{32\gamma_1\alpha_{pq}\eta L_g^2}{(\gamma_1-L_\rho)^2}\frac{4\alpha_r^2\eta^2}{(\gamma_2-L_\rho)^2}\frac{4\gamma_2^2\omega_3^2}{\mu(\gamma_1-L_\rho)}\le\frac{\alpha_x\eta}{72\times24}\,. \qquad (38)
\end{aligned}
$$

By solving these inequalities, we obtain

$$
\begin{aligned}
&\alpha_z=\alpha_x\underbrace{\frac{(\gamma_1-L_\rho)^2}{72L_g^2}}_{c_{\alpha_z}=O(1)}\,,\quad \alpha_{pq}=\alpha_x\underbrace{\frac{(\gamma_1-L_\rho)^4(\gamma_2-L_\rho)^2\mu}{72\times768\gamma_1L_g^2(5L_g^2\mu+16\gamma_2^2(\gamma_1-L_\rho))}}_{c_{\alpha_{pq}}=O(1/\kappa)}\,,\\
&\alpha_r=\alpha_x\underbrace{\frac{(\gamma_1-L_\rho)^4(\gamma_2-L_\rho)^4}{72\times24\times640\gamma_1\gamma_2^2L_g^4}}_{c_{\alpha_r}=O(1)}\,,\quad \alpha_x\le\min\Big\{\frac{9}{80\gamma_1}\,,\frac{\sqrt{\mu(\gamma_1-L_\rho)^3}(\gamma_2-L_\rho)^2}{24\times16\sqrt{6\gamma_1c_{\alpha_{pq}}}L_gc_{\alpha_r}\gamma_2^2}\Big\}\,.
\end{aligned}
$$

In addition, for the coefficient of $\|\nabla_z F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)-\bar{w}_t\|^2$ and $\|\bar{w}_t\|^2$, we obtain

$$
\begin{aligned}
&\frac{C_\rho}{(\gamma_2-L_\rho)^2}\le\frac{\alpha_x\eta(\gamma_1-L_\rho)^2}{72\times16L_g^2}=\frac{\alpha_z\eta}{16}\,,\\
&\alpha_z\eta+240\gamma_1\alpha_{pq}\eta\alpha_z^2\eta^2\omega_2^2+\frac{2C_\rho}{(\gamma_2-L_\rho)^2}\\
&\le\alpha_z\eta+\frac{3\alpha_x\eta(\gamma_1-L_\rho)^2}{72\times8L_g^2}+\frac{\alpha_x\eta(\gamma_1-L_\rho)^2}{72\times8L_g^2}\le\frac{3\alpha_z\eta}{2}\,.
\end{aligned}
$$

Hence, the proof is completed. □

E FEDERATED LEMMAS

Lemma E.1. *Given Assumptions 4.1-4.4, we have the following inequalities:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)}-x_t^{(k)}\|^2] \le 6\alpha_x^2\eta^2\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t}^{(k)}-\bar{u}_{1,t}\|^2] + 6\alpha_x^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t}^{(k)}-\bar{u}_{2,t}\|^2] + 6\alpha_x^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,t}^{(k)}-\bar{u}_{3,t}\|^2] + 6\alpha_x^2\eta^2\mathbb{E}[\|\bar{u}_t\|^2]\,,$$

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_{t+1}^{(k)}-y_t^{(k)}\|^2] \le 4\alpha_y^2\eta^2\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{1,t}^{(k)}-\bar{v}_{1,t}\|^2] + 4\alpha_y^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{2,t}^{(k)}-\bar{v}_{2,t}\|^2] + 4\alpha_y^2\eta^2\mathbb{E}[\|\bar{v}_t\|^2]\,,$$

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|z_{t+1}^{(k)}-z_t^{(k)}\|^2] \le 2\alpha_z^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|w_t^{(k)}-\bar{w}_t\|^2] + 2\alpha_z^2\eta^2\mathbb{E}[\|\bar{w}_t\|^2]\,.$$

Proof. Due to $x_{t+1}^{(k)} = x_t^{(k)} - \alpha_x\eta u_t^{(k)}$, we have

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)}-x_t^{(k)}\|^2] = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\alpha_x\eta u_t^{(k)}\|^2]\\
&\le 2\alpha_x^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_t^{(k)}-\bar{u}_t\|^2] + 2\alpha_x^2\eta^2\mathbb{E}[\|\bar{u}_t\|^2]\\
&= 2\alpha_x^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\rho u_{1,t}^{(k)} + (u_{2,t}^{(k)}-u_{3,t}^{(k)}) - \rho\bar{u}_{1,t} - (\bar{u}_{2,t}-\bar{u}_{3,t})\|^2] + 2\alpha_x^2\eta^2\mathbb{E}[\|\bar{u}_t\|^2]\\
&\le 6\alpha_x^2\eta^2\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t}^{(k)}-\bar{u}_{1,t}\|^2] + 6\alpha_x^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t}^{(k)}-\bar{u}_{2,t}\|^2]\\
&\quad + 6\alpha_x^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,t}^{(k)}-\bar{u}_{3,t}\|^2] + 6\alpha_x^2\eta^2\mathbb{E}[\|\bar{u}_t\|^2]\,.
\end{aligned}$$

Similarly, due to $y_{t+1}^{(k)} = y_t^{(k)} - \alpha_y\eta v_t^{(k)}$, we have

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_{t+1}^{(k)}-y_t^{(k)}\|^2] = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\alpha_y\eta v_t^{(k)}\|^2]\\
&\le 2\alpha_y^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_t^{(k)}-\bar{v}_t\|^2] + 2\alpha_y^2\eta^2\mathbb{E}[\|\bar{v}_t\|^2]\\
&= 2\alpha_y^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\rho v_{1,t}^{(k)} + v_{2,t}^{(k)} - \rho\bar{v}_{1,t} - \bar{v}_{2,t}\|^2] + 2\alpha_y^2\eta^2\mathbb{E}[\|\bar{v}_t\|^2]\\
&\le 4\alpha_y^2\eta^2\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{1,t}^{(k)}-\bar{v}_{1,t}\|^2] + 4\alpha_y^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{2,t}^{(k)}-\bar{v}_{2,t}\|^2] + 4\alpha_y^2\eta^2\mathbb{E}[\|\bar{v}_t\|^2]\,.
\end{aligned}$$

Moreover, due to $z_{t+1}^{(k)} = z_t^{(k)} - \alpha_z\eta w_t^{(k)}$, we have

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|z_{t+1}^{(k)}-z_t^{(k)}\|^2] = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\alpha_z\eta w_t^{(k)}\|^2]$$

$$\leq 2\alpha_z^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|w_t^{(k)}-\bar{w}_t\|^2]+2\alpha_z^2\eta^2\mathbb{E}[\|\bar{w}_t\|^2]\,.$$

□

Lemma E.2. *Given Assumptions 4.1-4.4, by denoting $s_t=\lfloor t/\tau\rfloor$, we have the following inequalities:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_t^{(k)}-\bar{x}_t\|^2]\leq 3\tau^2\alpha_x^2\eta^2\Bigg(\rho^2\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t}^{(k)}-\bar{u}_{1,t}\|^2]$$
$$+\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t}^{(k)}-\bar{u}_{2,t}\|^2]+\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,t}^{(k)}-\bar{u}_{3,t}\|^2]\Bigg)\,,$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_t^{(k)}-\bar{y}_t\|^2]\leq 2\tau^2\alpha_y^2\eta^2\Bigg(\rho^2\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{1,t}^{(k)}-\bar{v}_{1,t}\|^2]$$
$$+\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{2,t}^{(k)}-\bar{v}_{2,t}\|^2]\Bigg)\,,$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|z_t^{(k)}-\bar{z}_t\|^2]\leq\tau^2\alpha_z^2\eta^2\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|w_t^{(k)}-\bar{w}_t\|^2]\,.$$

Proof. By denoting $s_t=\lfloor t/\tau\rfloor$, we have

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_t^{(k)}-\bar{x}_t\|^2]$$
$$=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{s_t\tau}^{(k)}-\alpha_x\eta\sum_{t'=s_t\tau}^{t-1}u_{t'}^{(k)}-\bar{x}_{s_t\tau}+\alpha_x\eta\sum_{t'=s_t\tau}^{t-1}\bar{u}_{t'}\|^2]$$
$$=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|-\alpha_x\eta\sum_{t'=s_t\tau}^{t-1}u_{t'}^{(k)}+\alpha_x\eta\sum_{t'=s_t\tau}^{t-1}\bar{u}_{t'}\|^2]$$
$$\leq\tau\alpha_x^2\eta^2\sum_{t'=s_t\tau}^{t-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|-u_{t'}^{(k)}+\bar{u}_{t'}\|^2]$$
$$=\tau\alpha_x^2\eta^2\sum_{t'=s_t\tau}^{t-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\rho u_{1,t'}^{(k)}+(u_{2,t'}^{(k)}-u_{3,t'}^{(k)})-\rho\bar{u}_{1,t'}-(\bar{u}_{2,t'}-\bar{u}_{3,t'})\|^2]$$
$$\leq 3\tau\alpha_x^2\eta^2\rho^2\sum_{t'=s_t\tau}^{t-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t'}^{(k)}-\bar{u}_{1,t'}\|^2]+3\tau\alpha_x^2\eta^2\sum_{t'=s_t\tau}^{t-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t'}^{(k)}-\bar{u}_{2,t'}\|^2]$$
$$+3\tau\alpha_x^2\eta^2\sum_{t'=s_t\tau}^{t-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,t'}^{(k)}-\bar{u}_{3,t'}\|^2]\,.$$

The other two inequalities can be proved in the same way. Therefore, we omit their detailed proof.

□

Lemma E.3. *Given Assumptions 4.1-4.4, by setting $\eta\leq\min\{\frac{1}{\sqrt{\beta_x}},\frac{1}{\sqrt{\beta_y}},\frac{1}{\sqrt{\beta_z}}\}$, we have the following inequalities:*

$$\mathbb{E}[\|\mathcal{G}_{u,1,t+1}\|^2]\leq(1-\beta_x\eta^2)\mathbb{E}[\|\mathcal{G}_{u,1,t}\|^2]+2L_f^2\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)}-x_t^{(k)}\|^2]$$

$$+ 2L_f^2 \frac{1}{K^2} \sum_{k=1}^{K} \mathbb{E}[\|y_{t+1}^{(k)} - y_t^{(k)}\|^2] + 2\beta_x^2 \eta^4 \sigma^2 \frac{1}{K} ,$$

$$\mathbb{E}[\|\mathcal{G}_{v,1,t+1}\|^2] \leq (1 - \beta_y \eta^2) \mathbb{E}[\|\mathcal{G}_{v,1,t}\|^2] + 2L_f^2 \frac{1}{K^2} \sum_{k=1}^{K} \mathbb{E}[\|x_{t+1}^{(k)} - x_t^{(k)}\|^2]$$
$$+ 2L_f^2 \frac{1}{K^2} \sum_{k=1}^{K} \mathbb{E}[\|y_{t+1}^{(k)} - y_t^{(k)}\|^2] + 2\beta_y^2 \eta^4 \sigma^2 \frac{1}{K} ,$$

$$\mathbb{E}[\|\mathcal{G}_{u,3,t+1}\|^2] \leq (1 - \beta_x \eta^2) \mathbb{E}[\|\mathcal{G}_{u,3,t}\|^2] + 2L_g^2 \frac{1}{K^2} \sum_{k=1}^{K} \mathbb{E}[\|x_{t+1}^{(k)} - x_t^{(k)}\|^2]$$
$$+ 2L_g^2 \frac{1}{K^2} \sum_{k=1}^{K} \mathbb{E}[\|z_{t+1}^{(k)} - z_t^{(k)}\|^2] + 2\beta_x^2 \eta^4 \sigma^2 \frac{1}{K} ,$$

Proof. According to the update rule of $u_{1,t+1}^{(k)}$, we have

$$\mathbb{E}[\|\mathcal{G}_{u,1,t+1}\|^2] = \mathbb{E}[\| \frac{1}{K} \sum_{k=1}^{K} \nabla_1 f^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}) - \frac{1}{K} \sum_{k=1}^{K} u_{1,t+1}^{(k)} \|^2]$$

$$= \mathbb{E}[\| \frac{1}{K} \sum_{k=1}^{K} ((1 - \beta_x \eta^2)(u_{1,t}^{(k)} - \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_{t+1}^{(k)})) + \nabla_1 f^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; \xi_{t+1}^{(k)}))$$
$$- \frac{1}{K} \sum_{k=1}^{K} \nabla_1 f^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}) \|^2]$$

$$\leq (1 - \beta_x \eta^2)^2 \mathbb{E}[\| \frac{1}{K} \sum_{k=1}^{K} u_{1,t}^{(k)} - \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}) \|^2]$$
$$+ 2 \frac{1}{K^2} \sum_{k=1}^{K} \mathbb{E}[\|\nabla_1 f^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; \xi_{t+1}^{(k)}) - \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_{t+1}^{(k)})\|^2]$$
$$+ 2\beta_x^2 \eta^4 \frac{1}{K^2} \sum_{k=1}^{K} \mathbb{E}[\|\nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_{t+1}^{(k)}) - \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)})\|^2]$$

$$\leq (1 - \beta_x \eta^2) \mathbb{E}[\|\mathcal{G}_{u,1,t}\|^2] + 2L_f^2 \frac{1}{K^2} \sum_{n=1}^{N} \mathbb{E}[\|x_{t+1}^{(k)} - x_t^{(k)}\|^2]$$
$$+ 2L_f^2 \frac{1}{K^2} \sum_{n=1}^{N} \mathbb{E}[\|y_{t+1}^{(k)} - y_t^{(k)}\|^2] + 2\beta_x^2 \eta^4 \sigma^2 \frac{1}{K} . \tag{39}$$

Following the proof of Eq. (39), we can prove the other inequalities. Therefore, we omit the detailed proof to save space. □

Lemma E.4. *Given Assumptions 4.1-4.4, by setting* $\eta \leq \min\{\frac{1}{\sqrt{\beta_x}}, \frac{1}{\sqrt{\beta_y}}\}$*, we have the following inequalities:*

$$\mathbb{E}[\|\mathcal{G}_{u,2,t+1}\|^2] \leq (1 - \beta_x \eta^2) \mathbb{E}[\|\mathcal{G}_{u,2,t}\|^2] + 2L_g^2 \frac{1}{K^2} \sum_{k=1}^{K} \mathbb{E}[\|x_{t+1}^{(k)} - x_t^{(k)}\|^2]$$
$$+ 2L_g^2 \frac{1}{K^2} \sum_{k=1}^{K} \mathbb{E}[\|y_{t+1}^{(k)} - y_t^{(k)}\|^2] + 2\beta_x^2 \eta^4 \sigma^2 \frac{1}{K} ,$$

$$\mathbb{E}[\|\mathcal{G}_{v,2,t+1}\|^2] \le (1-\beta_y\eta^2)\mathbb{E}[\|\mathcal{G}_{v,2,t}\|^2] + 2L_g^2\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)} - x_t^{(k)}\|^2] + 2L_g^2\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}[\|y_{t+1}^{(k)} - y_t^{(k)}\|^2] + 2\beta_y^2\eta^4\sigma^2\frac{1}{K}\,.$$

$$\mathbb{E}[\|\mathcal{G}_{w,t+1}\|^2] \le (1-\beta_z\eta^2)\mathbb{E}[\|\mathcal{G}_{w,t}\|^2] + 2L_g^2\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)} - x_t^{(k)}\|^2] + 2L_g^2\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}[\|z_{t+1}^{(k)} - z_t^{(k)}\|^2] + 2\beta_z^2\eta^4\sigma^2\frac{1}{K}\,.$$

Proof. According to the update rule of $u_{2,t+1}^{(k)}$, we have

$$\begin{aligned}
&\mathbb{E}[\|\mathcal{G}_{u,2,t+1}\|^2] = \mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\nabla_1 g_{\gamma_1}^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; p_{t+1}^{(k)}, q_{t+1}^{(k)})) - \frac{1}{K}\sum_{k=1}^{K}u_{2,t+1}^{(k)}\|^2] \\
&= \mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}((1-\beta_x\eta^2)(u_{2,t}^{(k)} - \nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)}; \zeta_{t+1}^{(k)})) \\
&\qquad + \nabla_1 g_{\gamma_1}^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; p_{t+1}^{(k)}, q_{t+1}^{(k)}; \zeta_{t+1}^{(k)})) - \frac{1}{K}\sum_{k=1}^{K}\nabla_1 g_{\gamma_1}^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; p_{t+1}^{(k)}, q_{t+1}^{(k)})\|^2] \\
&\le (1-\beta_x\eta^2)^2\mathbb{E}[\|\bar{u}_{2,t} - \frac{1}{K}\sum_{k=1}^{K}\nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)})\|^2] \\
&\quad + 2\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)}) - \nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)}; \zeta_{t+1}^{(k)}) \\
&\qquad + \nabla_1 g_{\gamma_1}^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; p_{t+1}^{(k)}, q_{t+1}^{(k)}; \zeta_{t+1}^{(k)}) - \nabla_1 g_{\gamma_1}^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; p_{t+1}^{(k)}, q_{t+1}^{(k)})\|^2] \\
&\quad + 2\beta_x^2\eta^4\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)}; \zeta_{t+1}^{(k)}) - \nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)})\|^2]\,.
\end{aligned}$$

Follows Assumption 4.4, the last term can be bounded by $2\beta_x^2\eta^4\sigma^2\frac{1}{K}$.

In addition the second term can be bounded as follows

$$\begin{aligned}
&2\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)}) - \nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)}; \zeta_{t+1}^{(k)}) \\
&\qquad + \nabla_1 g_{\gamma_1}^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; p_{t+1}^{(k)}, q_{t+1}^{(k)}; \zeta_{t+1}^{(k)}) - \nabla_1 g_{\gamma_1}^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; p_{t+1}^{(k)}, q_{t+1}^{(k)})\|^2] \\
&= 2\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_t^{(k)}, y_t^{(k)}) + \gamma_1(x_t^{(k)} - p_t^{(k)}) - \nabla_1 g^{(k)}(x_t^{(k)}, y_t^{(k)}; \zeta_{t+1}^{(k)}) - \gamma_1(x_t^{(k)} - p_t^{(k)}) \\
&\qquad + \nabla_1 g^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; \zeta_{t+1}^{(k)}) + \gamma_1(x_{t+1}^{(k)} - p_{t+1}^{(k)}) - \nabla_1 g^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}) - \gamma_1(x_{t+1}^{(k)} - p_{t+1}^{(k)})\|^2] \\
&= 2\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_t^{(k)}, y_t^{(k)}) - \nabla_1 g^{(k)}(x_t^{(k)}, y_t^{(k)}; \zeta_{t+1}^{(k)}) \\
&\qquad + \nabla_1 g^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; \zeta_{t+1}^{(k)}) - \nabla_1 g^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)})\|^2] \\
&\le 2\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; \zeta_{t+1}^{(k)}) - \nabla_1 g^{(k)}(x_t^{(k)}, y_t^{(k)}; \zeta_{t+1}^{(k)})\|^2]
\end{aligned}$$

$$\leq 2L_g^2 \frac{1}{K^2} \sum_{k=1}^{K} \mathbb{E}[\|x_{t+1}^{(k)} - x_t^{(k)}\|^2] + 2L_g^2 \frac{1}{K^2} \sum_{k=1}^{K} \mathbb{E}[\|y_{t+1}^{(k)} - y_t^{(k)}\|^2] \,.$$

Combining these inequalities completes the proof.

Using a similar argument, we can establish the other inequalities. Hence, we omit the details. □

Lemma E.5. *Given Assumptions 4.1-4.5, by setting $\eta \leq \min\{\frac{1}{\sqrt{\beta_x}}, \frac{1}{\sqrt{\beta_y}}\}$, we have the following inequalities:*

$$\begin{aligned}
(i).\ &\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|u_{1,t+1}^{(k)} - \bar{u}_{1,t+1}\|^2] \leq (1 + \frac{1}{\tau} + 24\tau L_f^2 \alpha_x^2 \eta^2 \rho^2) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|u_{1,t}^{(k)} - \bar{u}_{1,t}\|^2] \\
&+ 24\tau L_f^2 \alpha_x^2 \eta^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|u_{2,t}^{(k)} - \bar{u}_{2,t}\|^2] + 24\tau L_f^2 \alpha_x^2 \eta^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|u_{3,t}^{(k)} - \bar{u}_{3,t}\|^2] \\
&+ 16\tau L_f^2 \alpha_y^2 \eta^2 \rho^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|v_{1,t}^{(k)} - \bar{v}_{1,t}\|^2] + 16\tau L_f^2 \alpha_y^2 \eta^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|v_{2,t}^{(k)} - \bar{v}_{2,t}\|^2] \\
&+ 24\tau \beta_x^2 \eta^4 L_f^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|x_t^{(k)} - \bar{x}_t\|^2] + 24\tau \beta_x^2 \eta^4 L_f^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|y_t^{(k)} - \bar{y}_t\|^2] \\
&+ 24\tau L_f^2 \alpha_x^2 \eta^2 \mathbb{E}[\|\bar{u}_t\|^2] + 16\tau L_f^2 \alpha_y^2 \eta^2 \mathbb{E}[\|\bar{v}_t\|^2] + 12\tau \beta_x^2 \eta^4 \delta^2 + 8\tau \beta_x^2 \eta^4 \sigma^2 \,,
\end{aligned}$$

$$\begin{aligned}
(ii).\ &\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|v_{1,t+1}^{(k)} - \bar{v}_{1,t+1}\|^2] \leq (1 + \frac{1}{\tau} + 16\tau L_f^2 \alpha_y^2 \eta^2 \rho^2) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|v_{1,t}^{(k)} - \bar{v}_{1,t}\|^2] \\
&+ 24\tau L_f^2 \alpha_x^2 \eta^2 \rho^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|u_{1,t}^{(k)} - \bar{u}_{1,t}\|^2] + 24\tau L_f^2 \alpha_x^2 \eta^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|u_{2,t}^{(k)} - \bar{u}_{2,t}\|^2] \\
&+ 24\tau L_f^2 \alpha_x^2 \eta^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|u_{3,t}^{(k)} - \bar{u}_{3,t}\|^2] + 16\tau L_f^2 \alpha_y^2 \eta^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|v_{2,t}^{(k)} - \bar{v}_{2,t}\|^2] \\
&+ 24\tau \beta_y^2 \eta^4 L_f^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|x_t^{(k)} - \bar{x}_t\|^2] + 24\tau \beta_y^2 \eta^4 L_f^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|y_t^{(k)} - \bar{y}_t\|^2] \\
&+ 24\tau L_f^2 \alpha_x^2 \eta^2 \mathbb{E}[\|\bar{u}_t\|^2] + 16\tau L_f^2 \alpha_y^2 \eta^2 \mathbb{E}[\|\bar{v}_t\|^2] + 12\tau \beta_y^2 \eta^4 \delta^2 + 8\tau \beta_y^2 \eta^4 \sigma^2 \,,
\end{aligned}$$

$$\begin{aligned}
(iii).\ &\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|u_{3,t+1}^{(k)} - \bar{u}_{3,t+1}\|^2] \leq (1 + \frac{1}{\tau} + 24\tau L_g^2 \alpha_x^2 \eta^2) \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|u_{3,t}^{(k)} - \bar{u}_{3,t}\|^2] \\
&+ 24\tau L_g^2 \alpha_x^2 \eta^2 \rho^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|u_{1,t}^{(k)} - \bar{u}_{1,t}\|^2] + 24\tau L_g^2 \alpha_x^2 \eta^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|u_{2,t}^{(k)} - \bar{u}_{2,t}\|^2] \\
&+ 8\tau L_g^2 \alpha_z^2 \eta^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|w_t^{(k)} - \bar{w}_t\|^2] + 24\tau L_g^2 \alpha_x^2 \eta^2 \mathbb{E}[\|\bar{u}_t\|^2] + 8\tau L_g^2 \alpha_z^2 \eta^2 \mathbb{E}[\|\bar{w}_t\|^2] \\
&+ 24\tau \beta_x^2 \eta^4 L_g^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|x_t^{(k)} - \bar{x}_t\|^2] + 24\tau \beta_x^2 \eta^4 L_g^2 \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|z_t^{(k)} - \bar{z}_t\|^2] \\
&+ 12\tau \beta_x^2 \eta^4 \delta^2 + 8\tau \beta_x^2 \eta^4 \sigma^2 \,,
\end{aligned}$$

Proof. According to the update rule of $u_{1,t+1}^{(k)}$ and $\bar{u}_{1,t+1}$, we have

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|u_{1,t+1}^{(k)} - \bar{u}_{1,t+1}\|^2]$$

$$
\begin{aligned}
&= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\Big((1-\beta_x\eta^2)(u_{1,t}^{(k)} - \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_{t+1}^{(k)})) + \nabla_1 f^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; \xi_{t+1}^{(k)})\Big) \\
&\quad - \frac{1}{K}\sum_{k'=1}^{K}\Big((1-\beta_x\eta^2)(u_{1,t}^{(k')} - \nabla_1 f^{(k')}(x_t^{(k')}, y_t^{(k')}; \xi_{t+1}^{(k')})) + \nabla_1 f^{(k')}(x_{t+1}^{(k')}, y_{t+1}^{(k')}; \xi_{t+1}^{(k')})\Big)\|^2] \\
&\le (1-\beta_x\eta^2)^2(1+\frac{1}{\tau})\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t}^{(k)} - \frac{1}{K}\sum_{k'=1}^{K}u_{1,t}^{(k')}\|^2] \\
&\quad + 2(1+\tau)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\| - \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_{t+1}^{(k)}) + \nabla_1 f^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; \xi_{t+1}^{(k)})\|^2] \\
&\quad + 2(1+\tau)\beta_x^2\eta^4\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_{t+1}^{(k)}) - \frac{1}{K}\sum_{k'=1}^{K}\nabla_1 f^{(k')}(x_t^{(k')}, y_t^{(k')}; \xi_{t+1}^{(k')})\|^2] \\
&\le (1-\beta_x\eta^2)^2(1+\frac{1}{\tau})\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t}^{(k)} - \bar{u}_{1,t}\|^2] \\
&\quad + 2(1+\tau)L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)} - x_t^{(k)}\|^2] + 2(1+\tau)L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_{t+1}^{(k)} - y_t^{(k)}\|^2] \qquad (40) \\
&\quad + 2(1+\tau)\beta_x^2\eta^4\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_{t+1}^{(k)}) - \frac{1}{K}\sum_{k'=1}^{K}\nabla_1 f^{(k')}(x_t^{(k')}, y_t^{(k')}; \xi_{t+1}^{(k')})\|^2] \,.
\end{aligned}
$$

For the last term in Eq. (40), we obtain

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_{t+1}^{(k)}) - \frac{1}{K}\sum_{k'=1}^{K}\nabla_1 f^{(k')}(x_t^{(k')}, y_t^{(k')}; \xi_{t+1}^{(k')})\|^2] \\
&\le \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_{t+1}^{(k)}) - \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}) + \nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}) \\
&\quad - \frac{1}{K}\sum_{k'=1}^{K}\nabla_1 f^{(k')}(x_t^{(k')}, y_t^{(k')}) + \frac{1}{K}\sum_{k'=1}^{K}\nabla_1 f^{(k')}(x_t^{(k')}, y_t^{(k')}) \\
&\quad - \frac{1}{K}\sum_{k'=1}^{K}\nabla_1 f^{(k')}(x_t^{(k')}, y_t^{(k')}; \xi_{t+1}^{(k')})\|^2] \\
&\le 2\sigma^2 + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}) - \nabla_1 f^{(k)}(\bar{x}, \bar{y}) + \nabla_1 f^{(k)}(\bar{x}, \bar{y}) - \frac{1}{K}\sum_{k'=1}^{K}\nabla_1 f^{(k')}(\bar{x}_t, \bar{y}_t) \\
&\qquad\qquad + \frac{1}{K}\sum_{k'=1}^{K}\nabla_1 f^{(k')}(\bar{x}_t, \bar{y}_t) - \frac{1}{K}\sum_{k'=1}^{K}\nabla_1 f^{(k')}(x_t^{(k')}, y_t^{(k')})\|^2] \\
&\le 2\sigma^2 + 3\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}) - \nabla_1 f^{(k)}(\bar{x}, \bar{y})\|^2] \\
&\quad + 3\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 f^{(k)}(\bar{x}, \bar{y}) - \frac{1}{K}\sum_{k'=1}^{K}\nabla_1 f^{(k')}(\bar{x}_t, \bar{y}_t)\|^2] \\
&\quad + 3\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\frac{1}{K}\sum_{k'=1}^{K}\nabla_1 f^{(k')}(\bar{x}_t, \bar{y}_t) - \frac{1}{K}\sum_{k'=1}^{K}\nabla_1 f^{(k')}(x_t^{(k')}, y_t^{(k')})\|^2] \\
&\le 2\sigma^2 + 6L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_t^{(k)} - \bar{x}_t\|^2] + 6L_f^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_t^{(k)} - \bar{y}_t\|^2] + 3\delta^2 \,. \qquad (41)
\end{aligned}
$$

Thus, by substituting these inequalities into Eq. (40) with $1+\tau \leq 2\tau$, and applying Lemma E.1, the proof is complete.

The other inequalities can be proved in the same way. Therefore, we omit the detailed proof. □

Lemma E.6. *Given Assumptions 4.1-4.5, by setting* $\eta \leq \min\{\frac{1}{\sqrt{\beta_x}}, \frac{1}{\sqrt{\beta_y}}\}$*, we have the following inequalities:*

$$
\begin{aligned}
(i).\ \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t+1}^{(k)}-\bar{u}_{2,t+1}\|^2] &\leq (1+\frac{1}{\tau}+72\tau(L_g^2+\gamma_1^2)\alpha_x^2\eta^2)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t}^{(k)}-\bar{u}_{2,t}\|^2] \\
&+ 72\tau(L_g^2+\gamma_1^2)\alpha_x^2\eta^2\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t}^{(k)}-\bar{u}_{1,t}\|^2] + 72\tau(L_g^2+\gamma_1^2)\alpha_x^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,t}^{(k)}-\bar{u}_{3,t}\|^2] \\
&+ 48\tau L_g^2\alpha_y^2\eta^2\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{1,t}^{(k)}-\bar{v}_{1,t}\|^2] + 48\tau L_g^2\alpha_y^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{2,t}^{(k)}-\bar{v}_{2,t}\|^2] \\
&+ 36\tau\gamma_1^2\alpha_p^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)}-\bar{x}_{t+1}\|^2] + 36\tau\gamma_1^2\mathbb{E}[\|\bar{p}_{t+1}-\bar{p}_t\|^2] \\
&+ 12\tau\beta_x^2\eta^4(6L_g^2+\gamma_1^2)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_t^{(k)}-\bar{x}_t\|^2] + 72\tau\beta_x^2\eta^4L_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_t^{(k)}-\bar{y}_t\|^2] \\
&+ 12\tau\gamma_1^2\Big(3\alpha_p^2\eta^2+\beta_x^2\eta^4\Big)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t-p_t^{(k)}\|^2] + 72\tau(L_g^2+\gamma_1^2)\alpha_x^2\eta^2\mathbb{E}[\|\bar{u}_t\|^2] \\
&+ 48\tau L_g^2\alpha_y^2\eta^2\mathbb{E}[\|\bar{v}_t\|^2] + 24\tau\beta_x^2\eta^4\sigma^2 + 36\tau\beta_x^2\eta^4\delta^2\ .
\end{aligned}
$$

$$
\begin{aligned}
(ii).\ \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{2,t+1}^{(k)}-\bar{v}_{2,t+1}\|^2] &\leq (1+\frac{1}{\tau}+48\tau(L_g^2+\gamma_1^2)\alpha_y^2\eta^2)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{2,t}^{(k)}-\bar{v}_{2,t}\|^2] \\
&+ 72\tau L_g^2\alpha_x^2\eta^2\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t}^{(k)}-\bar{u}_{1,t}\|^2] + 72\tau L_g^2\alpha_x^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t}^{(k)}-\bar{u}_{2,t}\|^2] \\
&+ 72\tau L_g^2\alpha_x^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,t}^{(k)}-\bar{u}_{3,t}\|^2] + 48\tau(L_g^2+\gamma_1^2)\alpha_y^2\eta^2\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{1,t}^{(k)}-\bar{v}_{1,t}\|^2] \\
&+ 36\tau\gamma_1^2\alpha_q^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_{t+1}^{(k)}-\bar{y}_{t+1}\|^2] + 36\tau\gamma_1^2\mathbb{E}[\|\bar{q}_{t+1}-\bar{q}_t\|^2] \\
&+ 72\tau\beta_y^2\eta^4L_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_t^{(k)}-\bar{x}_t\|^2] + 12\tau\beta_y^2\eta^4(6L_g^2+\gamma_1^2)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_t^{(k)}-\bar{y}_t\|^2] \\
&+ 12\tau\gamma_1^2\Big(3\alpha_q^2\eta^2+\beta_y^2\eta^4\Big)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{q}_t-q_t^{(k)}\|^2] + 72\tau L_g^2\alpha_x^2\eta^2\mathbb{E}[\|\bar{u}_t\|^2] \\
&+ 48\tau(L_g^2+\gamma_1^2)\alpha_y^2\eta^2\mathbb{E}[\|\bar{v}_t\|^2] + 24\tau\beta_y^2\eta^4\sigma^2 + 36\tau\beta_y^2\eta^4\delta^2\ .
\end{aligned}
$$

$$
\begin{aligned}
(iii).\ \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|w_{t+1}^{(k)}-\bar{w}_{t+1}\|^2] &\leq \Big(1+\frac{1}{\tau}+24\tau(L_g^2+\gamma_2^2)\alpha_z^2\eta^2\Big)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|w_t^{(k)}-\bar{w}_t\|^2] \\
&+ 72\tau L_g^2\alpha_x^2\eta^2\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t}^{(k)}-\bar{u}_{1,t}\|^2] + 72\tau L_g^2\alpha_x^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t}^{(k)}-\bar{u}_{2,t}\|^2] \\
&+ 72\tau L_g^2\alpha_x^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,t}^{(k)}-\bar{u}_{3,t}\|^2] + 36\tau\gamma_2^2\alpha_\tau^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|z_{t+1}^{(k)}-\bar{z}_{t+1}\|^2]
\end{aligned}
$$

$$+ 36\tau\gamma_2^2\mathbb{E}[\|\bar{r}_{t+1} - \bar{r}_t\|^2] + 12\tau\gamma_2^2\Big(3\alpha_r^2\eta^2 + \beta_z^2\eta^4\Big)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{r}_t - r_t^{(k)}\|^2]$$

$$+ 72\tau\beta_z^2\eta^4 L_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_t^{(k)} - \bar{x}_t\|^2] + 12\tau\beta_z^2\eta^4(6L_g^2 + \gamma_2^2)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|z_t^{(k)} - \bar{z}_t\|^2]$$

$$+ 72\tau L_g^2\alpha_x^2\eta^2\mathbb{E}[\|\bar{u}_t\|^2] + 24\tau(L_g^2 + \gamma_2^2)\alpha_z^2\eta^2\mathbb{E}[\|\bar{w}_t\|^2] + 24\tau\beta_z^2\eta^4\sigma^2 + 36\tau\beta_z^2\eta^4\delta^2\,.$$

Proof. According to the update rule of $u_{2,t+1}^{(k)}$ and $\bar{u}_{2,t+1}$, we have

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t+1}^{(k)} - \bar{u}_{2,t+1}\|^2] \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\Big((1-\beta_x\eta^2)(u_{2,t}^{(k)} - \nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)}; \zeta_{t+1}^{(k)})) \\
&\quad + \nabla_1 g_{\gamma_1}^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; p_{t+1}^{(k)}, q_{t+1}^{(k)}; \zeta_{t+1}^{(k)})\Big) - \frac{1}{K}\sum_{k=1}^{N}\Big((1-\beta_x\eta^2)(u_{2,t}^{(k')} \\
&\quad - \nabla_1 g_{\gamma_1}^{(k')}(x_t^{(k')}, y_t^{(k')}; p_t^{(k')}, q_t^{(k')}; \zeta_{t+1}^{(k')})) + \nabla_1 g_{\gamma_1}^{(k')}(x_{t+1}^{(k')}, y_{t+1}^{(k')}; p_{t+1}^{(k')}, q_{t+1}^{(k')}; \zeta_{t+1}^{(k')})\Big)\|^2] \\
&\le (1-\beta_x\eta^2)^2(1+\frac{1}{\tau})\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t}^{(k)} - \bar{u}_{2,t}\|^2] \\
&\quad + 2(1+\tau)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g_{\gamma_1}^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; p_{t+1}^{(k)}, q_{t+1}^{(k)}; \zeta_{t+1}^{(k)}) \\
&\quad\quad - \nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)}; \zeta_{t+1}^{(k)})\|^2] \\
&\quad + 2\beta_x^2\eta^4(1+\tau)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)}; \zeta_{t+1}^{(k)}) \\
&\quad\quad - \frac{1}{K}\sum_{k'=1}^{K}\nabla_1 g_{\gamma_1}^{(k')}(x_t^{(k')}, y_t^{(k')}; p_t^{(k')}, q_t^{(k')}; \zeta_{t+1}^{(k')})\|^2]\,.
\end{aligned} \tag{42}$$

For the second term in Eq. (42), we obtain

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g_{\gamma_1}^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; p_{t+1}^{(k)}, q_{t+1}^{(k)}; \zeta_{t+1}^{(k)}) - \nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)}; \zeta_{t+1}^{(k)})\|^2] \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; \zeta_{t+1}^{(k)}) + \gamma_1(x_{t+1}^{(k)} - p_{t+1}^{(k)}) - \nabla_1 g^{(k)}(x_t^{(k)}, y_t^{(k)}; \zeta_{t+1}^{(k)}) \\
&\quad\quad - \gamma_1(x_t^{(k)} - p_t^{(k)})\|^2] \\
&\le 3\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; \zeta_{t+1}^{(k)}) - \nabla_1 g^{(k)}(x_t^{(k)}, y_t^{(k)}; \zeta_{t+1}^{(k)})\|^2] \\
&\quad + 3\gamma_1^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)} - x_t^{(k)}\|^2] + 3\gamma_1^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_{t+1}^{(k)} - p_t^{(k)}\|^2] \\
&\le 3(L_g^2 + \gamma_1^2)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)} - x_t^{(k)}\|^2] + 3L_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)} - x_t^{(k)}\|^2] \\
&\quad + 3\gamma_1^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_{t+1}^{(k)} - p_t^{(k)}\|^2]\,.
\end{aligned}$$

Moreover, we have

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_{t+1}^{(k)}-p_t^{(k)}\|^2]=\alpha_p^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)}-p_t^{(k)}\|^2]$$
$$\leq 3\alpha_p^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)}-\bar{x}_{t+1}\|^2]+3\alpha_p^2\eta^2\mathbb{E}[\|\bar{x}_{t+1}-\bar{p}_t\|^2]+3\alpha_p^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t-p_t^{(k)}\|^2]$$
$$=3\alpha_p^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)}-\bar{x}_{t+1}\|^2]+3\mathbb{E}[\|\bar{p}_{t+1}-\bar{p}_t\|^2]+3\alpha_p^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t-p_t^{(k)}\|^2]\,.$$

For the third term in Eq. (42), we obtain

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)},y_t^{(k)};p_t^{(k)},q_t^{(k)};\zeta_{t+1}^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}\nabla_1 g_{\gamma_1}^{(k')}(x_t^{(k')},y_t^{(k')};p_t^{(k')},q_t^{(k')};\zeta_{t+1}^{(k')})\|^2]$$
$$\leq\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_t^{(k)},y_t^{(k)};\zeta_{t+1}^{(k)})+\gamma_1(x_t^{(k)}-p_t^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}\nabla_1 g^{(k')}(x_t^{(k')},y_t^{(k')};\zeta_{t+1}^{(k')})$$
$$-\frac{1}{K}\sum_{k'=1}^{K}\gamma_1(x_t^{(k')}-p_t^{(k')})\|^2]$$
$$\leq 3\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(x_t^{(k)},y_t^{(k)};\zeta_{t+1}^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}\nabla_1 g^{(k')}(x_t^{(k')},y_t^{(k')};\zeta_{t+1}^{(k')})\|^2]$$
$$+3\gamma_1^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_t^{(k)}-\bar{x}_t\|^2]+3\gamma_1^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_t^{(k)}-\bar{p}_t\|^2]$$
$$\leq 6\sigma^2+9\delta^2+(18L_g^2+3\gamma_1^2)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_t^{(k)}-\bar{x}_t\|^2]+18L_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_t^{(k)}-\bar{y}_t\|^2]$$
$$+3\gamma_1^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_t^{(k)}-\bar{p}_t\|^2]\,,$$

where the last step follows Eq. (41). Thus, by substituting these inequalities into Eq. (42) with $1+\tau\leq 2\tau$, and applying Lemma E.1, the proof is complete.

Using a similar argument, we can establish the other inequalities. Hence, we omit the details.

□

Lemma E.7. *Given Assumptions 4.1-4.5, by defining $L_{\rho,\gamma_1}^2=L_f^2\rho^2+4L_g^2+3\gamma_1^2$, and setting*

$$\eta\leq\frac{1}{360\tau L_{\rho,\gamma_1}}\,,\alpha_x\leq 10\,,\alpha_y\leq 10\,,\alpha_z\leq 10\,,\alpha_p\leq 10\gamma_1\,,\alpha_q\leq 10\gamma_1\,,$$
$$\beta_x\leq\frac{L_{\rho,\gamma_1}^2}{K}\,,\beta_y\leq\frac{L_{\rho,\gamma_1}^2}{K}\,,\beta_z\leq\frac{L_{\rho,\gamma_1}^2}{K}\,, \tag{43}$$

we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\Bigg(\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t}^{(k)}-\bar{u}_{1,t}\|^2]+\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t}^{(k)}-\bar{u}_{2,t}\|^2]+\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,t}^{(k)}-\bar{u}_{3,t}\|^2]$$
$$+\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{1,t}^{(k)}-\bar{v}_{1,t}\|^2]+\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{2,t}^{(k)}-\bar{v}_{2,t}\|^2]+\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|w_t^{(k)}-\bar{w}_t\|^2]\Bigg)$$
$$\leq 288\tau^2\alpha_x^2\eta^2L_{\rho,\gamma}^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{u}_t\|^2]+192\tau^2\alpha_y^2\eta^2L_{\rho,\gamma}^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{v}_t\|^2]+48\tau^2\alpha_z^2\eta^2L_{\rho,\gamma}^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{w}_t\|^2]$$

$$+ 432\tau^2\gamma_1^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{p}_{t+1}-\bar{p}_t\|^2] + 432\tau^2\gamma_1^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{q}_{t+1}-\bar{q}_t\|^2]$$
$$+ 432\tau^2\gamma_2^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{r}_{t+1}-\bar{r}_t\|^2] + 3\tau S\,,$$

where the definition of S is in Eq. (45).

Proof. From Lemma E.5, Lemma E.6, and define $L_{\rho,\gamma}^2 = L_f^2\rho^2 + 5L_g^2 + 3(\gamma_1^2+\gamma_2^2)$, we obtain

$$\begin{aligned}
&\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t+1}^{(k)}-\bar{u}_{1,t+1}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t+1}^{(k)}-\bar{u}_{2,t+1}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,t+1}^{(k)}-\bar{u}_{3,t+1}\|^2]\\
&+ \rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{1,t+1}^{(k)}-\bar{v}_{1,t+1}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{2,t+1}^{(k)}-\bar{v}_{2,t+1}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|w_{t+1}^{(k)}-\bar{w}_{t+1}\|^2]\\
\le{}& \Big(1+\frac{1}{\tau}+48\tau\alpha_x^2\eta^2L_{\rho,\gamma}^2\Big)\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t}^{(k)}-\bar{u}_{1,t}\|^2] + 48\tau\alpha_x^2\eta^2L_{\rho,\gamma}^2\mathbb{E}[\|\bar{u}_t\|^2]\\
&+ \Big(1+\frac{1}{\tau}+48\tau\alpha_x^2\eta^2L_{\rho,\gamma}^2\Big)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t}^{(k)}-\bar{u}_{2,t}\|^2] + 32\tau\alpha_y^2\eta^2L_{\rho,\gamma}^2\mathbb{E}[\|\bar{v}_t\|^2]\\
&+ \Big(1+\frac{1}{\tau}+48\tau\alpha_x^2\eta^2L_{\rho,\gamma}^2\Big)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,t}^{(k)}-\bar{u}_{3,t}\|^2] + 8\tau L_{\rho,\gamma}^2\alpha_z^2\eta^2\mathbb{E}[\|\bar{w}_t\|^2]\\
&+ \Big(1+\frac{1}{\tau}+32\tau\alpha_y^2\eta^2L_{\rho,\gamma}^2\Big)\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{1,t}^{(k)}-\bar{v}_{1,t}\|^2] + S_x\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_t^{(k)}-\bar{x}_t\|^2]\\
&+ \Big(1+\frac{1}{\tau}+32\tau\alpha_y^2\eta^2L_{\rho,\gamma}^2\Big)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{2,t}^{(k)}-\bar{v}_{2,t}\|^2] + S_y\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_t^{(k)}-\bar{y}_t\|^2]\\
&+ \Big(1+\frac{1}{\tau}+8\tau L_{\rho,\gamma}^2\alpha_z^2\eta^2\Big)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|w_t^{(k)}-\bar{w}_t\|^2] + S_z\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|z_t^{(k)}-\bar{z}_t\|^2]\\
&+ S_p\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t-p_t^{(k)}\|^2] + S_q\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{q}_t-q_t^{(k)}\|^2] + S_r\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{r}_t-r_t^{(k)}\|^2]\\
&+ 36\tau\gamma_1^2\alpha_p^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)}-\bar{x}_{t+1}\|^2] + 36\tau\gamma_1^2\alpha_q^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_{t+1}^{(k)}-\bar{y}_{t+1}\|^2]\\
&+ 36\tau\gamma_2^2\alpha_r^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|z_{t+1}^{(k)}-\bar{z}_{t+1}\|^2] + 36\tau\gamma_1^2\mathbb{E}[\|\bar{p}_{t+1}-\bar{p}_t\|^2]\\
&+ 36\tau\gamma_1^2\mathbb{E}[\|\bar{q}_{t+1}-\bar{q}_t\|^2] + 36\tau\gamma_2^2\mathbb{E}[\|\bar{r}_{t+1}-\bar{r}_t\|^2] + S\,. \qquad (44)
\end{aligned}$$

where we denote

$$\begin{aligned}
S_x &= 12\tau\eta^4(2\beta_x^2L_f^2\rho^2 + 2\beta_y^2L_f^2\rho^2 + 8\beta_x^2L_g^2 + 6\beta_y^2L_g^2 + 6\beta_z^2L_g^2 + \beta_x^2\gamma_1^2)\,,\\
S_y &= 12\tau\eta^4(2\beta_x^2L_f^2\rho^2 + 2\beta_y^2L_f^2\rho^2 + 6\beta_x^2L_g^2 + 6\beta_y^2L_g^2 + \beta_y^2\gamma_1^2)\,,\\
S_z &= 12\tau\eta^4(2\beta_x^2L_g^2 + 6\beta_z^2L_g^2 + \beta_z^2\gamma_2^2)\,,\\
S_p &= 36\tau\gamma_1^2\alpha_p^2\eta^2 + 12\tau\gamma_1^2\beta_x^2\eta^4\,, \qquad (45)\\
S_q &= 36\tau\gamma_1^2\alpha_q^2\eta^2 + 12\tau\gamma_1^2\beta_y^2\eta^4\,,\\
S_r &= 36\tau\gamma_2^2\alpha_r^2\eta^2 + 12\tau\gamma_2^2\beta_z^2\eta^4\,,\\
S &= 12\tau(\beta_x^2+\beta_y^2)\eta^4\delta^2\rho^2 + 8\tau(\beta_x^2+\beta_y^2)\eta^4\sigma^2\rho^2 + 12\tau(4\beta_x^2+3\beta_y^2+\beta_z^2)\eta^4\delta^2
\end{aligned}$$

$$+ 8\tau(4\beta_x^2 + 3\beta_y^2 + \beta_z^2)\eta^4\sigma^2 \,.$$

By setting $\alpha_x \leq 10$, $\alpha_y < 10$, $\alpha_z < 10$, $\eta \leq \frac{1}{360\tau L_{\rho,\gamma}}$, we obtain

$$48\tau\alpha_x^2\eta^2 L_{\rho,\gamma}^2 \leq 4800\tau\eta^2 L_{\rho,\gamma}^2 \leq \frac{4800\tau}{360 \times 360\tau^2} \leq \frac{1}{25\tau} \,,$$
$$32\tau\alpha_y^2\eta^2 L_{\rho,\gamma}^2 \leq 3200\tau\eta^2 L_{\rho,\gamma}^2 \leq \frac{3200\tau}{360 \times 360\tau^2} \leq \frac{1}{25\tau} \,,$$
$$8\tau L_{\rho,\gamma}^2\alpha_z^2\eta^2 \leq 800\tau L_{\rho,\gamma}^2\eta^2 \leq \frac{800\tau}{360 \times 360\tau^2} \leq \frac{1}{25\tau} \,.$$

Moreover, we derive

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)} - \bar{x}_{t+1}\|^2] = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t\tau}^{(k)} - \alpha_x\eta\sum_{t'=s_t\tau}^{t}u_{t'}^{(k)} - \bar{x}_{t\tau} + \alpha_x\eta\sum_{t'=s_t\tau}^{t}\bar{u}_{t'}\|^2]$$
$$\leq \tau\alpha_x^2\eta^2\sum_{\hat{t}=s_t\tau}^{t}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{\hat{t}}^{(k)} - \bar{u}_{\hat{t}}\|^2] \,, \tag{46}$$

and

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_t^{(k)} - \bar{p}_t\|^2] = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_{t-1}^{(k)} + \alpha_p\eta(x_t^{(k)} - p_{t-1}^{(k)}) - \bar{p}_{t-1} - \alpha_p\eta(\bar{x}_t - \bar{p}_{t-1})\|^2]$$
$$= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|(1-\alpha_p\eta)(p_{t-1}^{(k)} - \bar{p}_{t-1}) + \alpha_p\eta(x_t^{(k)} - \bar{x}_t)\|^2]$$
$$\leq (1 + \frac{\alpha_p\eta}{1-\alpha_p\eta})(1-\alpha_p\eta)^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_{t-1}^{(k)} - \bar{p}_{t-1}\|^2]$$
$$+ (1 + \frac{1-\alpha_p\eta}{\alpha_p\eta})\alpha_p^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_t^{(k)} - \bar{x}_t\|^2]$$
$$= (1-\alpha_p\eta)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|p_{t-1}^{(k)} - \bar{p}_{t-1}\|^2] + \alpha_p\eta\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_t^{(k)} - \bar{x}_t\|^2]$$
$$\leq \sum_{\hat{t}=s_t\tau}^{t}(1-\alpha_p\eta)^{t-\hat{t}}\alpha_p\eta\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{\hat{t}}^{(k)} - \bar{x}_{\hat{t}}\|^2] \leq \alpha_p\eta\sum_{\hat{t}=s_t\tau}^{t}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{\hat{t}}^{(k)} - \bar{x}_{\hat{t}}\|^2] \,, \tag{47}$$

where s_t denotes the index of the latest communication round before iteration t.

Similarly, we have

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_{t+1}^{(k)} - \bar{y}_{t+1}\|^2] \leq \tau\alpha_y^2\eta^2\sum_{\hat{t}=s_t\tau}^{t}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{\hat{t}}^{(k)} - \bar{v}_{\hat{t}}\|^2] \,,$$
$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|q_t^{(k)} - \bar{q}_t\|^2] \leq \alpha_q\eta\sum_{\hat{t}=s_t\tau}^{t}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_{\hat{t}}^{(k)} - \bar{y}_{\hat{t}}\|^2] \,,$$
$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|z_{t+1}^{(k)} - \bar{z}_{t+1}\|^2] \leq \tau\alpha_z^2\eta^2\sum_{\hat{t}=s_t\tau}^{t}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|w_{\hat{t}}^{(k)} - \bar{w}_{\hat{t}}\|^2] \,,$$
$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|r_t^{(k)} - \bar{r}_t\|^2] \leq \alpha_r\eta\sum_{\hat{t}=s_t\tau}^{t}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|z_{\hat{t}}^{(k)} - \bar{z}_{\hat{t}}\|^2] \,, \tag{48}$$

where s_t denotes the index of the latest communication round before iteration t.

Thus, by setting $s_t = \lfloor \frac{t}{\tau} \rfloor$, applying Eq.(46), (47), and (48), we obtain:

$$
\begin{aligned}
&\rho^2 \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t+1}^{(k)} - \bar{u}_{1,t+1}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t+1}^{(k)} - \bar{u}_{2,t+1}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,t+1}^{(k)} - \bar{u}_{3,t+1}\|^2] \\
&+ \rho^2 \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{1,t+1}^{(k)} - \bar{v}_{1,t+1}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{2,t+1}^{(k)} - \bar{v}_{2,t+1}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|w_{t+1}^{(k)} - \bar{w}_{t+1}\|^2] \\
\le\ & 48\tau\alpha_x^2\eta^2 L_{\rho,\gamma}^2 \sum_{t'=s_t\tau}^{t-1}(1+\frac{26}{25\tau})^{t-1-t'}\mathbb{E}[\|\bar{u}_{t'}\|^2] + 32\tau\alpha_y^2\eta^2 L_{\rho,\gamma}^2 \sum_{t'=s_t\tau}^{t-1}(1+\frac{26}{25\tau})^{t-1-t'}\mathbb{E}[\|\bar{v}_{t'}\|^2] \\
&+ 8\tau L_{\rho,\gamma}^2\alpha_z^2\eta^2 \sum_{t'=s_t\tau}^{t-1}(1+\frac{26}{25\tau})^{t-1-t'}\mathbb{E}[\|\bar{w}_{t'}\|^2] \\
&+ (\tau\alpha_p\eta S_p + S_x)\sum_{t'=s_t\tau}^{t-1}(1+\frac{26}{25\tau})^{t-1-t'}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t'}^{(k)} - \bar{x}_{t'}\|^2] \\
&+ (\tau\alpha_q\eta S_q + S_y)\sum_{t'=s_t\tau}^{t-1}(1+\frac{26}{25\tau})^{t-1-t'}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_{t'}^{(k)} - \bar{y}_{t'}\|^2] \\
&+ (\tau\alpha_r\eta S_r + S_z)\sum_{t'=s_t\tau}^{t-1}(1+\frac{26}{25\tau})^{t-1-t'}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|z_{t'}^{(k)} - \bar{z}_{t'}\|^2] \\
&+ 72\tau^3\gamma_1^2\alpha_x^2\alpha_p^2\eta^4 \sum_{t'=s_t\tau}^{t-1}(1+\frac{26}{25\tau})^{t-1-t'}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{t'}^{(k)} - \bar{u}_{t'}\|^2] \\
&+ 72\tau^3\gamma_1^2\alpha_y^2\alpha_q^2\eta^4 \sum_{t'=s_t\tau}^{t-1}(1+\frac{26}{25\tau})^{t-1-t'}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{t'}^{(k)} - \bar{v}_{t'}\|^2] \\
&+ 72\tau^3\gamma_2^2\alpha_z^2\alpha_r^2\eta^4 \sum_{t'=s_t\tau}^{t-1}(1+\frac{26}{25\tau})^{t-1-t'}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|w_{t'}^{(k)} - \bar{w}_{t'}\|^2] \\
&+ 72\tau\gamma_1^2 \sum_{t'=s_t\tau}^{t-1}(1+\frac{26}{25\tau})^{t-1-t'}\mathbb{E}[\|\bar{p}_{t'+1} - \bar{p}_{t'}\|^2] + 72\tau\gamma_1^2 \sum_{t'=s_t\tau}^{t-1}(1+\frac{26}{25\tau})^{t-1-t'}\mathbb{E}[\|\bar{q}_{t'+1} - \bar{q}_{t'}\|^2] \\
&+ 72\tau\gamma_2^2 \sum_{t'=s_t\tau}^{t-1}(1+\frac{26}{25\tau})^{t-1-t'}\mathbb{E}[\|\bar{r}_{t'+1} - \bar{r}_{t'}\|^2] + S\sum_{t'=s_t\tau}^{t-1}(1+\frac{26}{25\tau})^{t-1-t'}\ .
\end{aligned}
$$

Moreover, since $(1+\frac{26}{25\tau})^{t-1-t'} \le (1+\frac{26}{25\tau})^{\tau} \le 3$, denoting

$$
\begin{aligned}
S_x' &= 3\tau^3\alpha_x^2\eta^2\Big(\tau\alpha_p\eta S_p + S_x + 72\tau\gamma_1^2\alpha_p^2\eta^2\Big)\ , \\
S_y' &= 3\tau^3\alpha_y^2\eta^2\Big(\tau\alpha_q\eta S_q + S_y + 72\tau\gamma_1^2\alpha_q^2\eta^2\Big)\ , \\
S_z' &= 3\tau^3\alpha_z^2\eta^2\Big(\tau\alpha_r\eta S_r + S_y + 72\tau\gamma_2^2\alpha_r^2\eta^2\Big)\ ,
\end{aligned}
$$

applying Lemma E.2 and summing up t' from $s_t\tau$ to $(s_t+1)\tau - 1$, we obtain

$$
\begin{aligned}
&\sum_{t'=s_t\tau}^{(s_t+1)\tau-1}\Bigg(\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t'}^{(k)} - \bar{u}_{1,t'}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t'}^{(k)} - \bar{u}_{2,t'}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,t'}^{(k)} - \bar{u}_{3,t'}\|^2] \\
&\qquad + \rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{1,t'}^{(k)} - \bar{v}_{1,t'}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{2,t'}^{(k)} - \bar{v}_{2,t'}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|w_{t'}^{(k)} - \bar{w}_{t'}\|^2]\Bigg) \\
\le\ & 144\tau^2\alpha_x^2\eta^2 L_{\rho,\gamma}^2 \sum_{t'=s_t\tau}^{(s_t+1)\tau-1}\mathbb{E}[\|\bar{u}_{t'}\|^2] + 96\tau^2\alpha_y^2\eta^2 L_{\rho,\gamma}^2 \sum_{t'=s_t\tau}^{(s_t+1)\tau-1}\mathbb{E}[\|\bar{v}_{t'}\|^2]
\end{aligned}
$$

$$+ 24\tau^2\alpha_z^2\eta^2 L_{\rho,\gamma}^2 \sum_{t'=s_t\tau}^{t-1} \mathbb{E}[\|\bar{w}_{t'}\|^2] + 3\tau S \sum_{t'=s_t\tau}^{(s_t+1)\tau-1} + 216\tau^2\gamma_1^2 \sum_{t'=s_t\tau}^{(s_t+1)\tau-1} \mathbb{E}[\|\bar{p}_{t'+1} - \bar{p}_{t'}\|^2]$$
$$+ 216\tau^2\gamma_1^2 \sum_{t'=s_t\tau}^{(s_t+1)\tau-1} \mathbb{E}[\|\bar{q}_{t'+1} - \bar{q}_{t'}\|^2] + 216\tau^2\gamma_2^2 \sum_{t'=s_t\tau}^{(s_t+1)\tau-1} \mathbb{E}[\|\bar{r}_{t'+1} - \bar{r}_{t'}\|^2]$$
$$+ 3S_x'\rho^2 \sum_{t'=s_t\tau}^{(s_t+1)\tau-1} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\|u_{1,t'}^{(k)} - \bar{u}_{1,t'}\|^2] + 3S_x' \sum_{t'=s_t\tau}^{(s_t+1)\tau-1} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\|u_{2,t'}^{(k)} - \bar{u}_{2,t'}\|^2]$$
$$+ 3S_x' \sum_{t'=s_t\tau}^{(s_t+1)\tau-1} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\|u_{3,t'}^{(k)} - \bar{u}_{3,t'}\|^2] + 2S_y'\rho^2 \sum_{t'=s_t\tau}^{(s_t+1)\tau-1} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\|v_{1,t'}^{(k)} - \bar{v}_{1,t'}\|^2]$$
$$+ 2S_y' \sum_{t'=s_t\tau}^{(s_t+1)\tau-1} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\|v_{2,t'}^{(k)} - \bar{v}_{2,t'}\|^2] + S_z' \sum_{t'=s_t\tau}^{(s_t+1)\tau-1} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\|w_{t'}^{(k)} - \bar{w}_{t'}\|^2] .$$

Next, we would like to enforce $S_x' \le \frac{1}{6}$, and make sure $1 - 3S_x' \ge \frac{1}{2}$. In particular, we decompose S_x' as follows:

$$\begin{aligned} S_x' &= 3\tau^3\alpha_x^2\eta^2\Big(\tau\alpha_p\eta S_p + S_x + 72\tau\gamma_1^2\alpha_p^2\eta^2\Big) \\ &= 3\tau^4\alpha_x^2\eta^2\Big(\underbrace{\tau\alpha_p\eta 36\gamma_1^2\alpha_p^2\eta^2}_{S_{x,1}} + \underbrace{\tau\alpha_p\eta 12\gamma_1^2\beta_x^2\eta^4}_{S_{x,2}} \\ &\quad + \underbrace{24\eta^4(\beta_x^2L_f^2\rho^2 + \beta_y^2L_f^2\rho^2 + 4\beta_x^2L_g^2 + 3\beta_y^2L_g^2 + 3\beta_z^2L_g^2)}_{S_{x,3}} + \underbrace{12\gamma_1^2\eta^4\beta_x^2}_{S_{x,4}} + \underbrace{72\gamma_1^2\alpha_p^2\eta^2}_{S_{x,5}}\Big) . \end{aligned} \tag{49}$$

By imposing $\alpha_{pq} \le 10\gamma$, and recalling the definition of α_{pq} in Eq. (32), we obtain $\alpha_x \le \frac{96\times 1280\gamma_1^2L_\rho^2}{(\gamma_1-L_\rho)^3\mu} \le \frac{96\times 1280(\gamma_1+L_\rho)^4}{(\gamma-L_\rho)^3\mu}$. Since $\alpha_x \le 10$ holds is enforced by our parameter choice, the above inequality is automatically satisfied. Therefore, with $\eta \le \frac{1}{360\tau L_{\rho,\gamma}}$, we have

$$3\tau^4\alpha_x^2\eta^2 S_{x,1} = 3\tau^4\alpha_x^2\eta^2\tau\gamma_1^2 \times 36\alpha_p^3\eta^3 \le \frac{108\times 100\tau^5\gamma_1^2\times 1000\gamma_1^3}{360^5\tau^5L_{\rho,\gamma}^5} \le \frac{1}{30} .$$

Similarly, by setting $\beta_x \le \frac{L_{\rho,\gamma}^2}{K}$, we derive

$$3\tau^4\alpha_x^2\eta^2 S_{x,2} = 36\tau^4\alpha_x^2\eta^2\tau\gamma^2\beta_x^2\eta^4\alpha_p\eta \le \frac{36\times 100\tau^5\gamma_1^2\times 10\gamma L_{\rho,\gamma}^4}{360^7\tau^7L_{\rho,\gamma}^7\times K^2} \le \frac{1}{30} .$$

In addition, by setting $\beta_y \le \frac{L_{\rho,\gamma}}{K}$, $\beta_z \le \frac{L_{\rho,\gamma}}{K}$, we obtain

$$\begin{aligned} &3\tau^4\alpha_x^2\eta^2 S_{x,3} = 72\tau^4\alpha_x^2\eta^6(\beta_x^2L_f^2\rho^2 + \beta_y^2L_f^2\rho^2 + 4\beta_x^2L_g^2 + 3\beta_y^2L_g^2 + 3\beta_z^2L_g^2) \\ &\le \frac{72\times 200\tau^4(L_f^2\rho^2 + 5L_g^2)L_{\rho,\gamma}^4}{360^6\tau^6L_{\rho,\gamma}^6\times K^2} \le \frac{1}{30} . \end{aligned}$$

For the remaining two terms, we also derive

$$36\tau^4\alpha_x^2\eta^2\gamma_1^2\eta^4\beta_x^2 \le \frac{3600\tau^4\gamma_1^2L_{\rho,\gamma}^4}{360^6\tau^6L_{\rho,\gamma}^6\times K^2} \le \frac{1}{30} ,$$
$$216\tau^4\alpha_x^2\eta^2 6\gamma_1^2\alpha_p^2\eta^2 \le \frac{216\times 100\tau^4\gamma_1^2\times 100\gamma_1^2}{360^4\tau^4L_{\rho,\gamma}^4} \le \frac{1}{30} .$$

In a similar manner, we can also make sure $1 - 3S_y' \ge \frac{1}{2}$, $1 - 3S_z' \ge \frac{1}{2}$.

Therefore, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t}^{(k)} - \bar{u}_{1,t}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t}^{(k)} - \bar{u}_{2,t}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,t}^{(k)} - \bar{u}_{3,t}\|^2]\right.$$

$$+ \rho^2 \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{1,t}^{(k)} - \bar{v}_{1,t}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{2,t}^{(k)} - \bar{v}_{2,t}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|w_t^{(k)} - \bar{w}_t\|^2]\Big)$$

$$\leq 288\tau^2\alpha_x^2\eta^2 L_{\rho,\gamma}^2 \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{u}_t\|^2] + 192\tau^2\alpha_y^2\eta^2 L_{\rho,\gamma}^2 \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{v}_t\|^2] + 48\tau^2\alpha_z^2\eta^2 L_{\rho,\gamma}^2 \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{w}_t\|^2]$$

$$+ 432\tau^2\gamma_1^2 \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{p}_{t+1} - \bar{p}_t\|^2] + 432\tau^2\gamma_1^2 \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{q}_{t+1} - \bar{q}_t\|^2]$$

$$+ 432\tau^2\gamma_2^2 \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{r}_{t+1} - \bar{r}_t\|^2] + 3\tau S\,.$$

The proof is completed. □

F Proof of Theorem 4.8

We propose a novel potential function as follows:

$$\mathcal{P}_t = \Psi_t + c_1\rho^2\mathcal{G}_{u,1,t} + c_2\mathcal{G}_{u,2,t} + c_3\mathcal{G}_{u,3,t} + c_4\rho^2\mathcal{G}_{v,1,t} + c_5\mathcal{G}_{v,2,t} + c_6\mathcal{G}_{w,t}\,. \tag{50}$$

In particular, we have

$$\begin{aligned}
&\mathbb{E}[\|\nabla_x F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{u}_t\|^2] \\
&= \mathbb{E}[\|\rho\nabla_1 f(\bar{x}_t, \bar{y}_t) + \nabla_1 g_{\gamma_1}(\bar{x}_t, \bar{y}_t; \bar{p}_t, \bar{q}_t) - \nabla_1 g_{\gamma_2}(\bar{x}_t, \bar{z}_t; \bar{r}_t) - \bar{u}_t\|^2] \\
&\leq 2\mathbb{E}[\|\rho\nabla_1 f(\bar{x}_t, \bar{y}_t) + \nabla_1 g_{\gamma_1}(\bar{x}_t, \bar{y}_t; \bar{p}_t, \bar{q}_t) - \nabla_1 g_{\gamma_2}(\bar{x}_t, \bar{z}_t; \bar{r}_t) \\
&\quad \underbrace{-\frac{1}{K}\sum_{k=1}^{K}\Big(\rho\nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}) + \nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)}) - \nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, z_t^{(k)}; r_t^{(k)})\Big)\|^2]}_{T_1} \\
&\quad + \underbrace{2\,\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K}\Big(\rho\nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)}) + \nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)}) - \nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, z_t^{(k)}; r_t^{(k)})\Big) - \bar{u}_t\|^2]}_{T_2}\,.
\end{aligned}$$

The second term can be simply bounded as

$$T_2 \leq 3\rho^2\mathcal{G}_{u,1,t} + 3\mathcal{G}_{u,2,t} + 3\mathcal{G}_{u,3,t}\,.$$

For the first term, from the following inequality

$$\begin{aligned}
&\mathbb{E}[\|\nabla_1 g_{\gamma_1}(\bar{x}_t, \bar{y}_t; \bar{p}_t, \bar{q}_t) - \frac{1}{K}\sum_{k=1}^{K}\nabla_1 g_{\gamma_1}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)})\|^2] \\
&= \mathbb{E}[\|\nabla_1 g(\bar{x}_t, \bar{y}_t) + \gamma(\bar{x}_t - \bar{p}_t) - \frac{1}{K}\sum_{k=1}^{K}\nabla_1 g^{(k)}(x_t^{(k)}, y_t^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\gamma(x_t^{(k)} - p_t^{(k)})\|^2] \\
&\leq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 g^{(k)}(\bar{x}_t, \bar{y}_t) - \nabla_1 g(x_t^{(k)}, y_t^{(k)})\|^2] \\
&\leq L_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2] + L_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|^2]\,,
\end{aligned}$$

we obtain

$$\begin{aligned}
T_1 &\leq 3\rho^2\mathbb{E}[\|\nabla_1 f(\bar{x}_t, \bar{y}_t) - \frac{1}{K}\sum_{k=1}^{K}\nabla_1 f^{(k)}(x_t^{(k)}, y_t^{(k)})\|^2] \\
&\quad + 3\mathbb{E}[\|\nabla_1 g_{\gamma_1}(\bar{x}_t, \bar{y}_t; \bar{p}_t, \bar{q}_t) - \frac{1}{K}\sum_{k=1}^{K}\nabla_1 g_{\gamma_1}^{(k)}(x_t^{(k)}, y_t^{(k)}; p_t^{(k)}, q_t^{(k)})\|^2]
\end{aligned}$$

$$+ 3\mathbb{E}[\|\nabla_1 g_{\gamma_2}(\bar{x}_t, \bar{z}_t; \bar{r}_t) - \frac{1}{K}\sum_{k=1}^{K} \nabla_1 g_{\gamma_2}^{(k)}(x_t^{(k)}, z_t^{(k)}; r_t^{(k)})\|^2]$$
$$\leq 3(\rho^2 L_f^2 + 2L_g^2)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2] + 3(\rho^2 L_f^2 + L_g^2)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|^2]$$
$$+ 3L_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{z}_t - z_t^{(k)}\|^2] .$$

Thus, combing T_1 and T_2, we derive

$$\mathbb{E}[\|\nabla_x F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{u}_t\|^2] \leq 6(\rho^2 L_f^2 + 2L_g^2)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2]$$
$$+ 6(\rho^2 L_f^2 + L_g^2)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|^2] + 6L_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{z}_t - z_t^{(k)}\|^2]$$
$$+ 6\rho^2\mathcal{G}_{u,1,t} + 6\mathcal{G}_{u,2,t} + 6\mathcal{G}_{u,3,t} .$$

Similarly, we obtain

$$\mathbb{E}[\|\nabla_y F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{v}_t\|^2] = \mathbb{E}[\|\nabla_2 f(\bar{x}_t, \bar{y}_t) + \nabla_2 g_{\gamma_1}(\bar{x}_t, \bar{y}_t; \bar{p}_t, \bar{q}_t) - \bar{v}_t\|^2]$$
$$\leq 4(\rho^2 L_f^2 + L_g^2)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2] + 4(\rho^2 L_f^2 + L_g^2)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|^2]$$
$$+ 4\rho^2\mathcal{G}_{v,1,t} + 4\mathcal{G}_{v,2,t} ,$$

and

$$\mathbb{E}[\|\nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t) - \bar{w}_t\|^2] = \mathbb{E}[\|\nabla_2 g_{\gamma_2}(\bar{x}_t, \bar{z}_t; \bar{r}_t) - \bar{w}_t\|^2]$$
$$\leq 2L_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2] + 2L_g^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{z}_t - z_t^{(k)}\|^2] + 2\mathcal{G}_{w,t} .$$

Then, from Lemma D.8, E.3, E.4, by setting

$$c_1 = c_2 = c_3 = \frac{6\alpha_x}{\beta_x\eta} , \quad c_4 = c_5 = \frac{2\alpha_y}{\beta_y\eta} , \quad c_6 = \frac{3\alpha_z}{\beta_z\eta} , \tag{51}$$

we can eliminate the gradient error terms, with denotation

$$\begin{aligned}
\mathcal{M}_1 &= \frac{1}{K}(2L_f^2\rho^2 c_1 + 2L_g^2 c_2 + 2L_g^2 c_3 + 2L_f^2\rho^2 c_4 + 2L_g^2 c_5 + 2L_g^2 c_6) ,\\
\mathcal{M}_2 &= \frac{1}{K}(2L_f^2\rho^2 c_1 + 2L_g^2 c_2 + 2L_f^2\rho^2 c_4 + 2L_g^2 c_5) ,\\
\mathcal{M}_3 &= \frac{1}{K}(2L_g^2 c_3 + 2L_g^2 c_6) ,\\
\mathcal{M}_4 &= 6\alpha_x\eta(\rho^2 L_f^2 + 2L_g^2) + 2\alpha_y\eta(\rho^2 L_f^2 + L_g^2) + 3\alpha_z\eta L_g^2 ,\\
\mathcal{M}_5 &= 6\alpha_x\eta(\rho^2 L_f^2 + L_g^2) + 2\alpha_y\eta(\rho^2 L_f^2 + L_g^2) ,\\
\mathcal{M}_6 &= 6\alpha_x\eta L_g^2 + 3\alpha_z\eta L_g^2 ,\\
\mathcal{M}_0 &= \frac{2\sigma^2}{K}\Big(\beta_x^2\eta^4\rho^2 c_1 + \beta_x^2\eta^4 c_2 + \beta_x^2\eta^4 c_3 + \beta_y^2\eta^4\rho^2 c_4 + \beta_y^2\eta^4 c_5 + \beta_z^2\eta^4 c_6\Big) ,
\end{aligned} \tag{52}$$

we obtain

$$\mathcal{P}_{t+1} - \mathcal{P}_t \leq -\frac{\alpha_x\eta}{8}\mathbb{E}[\|\nabla_x F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2] - \frac{\alpha_y\eta}{2}\mathbb{E}[\|\nabla_y F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2]$$
$$- \frac{\alpha_z\eta}{2}\mathbb{E}[\|\nabla_z F(\bar{x}_t, \bar{y}_t, \bar{z}_t; \bar{p}_t, \bar{q}_t, \bar{r}_t)\|^2] + \Big(16\gamma_1\alpha_{pq}\eta\big(5\omega_2^2\omega_3^2 + C_r\big) - \frac{\gamma_2}{2\alpha_r\eta}\Big)\mathbb{E}[\|\bar{r}_{t+1} - \bar{r}_t\|^2]$$

$$
\begin{aligned}
&+\Big(1+\omega_1^2+8\alpha_{pq}\eta\omega_5^2\big(5\omega_2^2+\frac{2\gamma_2^2}{\mu(\gamma_1-L_\rho)}\big)-\frac{1}{4\alpha_{pq}\eta}\Big)\gamma_1\big(\mathbb{E}[\|\bar{p}_{t+1}-\bar{p}_t\|^2]+\mathbb{E}[\|\bar{q}_{t+1}-\bar{q}_t\|^2]\big)\\
&+\Big(4\alpha_x^2\alpha_z\eta^3L_g^2-\frac{\alpha_x\eta}{4}\Big)\mathbb{E}[\|\bar{u}_t\|^2]+\Big(\frac{3\alpha_x\alpha_y^2\eta^3L_\rho^2}{2}-\frac{\alpha_y\eta}{4}\Big)\mathbb{E}[\|\bar{v}_t\|^2]\\
&+\Big(\frac{\alpha_z^2\eta^2}{2}(L_{F,2}+2L_d+2C_\rho)-\frac{\alpha_z\eta}{16}\Big)\mathbb{E}[\|\bar{w}_t\|^2]\\
&+\mathcal{M}_1\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|x_{t+1}^{(k)}-x_t^{(k)}\|^2]+\mathcal{M}_2\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|y_{t+1}^{(k)}-y_t^{(k)}\|^2]+\mathcal{M}_3\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|z_{t+1}^{(k)}-z_t^{(k)}\|^2]\\
&+\mathcal{M}_4\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{x}_t-x_t^{(k)}\|^2]+\mathcal{M}_5\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{y}_t-y_t^{(k)}\|^2]+\mathcal{M}_6\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{z}_t-z_t^{(k)}\|^2]+\mathcal{M}_0\,.
\end{aligned}
$$

Then, by applying Lemma E.1, we obtain

$$
\begin{aligned}
&\mathcal{P}_{t+1}-\mathcal{P}_t\le-\frac{\alpha_x\eta}{8}\mathbb{E}[\|\nabla_xF(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]-\frac{\alpha_y\eta}{2}\mathbb{E}[\|\nabla_yF(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]\\
&-\frac{\alpha_z\eta}{2}\mathbb{E}[\|\nabla_zF(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]+\Big(16\gamma_1\alpha_{pq}\eta\big(5\omega_2^2\omega_3^2+C_r\big)-\frac{\gamma_2}{2\alpha_r\eta}\Big)\mathbb{E}[\|\bar{r}_{t+1}-\bar{r}_t\|^2]\\
&+\Big(1+\omega_1^2+8\alpha_{pq}\eta\omega_5^2\big(5\omega_2^2+\frac{2\gamma_2^2}{\mu(\gamma_1-L_\rho)}\big)-\frac{1}{4\alpha_{pq}\eta}\Big)\gamma_1\big(\mathbb{E}[\|\bar{p}_{t+1}-\bar{p}_t\|^2]+\mathbb{E}[\|\bar{q}_{t+1}-\bar{q}_t\|^2]\big)\\
&+\Big(4\alpha_x^2\alpha_z\eta^3L_g^2+6\alpha_x^2\eta^2\mathcal{M}_1-\frac{\alpha_x\eta}{4}\Big)\mathbb{E}[\|\bar{u}_t\|^2]+\Big(\frac{3\alpha_x\alpha_y^2\eta^3L_\rho^2}{2}+4\alpha_y^2\eta^2\mathcal{M}_2-\frac{\alpha_y\eta}{4}\Big)\mathbb{E}[\|\bar{v}_t\|^2]\\
&+\Big(\frac{\alpha_z^2\eta^2}{2}(L_{F,2}+2L_d+2C_\rho)+2\alpha_z^2\eta^2\mathcal{M}_3-\frac{\alpha_z\eta}{16}\Big)\mathbb{E}[\|\bar{w}_t\|^2]\\
&+6\alpha_x^2\eta^2\mathcal{M}_1\Big(\rho^2\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|u_{1,t}^{(k)}-\bar{u}_{1,t}\|^2]+\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|u_{2,t}^{(k)}-\bar{u}_{2,t}\|^2]+\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|u_{3,t}^{(k)}-\bar{u}_{3,t}\|^2]\Big)\\
&+4\alpha_y^2\eta^2\mathcal{M}_2\Big(\rho^2\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|v_{1,t}^{(k)}-\bar{v}_{1,t}\|^2]+\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|v_{2,t}^{(k)}-\bar{v}_{2,t}\|^2]\Big)\\
&+2\alpha_z^2\eta^2\mathcal{M}_3\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|w_t^{(k)}-\bar{w}_t\|^2]+\mathcal{M}_4\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{x}_t-x_t^{(k)}\|^2]\\
&+\mathcal{M}_5\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{y}_t-y_t^{(k)}\|^2]+\mathcal{M}_6\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{z}_t-z_t^{(k)}\|^2]+\mathcal{M}_0\,.
\end{aligned}
$$

By summing over t from 0 to $T-1$, applying Lemma E.2 we obtain

$$
\begin{aligned}
&\frac{\mathcal{P}_T-\mathcal{P}_0}{T}\le-\frac{\alpha_x\eta}{8}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_xF(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]\\
&-\frac{\alpha_y\eta}{2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_yF(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]-\frac{\alpha_z\eta}{2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_zF(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]\\
&+\Big(16\gamma_1\alpha_{pq}\eta\big(5\omega_2^2\omega_3^2+C_r\big)-\frac{\gamma_2}{2\alpha_r\eta}\Big)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{r}_{t+1}-\bar{r}_t\|^2]\\
&+\Big(1+\omega_1^2+8\alpha_{pq}\eta\omega_5^2\big(5\omega_2^2+\frac{2\gamma_2^2}{\mu(\gamma_1-L_\rho)}\big)-\frac{1}{4\alpha_{pq}\eta}\Big)\gamma_1\frac{1}{T}\sum_{t=0}^{T-1}\big(\mathbb{E}[\|\bar{p}_{t+1}-\bar{p}_t\|^2]+\mathbb{E}[\|\bar{q}_{t+1}-\bar{q}_t\|^2]\big)\\
&+\Big(4\alpha_x^2\alpha_z\eta^3L_g^2+6\alpha_x^2\eta^2\mathcal{M}_1-\frac{\alpha_x\eta}{4}\Big)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{u}_t\|^2]
\end{aligned}
$$

$$
\begin{aligned}
&+\Big(\frac{3\alpha_x\alpha_y^2\eta^3L_\rho^2}{2}+4\alpha_y^2\eta^2\mathcal{M}_2-\frac{\alpha_y\eta}{4}\Big)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{v}_t\|^2]\\
&+\Big(\frac{\alpha_z^2\eta^2}{2}(L_{F,2}+2L_d+2C_\rho)+2\alpha_z^2\eta^2\mathcal{M}_3-\frac{\alpha_z\eta}{16}\Big)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{w}_t\|^2]\\
&+\alpha_x^2\eta^2(6\mathcal{M}_1+3\tau^2\mathcal{M}_4)\frac{1}{T}\sum_{t=0}^{T-1}\Big(\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{1,t}^{(k)}-\bar{u}_{1,t}\|^2]+\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{2,t}^{(k)}-\bar{u}_{2,t}\|^2]\\
&\qquad+\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{3,t}^{(k)}-\bar{u}_{3,t}\|^2]\Big)\\
&+\alpha_y^2\eta^2(4\mathcal{M}_2+2\tau^2\mathcal{M}_5)\frac{1}{T}\sum_{t=0}^{T-1}\Big(\rho^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{1,t}^{(k)}-\bar{v}_{1,t}\|^2]+\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|v_{2,t}^{(k)}-\bar{v}_{2,t}\|^2]\Big)\\
&+\alpha_z^2\eta^2(2\mathcal{M}_3+\tau^2\mathcal{M}_6)\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|w_t^{(k)}-\bar{w}_t\|^2]+\mathcal{M}_0\,.
\end{aligned}
$$

Moreover, by defining

$$
\begin{aligned}
\mathcal{M}=&\frac{6(\alpha_x^2+\alpha_y^2+\alpha_z^2)\eta^2}{K}(2L_f^2\rho^2c_1+2L_g^2c_2+2L_g^2c_3+2L_f^2\rho^2c_4+2L_g^2c_5+2L_g^2c_6)\\
&+3\tau^2(\alpha_x^2+\alpha_y^2+\alpha_z^2)\eta^2\Big(6\alpha_x\eta(\rho^2L_f^2+2L_g^2)+2\alpha_y\eta(\rho^2L_f^2+L_g^2)+6\alpha_z\eta L_g^2\Big)\,,
\end{aligned}\tag{53}
$$

and applying Lemma E.7, we obtain

$$
\begin{aligned}
\frac{\mathcal{P}_T-\mathcal{P}_0}{T}\le&-\frac{\alpha_x\eta}{8}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_xF(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]\\
&-\frac{\alpha_y\eta}{2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_yF(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]-\frac{\alpha_z\eta}{2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_zF(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]\\
&+\Big(16\gamma_1\alpha_{pq}\eta\big(5\omega_2^2\omega_3^2+C_r\big)+432\tau^2\gamma_2^2\mathcal{M}-\frac{\gamma_2}{2\alpha_r\eta}\Big)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{r}_{t+1}-\bar{r}_t\|^2]\\
&+\Big(1+\omega_1^2+8\alpha_{pq}\eta\omega_5^2\big(5\omega_2^2+\frac{2\gamma_2^2}{\mu(\gamma_1-L_\rho)}\big)+432\tau^2\gamma_1\mathcal{M}-\frac{1}{4\alpha_{pq}\eta}\Big)\gamma_1\frac{1}{T}\sum_{t=0}^{T-1}\big(\mathbb{E}[\|\bar{p}_{t+1}-\bar{p}_t\|^2]\\
&\qquad+\mathbb{E}[\|\bar{q}_{t+1}-\bar{q}_t\|^2]\big)\\
&+\Big(4\alpha_x^2\alpha_z\eta^3L_g^2+6\alpha_x^2\eta^2\mathcal{M}_1+288\tau^2\alpha_x^2\eta^2L_{\rho,\gamma}^2\mathcal{M}-\frac{\alpha_x\eta}{4}\Big)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{u}_t\|^2]\\
&+\Big(\frac{3\alpha_x\alpha_y^2\eta^3L_\rho^2}{2}+4\alpha_y^2\eta^2\mathcal{M}_2+192\tau^2\alpha_y^2\eta^2L_{\rho,\gamma}^2\mathcal{M}-\frac{\alpha_y\eta}{4}\Big)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{v}_t\|^2]+\mathcal{M}_0\\
&+\Big(\frac{\alpha_z^2\eta^2}{2}(L_{F,2}+2L_d+2C_\rho)+2\alpha_z^2\eta^2\mathcal{M}_3+48\tau^2\alpha_z^2\eta^2L_g^2\mathcal{M}-\frac{\alpha_z\eta}{16}\Big)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{w}_t\|^2]+3\tau S\mathcal{M}\,.
\end{aligned}
$$

To eliminate $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{p}_{t+1}-\bar{p}_t\|^2]$ and $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{q}_{t+1}-\bar{q}_t\|^2]$, we enforce

$$
1+\omega_1^2+8\alpha_{pq}\eta\omega_5^2\big(5\omega_2^2+\frac{2\gamma_2^2}{\mu(\gamma_1-L_\rho)}\big)+432\tau^2\gamma_1^2\mathcal{M}-\frac{1}{4\alpha_{pq}\eta}\le-\frac{1}{8\alpha_{pq}\eta}\,.
$$

Specifically, we enforce

$$
1+\omega_1^2\le\frac{1}{24\alpha_{pq}\eta}\,,\quad 8\alpha_{pq}\eta\omega_5^2\big(5\omega_2^2+\frac{2\gamma_2^2}{\mu(\gamma_1-L_\rho)}\big)\le\frac{1}{24\alpha_{pq}\eta}\,,\quad 432\tau^2\gamma\mathcal{M}\le\frac{1}{24\alpha_{pq}\eta}\,.
$$

By solving the first two inequalities, we derive

$$\alpha_x \leq \min\{\frac{1}{24c_{\alpha_{pq}}(1+\omega_1^2)}\,,\frac{1}{8\sqrt{3}c_{\alpha_{pq}}\omega_5\sqrt{5\omega_2^2+\frac{2\gamma_2^2}{\mu(\gamma_1-L_\rho)}}}\}\,.$$

To solve the last inequality, from Eq. (43), we derive:

$$(\alpha_x^2+\alpha_y^2+\alpha_z^2)\eta^2 \leq \frac{300}{360^2\tau^2L_{\rho,\gamma}^2}\,,$$

and from the definition of $\mathcal{M}$ in Eq. (53), we enforce:

$$\begin{aligned}
&432\tau^2\gamma_1\frac{6\times 300(2L_f^2\rho^2c_1+2L_g^2c_2+2L_g^2c_3)}{360^2\tau^2L_{\rho,\gamma}^2K} \leq \frac{1}{24\times 6\alpha_{pq}\eta}\,,\\
&432\tau^2\gamma_1\frac{6\times 300(2L_f^2\rho^2c_4+2L_g^2c_5)}{360^2\tau^2L_{\rho,\gamma}^2K} \leq \frac{1}{24\times 6\alpha_{pq}\eta}\,,\\
&432\tau^2\gamma_1\frac{6\times 300\times 2L_g^2c_6}{360^2\tau^2L_{\rho,\gamma}^2K} \leq \frac{1}{24\times 6\alpha_{pq}\eta}\,,\\
&432\tau^2\gamma_1\frac{300\times 3\tau^2\times 6\alpha_x\eta(\rho^2L_f^2+2L_g^2)}{360^2\tau^2L_{\rho,\gamma}^2} \leq \frac{1}{24\times 6\alpha_{pq}\eta}\,,\\
&432\tau^2\gamma_1\frac{300\times 3\tau^2\times 2\alpha_y\eta(\rho^2L_f^2+L_g^2)}{360^2\tau^2L_{\rho,\gamma}^2} \leq \frac{1}{24\times 6\alpha_{pq}\eta}\,,\\
&432\tau^2\gamma_1\frac{300\times 3\tau^2\times 3\alpha_z\eta L_g^2}{360^2\tau^2L_{\rho,\gamma}^2} \leq \frac{1}{24\times 6\alpha_{pq}\eta}\,.
\end{aligned}\tag{54}$$

To solve these inequalities, we apply the value of $\{c_i\}_{i=1}^6$ in Eq. (51), with $\alpha_y=\alpha_x$, $\alpha_z=\alpha_x c_{\alpha_z}$, $\alpha_{pq}=\alpha_x c_{\alpha_{pq}}$, and obtain

$$\alpha_x \leq \min\{\frac{\sqrt{\beta_xK}}{72\sqrt{2\gamma_1c_{\alpha_{pq}}}}\,,\frac{\sqrt{\beta_yK}}{24\sqrt{6\gamma_1c_{\alpha_{pq}}}}\,,\frac{\sqrt{\beta_zK}}{72\sqrt{\gamma_1c_{\alpha_z}c_{\alpha_{pq}}}}\,,\frac{1}{36\tau\eta\sqrt{2\gamma_1c_{\alpha_{pq}}}}\,,\frac{1}{36\tau\eta\sqrt{\gamma_1c_{\alpha_z}c_{\alpha_{pq}}}}\}\,.$$

Similarly, to eliminate $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{r}_{t+1}-\bar{r}_t\|^2]$, we enforce

$$16\gamma_1\alpha_{pq}\eta\big(5\omega_2^2\omega_3^2+C_r\big)+432\tau^2\gamma_2^2\mathcal{M}-\frac{\gamma_2}{2\alpha_r\eta} \leq -\frac{\gamma_2}{4\alpha_r\eta}\,.$$

Specifically, from $C_r=\frac{4\gamma_2^2}{\mu(\gamma_1-L_\rho)}\big(\omega_3^2+\frac{(1-\alpha_r\eta)^2}{\alpha_r^2\eta^2}\big)$, we enforce

$$\begin{aligned}
&16\gamma_1\alpha_{pq}\eta\times 5\omega_2^2\omega_3^2 \leq \frac{\gamma_2}{16\alpha_r\eta}\,,\quad \frac{16\gamma_1\alpha_{pq}\eta\times 4\gamma_2^2}{\mu(\gamma_1-L_\rho)}\omega_3^2 \leq \frac{\gamma_2}{16\alpha_r\eta}\,,\\
&\frac{16\gamma_1\alpha_{pq}\eta\times 4\gamma_2^2}{\mu(\gamma_1-L_\rho)}\frac{(1-\alpha_r\eta)^2}{\alpha_r^2\eta^2} \leq \frac{\gamma_2}{16\alpha_r\eta}\,,\quad 432\tau^2\gamma_2^2\mathcal{M} \leq \frac{\gamma_2}{16\alpha_r\eta}\,.
\end{aligned}$$

To solve the first and second inequality, we obtain

$$\alpha_x \leq \Big\{\frac{(\gamma_1-L_\rho)(\gamma_2-L_\rho)}{16\sqrt{5\gamma_1\gamma_2c_{\alpha_{pq}}c_{\alpha_r}}L_g}\,,\frac{\sqrt{\mu(\gamma_1-L_\rho)}(\gamma_2-L_\rho)}{32\gamma_2\sqrt{\gamma_1\gamma_2c_{\alpha_{pq}}c_{\alpha_r}}}\Big\}\,.$$

To solve the third inequality, from the third inequality in Eq. (38), we derive

$$\frac{16\gamma_1\alpha_{pq}\eta\times 4\gamma_2^2}{\mu(\gamma_1-L_\rho)}\frac{(1-\alpha_r\eta)^2}{\alpha_r^2\eta^2} \leq \frac{\alpha_x\eta(\gamma_1-L_\rho)^2(\gamma_2-L_\rho)^2}{72\times 24L_g^2} \leq \frac{\gamma_2}{16\alpha_r\eta}\,,$$

thus, we obtain $\alpha_x \leq \frac{6\sqrt{3\gamma_2}L_g}{\sqrt{c_{\alpha_r}}(\gamma_1-L_\rho)(\gamma_2-L_\rho)}$. To solve the last inequality, following Eq. (54), we obtain

$$\alpha_x \leq \min\Big\{\frac{\sqrt{\beta_xK}}{48\sqrt{3\gamma_2c_{\alpha_r}}}\,,\frac{\sqrt{\beta_yK}}{48\sqrt{\gamma_2c_{\alpha_r}}}\,,\frac{\sqrt{\beta_zK}}{24\sqrt{6\gamma_2c_{\alpha_z}c_{\alpha_r}}}\,,\frac{1}{24\tau\eta\sqrt{3\gamma_2c_{\alpha_r}}}\,,\frac{1}{12\tau\eta\sqrt{6\gamma_2c_{\alpha_z}c_{\alpha_r}}}\Big\}\,.$$

Then, to eliminate $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{u}_t\|^2]$, we enforce

$$4\alpha_x^2\alpha_z\eta^3L_g^2+6\alpha_x^2\eta^2\mathcal{M}_1+288\tau^2\alpha_x^2\eta^2L_{\rho,\gamma}^2\mathcal{M}-\frac{\alpha_x\eta}{4}\leq-\frac{\alpha_x\eta}{8}\,.$$

In particular, we enforce

$$4\alpha_x^2\alpha_z\eta^3L_g^2\leq\frac{\alpha_x\eta}{48}\,,\quad 6\alpha_x^2\eta^2\mathcal{M}_1\leq\frac{\alpha_x\eta}{24}\,,\quad 288\tau^2\alpha_x^2\eta^2L_{\rho,\gamma}^2\mathcal{M}\leq\frac{\alpha_x\eta}{24}\,. \tag{55}$$

By solving the first inequality, we derive $\alpha_x\leq\frac{1}{8\sqrt{3c_{\alpha_z}}L_g}$. To solve the second inequality, from the definition of $\mathcal{M}_1$ in Eq. (52), we enforce:

$$\frac{6\alpha_x^2\eta^2}{K}(2L_f^2\rho^2c_1+2L_g^2c_2+2L_g^2c_3)\leq\frac{\alpha_x\eta}{72}\,,$$
$$\frac{6\alpha_x^2\eta^2}{K}(2L_f^2\rho^2c_4+2L_g^2c_5)\leq\frac{\alpha_x\eta}{72}\,,\qquad\frac{6\alpha_x^2\eta^2}{K}2L_g^2c_6\leq\frac{\alpha_x\eta}{72}\,.$$

To solve these inequalities, we apply the value of $\{c_i\}_{i=1}^6$ in Eq. (51), and obtain

$$\alpha_x\leq\min\{\frac{\sqrt{\beta_xK}}{72L_\rho}\,,\frac{\sqrt{\beta_yK}}{24\sqrt{3}L_\rho}\,,\frac{\sqrt{\beta_zK}}{36\sqrt{2c_{\alpha_z}}L_g}\}\,.$$

Similarly, to solve the third inequality, following Eq. (54), we obtain

$$\alpha_x\leq\{\frac{\sqrt{\beta_xK}}{48\sqrt{3}L_\rho}\,,\frac{\sqrt{\beta_yK}}{48L_\rho}\,,\frac{\sqrt{\beta_zK}}{24\sqrt{6c_{\alpha_z}}L_g}\,,\frac{1}{24\sqrt{3}\tau\eta L_\rho}\,,\frac{1}{12\sqrt{6c_{\alpha_z}}\tau\eta L_g}\}\,.$$

In addition, to eliminate $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{v}_t\|^2]$, we enforce

$$\frac{3\alpha_x\alpha_y^2\eta^3L_\rho^2}{2}+4\alpha_y^2\eta^2\mathcal{M}_2+192\tau^2\alpha_y^2\eta^2(L_f^2\rho^2+3L_g^2+3\gamma^2)\mathcal{M}-\frac{\alpha_y\eta}{4}\leq-\frac{\alpha_y\eta}{8}\,.$$

Specifically, we enforce

$$\frac{3\alpha_x\alpha_y^2\eta^3L_\rho^2}{2}\leq\frac{\alpha_y\eta}{48}\,,\quad 4\alpha_y^2\eta^2\mathcal{M}_2\leq\frac{\alpha_y\eta}{24}\,,$$
$$192\tau^2\alpha_y^2\eta^2(L_f^2\rho^2+3L_g^2+3\gamma^2)\mathcal{M}\leq\frac{\alpha_y\eta}{24}\,.$$

From the definitions of $\mathcal{M}_2$ and $\mathcal{M}$, together with the condition $\alpha_y=\alpha_x$, and noting that Eq. (55) holds, the last two inequalities above follow directly.

By solving the first inequality above, we derive $\alpha_x\leq\frac{1}{6\sqrt{2}L_\rho}$.

At last, to eliminate $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{w}_t\|^2]$, we enforce

$$\frac{\alpha_z^2\eta^2}{2}(L_{F,2}+2L_d+2C_\rho)+2\alpha_z^2\eta^2\mathcal{M}_3+48\tau^2\alpha_z^2\eta^2L_g^2\mathcal{M}-\frac{\alpha_z\eta}{16}\leq-\frac{\alpha_z\eta}{32}\,.$$

In particular, we enforce

$$\frac{\alpha_z^2\eta^2}{2}(L_F+2L_d)\leq\frac{\alpha_z\eta}{32\times6}\,,\quad\alpha_z^2\eta^2C_\rho\leq\frac{\alpha_z\eta}{32\times6}\,,$$
$$2\alpha_z^2\eta^2\mathcal{M}_3\leq\frac{\alpha_z\eta}{96}\,,\quad 48\tau^2\alpha_z^2\eta^2L_g^2\mathcal{M}\leq\frac{\alpha_z\eta}{96}\,.$$

By solving the first inequality above, we obtain $\alpha_x\leq\frac{1}{24c_{\alpha_z}(L_F+2L_d)}$.

To solve the second inequality, from Eq. (38), we obtain $\alpha_x\leq\frac{1}{2\sqrt{3}c_{\alpha_z}(\gamma_2-L_\rho)}$.

To solve the third inequality, from the definition of $\mathcal{M}_3$ in Eq. (52), we enforce:

$$\frac{2\alpha_z^2\eta^2}{K}2L_g^2c_3\leq\frac{\alpha_z\eta}{192}\,,\quad\frac{2\alpha_z^2\eta^2}{K}2L_g^2c_6\leq\frac{\alpha_z\eta}{192}\,,$$

and we obtain

$$\alpha_x \leq \min\{\frac{\sqrt{\beta_x K}}{48\sqrt{2c_{\alpha_z}}L_g}\,,\frac{\sqrt{\beta_z K}}{24\sqrt{2}c_{\alpha_z}L_g}\}\,.$$

Similarly, to solve the last inequality, following Eq. (54), we obtain

$$\alpha_x \leq \min\{\frac{\sqrt{\beta_x K}}{48\sqrt{2c_{\alpha_z}}L_g}\,,\frac{\sqrt{\beta_y K}}{16\sqrt{6c_{\alpha_z}}L_g}\,,\frac{\sqrt{\beta_z K}}{48c_{\alpha_z}L_g}\,,\frac{1}{24\sqrt{2c_{\alpha_z}}\tau\eta L_g}\,,\frac{1}{24c_{\alpha_z}\tau\eta L_g}\}\,.$$

In summary, by setting

$$\begin{aligned}
\alpha_x \leq \min\Big\{&\frac{9}{80\gamma_1}\,,\frac{\sqrt{\mu(\gamma_1-L_\rho)^3}(\gamma_2-L_\rho)^2}{24\times 16\sqrt{6\gamma_1 c_{\alpha_{pq}}}L_g c_{\alpha_r}\gamma_2^2}\,,\frac{1}{24c_{\alpha_{pq}}(1+\omega_1^2)}\,,\frac{1}{8\sqrt{3}c_{\alpha_{pq}}\omega_5\sqrt{5\omega_2^2+\frac{2\gamma_2^2}{\mu(\gamma_1-L_\rho)}}}\\
&\frac{\sqrt{\beta_x K}}{72\sqrt{2\gamma_1 c_{\alpha_{pq}}}}\,,\frac{\sqrt{\beta_y K}}{24\sqrt{6\gamma_1 c_{\alpha_{pq}}}}\,,\frac{\sqrt{\beta_z K}}{72\sqrt{\gamma_1 c_{\alpha_z}c_{\alpha_{pq}}}}\,,\frac{1}{36\tau\eta\sqrt{2\gamma_1 c_{\alpha_{pq}}}}\,,\frac{1}{36\tau\eta\sqrt{\gamma_1 c_{\alpha_z}c_{\alpha_{pq}}}}\,,\\
&\frac{(\gamma_1-L_\rho)(\gamma_2-L_\rho)}{16\sqrt{5\gamma_1\gamma_2 c_{\alpha_{pq}}c_{\alpha_r}}L_g}\,,\frac{\sqrt{\mu(\gamma_1-L_\rho)}(\gamma_2-L_\rho)}{32\gamma_2\sqrt{\gamma_1\gamma_2 c_{\alpha_{pq}}c_{\alpha_r}}}\,,\frac{6\sqrt{3\gamma_2}L_g}{\sqrt{c_{\alpha_r}}(\gamma_1-L_\rho)(\gamma_2-L_\rho)}\,,\\
&\frac{\sqrt{\beta_x K}}{48\sqrt{3\gamma_2 c_{\alpha_r}}}\,,\frac{\sqrt{\beta_y K}}{48\sqrt{\gamma_2 c_{\alpha_r}}}\,,\frac{\sqrt{\beta_z K}}{24\sqrt{6\gamma_2 c_{\alpha_z}c_{\alpha_r}}}\,,\frac{1}{24\tau\eta\sqrt{3\gamma_2 c_{\alpha_r}}}\,,\frac{1}{12\tau\eta\sqrt{6\gamma_2 c_{\alpha_z}c_{\alpha_r}}}\,,\\
&\frac{1}{8\sqrt{3c_{\alpha_z}}L_g}\,,\frac{\sqrt{\beta_x K}}{48\sqrt{3}L_\rho}\,,\frac{\sqrt{\beta_y K}}{48L_\rho}\,,\frac{\sqrt{\beta_z K}}{24\sqrt{6c_{\alpha_z}}L_g}\,,\frac{1}{24\sqrt{3}\tau\eta L_\rho}\,,\frac{1}{12\sqrt{6c_{\alpha_z}}\tau\eta L_g}\,,\\
&\frac{1}{6\sqrt{2}L_\rho}\,,\frac{1}{24c_{\alpha_z}(L_F+2L_d)}\,,\frac{1}{2\sqrt{3}c_{\alpha_z}(\gamma_2-L_\rho)}\,,\frac{\sqrt{\beta_x K}}{48\sqrt{2c_{\alpha_z}}L_g}\,,\frac{\sqrt{\beta_y K}}{16\sqrt{6c_{\alpha_z}}L_g}\,,\frac{\sqrt{\beta_z K}}{48c_{\alpha_z}L_g}\,,\\
&\frac{1}{24\sqrt{2c_{\alpha_z}}\tau\eta L_g}\,,\frac{1}{24c_{\alpha_z}\tau\eta L_g}\Big\}\,,\\
\alpha_y = \alpha_x\,,\quad &\alpha_z = \alpha_x\underbrace{\frac{(\gamma_1-L_\rho)^2}{72L_g^2}}_{c_{\alpha_z}=O(1)}\,,\quad \alpha_{pq} = \alpha_x\underbrace{\frac{(\gamma_1-L_\rho)^4(\gamma_2-L_\rho)^2\mu}{72\times 768\gamma_1 L_g^2(5L_g^2\mu+16\gamma_2^2(\gamma_1-L_\rho))}}_{c_{\alpha_{pq}}=O(1/\kappa)}\,,\\
\alpha_r = \alpha_x&\underbrace{\frac{(\gamma_1-L_\rho)^4(\gamma_2-L_\rho)^4}{72\times 24\times 640\gamma_1\gamma_2^2 L_g^4}}_{c_{\alpha_r}=O(1)}\,,
\end{aligned}\tag{56}$$

we obtain

$$\begin{aligned}
\frac{\mathcal{P}_T-\mathcal{P}_0}{T} \leq &-\frac{\alpha_x\eta}{8}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]\\
&-\frac{\alpha_y\eta}{2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_y F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]-\frac{\alpha_z\eta}{2}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_z F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]\\
&-\frac{\gamma_1}{8\alpha_{pq}\eta}\frac{1}{T}\sum_{t=0}^{T-1}\big(\mathbb{E}[\|\bar{p}_{t+1}-\bar{p}_t\|^2]+\mathbb{E}[\|\bar{q}_{t+1}-\bar{q}_t\|^2]\big)-\frac{\gamma_2}{4\alpha_r\eta}\mathbb{E}[\|\bar{r}_{t+1}-\bar{r}_t\|^2]\\
&+\mathcal{M}_0+3\tau S\mathcal{M}\,.
\end{aligned}$$

From

$$\begin{aligned}
\rho^2\mathbb{E}[\|\nabla_x f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2] &= 2\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]+2\gamma_1^2\mathbb{E}[\|\bar{x}_t-\bar{p}_t\|^2]\,,\\
\rho^2\mathbb{E}[\|\nabla_y f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2] &\leq 2\mathbb{E}[\|\nabla_y F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]+2\gamma_1^2\mathbb{E}[\|\bar{y}_t-\bar{q}_t\|^2]\,,\\
\rho^2\mathbb{E}[\|\nabla_z f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2] &\leq 2\mathbb{E}[\|\nabla_z F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]+2\gamma_2^2\mathbb{E}[\|\bar{z}_t-\bar{r}_t\|^2]\,,
\end{aligned}$$

we obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\Big(\rho^2\mathbb{E}[\|\nabla_x f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2]+\rho^2\mathbb{E}[\|\nabla_y f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2]+\rho^2\kappa\mathbb{E}[\|\nabla_z f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2]\Big)\\
&\le \frac{1}{T}\sum_{t=0}^{T-1}\Big(2\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]+2\mathbb{E}[\|\nabla_y F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]\\
&\qquad+2\kappa\mathbb{E}[\|\nabla_z F(\bar{x}_t,\bar{y}_t,\bar{z}_t;\bar{p}_t,\bar{q}_t,\bar{r}_t)\|^2]+2\gamma_1^2\mathbb{E}[\|\bar{x}_t-\bar{p}_t\|^2]+2\gamma_1^2\mathbb{E}[\|\bar{y}_t-\bar{q}_t\|^2]\\
&\qquad+2\gamma_2^2\kappa\mathbb{E}[\|\bar{z}_t-\bar{r}_t\|^2]\Big)\\
&\le \max\{\frac{16}{\alpha_x\eta},\frac{16\kappa}{\alpha_x\eta c_{\alpha_z}},\frac{16\gamma_1}{\alpha_x\eta c_{\alpha_{pq}}},\frac{16\kappa\gamma_2}{\alpha_r\eta c_{\alpha_r}}\}\Big(\frac{\mathcal{P}_0-\mathcal{P}_T}{T}+\mathcal{M}_0+3\tau S\mathcal{M}\Big)\,. \qquad (57)
\end{aligned}$$

When $t=0$, we have

$$\begin{aligned}
\mathcal{P}_0&=\Psi_0+\frac{6\alpha_x}{\beta_x\eta}\rho^2\mathcal{G}_{u,1,0}+\frac{6\alpha_x}{\beta_x\eta}\mathcal{G}_{u,2,0}+\frac{6\alpha_x}{\beta_x\eta}\mathcal{G}_{u,3,0}+\frac{2\alpha_y}{\beta_y\eta}\rho^2\mathcal{G}_{v,1,0}+\frac{2\alpha_y}{\beta_y\eta}\mathcal{G}_{v,2,0}+\frac{6\alpha_z}{\beta_z\eta}\mathcal{G}_{w,0}\\
&\le\Psi_0+\frac{6\alpha_x}{\beta_x\eta}\frac{\sigma^2}{BK}(\rho^2+2)+\frac{2\alpha_x}{\beta_y\eta}\frac{\sigma^2}{BK}(\rho^2+1)+\frac{3\alpha_x c_{\alpha_z}}{\beta_z\eta}\frac{\sigma^2}{BK}\,. \qquad (58)
\end{aligned}$$

Therefore, taking O notation, the first term in Eq. (57) can be shown as

$$\max\{\frac{16}{\alpha_x\eta},\frac{16\kappa}{\alpha_x\eta c_{\alpha_z}},\frac{16\gamma_1}{\alpha_x\eta c_{\alpha_{pq}}},\frac{16\kappa\gamma_2}{\alpha_r\eta c_{\alpha_r}}\}\frac{\mathcal{P}_0-\mathcal{P}_T}{T}\le O\Big(\frac{\Psi_0\kappa}{\alpha_x\eta T}\Big)+O\Big(\frac{\kappa}{\beta\eta^2BKT}\Big)\,.$$

Similarly, the second term in Eq. (57) can be shown as

$$\begin{aligned}
\mathcal{M}_0&=\frac{2\sigma^2}{K}\Big(\beta_x^2\eta^4\frac{6\alpha_x}{\beta_x\eta}(\rho^2+2)+\beta_y^2\eta^4(\rho^2+1)\frac{2\alpha_y}{\beta_y\eta}+\beta_z^2\eta^4\frac{3\alpha_z}{\beta_z\eta}\Big)\,,\\
&\max\{\frac{16}{\alpha_x\eta},\frac{16\kappa}{\alpha_x\eta c_{\alpha_z}},\frac{16\gamma_1}{\alpha_x\eta c_{\alpha_{pq}}},\frac{16\kappa\gamma_2}{\alpha_r\eta c_{\alpha_r}}\}\mathcal{M}_0\le O\Big(\frac{\kappa\beta\eta^2}{K}\Big)\,. \qquad (59)
\end{aligned}$$

At last, the third term in Eq. (57) can be shown as

$$\begin{aligned}
S&=12\tau(\beta_x^2+\beta_y^2)\eta^4\delta^2\rho^2+8\tau(\beta_x^2+\beta_y^2)\eta^4\sigma^2\rho^2+12\tau(4\beta_x^2+3\beta_y^2+\beta_z^2)\eta^4\delta^2\\
&\quad+8\tau(4\beta_x^2+3\beta_y^2+\beta_z^2)\eta^4\sigma^2=O\Big(\tau\beta^2\eta^4\Big)\\
\mathcal{M}&=\frac{6(\alpha_x^2+\alpha_y^2+\alpha_z^2)\eta^2}{K}(2L_f^2\rho^2c_1+2L_g^2c_2+2L_g^2c_3+2L_f^2\rho^2c_4+2L_g^2c_5+2L_g^2c_6)\\
&\quad+3\tau^2(\alpha_x^2+\alpha_y^2+\alpha_z^2)\eta^2\Big(6\alpha_x\eta(\rho^2L_f^2+2L_g^2)+2\alpha_y\eta(\rho^2L_f^2+L_g^2)+3\alpha_z\eta L_g^2\Big)\\
&=O\Big(\frac{\alpha_x^3\eta}{\beta K}(\rho^2+1)\Big)+O\Big(\tau^2\alpha_x^3\eta^3(\rho^2+1)\Big)\,. \qquad (60)
\end{aligned}$$

Therefore, we obtain

$$\max\{\frac{16}{\alpha_x\eta},\frac{16\kappa}{\alpha_x\eta c_{\alpha_z}},\frac{16\gamma_1}{\alpha_x\eta c_{\alpha_{pq}}},\frac{16\kappa\gamma_2}{\alpha_r\eta c_{\alpha_r}}\}3\tau S\mathcal{M}\le O\Big(\frac{\kappa\tau^2\beta\eta^4\alpha_x^2}{K}\Big)+O\Big(\kappa\tau^4\alpha_x^2\beta^2\eta^6\Big)\,. \qquad (61)$$

In summary, we establish the convergence rate

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\Big(\mathbb{E}[\|\nabla_x f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2]+\mathbb{E}[\|\nabla_y f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2]+\kappa\mathbb{E}[\|\nabla_z f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2]\Big)\\
&\le\frac{1}{\rho^2}\max\{\frac{16}{\alpha_x\eta},\frac{16\kappa}{\alpha_x\eta c_{\alpha_z}},\frac{16\gamma_1}{\alpha_x\eta c_{\alpha_{pq}}},\frac{16\kappa\gamma_2}{\alpha_r\eta c_{\alpha_r}}\}\Big(\frac{\mathcal{P}_0-\mathcal{P}_T}{T}+\mathcal{M}_0+3\tau S\mathcal{M}\Big) \qquad (62)\\
&\le O\Big(\frac{\kappa(\Psi_0-\Psi_T)}{\alpha_x\eta\rho^2T}\Big)+O\Big(\frac{\kappa}{\beta\eta^2\rho^2BKT}\Big)+O\Big(\frac{\kappa\beta\eta^2}{\rho^2K}\Big)+O\Big(\frac{\kappa\tau^2\beta\eta^4\alpha_x^2}{\rho^2K}\Big)+O\Big(\frac{\kappa\tau^4\alpha_x^2\beta^2\eta^6}{\rho^2}\Big)\,.
\end{aligned}$$

Since $F(x,y,z;p,q,z) = \rho f(x,y) + (g(x,y) - g(x,z)) + \frac{\gamma_1}{2}\|x-p\|^2 + \frac{\gamma_1}{2}\|y-q\|^2 - \frac{\gamma_2}{2}\|z-r\|^2$, we need to scale $F(x,y,z;p,q,r)$ by $\frac{1}{\rho}$ to study the original penalized function $f(x,y) + \frac{1}{\rho}(g(x,y) - g(x,z))$. Then according to Zhang et al. (2025), by initializing y_0 and z_0 such that $\mathbb{E}[\|y_0 - y^*_{1/\rho}(x_0)\|^2] \le O(\epsilon^2)$ and $\mathbb{E}[\|z_0 - y^*(x_0)\|^2] \le O(\epsilon^2)$ where $y^*_{1/\rho}(x) = \arg\min_y f(x,y) + \frac{1}{\rho}g(x,y)$, we can mitigate the influence of $\frac{1}{\rho}$ in the initial upper bound. As a result, we can obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}[\|\nabla_x f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2] + \mathbb{E}[\|\nabla_y f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2] + \kappa\mathbb{E}[\|\nabla_z f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2]\right) \\
&\le O\Big(\frac{\kappa}{\alpha_x\eta\rho T}\Big) + O\Big(\frac{\kappa}{\beta\eta^2\rho^2 BKT}\Big) + O\Big(\frac{\kappa\beta\eta^2}{\rho^2 K}\Big) + O\Big(\frac{\kappa\tau^2\beta\eta^4\alpha_x^2}{\rho^2 K}\Big) + O\Big(\frac{\kappa\tau^4\alpha_x^2\beta^2\eta^6}{\rho^2}\Big)\,. \quad (63)
\end{aligned}$$

By setting

$$\begin{aligned}
&\rho = O\left(\frac{\epsilon}{\kappa^3}\right)\,, \quad \alpha_x = O\left(1\right)\,, \\
&\beta_x = O\left(\frac{1}{K}\right)\,, \quad \beta_y = O\left(\frac{1}{K}\right)\,, \quad \beta_z = O\left(\frac{1}{K}\right)\,, \\
&\eta = O\left(\frac{K\epsilon^2}{\kappa^{7/2}}\right)\,, \tau = O\left(\frac{\kappa^{7/2}}{K\epsilon^2}\right)\,, \quad B = O\left(\frac{\kappa^{13/2}}{K\epsilon^3}\right)\,, \quad T = O\left(\frac{\kappa^{15/2}}{K\epsilon^5}\right)\,. \quad (64)
\end{aligned}$$

we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}[\|\nabla_x f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2] + \mathbb{E}[\|\nabla_y f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2] + \kappa\mathbb{E}[\|\nabla_z f_\rho(\bar{x}_t,\bar{y}_t,\bar{z}_t)\|^2]\right) \le O(\epsilon^2)\,.$$

G Translation between Stationarity Measures

Algorithm 3 LocalSGDA (Deng & Mahdavi, 2021)

Input: $\eta_\alpha > 0$, $\eta_\theta > 0$. Initialization on worker k: $\alpha_0^{(k)} = \alpha_0$, $\theta_0^{(k)} = \theta_0$,

1: **for** $t = 0, \cdots, T-1$, worker k **do**
2: Update α: $\alpha_{t+1}^{(k)} = \alpha_t^{(k)} - \eta_\alpha \nabla_1 \tilde{f}_\rho^{(k)}(\alpha_t^{(k)}, \theta_t^{(k)}; \xi_t^{(k)})$,
3: Update θ: $\theta_{t+1}^{(k)} = \theta_t^{(k)} + \eta_\theta \nabla_2 \tilde{f}_\rho^{(k)}(\alpha_t^{(k)}, \theta_t^{(k)}; \xi_t^{(k)})$,
4: **if** $\mathrm{mod}(t+1, \tau) == 0$ **then**
5: $\alpha_{t+1}^{(k)} = \frac{1}{K}\sum_{k'=1}^K \alpha_{t+1}^{(k')}$, $\theta_{t+1}^{(k)} = \frac{1}{K}\sum_{k'=1}^K \theta_{t+1}^{(k')}$,
6: **end if**
7: **end for**

In this section, we present a detailed treatment of the translation between the stationarity measures defined in Definition 4.7, and complete the proof of Remark 4.10. To establish this connection in the federated setting, we employ the LocalSGDA method (Algorithm 3) on an auxiliary two-sided PL minimax problem given by

$$\min_{\alpha \triangleq (x,y)} \max_{\theta \triangleq z} \tilde{f}_\rho(\alpha,\theta) \triangleq \frac{1}{K}\sum_{k=1}^K \underbrace{f_\rho^{(k)}(\alpha,\theta) + L\|\alpha - \tilde{\alpha}\|^2}_{\tilde{f}_\rho^{(k)}(\alpha,\theta)}\,, \quad (65)$$

where $\tilde{\alpha}$ is the output of by Algorithm 1. The resulting objective $\tilde{f}_\rho(\alpha,\theta)$ possesses several useful properties: it is L-strongly convex in α, satisfies the μ-PL condition with respect to θ, and is $\tilde{L}$-smooth with $\tilde{L} = 3L$. We next characterize the convergence behavior of Algorithm 3 under the strongly-convex–PL assumption and show that its convergence rate is dominated by that of Algorithm 1; therefore, the overall convergence speed is also dominated by Algorithm 1.

Theorem G.1. *Given the smoothness and strongly-convex-PL assumptions on $\tilde{f}_\rho(\alpha,\theta)$, choosing step sizes $\eta_\alpha = O(\epsilon^2/\kappa^5)$ and $\eta_\theta = 128\kappa^2\eta_\alpha$, then Algorithm 3 has a convergence rate of $T = O(\epsilon^{-2}\kappa^6\log(\kappa))$ to achieve the ϵ-accuracy solution: $\mathbb{E}[\|\alpha_T - \alpha^*\|^2] \le O(\epsilon^2)$, where $0 < \epsilon < 1$.*

Proof. To prove Theorem G.1, we introduce the following auxiliary functions:

$$\tilde{\Phi}(\alpha) = \max_{\alpha} \tilde{f}_\rho(\alpha, \theta) = \max_{\theta} f_\rho(\alpha, \theta) + L\|\alpha - \tilde{\alpha}\|^2 \,,$$
$$\tilde{\Psi}(\theta) = \min_{\alpha} \tilde{f}_\rho(\alpha, \theta) = \min_{\alpha} f_\rho(\alpha, \theta) + L\|\alpha - \tilde{\alpha}\|^2 \,, \quad \tilde{\Phi}^* = \max_{\theta} \tilde{\Psi}(\theta) \,.$$

In addition, we denote $\bar{\alpha}_t = \frac{1}{K}\sum_{k=1}^K \alpha_t^{(k)}$, $\bar{\theta}_t = \frac{1}{K}\sum_{k=1}^K \theta_t^{(k)}$, and:

$$\hat{m}_{\alpha,t}^{(k)} = \nabla_1 \tilde{f}_\rho^{(k)}(\alpha_t^{(k)}, \theta_t^{(k)}; \xi_t^{(k)}) \,, \quad \hat{m}_{\theta,t}^{(k)} = \nabla_2 \tilde{f}_\rho^{(k)}(\alpha_t^{(k)}, \theta_t^{(k)}; \xi_t^{(k)}) \,,$$
$$m_{\alpha,t}^{(k)} := \nabla_1 \tilde{f}_\rho^{(k)}(\alpha_t^{(k)}, \theta_t^{(k)}) \,, \quad m_{\theta,t}^{(k)} := \nabla_2 \tilde{f}_\rho^{(k)}(\alpha_t^{(k)}, \theta_t^{(k)}) \,.$$

Given that the $\tilde{L}$-smooth of $\tilde{f}_\rho(\alpha, \theta)$ with respect to α, we derive

$$\mathbb{E}[\tilde{f}_\rho(\bar{\alpha}_t, \bar{\theta}_t)] - \mathbb{E}[\tilde{f}_\rho(\bar{\alpha}_{t+1}, \bar{\theta}_t)] \le -\mathbb{E}[\langle \nabla_1 \tilde{f}_\rho(\bar{\alpha}_t, \bar{\theta}_t), \bar{\alpha}_{t+1} - \bar{\alpha}_t \rangle] + \frac{\tilde{L}}{2}\mathbb{E}[\|\bar{\alpha}_{t+1} - \bar{\alpha}_t\|^2]$$
$$\le \eta_\alpha \mathbb{E}[\langle \nabla_1 \tilde{f}_\rho(\bar{\alpha}_t, \bar{\theta}_t), \frac{1}{K}\sum_{k=1}^K m_{\alpha,t}^{(k)} \rangle] + \frac{\eta_\alpha^2 \tilde{L}}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K m_{\alpha,t}^{(k)}\|^2] + \frac{\eta_\alpha^2 \tilde{L}\sigma^2}{2}$$
$$\le \frac{\eta_\alpha}{2}\mathbb{E}[\|\nabla_1 \tilde{f}_\rho(\bar{\alpha}_t, \bar{\theta}_t)\|^2] + (\frac{\eta_\alpha}{2} + \frac{\eta_\alpha^2 \tilde{L}}{2})\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K m_{\alpha,t}^{(k)}\|^2] + \frac{\eta_\alpha^2 \tilde{L}\sigma^2}{2} \,.$$

Similarly, given that the $\tilde{L}$-smooth of $\tilde{f}_\rho(\alpha, \theta)$ with respect to θ, we derive

$$\mathbb{E}[\tilde{f}_\rho(\bar{\alpha}_{t+1}, \bar{\theta}_t)] - \mathbb{E}[\tilde{f}_\rho(\bar{\alpha}_{t+1}, \bar{\theta}_{t+1})] \le -\mathbb{E}[\langle \nabla_2 \tilde{f}_\rho(\bar{\alpha}_{t+1}, \bar{\theta}_t), \bar{\theta}_{t+1} - \bar{\theta}_t \rangle] + \frac{\tilde{L}}{2}\mathbb{E}[\|\bar{\theta}_{t+1} - \bar{\theta}_t\|^2]$$
$$\le -\eta_\theta \mathbb{E}[\langle \nabla_2 \tilde{f}_\rho(\bar{\alpha}_{t+1}, \bar{\theta}_t), \frac{1}{K}\sum_{k=1}^K m_{\theta,t}^{(k)} \rangle] + \frac{\eta_\theta^2 \tilde{L}}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K m_{\theta,t}^{(k)}\|^2] + \frac{\eta_\theta^2 \tilde{L}\sigma^2}{2}$$
$$\le -\frac{\eta_\theta}{2}\mathbb{E}[\|\nabla_2 \tilde{f}_\rho(\bar{\alpha}_{t+1}, \bar{\theta}_t)\|^2] - \frac{\eta_\theta}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K m_{\theta,t}^{(k)}\|^2] + \eta_\theta \mathbb{E}[\|\nabla_2 \tilde{f}_\rho(\bar{\alpha}_t, \bar{\theta}_t) - \frac{1}{K}\sum_{k=1}^K m_{\theta,t}^{(k)}\|^2]$$
$$+ \eta_\theta \mathbb{E}[\|\nabla_2 \tilde{f}_\rho(\bar{\alpha}_{t+1}, \bar{\theta}_t) - \nabla_2 \tilde{f}_\rho(\bar{\alpha}_t, \bar{\theta}_t)\|^2] + \frac{\eta_\theta^2 \tilde{L}}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K m_{\theta,t}^{(k)}\|^2] + \frac{\eta_\theta^2 \tilde{L}\sigma^2}{2} \,.$$

From $-\|\nabla_2 \tilde{f}_\rho(\bar{\alpha}_{t+1}, \bar{\theta}_t)\|^2 \le -\frac{1}{2}\|\nabla_2 \tilde{f}_\rho(\bar{\alpha}_t, \bar{\theta}_t)\|^2 + \|\nabla_2 \tilde{f}_\rho(\bar{\alpha}_{t+1}, \bar{\theta}_t) - \nabla_2 \tilde{f}_\rho(\bar{\alpha}_t, \bar{\theta}_t)\|$ and $\eta_\theta \le \frac{1}{2\tilde{L}}$, we obtain

$$\mathbb{E}[\tilde{f}_\rho(\bar{\alpha}_{t+1}, \bar{\theta}_t)] - \mathbb{E}[\tilde{f}_\rho(\bar{\alpha}_{t+1}, \bar{\theta}_{t+1})]$$
$$\le -\frac{\eta_\theta}{4}\mathbb{E}[\|\nabla_2 \tilde{f}_\rho(\bar{\alpha}_t, \bar{\theta}_t)\|^2] - \frac{\eta_\theta}{4}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K m_{\theta,t}^{(k)}\|^2] + \eta_\theta \mathbb{E}[\|\nabla_2 \tilde{f}_\rho(\bar{\alpha}_t, \bar{\theta}_t) - \frac{1}{K}\sum_{k=1}^K m_{\theta,t}^{(k)}\|^2]$$
$$+ \frac{3\eta_\theta}{2}\mathbb{E}[\|\nabla_2 \tilde{f}_\rho(\bar{\alpha}_{t+1}, \bar{\theta}_t) - \nabla_2 \tilde{f}_\rho(\bar{\alpha}_t, \bar{\theta}_t)\|^2] + \frac{\eta_\theta^2 \tilde{L}\sigma^2}{2} \,.$$

Then, combining

$$\mathbb{E}[\|\nabla_2 \tilde{f}_\rho(\bar{\alpha}_t, \bar{\theta}_t) - \frac{1}{K}\sum_{k=1}^K m_{\theta,t}^{(k)}\|^2] \le \tilde{L}^2 \frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{\alpha}_t - \alpha_t^{(k)}\|^2] + \tilde{L}^2 \frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{\theta}_t - \theta_t^{(k)}\|^2] \,,$$

and

$$\mathbb{E}[\|\nabla_2 \tilde{f}_\rho(\bar{\alpha}_{t+1}, \bar{\theta}_t) - \nabla_2 \tilde{f}_\rho(\bar{\alpha}_t, \bar{\theta}_t)\|^2] \le \eta_\alpha^2 \tilde{L}^2 \mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \hat{m}_{\alpha,t}^{(k)}\|^2]$$
$$\le \eta_\alpha^2 \tilde{L}^2 \mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K m_{\alpha,t}^{(k)}\|^2] + \eta_\alpha^2 \tilde{L}^2 \sigma^2 \,,$$

we obtain

$$\begin{aligned}
&\mathbb{E}[\tilde{f}_\rho(\bar{\alpha}_{t+1},\bar{\theta}_t)] - \mathbb{E}[\tilde{f}_\rho(\bar{\alpha}_{t+1},\bar{\theta}_{t+1})] \\
&\le -\frac{\eta_\theta}{4}\mathbb{E}[\|\nabla_2 \tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2] - \frac{\eta_\theta}{4}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K m_{\theta,t}^{(k)}\|^2] + \frac{3\eta_\theta\eta_\alpha^2\tilde{L}^2}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K m_{\alpha,t}^{(k)}\|^2] \\
&\quad + \eta_\theta\tilde{L}^2\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{\alpha}_t - \alpha_t^{(k)}\|^2] + \eta_\theta\tilde{L}^2\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{\theta}_t - \theta_t^{(k)}\|^2] + \frac{3\eta_\theta\eta_\alpha^2\tilde{L}^2\sigma^2}{2} + \frac{\eta_\theta^2\tilde{L}\sigma^2}{2}\,.
\end{aligned}$$

Combining the upper bound of $\mathbb{E}[\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)] - \mathbb{E}[\tilde{f}_\rho(\bar{\alpha}_{t+1},\bar{\theta}_t)]$ and $\mathbb{E}[\tilde{f}_\rho(\bar{\alpha}_{t+1},\bar{\theta}_t)] - \mathbb{E}[\tilde{f}_\rho(\bar{\alpha}_{t+1},\bar{\theta}_{t+1})]$ with $\eta_\theta \le \frac{1}{2\tilde{L}}$, we obtain

$$\begin{aligned}
&\mathbb{E}[\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)] - \mathbb{E}[\tilde{f}_\rho(\bar{\alpha}_{t+1},\bar{\theta}_{t+1})] \le \frac{\eta_\alpha}{2}\mathbb{E}[\|\nabla_1 \tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2] - \frac{\eta_\theta}{4}\mathbb{E}[\|\nabla_2 \tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2] \\
&+ (\frac{\eta_\alpha}{2} + \frac{5\eta_\alpha^2\tilde{L}}{4})\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K m_{\alpha,t}^{(k)}\|^2] - \frac{\eta_\theta}{4}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K m_{\theta,t}^{(k)}\|^2] + \eta_\theta\tilde{L}^2\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{\alpha}_t - \alpha_t^{(k)}\|^2] \\
&+ \eta_\theta\tilde{L}^2\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{\theta}_t - \theta_t^{(k)}\|^2] + \frac{5\eta_\alpha^2\tilde{L}\sigma^2}{4} + \frac{\eta_\theta^2\tilde{L}\sigma^2}{2}\,. \qquad (66)
\end{aligned}$$

Moreover, using the $L_{\tilde{\Phi}}$-smoothness of $\tilde{\Phi}(\alpha)$, with $L_{\tilde{\Phi}} = O(\kappa)$ Lin et al. (2020), we derive

$$\begin{aligned}
&\mathbb{E}[\tilde{\Phi}(\bar{\alpha}_{t+1})] - \mathbb{E}[\tilde{\Phi}(\bar{\alpha}_t)] \le \mathbb{E}[\langle\nabla\tilde{\Phi}(\bar{\alpha}_t), \bar{\alpha}_{t+1} - \bar{\alpha}_t\rangle] + \frac{L_{\tilde{\Phi}}}{2}\mathbb{E}[\|\bar{\alpha}_{t+1} - \bar{\alpha}_t\|^2] \\
&\le -\eta_\alpha\mathbb{E}[\langle\nabla\tilde{\Phi}(\bar{\alpha}_t), \frac{1}{K}\sum_{k=1}^K m_{\alpha,t}^{(k)}\rangle] + \frac{\eta_\alpha^2 L_{\tilde{\Phi}}}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K m_{\alpha,t}^{(k)}\|^2] + \frac{\eta_\alpha^2 L_{\tilde{\Phi}}\sigma^2}{2} \\
&\le 8\eta_\alpha\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{\alpha}_t) - \nabla_1\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2] + \frac{\eta_\alpha}{8}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K m_{\alpha,t}^{(k)}\|^2] \\
&\quad - \frac{\eta_\alpha}{2}\mathbb{E}[\|\nabla_1\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2] - \frac{\eta_\alpha}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K m_{\alpha,t}^{(k)}\|^2] + \frac{\eta_\alpha}{2}\mathbb{E}[\|\nabla_1\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t) - \frac{1}{K}\sum_{k=1}^K m_{\alpha,t}^{(k)}\|^2] \\
&\quad + \frac{\eta_\alpha^2 L_{\tilde{\Phi}}}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K m_{\alpha,t}^{(k)}\|^2] + \frac{\eta_\alpha^2 L_{\tilde{\Phi}}\sigma^2}{2} \\
&\le 8\eta_\alpha\frac{\tilde{L}^2}{\mu^2}\mathbb{E}[\|\nabla_2\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2] - \frac{\eta_\alpha}{2}\mathbb{E}[\|\nabla_1\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2] + (\frac{\eta_\alpha^2 L_{\tilde{\Phi}}}{2} - \frac{3\eta_\alpha}{8})\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K m_{\alpha,t}^{(k)}\|^2] \\
&\quad + \frac{\eta_\alpha\tilde{L}^2}{2}\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{\alpha}_t - \alpha_t^{(k)}\|^2] + \frac{\eta_\alpha\tilde{L}^2}{2}\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{\theta}_t - \theta_t^{(k)}\|^2] + \frac{\eta_\alpha^2 L_{\tilde{\Phi}}\sigma^2}{2}\,, \qquad (67)
\end{aligned}$$

where the last step follows from that $\tilde{f}_\rho(\alpha,\theta)$ satisfies the μ-PL condition with respect to θ.

We next define a potential function of the form:

$$\mathcal{L}_t = 2\tilde{\Phi}(\bar{\alpha}_t) - \tilde{\Phi}^* - \tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\,,$$

By invoking Eq. (66) and Eq. (67), and choosing $\eta_\alpha = \frac{\eta_\theta}{128\kappa^2}$ with $\kappa = \frac{\tilde{L}}{\mu} > 1$, and setting $\eta_\alpha \le \frac{1}{2(5\tilde{L}+4L_{\tilde{\Phi}})}$, we obtain

$$\begin{aligned}
&\mathcal{L}_{t+1} - \mathcal{L}_t \\
&= 2(\tilde{\Phi}(\bar{\alpha}_{t+1}) - \tilde{\Phi}(\bar{\alpha}_t)) + \tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t) - \tilde{f}_\rho(\bar{\alpha}_{t+1},\bar{\theta}_{t+1}) \\
&\le -\frac{\eta_\alpha}{2}\mathbb{E}[\|\nabla_1\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2] - \frac{\eta_\theta}{8}\mathbb{E}[\|\nabla_2\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2] - \frac{\eta_\alpha}{8}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K m_{\alpha,t}^{(k)}\|^2]
\end{aligned}$$

$$- \frac{\eta_\theta}{4}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^{K} m_{\theta,t}^{(k)}\|^2] + (\eta_\alpha + \eta_\theta)\tilde{L}^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\alpha}_t - \alpha_t^{(k)}\|^2]$$
$$+ (\eta_\alpha + \eta_\theta)\tilde{L}^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\theta}_t - \theta_t^{(k)}\|^2] + \eta_\alpha^2 L_{\tilde{\Phi}}\sigma^2 + \frac{5\eta_\alpha^2\tilde{L}\sigma^2}{4} + \frac{\eta_\theta^2\tilde{L}\sigma^2}{2}\,. \tag{68}$$

Given that $\tilde{f}_\rho(\alpha,\theta)$ satisfies the μ-PL condition with respect to θ, we obtain

$$\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{\alpha}_t)\|^2] \le 2\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{\alpha}_t) - \nabla_1\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2] + 2\mathbb{E}[\|\nabla_1\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2]$$
$$\le \frac{2\tilde{L}^2}{\mu^2}\mathbb{E}[\|\nabla_2\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2] + 2\mathbb{E}[\|\nabla_1\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2]\,.$$

Then, by setting $\eta_\alpha = \frac{\eta_\theta}{128\kappa^2}$ and using the fact that $\tilde{f}_\rho(\alpha,\theta)$ satisfies the μ-PL condition with respect to θ, we obtain

$$\frac{\eta_\alpha}{2}\mathbb{E}[\|\nabla_1\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2] + \frac{\eta_\theta}{8}\mathbb{E}[\|\nabla_2\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2]$$
$$\ge \frac{\eta_\alpha}{4}\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{\alpha}_t)\|^2] - \frac{\eta_\alpha}{2}\frac{\tilde{L}^2}{\mu^2}\mathbb{E}[\|\nabla_2\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2] + \frac{\eta_\theta}{8}\mathbb{E}[\|\nabla_2\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2]$$
$$\ge \frac{\eta_\alpha}{4}\mathbb{E}[\|\nabla\tilde{\Phi}(\bar{\alpha}_t)\|^2] + \frac{\eta_\theta}{16}\mathbb{E}[\|\nabla_2\tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t)\|^2]$$
$$\ge \frac{\eta_\alpha\mu}{2}(\tilde{\Phi}(\bar{\alpha}_t) - \tilde{\Phi}^*) + \frac{\eta_\theta\mu}{8}(\tilde{\Phi}(\alpha_t) - \tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t))$$
$$\ge \frac{\eta_\alpha\mu}{2}(\tilde{\Phi}(\bar{\alpha}_t) - \tilde{\Phi}^*) + \frac{\eta_\alpha\mu}{2}(\tilde{\Phi}(\alpha_t) - \tilde{f}_\rho(\bar{\alpha}_t,\bar{\theta}_t))$$
$$= \frac{\eta_\alpha\mu}{2}\mathcal{L}_t\,.$$

Rearranging Eq. (68) yields

$$\mathcal{L}_{t+1} \le (1 - \frac{\eta_\alpha\mu}{2})\mathcal{L}_t + \eta_\alpha^2 L_{\tilde{\Phi}}\sigma^2 + \frac{5\eta_\alpha^2\tilde{L}\sigma^2}{4} + \frac{\eta_\theta^2\tilde{L}\sigma^2}{2}$$
$$+ (\eta_\alpha + \eta_\theta)\tilde{L}^2\Big(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\alpha}_t - \alpha_t^{(k)}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\theta}_t - \theta_t^{(k)}\|^2]\Big)$$
$$\le (1 - \frac{\eta_\alpha\mu}{2})^{t+1}\mathcal{L}_0 + \frac{2\eta_\alpha L_{\tilde{\Phi}}\sigma^2}{\mu} + \frac{5\eta_\alpha\tilde{L}\sigma^2}{2\mu} + \frac{\eta_\theta^2\tilde{L}\sigma^2}{\eta_\alpha\mu}$$
$$+ \frac{2(\eta_\alpha + \eta_\theta)}{\eta_\alpha\mu}\tilde{L}^2\sum_{t'=0}^{t}\Big(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\alpha}_{t'} - \alpha_{t'}^{(k)}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\theta}_{t'} - \theta_{t'}^{(k)}\|^2]\Big)\,.$$

For the final term, we begin by establishing a bound on the consensus error associated with α:

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\alpha}_t - \alpha_t^{(k)}\|^2] = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\alpha}_{s_t\tau} - \eta_\alpha\sum_{t'=s_t\tau}^{t-1}\frac{1}{K}\sum_{k'=1}^{K}\hat{m}_{\alpha,t'}^{(k')} - (\alpha_{s_t\tau}^{(k)} - \eta_\alpha\sum_{t'=s_t\tau}^{t-1}\hat{m}_{\alpha,t'}^{(k)})\|^2]$$
$$= \tau\eta_\alpha^2\frac{1}{K}\sum_{k=1}^{K}\sum_{t'=s_t\tau}^{t-1}\mathbb{E}[\|\frac{1}{K}\sum_{k'=1}^{K}\hat{m}_{\alpha,t'}^{(k')} - \hat{m}_{\alpha,t'}^{(k)}\|^2]$$
$$= \tau\eta_\alpha^2\frac{1}{K}\sum_{k=1}^{K}\sum_{t'=s_t\tau}^{t-1}\mathbb{E}[\|\frac{1}{K}\sum_{k'=1}^{K}\hat{m}_{\alpha,t'}^{(k')} - \frac{1}{K}\sum_{k'=1}^{K}m_{\alpha,t'}^{(k')} + \frac{1}{K}\sum_{k'=1}^{K}m_{\alpha,t'}^{(k')} - m_{\alpha,t'}^{(k)} + m_{\alpha,t'}^{(k)} - \hat{m}_{\alpha,t'}^{(k)}\|^2]$$
$$\le \tau\eta_\alpha^2\frac{1}{K}\sum_{k=1}^{K}\sum_{t'=s_t\tau}^{t-1}\mathbb{E}[\|\frac{1}{K}\sum_{k'=1}^{K}\hat{m}_{\alpha,t'}^{(k')} - \frac{1}{K}\sum_{k'=1}^{K}m_{\alpha,t'}^{(k')}\|^2] + \tau\eta_\alpha^2\frac{1}{K}\sum_{k=1}^{K}\sum_{t'=s_t\tau}^{t-1}\mathbb{E}[\|m_{\alpha,t'}^{(k)} - \hat{m}_{\alpha,t'}^{(k)}\|^2]$$
$$+ 3\tau\eta_\alpha^2\frac{1}{K}\sum_{k=1}^{K}\sum_{t'=s_t\tau}^{t-1}\mathbb{E}[\|\frac{1}{K}\sum_{k'=1}^{K}m_{\alpha,t'}^{(k')} - \frac{1}{K}\sum_{k'=1}^{K}\nabla_1\tilde{f}_\rho^{(k')}(\bar{\alpha}_{t'},\bar{\theta}_{t'})\|^2]$$

$$
\begin{aligned}
&+ 3\tau\eta_\alpha^2 \frac{1}{K}\sum_{k=1}^{K}\sum_{t'=s_t\tau}^{t-1}\mathbb{E}[\|\frac{1}{K}\sum_{k'=1}^{K}\nabla_1 \tilde{f}_\rho^{(k')}(\bar{\alpha}_{t'},\bar{\theta}_{t'}) - \nabla_1 \tilde{f}_\rho^{(k)}(\bar{\alpha}_{t'},\bar{\theta}_{t'})\|^2] \\
&+ 3\tau\eta_\alpha^2 \frac{1}{K}\sum_{k=1}^{K}\sum_{t'=s_t\tau}^{t-1}\mathbb{E}[\|\nabla_1 \tilde{f}_\rho^{(k)}(\bar{\alpha}_{t'},\bar{\theta}_{t'}) - m_{\alpha,t'}^{(k)}\|^2] \\
\leq &\sum_{t'=s_t\tau}^{t-1} 2\tau\eta_\alpha^2\sigma^2 + \sum_{t'=s_t\tau}^{t-1} 3\tau\eta_\alpha^2\delta^2 + 6\tau\eta_\alpha^2\tilde{L}^2\frac{1}{K}\sum_{k=1}^{K}\sum_{t'=s_t\tau}^{t-1}\Big(\mathbb{E}[\|\bar{\alpha}_{t'} - \alpha_{t'}^{(k)}\|^2] + \mathbb{E}[\|\bar{\theta}_{t'} - \theta_{t'}^{(k)}\|^2]\Big)
\end{aligned}
$$

where the second step holds due to $\bar{\alpha}_{s_t\tau} = \alpha_{s_t\tau}^{(k)}$, the last step holds due to Assumptions 4.4-4.5. Similarly, we obtain a bound on the consensus error associated with θ:

$$
\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\theta}_t - \theta_t^{(k)}\|^2] &\leq 2\sum_{t'=s_t\tau}^{t-1}\tau\eta_\theta^2\sigma^2 + 3\sum_{t'=s_t\tau}^{t-1}\tau\eta_\theta^2\sigma^2 \\
&+ 6\tau\eta_\theta^2\tilde{L}^2\frac{1}{K}\sum_{k=1}^{K}\sum_{t'=s_t\tau}^{t-1}\Big(\mathbb{E}[\|\bar{\alpha}_{t'} - \alpha_{t'}^{(k)}\|^2] + \mathbb{E}[\|\bar{\theta}_{t'} - \theta_{t'}^{(k)}\|^2]\Big) .
\end{aligned}
$$

Combining together and by summing up t' from $s_t\tau$ to $(s_t+1)\tau - 1$, we derive

$$
\begin{aligned}
&\sum_{t'=s_t\tau}^{(s_t+1)\tau-1}\frac{1}{K}\sum_{k=1}^{K}\Big(\mathbb{E}[\|\bar{\alpha}_t - \alpha_t^{(k)}\|^2] + \mathbb{E}[\|\bar{\theta}_t - \theta_t^{(k)}\|^2]\Big) \leq \sum_{t'=s_t\tau}^{(s_t+1)\tau-1}\tau(\eta_\alpha^2+\eta_\theta^2)(2\sigma^2+3\delta^2) \\
&+ 6\tau^2(\eta_\alpha^2+\eta_\theta^2)\tilde{L}^2\frac{1}{K}\sum_{k=1}^{K}\sum_{t'=s_t\tau}^{(s_t+1)\tau-1}\Big(\mathbb{E}[\|\bar{\alpha}_{t'} - \alpha_{t'}^{(k)}\|^2] + \mathbb{E}[\|\bar{\theta}_{t'} - \theta_{t'}^{(k)}\|^2]\Big) .
\end{aligned}
$$

Setting $\eta_\alpha \leq \frac{1}{2\sqrt{6}\tau\tilde{L}}$, $\eta_\alpha \leq \frac{1}{2\sqrt{6}\tau\tilde{L}}$ yields

$$
\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\Big(\mathbb{E}[\|\bar{\alpha}_t - \alpha_t^{(k)}\|^2] + \mathbb{E}[\|\bar{\theta}_t - \theta_t^{(k)}\|^2]\Big) \leq \tau^2(\eta_\alpha^2+\eta_\theta^2)(4\sigma^2+6\sigma^2) .
$$

Plugging this upper bound to above inequality, we obtain

$$
\begin{aligned}
\mathcal{L}_T &\leq (1-\frac{\eta_\alpha\mu}{2})^T\mathcal{L}_0 + \frac{2\eta_\alpha L_{\tilde{\Phi}}\sigma^2}{\mu} + \frac{5\eta_\alpha\tilde{L}\sigma^2}{2\mu} + \frac{\eta_\theta^2\tilde{L}\sigma^2}{\eta_\alpha\mu} \\
&\quad + \frac{2}{\mu}(1+\frac{\eta_\theta}{\eta_\alpha})\tilde{L}^2\sum_{t'=0}^{T-1}\Big(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\alpha}_{t'} - \alpha_{t'}^{(k)}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\theta}_{t'} - \theta_{t'}^{(k)}\|^2]\Big) \\
&\leq (1-\frac{\eta_\alpha\mu}{2})^T\mathcal{L}_0 + \frac{2\eta_\alpha L_{\tilde{\Phi}}\sigma^2}{\mu} + \frac{5\eta_\alpha\tilde{L}\sigma^2}{2\mu} + \frac{\eta_\theta^2\tilde{L}\sigma^2}{\eta_\alpha\mu} + (1+\frac{\eta_\theta}{\eta_\alpha})\tilde{L}^2\frac{2\tau^2(\eta_\alpha^2+\eta_\theta^2)(4\sigma^2+6\delta^2)}{\mu} .
\end{aligned}
$$

Following Eq. (9) and its following inequality in Yang et al. (2022), we can obtain

$$
\begin{aligned}
\mathcal{L}_0 &\leq \tilde{\Phi}^* - \tilde{\Psi}(\tilde{\theta}) + \frac{1}{2L}\|\nabla_1 f(\tilde{\alpha},\tilde{\theta})\|^2 \leq \tilde{\Phi}^* - \tilde{\Psi}(\tilde{\theta}) + \frac{\epsilon^2}{2L} , \\
&\leq \frac{1}{2\mu}\|\nabla_2 f(\tilde{\alpha},\tilde{\theta})\|^2 + \frac{1}{2L}\|\nabla_1 f(\tilde{\alpha},\tilde{\theta})\|^2 + \frac{\epsilon^2}{2L} \leq \frac{\epsilon^2}{L} + \frac{\epsilon^2}{2L} \leq \frac{2\epsilon^2}{L} .
\end{aligned}
$$

By putting this upper bound into that of $\mathcal{L}_T$, we have

$$
\begin{aligned}
\mathcal{L}_T &\leq (1-\frac{\eta_\alpha\mu}{2})^T\frac{2\epsilon^2}{L} + \frac{2\eta_\alpha L_{\tilde{\Phi}}\sigma^2}{\mu} + \frac{5\eta_\alpha\tilde{L}\sigma^2}{2\mu} + \frac{\eta_\theta^2\tilde{L}\sigma^2}{\eta_\alpha\mu} \\
&\quad + (1+\frac{\eta_\theta}{\eta_\alpha})\tilde{L}^2\frac{2\tau^2(\eta_\alpha^2+\eta_\theta^2)(4\sigma^2+6\delta^2)}{\mu} .
\end{aligned}
$$

Following the argument in the proof of Proposition 2.1 in Yang et al. (2022) and choosing $\eta_\alpha = O(\eta_\theta/\kappa^2)$, one can establish that

$$\mathbb{E}[\|\alpha_T - \alpha^*\|^2] \leq O(\kappa)\mathcal{L}_T \leq O(\kappa\epsilon^2)(1 - \frac{\eta_\alpha \mu}{2})^T + O(\kappa^5\eta_\alpha) + O(\kappa^7\eta_\alpha^2)\ .$$

Finally, when $\eta_\alpha = O(\epsilon^2/\kappa^5)$ and $T = O(\epsilon^{-2}\kappa^6\log(\kappa))$, we obtain

$$\mathbb{E}[\|\alpha_T - \alpha^*\|^2] \leq O(\epsilon^2)\ . \tag{69}$$

This completes the proof. □

Based on Theorem G.1, we combine Theorem 4.8 and Theorem G.1 to obtain an $O(\epsilon)$-stationary guarantee for the original objective. Defining $\Phi_{1/2L}(\tilde{\alpha}) = \min_\beta \Phi(\beta) + L\|\beta - \tilde{\alpha}\|^2$, using the argument of Proposition 2.1 b) in Yang et al. (2022), one obtains

$$\begin{aligned}\|\nabla\Phi(\hat{\alpha})\| &\leq \|\nabla\Phi_{1/2L}(\tilde{\alpha})\| + 2\kappa L\|\hat{\alpha} - \alpha^*(\tilde{\alpha})\| \\ &\leq \underbrace{4\|\nabla_1 f_\rho(\tilde{\alpha},\tilde{\theta})\| + 4\kappa\|\nabla_2 f_\rho(\tilde{\alpha},\tilde{\theta})\|}_{\leq O(\epsilon)\text{ follows from Theorem 4.8}} + \underbrace{2\kappa L\|\hat{\alpha} - \alpha^*(\tilde{\alpha})\|}_{\leq O(\epsilon)\text{ follows from Theorem G.1}}\ .\end{aligned} \tag{70}$$

Here $(\tilde{\alpha},\tilde{\theta}) \triangleq (\tilde{x},\tilde{y},\tilde{z})$ is the output of Algorithm 1, while $\hat{\alpha}$ is obtained by running LocalSGDA initialized at $\tilde{\alpha}$. Then, it is easy to conclude that $\|\nabla\Phi(\hat{\alpha})\| \leq O(\epsilon)$.

Finally, we complete the Remark 4.10. The main algorithm requires $T = O\big(\frac{\kappa^{15/2}}{K\epsilon^5}\big)$ iterations, whereas the translation phase incurs $T = O\big(\epsilon^{-2}\kappa^6\log(\kappa)\big)$. After comparing these two number of iterations, it is easy to conclude that the overall complexity for obtaining an $O(\epsilon)$-stationary solution is $T = O\big(\frac{\kappa^{15/2}}{K\epsilon^5}\big)$.