A Doubly Smoothed Decentralized Stochastic Minimax Optimization Algorithm

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

Chiu C. Tan
Temple University
Philadelphia, PA, USA
cctan@temple.edu

Haibin Ling
Stony Brook University
New York, NY, USA
hling@cs.stonybrook.edu

Hongchang Gao*
Temple University
Philadelphia, PA, USA
hongchang.gao@temple.edu

Abstract

Decentralized stochastic minimax optimization has recently attracted significant attention due to its applications in machine learning. However, existing state-of-the-art methods use learning rates of different scales for the primal and dual variables, making them difficult to tune in practice. To address this problem, this paper proposes a novel doubly smoothed decentralized stochastic minimax algorithm. Specifically, in terms of algorithm design, we update both the primal and dual variables using smoothed gradients and introduce novel approaches to handle the computation and communication of the auxiliary variables introduced by the smoothing technique. On the theoretical side, for nonconvex-PL problems, our convergence analysis reveals that the learning rates for the primal and dual variables are of the same scale. Moreover, the order of the condition number in our convergence rate is improved to $O(\kappa^{3/2})$. To the best of our knowledge, this is the first time it has been improved to such a favorable order. Finally, extensive experimental results validate the effectiveness of our algorithm.

1 Introduction

In this paper, we focus on the following decentralized stochastic minimax optimization problem:

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

where $x \in \mathbb{R}^{d_1}$ is the primal variable, $y \in \mathbb{R}^{d_2}$ is the dual variable, $f^{(k)}(x,y) = \mathbb{E}[f^{(k)}(x,y;\xi^{(k)})]$ is the loss function on the k-th (where $k \in \{1,\cdots,K\}$) worker, and $\xi^{(k)}$ denotes the random sample on the k-th worker. Throughout this paper, it is assumed that $f(x,y)$ is nonconvex in x and satisfies the Polyak-Lojasiewicz (PL) condition in y.

Stochastic minimax optimization has attracted increasing attention in the machine learning community recently because it finds numerous applications, such as generative adversarial networks [8], adversarially robust learning [22], distributionally robust learning [5], imbalanced data classification [36], policy evaluation [41], etc. Moreover, in real-world machine learning applications, the training data is typically distributed on different devices. To take advantage of the distributed data to train

*Corresponding author.

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

Table 1: The communication complexity (i.e., iteration complexity) and computation complexity to achieve the $O(\epsilon)$-stationary solution of different decentralized stochastic minimax algorithms that use variance-reduced gradients. **BS**: batch size. **N-PL**: nonconvex-PL problems. **N-SCV**: nonconvex-strongly-concave problems. **LR Ratio**: the ratio between the learning rate of the primal variable and that of the dual variable. κ: condition number. $1-\lambda$: spectral gap. ϵ: solution accuracy. Note that Smoothed-SAGDA is a *single-machine* algorithm *without using variance-reduced gradients*. DGDA-VR and DREAM depend on the condition number, scaling as κ^2, *at the cost of a large batch size* $O\left(\frac{\kappa}{\epsilon}\right)$, which increases the per-iteration computational cost. Moreover, they *periodically require much larger batch sizes* $O\left(\frac{\kappa^2}{\epsilon^2}\right)$, limiting their practicality. DREAM achieves a better dependence on the spectral gap *at the cost of performing multi-round communication* in each iteration.

Algorithms	Communication	BS	Computation	Class	LR Ratio
Smoothed-SAGDA [34]	$O\left(\frac{\kappa^2}{\epsilon^4}\right)$	O(1)	$O\left(\frac{\kappa^2}{\epsilon^4}\right)$	N-PL	$O(1)$
DM-HSGD [31]	$O\left(\frac{\kappa^3}{(1-\lambda)^2\epsilon^3}\right)$	O(1)	$O\left(\frac{\kappa^3}{(1-\lambda)^2\epsilon^3}\right)$	N-SCV	$O(1/\kappa^2)$
DGDA-VR [42]	$O\left(\frac{\kappa^2}{(1-\lambda)^2\epsilon^2}\right)$	$O\left(\frac{\kappa}{\epsilon}\right)$	$O\left(\frac{\kappa^3}{(1-\lambda)^2\epsilon^3}\right)$	N-SCV	$O(1/\kappa^2)$
DREAM [3]	$O\left(\frac{\kappa^2}{\sqrt{1-\lambda}\epsilon^2}\right)$	$O\left(\frac{\kappa}{\epsilon}\right)$	$O\left(\frac{\kappa^3}{\epsilon^3}\right)$	N-SCV	$O(1/\kappa^2)$
DM-GDA [13]	$O\left(\frac{\kappa^3}{(1-\lambda)^2\epsilon^3}\right)$	O(1)	$O\left(\frac{\kappa^3}{(1-\lambda)^2\epsilon^3}\right)$	N-PL	$O(1/\kappa^2)$
Ours (Corollary 4.3)	$O\left(\frac{\kappa^{3/2}}{(1-\lambda)^2\epsilon^3}\right)$	O(1)	$O\left(\frac{\kappa^{3/2}}{(1-\lambda)^2\epsilon^3}\right)$	N-PL	$O(1)$

the aforementioned machine learning models, decentralized minimax optimization has been actively studied in recent years. For example, [31, 13] proposed decentralized stochastic variance-reduced gradient descent ascent algorithms based on the STORM gradient estimator [4], while [41, 42] proposed to use the SPIDER gradient estimator [6, 23]. Recently, [15] developed a decentralized adaptive minimax algorithm and established its convergence rate for nonconvex-strongly-concave problems.

However, most existing decentralized minimax optimization algorithms suffer from a significant limitation. Specifically, to ensure convergence, the learning rate for the primal variable is set on a different scale than that for the dual variable. For example, [31, 42, 3, 13] prove that the ratio between the learning rate of the primal variable and that of the dual variable has to be $O(1/\kappa^2)$ for nonconvex-strongly-concave (or nonconvex-PL) problems, where $\kappa > 1$ is the condition number. Since κ is an unknown parameter, it is difficult to tune their learning rates to ensure convergence in practice. To address this issue, a recent work [34] in the single-machine setting demonstrates that the smoothing technique proposed by [38] allows primal and dual variables to use learning rates of the same scale, that is, with a ratio of the order of $O(1)$. However, the convergence rate[2] $O(1/\epsilon^4)$ of [34] is inferior to $O(1/\epsilon^3)$ of [31, 13] because it just uses the standard stochastic gradient. Then, a natural question arises:

Can we develop a decentralized smoothed minimax optimization algorithm that achieves a better convergence rate while using the same-scale learning rates for the primal and dual variables?

Actually, there are unique challenges when applying the smoothing technique to decentralized minimax optimization in order to improve the convergence rate, as outlined below.

Challenge-1: How to incorporate the variance reduction technique into the smoothing technique to achieve a faster convergence rate? Existing minimax optimization algorithms with the smoothing technique in a single-machine setting are based on the *deterministic gradient* [38] or the *unbiased stochastic gradient* [34]. Directly extending their smoothing technique to decentralized *stochastic* minimax optimization will lead to a slow convergence rate. For example, [34] can only achieve an $O(1/\epsilon^4)$ convergence rate to achieve the ϵ-accuracy solution for a nonconvex-PL problem, while the existing decentralized minimax optimization algorithm [13] can achieve an $O(1/\epsilon^3)$ convergence rate for the same problem class by using the variance reduction technique. However, because smoothing introduces auxiliary variables, it is unclear how to leverage the variance reduction technique to

[2] In the introduction, we omit other factors in the convergence rate for clarity.

accelerate its convergence rate. For example, it is unclear which components of the smoothed gradient should incorporate variance reduction and how to control the gradient bias to guarantee the fast convergence rate.

Challenge-2: How should the auxiliary variables be updated and communicated, and how does this affect the communication complexity? Standard smoothing methods introduce an auxiliary variable for the primal variable to improve the geometry of the optimization subproblem and convergence rate. In a decentralized setting, however, each worker maintains a local copy of this auxiliary variable, raising the question of how to update and communicate these copies to control their consensus errors. This challenge becomes more involved when *smoothing is applied to both the primal and dual variables*, as in our algorithm, because two auxiliary variables must be coordinated with the original variables across workers. It is therefore unclear whether the resulting updates and communication can preserve favorable dependence on the spectral gap and condition number.

To answer the aforementioned questions, we develop a novel decentralized algorithm based on the smoothing technique: the doubly smoothed decentralized stochastic gradient descent ascent with momentum (Smoothed^2-DSGDAM) algorithm, which brings the following contributions:

- In terms of algorithm design, we apply the smoothing technique to both the primal and dual variables. Importantly, we propose a novel and feasible approach to incorporate the variance reduction technique into the smoothed gradient regarding both the primal and dual variables. More importantly, our algorithm demonstrates how to update and communicate the auxiliary variables introduced by the smoothing technique in the decentralized setting. **As far as we know, this is the first time to show how to handle the auxiliary variables and reduce the gradient variance for the decentralized smoothed minimax algorithm.**
- In terms of convergence analysis, we establish the convergence rate of our algorithm for nonconvex-PL minimax problems. In particular, on the one hand, for a nonconvex-PL minimax problem, the smoothing technique with a variance-reduced gradient can make the convergence rate enjoy a better dependence on the condition number κ , i.e., in the order of $O(\kappa^{3/2})$, which is better than the dependence $O(\kappa^3)$ in existing decentralized non-smoothed minimax algorithms [31, 13] and the dependence $O(\kappa^2)$ in smoothed minimax algorithms [34] in the single-machine setting [3]. **To the best of our knowledge, this is the first time the dependence on the condition number is improved to $O(\kappa^{3/2})$.** On the other hand, our convergence analysis shows that the ratio between the learning rate of the primal variable and that of the dual variable can be improved from $O(1/\kappa^2)$ of [31, 42, 3, 13] to $O(1)$, and the convergence rate can be improved from $O(1/\epsilon^4)$ of [34] to $O(1/\epsilon^3)$. **To the best of our knowledge, this is the first time that a decentralized stochastic minimax optimization algorithm can achieve such a fast convergence rate with the same-scale learning rate.** A detailed comparison between our algorithm and existing algorithms can be found in Table 1.

Finally, the extensive experimental results validate the performance of our proposed algorithm.

2 Related Works

2.1 Stochastic Minimax Optimization

Due to the widespread applications of stochastic minimax optimization in machine learning, numerous stochastic optimization algorithms [20, 21, 14, 26, 10, 33, 34, 2] have been developed recently. In particular, the nonconvex-strongly-concave and nonconvex-PL problems have been extensively studied. For the former, [20] established the convergence rate of the stochastic gradient descent ascent (SGDA) algorithm for nonconvex-strongly-concave problems. Following that, a couple of variance-reduced gradient methods [21, 14, 26, 10] have been developed to accelerate its convergence rate. Specifically, [14, 26] combined the STORM gradient estimator [4] with SGDA and established its convergence rate. [21] investigated the convergence rate when incorporating the SPIDER gradient estimator [6] into SGDA. For the latter, [33] investigated the convergence rate for the alternating stochastic gradient descent ascent (ASGDA) algorithm. [2] studied the convergence rate for the finite-sum minimax problem when combining the SPIDER gradient estimator with ASGDA.

[3] Here, to make a fair comparison, the existing methods considered use a batch size of $O(1)$, rather than large batch sizes.

The smoothing technique for the minimax problem was first studied for nonconvex-concave problems in [38]. Specifically, it established the convergence rate of the full-gradient alternating gradient descent ascent (AGDA) algorithm when incorporating the smoothing technique. Later, [34] applied this technique to nonconvex-PL problems and established its convergence rate for SGDA. Moreover, [39] incorporated the smoothing technique into stochastic multi-level compositional minimax optimization. In fact, due to the efficacy of the smoothing technique, it has been applied to various settings, such as nonconvex-nonconcave problems with the one-sided KŁ condition [48], constrained optimization problems [25], etc, which are beyond the scope of this paper.

2.2 Decentralized Stochastic Minimax Optimization

To facilitate decentralized optimization for minimax problems, a great amount of effort [30, 41, 31, 7, 42, 3, 32, 47] has recently been made. For example, [30] developed a decentralized gradient descent ascent algorithm by using the full gradient for local computation and the gradient tracking technique for communication. [31] proposed a decentralized minimax algorithm based on the STORM gradient estimator and established its convergence rate for the stochastic setting. [41] developed a decentralized minimax algorithm based on the SPIDER gradient estimator and established its convergence rate for the finite-sum setting. Later, its convergence rate for the stochastic setting was established in [42]. Moreover, [7] incorporated the ZeroSARAH gradient estimator into the decentralized minimax algorithm and provided convergence analysis for the finite-sum setting. Recently, [3] studied the convergence rate of the decentralized minimax algorithm when using the PAGE gradient estimator [19]. More recently, [15] introduced the adaptive learning rate to decentralized minimax optimization and established the corresponding convergence rate. Note that all these existing methods restrict their focus to the nonconvex-strongly-concave problem.

Recently, [13] developed a decentralized minimax algorithm for nonconvex-PL problems, where the STORM gradient estimator is used for local updates and the gradient tracking technique is used for communication between workers. To our knowledge, in the distributed setting, the smoothing technique has only been studied for federated centralized learning in [29, 45, 44]. For example, in [29], the smoothing technique is applied only on the central server, so the auxiliary variable behaves as in the single-machine setting and is easy to handle in the convergence analysis. [45] applied smoothing to local updates and introduced periodic communication of auxiliary variables, developing an algorithm that is more robust to the communication period. Subsequently, motivated by the studies of federated compositional minimax algorithms [43, 46], [44] applied the smoothing technique to the local updates on workers for federated stochastic multi-level compositional minimax optimization. All in all, the smoothing technique has not been studied for decentralized minimax optimization and it is unclear how to apply it from the algorithm design perspective and how to handle it from the convergence analysis perspective.

3 Method

3.1 Problem Setup

We introduce the following assumptions with respect to the loss function and communication topology, which have been widely used in the existing literature [34, 31, 13, 41, 42, 3].

Assumption 3.1. (**Smoothness**) For any $k \in \{1, 2, \cdots, K\}$, the loss function on the k-th worker satisfies the mean-squared Lipschitz smoothness, i.e., for any $(x_1, y_1) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ and $(x_2, y_2) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, there exists a constant value $L > 0$ such that $\mathbb{E}[\|\nabla_x f^{(k)}(x_1, y_1; \xi^{(k)}) - \nabla_x f^{(k)}(x_2, y_2; \xi^{(k)})\|^2] \leq L^2(\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2)$ and $\mathbb{E}[\|\nabla_y f^{(k)}(x_1, y_1; \xi^{(k)}) - \nabla_y f^{(k)}(x_2, y_2; \xi^{(k)})\|^2] \leq L^2(\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2)$.

Assumption 3.2. (**PL condition**) For any fixed $x \in \mathbb{R}^{d_1}$, the set of solutions of the optimization problem with respect to y, $\max_{y \in \mathbb{R}^{d_2}} f(x, y)$, is not empty and the optimal value is finite. Furthermore, for any $x \in \mathbb{R}^{d_1}$, there exists a constant value $\mu > 0$ such that $\|\nabla_y f(x, y)\|^2 \geq 2\mu(f(x, y^*) - f(x, y))$, where $y^* = \arg\max_{y \in \mathbb{R}^{d_2}} f(x, y)$.

Assumption 3.3. (**Variance**) For any $k \in \{1, 2, \cdots, K\}$, the stochastic gradients with respect to x and y of the loss function on the k-th worker are unbiased estimators and their variances are upper bounded as: $\mathbb{E}[\|\nabla_x f^{(k)}(x, y; \xi^{(k)}) - \nabla_x f^{(k)}(x, y)\|^2] \leq \sigma^2$ and $\mathbb{E}[\|\nabla_y f^{(k)}(x, y; \xi^{(k)}) - \nabla_y f^{(k)}(x, y)\|^2] \leq \sigma^2$, where $\sigma > 0$ is a constant value.

Assumption 3.4. (**Communication graph**) The element w_{ij} of the adjacency matrix $W \in \mathbb{R}^{K\times K}$ of the communication graph is non-negative, with a positive value indicating that worker-i is connected to worker-j, and a value of zero indicating they are disconnected. Moreover, W is doubly stochastic and symmetric, and its eigenvalues satisfy $|\lambda_K| \leq |\lambda_{K-1}| \leq \cdots \leq |\lambda_2| < |\lambda_1| = 1$.

By denoting $\lambda = |\lambda_2|$, the spectral gap of the adjacency matrix can be represented by $1-\lambda$. Moreover, we use $\mathcal{N}_k$ to denote the neighboring workers of the k-th worker, and use $\kappa = L/\mu$ to represent the condition number. In addition, we use $\bar{a}_t = \frac{1}{K}\sum_{k=1}^{K} a_t^{(k)}$ to denote the average value of any $\{a_t^{(k)}\}_{k=1}^K$ in the t-th iteration.

Algorithm 1 Doubly Smoothed Decentralized Stochastic Gradient Descent Ascent with Momentum (Smoothed^2-DSGDAM)

Input: $\eta > 0$ and $\rho_x, \rho_y, \beta_x, \beta_y, \hat{\beta}_x, \hat{\beta}_y > 0$, with $\rho_x\eta^2, \rho_y\eta^2, \hat{\beta}_x\eta, \hat{\beta}_y\eta < 1$.

Initialization on worker k: $x_0^{(k)} = x_0$, $y_0^{(k)} = y_0$, $\hat{x}_0^{(k)} = \hat{x}_0$, $\hat{y}_0^{(k)} = \hat{y}_0$,
$u_0^{(k)} = \nabla_x F^{(k)}(x_0^{(k)}, y_0^{(k)}; \hat{x}_0^{(k)}, \hat{y}_0^{(k)}; \xi_0^{(k)})$, $v_0^{(k)} = \nabla_y F^{(k)}(x_0^{(k)}, y_0^{(k)}; \hat{x}_0^{(k)}, \hat{y}_0^{(k)}; \xi_0^{(k)})$,
$p_0^{(k)} = u_0^{(k)}, q_0^{(k)} = v_0^{(k)}$.

for $t = 0, \cdots, T-1$, worker k **do**
  Update x: $\tilde{x}_{t+1}^{(k)} = \sum_{j\in\mathcal{N}_k} w_{kj} x_t^{(j)} - \beta_x p_t^{(k)}$, $x_{t+1}^{(k)} = x_t^{(k)} + \eta(\tilde{x}_{t+1}^{(k)} - x_t^{(k)})$,
  Update y: $\tilde{y}_{t+1}^{(k)} = \sum_{j\in\mathcal{N}_k} w_{kj} y_t^{(j)} + \beta_y q_t^{(k)}$, $y_{t+1}^{(k)} = y_t^{(k)} + \eta(\tilde{y}_{t+1}^{(k)} - y_t^{(k)})$,
  Update $\hat{x}$: $\tilde{\hat{x}}_{t+1}^{(k)} = \sum_{j\in\mathcal{N}_k} w_{kj} \hat{x}_t^{(j)} + \hat{\beta}_x(x_{t+1}^{(k)} - \hat{x}_t^{(k)})$, $\hat{x}_{t+1}^{(k)} = \hat{x}_t^{(k)} + \eta(\tilde{\hat{x}}_{t+1}^{(k)} - \hat{x}_t^{(k)})$,
  Update $\hat{y}$: $\tilde{\hat{y}}_{t+1}^{(k)} = \sum_{j\in\mathcal{N}_k} w_{kj} \hat{y}_t^{(j)} + \hat{\beta}_y(y_{t+1}^{(k)} - \hat{y}_t^{(k)})$, $\hat{y}_{t+1}^{(k)} = \hat{y}_t^{(k)} + \eta(\tilde{\hat{y}}_{t+1}^{(k)} - \hat{y}_t^{(k)})$,
  Compute variance-reduced gradients $u_t^{(k)}$ and $v_t^{(k)}$:
  $u_{t+1}^{(k)} = (1-\rho_x\eta^2)(u_t^{(k)} - \nabla_x F^{(k)}(x_t^{(k)}, y_t^{(k)}; \hat{x}_t^{(k)}, \hat{y}_t^{(k)}; \xi_{t+1}^{(k)})) + \nabla_x F^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; \hat{x}_{t+1}^{(k)}, \hat{y}_{t+1}^{(k)}; \xi_{t+1}^{(k)})$,
  $v_{t+1}^{(k)} = (1-\rho_y\eta^2)(v_t^{(k)} - \nabla_y F^{(k)}(x_t^{(k)}, y_t^{(k)}; \hat{x}_t^{(k)}, \hat{y}_t^{(k)}; \xi_{t+1}^{(k)})) + \nabla_y F^{(k)}(x_{t+1}^{(k)}, y_{t+1}^{(k)}; \hat{x}_{t+1}^{(k)}, \hat{y}_{t+1}^{(k)}; \xi_{t+1}^{(k)})$,
  Gradient tracking:
  $p_{t+1}^{(k)} = \sum_{j\in\mathcal{N}_k} w_{kj} p_t^{(j)} + u_{t+1}^{(k)} - u_t^{(k)}$, $q_{t+1}^{(k)} = \sum_{j\in\mathcal{N}_k} w_{kj} q_t^{(j)} + v_{t+1}^{(k)} - v_t^{(k)}$,
end for

3.2 Smoothed^2-DSGDAM

Smoothing introduces quadratic regularization around auxiliary variables to improve the geometry of the update subproblems. We apply this technique to both the primal and dual variables. For fixed auxiliary variables and sufficiently large regularization parameters, the resulting objective is strongly convex in the primal variable and strongly concave in the dual variable, facilitating coordinated updates. Specifically, we define the doubly smoothed loss function $F(x, y; \hat{x}, \hat{y})$ as follows:

$$\frac{1}{K}\sum_{k=1}^{K} \underbrace{f^{(k)}(x,y) + \frac{\gamma_1}{2}\|x-\hat{x}\|^2 - \frac{\gamma_2}{2}\|y-\hat{y}\|^2}_{F^{(k)}(x,y;\hat{x},\hat{y})}, \tag{2}$$

where $\gamma_1 > 0$ and $\gamma_2 > 0$ are hyperparameters, and $\hat{x} \in \mathbb{R}^{d_1}$, $\hat{y} \in \mathbb{R}^{d_2}$ **are the auxiliary variables for the primal and dual variables, respectively**. For fixed $\hat{x}$ and $\hat{y}$, γ_1 and γ_2 are set such that $F(x, y; \hat{x}, \hat{y})$ is strongly convex with respect to x and strongly concave with respect to y. For example, we can set $\gamma_1 = 2L$ and $\gamma_2 = 2L$. Note that most existing works in the single-machine setting, such as [38, 34], apply the smoothing technique to a single variable. Only a recent work [48] uses it for both variables, but it considers the deterministic single-machine setting (see the end of this section for a detailed comparison). Hence, a new algorithm design and convergence analysis are required to address the challenges raised in Eq. (2).

Based on the smoothed loss function in Eq. (2), the k-th worker can compute the stochastic gradient with respect to the primal and dual variables in the t-th iteration as follows:

$$\nabla_x F^{(k)}(x_t^{(k)}, y_t^{(k)}; \hat{x}_t^{(k)}, \hat{y}_t^{(k)}; \xi_t^{(k)}) = \nabla_x f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_t^{(k)}) + \gamma_1(x_t^{(k)} - \hat{x}_t^{(k)}),$$

$$\nabla_y F^{(k)}(x_t^{(k)}, y_t^{(k)}; \hat{x}_t^{(k)}, \hat{y}_t^{(k)}; \xi_t^{(k)}) = \nabla_y f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_t^{(k)}) - \gamma_2(y_t^{(k)} - \hat{y}_t^{(k)}) \,. \tag{3}$$

In terms of the smoothed loss function in Eq. (2) and the stochastic gradients in Eq. (3), we develop a novel doubly smoothed decentralized stochastic gradient descent ascent with momentum (**Smoothed^2-DSGDAM**) algorithm in Algorithm 1. Generally speaking, we apply the variance reduction technique, STORM [4] (see Step 6 of Algorithm 1), to the stochastic gradient on each worker to update the primal and dual variables, and use the gradient tracking technique to conduct communication between different workers. However, there are two unique challenges when designing our Smoothed^2-DSGDAM algorithm: **1) How to apply the variance reduction technique in the presence of the smoothing term? 2) How to update and communicate the auxiliary variables $\hat{x}$ and $\hat{y}$ to guarantee convergence in the decentralized setting?**

As for the first challenge regarding variance reduction, there are actually two ways to apply variance reduction. Specifically, we can apply it to the original stochastic gradient $\nabla_x f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_t^{(k)})$ or to the smoothed gradient $\nabla_x F^{(k)}(x_t^{(k)}, y_t^{(k)}; \hat{x}_t^{(k)}, \hat{y}_t^{(k)}; \xi_t^{(k)})$. However, computing the variance-reduced gradient $u_t^{(k)}$ for the original stochastic gradient $\nabla_x f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_t^{(k)})$ will complicate the convergence analysis, when bounding a critical term $\langle \nabla_x F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t), \bar{x}_{t+1} - \bar{x}_t \rangle$, where $\bar{x}_t$, $\bar{y}_t$, $\bar{\hat{x}}_t$, and $\bar{\hat{y}}_t$ denote the averaged variable across workers.

Specifically, when computing the STORM gradient estimator $u_t^{(k)}$ for the smoothed gradient $\nabla_x F^{(k)}(x_t^{(k)}, y_t^{(k)}; \hat{x}_t^{(k)}, \hat{y}_t^{(k)}; \xi_t^{(k)})$, we can bound it as follows:

$$\begin{aligned}
&\langle \nabla_x F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t), \bar{x}_{t+1} - \bar{x}_t \rangle = -\eta\beta_x \langle \nabla_x F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t), \bar{u}_t \rangle \\
&= -\frac{\eta\beta_x}{2}\|\bar{u}_t\|^2 - \frac{\eta\beta_x}{2}\|\nabla_x F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)\|^2 + \frac{\eta\beta_x}{2}\|\nabla_x F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \bar{u}_t\|^2 \,.
\end{aligned} \tag{4}$$

All three terms in the last step are straightforward to handle.

However, when computing the STORM gradient estimator $u_t^{(k)}$ for the original stochastic gradient $\nabla_x f^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_t^{(k)})$, we have

$$\begin{aligned}
&\langle \nabla_x F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t), \bar{x}_{t+1} - \bar{x}_t \rangle \\
&= -\eta\beta_x \langle \nabla_x F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t), \gamma_1(\bar{x}_t - \bar{\hat{x}}_t) \rangle - \eta\beta_x \langle \nabla_x F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t), \bar{u}_t \rangle \\
&\le \frac{\eta\beta_x}{4}\|\nabla_x F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)\|^2 + 4\eta\beta_x\gamma_1^2\|\bar{x}_t - \bar{\hat{x}}_t\|^2 \\
&\quad - \frac{\eta\beta_x}{2}\|\bar{u}_t\|^2 - \frac{\eta\beta_x}{2}\|\nabla_x F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)\|^2 + \frac{\eta\beta_x}{2}\|\nabla_x F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \bar{u}_t\|^2 \,.
\end{aligned} \tag{5}$$

Here, the last term should be further handled by $\|\nabla_x F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \bar{u}_t\|^2 \le 2\|\nabla_x f(\bar{x}_t, \bar{y}_t) - \bar{u}_t\|^2 + 2\gamma_1^2\|\bar{x}_t - \bar{\hat{x}}_t\|^2$. Then, **it can be seen that this approach introduces an additional term (marked in blue), which can make it more challenging to select hyperparameters to handle $\|\bar{x}_t - \bar{\hat{x}}_t\|^2$.**

In addition to this problem, if STORM is applied to the original stochastic gradient, whenever $\|\bar{x}_{t+1} - \bar{x}_t\|^2$ appears, it should be decomposed into two terms: $\|\bar{u}_t\|^2$ and $\|\bar{x}_t - \bar{\hat{x}}_t\|^2$. In contrast, the smoothed one only needs to replace $\|\bar{x}_{t+1} - \bar{x}_t\|^2$ with $\eta^2\beta_x^2\|\bar{u}_t\|^2$, which is much easier for the downstream proof. All in all, applying the variance reduction technique to the original stochastic gradient could significantly complicate the proof. Therefore, we apply it to the smoothed gradient $\nabla_x F^{(k)}(x_t^{(k)}, y_t^{(k)}; \hat{x}_t^{(k)}, \hat{y}_t^{(k)}; \xi_t^{(k)})$, which is shown in Step 6 of Algorithm 1.

We next consider how to update and communicate the auxiliary variables $\hat{x}$ and $\hat{y}$. A straightforward approach is to update $\hat{x}$ (and $\hat{y}$) locally without communication. However, in convergence analysis, we have to handle the negative term, $-\|\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t\|^2$ (see Lemma C.1), and positive term, $\|\hat{x}_{t+1}^{(k)} - \hat{x}_t^{(k)}\|^2$ (see Lemma D.1), simultaneously. Specifically, we need to convert $\|\hat{x}_{t+1}^{(k)} - \hat{x}_t^{(k)}\|^2$ to $\|\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t\|^2$ based on the consensus error $\|\hat{x}_t^{(k)} - \bar{\hat{x}}_t\|^2$. Without an appropriate communication mechanism, the consensus error may grow over iterations. To address this challenge, we propose the following approach for the update and communication of $\hat{x}$ (and $\hat{y}$):

$$\tilde{\hat{x}}_{t+1}^{(k)} = \sum_{j \in \mathcal{N}_k} w_{kj}\hat{x}_t^{(j)} + \hat{\beta}_x(x_{t+1}^{(k)} - \hat{x}_t^{(k)}) \,, \quad \hat{x}_{t+1}^{(k)} = \hat{x}_t^{(k)} + \eta(\tilde{\hat{x}}_{t+1}^{(k)} - \hat{x}_t^{(k)}) \,, \tag{6}$$

where $\hat{\beta}_x > 0$ and $\eta > 0$ are hyperparameters. The first step in Eq. (6) can be viewed as the update of the communicated variable $\sum_{j\in\mathcal{N}_k} w_{kj}\hat{x}_t^{(j)}$ with the local gradient $x_{t+1}^{(k)} - \hat{x}_t^{(k)}$, and the second step is a convex combination between the intermediate variable $\tilde{\hat{x}}_{t+1}^{(k)}$ and the local variable $\hat{x}_t^{(k)}$. With such an update and communication strategy, we can bound the consensus error regarding the auxiliary variable as shown in our Lemma D.3, where the coefficient $1 - \frac{\eta(1-\lambda^2)}{4}$ is important to shrink the consensus error.

These design choices also distinguish our method from the closest related approach, DS-GDA [48], which smooths both the primal and dual variables in the deterministic single-machine setting. The differences are threefold. First, DS-GDA uses full gradients, whereas we use stochastic gradients, so its deterministic convergence analysis does not directly apply. Even when $K = 1$, our analysis must account for both stochastic gradient noise and smoothing bias. Applying STORM to the smoothed gradients requires a new theoretical analysis of the coupled updates of the gradient estimators and the auxiliary variables $\hat{x}$ and $\hat{y}$. Second, the decentralized setting requires communicating these auxiliary variables across workers and controlling their consensus errors. Third, DS-GDA assumes that the domains of both variables have finite diameters and the loss function satisfies the Kurdyka-Łojasiewicz (KŁ) property with respect to y, whereas our analysis does not require these assumptions. In summary, the smoothing technique introduces new challenges in the decentralized stochastic setting, which motivate our novel algorithm design.

4 Convergence Analysis

Before presenting the convergence rate of our algorithm, we introduce the following stationary measures, which were introduced in [34].

Definition 4.1. A solution (x, y) is termed the (ϵ_1, ϵ_2)-stationary solution if $\|\nabla_x f(x, y)\| \leq \epsilon_1$ and $\|\nabla_y f(x, y)\| \leq \epsilon_2$. A solution x is termed the ϵ-stationary solution if $\|\nabla\Phi(x)\| \leq \epsilon$, where $\Phi(x) = f(x, y^*(x))$ and $y^*(x) = \arg\max_{y\in\mathbb{R}^{d_2}} f(x, y)$.

Based on the assumptions in Section 3, we establish the convergence rate of our Algorithm 1.

Theorem 4.2. *Given Assumptions 3.1-3.4, when $\rho_x > 0$, $\rho_y > 0$, $\gamma_1 = O(L)$, $\gamma_2 = O(L)$, the condition about η and β_x in Eq. (49), and those about β_y, $\hat{\beta}_x$, $\hat{\beta}_y$ in Eq. (26) hold, Algorithm 1 has the following convergence upper bound:*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}[\|\nabla_x f(\bar{x}_t, \bar{y}_t)\|^2] + \kappa\mathbb{E}[\|\nabla_y f(\bar{x}_t, \bar{y}_t)\|^2]\right) \leq O(\kappa\rho_x^2\eta^4\sigma^2) + O(\kappa\rho_y^2\eta^4\sigma^2) + O\left(\frac{\kappa\mathcal{P}_0}{\beta_x\eta T}\right) \\
&\quad + O\left(\frac{\kappa}{T}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_x f^{(k)}(x_0, y_0)\|^2]\right) + O\left(\frac{\kappa}{T}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_y f^{(k)}(x_0, y_0)\|^2]\right) \\
&\quad + O\left(\frac{\kappa\sigma^2}{\rho_x\eta^2 TB}\right) + O\left(\frac{\kappa\sigma^2}{\rho_y\eta^2 TB}\right) + O\left(\frac{\kappa\sigma^2}{T}\right) + O\left(\frac{\kappa\rho_x\eta^2\sigma^2}{K}\right) + O\left(\frac{\kappa\rho_y\eta^2\sigma^2}{K}\right), \qquad (7)
\end{aligned}$$

where $\mathcal{P}_0 = F(x_0, y_0; \hat{x}_0, \hat{y}_0) - 2F_d(y_0; \hat{x}_0, \hat{y}_0) + 2q(\hat{x}_0)$, whose definitions can be found in Eq. (14).

Corollary 4.3. *Given Assumptions 3.1-3.4, by setting $\beta_x = O((1-\lambda)^2)$, $\beta_y = O((1-\lambda)^2)$, $\hat{\beta}_x = O\left(\frac{(1-\lambda)^2}{\kappa}\right)$, $\hat{\beta}_y = O((1-\lambda)^2)$, $\eta = O\left(\frac{K\epsilon}{\kappa^{1/2}}\right)$, $\rho_x = O\left(\frac{1}{K}\right)$, $\rho_y = O\left(\frac{1}{K}\right)$, $B = O\left(\frac{\kappa^{1/2}}{\epsilon}\right)$, $T = O\left(\frac{\kappa^{3/2}}{K(1-\lambda)^2\epsilon^3}\right)$, Algorithm 1 can achieve the $O(\epsilon, \epsilon/\sqrt{\kappa})$-stationary solution, where $\epsilon > 0$ denotes the solution accuracy, and B is the batch size in the initial iteration.*

Remark 4.4. The actual learning rate of the primal variable is $\beta_x\eta = O\left(\frac{K(1-\lambda)^2\epsilon}{\kappa^{1/2}}\right)$, and that of the dual variable is $\beta_y\eta = O\left(\frac{K(1-\lambda)^2\epsilon}{\kappa^{1/2}}\right)$. Obviously, they are on the same scale in terms of condition number κ, solution accuracy ϵ, and spectral gap $1-\lambda$. In addition, the constant batch size based methods, including DM-HSGD [31] and DM-GDA [13], use the learning rate for the primal variable in the order of $O\left(\frac{K(1-\lambda)^2\epsilon}{\kappa^3}\right)$ and that for the dual variable is $O\left(\frac{K(1-\lambda)^2\epsilon}{\kappa^1}\right)$. Apparently, our algorithm can allow a larger learning rate. Moreover, when the number of workers $K = 1$, the spectral gap becomes $1 - \lambda = 1$. Our learning rates $O\left(\frac{\epsilon}{\kappa^{1/2}}\right)$ are larger than $O(\frac{\epsilon^2}{\kappa^1})$ of the single-machine method, Smoothed-SAGDA [34].

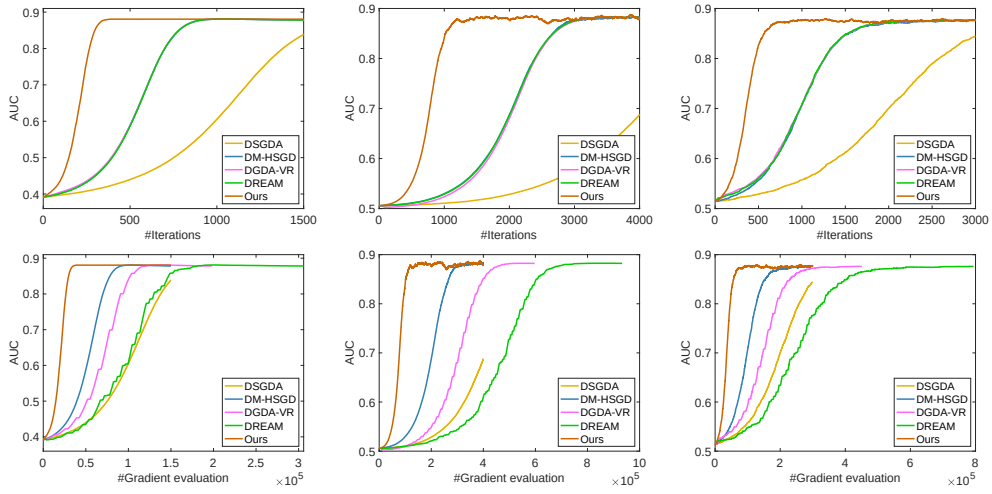
The primal–dual learning rate ratio is critical because the loss function often exhibits different properties for the two variables. In particular, when the loss is nonconvex in the primal variable but satisfies the PL condition in the dual variable, the primal update is more challenging and typically requires a smaller learning rate (see Table 1) to ensure stability. Consequently, existing decentralized minimax methods adopt different step sizes for primal and dual variables (e.g., $O(1/\kappa^2)$), which depend on the condition number κ and are difficult to tune, as improper balancing can lead to divergence. In contrast, with smoothing, the loss becomes strongly convex in the primal variable and strongly concave in the dual variable, yielding a well-behaved landscape that allows both variables to share the same learning rate scale, thereby simplifying hyperparameter tuning.

Remark 4.5. To compare the convergence rate of our algorithm in Corollary 4.3 with existing algorithms in Table 1, we need to translate the $O(\epsilon, \epsilon/\sqrt{\kappa})$-stationary solution to the $O(\epsilon)$-stationary solution, which is a standard technique in smoothed minimax optimization (e.g., Proposition 2.1 of [29]). In particular, we can use the decentralized stochastic gradient descent ascent method in Algorithm 2 in Appendix F to solve the optimization problem: $\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathbb{R}^{d_2}} \hat{f}(x,y) \triangleq f(x,y) + L\|x-\tilde{x}\|^2$, where $\tilde{x}$ is the output of our Algorithm 1. Since this problem satisfies *strong convexity in x and PL condition in y*, the iteration complexity for the translation is in the order of $\tilde{O}(\frac{1}{\epsilon^2})$ as shown in Theorem F.2 in Appendix F, which is dominated by $T = O\left(\frac{\kappa^{3/2}}{K(1-\lambda)^2\epsilon^3}\right)$. Therefore, the iteration complexity to find the $O(\epsilon)$-stationary solution is still $T = O\left(\frac{\kappa^{3/2}}{K(1-\lambda)^2\epsilon^3}\right)$.

The proof structure and all technical details are provided in Appendix B-F.

5 Experiments

We evaluate Algorithm 1 in terms of convergence efficiency, hyperparameter sensitivity, and robustness to data heterogeneity. The main text presents AUC maximization experiments, with the objective defined in Appendix A.1. Appendix A provides additional evaluations on fair classification and WGAN-GP training, together with ablation studies and measurements of runtime and communication overhead.



(a) a9a (b) ijcnn1 (c) w8a

Figure 1: Test AUC vs. Iterations and Gradient Evaluations (Random Graph).

5.1 AUC maximization

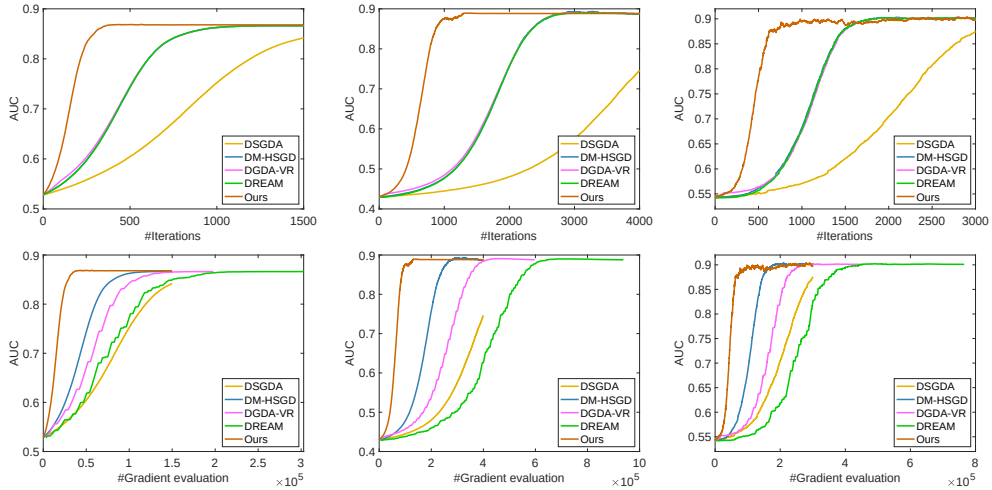
Experimental Settings We employ three benchmark datasets: a9a, w8a, and ijcnn1 from the LIBSVM repository[4]. In our experiments, 80% of samples are used as the training set, while the remaining 20% are used for testing. The training samples are randomly distributed across ten workers, where $K = 10$. To evaluate the performance of our algorithm, we compare it with the state-of-the-art decentralized optimization algorithms: DSGDA [30], DM-HSGD[5] [31], DGDA-VR [42], and

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

[5] Note that DM-GDA is the same as DM-HSGD; they differ only in their convergence analysis under different assumptions.

DREAM [3]. Notably, for DSGDA, we use the stochastic gradient descent ascent instead of the full gradient as described in their paper. For DM-HSGD, the STORM gradient estimator is employed. DGDA-VR leverages the SPIDER gradient estimator in the stochastic setting, while DREAM utilizes the PAGE estimator.

Specifically, we consider two types of communication networks: (1) an Erdős-Rényi random graph with an edge probability of 0.5, resulting in a spectral gap on the order of $O(1)$, and (2) a ring communication network in which each worker communicates with its two neighbors, leading to a spectral gap on the order of $O(1/K^2)$ [35]. Accordingly, we assume a spectral gap of 0.5 for the random graph and set it to 0.01 for the ring network. Throughout all experiments, we fix the solution accuracy ϵ at 0.01 and use a batch size b of 100. For the baseline methods, the learning rates are chosen according to their theoretical results and are carefully tuned. Specifically, in DSGDA, DM-HSGD, and DGDA-VR, the learning rate of the dual variable is scaled by $1/\kappa$, while that of the primal variable is scaled by $1/\kappa^3$. For DREAM, the scaling is 1 for the dual variable and $1/\kappa^2$ for the primal variable. In contrast, our method adopts a unified scaling of $1/\kappa^{1/2}$ for both primal and dual variables. In our experiments, we set $\kappa = 1.5$. Concretely, for the a9a and ijcnn1 datasets, this corresponds to an effective step size of 0.01, while for the w8a dataset, the effective step size is 0.05. For our method, we further set β_x, β_y, $\hat{\beta}_x$, and $\hat{\beta}_y$ to 0.1 (for a9a and ijcnn1) or 0.5 (for w8a), with $\eta = 0.1$, such that their product matches the corresponding effective step sizes. Additionally, in our method, γ_1 and γ_2 are assigned a value of 0.01. For DM-HSGD, the coefficient of the STORM estimator is set to 0.01. Additionally, DGDA-VR computes the full gradient every 100 iterations, while for DREAM, the probability of the PAGE estimator is set to $\frac{\sqrt{b}}{b\sqrt{K}}$.

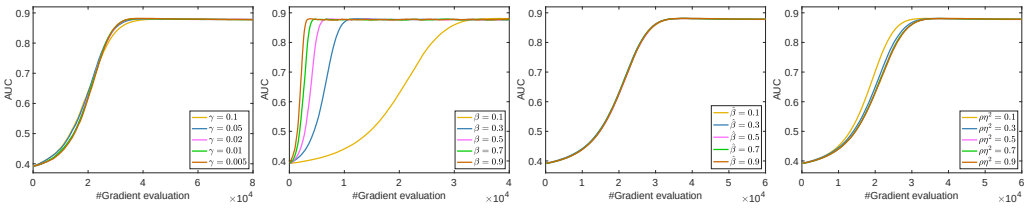


(a) a9a (b) ijcnn1 (c) w8a

Figure 2: Test AUC vs. Iterations and Gradient Evaluations (Ring Graph).

Experimental Results For the random communication graph, we present test AUC versus the number of iterations and gradient evaluations in Figure 1. Our algorithm converges significantly faster than all baselines in terms of the number of iterations, demonstrating its superior efficiency. Furthermore, Figure 1 also indicates that our method converges more quickly when measured by the number of gradient evaluations, highlighting its lower sample complexity. Notably, DGDA-VR and DREAM incur significantly higher computational cost due to periodic full-gradient computation. These results underscore the efficacy of our algorithm in optimizing performance while maintaining computational efficiency. For the ring communication graph, we also present test AUC versus the number of iterations and gradient evaluations in Figure 2. Our method continues to exhibit faster convergence compared to the baseline methods, further validating its effectiveness.

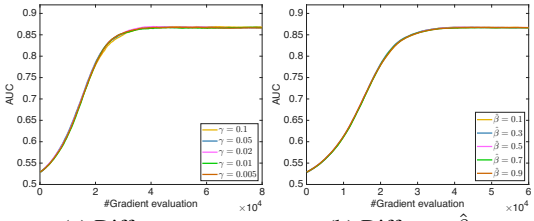
Next, we evaluate the performance of our method under different values of γ, β, $\hat{\beta}$, and $\rho\eta^2$ in Figure 3, where we set $\gamma_1 = \gamma_2 = \gamma$, $\beta_x = \beta_y = \beta$, $\hat{\beta}_x = \hat{\beta}_y = \hat{\beta}$, and $\rho_x = \rho_y = \rho$. Given that our method uses an $O(1)$ batch size, we report performance in terms of gradient evaluations; the iteration curves show the same behavior. Our results indicate that the method is robust to all hyperparameters except β, which primarily acts as a scaling factor for the learning rate. Thus, in practice, we fix



(a) Different γ (b) Different β (c) Different $\hat{\beta}$ (d) Different $\rho\eta^2$

Figure 3: Test AUC under hyperparameters (Random Graph, a9a).

β and tune the learning rate η while keeping the STORM coefficient $\rho\eta^2$ fixed by adjusting ρ accordingly. This allows us to vary the learning rate without changing the estimator coefficient. Compared to existing decentralized minimax baselines, which typically involve multiple hyperparameters (e.g., separate step sizes for primal and dual variables and momentum parameters), our method introduces only two additional hyperparameters, γ and $\hat{\beta}$. Importantly, both are robust across a wide range of values and therefore do not require careful tuning. Consequently, our method does not introduce a fundamentally higher tuning burden. Figure 4 further shows that our method is also robust to γ and $\hat{\beta}$ under the ring graph.



(a) Different γ (b) Different $\hat{\beta}$

Figure 4: Test AUC under hyperparameters (Ring Graph, a9a).

We additionally provide the statistical results of the AUC maximization experiment on a9a with three independent runs under both IID and non-IID settings. In the non-IID setting, different workers are assigned different class imbalance ratios ranging from 0.05 to 0.22, which leads to heterogeneous local data distributions. Table 2 reports the mean, standard deviation, and 95% confidence interval of the test AUC. Our method achieves the best performance in both settings, and its confidence intervals do not overlap with those of the baselines, indicating that the improvement is statistically significant. Moreover, the improvement is even larger in the non-IID setting, demonstrating the robustness of our method to data heterogeneity.

Table 2: Test AUC over multiple independent runs: mean, standard deviation (Std), and 95% confidence interval (CI) (Ring Graph, a9a).

	IID			Non-IID		
Method	Mean	Std	95% CI	Mean	Std	95% CI
DSGDA	0.861	0.004	[0.852, 0.871]	0.787	0.019	[0.740, 0.834]
DM-HSGD	0.862	0.003	[0.855, 0.869]	0.810	0.014	[0.775, 0.844]
DGDA-VR	0.862	0.005	[0.850, 0.873]	0.807	0.017	[0.765, 0.849]
DREAM	0.877	**0.0004**	[0.876, 0.878]	0.853	0.002	[0.847, 0.859]
Ours	**0.891**	0.002	**[0.885, 0.896]**	**0.880**	**0.0017**	**[0.876, 0.885]**

6 Conclusion

In this paper, we developed a doubly smoothed decentralized stochastic minimax algorithm. Our algorithm applies STORM variance reduction to the smoothed gradients and introduces update and communication strategies for the auxiliary variables in the decentralized setting. For nonconvex-PL problems, our convergence analysis permits primal and dual learning rates of the same order and establishes an $O(\kappa^{3/2})$ dependence on the condition number, improving upon existing decentralized minimax methods. Finally, extensive experimental results demonstrate the effectiveness of our algorithm and its robustness across the tested hyperparameter ranges.

Acknowledgements. We thank the anonymous reviewers for their constructive comments. X. Zhang and H. Gao were partially supported by U.S. NSF CAREER 2339545 and NSF IIS 2416607. H. Ling contributed to the work while was affiliated with Stony Brook University.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. Pmlr, 2017.

[2] L. Chen, B. Yao, and L. Luo. Faster stochastic algorithms for minimax optimization under polyak-{\L} ojasiewicz condition. In *Advances in Neural Information Processing Systems*, 2022.

[3] L. Chen, H. Ye, and L. Luo. An efficient stochastic algorithm for decentralized nonconvex-strongly-concave minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1990–1998. PMLR, 2024.

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

[5] J. C. Duchi, P. W. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.

[6] C. Fang, C. J. Li, Z. Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

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

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[9] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

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

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

[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

[14] F. Huang, S. Gao, J. Pei, and H. Huang. Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. *Journal of Machine Learning Research*, 23(36):1–70, 2022.

[15] Y. Huang, X. Li, Y. Shen, N. He, and J. Xu. Achieving near-optimal convergence for distributed minimax optimization with adaptive stepsizes. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[16] W. Jiang, M. Xu, W. Yang, Y. Wang, Z. Li, and L. Zhang. Convergence analysis of the lion optimizer in centralized and distributed settings. In *Forty-third International Conference on Machine Learning*, 2026.

[17] W. Jiang, S. Yang, W. Yang, and L. Zhang. Efficient sign-based optimization: Accelerating convergence via variance reduction. *Advances in Neural Information Processing Systems*, 37:33891–33932, 2024.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[19] Z. Li, H. Bao, X. Zhang, and P. Richtárik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International conference on machine learning*, pages 6286–6295. PMLR, 2021.

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

[21] L. Luo, H. Ye, Z. Huang, and T. Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in neural information processing systems*, 33:20566–20577, 2020.

[22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[23] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR, 2017.

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

[25] W. Pu, K. Sun, and J. Zhang. Smoothed proximal lagrangian method for nonlinear constrained programs. *arXiv preprint arXiv:2408.15047*, 2024.

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

[27] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

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

[29] W. Shen, M. Huang, J. Zhang, and C. Shen. Stochastic smoothed gradient descent ascent for federated minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3988–3996. PMLR, 2024.

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

[31] W. Xian, F. Huang, Y. Zhang, and H. Huang. A faster decentralized algorithm for nonconvex minimax problems. *Advances in Neural Information Processing Systems*, 34, 2021.

[32] Y. Xu. Decentralized gradient descent maximization method for composite nonconvex strongly-concave minimax problems. *SIAM Journal on Optimization*, 34(1):1006–1044, 2024.

[33] J. Yang, N. Kiyavash, and N. He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2002.09621*, 2020.

[34] J. Yang, A. Orvieto, A. Lucchi, and N. He. Faster single-loop algorithms for minimax optimization without strong concavity. In *International Conference on Artificial Intelligence and Statistics*, pages 5485–5517. PMLR, 2022.

[35] B. Ying, K. Yuan, Y. Chen, H. Hu, P. Pan, and W. Yin. Exponential graph is provably efficient for decentralized deep training. *Advances in Neural Information Processing Systems*, 34:13975–13987, 2021.

[36] Y. Ying, L. Wen, and S. Lyu. Stochastic online auc maximization. *Advances in neural information processing systems*, 29:451–459, 2016.

[37] P. Yu, G. Li, and T. K. Pong. Kurdyka–łojasiewicz exponent via inf-projection. *Foundations of Computational Mathematics*, 22(4):1171–1217, 2022.

[38] J. Zhang, P. Xiao, R. Sun, and Z. Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in neural information processing systems*, 33:7377–7389, 2020.

[39] X. Zhang and H. Gao. On the convergence of stochastic smoothed multi-level compositional gradient descent ascent. *Advances in Neural Information Processing Systems*, 38:161672–161724, 2025.

[40] X. Zhang and H. Gao. Sharpness-aware optimization through variance suppression on deep auc maximization. In *2025 IEEE International Conference on Data Mining (ICDM)*, pages 953–962. IEEE, 2025.

[41] X. Zhang, Z. Liu, J. Liu, Z. Zhu, and S. Lu. Taming communication and sample complexities in decentralized policy evaluation for cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:18825–18838, 2021.

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

[43] X. Zhang, A. Payani, M. Lee, R. Souvenir, and H. Gao. A federated stochastic multi-level compositional minimax algorithm for deep AUC maximization. In *Proceedings of the 41st International Conference on Machine Learning*, pages 59601–59640, 2024.

[44] X. Zhang, R. Souvenir, and H. Gao. LS^{2}MC-GDA: A smoothed algorithm for federated stochastic multi-level compositional minimax optimization. In *Forty-third International Conference on Machine Learning*, 2026.

[45] X. Zhang, L. Zhan, and H. Gao. Faster federated smoothed minimax optimization. In *2026 IEEE International Conference on Data Mining (ICDM)*, 2026. in press.

[46] X. Zhang, Y. Zhang, T. Yang, R. Souvenir, and H. Gao. Federated compositional deep auc maximization. *Advances in Neural Information Processing Systems*, 36:9648–9660, 2023.

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

[48] T. Zheng, L. Zhu, A. M.-C. So, J. Blanchet, and J. 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 AUC Maximization

Specifically, we focus on the AUC maximization problem [36, 40] for the binary classification task, which is formulated as the following minimax optimization problem (Note that we have included the smoothed term $\gamma_1/2\|x-\hat{x}\|^2$, $\gamma_2/2\|y-\hat{y}\|^2$):

$$\begin{aligned}\min_{x,\tilde{x}_1,\tilde{x}_2}\max_{y}\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}\Big((1-p)(x^Ta_i^{(k)}-\tilde{x}_1)^2\mathbb{I}_{[b_i^{(k)}=1]}+p(x^Ta_i^{(k)}-\tilde{x}_2)^2\mathbb{I}_{[b_i^{(k)}=-1]}\\ +2(1+y)\big(px^Ta_i^{(k)}\mathbb{I}_{[b_i^{(k)}=-1]}-(1-p)x^Ta_i^{(k)}\mathbb{I}_{[b_i^{(k)}=1]}\big)-p(1-p)y^2\\ +\rho\sum_{j=1}^{d}\frac{x_j^2}{1+x_j^2}+\frac{\gamma_1}{2}\|x-\hat{x}\|^2-\frac{\gamma_2}{2}\|y-\hat{y}\|^2\Big),\end{aligned}\tag{8}$$

where $x\in\mathbb{R}^d$ is the classifier's parameter, $\tilde{x}_1\in\mathbb{R}$, $\tilde{x}_2\in\mathbb{R}$, $y\in\mathbb{R}$ are the parameters to compute the AUC loss, $\hat{x}$ and $\hat{y}$ are the auxiliary variables. $(a_i^{(k)},b_i^{(k)})$ is the i-th sample's feature and label on the k-th worker, p is the prior probability of positive class, $\mathbb{I}$ is an indicator function, ρ is a hyperparameter for the regularization term, and $\gamma_1>0$, $\gamma_2>0$ are hyperparameters for the auxiliary variable. In our experiments, we set ρ to 0.001. Notably, this optimization problem is a nonconvex-PL optimization problem, which can be efficiently solved using our proposed Algorithm 1.

In our paper, the AUC maximization task is evaluated using the standard AUROC metric. In particular, testing AUC is a standard evaluation metric in AUC maximization [36], as it directly reflects the model's ranking performance on the target task and aligns with the optimization objective.

Although the true condition number $\kappa=L/\mu$ is determined by the underlying problem, it is useful to examine how its value in the theoretically motivated learning rate formulas affects empirical performance. We therefore evaluate AUC maximization on a9a using $\kappa=1.05$, 1.5, and 2.0 in these formulas. Larger values result in smaller learning rates for the methods whose step sizes scale inversely with κ. As shown in Table 3, our method maintains similar test AUC across the tested values, whereas the baselines exhibit larger performance drops, consistent with the milder dependence of our learning rates on κ.

Table 3: Test AUC under different condition numbers κ (Ring Graph, a9a).

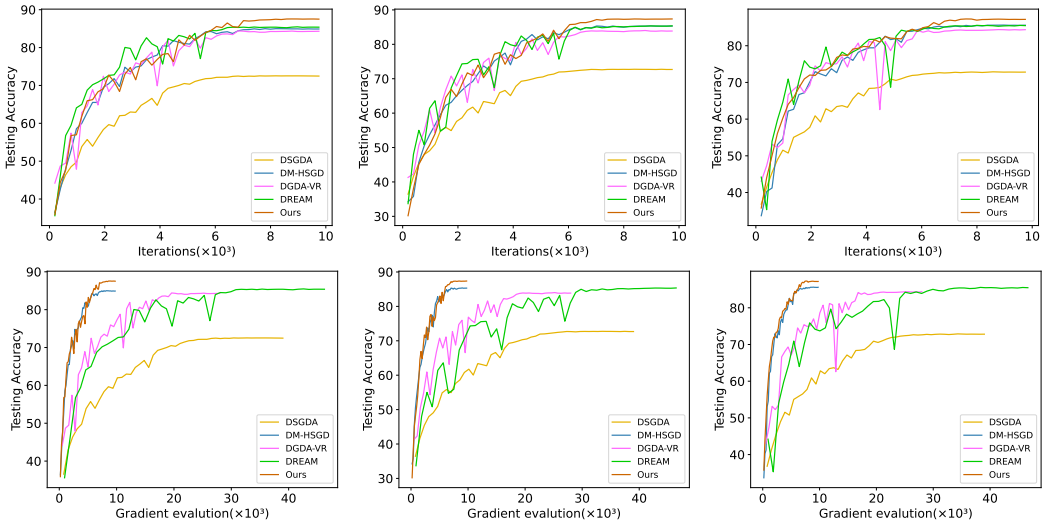
Method	$\kappa=1.05$	$\kappa=1.5$	$\kappa=2.0$
DSGDA	0.889	0.861	0.736
DM-HSGD	0.892	0.862	0.733
DGDA-VR	0.892	0.862	0.740
DREAM	0.892	0.877	0.853
Ours	**0.893**	**0.891**	**0.891**

A.2 Fair Classification

We consider the following nonconvex-PL minimax optimization problem [24]:

$$\min_{x}\max_{y\in\mathcal{Y}}\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}y_c\mathcal{L}_c^{(k)}(x)-\frac{\lambda}{2}\|y\|^2\quad s.t.\quad \mathcal{Y}=\{y\in\mathbb{R}^C|y_c\geq 0,\sum_{c=1}^{C}y_c=1\}\,.\tag{9}$$

This task serves as a standard benchmark for reweighting classes to improve worst-class performance and has been widely evaluated in federated learning algorithms [28]. The simplex constraint on y is enforced by projection and normalization operations in our implementation. We conduct the evaluation in a decentralized setting with eight workers on CIFAR-10 using ResNet-18 [11]. In this setup, $\mathcal{L}_c^{(k)}$ represents the cross-entropy loss functions corresponding to class c on worker k for the $C=10$ classes, and x denotes the model parameters of ResNet-18. We consider three types of communication networks: a random graph, a ring graph, and a torus graph. The learning rate is tuned via grid search and fixed at 0.1, the per-worker batch size is set to 32, and all other hyperparameters and baseline settings remain consistent with the earlier experiments.

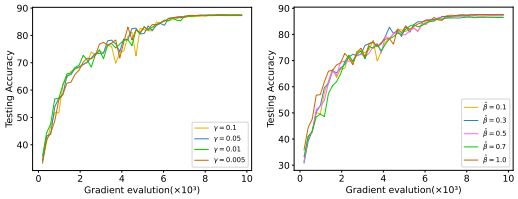


(a) Random Graph (b) Ring Graph (c) Torus Graph

Figure 5: Test Accuracy vs. Iterations and Gradient Evaluations on CIFAR-10.

Figure 5 reports the test accuracy versus the number of iterations and gradient evaluations. Our algorithm achieves the best overall test accuracy among all baselines and converges faster in terms of gradient evaluations. Although DREAM appears slightly faster in early iterations, its periodic large-batch updates substantially increase the total number of gradient computations, leading to a higher overall computational cost. Overall, our algorithm achieves both superior accuracy and more efficient convergence. We also report the overall, best-class, and worst-class accuracy in Table 4, where our method consistently outperforms the baselines across all three metrics.

In addition, we evaluate the sensitivity of our method to the hyperparameters γ and $\hat{\beta}$, which are the only new hyperparameters introduced compared with existing baselines. As shown in Figure 6, our method remains robust even under this more challenging task, further demonstrating that it does not require complicated hyperparameter tuning.



(a) Different γ (b) Different $\hat{\beta}$

Figure 6: Test Accuracy under hyperparameters (Random Graph, CIFAR-10).

Table 4: Overall, best-class, and worst-class test accuracy (%) on CIFAR-10, Ring Graph.

Method	Overall	Best-class	Worst-class
DSGDA	72.45	87.2	43.62
DM-HSGD	85.51	94.8	64.50
DGDA-VR	83.88	93.8	61.39
DREAM	84.98	94.6	61.81
Ours	**87.13**	**95.7**	**66.73**

Runtime and communication cost. We further evaluate the computational and communication overhead on the fair classification task in terms of wall-clock time (per epoch and per iteration), communication bandwidth, latency, and peak memory, as summarized in Table 5. DGDA-VR and DREAM incur higher runtime and higher peak memory due to their periodic large-batch gradient computations, which is consistent with our theoretical analysis. In contrast, our method is not the most expensive under any of these metrics. Overall, it achieves the best optimization performance while maintaining moderate communication and computational overhead.

Table 5: Runtime and communication cost on the fair classification task (CIFAR-10, Ring Graph). ↑/↓: higher/lower is better.

Method	Time/epoch (s) ↓	Time/iter (ms) ↓	Bandwidth (MB/s) ↑	Latency (ms) ↓	Peak memory (GB) ↓
DSGDA	36.4	187	1885	47.31	2.40
DM-HSGD	37.8	194	1887	47.41	2.66
DGDA-VR	52.5	269	1703	52.48	6.01
DREAM	68.4	351	1578	75.86	6.01
Ours	48.7	250	1915	46.67	2.70

Ablation study. To evaluate the contribution of each component of our method, we compare it with several ablation variants on the fair classification task: no smoothing, primal-only smoothing, dual-only smoothing, no gradient tracking, and no auxiliary-variable communication. Note that the no-smoothing variant is DM-HSGD, which employs the STORM estimator but removes the smoothing technique for both the primal and dual variables. Table 6 reports the test accuracy averaged over the last five epochs. Our method achieves the best performance among all variants, indicating that double smoothing, auxiliary-variable communication, and gradient tracking are all necessary for strong practical performance. In particular, smoothing only one of the two variables is inferior to smoothing both.

Table 6: Ablation study on the fair classification task (CIFAR-10, Ring Graph). Test accuracy (%) is averaged over the last five epochs.

Variant	Test accuracy (%)
No smoothing (DM-HSGD)	85.31
Primal-only smoothing	86.56
Dual-only smoothing	84.90
No gradient tracking	86.05
No auxiliary-variable communication	86.55
Ours	**87.34**

A.3 Generative Adversarial Networks

Additionally, we consider training a Wasserstein GAN with gradient penalty (WGAN-GP) [1, 9], which is formulated as the following minimax optimization problem:

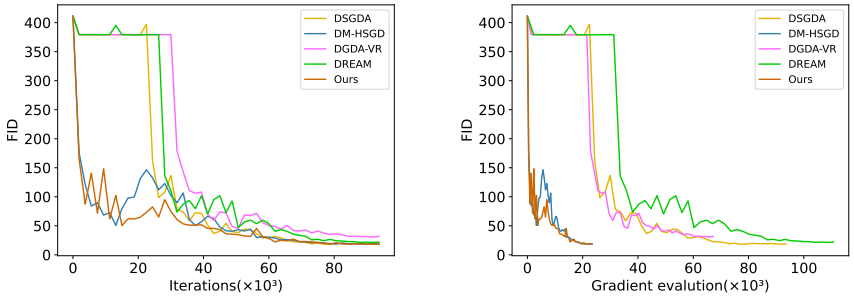
$$\begin{aligned}
&\min_{\theta}\max_{\phi}\ \frac{1}{K}\sum_{k=1}^{K} f^{(k)}(\theta,\phi)\ , \quad \text{where} \\
&f^{(k)}(\theta,\phi) = \mathbb{E}_{a\sim\mathcal{D}^{(k)}}\big[D_\phi(a)\big] - \mathbb{E}_{z\sim\mathcal{N}(0,I)}\big[D_\phi(G_\theta(z))\big] \\
&\qquad\qquad - \lambda\,\mathbb{E}_{\tilde{a}\sim\tilde{\mathcal{D}}^{(k)}}\big[(\|\nabla_{\tilde{a}} D_\phi(\tilde{a})\|_2 - 1)^2\big] - \frac{\rho}{2}\|\phi\|^2\ ,
\end{aligned} \tag{10}$$

where G_θ is the generator with the primal variable θ, D_ϕ is the critic with the dual variable ϕ, $\mathcal{D}^{(k)}$ is the local data distribution on the k-th worker, z is a latent noise vector, $\tilde{\mathcal{D}}^{(k)}$ denotes the distribution of points sampled uniformly along straight lines between real samples from $\mathcal{D}^{(k)}$ and generated samples $G_\theta(z)$, and $f^{(k)}$ is the local objective on the k-th worker. Each worker maintains local copies of θ and ϕ, which reach consensus through communication with neighbors. λ is the weight of the gradient penalty, and ρ is the coefficient of the ℓ_2 regularizer on the critic, which makes the inner maximization better conditioned. We conduct the evaluation on MNIST [18] in a decentralized setting with four workers connected by a ring graph, where the training set is evenly partitioned across workers. The generator adopts the DCGAN architecture [27] with a 100-dimensional latent input, and the critic is a four-layer convolutional network with instance normalization. We set $\lambda = 10$ and $\rho = 0.1$, the per-worker batch size to 32, the learning rate to 0.02 for all methods, and $\gamma = 0.01$ for our method, and train for 50 epochs. Generation quality is measured by the Fréchet Inception Distance (FID) [12], computed with Inception-v3 features between 2,000 generated samples and 2,000 MNIST test images after every epoch.

Table 7: Final test FID on MNIST (lower is better).

Method	DSGDA	DM-HSGD	DGDA-VR	DREAM	Ours
FID ↓	18.75	18.51	31.36	22.11	**18.18**

Figure 7 reports the FID versus the number of iterations and gradient evaluations. Table 7 reports the final test FID, where our method achieves the lowest FID. Moreover, DREAM and DGDA-VR require periodic large-batch gradient computations, which further increase their total gradient evaluations, while our method attains a better FID with a smaller computational cost.



(a) FID vs. iterations (b) FID vs. gradient evaluations

Figure 7: FID vs. iterations and gradient evaluations on MNIST, Ring Graph.

B The Structure of the Proof for Theorem 4.2

To make our proof easy to follow, we provide an overview diagram in Figure 8.

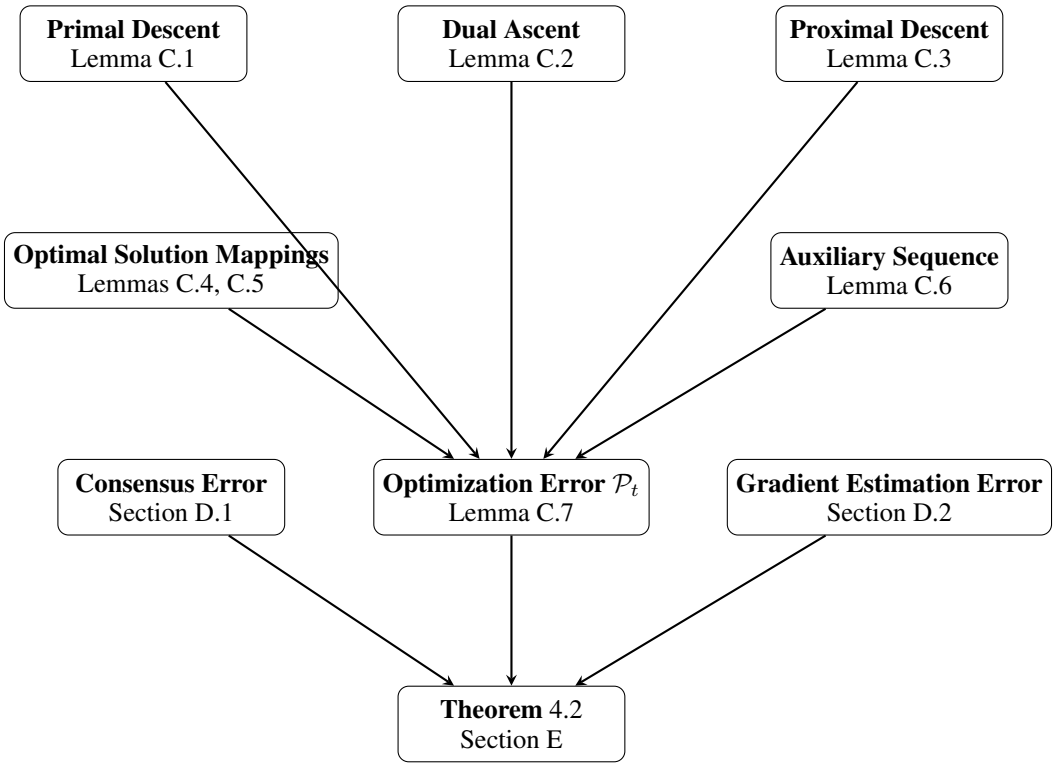


Figure 8: The structure of the proof for Theorem 4.2

It is worth noting that the STORM gradient estimator is a biased gradient estimator, so existing convergence analyses based on the deterministic gradient [38, 48] and the unbiased gradient estimator [34] cannot be applied directly to our algorithm. Moreover, most existing stochastic smoothing methods typically apply smoothing only to the primal variable, which makes their analysis insufficient for our algorithm.

In Figure 8, there are actually two key components in our proof: 1) **the optimization error related to doubly smoothing**, 2) **the consensus error and the gradient estimation error related to the decentralized setting**. In Section C, we provide the lemmas for bounding the optimization error. This includes:

- descent-ascent update lemmas (Lemma C.1, Lemma C.2, Lemma C.3);
- optimal solution mappings (Lemma C.4, Lemma C.5);
- auxiliary sequences (Lemma C.6).

These results are used in a potential function as Eq.(25):

$$\mathcal{P}_t = \mathbb{E}[F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)] - 2\mathbb{E}[F_d(\bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)] + 2\mathbb{E}[q(\bar{\hat{x}}_t)] \ ,$$

to establish the overall optimization error bound $\mathcal{P}_{t+1} - \mathcal{P}_t$ in Lemma C.7. It is worth noted that Lemma C.7 demonstrates that optimization error is affected by the consensus error caused by the decentralized setting and gradient estimation errors. Therefore, in Section D, we address two types of error in the decentralized setting:

- the consensus error, including that of auxiliary variables introduced by smoothing (Section D.1);
- the gradient estimation error from the STORM update (Section D.2).

After establishing all supporting lemmas, we proceed to derive the convergence rate through a novel potential function $\mathcal{L}_t$, which integrates the optimization error in Lemma C.7 and the consensus error and gradient estimation error together as follows:

$$\begin{aligned}
\mathcal{L}_t = & \underbrace{\mathcal{P}_t}_{\text{optimization error}} \underbrace{+ c_1\mathbb{E}[\|\bar{\mathcal{G}}_{x,t}\|^2] + c_2\mathbb{E}[\|\bar{\mathcal{G}}_{y,t}\|^2] + c_8\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\hat{\mathcal{G}}_{x,t}^{(k)}\|^2] + c_9\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\hat{\mathcal{G}}_{y,t}^{(k)}\|^2]}_{\text{gradient estimation error}} \\
& \underbrace{+ c_3\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2] + c_4\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|^2] + c_5\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\hat{x}}_t - \hat{x}_t^{(k)}\|^2]}_{\text{consensus error}} \\
& \underbrace{+ c_{10}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\hat{y}}_t - \hat{y}_t^{(k)}\|^2] + c_6\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t - p_t^{(k)}\|^2] + c_7\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{q}_t - q_t^{(k)}\|^2]}_{\text{consensus error}} \ .
\end{aligned}$$

By selecting appropriate hyperparameters, as detailed in Section E, we establish the convergence guarantee stated in Theorem 4.2. The construction of this proof framework is both technically intricate and conceptually non-trivial, underscoring the novelty and difficulty of our analysis.

B.1 Terminologies

To establish the convergence rate of Algorithm 1, we introduce the following symbols:

$$\begin{aligned}
& X_t = [x_t^{(1)}, x_t^{(2)}, \cdots, x_t^{(K)}] \in \mathbb{R}^{d_1 \times K} \ , \ \tilde{X}_t = [\tilde{x}_t^{(1)}, \tilde{x}_t^{(2)}, \cdots, \tilde{x}_t^{(K)}] \in \mathbb{R}^{d_1 \times K} \ , \\
& Y_t = [y_t^{(1)}, y_t^{(2)}, \cdots, y_t^{(K)}] \in \mathbb{R}^{d_2 \times K} \ , \ \tilde{Y}_t = [\tilde{y}_t^{(1)}, \tilde{y}_t^{(2)}, \cdots, \tilde{y}_t^{(K)}] \in \mathbb{R}^{d_2 \times K} \ , \\
& \hat{X}_t = [\hat{x}_t^{(1)}, \hat{x}_t^{(2)}, \cdots, \hat{x}_t^{(K)}] \in \mathbb{R}^{d_1 \times K} \ , \ \tilde{\hat{X}}_t = [\tilde{\hat{x}}_t^{(1)}, \tilde{\hat{x}}_t^{(2)}, \cdots, \tilde{\hat{x}}_t^{(K)}] \in \mathbb{R}^{d_1 \times K} \ , \\
& \hat{Y}_t = [\hat{y}_t^{(1)}, \hat{y}_t^{(2)}, \cdots, \hat{y}_t^{(K)}] \in \mathbb{R}^{d_2 \times K} \ , \ \tilde{\hat{Y}}_t = [\tilde{\hat{y}}_t^{(1)}, \tilde{\hat{y}}_t^{(2)}, \cdots, \tilde{\hat{y}}_t^{(K)}] \in \mathbb{R}^{d_2 \times K} \ , \\
& U_t = [u_t^{(1)}, u_t^{(2)}, \cdots, u_t^{(K)}] \in \mathbb{R}^{d_1 \times K} \ , \ V_t = [v_t^{(1)}, v_t^{(2)}, \cdots, v_t^{(K)}] \in \mathbb{R}^{d_2 \times K} \ ,
\end{aligned}$$

$$
\begin{aligned}
&P_t = [p_t^{(1)}, p_t^{(2)}, \cdots, p_t^{(K)}] \in \mathbb{R}^{d_1 \times K} \ , \ Q_t = [q_t^{(1)}, q_t^{(2)}, \cdots, q_t^{(K)}] \in \mathbb{R}^{d_2 \times K} \ , \\
&\bar{X}_t = \frac{1}{K} X_t \mathbf{1}\mathbf{1}^T \ , \ \bar{Y}_t = \frac{1}{K} Y_t \mathbf{1}\mathbf{1}^T \ , \ \bar{\hat{X}}_t = \frac{1}{K} \hat{X}_t \mathbf{1}\mathbf{1}^T \ , \ \bar{\hat{Y}}_t = \frac{1}{K} \hat{Y}_t \mathbf{1}\mathbf{1}^T \ , \\
&\bar{U}_t = \frac{1}{K} U_t \mathbf{1}\mathbf{1}^T \ , \ \bar{V}_t = \frac{1}{K} V_t \mathbf{1}\mathbf{1}^T \ , \ \bar{P}_t = \frac{1}{K} P_t \mathbf{1}\mathbf{1}^T \ , \ \bar{Q}_t = \frac{1}{K} Q_t \mathbf{1}\mathbf{1}^T \ ,
\end{aligned} \tag{11}
$$

where $\mathbf{1} = [1, 1, \cdots, 1]^T \in \mathbb{R}^K$. Based on these terminologies, the update of x, y, $\hat{x}$, $\hat{y}$, p, and q in Algorithm 1 is represented as follows:

$$
\begin{aligned}
&\tilde{X}_{t+1} = X_t W - \beta_x P_t \ , X_{t+1} = X_t + \eta(\tilde{X}_{t+1} - X_t) \ , \\
&\tilde{Y}_{t+1} = Y_t W + \beta_y Q_t \ , \ Y_{t+1} = Y_t + \eta(\tilde{Y}_{t+1} - Y_t) \ , \\
&\tilde{\hat{X}}_{t+1} = \hat{X}_t W + \hat{\beta}_x (X_{t+1} - \hat{X}_t) \ , \ \hat{X}_{t+1} = \hat{X}_t + \eta(\tilde{\hat{X}}_{t+1} - \hat{X}_t) \ , \\
&\tilde{\hat{Y}}_{t+1} = \hat{Y}_t W + \hat{\beta}_y (Y_{t+1} - \hat{Y}_t) \ , \ \hat{Y}_{t+1} = \hat{Y}_t + \eta(\tilde{\hat{Y}}_{t+1} - \hat{Y}_t) \ , \\
&P_{t+1} = P_t W + U_{t+1} - U_t \ , \ Q_{t+1} = Q_t W + V_{t+1} - V_t \ , \\
&\bar{X}_{t+1} = \bar{X}_t - \beta_x \eta \bar{U}_t \ , \ \bar{Y}_{t+1} = \bar{Y}_t + \beta_y \eta \bar{V}_t \ , \\
&\bar{\hat{X}}_{t+1} = \bar{\hat{X}}_t + \hat{\beta}_x \eta (\bar{X}_{t+1} - \bar{\hat{X}}_t) \ , \bar{\hat{Y}}_{t+1} = \bar{\hat{Y}}_t + \hat{\beta}_y \eta (\bar{Y}_{t+1} - \bar{\hat{Y}}_t) \ .
\end{aligned} \tag{12}
$$

Note that $\bar{P}_t = \bar{U}_t$ and $\bar{Q}_t = \bar{V}_t$.

In addition, to simplify notation, we provide the following gradient estimation errors

$$
\begin{aligned}
&\bar{\mathcal{G}}_{x,t} = \frac{1}{K} \sum_{k=1}^K u_t^{(k)} - \frac{1}{K} \sum_{k=1}^K \nabla_x F^{(k)}(x_t^{(k)}, y_t^{(k)}; \hat{x}_t^{(k)}, \hat{y}_t^{(k)}) \ , \\
&\bar{\mathcal{G}}_{y,t} = \frac{1}{K} \sum_{k=1}^K v_t^{(k)} - \frac{1}{K} \sum_{k=1}^K \nabla_y F^{(k)}(x_t^{(k)}, y_t^{(k)}; \hat{x}_t^{(k)}, \hat{y}_t^{(k)}) \ , \\
&\hat{\mathcal{G}}_{x,t}^{(k)} = u_t^{(k)} - \nabla_x F^{(k)}(x_t^{(k)}, y_t^{(k)}; \hat{x}_t^{(k)}, \hat{y}_t^{(k)}) \ , \\
&\hat{\mathcal{G}}_{y,t}^{(k)} = v_t^{(k)} - \nabla_y F^{(k)}(x_t^{(k)}, y_t^{(k)}; \hat{x}_t^{(k)}, \hat{y}_t^{(k)}) \ .
\end{aligned} \tag{13}
$$

Moreover, following [34, 48], we introduce the following auxiliary functions and variables for convergence analysis:

$$
\begin{aligned}
&F_d(y; \hat{x}, \hat{y}) = \min_{x \in \mathbb{R}^{d_1}} F(x, y; \hat{x}, \hat{y}) \ , \qquad \text{dual function} \\
&F_p(x; \hat{x}, \hat{y}) = \max_{y \in \mathbb{R}^{d_2}} F(x, y; \hat{x}, \hat{y}) \ , \quad \text{primal function} \\
&g(\hat{x}, \hat{y}) = \min_{x \in \mathbb{R}^{d_1}} \max_{y \in \mathbb{R}^{d_2}} F(x, y; \hat{x}, \hat{y}) \ , \\
&p(\hat{y}) = \min_{\hat{x} \in \mathbb{R}^{d_1}} g(\hat{x}, \hat{y}) \ , \qquad q(\hat{x}) = \max_{\hat{y} \in \mathbb{R}^{d_2}} g(\hat{x}, \hat{y}) \ , \\
&x^*(y; \hat{x}, \hat{y}) = \arg\min_{x \in \mathbb{R}^{d_1}} F(x, y; \hat{x}, \hat{y}) \ , \\
&y^*(x; \hat{x}, \hat{y}) = \arg\max_{y \in \mathbb{R}^{d_2}} F(x, y; \hat{x}, \hat{y}) \ , \\
&x^*(\hat{x}, \hat{y}) \triangleq x^*(y^*(\hat{x}, \hat{y}); \hat{x}, \hat{y}) = \arg\min_{x \in \mathbb{R}^{d_1}} F_p(x; \hat{x}, \hat{y}) \ , \\
&y^*(\hat{x}, \hat{y}) \triangleq y^*(x^*(\hat{x}, \hat{y}); \hat{x}, \hat{y}) = \arg\max_{y \in \mathbb{R}^{d_2}} F_d(y; \hat{x}, \hat{y}) \ , \\
&\hat{x}^*(\hat{y}) = \arg\min_{\hat{x} \in \mathbb{R}^{d_1}} g(\hat{x}, \hat{y}) \ , \qquad \hat{y}^*(\hat{x}) = \arg\max_{\hat{y} \in \mathbb{R}^{d_2}} g(\hat{x}, \hat{y}) \ , \\
&y^+(\hat{x}_t, \hat{y}_t) = y_t + \beta_y \eta \nabla_y F_d(y_t; \hat{x}_t, \hat{y}_t) \ , \\
&\hat{y}^+(\hat{x}_{t+1}) = \hat{y}_t + \hat{\beta}_y \eta (y^*(\hat{x}_t, \hat{y}_t) - \hat{y}_t) \ .
\end{aligned} \tag{14}
$$

B.2 Function Properties

Lemma B.1. *[48] Given Assumptions 3.1-3.4, then $F(x, y; \hat{x}, \hat{y})$ is $(\gamma_1 + L)$-smooth and $(\gamma_1 - L)$-strongly convex with respect to x. $F(x, y; \hat{x}, \hat{y})$ is $(\gamma_2 + L)$-smooth and $(\gamma_2 - L)$-strongly concave with respect to y.*

Lemma B.2. *[48] Given Assumptions 3.1-3.4, the following inequality holds:*

$$\begin{aligned}
&\|x^*(y_1; \hat{x}, \hat{y}) - x^*(y_2; \hat{x}, \hat{y})\| \leq C_{x^1_{y\hat{x}\hat{y}}} \|y_1 - y_2\| , \\
&\|x^*(y; \hat{x}_1, \hat{y}) - x^*(y; \hat{x}_2, \hat{y})\| \leq C_{x^2_{y\hat{x}\hat{y}}} \|\hat{x}_1 - \hat{x}_2\| , \\
&\|x^*(\hat{x}_1, \hat{y}) - x^*(\hat{x}_2, \hat{y})\| \leq C_{x^1_{\hat{x}\hat{y}}} \|\hat{x}_1 - \hat{x}_2\| , \\
&\|y^*(x_1; \hat{x}, \hat{y}) - y^*(x_2; \hat{x}, \hat{y})\| \leq C_{y^1_{x\hat{x}\hat{y}}} \|x_1 - x_2\| , \\
&\|y^*(x; \hat{x}, \hat{y}_1) - y^*(x; \hat{x}, \hat{y}_2)\| \leq C_{y^3_{x\hat{x}\hat{y}}} \|\hat{y}_1 - \hat{y}_2\| , \\
&\|y^*(\hat{x}_1, \hat{y}) - y^*(\hat{x}_2, \hat{y})\| \leq C_{y^1_{\hat{x}\hat{y}}} \|\hat{x}_1 - \hat{x}_2\| , \\
&\|y^*(\hat{x}, \hat{y}_1) - y^*(\hat{x}, \hat{y}_2)\| \leq C_{y^2_{\hat{x}\hat{y}}} \|\hat{y}_1 - \hat{y}_2\| ,
\end{aligned} \tag{15}$$

where

$$\begin{aligned}
&C_{x^1_{y\hat{x}\hat{y}}} = \frac{\gamma_1}{\gamma_1 - L} , \quad C_{x^2_{y\hat{x}\hat{y}}} = \frac{\gamma_1}{\gamma_1 - L} , \quad C_{x^1_{\hat{x}\hat{y}}} = \frac{\gamma_1}{\gamma_1 - L} , \\
&C_{y^1_{x\hat{x}\hat{y}}} = \frac{\gamma_2}{\gamma_2 - L} , \quad C_{y^3_{x\hat{x}\hat{y}}} = \frac{\gamma_2}{\gamma_2 - L} , \quad C_{y^1_{\hat{x}\hat{y}}} = \frac{\gamma_1}{\gamma_2 - L} C_{x^1_{y\hat{x}\hat{y}}} + 1 , C_{y^2_{\hat{x}\hat{y}}} = \frac{\gamma_2}{\gamma_2 - L} .
\end{aligned} \tag{16}$$

Lemma B.3. *[48] Given Assumptions 3.1-3.4, then $F_d(y; \hat{x}, \hat{y})$ is L_d-smooth, where $L_d = LC_{x^1_{y\hat{x}\hat{y}}} + L + \gamma_2$.*

Lemma B.4. *Given Assumptions 3.1-3.4, by defining $y^+(\hat{x}_t, \hat{y}_t) = y_t + \beta_y \eta \nabla_y F_d(y_t; \hat{x}_t, \hat{y}_t)$, the following inequality holds:*

$$\|y_t - y^*(\hat{x}_t, \hat{y}_t)\| \leq \frac{1}{\beta_y \eta(\gamma_2 - L)} \|y_t - y^+(\hat{x}_t, \hat{y}_t)\| . \tag{17}$$

Proof. Due to $y^*(\hat{x}_t, \hat{y}_t) = \arg\max_{y \in \mathbb{R}^{d_2}} F_d(y; \hat{x}_t, \hat{y}_t)$, for any $y \in \mathbb{R}^{d_2}$, we have

$$\langle y - y^*(\hat{x}_t, \hat{y}_t), \nabla_y F_d(y^*(\hat{x}_t, \hat{y}_t); \hat{x}_t, \hat{y}_t) \rangle \leq 0.$$

By taking $y = y_t$, we have

$$\langle y_t - y^*(\hat{x}_t, \hat{y}_t), \nabla_y F_d(y^*(\hat{x}_t, \hat{y}_t); \hat{x}_t, \hat{y}_t) \rangle \leq 0 .$$

In addition, because $F_d(y; \hat{x}, \hat{y})$ is $(\gamma_2 - L)$-strongly concave with respect to y, we have

$$\langle y_t - y^*(\hat{x}_t, \hat{y}_t), \nabla_y F_d(y_t; \hat{x}_t, \hat{y}_t) - \nabla_y F_d(y^*(\hat{x}_t, \hat{y}_t); \hat{x}_t, \hat{y}_t) \rangle + (\gamma_2 - L)\|y_t - y^*(\hat{x}_t, \hat{y}_t)\|^2 \leq 0 .$$

Combining the above two inequalities yields

$$\langle y_t - y^*(\hat{x}_t, \hat{y}_t), \nabla_y F_d(y_t; \hat{x}_t, \hat{y}_t) \rangle + (\gamma_2 - L)\|y_t - y^*(\hat{x}_t, \hat{y}_t)\|^2 \leq 0 .$$

Therefore, we obtain

$$\begin{aligned}
&(\gamma_2 - L)\|y_t - y^*(\hat{x}_t, \hat{y}_t)\|^2 \leq \langle y^*(\hat{x}_t, \hat{y}_t) - y_t, \nabla_y F_d(y_t; \hat{x}_t, \hat{y}_t) \rangle \\
&\leq \|y_t - y^*(\hat{x}_t, \hat{y}_t)\| \|\nabla_y F_d(y_t; \hat{x}_t, \hat{y}_t)\| = \|y_t - y^*(\hat{x}_t, \hat{y}_t)\| \|\frac{y^+(\hat{x}_t, \hat{y}_t) - y_t}{\beta_y \eta}\| .
\end{aligned}$$

As a result, we have

$$\|y_t - y^*(\hat{x}_t, \hat{y}_t)\| \leq \frac{1}{\beta_y \eta(\gamma_2 - L)} \|y^+(\hat{x}_t, \hat{y}_t) - y_t\| .$$

□

Lemma B.5. *Given Assumptions 3.1-3.4, then*

$$\|x_t - x^*(y_t; \hat{x}_t, \hat{y}_t)\| \leq \frac{1}{\gamma_1 - L} \|\nabla_x F(x_t, y_t; \hat{x}_t, \hat{y}_t)\|.$$

Proof. Due to $x^*(y_t;\hat{x}_t,\hat{y}_t) = \arg\min_{x\in\mathbb{R}^{d_1}} F(x,y_t;\hat{x}_t,\hat{y}_t)$, for any $x\in\mathbb{R}^{d_1}$, we have

$$\langle x - x^*(y_t;\hat{x}_t,\hat{y}_t), -\nabla_x F(x^*(y_t;\hat{x}_t,\hat{y}_t),y_t;\hat{x}_t,\hat{y}_t)\rangle \leq 0.$$

By taking $x = x_t$, we have

$$\langle x_t - x^*(y_t;\hat{x}_t,\hat{y}_t), \nabla_x F(x^*(y_t;\hat{x}_t,\hat{y}_t),y_t;\hat{x}_t,\hat{y}_t)\rangle \geq 0.$$

In addition, since $F(x,y;\hat{x},\hat{y})$ is $(\gamma_1 - L)$-strongly convex with respect to x, we have

$$\begin{aligned}&\langle x_t - x^*(y_t;\hat{x}_t,\hat{y}_t), \nabla_x F(x_t,y_t;\hat{x}_t,\hat{y}_t) - \nabla_x F(x^*(y_t;\hat{x}_t,\hat{y}_t),y_t;\hat{x}_t,\hat{y}_t)\rangle\\ &\geq (\gamma_1 - L)\|x_t - x^*(y_t;\hat{x}_t,\hat{y}_t)\|^2 .\end{aligned}$$

Combing the above two inequalities yields

$$\begin{aligned}&(\gamma_1 - L)\|x_t - x^*(y_t;\hat{x}_t,\hat{y}_t)\|^2 \leq \langle x_t - x^*(y_t;\hat{x}_t,\hat{y}_t), \nabla_x F(x_t,y_t;\hat{x}_t,\hat{y}_t)\rangle\\ &\leq \|x_t - x^*(y_t;\hat{x}_t,\hat{y}_t)\|\|\nabla_x F(x_t,y_t;\hat{x}_t,\hat{y}_t)\| .\end{aligned}$$

As a result, we have

$$\|x_t - x^*(y_t;\hat{x}_t,\hat{y}_t)\| \leq \frac{1}{\gamma_1 - L}\|\nabla_x F(x_t,y_t;\hat{x}_t,\hat{y}_t)\|.$$

□

C Optimization Errors

Lemma C.1. *Given Assumptions 3.1-3.4 and $\eta \leq \frac{1}{2\beta_x(\gamma_1+L)}$, the following inequality holds:*

$$\begin{aligned}&\mathbb{E}[F(\bar{x}_{t+1},\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_{t+1})] - \mathbb{E}[F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)] \leq -\frac{\beta_x\eta}{2}\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]\\ &+ \frac{\beta_y\eta}{2}\mathbb{E}[\|\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2] + \frac{\beta_x\eta}{2}\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t) - \bar{u}_t\|^2]\\ &+ (4\beta_y\eta\beta_x^2\eta^2L^2 - \frac{\beta_x\eta}{4})\mathbb{E}[\|\bar{u}_t\|^2] + (\frac{3\beta_y\eta}{4} + \frac{\beta_y^2\eta^2(\gamma_2+L)}{2})\mathbb{E}[\|\bar{v}_t\|^2]\\ &- \frac{\gamma_1(2-\hat{\beta}_x\eta)}{2\hat{\beta}_x\eta}\mathbb{E}[\|\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t\|^2] - \frac{\gamma_2(\hat{\beta}_y\eta - 2)}{2\hat{\beta}_y\eta}\mathbb{E}[\|\bar{\hat{y}}_{t+1} - \bar{\hat{y}}_t\|^2] .\end{aligned}$$

Proof. First, we decompose the formulation as

$$\begin{aligned}&\mathbb{E}[F(\bar{x}_{t+1},\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_{t+1})] - \mathbb{E}[F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)]\\ &= \mathbb{E}[F(\bar{x}_{t+1},\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_{t+1}) - F(\bar{x}_{t+1},\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)\\ &\quad + \Big(F(\bar{x}_{t+1},\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t) - F(\bar{x}_{t+1},\bar{y}_{t+1};\bar{\hat{x}}_t,\bar{\hat{y}}_t)\Big)\\ &\quad + \Big(F(\bar{x}_{t+1},\bar{y}_{t+1};\bar{\hat{x}}_t,\bar{\hat{y}}_t) - F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)]\Big) .\end{aligned}$$

Since $F(x,y;\hat{x},\hat{y})$ is $(L+\gamma_1)$-smooth with respect to x, we have

$$\begin{aligned}&\mathbb{E}[F(\bar{x}_{t+1},\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)] - \mathbb{E}[F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)]\\ &\leq \mathbb{E}[\langle\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t), \bar{x}_{t+1} - \bar{x}_t\rangle] + \frac{L+\gamma_1}{2}\mathbb{E}[\|\bar{x}_{t+1} - \bar{x}_t\|^2]\\ &= -\beta_x\eta\mathbb{E}[\langle\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t), \bar{u}_t\rangle] + \frac{\beta_x^2\eta^2(L+\gamma_1)}{2}\mathbb{E}[\|\bar{u}_t\|^2]\\ &= -\frac{\beta_x\eta}{2}\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2] - \frac{\beta_x\eta}{2}\mathbb{E}[\|\bar{u}_t\|^2] + \frac{\beta_x\eta}{2}\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t) - \bar{u}_t\|^2]\\ &\quad + \frac{\beta_x^2\eta^2(L+\gamma_1)}{2}\mathbb{E}[\|\bar{u}_t\|^2] \qquad (18)\\ &\leq -\frac{\beta_x\eta}{2}\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2] - \frac{\beta_x\eta}{4}\mathbb{E}[\|\bar{u}_t\|^2] + \frac{\beta_x\eta}{2}\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t) - \bar{u}_t\|^2] ,\end{aligned}$$

where the last step holds due to $\eta \leq \frac{1}{2\beta_x(\gamma_1+L)}$.

In addition, since $F(x,y;\hat{x},\hat{y})$ is $(L+\gamma_1)$-smooth with respect to y, we have

$$
\begin{aligned}
&\mathbb{E}[F(\bar{x}_{t+1},\bar{y}_{t+1};\bar{\hat{x}}_t,\bar{\hat{y}}_t)] - \mathbb{E}[F(\bar{x}_{t+1},\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)] \\
&\leq \mathbb{E}[\langle \nabla_y F(\bar{x}_{t+1},\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t), \bar{y}_{t+1}-\bar{y}_t\rangle] + \frac{\gamma_2+L}{2}\mathbb{E}[\|\bar{y}_{t+1}-\bar{y}_t\|^2] \\
&= \beta_y\eta\mathbb{E}[\langle \nabla_y F(\bar{x}_{t+1},\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t) - \nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t), \bar{v}_t\rangle] + \beta_y\eta\mathbb{E}[\langle \nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t), \bar{v}_t\rangle] \\
&\quad + \frac{\beta_y^2\eta^2(\gamma_2+L)}{2}\mathbb{E}[\|\bar{v}_t\|^2] \qquad (19) \\
&\leq 4\beta_y\eta\mathbb{E}[\|\nabla_y F(\bar{x}_{t+1},\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t) - \nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2] + \frac{\beta_y\eta}{4}\mathbb{E}[\|\bar{v}_t\|^2] \\
&\quad + \frac{\beta_y\eta}{2}\mathbb{E}[\|\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2] + \frac{\beta_y\eta}{2}\mathbb{E}[\|\bar{v}_t\|^2] + \frac{\beta_y^2\eta^2(\gamma_2+L)}{2}\mathbb{E}[\|\bar{v}_t\|^2] \\
&\leq \frac{\beta_y\eta}{2}\mathbb{E}[\|\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2] + 4\beta_y\eta\beta_x^2\eta^2L^2\mathbb{E}[\|\bar{u}_t\|^2] + \left(\frac{3\beta_y\eta}{4} + \frac{\beta_y^2\eta^2(\gamma_2+L)}{2}\right)\mathbb{E}[\|\bar{v}_t\|^2]\,,
\end{aligned}
$$

where the last step holds due to the following inequality.

$$
\begin{aligned}
&\mathbb{E}[\|\nabla_y F(\bar{x}_{t+1},\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t) - \nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2] \\
&= \mathbb{E}[\|\nabla_y f(\bar{x}_{t+1},\bar{y}_t) + \gamma_2(\bar{y}_t-\bar{\hat{y}}_t) - \nabla_y f(\bar{x}_t,\bar{y}_t) - \gamma_2(\bar{y}_t-\bar{\hat{y}}_t)\|^2] \\
&\leq L^2\mathbb{E}[\|\bar{x}_{t+1}-\bar{x}_t\|^2] \leq \beta_x^2\eta^2L^2\mathbb{E}[\|\bar{u}_t\|^2]\,.
\end{aligned}
$$

Combining Eq. (18) and Eq. (19) yields

$$
\begin{aligned}
&\mathbb{E}[F(\bar{x}_{t+1},\bar{y}_{t+1};\bar{\hat{x}}_t,\bar{\hat{y}}_t)] - \mathbb{E}[F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)] \\
&\leq -\frac{\beta_x\eta}{2}\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2] + \frac{\beta_y\eta}{2}\mathbb{E}[\|\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2] \\
&\quad + \frac{\beta_x\eta}{2}\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t) - \bar{u}_t\|^2] + (4\beta_y\eta\beta_x^2\eta^2L^2 - \frac{\beta_x\eta}{4})\mathbb{E}[\|\bar{u}_t\|^2] \\
&\quad + (\frac{3\beta_y\eta}{4} + \frac{\beta_y^2\eta^2(\gamma_2+L)}{2})\mathbb{E}[\|\bar{v}_t\|^2]\,.
\end{aligned}
$$

Moreover, according to the definition of $F(x,y;\hat{x},\hat{y})$, we have

$$
\begin{aligned}
&F(\bar{x}_{t+1},\bar{y}_{t+1};\bar{\hat{x}}_t,\bar{\hat{y}}_t) - F(\bar{x}_{t+1},\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t) \\
&= f(\bar{x}_{t+1},\bar{y}_{t+1}) + \frac{\gamma_1}{2}\|\bar{x}_{t+1}-\bar{\hat{x}}_t\|^2 - \frac{\gamma_2}{2}\|\bar{y}_{t+1}-\bar{\hat{y}}_t\|^2 \\
&\quad - f(\bar{x}_{t+1},\bar{y}_{t+1}) - \frac{\gamma_1}{2}\|\bar{x}_{t+1}-\bar{\hat{x}}_{t+1}\|^2 + \frac{\gamma_2}{2}\|\bar{y}_{t+1}-\bar{\hat{y}}_t\|^2 \\
&= \frac{\gamma_1}{2}\left(\|\bar{x}_{t+1}-\bar{\hat{x}}_t\|^2 - \|\bar{x}_{t+1}-\bar{\hat{x}}_{t+1}\|^2\right) = \frac{\gamma_1}{2}\left(\|\bar{x}_{t+1}-\bar{\hat{x}}_t\|^2 - \|(1-\hat{\beta}_x\eta)(\bar{x}_{t+1}-\bar{\hat{x}}_t)\|^2\right) \\
&= \frac{\gamma_1(1-(1-\hat{\beta}_x\eta)^2)}{2}\|\bar{x}_{t+1}-\bar{\hat{x}}_t\|^2 = \frac{\gamma_1(1-(1-\hat{\beta}_x\eta)^2)}{2\hat{\beta}_x^2\eta^2}\|\bar{\hat{x}}_{t+1}-\bar{\hat{x}}_t\|^2 \\
&= \frac{\gamma_1(2-\hat{\beta}_x\eta)}{2\hat{\beta}_x\eta}\|\bar{\hat{x}}_{t+1}-\bar{\hat{x}}_t\|^2\,, \qquad (20)
\end{aligned}
$$

where the third and fifth steps hold due to $\bar{\hat{x}}_{t+1} = \bar{\hat{x}}_t + \hat{\beta}_x\eta(\bar{x}_{t+1}-\bar{\hat{x}}_t)$.

Similarly, we have

$$
F(\bar{x}_{t+1},\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t) - F(\bar{x}_{t+1},\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_{t+1}) = \frac{\gamma_2(\hat{\beta}_y\eta-2)}{2\hat{\beta}_y\eta}\|\bar{\hat{y}}_{t+1}-\bar{\hat{y}}_t\|^2\,. \qquad (21)
$$

By combining the above three inequalities, the proof is complete. □

Lemma C.2. *Given Assumptions 3.1-3.4, the following inequality holds:*

$$\mathbb{E}[F_d(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_{t+1})]-\mathbb{E}[F_d(\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)]\geq\beta_y\eta\mathbb{E}[\langle\nabla_y F_d(\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t),\bar{v}_t\rangle]-\frac{\beta_y^2\eta^2L_d}{2}\mathbb{E}[\|\bar{v}_t\|^2]$$
$$+\frac{\gamma_2(2-\hat{\beta}_y\eta)}{2\hat{\beta}_y\eta}\mathbb{E}[\|\bar{\hat{y}}_{t+1}-\bar{\hat{y}}_t\|^2]+\frac{\gamma_1}{2}\mathbb{E}[\langle\bar{\hat{x}}_{t+1}-\bar{\hat{x}}_t,\bar{\hat{x}}_{t+1}+\bar{\hat{x}}_t-2x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)\rangle]\,.$$

Proof. First, we decompose the formulation as

$$\begin{aligned}&\mathbb{E}[F_d(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_{t+1})]-\mathbb{E}[F_d(\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)]\\&=\mathbb{E}[F_d(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_{t+1})-F_d(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)\\&\quad+\Big(F_d(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)-F_d(\bar{y}_{t+1};\bar{\hat{x}}_t,\bar{\hat{y}}_t)\Big)+\Big(F_d(\bar{y}_{t+1};\bar{\hat{x}}_t,\bar{\hat{y}}_t)-F_d(\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)]\Big)\,.\end{aligned}$$

According to the definition of $F_d(y;\hat{x},\hat{y})$, we have

$$\begin{aligned}&F_d(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_{t+1})-F_d(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)\\&=F(x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_{t+1}),\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_{t+1})-F(x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t),\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)\\&\geq F(x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_{t+1}),\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_{t+1})-F(x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_{t+1}),\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)\\&=\frac{\gamma_2(2-\hat{\beta}_y\eta)}{2\hat{\beta}_y\eta}\|\bar{\hat{y}}_{t+1}-\bar{\hat{y}}_t\|^2\,,\end{aligned}$$

where the second step holds due to $x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)=\arg\min_{x\in\mathbb{R}^{d_1}}F(x,\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)$, the last step holds as Eq. (21).

In addition, according to the definition of $F_d(y;\hat{x},\hat{y})$, we have

$$\begin{aligned}&F_d(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)-F_d(\bar{y}_{t+1};\bar{\hat{x}}_t,\bar{\hat{y}}_t)\\&=F(x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t),\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)-F(x^*(\bar{y}_{t+1};\bar{\hat{x}}_t,\bar{\hat{y}}_t),\bar{y}_{t+1};\bar{\hat{x}}_t,\bar{\hat{y}}_t)\\&\geq F(x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t),\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)-F(x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t),\bar{y}_{t+1};\bar{\hat{x}}_t,\bar{\hat{y}}_t)\\&=\frac{\gamma_1}{2}(\|x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)-\bar{\hat{x}}_{t+1}\|^2-\|x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)-\bar{\hat{x}}_t\|^2)\\&=\frac{\gamma_1}{2}\langle x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)-\bar{\hat{x}}_{t+1}-(x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)-\bar{\hat{x}}_t),\\&\qquad x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)-\bar{\hat{x}}_{t+1}+(x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)-\bar{\hat{x}}_t)\rangle\\&=\frac{\gamma_1}{2}\langle\bar{\hat{x}}_{t+1}-\bar{\hat{x}}_t,\bar{\hat{x}}_{t+1}+\bar{\hat{x}}_t-2x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)\rangle\,,\end{aligned}$$

where the second step holds due to $x^*(\bar{y}_{t+1};\bar{\hat{x}}_t,\bar{\hat{y}}_t)=\arg\min_{x\in\mathbb{R}^{d_1}}F(x,\bar{y}_{t+1};\bar{\hat{x}}_t,\bar{\hat{y}}_t)$, the fourth step holds due to the fact $a^2-b^2=(a-b)(a+b)$.

Moreover, since $F_d(y;\hat{x},\hat{y})$ is L_d-smooth, we have

$$\begin{aligned}F_d(\bar{y}_{t+1};\bar{\hat{x}}_t,\bar{\hat{y}}_t)&\geq F_d(\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)+\langle\nabla_y F_d(\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t),\bar{y}_{t+1}-\bar{y}_t\rangle-\frac{L_d}{2}\|\bar{y}_{t+1}-\bar{y}_t\|^2\\&=F_d(\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)+\beta_y\eta\langle\nabla_y F_d(\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t),\bar{v}_t\rangle-\frac{\beta_y^2\eta^2L_d}{2}\|\bar{v}_t\|^2\,.\end{aligned}\tag{22}$$

By combining the above three inequalities, the proof is complete. □

Lemma C.3. *Given Assumptions 3.1-3.4, the following inequality holds:*

$$q(\bar{\hat{x}}_{t+1})-q(\bar{\hat{x}}_t)\leq\frac{\gamma_1}{2}\langle\bar{\hat{x}}_{t+1}-\bar{\hat{x}}_t,\bar{\hat{x}}_{t+1}+\bar{\hat{x}}_t-2x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1}))\rangle\,.\tag{23}$$

Proof. First, we have

$$\begin{aligned}&q(\bar{\hat{x}}_{t+1})-q(\bar{\hat{x}}_t)=g(\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1}))-g(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_t))\\&=F_p(x^*(\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1}));\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1}))-F_p(x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_t));\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_t))\end{aligned}$$

$$\leq F_p(x^*(\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1})) - F_p(x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1}))$$
$$\leq \underbrace{F_p(x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1})) - F_p(x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1}))}_{T_1} .$$

where the second step holds due to $g(\hat{x},\hat{y}) = \min_{x\in\mathbb{R}^{d_1}} F_p(x;\hat{x},\hat{y})$, the three inequalities hold due to $y^*(x;\hat{x},\hat{y}) = \arg\max_y F(x,y;\hat{x},\hat{y})$ and $F_p(x;\hat{x},\hat{y}) = F(x,y^*(x;\hat{x},\hat{y});\hat{x},\hat{y})$.

Then, from the definition of F_p, we derive

$$\begin{aligned}
T_1 &= F(x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})),y^*(x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1})) \\
&\quad - F(x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})),y^*(x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})) \\
&\leq F(x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})),y^*(x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1})) \\
&\quad - F(x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})),y^*(x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})) \\
&= \frac{\gamma_1}{2}(\|x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})) - \bar{\hat{x}}_{t+1}\|^2 - \|x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})) - \bar{\hat{x}}_t\|^2) \\
&= \frac{\gamma_1}{2}\langle x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})) - \bar{\hat{x}}_{t+1} - (x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})) - \bar{\hat{x}}_t), \\
&\quad x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})) - \bar{\hat{x}}_{t+1} + (x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1})) - \bar{\hat{x}}_t)\rangle \\
&= \frac{\gamma_1}{2}\langle \bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t, \bar{\hat{x}}_{t+1} + \bar{\hat{x}}_t - 2x^*(\bar{\hat{x}}_t,\hat{y}^*(\bar{\hat{x}}_{t+1}))\rangle ,
\end{aligned}$$

the second to last step holds due to the fact $a^2 - b^2 = (a-b)(a+b)$. □

Lemma C.4. *Given Assumptions 3.1-3.4, the following inequality holds:*

$$\begin{aligned}
&\|x^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1})) - x^*(\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1}))\|^2 \qquad (24)\\
&\leq \frac{2}{\gamma_1 - L}\frac{2\gamma_2^2 C_{y^1_{\hat{x}\hat{y}}}^2}{\mu}\|\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t\|^2 + \frac{2}{\gamma_1 - L}\frac{2\gamma_2^2}{\mu}\left(C_{y^2_{\hat{x}\hat{y}}}^2 + \frac{(1-\hat{\beta}_y\eta)^2}{\hat{\beta}_y^2\eta^2}\right)\|\bar{\hat{y}}_t - \hat{y}^+(\bar{\hat{x}}_{t+1})\|^2 .
\end{aligned}$$

Proof. Since $F_p(x;\hat{x},\hat{y})$ is $(\gamma_1 - L)$-strongly convex with respect to x, we derive

$$\begin{aligned}
&\frac{\gamma_1 - L}{2}\|x^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1})) - x^*(\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1}))\|^2 \\
&\leq F_p(x^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1})) - F_p(x^*(\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1})) \\
&\leq \max_{\hat{y}\in\mathbb{R}^{d_2}} F_p(x^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_{t+1},\hat{y}) - F_p(x^*(\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1})) \\
&\leq \max_{\hat{y}\in\mathbb{R}^{d_2}} F_p(x^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_{t+1},\hat{y}) - F_p(x^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1})) \\
&\leq \underbrace{\frac{1}{2\mu}\|\nabla_{\hat{y}} F_p(x^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1}))\|^2}_{T_1} ,
\end{aligned}$$

where the last step holds due to Theorem 5.2 of [37] with PL property being a special KL property.

Then, from the definition of F_p, we derive

$$\begin{aligned}
T_1 &= \frac{\gamma_2^2}{2\mu}\|y^*(x^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1})) - \hat{y}^+(\bar{\hat{x}}_{t+1})\|^2 \\
&= \frac{\gamma_2^2}{2\mu}\|y^*(x^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1})),\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1})) - \bar{\hat{y}}_t - \hat{\beta}_y\eta(y^*(\bar{\hat{x}}_t,\bar{\hat{y}}_t) - \bar{\hat{y}}_t)\|^2 \\
&\leq \frac{\gamma_2^2}{\mu}\|y^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1})) - y^*(\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2 + (1-\hat{\beta}_y\eta)^2\frac{\gamma_2^2}{\mu}\|\bar{\hat{y}}_t - y^*(\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2 \\
&\leq \frac{2\gamma_2^2}{\mu}\|y^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1})) - y^*(\bar{\hat{x}}_t,\hat{y}^+(\bar{\hat{x}}_{t+1}))\|^2 + \frac{2\gamma_2^2}{\mu}\|y^*(\bar{\hat{x}}_t,\hat{y}^+(\bar{\hat{x}}_{t+1})) - y^*(\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2 \\
&\quad + (1-\hat{\beta}_y\eta)^2\frac{\gamma_2^2}{\mu}\|\bar{\hat{y}}_t - y^*(\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2
\end{aligned}$$

$$\leq \frac{2\gamma_2^2 C_{y^1_{\hat{x}\hat{y}}}^2}{\mu}\|\bar{\hat{x}}_{t+1}-\bar{\hat{x}}_t\|^2+\frac{2\gamma_2^2 C_{y^2_{\hat{x}\hat{y}}}^2}{\mu}\|\hat{y}^+(\bar{\hat{x}}_{t+1})-\bar{\hat{y}}_t\|^2+\frac{\gamma_2^2}{\mu}\frac{(1-\hat{\beta}_y\eta)^2}{\hat{\beta}_y^2\eta^2}\|\hat{y}^+(\bar{\hat{x}}_{t+1})-\bar{\hat{y}}_t\|^2$$
$$\leq \frac{2\gamma_2^2 C_{y^1_{\hat{x}\hat{y}}}^2}{\mu}\|\bar{\hat{x}}_{t+1}-\bar{\hat{x}}_t\|^2+\frac{2\gamma_2^2}{\mu}\left(C_{y^2_{\hat{x}\hat{y}}}^2+\frac{(1-\hat{\beta}_y\eta)^2}{\hat{\beta}_y^2\eta^2}\right)\|\bar{\hat{y}}_t-\hat{y}^+(\bar{\hat{x}}_{t+1})\|^2\,,$$

the second step and the last step hold due to the definition $\hat{y}^+(\bar{\hat{x}}_{t+1})=\bar{\hat{y}}_t+\hat{\beta}_y\eta(y^*(\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\bar{\hat{y}}_t)$, the second-to-last step follows from Lemma B.2. □

Lemma C.5. *Given Assumptions 3.1-3.4, the following inequality holds:*

$$\mathbb{E}[\|x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)-x^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1}))\|^2]\leq 10\beta_y^2\eta^2 C_{x^1_{y\hat{x}\hat{y}}}^2\mathbb{E}[\|\bar{v}_t-\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]$$
$$+10\beta_y^2\eta^2L^2C_{x^1_{y\hat{x}\hat{y}}}^2\mathbb{E}[\|\bar{x}_t-x^*(\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]+5C_{x^1_{y\hat{x}\hat{y}}}^2\left(1+\frac{1}{\beta_y^2\eta^2(\gamma_2-L)^2}\right)\mathbb{E}[\|y^+(\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\bar{y}_t\|^2]$$
$$+5C_{x^1_{y\hat{x}\hat{y}}}^2C_{y^1_{\hat{x}\hat{y}}}^2\mathbb{E}[\|\bar{\hat{x}}_t-\bar{\hat{x}}_{t+1}\|^2]+5C_{x^1_{y\hat{x}\hat{y}}}^2C_{y^2_{\hat{x}\hat{y}}}^2\mathbb{E}[\|\bar{\hat{y}}_t-\hat{y}^+(\bar{\hat{x}}_{t+1})\|^2]\,.$$

Proof. From Lemma B.2, we derive

$$\mathbb{E}[\|x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)-x^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1}))\|^2]\leq C_{x^1_{y\hat{x}\hat{y}}}^2\mathbb{E}[\|\bar{y}_{t+1}-y^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1}))\|^2]$$
$$=C_{x^1_{y\hat{x}\hat{y}}}^2\mathbb{E}[\|\bar{y}_{t+1}-y^+(\bar{\hat{x}}_t,\bar{\hat{y}}_t)+y^+(\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\bar{y}_t+\bar{y}_t-y^*(\bar{\hat{x}}_t,\bar{\hat{y}}_t)$$
$$+y^*(\bar{\hat{x}}_t,\bar{\hat{y}}_t)-y^*(\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)+y^*(\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)-y^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1}))\|^2]$$
$$\leq 5C_{x^1_{y\hat{x}\hat{y}}}^2\mathbb{E}[\|\bar{y}_{t+1}-y^+(\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]+5C_{x^1_{y\hat{x}\hat{y}}}^2\mathbb{E}[\|y^+(\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\bar{y}_t\|^2]+5C_{x^1_{y\hat{x}\hat{y}}}^2\mathbb{E}[\|\bar{y}_t-y^*(\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]$$
$$+5C_{x^1_{y\hat{x}\hat{y}}}^2\mathbb{E}[\|y^*(\bar{\hat{x}}_t,\bar{\hat{y}}_t)-y^*(\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)\|^2]+5C_{x^1_{y\hat{x}\hat{y}}}^2\mathbb{E}[\|y^*(\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)-y^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1}))\|^2]$$
$$\leq 10\beta_y^2\eta^2C_{x^1_{y\hat{x}\hat{y}}}^2\mathbb{E}[\|\bar{v}_t-\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]+10\beta_y^2\eta^2L^2C_{x^1_{y\hat{x}\hat{y}}}^2\mathbb{E}[\|\bar{x}_t-x^*(\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]$$
$$+5C_{x^1_{y\hat{x}\hat{y}}}^2\left(1+\frac{1}{\beta_y^2\eta^2(\gamma_2-L)^2}\right)\mathbb{E}[\|y^+(\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\bar{y}_t\|^2]$$
$$+5C_{x^1_{y\hat{x}\hat{y}}}^2C_{y^1_{\hat{x}\hat{y}}}^2\mathbb{E}[\|\bar{\hat{x}}_t-\bar{\hat{x}}_{t+1}\|^2]+5C_{x^1_{y\hat{x}\hat{y}}}^2C_{y^2_{\hat{x}\hat{y}}}^2\mathbb{E}[\|\bar{\hat{y}}_t-\hat{y}^+(\bar{\hat{x}}_{t+1})\|^2]\,,$$

where the last step holds due to the following inequality:

$$\mathbb{E}[\|\bar{y}_{t+1}-y^+(\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]=\mathbb{E}[\|\bar{y}_t+\beta_y\eta\bar{v}_t-\bar{y}_t-\beta_y\eta\nabla_y F(x^*(\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t),\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]$$
$$=\beta_y^2\eta^2\mathbb{E}[\|\bar{v}_t-\nabla_y F(x^*(\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t),\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]$$
$$\leq 2\beta_y^2\eta^2\mathbb{E}[\|\bar{v}_t-\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]$$
$$+2\beta_y^2\eta^2\mathbb{E}[\|\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\nabla_y F(x^*(\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t),\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]$$
$$\leq 2\beta_y^2\eta^2\mathbb{E}[\|\bar{v}_t-\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]+2\beta_y^2\eta^2L^2\mathbb{E}[\|\bar{x}_t-x^*(\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]\,.$$

□

Lemma C.6. *Given Assumptions 3.1-3.4, the following inequalities hold:*

$$\mathbb{E}[\|\bar{\hat{y}}_t-\hat{y}^+(\bar{\hat{x}}_{t+1})\|^2]\leq 2\mathbb{E}[\|\bar{\hat{y}}_{t+1}-\bar{\hat{y}}_t\|^2]+4\hat{\beta}_y^2\eta^2\beta_y^2\eta^2\mathbb{E}[\|\bar{v}_t\|^2]$$
$$+\frac{4\hat{\beta}_y^2}{\beta_y^2(\gamma_2-L)^2}\mathbb{E}[\|\bar{y}_t-y^+(\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]\,.$$
$$\mathbb{E}[\|\bar{y}_t-y^+(\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]\leq 4\beta_y^2\eta^2L^2\mathbb{E}[\|x^*(\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\bar{x}_t\|^2]+4\beta_y^2\eta^2\mathbb{E}[\|\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\bar{v}_t\|^2]$$
$$+2\beta_y^2\eta^2\mathbb{E}[\|\bar{v}_t\|^2]\,.$$

Proof. From the definition of $\hat{y}^+(\bar{\hat{x}}_{t+1})$, we derive

$$\frac{1}{2}\mathbb{E}[\|\bar{\hat{y}}_t-\hat{y}^+(\bar{\hat{x}}_{t+1})\|^2]\leq\mathbb{E}[\|\bar{\hat{y}}_{t+1}-\bar{\hat{y}}_t\|^2]+\mathbb{E}[\|\bar{\hat{y}}_{t+1}-\hat{y}^+(\bar{\hat{x}}_{t+1})\|^2]$$
$$\leq\mathbb{E}[\|\bar{\hat{y}}_{t+1}-\bar{\hat{y}}_t\|^2]+\mathbb{E}[\|\bar{\hat{y}}_t+\hat{\beta}_y\eta(\bar{y}_{t+1}-\bar{\hat{y}}_t)-\bar{\hat{y}}_t-\hat{\beta}_y\eta(y^*(\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\bar{\hat{y}}_t)\|^2]$$

$$
\begin{aligned}
&= \mathbb{E}[\|\bar{\hat{y}}_{t+1} - \bar{\hat{y}}_t\|^2] + \hat{\beta}_y^2\eta^2\mathbb{E}[\|\bar{y}_{t+1} - y^*(\bar{\hat{x}}_t, \bar{\hat{y}}_t)\|^2] \\
&\leq \mathbb{E}[\|\bar{\hat{y}}_{t+1} - \bar{\hat{y}}_t\|^2] + 2\hat{\beta}_y^2\eta^2\mathbb{E}[\|\bar{y}_{t+1} - \bar{y}_t\|^2] + 2\hat{\beta}_y^2\eta^2\mathbb{E}[\|\bar{y}_t - y^*(\bar{\hat{x}}_t, \bar{\hat{y}}_t)\|^2] \\
&\leq \mathbb{E}[\|\bar{\hat{y}}_{t+1} - \bar{\hat{y}}_t\|^2] + 2\hat{\beta}_y^2\eta^2\beta_y^2\eta^2\mathbb{E}[\|\bar{v}_t\|^2] + \frac{2\hat{\beta}_y^2}{\beta_y^2(\gamma_2 - L)^2}\mathbb{E}[\|\bar{y}_t - y^+(\bar{\hat{x}}_t, \bar{\hat{y}}_t)\|^2] \,.
\end{aligned}
$$

Similarly, we derive

$$
\begin{aligned}
&\mathbb{E}[\|y^+(\bar{\hat{x}}_t, \bar{\hat{y}}_t) - \bar{y}_t\|^2] \leq 2\mathbb{E}[\|y^+(\bar{\hat{x}}_t, \bar{\hat{y}}_t) - \bar{y}_{t+1}\|^2] + 2\mathbb{E}[\|\bar{y}_{t+1} - \bar{y}_t\|^2] \\
&= 2\mathbb{E}[\|\bar{y}_t + \beta_y\eta\nabla_y F_d(\bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \bar{y}_t - \beta_y\eta\bar{v}_t\|^2] + 2\beta_y^2\eta^2\mathbb{E}[\|\bar{v}_t\|^2] \\
&= 2\beta_y^2\eta^2\mathbb{E}[\|\nabla_y F_d(\bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \bar{v}_t\|^2] + 2\beta_y^2\eta^2\mathbb{E}[\|\bar{v}_t\|^2] \\
&\leq 4\beta_y^2\eta^2\mathbb{E}[\|\nabla_y F_d(\bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \nabla_y F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)\|^2] + 4\beta_y^2\eta^2\mathbb{E}[\|\nabla_y F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \bar{v}_t\|^2] \\
&\quad + 2\beta_y^2\eta^2\mathbb{E}[\|\bar{v}_t\|^2] \\
&\leq 4\beta_y^2\eta^2L^2\mathbb{E}[\|x^*(\bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \bar{x}_t\|^2] + 4\beta_y^2\eta^2\mathbb{E}[\|\nabla_y F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \bar{v}_t\|^2] + 2\beta_y^2\eta^2\mathbb{E}[\|\bar{v}_t\|^2] \,.
\end{aligned}
$$

□

Lemma C.7. *Given Assumptions 3.1-3.4, by defining*

$$
\mathcal{P}_t = \mathbb{E}[F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)] - 2\mathbb{E}[F_d(\bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)] + 2\mathbb{E}[q(\bar{\hat{x}}_t)] \,, \tag{25}
$$

by setting $\eta \leq \frac{1}{\hat{\beta}_x}$, $\eta \leq \frac{1}{\hat{\beta}_y}$, *and* $\beta_x \leq \min\{\frac{L^2}{120\gamma_1^3}, \frac{\sqrt{\mu(\gamma_1 - L)^3}(\gamma_2 - L)^2}{512\sqrt{6\gamma_1 c_{\hat{\beta}_x}}\gamma_2 c_{\hat{\beta}_y}}\}$, *then the following inequality holds:*

$$
\begin{aligned}
&\mathcal{P}_{t+1} - \mathcal{P}_t \leq -\frac{\beta_x\eta}{4}\mathbb{E}[\|\nabla_x F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)\|^2] - \frac{\beta_y\eta}{2}\mathbb{E}[\|\nabla_y F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)\|^2] \\
&+ \frac{\beta_x\eta}{2}\mathbb{E}[\|\nabla_x F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \bar{u}_t\|^2] + A_3\mathbb{E}[\|\nabla_y F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \bar{v}_t\|^2] \\
&+ \left(4\beta_y\eta\beta_x^2\eta^2L^2 - \frac{\beta_x\eta}{4}\right)\mathbb{E}[\|\bar{u}_t\|^2] \\
&+ \left(\beta_y^2\eta^2L_d + \frac{3\beta_y\eta}{4} + \frac{\beta_y^2\eta^2(\gamma_2 + L)}{2} + 4A_1\hat{\beta}_y^2\eta^2\beta_y^2\eta^2 + 2A_2\beta_y^2\eta^2 - \frac{7}{8}\beta_y\eta\right)\mathbb{E}[\|\bar{v}_t\|^2] \\
&+ \left(2\gamma_1 C_{x^1_{\hat{x}\hat{y}}} + \frac{\gamma_1}{6\hat{\beta}_x\eta} + 6\gamma_1\hat{\beta}_x\eta\left(10C^2_{x^1_{y\hat{x}\hat{y}}}C^2_{y^1_{\hat{x}\hat{y}}} + \frac{4}{\gamma_1 - L}\frac{2\gamma_2^2C^2_{y^1_{\hat{x}\hat{y}}}}{\mu}\right) - \frac{\gamma_1(2 - \hat{\beta}_x\eta)}{2\hat{\beta}_x\eta}\right)\mathbb{E}[\|\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t\|^2] \\
&+ \left(2A_1 - \frac{\gamma_2(2 - \hat{\beta}_y\eta)}{2\hat{\beta}_y\eta}\right)\mathbb{E}[\|\bar{\hat{y}}_{t+1} - \bar{\hat{y}}_t\|^2] \,,
\end{aligned}
$$

where

$$
\begin{aligned}
A_1 &= 6\gamma_1\hat{\beta}_x\eta\left(10C^2_{x^1_{y\hat{x}\hat{y}}}C^2_{y^2_{\hat{x}\hat{y}}} + \frac{4}{\gamma_1 - L}\frac{2\gamma_2^2}{\mu}\left(C^2_{y^2_{\hat{x}\hat{y}}} + \frac{(1 - \hat{\beta}_y\eta)^2}{\hat{\beta}_y^2\eta^2}\right)\right) \,, \\
A_2 &= 60\gamma_1\hat{\beta}_x\eta C^2_{x^1_{y\hat{x}\hat{y}}}\left(1 + \frac{1}{\beta_y^2\eta^2(\gamma_2 - L)^2}\right) + A_1\frac{4\hat{\beta}_y^2}{\beta_y^2(\gamma_2 - L)^2} \,, \\
A_3 &= \beta_y\eta + 120\gamma_1\hat{\beta}_x\eta\beta_y^2\eta^2C^2_{x^1_{y\hat{x}\hat{y}}} + 4A_2\beta_y^2\eta^2 \,.
\end{aligned}
$$

and

$$
\begin{aligned}
&\beta_y = \beta_x\underbrace{\frac{(\gamma_1 - L)^2}{64L^2}}_{c_{\beta_y} = O(1)} \,, \qquad \hat{\beta}_x = \beta_x\underbrace{\frac{(\gamma_1 - L)^4(\gamma_2 - L)^2\mu}{24 \times 64^2\gamma_1L^2\left(5\gamma_1^2\mu + 16\gamma_2^2(\gamma_1 - L)\right)}}_{c_{\hat{\beta}_x} = O(1/\kappa)} \,, \\
&\hat{\beta}_y = \beta_x\underbrace{\frac{(\gamma_1 - L)^4(\gamma_2 - L)^4}{64^2 \times 480\gamma_1^3\gamma_2^2L^2}}_{c_{\hat{\beta}_y} = O(1)} \,.
\end{aligned} \tag{26}
$$

Proof. Based on Lemmas C.1, C.2, C.3, we have

$$\begin{aligned}
&\mathcal{P}_{t+1} - \mathcal{P}_t \leq -\frac{\beta_x\eta}{2}\mathbb{E}[\|\nabla_x F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)\|^2] + \frac{\beta_y\eta}{2}\mathbb{E}[\|\nabla_y F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)\|^2] \\
&+ \frac{\beta_x\eta}{2}\mathbb{E}[\|\nabla_x F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \bar{u}_t\|^2] + \left(4\beta_y\eta\beta_x^2\eta^2L^2 - \frac{\beta_x\eta}{4}\right)\mathbb{E}[\|\bar{u}_t\|^2] \\
&+ \left(\beta_y^2\eta^2L_d + \frac{3\beta_y\eta}{4} + \frac{\beta_y^2\eta^2(\gamma_2+L)}{2}\right)\mathbb{E}[\|\bar{v}_t\|^2] + \left(-\frac{\gamma_1(2-\hat{\beta}_x\eta)}{2\hat{\beta}_x\eta}\right)\mathbb{E}[\|\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t\|^2] \\
&+ \left(-\frac{\gamma_2(2-\hat{\beta}_y\eta)}{2\hat{\beta}_y\eta}\right)\mathbb{E}[\|\bar{\hat{y}}_{t+1} - \bar{\hat{y}}_t\|^2] - 2\beta_y\eta\mathbb{E}[\langle\nabla_y F_d(\bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t), \bar{v}_t\rangle] \\
&+ 2\gamma_1\mathbb{E}[\langle\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t, x^*(\bar{y}_{t+1}; \bar{\hat{x}}_{t+1}, \bar{\hat{y}}_t) - x^*(\bar{\hat{x}}_t, \hat{y}^*(\bar{\hat{x}}_{t+1}))\rangle] \,. \qquad (27)
\end{aligned}$$

For $-2\beta_y\eta\mathbb{E}[\langle\nabla_y F_d(\bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t), \bar{v}_t\rangle]$, we have

$$\begin{aligned}
&-2\beta_y\eta\mathbb{E}[\langle\nabla_y F_d(\bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t), \bar{v}_t\rangle] \\
&= -2\beta_y\eta\mathbb{E}[\langle\nabla_y F_d(\bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \nabla_y F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t), \bar{v}_t\rangle] - 2\beta_y\eta\mathbb{E}[\langle\nabla_y F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t), \bar{v}_t\rangle] \\
&= -2\beta_y\eta\mathbb{E}[\langle\nabla_y F_d(\bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \nabla_y F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t), \bar{v}_t\rangle] \\
&\quad - \beta_y\eta\mathbb{E}[\|\nabla_y F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)\|^2] - \beta_y\eta\mathbb{E}[\|\bar{v}_t\|^2] + \beta_y\eta\mathbb{E}[\|\nabla_y F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \bar{v}_t\|^2] \,.
\end{aligned}$$

For the first term, with Young's inequality $2a^Tb \leq \frac{1}{\nu}\|a\|^2 + \nu\|b\|^2$ with $\nu > 0$ being a constant, we derive

$$\begin{aligned}
&-2\mathbb{E}[\langle\nabla_y F_d(\bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \nabla_y F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t), \bar{v}_t\rangle] \\
&\leq \frac{1}{\nu}\mathbb{E}[\|\nabla_y F_d(\bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \nabla_y F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)\|^2] + \nu\mathbb{E}[\|\bar{v}_t\|^2] \\
&= \frac{1}{\nu}\mathbb{E}[\|\nabla_y F(x^*(\bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t), \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \nabla_y F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)\|^2] + \nu\mathbb{E}[\|\bar{v}_t\|^2] \\
&\leq L^2\frac{1}{\nu}\mathbb{E}[\|x^*(\bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \bar{x}_t\|^2] + \nu\mathbb{E}[\|\bar{v}_t\|^2] \,.
\end{aligned}$$

Combining these inequalities yields

$$\begin{aligned}
&-2\mathbb{E}[\langle\nabla_y F_d(\bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t), \bar{v}_t\rangle] \leq L^2\frac{1}{\nu}\mathbb{E}[\|x^*(\bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \bar{x}_t\|^2] - \mathbb{E}[\|\nabla_y F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)\|^2] \\
&- (1-\nu)\mathbb{E}[\|\bar{v}_t\|^2] + \mathbb{E}[\|\nabla_y F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t) - \bar{v}_t\|^2] \,. \qquad (28)
\end{aligned}$$

For the last term in Eq. (27), applying Young's inequality $2a^Tb \leq \frac{1}{\nu}\|a\|^2 + \nu\|b\|^2$ with $\nu = 6\hat{\beta}_x\eta$ and $a^Tb \leq \|a\|\|b\|$, we have

$$\begin{aligned}
&2\gamma_1\mathbb{E}[\langle\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t, x^*(\bar{y}_{t+1}; \bar{\hat{x}}_{t+1}, \bar{\hat{y}}_t) - x^*(\bar{\hat{x}}_t, \hat{y}^*(\bar{\hat{x}}_{t+1}))\rangle] \\
&= 2\gamma_1\mathbb{E}[\langle\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t, x^*(\bar{y}_{t+1}; \bar{\hat{x}}_{t+1}, \bar{\hat{y}}_t) - x^*(\bar{\hat{x}}_{t+1}, \hat{y}^*(\bar{\hat{x}}_{t+1}))\rangle] \\
&\quad + 2\gamma_1\mathbb{E}[\langle\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t, x^*(\bar{\hat{x}}_{t+1}, \hat{y}^*(\bar{\hat{x}}_{t+1})) - x^*(\bar{\hat{x}}_t, \hat{y}^*(\bar{\hat{x}}_{t+1}))\rangle] \\
&\leq \frac{\gamma_1}{6\hat{\beta}_x\eta}\mathbb{E}[\|\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t\|^2] + 6\gamma_1\hat{\beta}_x\eta\mathbb{E}[\|x^*(\bar{y}_{t+1}; \bar{\hat{x}}_{t+1}, \bar{\hat{y}}_t) - x^*(\bar{\hat{x}}_{t+1}, \hat{y}^*(\bar{\hat{x}}_{t+1}))\|^2] \qquad (29) \\
&\quad + 2\gamma_1\mathbb{E}[\|\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t\|\|x^*(\bar{\hat{x}}_{t+1}, \hat{y}^*(\bar{\hat{x}}_{t+1})) - x^*(\bar{\hat{x}}_t, \hat{y}^*(\bar{\hat{x}}_{t+1}))\|] \\
&\leq \frac{\gamma_1}{6\hat{\beta}_x\eta}\mathbb{E}[\|\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t\|^2] + 6\gamma_1\hat{\beta}_x\eta\mathbb{E}[\|x^*(\bar{y}_{t+1}; \bar{\hat{x}}_{t+1}, \bar{\hat{y}}_t) - x^*(\bar{\hat{x}}_{t+1}, \hat{y}^*(\bar{\hat{x}}_{t+1}))\|^2] \\
&\quad + 2\gamma_1 C_{x^1_{\hat{x}\hat{y}}}\mathbb{E}[\|\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t\|^2] \,,
\end{aligned}$$

where the last step holds due to Lemma B.2.

Then, by plugging Eq. (28) and Eq. (29) into Eq. (27) with $\nu = \frac{1}{8}$, we have

$$\mathcal{P}_{t+1} - \mathcal{P}_t \leq -\frac{\beta_x\eta}{2}\mathbb{E}[\|\nabla_x F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)\|^2] - \frac{\beta_y\eta}{2}\mathbb{E}[\|\nabla_y F(\bar{x}_t, \bar{y}_t; \bar{\hat{x}}_t, \bar{\hat{y}}_t)\|^2]$$

$$+\frac{\beta_x\eta}{2}\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\bar{u}_t\|^2]+\beta_y\eta\mathbb{E}[\|\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\bar{v}_t\|^2] \tag{30}$$
$$+\left(4\beta_y\eta\beta_x^2\eta^2L^2-\frac{\beta_x\eta}{4}\right)\mathbb{E}[\|\bar{u}_t\|^2]+\left(\beta_y^2\eta^2L_d+\frac{3\beta_y\eta}{4}+\frac{\beta_y^2\eta^2(\gamma_2+L)}{2}-\frac{7}{8}\beta_y\eta\right)\mathbb{E}[\|\bar{v}_t\|^2]$$
$$+\left(2\gamma_1C_{x^1_{\hat{x}\hat{y}}}+\frac{\gamma_1}{6\hat{\beta}_x\eta}-\frac{\gamma_1(2-\hat{\beta}_x\eta)}{2\hat{\beta}_x\eta}\right)\mathbb{E}[\|\bar{\hat{x}}_{t+1}-\bar{\hat{x}}_t\|^2]+\left(-\frac{\gamma_2(2-\hat{\beta}_y\eta)}{2\hat{\beta}_y\eta}\right)\mathbb{E}[\|\bar{\hat{y}}_{t+1}-\bar{\hat{y}}_t\|^2]$$
$$+8\beta_y\eta L^2\mathbb{E}[\|x^*(\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\bar{x}_t\|^2]+6\gamma_1\hat{\beta}_x\eta\mathbb{E}[\|x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)-x^*(\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1}))\|^2]\,.$$

For the last term in Eq. (30), applying Lemma C.4 and Lemma C.5, we have

$$\begin{aligned}
&\mathbb{E}[\|x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)-x^*(\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1}))\|^2]\\
&\le 2\mathbb{E}[\|x^*(\bar{y}_{t+1};\bar{\hat{x}}_{t+1},\bar{\hat{y}}_t)-x^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1}))\|^2]+2\mathbb{E}[\|x^*(\bar{\hat{x}}_{t+1},\hat{y}^+(\bar{\hat{x}}_{t+1}))-x^*(\bar{\hat{x}}_{t+1},\hat{y}^*(\bar{\hat{x}}_{t+1}))\|^2]\\
&\le 20\beta_y^2\eta^2C^2_{x^1_{y\hat{x}\hat{y}}}\mathbb{E}[\|\bar{v}_t-\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]+20\beta_y^2\eta^2L^2C^2_{x^1_{y\hat{x}\hat{y}}}\mathbb{E}[\|\bar{x}_t-x^*(\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]\\
&\quad+10C^2_{x^1_{y\hat{x}\hat{y}}}\left(1+\frac{1}{\beta_y^2\eta^2(\gamma_2-L)^2}\right)\mathbb{E}[\|y^+(\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\bar{y}_t\|^2]\\
&\quad+\Big(10C^2_{x^1_{y\hat{x}\hat{y}}}C^2_{y^1_{\hat{x}\hat{y}}}+\frac{4}{\gamma_1-L}\frac{2\gamma_2^2C^2_{y^1_{\hat{x}\hat{y}}}}{\mu}\Big)\mathbb{E}[\|\bar{\hat{x}}_t-\bar{\hat{x}}_{t+1}\|^2]\\
&\quad+\left(10C^2_{x^1_{y\hat{x}\hat{y}}}C^2_{y^2_{\hat{x}\hat{y}}}+\frac{4}{\gamma_1-L}\frac{2\gamma_2^2}{\mu}\left(C^2_{y^2_{\hat{x}\hat{y}}}+\frac{(1-\hat{\beta}_y\eta)^2}{\hat{\beta}_y^2\eta^2}\right)\right)\mathbb{E}[\|\bar{\hat{y}}_t-\hat{y}^+(\bar{\hat{x}}_{t+1})\|^2]\,.
\end{aligned}\tag{31}$$

Combining Eq. (31) with the inequalities in Lemma C.6, and Lemma B.5, and collecting like terms, we obtain

$$\begin{aligned}
\mathcal{P}_{t+1}-\mathcal{P}_t&\le\left(\frac{(A_3+7\beta_y\eta)L^2}{(\gamma_1-L)^2}-\frac{\beta_x\eta}{2}\right)\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]-\frac{\beta_y\eta}{2}\mathbb{E}[\|\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]\\
&+\frac{\beta_x\eta}{2}\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\bar{u}_t\|^2]+A_3\mathbb{E}[\|\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\bar{v}_t\|^2]+\left(4\beta_y\eta\beta_x^2\eta^2L^2-\frac{\beta_x\eta}{4}\right)\mathbb{E}[\|\bar{u}_t\|^2]\\
&+\left(\beta_y^2\eta^2L_d+\frac{3\beta_y\eta}{4}+\frac{\beta_y^2\eta^2(\gamma_2+L)}{2}+4A_1\hat{\beta}_y^2\eta^2\beta_y^2\eta^2+2A_2\beta_y^2\eta^2-\frac{7}{8}\beta_y\eta\right)\mathbb{E}[\|\bar{v}_t\|^2]\\
&+\left(2\gamma_1C_{x^1_{\hat{x}\hat{y}}}+\frac{\gamma_1}{6\hat{\beta}_x\eta}+6\gamma_1\hat{\beta}_x\eta\Big(10C^2_{x^1_{y\hat{x}\hat{y}}}C^2_{y^1_{\hat{x}\hat{y}}}+\frac{4}{\gamma_1-L}\frac{2\gamma_2^2C^2_{y^1_{\hat{x}\hat{y}}}}{\mu}\Big)-\frac{\gamma_1(2-\hat{\beta}_x\eta)}{2\hat{\beta}_x\eta}\right)\mathbb{E}[\|\bar{\hat{x}}_{t+1}-\bar{\hat{x}}_t\|^2]\\
&+\Big(2A_1-\frac{\gamma_2(2-\hat{\beta}_y\eta)}{2\hat{\beta}_y\eta}\Big)\mathbb{E}[\|\bar{\hat{y}}_{t+1}-\bar{\hat{y}}_t\|^2]\,,
\end{aligned}\tag{32}$$

where

$$\begin{aligned}
A_1&=6\gamma_1\hat{\beta}_x\eta\left(10C^2_{x^1_{y\hat{x}\hat{y}}}C^2_{y^2_{\hat{x}\hat{y}}}+\frac{4}{\gamma_1-L}\frac{2\gamma_2^2}{\mu}\Big(C^2_{y^2_{\hat{x}\hat{y}}}+\frac{(1-\hat{\beta}_y\eta)^2}{\hat{\beta}_y^2\eta^2}\Big)\right)\,,\\
A_2&=60\gamma_1\hat{\beta}_x\eta C^2_{x^1_{y\hat{x}\hat{y}}}\left(1+\frac{1}{\beta_y^2\eta^2(\gamma_2-L)^2}\right)+A_1\frac{4\hat{\beta}_y^2}{\beta_y^2(\gamma_2-L)^2}\,,\\
A_3&=\beta_y\eta+120\gamma_1\hat{\beta}_x\eta\beta_y^2\eta^2C^2_{x^1_{y\hat{x}\hat{y}}}+4A_2\beta_y^2\eta^2\,.
\end{aligned}$$

Then, for $\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]$, we set

$$\begin{aligned}
&\frac{(A_3+7\beta_y\eta)L^2}{(\gamma_1-L)^2}-\frac{\beta_x\eta}{2}\\
&=\frac{1}{(\gamma_1-L)^2}\left(8\beta_y\eta L^2+120\gamma_1\hat{\beta}_x\eta\beta_y^2\eta^2L^2C^2_{x^1_{y\hat{x}\hat{y}}}+4A_2\beta_y^2\eta^2L^2\right)-\frac{\beta_x\eta}{2}\le-\frac{\beta_x\eta}{4}\,.
\end{aligned}\tag{33}$$

Specifically, we enforce

$$\frac{8\beta_y\eta L^2}{(\gamma_1-L)^2}\le\frac{\beta_x\eta}{8}\,,\qquad\frac{4\beta_y^2\eta^2L^2}{(\gamma_1-L)^2}A_2\le\frac{\beta_x\eta}{32\times16}$$

$$\frac{120\gamma_1\hat{\beta}_x\eta\beta_y^2\eta^2L^2C_{x^1_{y\hat{x}\hat{y}}}^2}{(\gamma_1-L)^2} \leq \frac{\beta_x\eta}{32\times 16} \,. \tag{34}$$

For the first inequality, we set $\beta_y = \beta_x \underbrace{\frac{(\gamma_1-L)^2}{64L^2}}_{c_{\beta_y}=O(1)}$.

For the second inequality in Eq. (34), from the definition of A_1 and A_2, we enforce

$$\begin{aligned}
&\frac{4\beta_y^2\eta^2L^2}{(\gamma_1-L)^2}60\gamma_1\hat{\beta}_x\eta C_{x^1_{y\hat{x}\hat{y}}}^2 \leq \frac{\beta_x\eta}{32\times 64} \,,\\
&\frac{4L^2}{(\gamma_1-L)^2}\left(\frac{60\gamma_1\hat{\beta}_x\eta C_{x^1_{y\hat{x}\hat{y}}}^2}{(\gamma_2-L)^2} + \frac{4\hat{\beta}_y^2\eta^2}{(\gamma_2-L)^2}\frac{24\gamma_1\hat{\beta}_x\eta}{\gamma_1-L}\frac{2\gamma_2^2}{\mu}\frac{(1-\hat{\beta}_y\eta)^2}{\hat{\beta}_y^2\eta^2}\right) \leq \frac{\beta_x\eta}{32\times 64} \,,\\
&\frac{4L^2}{(\gamma_1-L)^2}60\gamma_1\hat{\beta}_x\eta C_{x^1_{y\hat{x}\hat{y}}}^2 C_{y^2_{\hat{x}\hat{y}}}^2\frac{4\hat{\beta}_y^2\eta^2}{(\gamma_2-L)^2} \leq \frac{\beta_x\eta}{32\times 64} \,,\\
&\frac{4L^2}{(\gamma_1-L)^2}6\gamma_1\hat{\beta}_x\eta\frac{4}{\gamma_1-L}\frac{2\gamma_2^2}{\mu}C_{y^2_{\hat{x}\hat{y}}}^2\frac{4\hat{\beta}_y^2\eta^2}{(\gamma_2-L)^2} \leq \frac{\beta_x\eta}{32\times 64} \,.
\end{aligned} \tag{35}$$

To solve the first inequality in Eq. (35), since $\hat{\beta}_x\eta \leq 1$ and $\eta < 1$, from $C_{x^1_{y\hat{x}\hat{y}}} = \frac{\gamma_1}{\gamma_1-L}$, we obtain $\beta_x \leq \frac{L^2}{120\gamma_1^3}$. Here, we have also shown that the last inequality in Eq. (34) holds.

Then, to address the second inequality in Eq. (35), note that since $\hat{\beta}_y\eta \leq 1$, it follows that $1-\hat{\beta}_y\eta \leq 1$. Consequently, we obtain

$$\hat{\beta}_x = \beta_x \underbrace{\frac{(\gamma_1-L)^4(\gamma_2-L)^2\mu}{24\times 64^2\gamma_1L^2\Big(5\gamma_1^2\mu+16\gamma_2^2(\gamma_1-L)\Big)}}_{c_{\hat{\beta}_x}=O(1/\kappa)} \,. \tag{36}$$

Similarly, for the third inequality in Eq. (35), from $C_{y^2_{\hat{x}\hat{y}}} = \frac{\gamma_2}{\gamma_2-L}$, we obtain

$$\hat{\beta}_y = \beta_x \underbrace{\frac{(\gamma_1-L)^4(\gamma_2-L)^4}{64^2\times 480\gamma_1^3\gamma_2^2L^2}}_{c_{\hat{\beta}_y}=O(1)} \,.$$

Moreover, to solve the last inequality in Eq. (35), we obtain

$$\beta_x \leq \frac{\sqrt{\mu(\gamma_1-L)^3}(\gamma_2-L)^2}{512\sqrt{6\gamma_1c_{\hat{\beta}_x}}\gamma_2c_{\hat{\beta}_y}} = O(1) \,.$$

Finally, by plugging Eq. (33) into Eq. (32), the proof is complete. □

D Key Lemmas Related to the Decentralized Setting

D.1 Consensus Errors

Lemma D.1. *Given Assumptions 3.1-3.4, the following inequality holds:*

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{p}_{t+1}-p_{t+1}^{(k)}\|^2] \leq \lambda\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\bar{p}_t-p_t^{(k)}\|^2] + 3\rho_x^2\eta^4\frac{1}{1-\lambda}\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\hat{\mathcal{G}}_{x,t}^{(k)}\|^2]\\
&+\frac{9(L^2+\gamma_1^2)}{1-\lambda}\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|x_{t+1}^{(k)}-x_t^{(k)}\|^2] + \frac{9L^2}{1-\lambda}\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|y_{t+1}^{(k)}-y_t^{(k)}\|^2]\\
&+\frac{9\gamma_1^2}{1-\lambda}\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\|\hat{x}_{t+1}^{(k)}-\hat{x}_t^{(k)}\|^2] + 3\rho_x^2\eta^4\sigma^2\frac{1}{1-\lambda} \,.
\end{aligned}$$

Proof. From the update rule of $\bar{p}_{t+1}$, we derive

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_{t+1}-p_{t+1}^{(k)}\|^2]=\frac{1}{K}\mathbb{E}[\|\bar{P}_{t+1}-P_{t+1}\|_F^2]\\
&=\frac{1}{K}\mathbb{E}[\|\bar{P}_t-\bar{U}_t+\bar{U}_{t+1}-P_tW+U_t-U_{t+1}\|_F^2]\\
&\le(1+a)\frac{1}{K}\mathbb{E}[\|\bar{P}_t-P_tW\|_F^2]+(1+1/a)\frac{1}{K}\mathbb{E}[\|-\bar{U}_t+\bar{U}_{t+1}+U_t-U_{t+1}\|_F^2]\\
&\le(1+a)\lambda^2\frac{1}{K}\mathbb{E}[\|\bar{P}_t-P_t\|_F^2]+(1+1/a)\frac{1}{K}\mathbb{E}[\|U_t-U_{t+1}\|_F^2]\\
&\le\lambda\frac{1}{K}\mathbb{E}[\|\bar{P}_t-P_t\|_F^2]+\frac{1}{1-\lambda}\frac{1}{K}\mathbb{E}[\|U_t-U_{t+1}\|_F^2]\,,
\end{aligned}\tag{37}$$

where the second step follows from Young's inequality, and the last step follows by setting $a=\frac{1-\lambda}{\lambda}$. In addition, for the second term, we derive:

$$\begin{aligned}
&\frac{1}{K}\mathbb{E}[\|U_t-U_{t+1}\|_F^2]=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_{t+1}^{(k)}-u_t^{(k)}\|^2]\\
&=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|(1-\rho_x\eta^2)(u_t^{(k)}-\nabla_xF^{(k)}(x_t^{(k)},y_t^{(k)};\hat{x}_t^{(k)},\hat{y}_t^{(k)};\xi_{t+1}^{(k)}))\\
&\qquad+\nabla_xF^{(k)}(x_{t+1}^{(k)},y_{t+1}^{(k)};\hat{x}_{t+1}^{(k)},\hat{y}_{t+1}^{(k)};\xi_{t+1}^{(k)})-u_t^{(k)}\|^2]\\
&\le3\rho_x^2\eta^4\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|u_t^{(k)}-\nabla_xF^{(k)}(x_t^{(k)},y_t^{(k)};\hat{x}_t^{(k)},\hat{y}_t^{(k)})\|^2]+3\rho_x^2\eta^4\sigma^2\\
&\quad+3\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|-\nabla_xF^{(k)}(x_t^{(k)},y_t^{(k)};\hat{x}_t^{(k)},\hat{y}_t^{(k)};\xi_{t+1}^{(k)})+\nabla_xF^{(k)}(x_{t+1}^{(k)},y_{t+1}^{(k)};\hat{x}_{t+1}^{(k)},\hat{y}_{t+1}^{(k)};\xi_{t+1}^{(k)})\|^2]\\
&\le3\rho_x^2\eta^4\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\hat{\mathcal{G}}_{x,t}^{(k)}\|^2]+3\rho_x^2\eta^4\sigma^2+9(L^2+\gamma_1^2)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)}-x_t^{(k)}\|^2]\\
&\quad+9L^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_{t+1}^{(k)}-y_t^{(k)}\|^2]+9\gamma_1^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\hat{x}_{t+1}^{(k)}-\hat{x}_t^{(k)}\|^2]\,,
\end{aligned}$$

where the last step holds due to the following inequality:

$$\begin{aligned}
&\mathbb{E}[\|-\nabla_xF^{(k)}(x_t^{(k)},y_t^{(k)};\hat{x}_t^{(k)},\hat{y}_t^{(k)};\xi_{t+1}^{(k)})+\nabla_xF^{(k)}(x_{t+1}^{(k)},y_{t+1}^{(k)};\hat{x}_{t+1}^{(k)},\hat{y}_{t+1}^{(k)};\xi_{t+1}^{(k)})\|^2]\\
&=\mathbb{E}[\|-\nabla_xf^{(k)}(x_t^{(k)},y_t^{(k)};\xi_{t+1}^{(k)})-\gamma_1(x_t^{(k)}-\hat{x}_t^{(k)})\\
&\quad+\nabla_xf^{(k)}(x_{t+1}^{(k)},y_{t+1}^{(k)};\xi_{t+1}^{(k)})+\gamma_1(x_{t+1}^{(k)}-\hat{x}_{t+1}^{(k)})\|^2]\\
&\le3\mathbb{E}[\|\nabla_xf^{(k)}(x_{t+1}^{(k)},y_{t+1}^{(k)};\xi_{t+1}^{(k)})-\nabla_xf^{(k)}(x_t^{(k)},y_t^{(k)};\xi_{t+1}^{(k)})\|^2]\\
&\quad+3\gamma_1^2\mathbb{E}[\|x_{t+1}^{(k)}-x_t^{(k)}\|^2]+3\gamma_1^2\mathbb{E}[\|\hat{x}_{t+1}^{(k)}-\hat{x}_t^{(k)}\|^2]\\
&\le3(L^2+\gamma_1^2)\mathbb{E}[\|x_{t+1}^{(k)}-x_t^{(k)}\|^2]+3L^2\mathbb{E}[\|y_{t+1}^{(k)}-y_t^{(k)}\|^2]+3\gamma_1^2\mathbb{E}[\|\hat{x}_{t+1}^{(k)}-\hat{x}_t^{(k)}\|^2]\,.
\end{aligned}$$

□

Lemma D.2. *Given Assumptions 3.1-3.4, the following inequality holds:*

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{q}_{t+1}-q_{t+1}^{(k)}\|^2]\le\lambda\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{q}_t-q_t^{(k)}\|^2]+3\rho_y^2\eta^4\frac{1}{1-\lambda}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\hat{\mathcal{G}}_{y,t}^{(k)}\|^2]\\
&+\frac{9L^2}{1-\lambda}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)}-x_t^{(k)}\|^2]+\frac{9(L^2+\gamma_2^2)}{1-\lambda}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_{t+1}^{(k)}-y_t^{(k)}\|^2]
\end{aligned}$$

$$+ \frac{9\gamma_2^2}{1-\lambda}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\hat{y}_{t+1}^{(k)} - \hat{y}_t^{(k)}\|^2] + 3\rho_y^2\eta^4\sigma^2\frac{1}{1-\lambda}\,.$$

This lemma can be proved by following Lemma D.1. Thus, we omit its proof.

Lemma D.3. *Given Assumptions 3.1-3.4, when $\hat{\beta}_x \leq \frac{1-\lambda}{4}$, the following inequality holds:*

$$\frac{1}{K}\mathbb{E}[\|\hat{X}_{t+1} - \bar{\hat{X}}_{t+1}\|_F^2] \leq \left(1 - \frac{\eta(1-\lambda^2)}{4}\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\hat{x}}_t - \hat{x}_t^{(k)}\|^2]$$
$$+ \frac{4\eta\hat{\beta}_x^2}{1-\lambda^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2] + \frac{4\eta\hat{\beta}_x^2}{1-\lambda^2}\frac{2\eta\beta_x^2}{1-\lambda^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t - p_t^{(k)}\|^2]\,.$$

Proof. From the update rule of $\hat{X}_{t+1}$, we derive

$$\begin{aligned}
&\|\hat{X}_{t+1} - \bar{\hat{X}}_{t+1}\|_F^2 = \|\hat{X}_t + \eta(\tilde{\hat{X}}_{t+1} - \hat{X}_t) - \bar{\hat{X}}_t - \eta\hat{\beta}_x(\bar{X}_{t+1} - \bar{\hat{X}}_t)\|_F^2\\
&= \|\hat{X}_t + \eta(\hat{X}_tW + \hat{\beta}_x(X_{t+1} - \hat{X}_t) - \hat{X}_t) - \bar{\hat{X}}_t - \eta\hat{\beta}_x(\bar{X}_{t+1} - \bar{\hat{X}}_t)\|_F^2\\
&= \|(1-\eta)(\hat{X}_t - \bar{\hat{X}}_t) + \eta(\hat{X}_tW - \bar{\hat{X}}_t) + \eta\hat{\beta}_x(X_{t+1} - \hat{X}_t) - \eta\hat{\beta}_x(\bar{X}_{t+1} - \bar{\hat{X}}_t)\|_F^2\\
&\leq (1-\eta)\|\hat{X}_t - \bar{\hat{X}}_t\|_F^2 + \eta\|\hat{X}_tW - \bar{\hat{X}}_t + \hat{\beta}_x(X_{t+1} - \hat{X}_t) - \hat{\beta}_x(\bar{X}_{t+1} - \bar{\hat{X}}_t)\|_F^2\\
&\leq (1-\eta)\|\hat{X}_t - \bar{\hat{X}}_t\|_F^2 + (1+c)\eta\lambda^2\|\hat{X}_t - \bar{\hat{X}}_t\|_F^2 + 2(1+1/c)\eta\hat{\beta}_x^2\|X_{t+1} - \bar{X}_{t+1}\|_F^2\\
&\quad + 2(1+1/c)\eta\hat{\beta}_x^2\|\hat{X}_t - \bar{\hat{X}}_t\|_F^2\\
&\leq \left(1 - \frac{\eta(1-\lambda^2)}{4}\right)\|\hat{X}_t - \bar{\hat{X}}_t\|_F^2 + \frac{4\eta\hat{\beta}_x^2}{1-\lambda^2}\|X_{t+1} - \bar{X}_{t+1}\|_F^2\\
&\leq \left(1 - \frac{\eta(1-\lambda^2)}{4}\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\hat{x}}_t - \hat{x}_t^{(k)}\|^2] + \frac{4\eta\hat{\beta}_x^2}{1-\lambda^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2]\\
&\quad + \frac{4\eta\hat{\beta}_x^2}{1-\lambda^2}\frac{2\eta\beta_x^2}{1-\lambda^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t - p_t^{(k)}\|^2]\,,
\end{aligned}$$

where the fourth step follows from Young's inequality, the second to last inequality holds due to $\hat{\beta}_x \leq \frac{1-\lambda}{4}$, with $c = \frac{1-\lambda^2}{2\lambda^2}$, and the last step holds due to Lemma D.5. □

Lemma D.4. *Given Assumptions 3.1-3.4, when $\beta_{\hat{y}} \leq \frac{1-\lambda}{4}$, the following inequality holds:*

$$\frac{1}{K}\mathbb{E}[\|\hat{Y}_{t+1} - \bar{\hat{Y}}_{t+1}\|_F^2] \leq \left(1 - \frac{\eta(1-\lambda^2)}{4}\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\hat{y}}_t - \hat{y}_t^{(k)}\|^2]$$
$$+ \frac{4\eta\hat{\beta}_y^2}{1-\lambda^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|^2] + \frac{4\eta\hat{\beta}_y^2}{1-\lambda^2}\frac{2\eta\beta_y^2}{1-\lambda^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{q}_t - q_t^{(k)}\|^2]\,.$$

Lemma D.5. *Given Assumptions 3.1-3.4, the following inequality holds:*

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

Lemma D.6. *Given Assumptions 3.1-3.4, the following inequality holds:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_{t+1} - y_{t+1}^{(k)}\|^2] \leq \left(1 - \frac{\eta(1-\lambda^2)}{2}\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|^2]$$

$$+\frac{2\eta\beta_y^2}{1-\lambda^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{q}_t-q_t^{(k)}\|^2]\,.$$

These three lemmas can be proved by following Lemma D.3. Thus, we omit its proof.

D.2 Gradient Estimation Errors

Lemma D.7. *Given Assumptions 3.1-3.4, when $\eta \le \frac{1}{\sqrt{\rho_x}}$, the following inequality holds:*

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

Proof. From the update rule of $u_{t+1}^{(k)}$, we derive

$$\begin{aligned}
&\mathbb{E}[\|\bar{\mathcal{G}}_{x,t+1}\|^2] = \mathbb{E}[\|(1-\rho_x\eta^2)(\frac{1}{K}\sum_{k=1}^{K}u_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\nabla_x F^{(k)}(x_t^{(k)},y_t^{(k)};\hat{x}_t^{(k)},\hat{y}_t^{(k)};\xi_{t+1}^{(k)}))\\
&\quad+\frac{1}{K}\sum_{k=1}^{K}\nabla_x F^{(k)}(x_{t+1}^{(k)},y_{t+1}^{(k)};\hat{x}_{t+1}^{(k)},\hat{y}_{t+1}^{(k)};\xi_{t+1}^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\nabla_x F^{(k)}(x_{t+1}^{(k)},y_{t+1}^{(k)};\hat{x}_{t+1}^{(k)},\hat{y}_{t+1}^{(k)})\|^2]\\
&\le (1-\rho_x\eta^2)^2\mathbb{E}[\|\bar{\mathcal{G}}_{x,t}\|^2]\\
&\quad+2(1-\rho_x\eta^2)^2\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_x F^{(k)}(x_t^{(k)},y_t^{(k)};\hat{x}_t^{(k)},\hat{y}_t^{(k)}) - \nabla_x F^{(k)}(x_t^{(k)},y_t^{(k)};\hat{x}_t^{(k)},\hat{y}_t^{(k)};\xi_{t+1}^{(k)})\\
&\qquad\qquad+\nabla_x F^{(k)}(x_{t+1}^{(k)},y_{t+1}^{(k)};\hat{x}_{t+1}^{(k)},\hat{y}_{t+1}^{(k)};\xi_{t+1}^{(k)}) - \nabla_x F^{(k)}(x_{t+1}^{(k)},y_{t+1}^{(k)};\hat{x}_{t+1}^{(k)},\hat{y}_{t+1}^{(k)})\|^2]\\
&\quad+2\rho_x^2\eta^4\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_x F^{(k)}(x_{t+1}^{(k)},y_{t+1}^{(k)};\hat{x}_{t+1}^{(k)},\hat{y}_{t+1}^{(k)};\xi_{t+1}^{(k)}) - \nabla_x F^{(k)}(x_{t+1}^{(k)},y_{t+1}^{(k)};\hat{x}_{t+1}^{(k)},\hat{y}_{t+1}^{(k)})\|^2]\\
&\le (1-\rho_x\eta^2)\mathbb{E}[\|\bar{\mathcal{G}}_{x,t}\|^2] + 2\rho_x^2\eta^4\sigma^2\frac{1}{K}\\
&\quad+2\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_x F^{(k)}(x_{t+1}^{(k)},y_{t+1}^{(k)};\hat{x}_{t+1}^{(k)},\hat{y}_{t+1}^{(k)};\xi_{t+1}^{(k)}) - \nabla_x F^{(k)}(x_t^{(k)},y_t^{(k)};\hat{x}_t^{(k)},\hat{y}_t^{(k)};\xi_{t+1}^{(k)})\|^2]\\
&\le (1-\rho_x\eta^2)\mathbb{E}[\|\bar{\mathcal{G}}_{x,t}\|^2] + 2\rho_x^2\eta^4\sigma^2\frac{1}{K} + 4L^2\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)}-x_t^{(k)}\|^2]\\
&\quad+4L^2\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}[\|y_{t+1}^{(k)}-y_t^{(k)}\|^2]\,,
\end{aligned}$$

where the second-to-last step holds from Assumption 3.3, the last step holds due to the following inequality:

$$\begin{aligned}
&\mathbb{E}[\|\nabla_x F^{(k)}(x_t^{(k)},y_t^{(k)};\hat{x}_t^{(k)},\hat{y}_t^{(k)}) - \nabla_x F^{(k)}(x_t^{(k)},y_t^{(k)};\hat{x}_t^{(k)},\hat{y}_t^{(k)};\xi_{t+1}^{(k)})\\
&\quad+\nabla_x F^{(k)}(x_{t+1}^{(k)},y_{t+1}^{(k)};\hat{x}_{t+1}^{(k)},\hat{y}_{t+1}^{(k)};\xi_{t+1}^{(k)}) - \nabla_x F^{(k)}(x_{t+1}^{(k)},y_{t+1}^{(k)};\hat{x}_{t+1}^{(k)},\hat{y}_{t+1}^{(k)})\|^2]\\
&= \mathbb{E}[\|\nabla_x f(x_t^{(k)},y_t^{(k)}) + \gamma_1(x_t^{(k)}-\hat{x}_t^{(k)}) - \nabla_x f(x_t^{(k)},y_t^{(k)};\xi_{t+1}^{(k)}) - \gamma_1(x_t^{(k)}-\hat{x}_t^{(k)})\\
&\qquad+\nabla_x f(x_{t+1}^{(k)},y_{t+1}^{(k)};\xi_{t+1}^{(k)}) + \gamma_1(x_{t+1}^{(k)}-\hat{x}_{t+1}^{(k)}) - \nabla_x f(x_{t+1}^{(k)},y_{t+1}^{(k)}) - \gamma_1(x_{t+1}^{(k)}-\hat{x}_{t+1}^{(k)})\|^2]\\
&= \mathbb{E}[\|\nabla_x f(x_t^{(k)},y_t^{(k)}) - \nabla_x f(x_t^{(k)},y_t^{(k)};\xi_{t+1}^{(k)}) + \nabla_x f(x_{t+1}^{(k)},y_{t+1}^{(k)};\xi_{t+1}^{(k)}) - \nabla_x f(x_{t+1}^{(k)},y_{t+1}^{(k)})\|^2]\\
&\le \mathbb{E}[\|\nabla_x f(x_{t+1}^{(k)},y_{t+1}^{(k)};\xi_{t+1}^{(k)}) - \nabla_x f(x_t^{(k)},y_t^{(k)};\xi_{t+1}^{(k)})\|^2]
\end{aligned}$$

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

□

Lemma D.8. *Given Assumptions 3.1-3.4, when $\eta \leq \frac{1}{\sqrt{\rho_x}}$, the following inequality holds:*

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

Lemma D.9. *Given Assumptions 3.1-3.4, when $\eta \leq \frac{1}{\sqrt{\rho_y}}$, the following inequality holds:*

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

Lemma D.10. *Given Assumptions 3.1-3.4, when $\eta \leq \frac{1}{\sqrt{\rho_y}}$, the following inequality holds:*

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

These three lemmas can be proved by following Lemma D.7. Thus, we omit its proof.

D.3 Other Auxiliary Lemmas

Lemma D.11. *Given Assumptions 3.1-3.4, the following inequality holds:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\hat{x}_{t+1}^{(k)} - \hat{x}_t^{(k)}\|^2] \leq 3\mathbb{E}[\|\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t\|^2] + 6\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\hat{x}}_t - \hat{x}_t^{(k)}\|^2]$$
$$+ \frac{12\eta\hat{\beta}_x^2}{1-\lambda^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2] + \frac{12\eta\hat{\beta}_x^2}{1-\lambda^2}\frac{2\eta\beta_x^2}{1-\lambda^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t - p_t^{(k)}\|^2] \,.$$

Proof. From the update rule of $\hat{x}_{t+1}^{(k)}$, we derive

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\hat{x}_{t+1}^{(k)} - \hat{x}_t^{(k)}\|^2] = \frac{1}{K}\mathbb{E}[\|\hat{X}_{t+1} - \hat{X}_t\|_F^2]$$
$$\leq 3\frac{1}{K}\mathbb{E}[\|\hat{X}_{t+1} - \bar{\hat{X}}_{t+1}\|_F^2] + 3\frac{1}{K}\mathbb{E}[\|\hat{X}_t - \bar{\hat{X}}_t\|_F^2] + 3\frac{1}{K}\mathbb{E}[\|\bar{\hat{X}}_{t+1} - \bar{\hat{X}}_t\|_F^2]$$
$$\leq 3\mathbb{E}[\|\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t\|^2] + 6\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\hat{x}}_t - \hat{x}_t^{(k)}\|^2]$$
$$+ \frac{12\eta\hat{\beta}_x^2}{1-\lambda^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2] + \frac{12\eta\hat{\beta}_x^2}{1-\lambda^2}\frac{2\eta\beta_x^2}{1-\lambda^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t - p_t^{(k)}\|^2] \,,$$

where the last step holds due to Lemma D.3. □

Lemma D.12. *Given Assumptions 3.1-3.4, the following inequality holds:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\hat{y}_{t+1}^{(k)}-\hat{y}_t^{(k)}\|^2] \le 3\mathbb{E}[\|\bar{\hat{y}}_{t+1}-\bar{\hat{y}}_t\|^2] + 6\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\hat{y}}_t-\hat{y}_t^{(k)}\|^2] + \frac{12\eta\hat{\beta}_y^2}{1-\lambda^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t-y_t^{(k)}\|^2] + \frac{12\eta\hat{\beta}_y^2}{1-\lambda^2}\frac{2\eta\beta_y^2}{1-\lambda^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{q}_t-q_t^{(k)}\|^2]\,. \tag{38}$$

This lemma can be proved by following Lemma D.11. Thus, we omit its proof.

Lemma D.13. *Given Assumptions 3.1-3.4, the following inequality holds:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t+1}^{(k)}-x_t^{(k)}\|^2] \le 12\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t-x_t^{(k)}\|^2] + 3\beta_x^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t-p_t^{(k)}\|^2] + 3\beta_x^2\eta^2\mathbb{E}[\|\bar{u}_t\|^2]\,. \tag{39}$$

Lemma D.14. *Given Assumptions 3.1-3.4, the following inequality holds:*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_{t+1}^{(k)}-y_t^{(k)}\|^2] \le 12\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t-y_t^{(k)}\|^2] + 3\beta_y^2\eta^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{q}_t-q_t^{(k)}\|^2] + 3\beta_y^2\eta^2\mathbb{E}[\|\bar{v}_t\|^2]\,. \tag{40}$$

Lemmas D.13, D.14 can be proved by following [7].

E Proof of Theorem 4.2

We first propose a novel potential function as follows:

$$\begin{aligned}\mathcal{L}_t = \mathcal{P}_t &+ c_1\mathbb{E}[\|\bar{\mathcal{G}}_{x,t}\|^2] + c_2\mathbb{E}[\|\bar{\mathcal{G}}_{y,t}\|^2] + c_8\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\hat{\mathcal{G}}_{x,t}^{(k)}\|^2] + c_9\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\hat{\mathcal{G}}_{y,t}^{(k)}\|^2]\\ &+ c_3\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t-x_t^{(k)}\|^2] + c_4\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t-y_t^{(k)}\|^2] + c_5\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\hat{x}}_t-\hat{x}_t^{(k)}\|^2]\\ &+ c_{10}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\hat{y}}_t-\hat{y}_t^{(k)}\|^2] + c_6\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t-p_t^{(k)}\|^2] + c_7\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{q}_t-q_t^{(k)}\|^2]\,,\end{aligned} \tag{41}$$

where the coefficient $\{c_i\}_{i=1}^{9}$ are positive.

Since

$$\begin{aligned}&\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\bar{u}_t\|^2]\\ &\le 2\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\frac{1}{K}\sum_{k=1}^{K}\nabla_x F^{(k)}(x_t^{(k)},y_t^{(k)};\hat{x}_t^{(k)},\hat{y}_t^{(k)})\|^2] + 2\mathbb{E}[\|\bar{\mathcal{G}}_{x,t}\|^2]\\ &\le 2L^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t-x_t^{(k)}\|^2] + 2L^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t-y_t^{(k)}\|^2] + 2\mathbb{E}[\|\bar{\mathcal{G}}_{x,t}\|^2]\,,\end{aligned}$$

and

$$\mathbb{E}[\|\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)-\bar{v}_t\|^2] \le 2L^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t-x_t^{(k)}\|^2] + 2L^2\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t-y_t^{(k)}\|^2] + 2\mathbb{E}[\|\bar{\mathcal{G}}_{y,t}\|^2]\,,$$

we apply Lemma C.7, consensus error lemmas in section D.1, gradient estimation error lemmas in section D.2, with the property of communication network that $\lambda < 1$, we obtain

$$\mathcal{L}_{t+1}-\mathcal{L}_t \le -\frac{\beta_x\eta}{4}\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2] - \frac{\beta_y\eta}{2}\mathbb{E}[\|\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2]$$

$$+(\beta_x\eta - c_1\rho_x\eta^2)\mathbb{E}[\|\bar{\mathcal{G}}_{x,t}\|^2] + \left(2\beta_y\eta + 240\gamma_1\hat{\beta}_x\eta\beta_y^2\eta^2 C^2_{x^1_{y\hat{x}\hat{y}}} + 8A_2\beta_y^2\eta^2 - \rho_y\eta^2 c_2\right)\mathbb{E}[\|\bar{\mathcal{G}}_{y,t}\|^2]$$
$$+\left(\frac{3\rho_x^2\eta^4}{1-\lambda}c_6 - \rho_x\eta^2 c_8\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\hat{\mathcal{G}}^{(k)}_{x,t}\|^2] + \left(\frac{3\rho_y^2\eta^4}{1-\lambda}c_7 - \rho_y\eta^2 c_9\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\hat{\mathcal{G}}^{(k)}_{y,t}\|^2]$$
$$+\left(2\gamma_1 C_{x^1_{\hat{x}\hat{y}}} + \frac{\gamma_1}{6\hat{\beta}_x\eta} + 6\gamma_1\hat{\beta}_x\eta\left(10C^2_{x^1_{y\hat{x}\hat{y}}}C^2_{y^1_{\hat{x}\hat{y}}} + \frac{4}{\gamma_1 - L}\frac{2\gamma_2^2 C^2_{y^1_{\hat{x}\hat{y}}}}{\mu}\right) + \frac{27\gamma_1^2}{1-\lambda}c_6\right.$$
$$\left. - \frac{\gamma_1(2-\hat{\beta}_x\eta)}{2\hat{\beta}_x\eta}\right)\mathbb{E}[\|\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t\|^2] + \left(2A_1 + \frac{27\gamma_2^2}{1-\lambda}c_7 - \frac{\gamma_2(2-\hat{\beta}_y\eta)}{2\hat{\beta}_y\eta}\right)\mathbb{E}[\|\bar{\hat{y}}_{t+1} - \bar{\hat{y}}_t\|^2]$$
$$+\left(\frac{54\gamma_1^2}{1-\lambda}c_6 - \frac{\eta(1-\lambda^2)}{4}c_5\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\hat{x}}_t - \hat{x}_t^{(k)}\|^2] + \frac{2c_1}{K}\rho_x^2\eta^4\sigma^2 + \frac{2c_2}{K}\rho_y^2\eta^4\sigma^2 + \frac{3c_6\rho_x^2\eta^4\sigma^2}{1-\lambda}$$
$$+\left(\frac{54\gamma_2^2}{1-\lambda}c_7 - \frac{\eta(1-\lambda^2)}{4}c_{10}\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\hat{y}}_t - \hat{y}_t^{(k)}\|^2] + \frac{3c_7\rho_y^2\eta^4\sigma^2}{1-\lambda} + 2c_8\rho_x^2\eta^4\sigma^2 + 2c_9\rho_y^2\eta^4\sigma^2$$
$$+ T_{\text{momentum}} + T_{\text{consensus error}}\ ,$$

where T_{momentum} and $T_{\text{consensus error}}$ are deferred to the subsequent proof (defined in Eq. (45)) and thus omitted here.

To cancel out $\mathbb{E}[\|\bar{\mathcal{G}}_{x,t}\|^2]$ and $\mathbb{E}[\|\bar{\mathcal{G}}_{y,t}\|^2]$, i.e.,

$$\beta_x\eta - \rho_x\eta^2 c_1 \le 0\ .$$
$$2\beta_y\eta + 240\gamma_1\hat{\beta}_x\eta\beta_y^2\eta^2 C^2_{x^1_{y\hat{x}\hat{y}}} + 8A_2\beta_y^2\eta^2 - \rho_y\eta^2 c_2 \le 0\ . \tag{42}$$

For the first inequality, we set $c_1 = \frac{\beta_x}{\rho_x\eta}$. For the second one, since the second and last inequality in Eq. (34) holds, we have

$$240\gamma_1\hat{\beta}_x\eta\beta_y^2\eta^2 C^2_{x^1_{y\hat{x}\hat{y}}} \le \frac{2\beta_x\eta}{32\times 16}\frac{(\gamma_1 - L)^2}{L^2}\ ,\quad 8A_2\beta_y^2\eta^2 \le \frac{2\beta_x\eta}{32\times 16}\frac{(\gamma_1 - L)^2}{L^2}\ .$$

Then, by the definition of c_{β_y}, i.e., $c_{\beta_y} = \frac{(\gamma_1 - L)^2}{64L^2}$, we set

$$240\gamma_1\hat{\beta}_x\eta\beta_y^2\eta^2 C^2_{x^1_{y\hat{x}\hat{y}}} \le \frac{2\beta_x\eta}{32\times 16}64c_{\beta_y} = \frac{1}{4}\beta_y\eta\ ,\quad 8A_2\beta_y^2\eta^2 \le \frac{2\beta_x\eta}{32\times 16}64c_{\beta_y} = \frac{\beta_y\eta}{4}\ . \tag{43}$$

Therefore, we obtain $c_2 = \frac{5\beta_y}{2\rho_y\eta}$.

To cancel out $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\hat{\mathcal{G}}^{(k)}_{x,t}\|^2]$ and $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\hat{\mathcal{G}}^{(k)}_{y,t}\|^2]$, i.e.,

$$\frac{3\rho_x^2\eta^4}{1-\lambda}c_6 - \rho_x\eta^2 c_8 \le 0\ ,\quad \frac{3\rho_y^2\eta^4}{1-\lambda}c_7 - \rho_y\eta^2 c_9 \le 0\ .$$

Since $\rho_x\eta^2 < 1$, $\rho_y\eta^2 < 1$, we set

$$c_6 = \beta_x\eta(1-\lambda)\ ,\quad c_7 = \beta_y\eta(1-\lambda)\ ,\quad c_8 = 3\beta_x\eta\ ,\quad c_9 = 3\beta_y\eta\ .$$

To cancel out $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\hat{x}}_t - \hat{x}_t^{(k)}\|^2]$ and $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{\hat{y}}_t - \hat{y}_t^{(k)}\|^2]$, i.e.,

$$\frac{54\gamma_1^2}{1-\lambda}c_6 - \frac{\eta(1-\lambda^2)}{4}c_5 \le 0\ ,\quad \frac{54\gamma_2^2}{1-\lambda}c_7 - \frac{\eta(1-\lambda^2)}{4}c_{10} \le 0\ ,$$

we set

$$c_5 = \frac{216\beta_x\gamma_1^2}{(1-\lambda)}\ ,\quad c_{10} = \frac{216\beta_y\gamma_2^2}{(1-\lambda)}\ .$$

For $\mathbb{E}[\|\bar{\hat{x}}_{t+1} - \bar{\hat{x}}_t\|^2]$, by setting

$$2\gamma_1 C_{x^1_{\hat{x}\hat{y}}} + 6\gamma_1\hat{\beta}_x\eta\left(10C^2_{x^1_{y\hat{x}\hat{y}}}C^2_{y^1_{\hat{x}\hat{y}}} + \frac{4}{\gamma_1 - L}\frac{2\gamma_2^2 C^2_{y^1_{\hat{x}\hat{y}}}}{\mu}\right) + \frac{27\gamma_1^2}{1-\lambda}c_6 - \frac{\gamma_1}{3\hat{\beta}_x\eta} \le -\frac{\gamma_1}{4\hat{\beta}_x\eta}\ .$$

Specifically, we enforce

$$2\gamma_1 C_{x^1_{\hat{x}\hat{y}}} \leq \frac{\gamma_1}{36\hat{\beta}_x\eta}\,, \quad \frac{27\gamma_1^2}{1-\lambda}c_6 \leq \frac{\gamma_1}{36\hat{\beta}_x\eta}\,,$$
$$6\gamma_1\hat{\beta}_x\eta\Big(10C^2_{x^1_{\hat{x}\hat{y}}}C^2_{y^1_{\hat{x}\hat{y}}} + \frac{4}{\gamma_1 - L}\frac{2\gamma_2^2 C^2_{y^1_{\hat{x}\hat{y}}}}{\mu}\Big) \leq \frac{\gamma_1}{36\hat{\beta}_x\eta}\,.$$

Since $C_{x^1_{\hat{x}\hat{y}}} = \frac{\gamma_1}{\gamma_1 - L}$, $C_{x^1_{y\hat{x}\hat{y}}} = \frac{\gamma_1}{\gamma_1 - L}$, and $C_{y^1_{\hat{x}\hat{y}}} = \frac{\gamma_1}{\gamma_2 - L}$, we obtain

$$\beta_x \leq \min\left\{\frac{\gamma_1 - L}{72c_{\hat{\beta}_x}\gamma_1}\,, \frac{(\gamma_1 - L)(\gamma_2 - L)\sqrt{\mu}}{6\gamma_1 c_{\hat{\beta}_x}\sqrt{6(10\gamma_1^2\mu + 8\gamma_2^2(\gamma_1 - L))}}\,, \frac{1}{18\sqrt{3c_{\hat{\beta}_x}\gamma_1}}\right\}\,.$$

For $\mathbb{E}[\|\bar{\tilde{y}}_{t+1} - \bar{\tilde{y}}_t\|^2]$, by setting

$$2A_1 + \frac{27\gamma_2^2}{1-\lambda}c_7 - \frac{\gamma_2}{2\hat{\beta}_y\eta} \leq -\frac{\gamma_2}{4\hat{\beta}_y\eta}\,.$$

Specifically, from the definition of A_1, we enforce

$$12\gamma_1\hat{\beta}_x\eta C^2_{y^2_{\hat{x}\hat{y}}}\left(10C^2_{x^1_{y\hat{x}\hat{y}}} + \frac{8\gamma_2^2}{(\gamma_1 - L)\mu}\right) \leq \frac{\gamma_2}{16\hat{\beta}_y\eta}\,,$$
$$12\gamma_1\hat{\beta}_x\eta\frac{8\gamma_2^2}{(\gamma_1 - L)\mu}\frac{(1-\hat{\beta}_y\eta)^2}{\hat{\beta}_y^2\eta^2} \leq \frac{\gamma_2}{16\hat{\beta}_y\eta}\,, \quad \frac{27\gamma_2^2}{1-\lambda}c_7 \leq \frac{\gamma_2}{8\hat{\beta}_y\eta}\,. \tag{44}$$

To solve the second inequality, we use the second inequality in Eq. (35) to obtain the following:

$$\frac{4}{(\gamma_2 - L)^2}\frac{24\gamma_1\hat{\beta}_x\eta}{\gamma_1 - L}\frac{2\gamma_2^2}{\mu}\frac{(1-\hat{\beta}_y\eta)^2}{\hat{\beta}_y^2\eta^2} \leq \frac{\beta_x\eta}{32\times 64}\frac{(\gamma_1 - L)^2}{4L^2}\,.$$

Then, we derive

$$12\gamma_1\hat{\beta}_x\eta\frac{8\gamma_2^2}{(\gamma_1 - L)\mu}\frac{(1-\hat{\beta}_y\eta)^2}{\hat{\beta}_y^2\eta^2} \leq \frac{\beta_x\eta}{32\times 64}\frac{(\gamma_1 - L)^2}{4L^2}\frac{(\gamma_2 - L)^2}{2}\,.$$

Therefore, it leads us to solve

$$\frac{\beta_x\eta}{32\times 64}\frac{(\gamma_1 - L)^2}{4L^2}\frac{(\gamma_2 - L)^2}{2} \leq \frac{\gamma_2}{16\hat{\beta}_y\eta}\,,$$

and we obtain $\beta_x \leq \frac{32L}{\sqrt{c_{\hat{\beta}_y}}(\gamma_1 - L)(\gamma_2 - L)}$. Finally, to solve the first and last inequality in Eq. (44), from $C_{x^1_{y\hat{x}\hat{y}}} = \frac{\gamma_1}{\gamma_1 - L}$ and $C_{y^2_{\hat{x}\hat{y}}} = \frac{\gamma_2}{\gamma_2 - L}$, we obtain

$$\beta_x \leq \min\Big\{\frac{\sqrt{\mu}(\gamma_1 - L)(\gamma_2 - L)}{8\sqrt{3c_{\hat{\beta}_x}c_{\hat{\beta}_y}\gamma_1\gamma_2(10\gamma_1^2\mu + 8\gamma_2^2(\gamma_1 - L))}}\,, \frac{1}{6\sqrt{6\gamma_2 c_{\beta_y}c_{\hat{\beta}_y}}}\Big\}\,.$$

Next, we solve the T_{momentum} and $T_{\text{consensus error}}$:

$$T_{\text{momentum}} + T_{\text{consensus error}} = \left(4\beta_y\eta\beta_x^2\eta^2L^2 + 3\beta_x^2\eta^2\mathcal{X} - \frac{\beta_x\eta}{4}\right)\mathbb{E}[\|\bar{u}_t\|^2]$$
$$+ \left(\beta_y^2\eta^2L_d + \frac{3\beta_y\eta}{4} + \frac{\beta_y^2\eta^2(\gamma_2 + L)}{2} + 4A_1\hat{\beta}_y^2\eta^2\beta_y^2\eta^2 + 2A_2\beta_y^2\eta^2 + 3\beta_y^2\eta^2\mathcal{Y} - \frac{7}{8}\beta_y\eta\right)\mathbb{E}[\|\bar{v}_t\|^2]$$
$$+ \left(\beta_x\eta L^2 + 2L^2A_3 + \frac{4\eta\hat{\beta}_x^2}{1-\lambda}c_5 + \frac{108\eta\hat{\beta}_x^2\gamma_1^2}{(1-\lambda)^2}c_6 + 12\eta^2\mathcal{X} - \frac{\eta(1-\lambda^2)}{2}c_3\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2]$$
$$+ \left(\beta_x\eta L^2 + 2L^2A_3 + \frac{4\eta\hat{\beta}_y^2}{1-\lambda}c_{10} + \frac{108\eta\hat{\beta}_y^2\gamma_2^2}{(1-\lambda)^2}c_7 + 12\eta^2\mathcal{Y} - \frac{\eta(1-\lambda^2)}{2}c_4\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|^2]$$

$$+\left(\frac{2\eta\beta_x^2}{1-\lambda}c_3+\frac{8\eta^2\beta_x^2\hat{\beta}_x^2}{(1-\lambda)^2}c_5+\frac{216\eta^2\beta_x^2\hat{\beta}_x^2\gamma_1^2}{(1-\lambda)^3}c_6+3\beta_x^2\eta^2\mathcal{X}-(1-\lambda)c_6\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t-p_t^{(k)}\|^2]$$
$$+\left(\frac{2\eta\beta_y^2}{1-\lambda}c_4+\frac{8\eta^2\beta_y^2\hat{\beta}_y^2}{(1-\lambda)^2}c_{10}+\frac{216\eta^2\beta_y^2\hat{\beta}_y^2\gamma_2^2}{(1-\lambda)^3}c_7+3\beta_y^2\eta^2\mathcal{Y}-(1-\lambda)c_7\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{q}_t-q_t^{(k)}\|^2]\,. \tag{45}$$

where

$$\mathcal{X}=\frac{4L^2}{K}\frac{\beta_x}{\rho_x\eta}+\frac{10L^2}{K}\frac{\beta_y}{\rho_y\eta}+(21L^2+9\gamma_1^2)\beta_x\eta+21L^2\beta_y\eta\,,$$
$$\mathcal{Y}=\frac{4L^2}{K}\frac{\beta_x}{\rho_x\eta}+\frac{10L^2}{K}\frac{\beta_y}{\rho_y\eta}+21L^2\beta_x\eta+(21L^2+9\gamma_2^2)\beta_y\eta\,.$$

To cancel out $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t-x_t^{(k)}\|^2]$, i.e.,

$$\beta_x\eta L^2+2L^2A_3+\frac{4\eta\hat{\beta}_x^2}{1-\lambda}c_5+\frac{108\eta\hat{\beta}_x^2\gamma_1^2}{(1-\lambda)^2}c_6+12\eta^2\mathcal{X}-\frac{\eta(1-\lambda^2)}{2}c_3\le 0\,.$$

Moreover, from the definition of A_3 and Eq. (42), we have

$$\beta_y\eta+120\gamma_1\hat{\beta}_x\eta\beta_y^2\eta^2C^2_{x^1_{y\hat{x}\hat{y}}}+4A_2\beta_y^2\eta^2\le\frac{5}{4}\beta_y\eta\,.$$

Therefore, we set

$$\begin{aligned}
&\beta_x\eta L^2+2L^2A_3+\frac{4\eta\hat{\beta}_x^2}{1-\lambda}c_5+\frac{108\eta\hat{\beta}_x^2\gamma_1^2}{(1-\lambda)^2}c_6+12\eta^2\mathcal{X}\\
&\le\beta_x\eta L^2+\frac{5}{2}\beta_y\eta L^2+\frac{4\eta\hat{\beta}_x^2}{1-\lambda}\frac{216\beta_x\gamma_1^2}{(1-\lambda)}+\frac{108\gamma_1^2\eta\hat{\beta}_x^2}{(1-\lambda)}\beta_x\eta\\
&\quad+12\eta^2\left(\frac{4L^2}{K}\frac{\beta_x}{\rho_x\eta}+\frac{10L^2}{K}\frac{\beta_y}{\rho_y\eta}+(21L^2+9\gamma_1^2)\beta_x\eta+21L^2\beta_y\eta\right)\\
&\le\beta_x\eta L^2+\beta_x\eta L^2\frac{5\beta_y}{2\beta_x}+\beta_x\eta\hat{\beta}_x^2\frac{864\gamma_1^2}{(1-\lambda)^2}+\beta_x\eta\hat{\beta}_x\frac{108\gamma_1^2}{(1-\lambda)}\\
&\quad+12\eta\left(\frac{4L^2}{\rho_xK}\beta_x+\frac{10L^2}{\rho_yK}\beta_y+(21L^2+9\gamma_1^2)\beta_x+21L^2\beta_y\right)\\
&=\beta_x\eta L^2+\beta_x\eta L^2\frac{5}{2}c_{\beta_y}+\beta_x^3\eta c^2_{\hat{\beta}_x}\frac{864\gamma_1^2}{(1-\lambda)^2}+\beta_x^2\eta c_{\hat{\beta}_x}\frac{108\gamma_1^2}{(1-\lambda)}\\
&\quad+12\beta_x\eta\left(\frac{4L^2}{\rho_xK}+\frac{10L^2}{\rho_yK}c_{\beta_y}+(21L^2+9\gamma_1^2)+21L^2c_{\beta_y}\right)\\
&\le\frac{\eta(1-\lambda)}{2}c_3\,,
\end{aligned}$$

where the second step holds due to $\hat{\beta}_x\eta<1$ and $\eta<1$, the third step holds due to Eq. (26) . By solving this inequality, we obtain

$$c_3\ge\frac{2\beta_x}{(1-\lambda)}\left(\frac{48L^2}{\rho_xK}+\frac{120L^2}{\rho_yK}c_{\beta_y}+253L^2+108\gamma_1^2+255L^2c_{\beta_y}+\beta_x^2c^2_{\hat{\beta}_x}\frac{864\gamma_1^2}{(1-\lambda)^2}+\beta_xc_{\hat{\beta}_x}\frac{108\gamma_1^2}{(1-\lambda)}\right).$$

Then, we set

$$\begin{aligned}
c_3=&\frac{2\beta_x}{(1-\lambda)}\Bigg(\underbrace{\frac{48L^2}{\rho_xK}+\frac{120L^2}{\rho_yK}c_{\beta_y}+253L^2+108\gamma_1^2+255L^2c_{\beta_y}}_{c_{3,1}}\Bigg)\\
&+\frac{2\beta_x^3}{(1-\lambda)^3}\underbrace{864\gamma_1^2c^2_{\hat{\beta}_x}}_{c_{3,2}}+\frac{2\beta_x^2}{(1-\lambda)^2}\underbrace{108\gamma_1^2c_{\hat{\beta}_x}}_{c_{3,3}}\,.
\end{aligned}$$

Here, it is easy to know that $c_{3,1} = O(1)$ when $\rho_x = O(1/K)$ and $\rho_y = O(1/K)$, $c_{3,2} = O(1/\kappa^2)$ and $c_{3,3} = O(1/\kappa)$ due to $c_{\hat{\beta}_x} = O(1/\kappa)$.

To cancel out $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|^2]$, i.e.,

$$\beta_x \eta L^2 + 2L^2 A_3 + \frac{4\eta\hat{\beta}_y^2}{1-\lambda}c_{10} + \frac{108\eta\hat{\beta}_y^2\gamma_2^2}{(1-\lambda)^2}c_7 + 12\eta^2\mathcal{Y} - \frac{\eta(1-\lambda^2)}{2}c_4 \leq 0\,.$$

Therefore, we set

$$\begin{aligned}
&\beta_x \eta L^2 + 2L^2 A_3 + \frac{4\eta\hat{\beta}_y^2}{1-\lambda}c_{10} + \frac{108\eta\hat{\beta}_y^2\gamma_2^2}{(1-\lambda)^2}c_7 + 12\eta^2\mathcal{Y} \\
&\leq \beta_x \eta L^2 + \beta_x \eta L^2 \frac{5}{2}c_{\beta_y} + \beta_x^3 \eta c_{\beta_y} c_{\hat{\beta}_y}^2 \frac{864\gamma_2^2}{(1-\lambda)^2} + \beta_x^2 \eta c_{\beta_y} c_{\hat{\beta}_y} \frac{108\gamma_2^2}{(1-\lambda)} \\
&\quad + 12\beta_x\eta\left(\frac{4L^2}{\rho_x K} + \frac{10L^2}{\rho_y K}c_{\beta_y} + 21L^2 + (21L^2 + 9\gamma_2^2)c_{\beta_y}\right) \\
&\leq \frac{\eta(1-\lambda)}{2}c_4\,,
\end{aligned}$$

where the first step holds due to $\hat{\beta}_y\eta < 1$, $\eta < 1$, and Eq. (26). By solving this inequality, we set

$$\begin{aligned}
c_4 = \frac{2\beta_x}{(1-\lambda)}\Bigg(&\underbrace{\frac{48L^2}{\rho_x K} + \frac{120L^2}{\rho_y K}c_{\beta_y} + 253L^2 + 255L^2 c_{\beta_y} + 108\gamma_2^2 c_{\beta_y}}_{c_{4,1}}\Bigg) \\
&+ \frac{2\beta_x^3}{(1-\lambda)^3}\underbrace{864\gamma_2^2 c_{\beta_y} c_{\hat{\beta}_y}^2}_{c_{4,2}} + \frac{2\beta_x^2}{(1-\lambda)^2}\underbrace{108\gamma_2^2 c_{\beta_y} c_{\hat{\beta}_y}}_{c_{4,3}}\,.
\end{aligned}$$

Similarly, it is easy to know that $c_{4,1} = O(1)$ when $\rho_x = O(1/K)$ and $\rho_y = O(1/K)$, $c_{4,2} = O(1)$ and $c_{4,3} = O(1)$ due to $c_{\beta_y} = O(1)$ and $c_{\hat{\beta}_y} = O(1)$.

To cancel out $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_t - p_t^{(k)}\|^2]$, i.e.,

$$\frac{2\eta\beta_x^2}{1-\lambda}c_3 + \frac{8\eta^2\beta_x^2\hat{\beta}_x^2}{(1-\lambda)^2}c_5 + \frac{216\eta^2\beta_x^2\hat{\beta}_x^2\gamma_1^2}{(1-\lambda)^3}c_6 + 3\beta_x^2\eta^2\mathcal{X} - (1-\lambda)c_6 \leq 0\,.$$

Firstly, we enforce

$$\frac{216\eta^2\beta_x^2\hat{\beta}_x^2\gamma_1^2}{(1-\lambda)^3}c_6 \leq \frac{(1-\lambda)}{4}c_6\,.$$

Then, based on Eq. (26) , we obtain $\beta_x \leq \frac{(1-\lambda)}{6\sqrt{\gamma_1 c_{\hat{\beta}_x}}}$.

Then, we enforce

$$c_3\frac{2\eta\beta_x^2}{1-\lambda} \leq \frac{\beta_x\eta}{4}(1-\lambda)^2\,, \quad c_5\frac{8\eta^2\beta_x^2\hat{\beta}_x^2}{(1-\lambda)^2} \leq \frac{\beta_x\eta}{4}(1-\lambda)^2\,, \quad 3\beta_x^2\eta^2\mathcal{X} \leq \frac{\beta_x\eta}{16}(1-\lambda)^2\,. \tag{46}$$

To solve the first inequality in Eq. (46), we enforce

$$\begin{aligned}
&\frac{2\eta\beta_x^2}{1-\lambda}\frac{2\beta_x}{(1-\lambda)}c_{3,1} \leq \frac{\beta_x\eta}{12}(1-\lambda)^2\,, \quad \frac{2\eta\beta_x^2}{1-\lambda}\frac{2\beta_x^3}{(1-\lambda)^3}c_{3,2} \leq \frac{\beta_x\eta}{12}(1-\lambda)^2\,, \\
&\frac{2\eta\beta_x^2}{1-\lambda}\frac{2\beta_x^2}{(1-\lambda)^2}c_{3,3} \leq \frac{\beta_x\eta}{12}(1-\lambda)^2\,.
\end{aligned}$$

Therefore, we obtain

$$\beta_x \leq \min\left\{\frac{(1-\lambda)^2}{4\sqrt{3c_{3,1}}}\,, \frac{(1-\lambda)^{3/2}}{2(3c_{3,2})^{1/4}}\,, \frac{(1-\lambda)^{5/3}}{2(6c_{3,3})^{1/3}}\right\}\,.$$

To solve the second inequality in Eq. (46), we obtain $\beta_x \leq \frac{(1-\lambda)^{5/4}}{12\sqrt{\gamma_1 c_{\hat{\beta}_x}}}$.

To solve the last inequality in Eq. (46), we enforce

$$3\beta_x\eta\frac{4L^2}{K}\frac{\beta_x}{\rho_x\eta} \leq \frac{1}{16\times 4}(1-\lambda)^2 \;, \quad 3\beta_x\eta\frac{10L^2}{K}\frac{\beta_y}{\rho_y\eta} \leq \frac{1}{16\times 4}(1-\lambda)^2 \;,$$
$$3\beta_x\eta(21L^2+9\gamma_1^2)\beta_x\eta \leq \frac{1}{16\times 4}(1-\lambda)^2 \;, \quad 3\beta_x\eta 21L^2\beta_y\eta \leq \frac{1}{16\times 4}(1-\lambda)^2 \;.$$

We obtain

$$\beta_x \leq \left\{\frac{\sqrt{\rho_x K}(1-\lambda)}{16\sqrt{3}L} \;, \frac{\sqrt{\rho_y K}(1-\lambda)}{8L\sqrt{30c_{\beta_y}}} \;, \frac{(1-\lambda)}{8\sqrt{3(21L^2+9\gamma_1^2)}} \;, \frac{(1-\lambda)}{24L\sqrt{7c_{\beta_y}}}\right\} .$$

To cancel out $\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{q}_t - q_t^{(k)}\|^2]$, i.e.,

$$\frac{2\eta\beta_y^2}{1-\lambda}c_4 + \frac{8\eta^2\beta_y^2\hat{\beta}_y^2}{(1-\lambda)^2}c_{10} + \frac{216\eta^2\beta_y^2\hat{\beta}_y^2\gamma_2^2}{(1-\lambda)^3}c_7 + 3\beta_y^2\eta^2\mathcal{Y} - (1-\lambda)c_7 \leq 0 \;.$$

Firstly, we enforce

$$\frac{216\eta^2\beta_y^2\hat{\beta}_y^2\gamma_2^2}{(1-\lambda)^3}c_7 \leq \frac{(1-\lambda)}{6}c_7 \;.$$

Then, based on Eq. (26), we obtain $\beta_x \leq \frac{(1-\lambda)}{6\sqrt{\gamma_2 c_{\beta_y} c_{\hat{\beta}_y}}}$.

Then, we enforce

$$c_4\frac{2\eta\beta_y^2}{1-\lambda} \leq \frac{\beta_y\eta}{4}(1-\lambda)^2 \;, \quad c_{10}\frac{8\eta^2\beta_y^2\hat{\beta}_y^2}{(1-\lambda)^2} \leq \frac{\beta_y\eta}{4}(1-\lambda)^2 \;, \quad 3\beta_y^2\eta^2\mathcal{Y} \leq \frac{\beta_y\eta}{96}(1-\lambda)^2 \;. \tag{47}$$

To solve the first inequality in Eq. (47), we enforce

$$\frac{2\eta\beta_y^2}{1-\lambda}\frac{2\beta_x}{(1-\lambda)}c_{4,1} \leq \frac{\beta_y\eta}{12}(1-\lambda)^2 \;, \quad \frac{2\eta\beta_y^2}{1-\lambda}\frac{2\beta_x^3}{(1-\lambda)^3}c_{4,2} \leq \frac{\beta_y\eta}{12}(1-\lambda)^2 \;,$$
$$\frac{2\eta\beta_y^2}{1-\lambda}\frac{2\beta_x^2}{(1-\lambda)^2}c_{4,3} \leq \frac{\beta_y\eta}{12}(1-\lambda)^2 \;.$$

Therefore, we obtain

$$\beta_x \leq \min\left\{\frac{(1-\lambda)^2}{4\sqrt{3c_{\beta_y}c_{4,1}}} \;, \frac{(1-\lambda)^{3/2}}{2(3c_{\beta_y}c_{4,2})^{1/4}} \;, \frac{(1-\lambda)^{5/3}}{2(6c_{\beta_y}c_{4,3})^{1/3}}\right\} .$$

To solve the second inequality in Eq. (47), we obtain $\beta_x \leq \frac{(1-\lambda)^{5/4}}{12\sqrt{\gamma_2 c_{\beta_y} c_{\hat{\beta}_y}}}$.

To solve the last inequality in Eq. (47), we enforce

$$3\beta_y\eta\frac{4L^2}{K}\frac{\beta_x}{\rho_x\eta} \leq \frac{1}{96\times 4}(1-\lambda)^2 \;, \quad 3\beta_y\eta\frac{10L^2}{K}\frac{\beta_y}{\rho_y\eta} \leq \frac{1}{96\times 4}(1-\lambda)^2 \;,$$
$$3\beta_y\eta 21L^2\beta_x\eta \leq \frac{1}{96\times 4}(1-\lambda)^2 \;, \quad 3\beta_y\eta(21L^2+9\gamma_2^2)\beta_y\eta \leq \frac{1}{96\times 4}(1-\lambda)^2 \;.$$

We obtain

$$\beta_x \leq \left\{\frac{\sqrt{\rho_x K}(1-\lambda)}{48\sqrt{2c_{\beta_y}}L} \;, \frac{\sqrt{\rho_y K}(1-\lambda)}{48c_{\beta_y}L\sqrt{5}} \;, \frac{(1-\lambda)}{24L\sqrt{42c_{\beta_y}}} \;, \frac{(1-\lambda)}{24c_{\beta_y}\sqrt{2(21L^2+9\gamma_2^2)}}\right\} .$$

For $\mathbb{E}[\|\bar{u}_t\|^2]$, we enforce

$$4\beta_y\eta\beta_x^2\eta^2L^2 + 3\beta_x^2\eta^2\mathcal{X} - \frac{\beta_x\eta}{4} \leq -\frac{\beta_x\eta}{8} \;.$$

Specifically, we enforce

$$4\beta_y\eta\beta_x^2\eta^2L^2 \leq \frac{\beta_x\eta}{16}\,, \quad 3\beta_x^2\eta^2\mathcal{X} \leq \frac{\beta_x\eta}{16}\,.$$

To solve the first inequality, we obtain $\beta_x \leq \frac{1}{8L\sqrt{c_{\beta_y}}}$. For the last inequality, we use the last inequality in Eq. (46) along with the fact that $1-\lambda < 1$, the inequality is already satisfied.

For $\mathbb{E}[\|\bar{v}_t\|^2]$, we enforce

$$\beta_y^2\eta^2L_d + \frac{3\beta_y\eta}{4} + \frac{\beta_y^2\eta^2(\gamma_2+L)}{2} + 4A_1\hat{\beta}_y^2\eta^2\beta_y^2\eta^2 + 2A_2\beta_y^2\eta^2 + 3\beta_y^2\eta^2\mathcal{Y} - \frac{7}{8}\beta_y\eta \leq -\frac{1}{32}\beta_y\eta\,.$$

Firstly, from Eq. (43) and the definition of A_2, we obtain $2A_2\beta_y^2\eta^2 \leq \frac{\beta_y\eta}{4\times 4}$, and

$$2\beta_y^2\eta^2A_1\frac{4\hat{\beta}_y^2}{\beta_y^2(\gamma_2-L)^2} \leq \frac{\beta_y\eta}{4\times 4}\,.$$

By reformulating the above inequality, we obtain

$$4A_1\hat{\beta}_y^2\eta^2\beta_y^2\eta^2 \leq \frac{\beta_y^3\eta^3(\gamma_2-L)^2}{4\times 8}\,.$$

Therefore, we enforce

$$\beta_y^2\eta^2L_d + \frac{\beta_y^2\eta^2(\gamma_2+L)}{2} + \frac{\beta_y^3\eta^3(\gamma_2-L)^2}{32} + 3\beta_y^2\eta^2\mathcal{Y} \leq \frac{1}{32}\beta_y\eta\,.$$

Specifically, we enforce

$$\beta_y^2\eta^2L_d + \frac{\beta_y^2\eta^2(\gamma_2+L)}{2} \leq \frac{1}{96}\beta_y\eta\,, \quad \frac{\beta_y^3\eta^3(\gamma_2-L)^2}{32} \leq \frac{1}{96}\beta_y\eta\,, \quad 3\beta_y^2\eta^2\mathcal{Y} \leq \frac{1}{96}\beta_y\eta\,.$$

To solve the first and second inequality, we obtain

$$\beta_x \leq \min\Big\{\frac{1}{48c_{\beta_y}(2L_d+\gamma_2+L)}\,, \frac{1}{\sqrt{3}c_{\beta_y}(\gamma_2-L)}\Big\}\,.$$

For the last inequality, we use the last inequality in Eq. (47) along with the fact that $1-\lambda<1$, the inequality is already satisfied.

In summary, by setting

$$\begin{aligned}
&c_1 = \frac{\beta_x}{\rho_x\eta}\,, \quad c_2 = \frac{5\beta_y}{2\rho_y\eta}\,,\\
&c_3 \triangleq \frac{2\beta_x}{(1-\lambda)}c_{3,1} + \frac{2\beta_x^3}{(1-\lambda)^3}c_{3,2} + \frac{2\beta_x^2}{(1-\lambda)^2}c_{3,3}\,,\\
&c_4 \triangleq \frac{2\beta_x}{(1-\lambda)}c_{4,1} + \frac{2\beta_x^3}{(1-\lambda)^3}c_{4,2} + \frac{2\beta_x^2}{(1-\lambda)^2}c_{4,3}\,,\\
&c_5 = \frac{216\beta_x\gamma_1^2}{(1-\lambda)}\,, \quad c_{10} = \frac{216\beta_y\gamma_2^2}{(1-\lambda)}\,,\\
&c_6 = \beta_x\eta(1-\lambda)\,, \quad c_7 = \beta_y\eta(1-\lambda)\,, \quad c_8 = 3\beta_x\eta\,, \quad c_9 = 3\beta_y\eta\,,
\end{aligned} \tag{48}$$

with

$$\begin{aligned}
\beta_x \leq \min\Big\{&\frac{(1-\lambda)}{6\sqrt{\gamma_1 c_{\hat{\beta}_x}}}\,, \frac{(1-\lambda)^2}{4\sqrt{3c_{3,1}}}\,, \frac{(1-\lambda)^{3/2}}{2(3c_{3,2})^{1/4}}\,, \frac{(1-\lambda)^{5/3}}{2(6c_{3,3})^{1/3}}\,, \frac{(1-\lambda)^{5/4}}{12\sqrt{\gamma_1 c_{\hat{\beta}_x}}}\,, \frac{\sqrt{\rho_x K}(1-\lambda)}{16\sqrt{3}L}\,,\\
&\frac{\sqrt{\rho_y K}(1-\lambda)}{8L\sqrt{30c_{\beta_y}}}\,, \frac{(1-\lambda)}{8\sqrt{3(21L^2+9\gamma_1^2)}}\,, \frac{(1-\lambda)}{6\sqrt{\gamma_2 c_{\beta_y}c_{\hat{\beta}_y}}}\,, \frac{(1-\lambda)^2}{4\sqrt{3c_{\beta_y}c_{4,1}}}\,, \frac{(1-\lambda)^{3/2}}{2(3c_{\beta_y}c_{4,2})^{1/4}}\,,\\
&\frac{(1-\lambda)^{5/3}}{2(6c_{\beta_y}c_{4,3})^{1/3}}\,, \frac{(1-\lambda)^{5/4}}{12\sqrt{\gamma_2 c_{\beta_y}c_{\hat{\beta}_y}}}\,, \frac{\sqrt{\rho_x K}(1-\lambda)}{48\sqrt{2c_{\beta_y}}L}\,, \frac{\sqrt{\rho_y K}(1-\lambda)}{48c_{\beta_y}L\sqrt{5}}\,, \frac{(1-\lambda)}{24L\sqrt{42c_{\beta_y}}}\,,
\end{aligned}$$

$$\frac{(1-\lambda)}{24c_{\beta_y}\sqrt{2(21L^2+9\gamma_2^2)}}, \frac{\gamma_1-L}{72c_{\hat{\beta}_x}\gamma_1}, \frac{(\gamma_1-L)(\gamma_2-L)\sqrt{\mu}}{6\gamma_1 c_{\hat{\beta}_x}\sqrt{6(10\gamma_1^2\mu+8\gamma_2^2(\gamma_1-L))}}, \frac{1}{18\sqrt{3c_{\hat{\beta}_x}\gamma_1}},$$
$$\frac{4L}{\sqrt{c_{\hat{\beta}_y}}(\gamma_1-L)(\gamma_2-L)}, \frac{\sqrt{\mu}(\gamma_1-L)(\gamma_2-L)}{8\sqrt{3c_{\hat{\beta}_x}c_{\hat{\beta}_y}\gamma_1\gamma_2(10\gamma_1^2\mu+8\gamma_2^2(\gamma_1-L))}}, \frac{1}{6\sqrt{6\gamma_2 c_{\beta_y}c_{\hat{\beta}_y}}},$$
$$\frac{L^2}{120\gamma_1^3}, \frac{\sqrt{\mu(\gamma_1-L)^3}(\gamma_2-L)^2}{512\sqrt{6\gamma_1 c_{\hat{\beta}_x}}\gamma_2 c_{\hat{\beta}_y}}, \frac{1}{8L\sqrt{c_{\beta_y}}}, \frac{1}{48c_{\beta_y}(2L_d+\gamma_2+L)}, \frac{1}{\sqrt{3}c_{\beta_y}(\gamma_2-L)}\Bigg\}$$
$$\eta \leq \min\left\{1, \frac{1}{\sqrt{\rho_x}}, \frac{1}{\sqrt{\rho_y}}, \frac{1}{\hat{\beta}_x}, \frac{1}{\hat{\beta}_y}, \frac{1}{\beta_x}, \frac{1}{\beta_y}, \frac{1}{2\beta_x(\gamma_1+L)}\right\}, \tag{49}$$

we obtain

$$\begin{aligned}
&\mathcal{L}_{t+1}-\mathcal{L}_t \leq -\frac{\beta_x\eta}{4}\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2] - \frac{\beta_x c_{\beta_y}\eta}{2}\mathbb{E}[\|\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2] \\
&- \gamma_1 c_{\hat{\beta}_x}\frac{\beta_x\eta}{4}\mathbb{E}[\|\bar{x}_{t+1}-\bar{\hat{x}}_t\|^2] - \gamma_2 c_{\hat{\beta}_y}\frac{\beta_x\eta}{4}\mathbb{E}[\|\bar{y}_{t+1}-\bar{\hat{y}}_t\|^2] - \frac{\beta_x\eta}{8}\mathbb{E}[\|\bar{u}_t\|^2] - \frac{\beta_x c_{\beta_y}\eta}{32}\mathbb{E}[\|\bar{v}_t\|^2] \\
&+ 2\beta_x\rho_x\eta^3\sigma^2\frac{1}{K} + 5\beta_x c_{\beta_y}\rho_y\eta^3\sigma^2\frac{1}{K} + 9\beta_x\rho_x^2\eta^5\sigma^2 + 9c_{\beta_y}\beta_x\rho_y^2\eta^5\sigma^2 \,.
\end{aligned} \tag{50}$$

Because

$$\begin{aligned}
\|\nabla_x f(\bar{x}_t,\bar{y}_t)\|^2 &\leq 2\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2 + 4\gamma_1^2\|\bar{x}_{t+1}-\bar{\hat{x}}_t\|^2 + 4\gamma_1^2\|\bar{x}_{t+1}-\bar{x}_t\|^2 \\
&\leq 4\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2 + 4\gamma_1^2\|\bar{x}_{t+1}-\bar{\hat{x}}_t\|^2 + 4\beta_x^2\eta^2\gamma_1^2\|\bar{u}_t\|^2 \,, \\
&\leq 4\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2 + 4\gamma_1^2\|\bar{x}_{t+1}-\bar{\hat{x}}_t\|^2 + 4\gamma_1^2\|\bar{u}_t\|^2 \,, \\
\|\nabla_y f(\bar{x}_t,\bar{y}_t)\|^2 &\leq 2\|\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2 + 2\gamma_2^2\|\bar{y}_{t+1}-\bar{\hat{y}}_t\|^2 \\
&\leq 4\|\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2 + 4\gamma_2^2\|\bar{y}_{t+1}-\bar{\hat{y}}_t\|^2 + 4\gamma_2^2\|\bar{v}_t\|^2 \,,
\end{aligned}$$

where the last step follows from $\eta \leq \frac{1}{\beta_x}$ and $\eta \leq \frac{1}{\beta_y}$, we obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}[\|\nabla_x f(\bar{x}_t,\bar{y}_t)\|^2] + \kappa\mathbb{E}[\|\nabla_y f(\bar{x}_t,\bar{y}_t)\|^2]\right) \\
&\leq \frac{1}{T}\sum_{t=0}^{T-1}\Big(4\mathbb{E}[\|\nabla_x F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2] + 4\kappa\mathbb{E}[\|\nabla_y F(\bar{x}_t,\bar{y}_t;\bar{\hat{x}}_t,\bar{\hat{y}}_t)\|^2] \\
&\qquad + 4\gamma_1^2\mathbb{E}[\|\bar{x}_{t+1}-\bar{\hat{x}}_t\|^2] + 4\kappa\gamma_2^2\mathbb{E}[\|\bar{y}_{t+1}-\bar{\hat{y}}_t\|^2] + 4\gamma_1^2\mathbb{E}[\|\bar{u}_t\|^2] + 4\kappa\gamma_2^2\mathbb{E}[\|\bar{v}_t\|^2]\Big) \\
&\leq \max\left\{\frac{128}{\beta_x\eta}, \frac{128\kappa}{\beta_x\eta c_{\beta_y}}, \frac{128\gamma_1}{\beta_x\eta c_{\hat{\beta}_x}}, \frac{128\kappa\gamma_2}{\beta_x\eta c_{\hat{\beta}_y}}, \frac{128\gamma_1^2}{\beta_x\eta}, \frac{128\kappa\gamma_2^2}{\beta_x\eta c_{\beta_y}}\right\}\Bigg(\frac{\mathcal{L}_0-\mathcal{L}_T}{T} \\
&\qquad + 2\beta_x\rho_x\eta^3\sigma^2\frac{1}{K} + 5\beta_x c_{\beta_y}\rho_y\eta^3\sigma^2\frac{1}{K} + 9\beta_x\rho_x^2\eta^5\sigma^2 + 9c_{\beta_y}\beta_x\rho_y^2\eta^5\sigma^2\Bigg) \,.
\end{aligned} \tag{51}$$

By setting $\gamma_1 = O(L)$, $\gamma_2 = O(L)$, we obtain

$$c_{\beta_y} = O(1)\,, \quad c_{\hat{\beta}_x} = O\left(\frac{1}{L^2\kappa}\right)\,, \quad c_{\hat{\beta}_y} = O(1)\,.$$

Since

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_0 - p_0^{(k)}\|^2] \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\frac{1}{K}\sum_{j=1}^{K}\nabla_x F^{(j)}(x_0,y_0;\hat{x}_0,\hat{y}_0;\xi_0^{(j)}) - \nabla_x F^{(k)}(x_0,y_0;\hat{x}_0,\hat{y}_0;\xi_0^{(k)})\|^2]
\end{aligned}$$

$$
\begin{aligned}
&\leq 3\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\frac{1}{K}\sum_{j=1}^{K}\nabla_x F^{(j)}(x_0,y_0;\hat{x}_0,\hat{y}_0;\xi_0^{(j)})-\frac{1}{K}\sum_{j=1}^{K}\nabla_x F^{(j)}(x_0,y_0;\hat{x}_0,\hat{y}_0)\|^2]\\
&\quad+3\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\frac{1}{K}\sum_{j=1}^{K}\nabla_x F^{(j)}(x_0,y_0;\hat{x}_0,\hat{y}_0)-\nabla_x F^{(k)}(x_0,y_0;\hat{x}_0,\hat{y}_0)\|^2]\\
&\quad+3\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_x F^{(k)}(x_0,y_0;\hat{x}_0,\hat{y}_0)-\nabla_x F^{(k)}(x_0,y_0;\hat{x}_0,\hat{y}_0;\xi_0^{(k)})\|^2]\\
&\leq 6\sigma^2+6\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_x f^{(k)}(x_0,y_0)\|^2]\,,
\end{aligned}
$$

and

$$
\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{q}_0-q_0^{(k)}\|^2]\leq 6\sigma^2+6\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_y f^{(k)}(x_0,y_0)\|^2]\,,
$$

we have

$$
\begin{aligned}
\mathcal{L}_0&=\mathcal{P}_0+\frac{\beta_x}{\rho_x\eta}\mathbb{E}[\|\bar{\mathcal{G}}_{x,0}\|^2]+\frac{5\beta_y}{2\rho_y\eta}\mathbb{E}[\|\bar{\mathcal{G}}_{x,0}\|^2]+3\beta_x\eta\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\hat{\mathcal{G}}_{x,0}^{(k)}\|^2]+3\beta_y\eta\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\hat{\mathcal{G}}_{y,0}^{(k)}\|^2]\\
&\quad+\beta_x\eta(1-\lambda)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{p}_0-p_0^{(k)}\|^2]+\beta_y\eta(1-\lambda)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{q}_0-q_0^{(k)}\|^2]\\
&\leq\mathcal{P}_0+\frac{\beta_x}{\rho_x\eta}\frac{\sigma^2}{B}+\frac{5\beta_y}{2\rho_y\eta}\frac{\sigma^2}{B}+9\beta_x\eta\sigma^2+9\beta_y\eta\sigma^2+6\beta_x\eta\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_x f^{(k)}(x_0,y_0)\|^2]\\
&\quad+6\beta_y\eta\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_y f^{(k)}(x_0,y_0)\|^2]\,.
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}[\|\nabla_x f(\bar{x}_t,\bar{y}_t)\|^2]+\kappa\mathbb{E}[\|\nabla_y f(\bar{x}_t,\bar{y}_t)\|^2]\right)&\leq O(\kappa\rho_x^2\eta^4\sigma^2)+O(\kappa\rho_y^2\eta^4\sigma^2)+O\left(\frac{\kappa\mathcal{P}_0}{\beta_x\eta T}\right)\\
&\quad+O\left(\frac{\kappa}{T}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_x f^{(k)}(x_0,y_0)\|^2]\right)+O\left(\frac{\kappa}{T}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_y f^{(k)}(x_0,y_0)\|^2]\right)\\
&\quad+O\left(\frac{\kappa\sigma^2}{\rho_x\eta^2TB}\right)+O\left(\frac{\kappa\sigma^2}{\rho_y\eta^2TB}\right)+O\left(\frac{\kappa\sigma^2}{T}\right)+O\left(\frac{\kappa\rho_x\eta^2\sigma^2}{K}\right)+O\left(\frac{\kappa\rho_y\eta^2\sigma^2}{K}\right)\,.
\end{aligned}
$$

By setting $\beta_x=O((1-\lambda)^2)$, $\eta=O(\frac{K\epsilon}{\kappa^{1/2}})$, $\rho_x=O(\frac{1}{K})$, $\rho_y=O(\frac{1}{K})$, $B=O(\frac{\kappa^{1/2}}{\epsilon})$, $T=O(\frac{\kappa^{3/2}}{K(1-\lambda)^2\epsilon^3})$, we have

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

F Translation of Stationarity Measures

To perform translation between different stationarity measures defined in Definition 4.1 in the decentralized setting, we use the Decentralized Stochastic Gradient Descent Ascent (DSGDA) algorithm in Algorithm 2 to solve the following two-sided PL minimax problem:

$$\min_{x\in\mathbb{R}^{d_1}}\max_{y\in\mathbb{R}^{d_2}}\hat{f}(x,y)\triangleq\frac{1}{K}\sum_{k=1}^{K}\underbrace{f^{(k)}(x,y)+L\|x-\tilde{x}\|^2}_{\hat{f}^{(k)}(x,y)}, \tag{52}$$

where $\tilde{x}$ is the output of Algorithm 1. Note that, $\hat{f}(x,y)$ is L-strongly convex about x, μ-PL about y and $\hat{L}$-smooth, where $\hat{L}=3L$. In this section, we will establish the convergence rate of Algorithm 2 for this strongly-convex-PL minimax problem, showing that the translation stage does not hurt the convergence rate of Algorithm 1.

Algorithm 2 Decentralized Stochastic Gradient Descent Ascent Algorithm

Input: $\eta_x>0$, $\eta_y>0$. Initialization on worker k: $x_0^{(k)}=x_0$, $y_0^{(k)}=y_0$,

1: **for** $t=0,\cdots,T-1$, worker k **do**
2: Update x: $x_{t+1}^{(k)}=\sum_{j\in\mathcal{N}_k}w_{kj}x_t^{(j)}-\eta_x\nabla_1\hat{f}^{(k)}(x_t^{(k)},y_t^{(k)};\xi_t^{(k)})$,
3: Update y: $y_{t+1}^{(k)}=\sum_{j\in\mathcal{N}_k}w_{kj}y_t^{(j)}+\eta_y\nabla_2\hat{f}^{(k)}(x_t^{(k)},y_t^{(k)};\xi_t^{(k)})$,
4: **end for**

In the first place, we introduce the heterogeneity assumption for the loss function gradient in Eq. (52).

Assumption F.1. The gradient heterogeneity of the loss function gradient in Eq. (52) is upper bounded as follows:

$$\|\nabla_1\hat{f}^{(k)}(x,y)-\nabla_1\hat{f}(x,y)\|^2\le\delta^2,\quad\|\nabla_2\hat{f}^{(k)}(x,y)-\nabla_2\hat{f}(x,y)\|^2\le\delta^2,$$

where $\delta>0$ is a constant.

Theorem F.2. *Given the Assumption F.1 and the smoothness and strongly-convex-PL conditions of $\hat{f}(x,y)$, setting $\eta_x=O(\epsilon^2/\kappa^5)$, $\eta_y=O(\kappa^2)\eta_x$, and $T=O(\epsilon^{-2}\kappa^6\log(\kappa))$, Algorithm 2 guarantees*

$$\mathbb{E}[\|x_T-x^*\|^2]\le O(\epsilon^2), \tag{53}$$

where $0<\epsilon<1$ is a constant.

Based on Theorem F.2, we will derive that Theorem 4.2 and Theorem F.2 guarantee the stationary point satisfying $\mathbb{E}[\|\nabla\Phi(\hat{x})\|]\le\epsilon$, where $\hat{x}$ is the output of Algorithm 2. Specifically, by defining $\Phi_{1/2L}(\tilde{x})=\min_z\Phi(z)+L\|z-\tilde{x}\|^2$, Part (b) of Proof of Proposition 2.1 in [34] shows that

$$\begin{aligned}\|\nabla\Phi_{1/2L}(\tilde{x})\|^2&\le4\|\nabla_1 f(\tilde{x},\tilde{y})\|^2+4\kappa\|\nabla_2 f(\tilde{x},\tilde{y})\|^2,\\ \|\nabla\Phi(\hat{x})\|&\le\|\nabla\Phi_{1/2L}(\tilde{x})\|+2\kappa L\|\hat{x}-x^*(\tilde{x})\|,\end{aligned} \tag{54}$$

where $(\tilde{x},\tilde{y})$ is the output of Algorithm 1, $\hat{x}$ is the output of Algorithm 2 given the initial point $\tilde{x}$.

As our Theorem 4.2 guarantees $4\|\nabla_1 f(\tilde{x},\tilde{y})\|^2+4\kappa\|\nabla_2 f(\tilde{x},\tilde{y})\|^2\le O(\epsilon^2)$ and Theorem F.2 guarantees $\|\hat{x}-x^*(\tilde{x})\|^2\le O(\epsilon^2)$, we can conclude that $\|\nabla\Phi(\hat{x})\|\le O(\epsilon)$. The convergence rate of Algorithm 1, $T=O\left(\frac{\kappa^{3/2}}{K(1-\lambda)^2\epsilon^3}\right)$, dominates the convergence rate of the translation Algorithm 2, $T=O\left(\frac{\kappa^6\log(\kappa)}{\epsilon^2}\right)$. Therefore, the iteration complexity to find the $O(\epsilon)$-stationary solution is still $T=O\left(\frac{\kappa^{3/2}}{K(1-\lambda)^2\epsilon^3}\right)$.

To prove Theorem F.2, we introduce the following auxiliary function:

$$\begin{aligned}\hat{\Phi}(x)&=\max_y\hat{f}(x,y)=\max_y f(x,y)+L\|x-\tilde{x}\|^2,\\ \hat{\Psi}(y)&=\min_x\hat{f}(x,y)=\min_x f(x,y)+L\|x-\tilde{x}\|^2,\\ \hat{\Phi}^*&=\max_y\hat{\Psi}(y).\end{aligned} \tag{55}$$

Then, we are ready to prove Theorem F.2.

Proof. From the $\hat{L}$-smooth of $\hat{f}(x,y)$ with respect to y, we derive

$$\begin{aligned}
&\mathbb{E}[\hat{f}(\bar{x}_{t+1},\bar{y}_t)] \le \mathbb{E}[\hat{f}(\bar{x}_{t+1},\bar{y}_{t+1})] - \mathbb{E}[\langle \nabla_2 \hat{f}(\bar{x}_{t+1},\bar{y}_t), \bar{y}_{t+1}-\bar{y}_t\rangle] + \frac{\hat{L}}{2}\mathbb{E}[\|\bar{y}_{t+1}-\bar{y}_t\|^2] \\
&= \mathbb{E}[\hat{f}(\bar{x}_{t+1},\bar{y}_{t+1})] - \eta_y \mathbb{E}[\langle \nabla_2 \hat{f}(\bar{x}_{t+1},\bar{y}_t), \frac{1}{K}\sum_{k=1}^K \nabla_2 \hat{f}^{(k)}(x_t^{(k)},y_t^{(k)};\xi_t^{(k)})\rangle] \\
&\quad + \frac{\eta_y^2\hat{L}}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_2 \hat{f}^{(k)}(x_t^{(k)},y_t^{(k)};\xi_t^{(k)})\|^2] \\
&= \mathbb{E}[\hat{f}(\bar{x}_{t+1},\bar{y}_{t+1})] - \eta_y \mathbb{E}[\langle \nabla_2 \hat{f}(\bar{x}_{t+1},\bar{y}_t), \frac{1}{K}\sum_{k=1}^K \nabla_2 \hat{f}^{(k)}(x_t^{(k)},y_t^{(k)})\rangle] \\
&\quad + \frac{\eta_y^2\hat{L}}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_2 \hat{f}^{(k)}(x_t^{(k)},y_t^{(k)})\|^2] + \frac{\eta_y^2\hat{L}\sigma^2}{2} \\
&\le \mathbb{E}[\hat{f}(\bar{x}_{t+1},\bar{y}_{t+1})] - \frac{\eta_y}{2}\mathbb{E}[\|\nabla_2 \hat{f}(\bar{x}_{t+1},\bar{y}_t)\|^2] - \frac{\eta_y}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_2 \hat{f}^{(k)}(x_t^{(k)},y_t^{(k)})\|^2] \\
&\quad + \eta_y \mathbb{E}[\|\nabla_2 \hat{f}(\bar{x}_{t+1},\bar{y}_t) - \nabla_2 \hat{f}(\bar{x}_t,\bar{y}_t)\|^2] + \eta_y \mathbb{E}[\|\nabla_2 \hat{f}(\bar{x}_t,\bar{y}_t) - \frac{1}{K}\sum_{k=1}^K \nabla_2 \hat{f}^{(k)}(x_t^{(k)},y_t^{(k)})\|^2] \\
&\quad + \frac{\eta_y^2\hat{L}}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_2 \hat{f}^{(k)}(x_t^{(k)},y_t^{(k)})\|^2] + \frac{\eta_y^2\hat{L}\sigma^2}{2}\,.
\end{aligned}$$

With $\eta_y \le \frac{1}{2\hat{L}}$, and applying

$$\begin{aligned}
&-\frac{\eta_y}{2}\mathbb{E}[\|\nabla_2 \hat{f}(\bar{x}_{t+1},\bar{y}_t)\|^2] + \eta_y \mathbb{E}[\|\nabla_2 \hat{f}(\bar{x}_{t+1},\bar{y}_t) - \nabla_2 \hat{f}(\bar{x}_t,\bar{y}_t)\|^2] \\
&\le -\frac{\eta_y}{4}\mathbb{E}[\|\nabla_2 \hat{f}(\bar{x}_t,\bar{y}_t)\|^2] + \frac{3\eta_y}{2}\mathbb{E}[\|\nabla_2 \hat{f}(\bar{x}_{t+1},\bar{y}_t) - \nabla_2 \hat{f}(\bar{x}_t,\bar{y}_t)\|^2]\,, \\
&\le -\frac{\eta_y}{4}\mathbb{E}[\|\nabla_2 \hat{f}(\bar{x}_t,\bar{y}_t)\|^2] + \frac{3\eta_y\eta_x^2\hat{L}^2}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_1 \hat{f}(x_t^{(k)},y_t^{(k)};\xi_t^{(k)})\|^2] \\
&\le -\frac{\eta_y}{4}\mathbb{E}[\|\nabla_2 \hat{f}(\bar{x}_t,\bar{y}_t)\|^2] + \frac{3\eta_y\eta_x^2\hat{L}^2}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_1 \hat{f}(x_t^{(k)},y_t^{(k)})\|^2] + \frac{3\eta_y\eta_x^2\hat{L}^2\sigma^2}{2}\,,
\end{aligned}$$

and

$$\begin{aligned}
&\eta_y \mathbb{E}[\|\nabla_2 \hat{f}(\bar{x}_t,\bar{y}_t) - \frac{1}{K}\sum_{k=1}^K \nabla_2 \hat{f}^{(k)}(x_t^{(k)},y_t^{(k)})\|^2] \\
&\le \eta_y \hat{L}^2 \frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2] + \eta_y \hat{L}^2 \frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|^2]\,,
\end{aligned}$$

we obtain

$$\begin{aligned}
&\mathbb{E}[\hat{f}(\bar{x}_{t+1},\bar{y}_t)] \le \mathbb{E}[\hat{f}(\bar{x}_{t+1},\bar{y}_{t+1})] - \frac{\eta_y}{4}\mathbb{E}[\|\nabla_2 \hat{f}(\bar{x}_t,\bar{y}_t)\|^2] + \frac{3\eta_y\eta_x^2\hat{L}^2}{2}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_1 \hat{f}(x_t^{(k)},y_t^{(k)})\|^2] \\
&- \frac{\eta_y}{4}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_2 \hat{f}^{(k)}(x_t^{(k)},y_t^{(k)})\|^2] + \eta_y \hat{L}^2 \frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2] + \eta_y \hat{L}^2 \frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|^2] \\
&+ \frac{3\eta_y\eta_x^2\hat{L}^2\sigma^2}{2} + \frac{\eta_y^2\hat{L}\sigma^2}{2}\,.
\end{aligned}$$

Similarly, from the $\hat{L}$-smooth of $\hat{f}(x,y)$ with respect to x, we derive

$$\mathbb{E}[\hat{f}(\bar{x}_t,\bar{y}_t)] \le \mathbb{E}[\hat{f}(\bar{x}_{t+1},\bar{y}_t)] - \mathbb{E}[\langle \nabla_1 \hat{f}(\bar{x}_t,\bar{y}_t), \bar{x}_{t+1}-\bar{x}_t\rangle] + \frac{\hat{L}}{2}\mathbb{E}[\|\bar{x}_{t+1}-\bar{x}_t\|^2]$$

$$= \mathbb{E}[\hat{f}(\bar{x}_{t+1}, \bar{y}_t)] + \eta_x \mathbb{E}[\langle \nabla_1 \hat{f}(\bar{x}_t, \bar{y}_t), \frac{1}{K}\sum_{k=1}^{K} \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_t^{(k)})\rangle]$$

$$+ \frac{\eta_x^2 \hat{L}}{2} \mathbb{E}[\| \frac{1}{K}\sum_{k=1}^{K} \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_t^{(k)})\|^2]$$

$$= \mathbb{E}[\hat{f}(\bar{x}_{t+1}, \bar{y}_t)] + \eta_x \mathbb{E}[\langle \nabla_1 \hat{f}(\bar{x}_t, \bar{y}_t), \frac{1}{K}\sum_{k=1}^{K} \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)})\rangle]$$

$$+ \frac{\eta_x^2 \hat{L}}{2} \mathbb{E}[\| \frac{1}{K}\sum_{k=1}^{K} \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)})\|^2] + \frac{\eta_x^2 \hat{L}\sigma^2}{2}$$

$$\leq \mathbb{E}[\hat{f}(\bar{x}_{t+1}, \bar{y}_t)] + \frac{\eta_x}{2} \mathbb{E}[\|\nabla_1 \hat{f}(\bar{x}_t, \bar{y}_t)\|^2] + (\frac{\eta_x}{2} + \frac{\eta_x^2 \hat{L}}{2}) \mathbb{E}[\| \frac{1}{K}\sum_{k=1}^{K} \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)})\|^2] + \frac{\eta_x^2 \hat{L}\sigma^2}{2} .$$

Combining these two inequalities with $\frac{3\eta_y \eta_x^2 \hat{L}^2}{2} \leq \frac{3\eta_x^2 \hat{L}}{4}$, we have

$$\mathbb{E}[\hat{f}(\bar{x}_t, \bar{y}_t)] \leq \mathbb{E}[\hat{f}(\bar{x}_{t+1}, \bar{y}_{t+1})] + \frac{\eta_x}{2} \mathbb{E}[\|\nabla_1 \hat{f}(\bar{x}_t, \bar{y}_t)\|^2] - \frac{\eta_y}{4} \mathbb{E}[\|\nabla_2 \hat{f}(\bar{x}_t, \bar{y}_t)\|^2]$$

$$+ (\frac{\eta_x}{2} + \frac{5\eta_x^2 \hat{L}}{4}) \mathbb{E}[\| \frac{1}{K}\sum_{k=1}^{K} \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)})\|^2] - \frac{\eta_y}{4} \mathbb{E}[\| \frac{1}{K}\sum_{k=1}^{K} \nabla_2 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)})\|^2]$$

$$+ \eta_y \hat{L}^2 \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2] + \eta_y \hat{L}^2 \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|^2] + \frac{5\eta_x^2 \hat{L}\sigma^2}{4} + \frac{\eta_y^2 \hat{L}\sigma^2}{2} . \quad (56)$$

From the $L_{\hat{\Phi}}$-smoothness of $\hat{\Phi}(x)$, where $L_{\hat{\Phi}} = O(\kappa)$ [20], we derive

$$\mathbb{E}[\hat{\Phi}(\bar{x}_{t+1})] \leq \mathbb{E}[\hat{\Phi}(\bar{x}_t)] + \mathbb{E}[\langle \nabla \hat{\Phi}(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t\rangle] + \frac{L_{\hat{\Phi}}}{2} \mathbb{E}[\|\bar{x}_{t+1} - \bar{x}_t\|^2]$$

$$= \mathbb{E}[\hat{\Phi}(\bar{x}_t)] - \eta_x \mathbb{E}[\langle \nabla \hat{\Phi}(\bar{x}_t), \frac{1}{K}\sum_{k=1}^{K} \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_t^{(k)})\rangle]$$

$$+ \frac{\eta_x^2 L_{\hat{\Phi}}}{2} \mathbb{E}[\| \frac{1}{K}\sum_{k=1}^{K} \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_t^{(k)})\|^2]$$

$$= \mathbb{E}[\hat{\Phi}(\bar{x}_t)] - \eta_x \mathbb{E}[\langle \nabla \hat{\Phi}(\bar{x}_t), \frac{1}{K}\sum_{k=1}^{K} \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)})\rangle] + \frac{\eta_x^2 L_{\hat{\Phi}}}{2} \mathbb{E}[\| \frac{1}{K}\sum_{k=1}^{K} \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)})\|^2]$$

$$+ \frac{\eta_x^2 L_{\hat{\Phi}} \sigma^2}{2} .$$

For the second term, we obtain

$$- \eta_x \mathbb{E}[\langle \nabla \hat{\Phi}(\bar{x}_t), \frac{1}{K}\sum_{k=1}^{K} \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)})\rangle]$$

$$= -\eta_x \mathbb{E}[\langle \nabla \hat{\Phi}(\bar{x}_t) - \nabla_1 \hat{f}(\bar{x}_t, \bar{y}_t), \frac{1}{K}\sum_{k=1}^{K} \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)})\rangle]$$

$$- \eta_x \mathbb{E}[\langle \nabla_1 \hat{f}(\bar{x}_t, \bar{y}_t), \frac{1}{K}\sum_{k=1}^{K} \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)})\rangle]$$

$$\leq 8\eta_x \mathbb{E}[\|\nabla \hat{\Phi}(\bar{x}_t) - \nabla_1 \hat{f}(\bar{x}_t, \bar{y}_t)\|^2] + \frac{\eta_x}{8} \mathbb{E}[\| \frac{1}{K}\sum_{k=1}^{K} \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)})\|^2] - \frac{\eta_x}{2} \mathbb{E}[\|\nabla_1 \hat{f}(\bar{x}_t, \bar{y}_t)\|^2]$$

$$- \frac{\eta_x}{2} \mathbb{E}[\| \frac{1}{K}\sum_{k=1}^{K} \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)})\|^2] + \frac{\eta_x}{2} \mathbb{E}[\|\nabla_1 \hat{f}(\bar{x}_t, \bar{y}_t) - \frac{1}{K}\sum_{k=1}^{K} \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)})\|^2] .$$

Since $\hat{f}(x,y)$ is μ-PL about y, combining these inequalities yields

$$\begin{aligned}
&\mathbb{E}[\hat{\Phi}(\bar{x}_{t+1})] \leq \mathbb{E}[\hat{\Phi}(\bar{x}_t)] + 8\eta_x \frac{\hat{L}^2}{\mu^2}\mathbb{E}[\|\nabla_2 \hat{f}(\bar{x}_t,\bar{y}_t)\|^2] - \frac{\eta_x}{2}\mathbb{E}[\|\nabla_1 \hat{f}(\bar{x}_t,\bar{y}_t)\|^2] \\
&+ (\frac{\eta_x^2 L_{\hat{\Phi}}}{2} - \frac{3\eta_x}{8})\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)})\|^2] + \frac{\eta_x \hat{L}^2}{2}\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2] \\
&+ \frac{\eta_x \hat{L}^2}{2}\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|^2] + \frac{\eta_x^2 L_{\hat{\Phi}}\sigma^2}{2} \,. \qquad (57)
\end{aligned}$$

Then, we introduce a potential function as follows:

$$\mathcal{P}_t = \hat{\Phi}(\bar{x}_t) - \hat{\Phi}^* + \hat{\Phi}(\bar{x}_t) - \hat{f}(\bar{x}_t,\bar{y}_t) = 2\hat{\Phi}(\bar{x}_t) - \hat{\Phi}^* - \hat{f}(\bar{x}_t,\bar{y}_t) \,,$$

Applying Eq. (56) and Eq. (57), and setting $\eta_x = \frac{\eta_y}{128\kappa^2}$ where $\kappa = \frac{\hat{L}}{\mu} > 1$, we obtain

$$\begin{aligned}
&\mathcal{P}_{t+1} - \mathcal{P}_t = 2\hat{\Phi}(\bar{x}_{t+1}) - \hat{\Phi}^* - \hat{f}(\bar{x}_{t+1},\bar{y}_{t+1}) - (2\hat{\Phi}(\bar{x}_t) - \hat{\Phi}^* - \hat{f}(\bar{x}_t,\bar{y}_t)) \\
&= 2(\hat{\Phi}(\bar{x}_{t+1}) - \hat{\Phi}(\bar{x}_t)) + \hat{f}(\bar{x}_t,\bar{y}_t) - \hat{f}(\bar{x}_{t+1},\bar{y}_{t+1}) \\
&\leq -\frac{\eta_x}{2}\mathbb{E}[\|\nabla_1 \hat{f}(\bar{x}_t,\bar{y}_t)\|^2] - \frac{\eta_y}{8}\mathbb{E}[\|\nabla_2 \hat{f}(\bar{x}_t,\bar{y}_t)\|^2] + \eta_x^2 L_{\hat{\Phi}}\sigma^2 + \frac{5\eta_x^2\hat{L}\sigma^2}{4} + \frac{\eta_y^2\hat{L}\sigma^2}{2} \\
&\quad + (\eta_x+\eta_y)\hat{L}^2\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2] + (\eta_x+\eta_y)\hat{L}^2\frac{1}{K}\sum_{k=1}^K \mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|^2] \\
&\quad + (\frac{\eta_x}{2} + \frac{5\eta_x^2\hat{L}}{4} + \eta_x^2 L_{\hat{\Phi}} - \frac{3\eta_x}{4})\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)})\|^2] \\
&\quad - \frac{\eta_y}{4}\mathbb{E}[\|\frac{1}{K}\sum_{k=1}^K \nabla_2 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)})\|^2] \,. \qquad (58)
\end{aligned}$$

By setting $\eta_x \leq \frac{1}{2(5\hat{L}+4L_{\hat{\Phi}})}$, we have $\frac{5\eta_x^2\hat{L}}{4} + \eta_x^2 L_{\hat{\Phi}} - \frac{\eta_x}{4} \leq -\frac{\eta_x}{8}$.

In particular, since $\hat{f}(x,y)$ is μ-PL about y, we obtain

$$\begin{aligned}
&\mathbb{E}[\|\nabla\hat{\Phi}(\bar{x}_t)\|^2] \leq 2\mathbb{E}[\|\nabla\hat{\Phi}(\bar{x}_t) - \nabla_1\hat{f}(\bar{x}_t,\bar{y}_t)\|^2] + 2\mathbb{E}[\|\nabla_1\hat{f}(\bar{x}_t,\bar{y}_t)\|^2] \\
&\leq 2\hat{L}^2\mathbb{E}[\|y^*(\bar{x}_t) - \bar{y}_t\|^2] + 2\mathbb{E}[\|\nabla_1\hat{f}(\bar{x}_t,\bar{y}_t)\|^2] \\
&\leq \frac{2\hat{L}^2}{\mu^2}\mathbb{E}[\|\nabla_2\hat{f}(\bar{x}_t,\bar{y}_t)\|^2] + 2\mathbb{E}[\|\nabla_1\hat{f}(\bar{x}_t,\bar{y}_t)\|^2] \,.
\end{aligned}$$

Therefore, we derive

$$\begin{aligned}
&\frac{\eta_x}{2}\mathbb{E}[\|\nabla_1\hat{f}(\bar{x}_t,\bar{y}_t)\|^2] + \frac{\eta_y}{8}\mathbb{E}[\|\nabla_2\hat{f}(\bar{x}_t,\bar{y}_t)\|^2] \\
&\geq \frac{\eta_x}{4}\mathbb{E}[\|\nabla\hat{\Phi}(\bar{x}_t)\|^2] - \frac{\eta_x}{2}\frac{\hat{L}^2}{\mu^2}\mathbb{E}[\|\nabla_2\hat{f}(\bar{x}_t,\bar{y}_t)\|^2] + \frac{\eta_y}{8}\mathbb{E}[\|\nabla_2\hat{f}(\bar{x}_t,\bar{y}_t)\|^2] \\
&\geq \frac{\eta_x}{16}\mathbb{E}[\|\nabla\hat{\Phi}(\bar{x}_t)\|^2] + \frac{\eta_y}{16}\mathbb{E}[\|\nabla_2\hat{f}(\bar{x}_t,\bar{y}_t)\|^2] \\
&\geq \frac{\eta_x\mu}{8}(\hat{\Phi}(\bar{x}_t) - \Phi^*) + \frac{\eta_y\mu}{8}(\hat{\Phi}(x_t) - \hat{f}(\bar{x}_t,\bar{y}_t)) \\
&\geq \frac{\eta_x\mu}{8}(\hat{\Phi}(\bar{x}_t) - \Phi^*) + \frac{\eta_x\mu}{8}(\hat{\Phi}(x_t) - \hat{f}(\bar{x}_t,\bar{y}_t)) \\
&= \frac{\eta_x\mu}{8}\mathcal{P}_t \,,
\end{aligned}$$

where the second inequality from $\eta_x = \frac{\eta_y}{128\kappa^2}$, the third inequality holds due to $\hat{f}(x,y)$ is μ-PL about y, and the fourth inequality from $\kappa > 1$. Then, by reformulating Eq. (58), we derive

$$\begin{aligned}
\frac{\eta_x \mu}{8}\mathcal{P}_t &\leq \frac{\eta_x}{2}\mathbb{E}[\|\nabla_1 \hat{f}(\bar{x}_t, \bar{y}_t)\|^2] + \frac{\eta_y}{8}\mathbb{E}[\|\nabla_2 \hat{f}(\bar{x}_t, \bar{y}_t)\|^2] \\
&\leq \mathcal{P}_t - \mathcal{P}_{t+1} + \eta_x^2 L_{\hat{\Phi}}\sigma^2 + \frac{5\eta_x^2 \hat{L}\sigma^2}{4} + \frac{\eta_y^2 \hat{L}\sigma^2}{2} \\
&\quad + (\eta_x + \eta_y)\hat{L}^2\Big(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|^2]\Big) .
\end{aligned} \tag{59}$$

In the following, we provide the upper bound of the last term. At first, from the update rule of X_t, we derive

$$\begin{aligned}
&\mathbb{E}[\|X_t - \bar{X}_t\|_F^2] = \mathbb{E}[\|X_{t-1}W - \eta_x P_{t-1} - \bar{X}_{t-1} + \eta_x \bar{P}_{t-1}\|_F^2] \\
&\leq (1+a)\mathbb{E}[\|X_{t-1}W - \bar{X}_{t-1}\|_F^2] + \eta_x^2(1+1/a)\mathbb{E}[\|P_{t-1} - \bar{P}_{t-1}\|_F^2] \\
&\leq (1+a)\lambda^2\mathbb{E}[\|X_{t-1} - \bar{X}_{t-1}\|_F^2] + \eta_x^2(1+1/a)\mathbb{E}[\|P_{t-1} - \bar{P}_{t-1}\|_F^2] \\
&\leq \lambda\mathbb{E}[\|X_{t-1} - \bar{X}_{t-1}\|_F^2] + \frac{\eta_x^2}{1-\lambda}\mathbb{E}[\|P_{t-1} - \bar{P}_{t-1}\|_F^2] ,
\end{aligned}$$

where

$$\begin{aligned}
P_t &= [\nabla_1 \hat{f}^{(1)}(x_t^{(1)}, y_t^{(1)}; \xi_t^{(1)}), \cdots, \nabla_1 \hat{f}^{(K)}(x_t^{(K)}, y_t^{(K)}; \xi_t^{(K)})] , \\
\bar{P}_t &= [\frac{1}{K}\sum_{k=1}^{K}\nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_t^{(k)}), \cdots, \frac{1}{K}\sum_{k=1}^{K}\nabla_1 \hat{f}^{(k)}(x_t^{(k)}, y_t^{(k)}; \xi_t^{(k)})] .
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2] \leq \lambda\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t-1}^{(k)} - \bar{x}_{t-1}\|_F^2] \\
&\qquad + \frac{\eta_x^2}{1-\lambda}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 \hat{f}^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}; \xi_{t-1}^{(k)}) - \frac{1}{K}\sum_{k'=1}^{K}\nabla_1 \hat{f}^{(k')}(x_{t-1}^{(k')}, y_{t-1}^{(k')}; \xi_{t-1}^{(k')})\|^2] \\
&\leq \lambda\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t-1}^{(k)} - \bar{x}_{t-1}\|_F^2] + \frac{2\eta_x^2\sigma^2}{1-\lambda} \\
&\quad + \frac{\eta_x^2}{1-\lambda}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 \hat{f}^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}) - \frac{1}{K}\sum_{k'=1}^{K}\nabla_1 \hat{f}^{(k')}(x_{t-1}^{(k')}, y_{t-1}^{(k')})\|^2] \\
&\leq \lambda\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t-1}^{(k)} - \bar{x}_{t-1}\|_F^2] + \frac{2\eta_x^2\sigma^2}{1-\lambda} \\
&\quad + \frac{3\eta_x^2}{1-\lambda}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 \hat{f}^{(k)}(x_{t-1}^{(k)}, y_{t-1}^{(k)}) - \nabla_1 \hat{f}^{(k)}(\bar{x}_{t-1}, \bar{y}_{t-1})\|^2] \\
&\quad + \frac{3\eta_x^2}{1-\lambda}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 \hat{f}^{(k)}(\bar{x}_{t-1}, \bar{y}_{t-1}) - \nabla_1 \hat{f}(\bar{x}_{t-1}, \bar{y}_{t-1})\|^2] \\
&\quad + \frac{3\eta_x^2}{1-\lambda}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla_1 \hat{f}(\bar{x}_{t-1}, \bar{y}_{t-1}) - \frac{1}{K}\sum_{k'=1}^{K}\nabla_1 \hat{f}^{(k')}(x_{t-1}^{(k')}, y_{t-1}^{(k')})\|^2] \\
&\leq \lambda\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t-1}^{(k)} - \bar{x}_{t-1}\|_F^2] + \frac{3\eta_x^2\delta^2}{1-\lambda} + \frac{2\eta_x^2\sigma^2}{1-\lambda} \\
&\quad + \frac{6\eta_x^2\hat{L}^2}{1-\lambda}\Big(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t-1}^{(k)} - \bar{x}_{t-1}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_{t-1}^{(k)} - \bar{y}_{t-1}\|^2]\Big) .
\end{aligned}$$

Similarly, we have

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|^2] \le \lambda\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_{t-1}^{(k)} - \bar{y}_{t-1}\|_F^2] + \frac{3\eta_y^2\delta^2}{1-\lambda} + \frac{3\eta_y^2\sigma^2}{1-\lambda} + \frac{6\eta_y^2\hat{L}^2}{1-\lambda}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t-1}^{(k)} - \bar{x}_{t-1}\|^2] + \frac{6\eta_y^2\hat{L}^2}{1-\lambda}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_{t-1}^{(k)} - \bar{y}_{t-1}\|^2]\,.$$

Combining these two inequalities yields

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{x}_t - x_t^{(k)}\|^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\bar{y}_t - y_t^{(k)}\|^2] \\
&\le \left(\lambda + \frac{6(\eta_x^2+\eta_y^2)\hat{L}^2}{1-\lambda}\right)\left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|x_{t-1}^{(k)} - \bar{x}_{t-1}\|_F^2] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|y_{t-1}^{(k)} - \bar{y}_{t-1}\|_F^2]\right) \\
&\quad + \frac{3(\eta_x^2+\eta_y^2)(\delta^2+\sigma^2)}{1-\lambda} \\
&\le \left(\frac{3(\eta_x^2+\eta_y^2)(\delta^2+\sigma^2)}{1-\lambda}\right)\sum_{i=0}^{t-1}\left(\lambda + \frac{6(\eta_x^2+\eta_y^2)\hat{L}^2}{1-\lambda}\right)^{t-1-i} \\
&\le \left(\frac{3(\eta_x^2+\eta_y^2)(\delta^2+\sigma^2)}{1-\lambda}\right)\frac{1}{1-\lambda-\frac{6(\eta_x^2+\eta_y^2)\hat{L}^2}{1-\lambda}} \\
&\le \left(\frac{3(\eta_x^2+\eta_y^2)(\delta^2+\sigma^2)}{1-\lambda}\right)\frac{1-\lambda}{(1-\lambda)^2 - 6(\eta_x^2+\eta_y^2)\hat{L}^2} \\
&\le \left(\frac{6(\eta_x^2+\eta_y^2)(\delta^2+\sigma^2)}{1-\lambda}\right)\frac{1-\lambda}{(1-\lambda)^2} \\
&= \frac{6(\eta_x^2+\eta_y^2)(\delta^2+\sigma^2)}{(1-\lambda)^2}\,,
\end{aligned}$$

where the second-to-last step follows from $\eta_x \le \frac{1-\lambda}{5\hat{L}}$, and $\eta_y \le \frac{1-\lambda}{5\hat{L}}$, which implies $(1-\lambda)^2 - 6(\eta_x^2+\eta_y^2)\hat{L}^2 \ge \frac{(1-\lambda)^2}{2}$.

Therefore, plugging this upper bound into Eq. (59), we obtain

$$\begin{aligned}
&\mathcal{P}_{t+1} \le (1-\frac{\eta_x\mu}{8})\mathcal{P}_t + \eta_x^2 L_{\hat{\Phi}}\sigma^2 + \frac{5\eta_x^2\hat{L}\sigma^2}{4} + \frac{\eta_y^2\hat{L}\sigma^2}{2} + (\eta_x+\eta_y)\hat{L}^2\frac{6(\eta_x^2+\eta_y^2)(\delta^2+\sigma^2)}{(1-\lambda)^2} \\
&\le (1-\frac{\eta_x\mu}{8})^{t+1}\mathcal{P}_0 + \frac{8\eta_x^2 L_{\hat{\Phi}}\sigma^2}{\eta_x\mu} + \frac{8}{\eta_x\mu}\frac{5\eta_x^2\hat{L}\sigma^2}{4} + \frac{8}{\eta_x\mu}\frac{\eta_y^2\hat{L}\sigma^2}{2} + \frac{8}{\eta_x\mu}(\eta_x+\eta_y)\hat{L}^2\frac{6(\eta_x^2+\eta_y^2)(\delta^2+\sigma^2)}{(1-\lambda)^2} \\
&= (1-\frac{\eta_x\mu}{8})^{t+1}\mathcal{P}_0 + \frac{8\eta_x L_{\hat{\Phi}}\sigma^2}{\mu} + \frac{10\eta_x\hat{L}\sigma^2}{\mu} + \frac{4\eta_y^2\hat{L}\sigma^2}{\eta_x\mu} + (1+\frac{\eta_y}{\eta_x})\hat{L}^2\frac{48(\eta_x^2+\eta_y^2)(\delta^2+\sigma^2)}{(1-\lambda)^2\mu}\,.
\end{aligned}$$

Regarding $\mathcal{P}_0$, we have

$$\begin{aligned}
\mathcal{P}_0 &= 2\hat{\Phi}(\tilde{x}) - \hat{\Phi}^* - \hat{f}(\tilde{x},\tilde{y}) \\
&= 2\big(\hat{\Phi}(\tilde{x}) - \hat{f}(\tilde{x},\tilde{y})\big) + \big(\hat{f}(\tilde{x},\tilde{y}) - \hat{\Psi}(\tilde{y})\big) + \hat{\Psi}(\tilde{y}) - \hat{\Phi}^* \\
&\le 2\big(\hat{\Phi}(\tilde{x}) - \hat{f}(\tilde{x},\tilde{y})\big) + \big(\hat{f}(\tilde{x},\tilde{y}) - \hat{\Psi}(\tilde{y})\big) \\
&\le \frac{1}{\mu}\|\nabla_2\hat{f}(\tilde{x},\tilde{y})\|^2 + \frac{1}{2L}\|\nabla_1\hat{f}(\tilde{x},\tilde{y})\|^2 \\
&= \frac{1}{\mu}\|\nabla_2 f(\tilde{x},\tilde{y})\|^2 + \frac{1}{2L}\|\nabla_1 f(\tilde{x},\tilde{y})\|^2
\end{aligned}$$

$$\leq \frac{\epsilon^2}{\mu\kappa} + \frac{\epsilon^2}{2L} = \frac{3\epsilon^2}{2L} \leq \frac{2\epsilon^2}{L} \tag{60}$$

where the second inequality follows from the μ-PL condition of $\hat{f}$ with respect to y and the L-strong convexity of $\hat{f}$ with respect to x, and the third inequality holds due to Theorem 4.2.

Then, we have

$$\mathcal{P}_T \leq (1 - \frac{\eta_x\mu}{8})^T \frac{2\epsilon^2}{L} + \frac{8\eta_x L_{\hat{\Phi}}\sigma^2}{\mu} + \frac{10\eta_x\hat{L}\sigma^2}{\mu} + \frac{4\eta_y^2\hat{L}\sigma^2}{\eta_x\mu} + (1 + \frac{\eta_y}{\eta_x})\hat{L}^2\frac{48(\eta_x^2 + \eta_y^2)(\delta^2 + \sigma^2)}{(1-\lambda)^2\mu} .$$

According to the proof of Proposition 2.1 in [34], we have

$$\mathbb{E}[\|x_T - x^*\|^2] \leq O(\kappa)\mathcal{P}_T .$$

As a result, by setting $\eta_x = O(\eta_y/\kappa^2)$, we have

$$\mathbb{E}[\|x_T - x^*\|^2] \leq O(\kappa\epsilon^2)(1 - \frac{\eta_x\mu}{8})^T + O(\kappa^5\eta_x) + O(\kappa^7\eta_x^2/(1-\lambda)^2) .$$

By setting $\eta_x = O(\epsilon^2/\kappa^5)$ and $T = O(\epsilon^{-2}\kappa^6\log(\kappa))$, we can obtain

$$\mathbb{E}[\|x_T - x^*\|^2] \leq O(\epsilon^2) . \tag{61}$$

□

G Limitations and Future Work

Our theoretical analysis relies on standard assumptions in decentralized stochastic minimax optimization, including smoothness, the PL condition, bounded stochastic gradient variance (Assumptions 3.1-3.3). However, recent works have relaxed these assumptions to better capture practical applications, such as compressed communication [17, 16] and nonconvex-nonconcave objectives [48]. Moreover, our algorithm communicates the auxiliary variables $\hat{x}$ and $\hat{y}$ in each round, which increases the per-round communication cost. Its communication complexity also has a worse dependence on the spectral gap, i.e., $O(1/(1-\lambda)^2)$, than $O(1/\sqrt{1-\lambda})$ of DREAM [3], although DREAM requires multi-round communication in each iteration (see Table 1). Extending our algorithm and analysis to these broader settings and further reducing its communication cost are interesting directions for future work.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: This paper has completed it.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: This paper has completed it.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: This paper has completed it.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: This paper has completed it.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: This paper has completed it.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: This paper has completed it.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: This paper has completed it.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: This paper has completed it.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: This paper has completed it.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: It is not applicable to this paper.

    Guidelines:

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

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: It is not applicable to this paper.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: This paper has completed it.

    Guidelines:

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

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

13. **New assets**

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

    Answer: [N/A]

    Justification: It is not applicable to this paper.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: It is not applicable to this paper.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: It is not applicable to this paper.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: It is not applicable to this paper.

Guidelines:

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