Enhancing Privacy in Decentralized Min-Max Optimization: A Differentially Private Approach

Yueyang Quan
University of North Texas
Denton, USA

Chang Wang*
University of Nevada, Las Vegas
Las Vegas, USA

Shengjie Zhai
University of Nevada, Las Vegas
Las Vegas, USA

Minghong Fang
University of Louisville
Louisville, USA

Zhuqing Liu
University of North Texas
Denton, USA

Abstract

Decentralized min-max optimization allows multi-agent systems to collaboratively solve global min-max optimization problems by facilitating the exchange of model updates among neighboring agents, eliminating the need for a central server. However, sharing model updates in such systems carry a risk of exposing sensitive data to inference attacks, raising significant privacy concerns. To mitigate these privacy risks, differential privacy (DP) has become a widely adopted technique for safeguarding individual data. Despite its advantages, implementing DP in decentralized min-max optimization poses challenges, as the added noise can hinder convergence, particularly in non-convex scenarios with complex agent interactions in min-max optimization problems. In this work, we propose an algorithm called DPMixSGD (Differential Private Minmax Hybrid Stochastic Gradient Descent), a novel privacy-preserving algorithm specifically designed for non-convex decentralized min-max optimization. Our method builds on the state-of-the-art STORM-based algorithm, one of the fastest decentralized min-max solutions. We rigorously prove that the noise added to local gradients does not significantly compromise convergence performance, and we provide theoretical bounds to ensure privacy guarantees. To validate our theoretical findings, we conduct extensive experiments across various tasks and models, demonstrating the effectiveness of our approach.

CCS Concepts

• **Computing methodologies** → **Distributed algorithms**.

Keywords

Min-max optimization, decentralized learning, differential privacy

ACM Reference Format:

Yueyang Quan, Chang Wang, Shengjie Zhai, Minghong Fang, and Zhuqing Liu. 2025. Enhancing Privacy in Decentralized Min-Max Optimization: A Differentially Private Approach. In *Proceedings of the 26th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc '25).* ACM, New York, NY, USA, 30 pages. https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

*Chang Wang conducted this research while he was an intern under the supervision of Zhuqing Liu.

MobiHoc 2025, Houston, USA

ACM ISBN XXX-X-XXXX-XXXX-X
https://doi.org/XX.XXXX/XXXXXXX.XXXXXXX

1 Introduction

Min-max optimization has been widely applied in various machine learning (ML) domains, such as in-context learning [43, 47], generative adversarial networks (GANs) [29, 30, 50], and adversarial reinforcement learning [32, 34, 69]. Traditionally, ML models have been trained using high-performance clusters within large data centers. However, the growing range of ML applications has led to an increasing shift toward deploying models on edge computing networks. This change is driven by the need to process data from geographically distributed sources (e.g., smart devices, vehicles, and sensors) and the high costs or impracticality of transmitting raw training data to centralized cloud servers due to communication bandwidth limitations or privacy concerns [48, 58]. This paradigm is particularly beneficial in scenarios such as multi-agent pretraining and fine-tuning of large language models (LLMs) [11, 22, 23, 33, 53, 60], where collaborative efforts are needed due to the sensitivity of fine-tuning data. It also benefits decentralized min-max optimization applications, such as decentralized AUC maximization [26], multi-agent meta-learning [53, 61], and multi-agent reinforcement learning [54, 83], as the decentralized framework reduces communication overhead and mitigates privacy risks by limiting direct data sharing among agents.

Despite the perceived privacy advantages of decentralized learning, which limit direct data sharing among agents, recent research has shown that it remains vulnerable to privacy breaches due to indirect leakage through model updates or gradient information [31, 41, 62, 78]. An attacker can exploit shared model updates to infer sensitive information from agents, and in some cases, even reconstruct the original training data [24, 35, 57]. These vulnerabilities introduce substantial privacy risks, compromising the anticipated benefits of decentralized systems. To mitigate these challenges, researchers have employed differential privacy (DP) [7, 15, 17, 51], a method that strengthens privacy by adding strategically designed noise to local updates before sharing, thereby offering increased protection against data breaches. While existing works primarily focus on incorporating DP into centralized learning or standard decentralized frameworks such as federated learning, applying DP to decentralized stochastic min-max optimization remains largely unexplored and presents unique challenges:

- Adding noise for differential privacy in decentralized min-max optimization introduces randomness that degrades gradient accuracy, slowing convergence and destabilizing the optimization. In such problems, even slight noise can disrupt the delicate min-max balance and cause oscillations near saddle points. Privacy noise thus poses unique challenges by destabilizing complex saddle-point dynamics, complicating convergence in decentralized, privacy-preserving settings.
- Decentralized min-max optimization faces additional difficulty due to non-IID data across agents, causing local gradients to diverge and hindering consensus on saddle points. Adding privacy noise worsens this by obscuring useful signals. The heterogeneity of non-IID data intensifies coordination challenges while preserving privacy.
- Privacy analysis is especially challenging in decentralized min-max setups because each agent adds noise locally and communicates iteratively, complicating cumulative privacy accounting. Ensuring rigorous differential privacy without harming convergence remains difficult.

In this paper, we bridge a critical gap by introducing DPMixSGD, an innovative and efficient algorithm for differentially private decentralized min-max optimization. Our method leverages the STORM framework [16] to reduce gradient variance, crucial for controlling noise under differential privacy. Its single-loop design eases implementation and privacy analysis. Unlike prior non-private decentralized min-max uses [74], we adapt STORM with privacy-preserving updates and a noise-aware convergence proof. The core mechanism of DPMixSGD involves each agent perturbing its gradients with carefully calibrated noise to ensure differential privacy. These perturbed gradients are then shared with neighboring agents, enabling decentralized collaboration while maintaining privacy. To demonstrate the effectiveness of our DPMixSGD, we conduct a thorough convergence analysis and assess the privacy guarantees of DPMixSGD under practical and reasonable assumptions.

Our key contributions are as follows:

- We introduce DPMixSGD, a new algorithm that guarantees differential privacy in non-convex-strongly-concave decentralized min-max optimization. DPMixSGD is built upon STORM-based algorithms tailored for min-max problems, providing robust privacy protection while maintaining high optimization performance.
- We establish rigorous theoretical convergence guarantee and privacy guarantees for our proposed algorithm, DPMixSGD. Our proof shows that even with Gaussian noise added to the communication process of local gradients, DPMixSGD can still maintain strict convergence. Meanwhile, by strategically designing the noise added to the communication of local gradients in decentralized min-max optimization framework, we achieve DP without significantly degrading the algorithm's performance. This approach effectively balances privacy preservation with strong optimization results.
- We empirically evaluate the DPMixSGD algorithm on logistic regression and AUROC min-max optimization tasks. To assess its performance, we compare DPMixSGD with several state-of-the-art methods, including DM-HSGD [74], SGDA [6], and DP-SGDA [79]. The results show that DPMixSGD performs robustly, naturally preserving privacy while achieving results on par with other methods. Moreover, we conduct a comparative experiment between our algorithm and DM-HSGD to demonstrate that our method significantly improves privacy robustness against Deep Leakage from Gradients (DLG) attacks.

2 Related works

2.1 Decentralized min-max optimization

Numerous methods have been proposed to address min-max optimization problems, including gradient descent techniques [77], momentum-based approaches [3, 37], mirror descent ascent methods [38], and stochastic gradient methods [14, 25, 55]. Building upon these foundational techniques, various algorithms have been developed specifically for decentralized min-max optimization, such as those in [44, 52, 56, 68]. Our approach draws primarily from the decentralized min-max hybrid stochastic gradient descent (DM-HSGD) algorithm [74], which achieves a stochastic first-order oracle (SFO) complexity of $O(\kappa^3\epsilon^{-3})$. Here, $\kappa = \frac{L}{\mu}$ denotes the condition number of the problem, defined as the ratio between the smoothness constant L and the strong convexity constant μ, and ϵ represents the target accuracy level for the optimization error. These developments have significantly expanded the applications of min-max optimization, especially in the context of machine learning.

2.2 Differential privacy (DP)

Differential Privacy (DP) [19] is a rigorous mathematical framework that ensures strong privacy guarantees when analyzing and sharing data. Many algorithms have been designed to provide these guarantees for minimization problems [13, 70, 72, 82], and some have been further adapted to handle min-max problems [39, 80]. Several recent works have explored differential privacy (DP) in the context of variational inequalities and saddle point problems. Boob et al. [8] investigated DP stochastic variational inequalities and saddle point problems, achieving optimal weak gap guarantees; González et al. [28] proposed DP mirror descent methods with nearly dimension-independent utility guarantees for stochastic saddle-point problems. However, their analysis is primarily limited to centralized settings and does not extend to decentralized regime. Bassily et al. [4] refined strong gap analysis using recursive regularization techniques, but their methods require strong convexity and are not directly applicable to min-max formulations in decentralized environments. Zhou et al. [86] addressed worst-group risk minimization through a stability-based lens, but did not consider interactive or game-theoretic settings such as saddle-point optimization. In the nonconvex-strongly-concave case, Zhao et al. [84] introduced a DP temporal difference learning algorithm; nonetheless, their focus lies in reinforcement learning rather than generic decentralized optimization. Our algorithm focuses on addressing the current limitations of these prior works.

Privacy concerns are particularly prominent in distributed systems, such as federated learning [2, 36, 67, 73] and multi-party computation, where nodes often need to exchange sensitive information. The main challenge in these scenarios is to safeguard

individual data privacy while ensuring effective model training and robust performance, especially in the face of potential threats from malicious actors that could compromise the integrity of the learning process. Unlike centralized settings, decentralized optimization lacks a trusted central authority and requires nodes to frequently communicate gradients or model updates over untrusted networks, significantly increasing privacy risks. For instance, adversaries can reconstruct training data from shared gradients, as demonstrated in [87], making privacy preservation even more challenging in decentralized settings. To address this, it is crucial to ensure differential privacy (DP) at each node. While existing work such as DP-SGDA [79] has explored DP in centralized min-max optimization, the decentralized case remains largely unaddressed. In this paper, we propose a novel approach that enhances privacy by injecting noise directly into local gradients at each node in a decentralized setting. While DP-SGDA [79] ensures privacy in centralized min-max optimization through gradient perturbation, it is not directly applicable to decentralized scenarios due to its reliance on centralized data access and coordination. In contrast, our method introduces noise locally in a distributed network, eliminating the need for central coordination and enabling efficient and privacy-preserving updates even in multi-agent systems. This design is naturally compatible with decentralized architectures and leads to improved scalability and communication efficiency. Following the strategy in [18], our algorithm operates by exchanging model variables $\mathbf{x}_t$ and $\mathbf{y}_t$, which depend on gradients. However, instead of perturbing the variables themselves, we perturb the local gradients to preserve privacy while maintaining optimization performance. This design makes our method particularly suitable for privacy-sensitive decentralized applications, such as federated adversarial training [64, 88]. We provide the Table 1 to intuitively illustrate the differences between our algorithm and the baselines we use in our paper. Our work thus takes a critical step toward bridging the gap between differential privacy and decentralized min-max optimization.

Table 1: Comparison of DPMixSGD with baselines.

Method	DP Guarantee	Variance Reduction	Min-Max Setting
DM-HSGD	×	✓(STORM)	✓
SGDA	×	×	✓
DP-SGDA	✓	×	✓
DPMixSGD (ours)	✓	✓(STORM)	✓

3 Problem Formulation and Motivation

3.1 Problem Formulation

Before presenting our problem formulation, we first introduce the mixing matrix $\mathbf{W}$, which represents the averaging weights in the communication network. The matrix $\mathbf{W} = \{\mathbf{w}_{ij}\} \in \mathbb{R}^{m\times m}$ is doubly stochastic and satisfies the following conditions:

$$\mathbf{W}\mathbf{1} = \mathbf{W}^\top \mathbf{1} = \mathbf{1} \tag{1}$$

where $\mathbf{1}$ is an all-ones matrix, and $\mathbf{W}^\top$ is the transpose of $\mathbf{W}$. Note that in this paper $\mathbf{W}$ is required to be symmetric, allowing the communication network to represent undirected graphs.

Decentralized min-max problems are typically formulated in a multi-agent environment, where each agent has access only to its local data and collaborates with other agents via limited communication to optimize a non-convex strongly concave min-max objective function. The mathematical form of a decentralized min-max problem can be expressed as follows:

$$\begin{aligned} \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}) &= \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x},\mathbf{y}), \\ f_i(\mathbf{x},\mathbf{y}) &:= \mathbb{E}_{\mathbf{z}^{(i)}\sim D_i} F_i\left(\mathbf{x},\mathbf{y};\mathbf{z}^{(i)}\right) \end{aligned} \tag{2}$$

where m is the total number of agents; $\mathcal{X} \subseteq \mathbb{R}^{d_1}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_2}$ represent the decision spaces for $\mathbf{x}$ and $\mathbf{y}$, respectively; The local objective function $F_i(\mathbf{x},\mathbf{y};\mathbf{z}^{(i)})$ is L-smooth, non-convex with respect to $\mathbf{x}$, and strongly concave with respect to $\mathbf{y}$; D_i denotes the data distribution on the i-th agent; $\mathbf{z}^{(i)}$ is a random vector sampled from the local dataset $\mathcal{Z}$.

In order to simplify the min-max problem, we often introduce $\Phi(\mathbf{x}) = \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y})$, reducing the problem to one that involves optimizing only over $\mathbf{x}$:

$$\min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}) = \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}). \tag{3}$$

Additionally, we introduce noise to the local gradients. To establish the privacy guarantees of our algorithm, we now present the definition of differential privacy in the context of stochastic decentralized min-max problems.

Definition 1. [Differential Privacy [20]] An algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$ is said to be (θ,γ)-differentially private if, for any adjacent datasets $\mathbf{z}^{(i)} \sim \mathbf{z}^{(i)\prime}$ (on the i-th agent) and for all output events $O \subseteq \text{range}(\mathcal{A})$, the following holds:

$$\mathbb{P}\left[\mathcal{A}\left(\mathbf{z}^{(i)}\right)\in O\right] \le e^{\theta}\mathbb{P}\left[\mathcal{A}\left(\mathbf{z}^{(i)\prime}\right)\in O\right] + \gamma, \tag{4}$$

where $\mathcal{A}\left(\mathbf{z}^{(i)}\right)$ is the output of the decentralized algorithm based on i-th agent datasets $\mathbf{z}^{(i)}$. $\mathbb{P}$ denotes the probability of the algorithm's output in the corresponding event. Note that two datasets are said to be adjacent if they differ in at most one data sample.

Empirical risk plays a crucial role in differential privacy. By optimizing the empirical risk across all agents, the algorithm can effectively train a global model while protecting individual data privacy. Furthermore, empirical risk minimization helps evaluate the impact of privacy-preserving mechanisms on the overall system, ensuring that the algorithm can still converge correctly and produce a meaningful model even after the addition of noise. By combining STORM with gradient tracking, our algorithm effectively mitigates consensus errors and ensures convergence in decentralized settings with non-identical data. This makes it more robust and suitable for complex distributed scenarios. In this paper, we define the average empirical risk as the mean of the local gradients estimators $\bar{\mathbf{g}}_t$ across all agents, expressed as follows:

$$\begin{aligned} \nabla_{\mathbf{x}} f_S(\bar{\mathbf{x}},\bar{\mathbf{y}}) = \bar{\mathbf{g}}_t &= \frac{1}{m}\sum_{i=1}^{m} \nabla_{\mathbf{x}} F_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_t^{(i)}\right) \\ &+ (1-\beta_{\mathbf{x}})\left(\bar{\mathbf{g}}_{t-1} - \frac{1}{m}\sum_{i=1}^{m} \nabla_{\mathbf{x}} F_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)};\mathbf{z}_t^{(i)}\right)\right) \end{aligned} \tag{5}$$

The empirical risk is constructed by $F_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}; \mathbf{z}^{(i)}\right)$, which is the local loss functions across all m agents, where $\mathbf{z}^{(i)}$ denotes the data of the i-th agent.

3.2 Motivating Applications

Our motivation stems from concerns about privacy leaks in the real-world applications of decentralized learning. Here, we present two motivating applications to illustrate the practical relevance of our work:

- **Decentralized Min-Max Learning in Healthcare**: Decentralized learning is widely used in healthcare to enable collaborative model training across institutions without sharing sensitive patient data [41, 46, 65, 66]. Each hospital trains a model locally using its own records and communicates model updates—such as gradients—with its neighbors. However, this process introduces privacy risks, as shared updates may leak confidential information. Beyond privacy, healthcare itself presents intrinsic min-max structures: resource allocation problems—such as distributing ICU beds, vaccines, or staff—often aim to optimize system-wide performance under limited capacity by minimizing the worst-case delay or maximizing the earliest service availability [59]. Motivated by this, we propose a decentralized min-max optimization framework, where each hospital solves a local min-max problem that captures both learning objectives and operational constraints inherent to healthcare. To ensure patient confidentiality during collaboration, we incorporate DP by injecting calibrated noise into local updates. This mechanism prevents sensitive information from being inferred from shared gradients, enabling secure and privacy-preserving model training across institutions.
- **Decentralized Min-Max Learning for Financial Systems**: In financial systems [63, 85], institutions often need to collaboratively train models—for tasks like risk assessment or market forecasting—without sharing sensitive data. Many of these problems naturally follow a min-max structure, as each agent seeks to minimize risk or loss under worst-case scenarios or regulatory constraints. Since financial data is highly sensitive and regulated, differential privacy (DP) is critical to prevent leakage of proprietary or customer information through shared model updates. By introducing noise into local computations, DP enables institutions to collaborate securely without compromising data confidentiality.

4 Solution Approach

In this section, we first outline the necessary preparations for the algorithm and then proceed to present the algorithm along with detailed explanations.

4.1 Preliminaries

Before detailing the proposed algorithms, we define the notations and key concepts used throughout this paper. Let $\mathbf{x}_t^{(i)}$ and $\mathbf{y}_t^{(i)}$ represent the column vector parameters on the i-th agent at t-th iteration. The matrices X_t and Y_t are defined by stacking the vectors $\mathbf{x}_t^{(i)}$ and $\mathbf{y}_t^{(i)}$ across all m agents, i.e., $X_t = \left[\mathbf{x}_t^{(1)}, \mathbf{x}_t^{(2)}, \ldots, \mathbf{x}_t^{(m)}\right]$, $Y_t = \left[\mathbf{y}_t^{(1)}, \mathbf{y}_t^{(2)}, \ldots, \mathbf{y}_t^{(m)}\right]$. The gradient estimators $\mathbf{g}_t^{(i)}$ and $\mathbf{h}_t^{(i)}$ are the local gradient estimators for $\mathbf{x}$ and $\mathbf{y}$ at the i-th agent, respectively, while $\mathbf{v}_t^{(i)}$ and $\mathbf{u}_t^{(i)}$ denote their aggregated counterparts across the network. The matrices G_t, H_t, V_t, and U_t are constructed by stacking the corresponding column vectors $\mathbf{g}_t^{(i)}, \mathbf{h}_t^{(i)}, \mathbf{v}_t^{(i)}$, and $\mathbf{u}_t^{(i)}$ from all agents. Additionally, the matrices G_t^* and H_t^* represent the gradient estimators with noise components included in $\mathbf{g}_t^{(i)}$ and $\mathbf{h}_t^{(i)}$, respectively.

For the mean vectors, we denote the lower-case variable with a bar to represent it, and the upper-case variables with a bar to represent matrices where each column is the corresponding mean vector. Specifically, the mean of $\mathbf{x}_t^{(i)}$ is given by $\bar{\mathbf{x}}_t = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_t^{(i)}$, and the matrix $\bar{X}_t$ is defined as $\bar{X}_t = [\bar{\mathbf{x}}_t, \bar{\mathbf{x}}_t, \ldots, \bar{\mathbf{x}}_t]$. Meanwhile, the added noise terms $n_{\mathbf{x},t}^{(i)} \sim \mathcal{N}(0, \sigma_{\mathbf{x}}^2 I_{d_1})$ and $n_{\mathbf{y},t}^{(i)} \sim \mathcal{N}(0, \sigma_{\mathbf{y}}^2 I_{d_2})$ for $\forall i$ are applied to the respective gradients. We define their mean values as follows:

$$\mathcal{N}_{\mathbf{x},t} = \frac{\sum_{i=1}^m n_{\mathbf{x},t}^{(i)}}{m}, \quad \mathcal{N}_{\mathbf{y},t} = \frac{\sum_{i=1}^m n_{\mathbf{y},t}^{(i)}}{m}. \tag{6}$$

Next, we define the optimal solution for $\mathbf{y}$ as:

$$\mathbf{y}^*(\cdot) = \arg\max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y}), \quad \hat{\mathbf{y}}_t = \arg\max_{\mathbf{y}\in\mathcal{Y}} f\left(\bar{\mathbf{x}}_t, \mathbf{y}\right), \tag{7}$$

where, under the condition that f is strongly concave in $\mathbf{y}$, $\hat{\mathbf{y}}_t$ is unique. We further define the deviation as

$$\delta_t = \|\hat{\mathbf{y}}_t - \bar{\mathbf{y}}_t\|^2. \tag{8}$$

The vectors $\mathbf{0}$ and $\mathbf{1}$ denote $m \times 1$ column vectors of all zeros and ones, respectively. The frobenius norm is denoted by $\|\cdot\|_F$, and the spectral norm by $\|\cdot\|_2$. Partial derivatives with respect to $\mathbf{x}$ and $\mathbf{y}$ are represented by $\nabla_{\mathbf{x}}$ and $\nabla_{\mathbf{y}}$.

4.2 DP in decentralized min-max problem

In this subsection, we will explain our new algorithm step by step. The overall procedure is similar to the STORM-based algorithm, however, it is important to note that we introduce gradient perturbation in the algorithm to ensure privacy protection.

The original values of the parameters at all agents are set to be identical, that is, $\mathbf{x}_0^{(i)} = \mathbf{x}_0$ and $\mathbf{y}_0^{(i)} = \mathbf{y}_0$ for every agent i. The quantities $\mathbf{g}_t^{(i)}$ and $\mathbf{h}_t^{(i)}$ represent the gradient estimators at the i-th agent with respect to $\mathbf{x}$ and $\mathbf{y}$, respectively. These estimators are computed following the STORM [16] method used on DM-HSGD [74]. Specifically, at $t = 0$, a large batch size of b_0 is used to estimate the stochastic gradient (see Initialize in Algorithm 1). For $t > 0$, the gradient estimators can be computed using either a single data point or a mini-batch (refer to lines 2 and 3 in Algorithm 1).

The update rule for the gradient estimator $\mathbf{g}_t^{(i)}$ across all agents can be expressed as follows:

$$\begin{aligned}\bar{\mathbf{g}}_t = &\frac{1}{m}\sum_{i=1}^m \nabla_{\mathbf{x}} F_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}, \mathbf{z}_t^{(i)}\right) \\ &+ (1-\beta_{\mathbf{x}})\left(\bar{\mathbf{g}}_{t-1} - \frac{1}{m}\sum_{i=1}^m \nabla_{\mathbf{x}} F_i\left(\mathbf{x}_{t-1}^{(i)}, \mathbf{y}_{t-1}^{(i)}, \mathbf{z}_t^{(i)}\right)\right)\end{aligned} \tag{9}$$

where the mean gradient $\bar{\mathbf{g}}_t$ is updated by combining the current and previous gradients weighted by $\beta_{\mathbf{x}}$.

Algorithm 1 DPMixSGD on the i-th agent.

Initialize: Mixing matrix $\mathbf{W}$, initial value $\mathbf{x}_0^{(i)} = \mathbf{x}_0, \mathbf{y}_0^{(i)} = \mathbf{y}_0, \mathbf{v}_{-1}^{(i)} = \mathbf{g}_{-1}^{(i)*} = \mathbf{0}, \mathbf{u}_{-1}^{(i)} = \mathbf{h}_{-1}^{(i)*} = \mathbf{0}$, when the algorithm can reach the optimal solution with 0 iteration. We set $\mathbf{g}_0^{(i)} = \nabla_{\mathbf{x}} F_i\left(\mathbf{x}_0^{(i)}, \mathbf{y}_0^{(i)}; \mathbf{z}_0^{(i)}\right)$ and $\mathbf{h}_0^{(i)} = \nabla_{\mathbf{y}} F_i\left(\mathbf{x}_0^{(i)}, \mathbf{y}_0^{(i)}; \mathbf{z}_0^{(i)}\right), \quad \left|\mathbf{z}_0^{(i)}\right| = b_0$.

Parameter: Privacy budgets θ, γ, learning rate $\eta_{\mathbf{x}}, \eta_{\mathbf{y}}$, weight $\beta_{\mathbf{x}}, \beta_{\mathbf{y}}$, batch size b_0, epoch T.

Output: $\bar{x}_\zeta$, where ζ is chosen randomly from $\{1, 2, \cdots, T\}$

1: **for** each $t = 1, \ldots, T-1$ **do**

2: $\mathbf{g}_t^{(i)} = (1-\beta_{\mathbf{x}})\left(\mathbf{g}_{t-1}^{(i)} - \nabla_{\mathbf{x}} F_i\left(\mathbf{x}_{t-1}^{(i)}, \mathbf{y}_{t-1}^{(i)}; \mathbf{z}_t^{(i)}\right)\right) + \nabla_{\mathbf{x}} F_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}; \mathbf{z}_t^{(i)}\right)$

3: $\mathbf{h}_t^{(i)} = (1-\beta_{\mathbf{y}})\left(\mathbf{h}_{t-1}^{(i)} - \nabla_{\mathbf{y}} F_i\left(\mathbf{x}_{t-1}^{(i)}, \mathbf{y}_{t-1}^{(i)}; \mathbf{z}_t^{(i)}\right)\right) + \nabla_{\mathbf{y}} F_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}; \mathbf{z}_t^{(i)}\right)$

4: //Encrypt gradients when communicating with other agents.

5: Sample noise $n_{\mathbf{x},t}^{(i)} \sim \mathcal{N}(0, \sigma_{\mathbf{x}}^2 I_{d_1})$ and $n_{\mathbf{y},t}^{(i)} \sim \mathcal{N}(0, \sigma_{\mathbf{y}}^2 I_{d_2})$.

6: $\mathbf{g}_t^{(i)*} = \mathbf{g}_t^{(i)} + n_{\mathbf{x},t}^{(i)}$

7: $\mathbf{h}_t^{(i)*} = \mathbf{h}_t^{(i)} + n_{\mathbf{y},t}^{(i)}$

8: //At the i-th agent, the encrypted gradient is received and calculated.

9: $\mathbf{v}_t^{(i)} = \sum_{j=1}^m w_{ij}\left(\mathbf{v}_{t-1}^{(j)} + \mathbf{g}_t^{(j)*} - \mathbf{g}_{t-1}^{(j)*}\right)$

10: $\mathbf{u}_t^{(i)} = \sum_{j=1}^m w_{ij}\left(\mathbf{u}_{t-1}^{(j)} + \mathbf{h}_t^{(j)*} - \mathbf{h}_{t-1}^{(j)*}\right)$

11: //Send the computation result to the respective agent and perform the mixed information calculation at that agent.

12: $\mathbf{x}_{t+1}^{(i)} = \sum_{j=1}^m w_{ij}\left(\mathbf{x}_t^{(j)} - \eta_{\mathbf{x}} \mathbf{v}_t^{(j)}\right)$

13: $\mathbf{y}_{t+1}^{(i)} = \sum_{j=1}^m w_{ij}\left(\mathbf{y}_t^{(j)} + \eta_{\mathbf{y}} \mathbf{u}_t^{(j)}\right)$

14: **end for**

Similarly, the computation for $\mathbf{h}_t^{(i)}$ follows the same procedure as for $\mathbf{g}_t^{(i)}$. Once the local gradient estimators $\mathbf{g}_t^{(i)}$ and $\mathbf{h}_t^{(i)}$ are computed, each agent communicates with its neighboring agents to aggregate the estimates and compute the new gradient estimators $\mathbf{u}_t^{(i)}$ and $\mathbf{v}_t^{(i)}$. To ensure differential privacy, noise is added to the local gradients during communication with neighboring agents (see lines 5 to 7 in Algorithm 1). This guarantees that our algorithm meets privacy requirements. To mitigate the consensus error, gradient tracking is employed (see lines 9 and 10 in Algorithm 1). After obtaining the updated gradient estimators $\mathbf{u}_t^{(i)}$ and $\mathbf{v}_t^{(i)}$, each agent communicates with its neighbors again to update the model parameters $\mathbf{x}$ and $\mathbf{y}$.

5 Theoretical Analysis

In this section, we present the convergence analysis and discuss the privacy guarantees of our DPMixSGD algorithm under certain mild assumptions. All relevant proofs are provided in Appendix. We begin by reviewing some essential assumptions and definitions.

5.1 Convergence analysis

In our proposed DPMixSGD, each agent introduces noise to the local gradients to ensure privacy during agent communication. Therefore, before presenting the privacy guarantees, we first provide a rigorous proof of convergence to demonstrate that the added noise does not affect the original algorithm's convergence. To support this proof, we introduce several mild assumptions.

Assumption 1. *(Lipschitz continuity of the gradient) Each local function $F_i\left(\mathbf{x}, \mathbf{y}; \mathbf{z}^{(i)}\right)$ is lipschitz smooth, meaning there exists a constant L such that for any two pairs $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{x}', \mathbf{y}')$, we have:*

$$\begin{aligned} &\left\|\nabla F_i(\mathbf{x}, \mathbf{y}; \mathbf{z}^{(i)}) - \nabla F_i\left(\mathbf{x}', \mathbf{y}'; \mathbf{z}^{(i)}\right)\right\|^2 \\ &\leq L^2\left(\|\mathbf{x} - \mathbf{x}'\|^2 + \|\mathbf{y} - \mathbf{y}'\|^2\right). \end{aligned} \tag{10}$$

Assumption 2. *(Bounded gradient variance) The gradient of each local function $F_i\left(\mathbf{x}, \mathbf{y}; \mathbf{z}^{(i)}\right)$ is an unbiased estimate of $\nabla f_i(\mathbf{x}, \mathbf{y})$ and has bounded variance, i.e.,*

$$\mathbb{E}\left\|\nabla F_i(\mathbf{x}, \mathbf{y}; \mathbf{z}^{(i)}) - \nabla f_i(\mathbf{x}, \mathbf{y})\right\|^2 \leq \sigma < +\infty. \tag{11}$$

Assumption 3. *(Lower bound of the objective) The global objective function $\Phi(\cdot)$ is lower bounded, i.e., $\inf_{\mathbf{x}} \Phi(\mathbf{x}) = \Phi^* > -\infty$.*

Remark. *All the aforementioned assumptions are standard assumptions in optimization analysis [16, 21, 40, 42, 81].*

Assumption 4. *(Spectral gap of the mixing matrix) The doubly stochastic mixing matrix $\mathbf{W}$ satisfies the following spectral gap condition: $\left\|\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_2 = \lambda \in [0, 1)$.*

Remark. *The spectral gap assumption plays a crucial role in ensuring effective information transfer across the network, allowing each agent to achieve global convergence by communicating with its neighboring agents, as highlighted in prior works [72, 76]. A typical spectral gap assumption requires the mixing matrix $\mathbf{W}$ to be symmetric and doubly stochastic, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ such that $|\lambda_2| < 1$ and $|\lambda_n| < 1$. This condition guarantees that the communication graph remains connected, preventing both excessive diffusion and slow propagation of information within the system.*

We adopt this symmetric and doubly stochastic matrix setting because undirected graphs are standard and widely adopted in decentralized learning. In many practical scenarios, such as decentralized federated learning over peer-to-peer networks [5], sensor networks [27], or cooperative robotics systems [9], communication between agents is naturally bidirectional—each agent can both send and receive information from its neighbors. This symmetric communication structure simplifies the design and analysis of algorithms and has been shown to yield stable and efficient convergence in numerous studies. Furthermore, undirected graphs with symmetric weight matrices allow for well-established consensus-based protocols and spectral methods to be employed, making them a natural choice for studying theoretical properties such as convergence and privacy guarantees in decentralized settings.

Assumption 5. *(Strong concavity) The function $f_i(\mathbf{x}, \mathbf{y})$ is μ-strongly concave in $\mathbf{y}$. That is, there exists a constant $\mu > 0$ such that for any*

$\mathbf{x}, \mathbf{y}$ *and* $\mathbf{y}'$*, we have:*

$$f_i(\mathbf{x}, \mathbf{y}) \leq f_i\left(\mathbf{x}, \mathbf{y}'\right) + \left\langle \nabla_{\mathbf{y}} f_i\left(\mathbf{x}, \mathbf{y}'\right), \mathbf{y} - \mathbf{y}' \right\rangle - \frac{\mu}{2}\left\|\mathbf{y} - \mathbf{y}'\right\|^2. \quad (12)$$

Remark. *The assumption of strong concavity in* $\mathbf{y}$ *is crucial for ensuring the well-posedness of the min-max problem. Specifically,* μ*-strong concavity ensures the uniqueness of the solution during the* $\mathbf{y}$*-update step, thereby preventing ambiguity in the optimization process. This assumption is standard in the analysis of min-max optimization problems and is essential for deriving theoretical guarantees related to convergence rates and stability. Many previous works in the field of decentralized min-max optimization [14] adopt this assumption to enhance their algorithms' convergence, stability, and efficiency.*

Similar to standard nonconvex-strongly-concave problems, we continue to use the ϵ-stationary point as the convergence criterion, i.e., $\|\nabla\Phi(\mathbf{x})\| \leq \epsilon$. From the Lemma 4.3 in [49], it is established that the function $\Phi(\mathbf{x})$ is differentiable and satisfies the $(L + \kappa L)$-smoothness condition. And it also mention that $\mathbf{y}^*(\cdot)$ is κ-Lipschitz continuous, meaning for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^{d_1}$, the inequality $\|\mathbf{y}^*(\mathbf{x}_1) - \mathbf{y}^*(\mathbf{x}_2)\| \leq \kappa \|\mathbf{x}_1 - \mathbf{x}_2\|$ holds. This indicates that the variation of $\mathbf{y}(\cdot)$ is bounded by κ. Consequently, we have:

$$\nabla\Phi\left(\bar{\mathbf{x}}_t\right) = \nabla_{\mathbf{x}} f\left(\bar{\mathbf{x}}_t, \hat{\mathbf{y}}_t\right) + \nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}_t, \hat{\mathbf{y}}_t\right) \cdot \partial \mathbf{y}^*\left(\bar{\mathbf{x}}_t\right), \quad (13)$$

where we use $\nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}_t, \hat{\mathbf{y}}_t\right) = 0$ as defined earlier. Thus we have $\nabla\Phi\left(\bar{\mathbf{x}}_t\right) = \nabla_{\mathbf{x}} f\left(\bar{\mathbf{x}}_t, \hat{\mathbf{y}}_t\right)$. With this, we now present the main theorem that show our algorithm maintains convergence despite the added noise for privacy preservation.

Theorem 1. *Let Assumptions 1 to 5 hold, our Algorithm 1 satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_t\right)\right\|^2 = O\left(\epsilon^2\right) + O\left(m\epsilon^2\right) + O\left(\sigma_{\mathbf{x}}^2 d_1 + \sigma_{\mathbf{y}}^2 d_2\right),$$

when we set $T = \frac{1500\kappa^3}{(1-\lambda)^2\epsilon\beta_{\mathbf{x}}}$ *and the other parameters satisfy* $\beta_{\mathbf{y}} = \frac{\beta_{\mathbf{x}}}{25\kappa^2}$, $\eta_{\mathbf{x}} = \frac{(1-\lambda)^2\beta_{\mathbf{x}}}{750\kappa^3 L\epsilon}$, $\eta_{\mathbf{y}} = \frac{(1-\lambda)^2\beta_{\mathbf{x}}}{75\kappa L\epsilon}$, $b_0 = \frac{20\kappa\epsilon}{\beta_{\mathbf{x}}}$, $\beta_x = \frac{\epsilon\min\{1, m\epsilon\}}{20}$. *And we have,*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_t\right)\right\|^2 &= O\left(\frac{1}{(mT_0)^{2/3}}\right) + O\left(\frac{1}{T_0}\right) \\ &+ O\left(\frac{m^{1/3}}{T_0^{2/3}}\right) + O\left(\sigma_{\mathbf{x}}^2 d_1 + \sigma_{\mathbf{y}}^2 d_2\right),\end{aligned} \quad (14)$$

when we set $T = \frac{30000\kappa^3 T_0}{(1-\lambda)^2}$*, and the other parameters satisfy* $\beta_{\mathbf{y}} = \frac{\beta_{\mathbf{x}}}{25\kappa^2}$, $\eta_{\mathbf{x}} = \frac{(1-\lambda)^2\beta_{\mathbf{x}}}{750\kappa^3 L\epsilon}$, $\eta_{\mathbf{y}} = \frac{(1-\lambda)^2\beta_{\mathbf{x}}}{75\kappa L\epsilon}$, $b_0 = \frac{20\kappa\epsilon}{\beta_{\mathbf{x}}}$, $T_0 \geq 10m^2$, *and* $\beta_x = \frac{m^{1/3}}{20T_0^{2/3}}$.

Remark. *We build upon the convergence analysis from Xian's work [74], however, by introducing noise into the local gradients during communication with neighboring agents, additional terms, namely* $n_{\mathbf{x},t}^{(i)}$ *and* $n_{\mathbf{y},t}^{(i)}$*, are introduced into* $\mathbf{g}_t$ *in our analysis. These additional terms are amplified during the proof process and require adjustments to the entire proof. As a result, we recalculated the bounds for all theorems and lemmas involved. To simplify our results, we applied novel scaling and bounding techniques. (e.g., for* $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_t\right)\right\|^2$*, we adopt a similar approach and found that our results, including the additional terms, are no greater than twice the original results in Xian's work. Since this already provides a tight bound, we opt to double some constants in the related terms for simplicity.) From our results, we conclude that the added noise does not impact the SFO complexity of the DM-HSGD algorithm. Specifically, when* T *is determined by* ϵ *which is shown in Eq. (14), if* $m \leq O\left(\epsilon^{-1}\right)$*, the SFO complexity of Algorithm 1 is* $O\left(\kappa^3\epsilon^{-3}\right)$*. For* $m > O\left(\epsilon^{-1}\right)$*, the SFO complexity is* $O\left(\kappa^3 m\epsilon^{-2}\right)$*. When* T *is independent of* ϵ *as we show in Eq. (14), the leading term in the convergence rate remains* $O\left(\frac{1}{(mT_0)^{2/3}}\right)$*, thus preserving the linear speedup characteristic of the original algorithm. Detailed proof of Theorem 1 is provided in Appendix A.*

5.2 Privacy analysis

Assumption 6. *[39] For the min-max problem, we say* f *is* ρ*-strongly-convex-strongly-concave (* $\rho - SC - SC$ *) if for each fixed* $\mathbf{y} \in \mathcal{Y}$*, the function* $f_i(\mathbf{x}, \mathbf{y}; \cdot)$ *is* ρ*-strongly-convex in* $\mathbf{x}$ *for all* i*. And for each fixed* $\mathbf{x} \in \mathcal{X}$*, the function* $f_i(\mathbf{x}, \mathbf{y}; \cdot)$ *is* ρ*-strongly-concave in* $\mathbf{y}$ *for all* i*. In this paper, we focus on the* $\rho - SC - SC$ *problem.*

Remark. *While this assumption may appear restrictive, it captures a number of important practical scenarios where strong convexity can be induced through regularization. Examples include robust federated learning, adversarial training, and resource allocation. We will clarify this in the revised version and explicitly mention it as a key direction for future work, to either relax this assumption or extend our analysis to broader settings.*

Assumption 7 (Bounded Gradient). *There exists a constant* $L_g > 0$ *such that, for any* $\mathbf{x}, \mathbf{y}$ *and* $\mathbf{z}$*,*

$$\left\|\nabla_{\mathbf{x}} F_i(\mathbf{x}, \mathbf{y}; \mathbf{z})\right\|_2 \leq L_g, \quad \left\|\nabla_{\mathbf{y}} F_i(\mathbf{x}, \mathbf{y}; \mathbf{z})\right\|_2 \leq L_g. \quad (15)$$

Now let's review the proof for convergence part (Appendix A), we have already known that $\bar{\mathbf{v}}_t = \bar{\mathbf{g}}_t = (1-\beta_{\mathbf{x}})\left(\bar{\mathbf{g}}_{t-1} - \frac{1}{m}\sum_{i=1}^m \nabla_{\mathbf{x}} F_i\left(\mathbf{x}_{t-1}^{(i)}, \mathbf{y}_{t-1}^{(i)}, \mathbf{z}_t^{(i)}\right)\right) + \frac{1}{m}\sum_{i=1}^m \nabla_{\mathbf{x}} F_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}, \mathbf{z}_t^{(i)}\right)$. And by the definition of $\mathbf{g}_t^{(i)}$, we can obtain this recursively:

$$\begin{aligned}\bar{\mathbf{g}}_t = \frac{1}{m}\sum_{j=1}^m \Bigg(\sum_{k=0}^t (1-\beta_{\mathbf{x}})^{t-k}\Big[\nabla_{\mathbf{x}} F_j\left(\mathbf{x}_k^{(i)}, \mathbf{y}_k^{(i)}; \mathbf{z}_k^{(i)}\right) \\ -\nabla_{\mathbf{x}} F_j\left(\mathbf{x}_k^{(i)}, \mathbf{y}_k^{(i)}; \mathbf{z}_{k+1}^{(i)}\right)\Big] + \nabla_{\mathbf{x}} F_j\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}; \mathbf{z}_{t+1}^{(i)}\right)\Bigg).\end{aligned} \quad (16)$$

From the definition of $\bar{\mathbf{x}}_t$, we know $\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{x}}_t - \eta_{\mathbf{x}}\bar{\mathbf{v}}_t$, so we have:

$$\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{x}}_t - \eta_{\mathbf{x}}\left(\bar{\mathbf{g}}_t + \mathcal{N}_{\mathbf{x},t}\right), \quad (17)$$

where $\mathcal{N}_{\mathbf{x},t} \sim \mathcal{N}\left(0, \frac{\sigma_{\mathbf{x}}^2}{m} I_{d_1}\right)$.

Lemma 1. *[71] In single parameter DP-GD paradigm whose model updates as* $\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{x}}_t - \eta_{\mathbf{x}}\left(\bar{\mathbf{g}}_t + \mathcal{N}_{\mathbf{x},t}\right)$*, meanwhile the loss function is* G*-lipschitz, for* $\theta, \gamma > 0$*, for some constant* c*, it is* (θ, γ)*-DP if the random noise is zero mean gaussian noise, i.e.,* $\mathcal{N}_{\mathbf{x},t} \sim \mathcal{N}\left(0, \frac{\sigma_{\mathbf{x}}^2}{m} I_{d_1}\right)$*, and* $\sigma_{\mathbf{x}}^2 = c\frac{G^2 T\log(1/\gamma)}{m\theta^2}$.

However, we use a momentum gradient descent method, so the parameter updates involve $\bar{\mathbf{g}}_t$ instead of just $\nabla_{\mathbf{x}} F_i(\cdot, \cdot; \cdot)$, according to Assumption 6, since each gradient term $\nabla_{\mathbf{x}} F_i(\cdot, \cdot; \cdot)$ is L-lipschitz, the weighted sum operation does not change the lipschitz constant. Therefore, $\mathbf{g}_t$ is G-lipschitz where G is derived from L, we

Table 2: AUROC score of each algorithm over epochs during the robust logistic regression experiments on 'a8a', 'a9a' and CIFAR-10 datasets.

(a) Impact of total number of agents m.

m	$m=5$			$m=10$			$m=15$			$m=20$		
Method	a8a	a9a	CIFAR-10	a8a	a9a	CIFAR-10	a8a	a9a	CIFAR-10	a8a	a9a	CIFAR-10
SGDA	0.7590	0.7164	0.6648	0.7801	0.7029	0.6705	0.7626	0.6968	0.6762	0.7625	0.6887	0.6778
DP-SGDA	0.7383	0.7037	0.5910	0.7417	0.7047	0.6041	0.7453	0.7453	0.6161	0.7302	0.6945	0.6332
DM-HSGD	0.7519	0.6708	0.6644	0.7420	0.7169	0.6695	0.7853	0.7053	0.6767	0.7702	0.6977	0.6772
DPMixSGD	0.8094	0.7003	0.5927	0.7392	0.6692	0.6099	0.7751	0.6970	0.6259	0.7457	0.6926	0.6367

(b) Impact of sparsity level p.

p	$t=0.2$			$p=0.5$			$p=0.8$			$p=1$		
Method	a8a	a9a	CIFAR-10	a8a	a9a	CIFAR-10	a8a	a9a	CIFAR-10	a8a	a9a	CIFAR-10
SGDA	0.7388	0.6778	0.6624	0.7270	0.6357	0.6648	0.7373	0.6965	0.6612	0.7276	0.6971	0.6601
DP-SGDA	0.7500	0.6582	0.5914	0.7374	0.6591	0.5910	0.7181	0.7096	0.5937	0.7380	0.7178	0.5969
DM-HSGD	0.7674	0.6588	0.6632	0.7018	0.7059	0.6644	0.7247	0.6753	0.6622	0.6888	0.6716	0.6619
DPMixSGD	0.7272	0.5971	0.5987	0.7825	0.7039	0.5927	0.7696	0.6504	0.5910	0.7666	0.6758	0.5906

(c) Impact of θ.

θ	$\theta=0.005$			$\theta=0.01$			$\theta=0.05$			$\theta=0.1$		
Method	a8a	a9a	CIFAR-10	a8a	a9a	CIFAR-10	a8a	a9a	CIFAR-10	a8a	a9a	CIFAR-10
SGDA	0.7719	0.6957	0.6648	0.7719	0.6957	0.6648	0.7719	0.6957	0.6648	0.7719	0.6957	0.6648
DP-SGDA	0.7595	0.6691	0.5918	0.7555	0.6673	0.5910	0.7257	0.6778	0.5965	0.7321	0.6750	0.6127
DM-HSGD	0.7941	0.7142	0.6644	0.7941	0.7142	0.6644	0.7941	0.7142	0.6644	0.7941	0.7142	0.6644
DPMixSGD	0.6653	0.5644	0.5932	0.6991	0.6026	0.5927	0.7651	0.7011	0.6000	0.7658	0.6170	0.5978

(d) Impact of γ.

γ	$\gamma=1/60000$			$\gamma=1/30000$			$\gamma=1/5000$			$\gamma=1/1000$		
Method	a8a	a9a	CIFAR-10	a8a	a9a	CIFAR-10	a8a	a9a	CIFAR-10	a8a	a9a	CIFAR-10
SGDA	0.7719	0.6957	0.6644	0.7719	0.6957	0.6644	0.7719	0.6957	0.6644	0.7719	0.6957	0.6644
DP-SGDA	0.7325	0.6507	0.5922	0.7564	0.7112	0.5910	0.7383	0.6990	0.6168	0.7757	0.7102	0.6007
DM-HSGD	0.7941	0.7142	0.6644	0.7941	0.7142	0.6644	0.7941	0.7142	0.6644	0.7941	0.7142	0.6644
DPMixSGD	0.7741	0.6927	0.5962	0.7979	0.6692	0.5927	0.7444	0.7023	0.5948	0.7719	0.6859	0.5927

clarify that the Lipschitz constant G follows from the L-Lipschitz continuity of ∇F and the structure of Eq. (9), where g_t is a linear combination of Lipschitz-smooth gradients. Using the recursion $G \le L + (1-\beta_{\mathbf{x}})G$, we get $G \le \frac{L}{\beta_{\mathbf{x}}}$.

The authors in [71] provide a tight noise bound for differentially private gradient descent under a single-parameter condition. However, in the min-max paradigm, privacy leakage also arises from the gradient information, regardless of whether it is used for minimization or maximization. Since the updates for $\mathbf{y}$ share the same structure as those for $\mathbf{x}$, the noise variance derived in [71] can be symmetrically applied to $\mathbf{y}$. Notably, the privacy cost is independent of whether the process involves minimization or maximization. Therefore, by injecting the noise proposed in [71] into both $\mathbf{x}$ and $\mathbf{y}$, the DP guarantee can still be ensured. Since the proof process is nearly identical (with the only difference being its application to $\mathbf{y}$ as well), we directly adopt the result in our theorem. Therefore, by Lemma 1 we give the privacy guarantees of DPMixSGD.

Theorem 2. *If $F_i(\cdot,\cdot;\cdot)$ satisfies Assumption 6 then for some privacy budget $\theta = \Omega\left(\frac{L_g d^{1/2}\log(1/\gamma)^{1/2}}{m^{1/2}\epsilon^4}\right), \gamma > 0$, we get a utility for DPMixSGD to be (θ,γ)-DP if*

$$\sigma_x, \sigma_y = O\left(\frac{L_g\sqrt{\left(\frac{8T(T+1)(2T+1)}{3}+4T\right)\log(1/\gamma)}}{2\theta\sqrt{m}}\right). \tag{18}$$

Remark. *While σ_x and σ_y scale with T as given in Theorem 2, the additional noise-induced error remains controlled under an appropriate choice of θ. Specifically, by ensuring*

$$\theta = \Omega\left(\frac{L_g d^{1/2}\log(1/\gamma)^{1/2}}{m^{1/2}\epsilon^4}\right) \tag{19}$$

we obtain:

$$\frac{L_g^2 d\log(1/\gamma)}{\theta^2 m\epsilon^6} = O(\epsilon^2) \tag{20}$$

Then we can get an optimization error bound of

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\Phi(\bar{\mathbf{x}}_t)\|^2 = O(\epsilon^2) \tag{21}$$

which ensures that our algorithm maintains the desired convergence rate without being dominated by noise.

6 Experiments

6.1 Robust logistic regression in decentralized min-max problem

In this section, we conduct the experiment of decentralized robust logistic regression based on "a8a" [12] ,"a9a" [12], and CIFAR-10 [45] datasets. In these experiments, we compare the DPMixSGD, DM-HSGD [74], SGDA [6], and DP-SGDA [79] algorithms. We partition the given dataset as $\{(a_i, b_i)\}_{i=1}^m$, where each feature vector $a_i \in \mathbb{R}^d$ and each label $b_i \in \{-1, 1\}$. Robust logistic regression is formulated as the following min-max problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \max_{\mathbf{y}\in\Delta_m} f(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{m} \mathbf{y}_i l_i(\mathbf{x}) - V(\mathbf{y}) + g(\mathbf{x}), \tag{22}$$

where m is the total number of agents, $\mathbf{y}_i$ represents the i-th component of the variable $\mathbf{y}$. The logistic loss function is defined by $l_i(\mathbf{x}) = \log\left(1 + \exp\left(-b_i a_i^\top \mathbf{x}\right)\right)$. The divergence measure $V(\mathbf{y})$ is given by $V(\mathbf{y}) = \frac{1}{2}\lambda_1 \|m\mathbf{y} - \mathbf{1}\|^2$. The simplex Δ_m in $\mathbb{R}^m$ is defined as $\Delta_m = \left\{\mathbf{y} \in \mathbb{R}^m \,\middle|\, 0 \le \mathbf{y}_i \le 1 \text{ for all } i,\ \sum_{i=1}^m \mathbf{y}_i = 1\right\}$. Additionally, the nonconvex regularization term $g(\mathbf{x})$ is formulated as $g(\mathbf{x}) = \lambda_2 \sum_{i=1}^d \frac{\alpha \mathbf{x}_i^2}{1+\alpha \mathbf{x}_i^2}$. Following the experimental configurations outlined here, we set the parameters to $\lambda_1 = \frac{1}{m^2}$, $\lambda_2 = 0.001$, and $\alpha = 10$ in our experiments.

For the evaluation of the DPMixSGD, DM-HSGD, SGDA, and DP-SGDA algorithms, we show the results of our experiment in Table 2. Regarding the optimization parameters within the neural network, the learning rates for the model parameters $\mathbf{x}$ and their dual variables $\mathbf{y}$ are selected from the set $\{1.0, 0.1, 0.01, 0.001\}$. The mini-batch size is fixed at 20. Specifically for the DPMixSGD and DM-HSGD algorithms, the batch size for the initial iteration is set to $b_0 = 10{,}000$. Additionally, the gradient weight adjustment parameters β_x and β_y are chosen from the set $\{0.5, 0.1, 0.01\}$.

Remark. *It is worth noting that our experiments did not apply gradient clipping, although gradient clipping is common in differential privacy (DP) training but not strictly required. In our case, the DP noise level is moderate, and a well-tuned learning rate ensures stable convergence without clipping. Experimental results show no signs of instability. Additional experiment on gradient clipping is provided in Appendix C, showing similar trends.*

In the experiment, the communication topology among agents is modeled using an Erdős–Rényi random graph $\mathcal{G}(m, p)$, where m is the number of agents and $p \in [0, 1]$ denotes the sparsity level, i.e., the probability that an edge exists between any two agents. A higher p implies a denser communication network. Formally, each edge is included in the graph independently with probability:

$$\mathbb{P}[(i, j) \in \mathcal{E}] = p, \quad \forall i \neq j, \tag{23}$$

where $\mathcal{E}$ denotes the edge set of the communication graph. The expected degree of each node is $(m-1)p$, and the total expected number of edges is $\frac{pm(m-1)}{2}$. Therefore, we have the definition of sparsity level p:

$$p = \frac{2|\mathcal{E}|}{m(m-1)}, \tag{24}$$

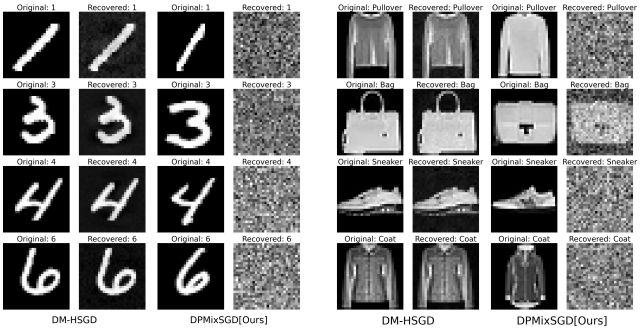
where $|\mathcal{E}|$ is the total number of edges or links in the system.

We conduct control group experiments on robust logistic regression, examined the impact of several factors. These include the number of agents in the network, the sparsity level p of the connectivity matrix, and the adding noise is affected by θ and γ. Table 2 illustrates the AUROC score of each algorithm over epochs during the robust logistic regression experiments.

From the Table 2, we observe that our algorithm performs better with a small number of agents. This can be attributed to the injection of noise into the local gradients before the model updates. However, as the communication frequency increases, the performance inevitably declines. Furthermore, in the experiments regarding sparsity levels, our algorithm exhibits superior performance compared to other algorithms under high sparsity conditions. The experimental data on the two parameters affecting noise show that our algorithm demonstrates greater stability when adding noise at different levels. Therefore, this indicates that its theoretical design is effective in practical applications, outperforming existing methods such as DP-SGDA, and in certain cases, approaching or even surpassing non-private mechanisms like SGDA and DM-HSGD. To further assess the applicability of our algorithm in non-convex deep learning settings, we also conduct experiments on image classification tasks using a multilayer perceptron (MLP) with the Fashion-MNIST dataset. These results, which demonstrate consistent advantages of our method over baselines, are provided in Appendix C.

6.2 Robustness to DLG Attacks

To further evaluate the privacy protection capabilities of our proposed DPMixSGD algorithm, we conduct additional experiments on the MNIST and Fashion-MNIST datasets using a multi-layer perceptron (MLP) model. These experiments focus on the algorithm's robustness against Deep Leakage from Gradients (DLG) attacks [87], a gradient inversion method that can potentially recover private training data from shared gradients. We assess the reconstruction quality of the DLG attack under a given noise level $\sigma = 1$. The results shown in Figure 1 demonstrate that our differentially private algorithm effectively mitigates the risk of visual identity recovery, significantly enhancing privacy robustness. For transparency and reproducibility, the experimental settings are specified as follows: the learning rate is set to 0.01 for the primal variable $\mathbf{x}$ and 0.001 for the dual variable $\mathbf{y}$, and the mini-batch size is fixed at 128. The evaluation confirms that DPMixSGD not only maintains strong privacy guarantees but also resists gradient leakage attacks in practical decentralized learning scenarios.



(a) MNIST dataset **(b) Fashion-MNIST dataset**

Figure 1: DLG Attack Reconstruction Results.

7 Conclusion

In this paper, we addressed the challenges of privacy protection in decentralized min-max learning problem by proposing a novel DPMixSGD algorithm. Our theoretical analysis proves that DPMixSGD ensures rigorous privacy guarantees while maintaining provable convergence.Empirical results demonstrate that out proposed method DPMixSGD not only upholds strong privacy guarantees but also effectively resists gradient leakage in practical decentralized learning scenarios. This work contributes to advancing decentralized learning by effectively balancing the need for privacy and efficient communication in distributed systems, providing a robust framework for future applications in privacy-sensitive domains.

Acknowledgments

We thank the anonymous reviewers for their helpful comments.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.

[2] Mohammed Adnan, Shivam Kalra, Jesse C Cresswell, Graham W Taylor, and Hamid R Tizhoosh. 2022. Federated learning and differential privacy for medical image analysis. *Scientific reports* 12, 1 (2022), 1953.

[3] Babak Barazandeh, Tianjian Huang, and George Michailidis. 2021. A decentralized adaptive momentum method for solving a class of min-max optimization problems. *Signal Processing* 189 (2021), 108245.

[4] Raef Bassily, Cristóbal Guzmán, and Michael Menart. 2023. Differentially private algorithms for the stochastic saddle point problem with optimal rates for the strong gap. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, 2482–2508.

[5] Monik Raj Behera, Suresh Shetty, Robert Otter, et al. 2021. Federated learning using peer-to-peer network for decentralized orchestration of model weights. *Authorea Preprints* (2021).

[6] Aleksandr Beznosikov, Eduard Gorbunov, Hugo Berard, and Nicolas Loizou. 2023. Stochastic gradient descent-ascent: Unified theory and new efficient methods. In *International conference on artificial intelligence and statistics*. PMLR, 172–235.

[7] Sayan Biswas, Mathieu Even, Laurent Massoulié, Anne-Marie Kermarrec, Rafael Pereira Pires, Rishi Sharma, and Martijn de Vos. 2024. Noiseless privacy-preserving decentralized learning. In *The 25th Privacy Enhancing Technologies Symposium*, Vol. 2025. Privacy Enhancing Technologies Symposium Advisory Board, 824–844.

[8] Digvijay Boob and Cristóbal Guzmán. 2024. Optimal algorithms for differentially private stochastic monotone variational inequalities and saddle-point problems. *Mathematical Programming* 204, 1 (2024), 255–297.

[9] Francesco Bullo, Jorge Cortés, and Sonia Martinez. 2009. *Distributed control of robotic networks: a mathematical approach to motion coordination algorithms*. Princeton University Press.

[10] Mark Bun and Thomas Steinke. 2016. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of cryptography conference*. Springer, 635–658.

[11] Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. 2021. Fltrust: Byzantine-robust federated learning via trust bootstrapping. In *NDSS*.

[12] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. In *ACM transactions on intelligent systems and technology (TIST)*.

[13] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. 2011. Differentially private empirical risk minimization. *Journal of Machine Learning Research* 12, 3 (2011).

[14] Lesi Chen, Haishan Ye, and Luo Luo. 2024. An Efficient Stochastic Algorithm for Decentralized Nonconvex-Strongly-Concave Minimax Optimization. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1990–1998.

[15] Hsin-Pai Cheng, Patrick Yu, Haojing Hu, Syed Zawad, Feng Yan, Shiyu Li, Hai Li, and Yiran Chen. 2019. Towards decentralized deep learning with differential privacy. In *International Conference on Cloud Computing*. Springer, 130–145.

[16] Ashok Cutkosky and Francesco Orabona. 2019. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems* 32 (2019).

[17] Edwige Cyffers and Aurélien Bellet. 2022. Privacy amplification by decentralization. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 5334–5353.

[18] Edwige Cyffers, Mathieu Even, Aurélien Bellet, and Laurent Massoulié. 2022. Muffliato: Peer-to-peer privacy amplification for decentralized optimization and averaging. *Advances in Neural Information Processing Systems* 35 (2022), 15889–15902.

[19] Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*. Springer, 1–12.

[20] C. Dwork and A. Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.

[21] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. 2018. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems* 31 (2018).

[22] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. 2020. Local model poisoning attacks to Byzantine-robust federated learning. In *USENIX Security Symposium*.

[23] Minghong Fang, Zifan Zhang, Hairi, Prashant Khanduri, Jia Liu, Songtao Lu, Yuchen Liu, and Neil Gong. 2024. Byzantine-robust decentralized federated learning. In *CCS*.

[24] Chong Fu, Xuhong Zhang, Shouling Ji, Jinyin Chen, Jingzheng Wu, Shanqing Guo, Jun Zhou, Alex X Liu, and Ting Wang. 2022. Label inference attacks against vertical federated learning. In *31st USENIX security symposium (USENIX Security 22)*. 1397–1414.

[25] Hongchang Gao. 2022. Decentralized stochastic gradient descent ascent for finite-sum minimax problems. *arXiv preprint arXiv:2212.02724* (2022).

[26] Hongchang Gao, Yubin Duan, Yihan Zhang, and Jie Wu. 2024. Decentralized Stochastic Compositional Gradient Descent for AUPRC Maximization. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*. SIAM, 226–234.

[27] Arvind Giridhar and PR Kumar. 2006. Toward a theory of in-network computation in wireless sensor networks. *IEEE Communications magazine* 44, 4 (2006), 98–107.

[28] Tomás González, Cristóbal Guzmán, and Courtney Paquette. 2024. Mirror descent algorithms with nearly dimension-independent rates for differentially-private stochastic saddle-point problems. *arXiv preprint arXiv:2403.02912* (2024).

[29] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[30] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. *Advances in neural information processing systems* 30 (2017).

[31] Ehsan Hallaji, Roozbeh Razavi-Far, Mehrdad Saif, Boyu Wang, and Qiang Yang. 2024. Decentralized federated learning: A survey on security and privacy. *IEEE Transactions on Big Data* (2024).

[32] Songyang Han, Sanbao Su, Sihong He, Shuo Han, Haizhao Yang, and Fei Miao. 2022. What is the Solution for State-Adversarial Multi-Agent Reinforcement Learning? *arXiv preprint arXiv:2212.02705* (2022).

[33] Diba Hashemi, Lie He, and Martin Jaggi. 2024. CoBo: Collaborative Learning via Bilevel Optimization. arXiv:2409.05539 [cs.LG] https://arxiv.org/abs/2409.05539

[34] Sihong He, Songyang Han, Sanbao Su, Shuo Han, Shaofeng Zou, and Fei Miao. 2023. Robust multi-agent reinforcement learning with state uncertainty. *arXiv preprint arXiv:2307.16212* (2023).

[35] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, and Xuyun Zhang. 2021. Source inference attacks in federated learning. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1102–1107.

[36] Rui Hu, Yuanxiong Guo, Hongning Li, Qingqi Pei, and Yanmin Gong. 2020. Personalized federated learning with differential privacy. *IEEE Internet of Things Journal* 7, 10 (2020), 9530–9539.

[37] Feihu Huang and Songcan Chen. 2023. Near-optimal decentralized momentum method for nonconvex-PL minimax problems. *arXiv preprint arXiv:2304.10902* (2023).

[38] Feihu Huang, Xidong Wu, and Heng Huang. 2021. Efficient mirror descent ascent methods for nonsmooth minimax problems. *Advances in Neural Information Processing Systems* 34 (2021), 10431–10443.

[39] Yilin Kang, Yong Liu, Jian Li, and Weiping Wang. 2022. Stability and generalization of differentially private minimax problems. *arXiv preprint arXiv:2204.04858* (2022).

[40] Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. 2020. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606* (2020).

[41] Harsh Kasyap and Somanath Tripathy. 2021. Privacy-preserving decentralized learning framework for healthcare system. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 17, 2s (2021), 1–24.

[42] Prashant Khanduri, Pranay Sharma, Haibo Yang, Mingyi Hong, Jia Liu, Ketan Rajawat, and Pramod Varshney. 2021. Stem: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning. *Advances in Neural Information Processing Systems* 34 (2021), 6050–6061.

[43] Juno Kim, Tai Nakamaki, and Taiji Suzuki. 2024. Transformers are Minimax Optimal Nonparametric In-Context Learners. *arXiv preprint arXiv:2408.12186*

(2024).
[44] Alec Koppel, Felicia Y Jakubiec, and Alejandro Ribeiro. 2015. A saddle point algorithm for networked online convex optimization. *IEEE Transactions on Signal Processing* 63, 19 (2015), 5149–5164.
[45] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).
[46] Tsung-Ting Kuo and Lucila Ohno-Machado. 2018. Modelchain: Decentralized privacy-preserving healthcare predictive modeling framework on private blockchain networks. *arXiv preprint arXiv:1802.01746* (2018).
[47] Jonathan Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. 2024. Supervised pretraining can learn in-context reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2024).
[48] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. 2017. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*. 5330–5340.
[49] Tianyi Lin, Chi Jin, and Michael Jordan. 2020. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*. PMLR, 6083–6093.
[50] Tianyi Lin, Chi Jin, and Michael I. Jordan. 2020. Near-Optimal Algorithms for Minimax Optimization. In *Proceedings of Thirty Third Conference on Learning Theory (Proceedings of Machine Learning Research, Vol. 125)*, Jacob Abernethy and Shivani Agarwal (Eds.). PMLR, 2738–2779. https://proceedings.mlr.press/v125/lin20a.html
[51] Wanyu Lin, Baochun Li, and Cong Wang. 2022. Towards private learning on decentralized graphs with local differential privacy. *IEEE Transactions on Information Forensics and Security* 17 (2022), 2936–2946.
[52] Weijie Liu, Aryan Mokhtari, Asuman Ozdaglar, Sarath Pattathil, Zebang Shen, and Nenggan Zheng. 2019. A decentralized proximal point-type method for saddle point problems. *arXiv preprint arXiv:1910.14380* (2019).
[53] Y. Liu and R. Liu. 2021. BOML: A modularized bilevel optimization library in Python for meta-learning. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 1–2.
[54] Songtao Lu, Siliang Zeng, Xiaodong Cui, Mark Squillante, Lior Horesh, Brian Kingsbury, Jia Liu, and Mingyi Hong. 2022. A stochastic linearized augmented Lagrangian method for decentralized bilevel optimization. *Advances in Neural Information Processing Systems* (2022).
[55] Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. 2020. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems* 33 (2020), 20566–20577.
[56] David Mateos-Núnez and Jorge Cortés. 2015. Distributed subgradient methods for saddle-point problems. In *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE, 5462–5467.
[57] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*. IEEE, 739–753.
[58] Angelia Nedic and Asuman Ozdaglar. 2009. Distributed subgradient methods for multi-agent optimization. *IEEE Trans. Automat. Control* 54, 1 (2009), 48.
[59] Anh Ninh and Zuo-Jun Max Shen. 2024. Stochastic Resource Allocation Problems: Minmax and Maxmin Solutions. *Manufacturing & Service Operations Management* 26, 6 (2024), 2322–2335.
[60] C. Poon and G. Peyré. 2021. Smooth bilevel programming for sparse regularization. In *Advances in Neural Information Processing Systems*, Vol. 34. 1543–1555.
[61] A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine. 2019. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, Vol. 32.
[62] Christopher Regan, Mohammad Nasajpour, Reza M Parizi, Seyedamin Pouriyeh, Ali Dehghantanha, and Kim-Kwang Raymond Choo. 2022. Federated IoT attack detection using decentralized edge data. *Machine Learning with Applications* 8 (2022), 100263.
[63] Shoupeng Ren, Lipeng He, Tianyu Tu, Di Wu, Jian Liu, Kui Ren, and Chun Chen. 2024. LookAhead: Preventing DeFi Attacks via Unveiling Adversarial Contracts. *arXiv preprint arXiv:2401.07261* (2024).
[64] Devansh Shah, Parijat Dube, Supriyo Chakraborty, and Ashish Verma. 2021. Adversarial training in communication constrained federated learning. *arXiv preprint arXiv:2103.01319* (2021).
[65] Chamani Shiranthika, Parvaneh Saeedi, and Ivan V Bajić. 2023. Decentralized learning in healthcare: a review of emerging techniques. *IEEE Access* 11 (2023), 54188–54209.
[66] Bernardo Camajori Tedeschini, Stefano Savazzi, Roman Stoklasa, Luca Barbieri, Ioannis Stathopoulos, Monica Nicoli, and Luigi Serio. 2022. Decentralized federated learning for healthcare networks: A case study on tumor segmentation. *IEEE access* 10 (2022), 8693–8708.
[67] Stacey Truex, Ling Liu, Ka-Ho Chow, Mehmet Emre Gursoy, and Wenqi Wei. 2020. LDP-Fed: Federated learning with local differential privacy. In *Proceedings of the third ACM international workshop on edge systems, analytics and networking*. 61–66.
[68] Ioannis Tsaknakis, Mingyi Hong, and Sijia Liu. 2020. 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)*. IEEE, 5755–5759.
[69] Hoi-To Wai, Mingyi Hong, Zhuoran Yang, Zhaoran Wang, and Kexin Tang. 2019. Variance reduced policy evaluation with smooth function approximation. *Advances in Neural Information Processing Systems* 32 (2019).
[70] Di Wang, Marco Gaboardi, and Jinhui Xu. 2018. Empirical risk minimization in non-interactive local differential privacy revisited. *Advances in Neural Information Processing Systems* 31 (2018).
[71] Di Wang, Minwei Ye, and Jinhui Xu. 2017. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems* 30 (2017).
[72] Yongqiang Wang and Tamer Başar. 2023. Decentralized nonconvex optimization with guaranteed privacy and accuracy. *Automatica* 150 (2023), 110858.
[73] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. 2020. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE transactions on information forensics and security* 15 (2020), 3454–3469.
[74] Wenhan Xian, Feihu Huang, Yanfu Zhang, and Heng Huang. 2021. A faster decentralized algorithm for nonconvex minimax problems. *Advances in Neural Information Processing Systems* 34 (2021), 25865–25877.
[75] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).
[76] Ran Xin, Usman Khan, and Soummya Kar. 2021. A hybrid variance-reduced method for decentralized stochastic non-convex optimization. In *International Conference on Machine Learning*. PMLR, 11459–11469.
[77] Yangyang Xu. 2024. Decentralized gradient descent maximization method for composite nonconvex strongly-concave minimax problems. *SIAM Journal on Optimization* 34, 1 (2024), 1006–1044.
[78] Zhixiong Yang, Arpita Gang, and Waheed U Bajwa. 2019. Adversary-resilient inference and machine learning: From distributed to decentralized. *stat* 1050 (2019), 23.
[79] Zhenhuan Yang, Shu Hu, Yunwen Lei, Kush R Vashney, Siwei Lyu, and Yiming Ying. 2022. Differentially private sgda for minimax problems. In *Uncertainty in Artificial Intelligence*. PMLR, 2192–2202.
[80] Liang Zhang, Kiran K Thekumparampil, Sewoong Oh, and Niao He. 2022. Bring your own algorithm for optimal differentially private stochastic minimax optimization. *Advances in Neural Information Processing Systems* 35 (2022), 35174–35187.
[81] Xinwei Zhang, Xiangyi Chen, Mingyi Hong, Zhiwei Steven Wu, and Jinfeng Yi. 2022. Understanding clipping for federated learning: Convergence and client-level differential privacy. In *International Conference on Machine Learning, ICML 2022*.
[82] Xin Zhang, Minghong Fang, Zhuqing Liu, Haibo Yang, Jia Liu, and Zhengyuan Zhu. 2022. Net-fleet: Achieving linear convergence speedup for fully decentralized federated learning with heterogeneous data. In *MobiHoc*.
[83] Xin Zhang, Zhuqing Liu, Jia Liu, Zhengyuan Zhu, and Songtao Lu. 2021. Taming communication and sample complexities in decentralized policy evaluation for cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 18825–18838.
[84] Canzhe Zhao, Yanjie Ze, Jing Dong, Baoxiang Wang, and Shuai Li. 2023. Differentially private temporal difference learning with stochastic nonconvex-strongly-concave optimization. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 985–993.
[85] Liyi Zhou, Xihan Xiong, Jens Ernstberger, Stefanos Chaliasos, Zhipeng Wang, Ye Wang, Kaihua Qin, Roger Wattenhofer, Dawn Song, and Arthur Gervais. 2023. Sok: Decentralized finance (defi) attacks. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2444–2461.
[86] Xinyu Zhou and Raef Bassily. 2024. Differentially private worst-group risk minimization. *arXiv preprint arXiv:2402.19437* (2024).
[87] Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep leakage from gradients. *Advances in neural information processing systems* 32 (2019).
[88] Giulio Zizzo, Ambrish Rawat, Mathieu Sinn, and Beat Buesser. 2020. Fat: Federated adversarial training. *arXiv preprint arXiv:2012.01791* (2020).

A Proof of convergence

From the algorithm, we obtain that:

$$\begin{aligned}\bar{\mathbf{g}}_t^* &= \bar{\mathbf{g}}_t + \frac{1}{m}\sum_{i=1}^{m} n_{\mathbf{x},t}^{(i)} \\ &= \frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}} F_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}, \mathbf{z}_t^{(i)}\right) + (1-\beta_{\mathbf{x}})\left(\bar{\mathbf{g}}_{t-1} - \frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}} F_i\left(\mathbf{x}_{t-1}^{(i)}, \mathbf{y}_{t-1}^{(i)}, \mathbf{z}_t^{(i)}\right)\right) + \mathcal{N}_{\mathbf{x},t}\end{aligned} \tag{25}$$

Where $\mathcal{N}_{\mathbf{x},t} = \frac{\sum_{i=1}^{m} n_{\mathbf{x},t}^{(i)}}{m}$

A.1 Basic Lemmas and Important Conclusions

First, we introduce following basic lemmas, which are broadly used in the convergence analysis of optimization algorithms.

Lemma 2. Let vector X be a stochastic variable. Then we have

$$0 \le \mathbb{E}\|X - \mathbb{E}X\|^2 = \mathbb{E}\|X\|^2 - \|\mathbb{E}X\|^2 \le \mathbb{E}\|X\|^2 \tag{26}$$

Lemma 3. Let $X_1, X_2, \cdots, X_n$ be m independent stochastic variables of which the means are 0. Then we have

$$\mathbb{E}\left\|\sum_{i=1}^{m} X_i\right\|^2 = \sum_{i=1}^{m}\mathbb{E}\|X_i\|^2 \tag{27}$$

Lemma 4. Suppose A and B are two matrices. Then it satisfies

$$\|AB\|_F \le \|A\|_2\|B\|_F \tag{28}$$

Lemma 5. (Lemma 4.3 from [49]) $\Phi(\mathbf{x})$ is $(L+\kappa L)$-smooth and $\mathbf{y}^*(\cdot)$ is κ-Lipschitz, which means $\|\mathbf{y}^*(\mathbf{x}_1) - \mathbf{y}^*(\mathbf{x}_2)\| \le \kappa\|\mathbf{x}_1 - \mathbf{x}_2\|$ for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^{d_1}$.

Lemma 6. When $\eta_{\mathbf{y}} \le \frac{1}{5L}$ we have following estimation for δ_t.

$$\begin{aligned}\sum_{t=0}^{T-1}\delta_t \le{}& \frac{4\kappa}{L\eta_{\mathbf{y}}}\delta_0 + \frac{10\eta_{\mathbf{y}}}{\mu}\sum_{t=1}^{T-1}\left(1-\frac{\mu\eta_{\mathbf{y}}}{4}\right)^{T-t-1}\sum_{s=0}^{t-1}\left\|\bar{\mathbf{u}}_s - \frac{1}{m}\sum_{i=1}^{m}\nabla f_i\left(\mathbf{x}_s^{(i)}, \mathbf{y}_s^{(i)}\right)\right\|^2 + \frac{40\kappa^2}{m}\sum_{t=0}^{T-1} \\ &\left(\left\|X_t - \bar{X}_t\right\|_F^2 + \left\|Y_t - \bar{Y}_t\right\|_F^2\right) + \frac{20\kappa^4\eta_{\mathbf{x}}^2}{L^2\eta_{\mathbf{y}}^2}\sum_{t=0}^{T-1}\|\bar{\mathbf{v}}_t\|^2 - \frac{14\eta_{\mathbf{y}}}{5\mu}\sum_{t=0}^{T-1}\left(1-\left(1-\frac{\mu\eta_{\mathbf{y}}}{4}\right)^{T-t}\right)\|\bar{\mathbf{u}}_t\|^2\end{aligned} \tag{29}$$

Proof: As we defined before, $\hat{\mathbf{y}}_t = \arg\max_{\mathbf{y}\in\mathcal{Y}} f(\bar{\mathbf{x}}_t, \mathbf{y})$, we obatin:

$$\|\bar{\mathbf{y}}_{t+1} - \hat{\mathbf{y}}_t\|^2 = \left\|\bar{\mathbf{y}}_t + \eta_{\mathbf{y}}\bar{\mathbf{u}}_t - \hat{\mathbf{y}}_t\right\|^2 = \|\bar{\mathbf{y}}_t - \hat{\mathbf{y}}_t\|^2 + \eta_{\mathbf{y}}^2\|\bar{\mathbf{u}}_t\|^2 + 2\eta_{\mathbf{y}}\langle\bar{\mathbf{y}}_t - \hat{\mathbf{y}}_t, \bar{\mathbf{u}}_t\rangle \tag{30}$$

As function f is strongly-concave in $\mathbf{y}$, we get:

$$\begin{aligned}f(\bar{\mathbf{x}}_t, \hat{\mathbf{y}}_t) &\le f(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t) + \left\langle\nabla_{\mathbf{y}} f(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t), \hat{\mathbf{y}}_t - \bar{\mathbf{y}}_t\right\rangle - \frac{\mu}{2}\|\hat{\mathbf{y}}_t - \bar{\mathbf{y}}_t\|^2 \\ &= f(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t) - \frac{\mu}{2}\|\hat{\mathbf{y}}_t - \bar{\mathbf{y}}_t\|^2 + \langle\bar{\mathbf{u}}_t, \hat{\mathbf{y}}_t - \bar{\mathbf{y}}_{t+1}\rangle + \left\langle\nabla_{\mathbf{y}} f(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t) - \bar{\mathbf{u}}_t, \hat{\mathbf{y}}_t - \bar{\mathbf{y}}_{t+1}\right\rangle \\ &\quad + \left\langle\nabla_{\mathbf{y}} f(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t), \bar{\mathbf{y}}_{t+1} - \bar{\mathbf{y}}_t\right\rangle\end{aligned} \tag{31}$$

From assumption 1, and let $L\eta_{\mathbf{y}} \le \frac{1}{5}$ we know:

$$\begin{aligned}-\frac{1}{10\eta_{\mathbf{y}}}\|\bar{\mathbf{y}}_{t+1} - \bar{\mathbf{y}}_t\|^2 &\le -\frac{L}{2}\|\bar{\mathbf{y}}_{t+1} - \bar{\mathbf{y}}_t\|^2 \\ &\le f(\bar{\mathbf{x}}_t, \mathbf{y}_{t+1}) - f(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t) - \left\langle\nabla_{\mathbf{y}} f(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t), \bar{\mathbf{y}}_{t+1} - \bar{\mathbf{y}}_t\right\rangle\end{aligned} \tag{32}$$

Adding Eq.(31) and (32) together, and from the algorithm, we know that $\bar{\mathbf{y}}_{t+1} - \bar{\mathbf{y}}_t = \mathbf{u}_t$.

$$
\begin{aligned}
& f\left(\bar{\mathbf{x}}_t, \hat{\mathbf{y}}_t\right) - f\left(\bar{\mathbf{x}}_t, \mathbf{y}_{t+1}\right) + \frac{\mu}{2}\left\|\hat{\mathbf{y}}_t - \bar{\mathbf{y}}_t\right\|^2 \\
& \leq \left\langle \bar{\mathbf{u}}_t, \hat{\mathbf{y}}_t - \bar{\mathbf{y}}_{t+1}\right\rangle + \left\langle \nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{u}}_t, \hat{\mathbf{y}}_t - \bar{\mathbf{y}}_{t+1}\right\rangle + \frac{1}{10\eta_{\mathrm{y}}}\left\|\bar{\mathbf{y}}_{t+1} - \bar{\mathbf{y}}_t\right\|^2 \\
& = \left\langle \bar{\mathbf{u}}_t, \bar{\mathbf{y}}_t - \bar{\mathbf{y}}_t\right\rangle + \left\langle \bar{\mathbf{u}}_t, \hat{\mathbf{y}}_t - \bar{\mathbf{y}}_{t+1}\right\rangle + \left\langle \nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{u}}_t, \hat{\mathbf{y}}_t - \bar{\mathbf{y}}_{t+1}\right\rangle + \frac{\eta_{\mathrm{y}}}{10}\left\|\bar{\mathbf{u}}_t\right\|^2 \\
& = \left\langle \bar{\mathbf{u}}_t, \hat{\mathbf{y}}_t - \bar{\mathbf{y}}_t\right\rangle - \eta_{\mathrm{y}}\left\|\bar{\mathbf{u}}_t\right\|^2 + \left\langle \nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{u}}_t, \hat{\mathbf{y}}_t - \bar{\mathbf{y}}_{t+1}\right\rangle + \frac{\eta_{\mathrm{y}}}{10}\left\|\bar{\mathbf{u}}_t\right\|^2 \\
& = \left\langle \bar{\mathbf{u}}_t, \hat{\mathbf{y}}_t - \bar{\mathbf{y}}_t\right\rangle + \left\langle \nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{u}}_t, \hat{\mathbf{y}}_t - \bar{\mathbf{y}}_{t+1}\right\rangle - \frac{9\eta_{\mathrm{y}}}{10}\left\|\bar{\mathbf{u}}_t\right\|^2 \\
& \leq \left\langle \bar{\mathbf{u}}_t, \hat{\mathbf{y}}_t - \bar{\mathbf{y}}_t\right\rangle + \frac{2}{\mu}\left\|\nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{u}}_t\right\|^2 + \frac{\mu}{8}\left\|\hat{\mathbf{y}}_t - \bar{\mathbf{y}}_{t+1}\right\|^2 - \frac{9\eta_{\mathrm{y}}}{10}\left\|\bar{\mathbf{u}}_t\right\|^2 \\
& \leq \left\langle \bar{\mathbf{u}}_t, \hat{\mathbf{y}}_t - \bar{\mathbf{y}}_t\right\rangle + \frac{2}{\mu}\left\|\nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{u}}_t\right\|^2 + \frac{\mu}{4}\left\|\hat{\mathbf{y}}_t - \bar{\mathbf{y}}_t\right\|^2 + \frac{\mu}{4}\left\|\bar{\mathbf{y}}_t - \bar{\mathbf{y}}_{t+1}\right\|^2 - \frac{9\eta_{\mathrm{y}}}{10}\left\|\bar{\mathbf{u}}_t\right\|^2 \\
& = \left\langle \bar{\mathbf{u}}_t, \hat{\mathbf{y}}_t - \bar{\mathbf{y}}_t\right\rangle + \frac{2}{\mu}\left\|\nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{u}}_t\right\|^2 + \frac{\mu}{4}\left\|\hat{\mathbf{y}}_t - \bar{\mathbf{y}}_t\right\|^2 - \left(\frac{9\eta_{\mathrm{y}}}{10} - \frac{\mu\eta_{\mathrm{y}}^2}{4}\right)\left\|\bar{\mathbf{u}}_t\right\|^2
\end{aligned}
\tag{33}
$$

Where in the second inequality, we use Young's inequality, and in the last inequality we use Cauchy-Schwartz inequality. As we defined $\hat{\mathbf{y}}_t$, so $f\left(\bar{\mathbf{x}}_t, \hat{\mathbf{y}}_t\right) \geq f\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_{t+1}\right)$.

$$
\frac{\mu\eta_{\mathrm{y}}}{2}\left\|\hat{\mathbf{y}}_t - \bar{\mathbf{y}}_t\right\|^2 \leq 2\eta_{\mathrm{y}}\left\langle \bar{\mathbf{u}}_t, \hat{\mathbf{y}}_t - \bar{\mathbf{y}}_t\right\rangle + \frac{4\eta_{\mathrm{y}}}{\mu}\left\|\nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{u}}_t\right\|^2 - \left(\frac{9\eta_{\mathrm{y}^2}}{5} - \frac{\mu\eta_{\mathrm{y}}^3}{2}\right)\left\|\bar{\mathbf{u}}_t\right\|^2 \tag{34}
$$

Combining Eq.(30) and (34), and we set $\mu\eta_{\mathrm{y}} \leq L\eta_{\mathrm{y}} \leq \frac{1}{5}$ we can get:

$$
\left\|\bar{\mathbf{y}}_{t+1} - \hat{\mathbf{y}}_t\right\|^2 \leq \left(1 - \frac{\mu\eta_{\mathrm{y}}}{2}\right)\left\|\hat{\mathbf{y}}_t - \bar{\mathbf{y}}_t\right\|^2 + \frac{4\eta_{\mathrm{y}}}{\mu}\left\|\nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{u}}_t\right\|^2 - \frac{7\eta_{\mathrm{y}}^2}{10}\left\|\bar{\mathbf{u}}_t\right\|^2 \tag{35}
$$

By Young's inequality we have:

$$
\begin{aligned}
& \left\|\bar{\mathbf{y}}_{t+1} - \hat{\mathbf{y}}_{t+1}\right\|^2 \\
& \leq \left(1 + \frac{\mu\eta_{\mathrm{y}}}{4}\right)\left\|\bar{\mathbf{y}}_{t+1} - \hat{\mathbf{y}}_t\right\|^2 + \left(1 + \frac{4}{\mu\eta_{\mathrm{y}}}\right)\left\|\hat{\mathbf{y}}_{t+1} - \hat{\mathbf{y}}_t\right\|^2 \\
& \leq \left(1 - \frac{\mu\eta_{\mathrm{y}}}{4} - \frac{\mu^2\eta_{\mathrm{y}}^2}{8}\right)\left\|\bar{\mathbf{y}}_t - \hat{\mathbf{y}}_t\right\|^2 + \left(\frac{4\eta_{\mathrm{y}}}{\mu} + \eta_{\mathrm{y}}^2\right)\left\|\nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{u}}_t\right\|^2 \\
& \quad + \frac{\mu\eta_{\mathrm{y}} + 4}{\mu\eta_{\mathrm{y}}}\left\|\hat{\mathbf{y}}_{t+1} - \hat{\mathbf{y}}_t\right\|^2 - \left(1 + \frac{\mu\eta_{\mathrm{y}}}{4}\right)\frac{7\eta_{\mathrm{y}}^2}{10}\left\|\bar{\mathbf{u}}_t\right\|^2 \\
& \leq \left(1 - \frac{\mu\eta_{\mathrm{y}}}{4}\right)\left\|\bar{\mathbf{y}}_t - \hat{\mathbf{y}}_t\right\|^2 + \frac{5\eta_{\mathrm{y}}}{\mu}\left\|\nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{u}}_t\right\|^2 + \frac{5\kappa^3\eta_{\mathrm{x}}^2}{L\eta_{\mathrm{y}}}\left\|\bar{\mathbf{v}}_t\right\|^2 - \frac{7\eta_{\mathrm{y}}{}^2}{10}\left\|\bar{\mathbf{u}}_t\right\|^2
\end{aligned}
\tag{36}
$$

Using Eq. (35) and the fact that $L\eta_{\mathrm{y}} \leq \frac{1}{5}$ in the calculation of the second inequality, we obtain $-\frac{\mu^2\eta_{\mathrm{y}}^2}{8} \geq 0$. Additionally, $\frac{4\eta_{\mathrm{y}}}{\mu} + \eta_{\mathrm{y}}^2 \leq \frac{4\eta_{\mathrm{y}}}{\mu} + \frac{1\eta_{\mathrm{y}}}{5\mu} \leq \frac{5\eta_{\mathrm{y}}}{\mu}$, and $-\left(1 + \frac{\mu\eta_{\mathrm{y}}}{4}\right) \leq -1$. We simplify the inequality using an approximation method, and the last inequality holds because the function $\mathbf{y}^*(\cdot)$ is κ-Lipschitz, therefore, we have $\left\|\hat{\mathbf{y}}_{t+1} - \hat{\mathbf{y}}_t\right\|^2 \leq \kappa^2\eta_{\mathrm{x}}^2\left\|\bar{\mathbf{v}}_t^2\right\|$ In combination with the previously provided conditions, we have $\frac{\mu\eta_{\mathrm{y}}+4}{\mu\eta_{\mathrm{y}}} \leq \frac{5}{\mu\eta_{\mathrm{y}}} = \frac{5\kappa}{L\eta_{\mathrm{y}}}$. By the Cauchy-Schwarz inequality and Assumption 1, we also have:

$$
\left\|\nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right) - \bar{\mathbf{u}}_t\right\|^2 \leq 2\left\|\bar{\mathbf{u}}_t - \frac{1}{m}\sum_{i=1}^{m}\nabla f_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}\right)\right\|^2 + \frac{2L^2}{m}\left(\left\|X_t - \bar{X}_t\right\|_F^2 + \left\|Y_t - \bar{Y}_t\right\|_F^2\right) \tag{37}
$$

By definition of δ_t and the recursion in Eq.(36) we obtain:

$$
\begin{aligned}
\delta_t \leq & \left(1 - \frac{\mu\eta_{\mathrm{y}}}{4}\right)^t \delta_0 + \frac{5\eta_{\mathrm{y}}}{\mu}\sum_{s=0}^{t-1}\left(1 - \frac{\mu\eta_{\mathrm{y}}}{4}\right)^{t-s-1}\left\|\bar{\mathbf{u}}_s - \nabla_{\mathbf{y}} f\left(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t\right)\right\|^2 \\
& + \frac{5\kappa^3\eta_{\mathrm{x}}^2}{L\eta_{\mathrm{y}}}\sum_{s=0}^{t-1}\left(1 - \frac{\mu\eta_{\mathrm{y}}}{4}\right)^{t-s-1}\left\|\bar{\mathbf{v}}_s\right\|^2 - \frac{7\eta_{\mathrm{y}}^2}{10}\sum_{s=0}^{t-1}\left(1 - \frac{\mu\eta_{\mathrm{y}}}{4}\right)^{t-s-1}\left\|\bar{\mathbf{u}}_s\right\|^2
\end{aligned}
\tag{38}
$$

Using Eq.(37) to sum above equation we have:

$$\sum_{t=0}^{T-1}\delta_t \le \frac{4\kappa}{L\eta_{\mathbf{y}}}\delta_0 + \frac{10\eta_{\mathbf{y}}}{\mu}\sum_{t=1}^{T-1}\left(1-\frac{\mu\eta_{\mathbf{y}}}{4}\right)^{T-t-1}\sum_{s=0}^{t-1}\left\|\bar{\mathbf{u}}_s - \frac{1}{m}\sum_{i=1}^{m}\nabla f_i\left(\mathbf{x}_s^{(i)},\mathbf{y}_s^{(i)}\right)\right\|^2 + \frac{40\kappa^2}{m}\sum_{t=0}^{T-1}$$
$$\left(\left\|X_t-\bar{X}_t\right\|_F^2 + \left\|Y_t-\bar{Y}_t\right\|_F^2\right) + \frac{20\kappa^4\eta_{\mathbf{x}}^2}{L^2\eta_{\mathbf{y}}^2}\sum_{t=0}^{T-1}\|\bar{\mathbf{v}}_t\|^2 - \frac{14\eta_{\mathbf{y}}}{5\mu}\sum_{t=0}^{T-1}\left(1-\left(1-\frac{\mu\eta_{\mathbf{y}}}{4}\right)^{T-t}\right)\|\bar{\mathbf{u}}_t\|^2 \tag{39}$$

Lemma 7. For all $t \in \{0,1,\cdots,T\}$ we have $\bar{\mathbf{v}}_t = \bar{\mathbf{g}}_t^*$ and $\bar{\mathbf{u}}_t = \bar{\mathbf{h}}_t^*$.

Proof: As matrix W is doubly stochastic, we have:

$$\bar{\mathbf{v}}_t = \bar{\mathbf{v}}_{t-1} + \bar{\mathbf{g}}_t^* - \bar{\mathbf{g}}_{t-1}^* \tag{40}$$

which is equivalent to $\bar{\mathbf{v}}_t - \bar{\mathbf{g}}_t^* = \bar{\mathbf{v}}_{t-1} - \bar{\mathbf{g}}_{t-1}^*$. Additionally, $\bar{\mathbf{v}}_{-1} = \bar{\mathbf{g}}_{-1}^*$, so $\bar{\mathbf{v}}_t = \bar{\mathbf{g}}_t^*$. Thus, from the above: $\bar{\mathbf{v}}_t = \bar{\mathbf{g}}_t + \mathcal{N}_{\mathbf{x},t}$.

Lemma 8. Let A_t, B_t be positive sequences satisfying

$$A_{t+1} \le (1-c)A_t + B_t \tag{41}$$

for some constant $c \in (0,1)$. Then for any positive integer T we have

$$\sum_{t=0}^{T} A_t \le \frac{1}{c}A_0 + \frac{1}{c}\sum_{t=0}^{T-1} B_t \tag{42}$$

Proof: Using recursion on Eq.(41) we can obtain

$$A_t \le (1-c)^t A_0 + \sum_{s=0}^{t-1}(1-c)^{t-s-1}B_s \tag{43}$$

for $\forall t \ge 0$. Sum above inequality and we achieve the desired conclusion Eq.(42), where we use the condtion A_t, B_t are positive and the fact that $\sum_{t=0}^{\infty}(1-c)^t = \frac{1}{c}$.

Lemma 9. We can prove the following bound for gradient estimator $\bar{\mathbf{v}}_t$ and $\bar{\mathbf{u}}_t$.

$$\begin{aligned}
&\sum_{s=0}^{t-1}\mathbb{E}\left\|\bar{\mathbf{v}}_s - \frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}} f_i\left(\mathbf{x}_s^{(i)},\mathbf{y}_s^{(i)}\right)\right\|^2 \\
&\le \frac{2\sigma^2}{mb_0\beta_{\mathbf{x}}} + \frac{2\beta_{\mathbf{x}}\sigma^2 t}{m} + \frac{12L^2}{m^2\beta_{\mathbf{x}}}\sum_{s=0}^{t-1}\left(\mathbb{E}\left\|X_s-\bar{X}_s\right\|_F^2 + \mathbb{E}\left\|Y_s-\bar{Y}_s\right\|_F^2\right) + \frac{6L^2}{m\beta_{\mathbf{x}}}\sum_{s=0}^{t-2} \\
&\quad\left(\eta_{\mathbf{x}}^2\mathbb{E}\|\bar{\mathbf{v}}_s\|^2 + \eta_{\mathbf{y}}^2\mathbb{E}\|\bar{\mathbf{u}}_s\|^2\right) + 2\sum_{s=0}^{t-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s} - \mathcal{N}_{\mathbf{x},s-1}\right\|^2 \\
&\sum_{s=0}^{t-1}\mathbb{E}\left\|\bar{\mathbf{u}}_s - \frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{y}} f_i\left(\mathbf{x}_s^{(i)},\mathbf{y}_s^{(i)}\right)\right\|^2 \\
&\le \frac{2\sigma^2}{mb_0\beta_{\mathbf{y}}} + \frac{2\beta_{\mathbf{y}}\sigma^2 t}{m} + \frac{12L^2}{m^2\beta_{\mathbf{y}}}\sum_{s=0}^{t-1}\left(\mathbb{E}\left\|X_s-\bar{X}_s\right\|_F^2 + \mathbb{E}\left\|Y_s-\bar{Y}_s\right\|_F^2\right) + \frac{6L^2}{m\beta_{\mathbf{y}}}\sum_{s=0}^{t-2} \\
&\quad\left(\eta_{\mathbf{x}}^2\mathbb{E}\|\bar{\mathbf{v}}_s\|^2 + \eta_{\mathbf{y}}^2\mathbb{E}\|\bar{\mathbf{u}}_s\|^2\right) + 2\sum_{s=0}^{t-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s} - \mathcal{N}_{\mathbf{y},s-1}\right\|^2
\end{aligned} \tag{44}$$

for all $t \in \{1,2,\cdots,T\}$.

Proof: By the definition of $\mathbf{g}_t^{(i)}$ and Lemma 7, now we have

$$\begin{aligned}
&\bar{\mathbf{v}}_t - \mathcal{N}_{\mathbf{x},t} - \frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}} f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right) \\
&= (1-\beta_{\mathbf{x}})\left(\bar{\mathbf{v}}_{t-1} - \mathcal{N}_{\mathbf{x},t-1} - \frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}} f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)\right) + \frac{\beta_{\mathbf{x}}}{m}\sum_{i=1}^{m}\left(\nabla_{\mathbf{x}} F_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_t^{(i)}\right)\right. \\
&\quad\left.-\nabla_{\mathbf{x}} f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right) + (1-\beta_{\mathbf{x}})\frac{1}{m}\sum_{i=1}^{m}\left(\nabla_{\mathbf{x}} F_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_t^{(i)}\right) - \nabla_{\mathbf{x}} F_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)};\mathbf{z}_t^{(i)}\right)\right. \\
&\quad\left.+\nabla_{\mathbf{x}} f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right) - \nabla_{\mathbf{x}} f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right)
\end{aligned} \tag{45}$$

Taking expectation $\left(\mathbf{z}_t^{(i)}\right)$, the last two terms of Equation above are 0. Therefore, By using Cauchy-Schwarz inequality, we obtain:

$$\begin{aligned}
&\mathbb{E}\left\|\bar{\mathbf{v}}_t-\frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right\|^2\\
&\leq\mathbb{E}\left\|(1-\beta_{\mathbf{x}})\left(\bar{\mathbf{v}}_{t-1}-\frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)\right)+\mathcal{N}_{\mathbf{x},t}-(1-\beta_{\mathbf{x}})\mathcal{N}_{\mathbf{x},t-1}\right\|^2\\
&\quad+\mathbb{E}\left\|\frac{\beta_{\mathbf{x}}}{m}\sum_{i=1}^{m}\left(\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_t^{(i)}\right)-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right)\right.\\
&\quad+2(1-\beta_{\mathbf{x}})\frac{1}{m}\sum_{i=1}^{m}\left(\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_t^{(i)}\right)-\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)};\mathbf{z}_t^{(i)}\right)\right.\\
&\quad\left.\left.+\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right)\right\|^2\\
&\leq 2(1-\beta_{\mathbf{x}})^2\mathbb{E}\left\|\bar{\mathbf{v}}_{t-1}-\frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)\right\|^2+\frac{2\beta_{\mathbf{x}}^2}{m^2}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_t^{(i)}\right)\right.\\
&\quad\left.-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right\|^2+\frac{2(1-\beta_{\mathbf{x}})^2}{m^2}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_t^{(i)}\right)-\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)};\mathbf{z}_t^{(i)}\right)\right.\\
&\quad\left.+\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right\|^2+2\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},t}-(1-\beta_{\mathbf{x}})\mathcal{N}_{\mathbf{x},t-1}\right\|^2
\end{aligned}\tag{46}$$

The first inequality is obtained by Cauchy-Schwartz inequality. Then we use Lemma 3 on the last two terms, and then use Assumption 2, Lemma 2 and Assumption 1, we can obtain.

$$\begin{aligned}
&\mathbb{E}\left\|\bar{\mathbf{v}}_t-\frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right\|^2\\
&\leq 2(1-\beta_{\mathbf{x}})^2\mathbb{E}\left\|\bar{\mathbf{v}}_{t-1}-\frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)\right\|^2+2\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},t}-(1-\beta_{\mathbf{x}})\mathcal{N}_{\mathbf{x},t-1}\right\|^2\\
&\quad+\frac{2\beta_{\mathbf{x}}^2\sigma^2}{m}+\frac{2L^2(1-\beta_{\mathbf{x}})^2}{m^2}\left(\mathbb{E}\left\|X_t-X_{t-1}\right\|_F^2+\mathbb{E}\left\|Y_t-Y_{t-1}\right\|_F^2\right)
\end{aligned}\tag{47}$$

At the same time, by using Cauchy-Schwarz inequality, we have a rewritten form for $X_t - X_{t-1}$, $Y_t - Y_{t-1}$.

$$\begin{aligned}
\left\|X_t-X_{t-1}\right\|_F^2&\leq 3\left\|X_t-\bar{X}_t\right\|_F^2+3m\eta_{\mathbf{x}}^2\left\|\bar{\mathbf{v}}_{t-1}\right\|^2+3\left\|X_{t-1}-\bar{X}_{t-1}\right\|_F^2\\
\left\|Y_t-Y_{t-1}\right\|_F^2&\leq 3\left\|Y_t-\bar{Y}_t\right\|_F^2+3m\eta_{\mathbf{y}}^2\left\|\bar{\mathbf{u}}_{t-1}\right\|^2+3\left\|Y_{t-1}-\bar{Y}_{t-1}\right\|_F^2
\end{aligned}\tag{48}$$

Combining above two inequalities with Eq.(47) and Lemma 8, we have:

$$\begin{aligned}
&\sum_{s=0}^{t-1}\mathbb{E}\left\|\bar{\mathbf{v}}_s-\frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_s^{(i)},\mathbf{y}_s^{(i)}\right)\right\|^2\\
&\leq\frac{2}{\beta_{\mathbf{x}}}\mathbb{E}\left\|\bar{\mathbf{v}}_0-\nabla_{\mathbf{x}}f\left(\mathbf{x}_0,\mathbf{y}_0\right)\right\|^2+\frac{2\beta_{\mathbf{x}}\sigma^2t}{m}+\frac{12L^2}{m^2\beta_{\mathbf{x}}}\sum_{s=0}^{t-1}\left(\mathbb{E}\left\|X_s-\bar{X}_s\right\|_F^2+\mathbb{E}\left\|Y_s-\bar{Y}_s\right\|_F^2\right)\\
&\quad+\frac{6L^2}{m\beta_{\mathbf{x}}}\sum_{s=0}^{t-2}\left(\eta_{\mathbf{x}}^2\mathbb{E}\left\|\bar{\mathbf{v}}_s\right\|^2+\eta_{\mathbf{y}}^2\mathbb{E}\left\|\bar{\mathbf{u}}_s\right\|^2\right)+2\sum_{s=0}^{t-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s}-(1-\beta_{\mathbf{x}})\mathcal{N}_{\mathbf{x},s-1}\right\|^2\\
&\leq\frac{2\sigma^2}{mb_0\beta_{\mathbf{x}}}+\frac{2\beta_{\mathbf{x}}\sigma^2t}{m}+\frac{12L^2}{m^2\beta_{\mathbf{x}}}\sum_{s=0}^{t-1}\left(\mathbb{E}\left\|X_s-\bar{X}_s\right\|_F^2+\mathbb{E}\left\|Y_s-\bar{Y}_s\right\|_F^2\right)\\
&\quad+\frac{6L^2}{m\beta_{\mathbf{x}}}\sum_{s=0}^{t-2}\left(\eta_{\mathbf{x}}^2\mathbb{E}\left\|\bar{\mathbf{v}}_s\right\|^2+\eta_{\mathbf{y}}^2\mathbb{E}\left\|\bar{\mathbf{u}}_s\right\|^2\right)+2\sum_{s=0}^{t-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s}-(1-\beta_{\mathbf{x}})\mathcal{N}_{\mathbf{x},s-1}\right\|^2
\end{aligned}\tag{49}$$

for all $t\in\{1,2,\cdots,T\}$. In the first inequality we use the fact $\frac{1}{1-(1-\beta_{\mathbf{x}})^2}\leq\frac{1}{\beta_{\mathbf{x}}}$ when $\beta_{\mathbf{x}}\leq 1$. Where $\mathbb{E}\left\|\bar{\mathbf{v}}_0-\nabla_{\mathbf{x}}f\left(\mathbf{x}_0,\mathbf{y}_0\right)\right\|^2\leq\frac{\sigma^2}{mb_0}$ holds because of Assumption 2 and Lemma 3. Mimic above steps we can also prove the second conclusion.

Lemma 10. The consensus error satisfies the following recursive relation

$$\begin{aligned}\left\|X_{t+1}-\bar{X}_{t+1}\right\|_F^2 &\leq \frac{1+\lambda^2}{2}\left\|X_t-\bar{X}_t\right\|_F^2+\frac{2\lambda^2\eta_{\mathbf{x}}^2}{1-\lambda^2}\left\|V_t-\bar{V}_t\right\|_F^2\\ \left\|Y_{t+1}-\bar{Y}_{t+1}\right\|_F^2 &\leq \frac{1+\lambda^2}{2}\left\|Y_t-\bar{Y}_t\right\|_F^2+\frac{2\lambda^2\eta_{\mathbf{y}}^2}{1-\lambda^2}\left\|U_t-\bar{U}_t\right\|_F^2\end{aligned} \tag{50}$$

Proof. As we set $J=\frac{11^T}{n}$, then we obtain:

$$\begin{aligned}\left\|X_{t+1}-\bar{X}_{t+1}\right\|_F^2 &= \left\|\left(X_t-\eta_{\mathbf{x}}V_t\right)W-\left(\bar{X}_t-\eta_{\mathbf{x}}\bar{V}_t\right)\right\|_F^2\\ &= \left\|\left(X_t-\bar{X}_t\right)(W-J)-\eta_{\mathbf{x}}\left(V_t-\bar{V}_t\right)(W-J)\right\|_F^2\end{aligned} \tag{51}$$

Now, we use Lemma 4 and Assumption 4 on Eq.(51)

$$\begin{aligned}&\left\|X_{t+1}-\bar{X}_{t+1}\right\|_F^2\\ &\leq \lambda^2\left\|X_t-\bar{X}_t\right\|_F^2+\lambda^2\eta_{\mathbf{x}}^2\left\|V_t-\bar{V}_t\right\|_F^2-2\left\langle\left(X_t-\bar{X}_t\right)(W-J),\eta_{\mathbf{x}}\left(V_t-\bar{V}_t\right)(W-J)\right\rangle\end{aligned} \tag{52}$$

We use the Young's inequality to eliminate the last term above, and we set the constant $\alpha=\frac{1-\lambda^2}{2\lambda^2}$.

$$\begin{aligned}\left\|X_{t+1}-\bar{X}_{t+1}\right\|_F^2 &\leq \left(\lambda^2+\alpha\lambda^2\right)\left\|X_t-\bar{X}_t\right\|_F^2+\left(\frac{\lambda^2\eta_{\mathbf{x}}^2}{\alpha}+\lambda^2\eta_{\mathbf{x}}^2\right)\left\|V_t-\bar{V}_t\right\|_F^2\\ &\leq \frac{1+\lambda^2}{2}\left\|X_t-\bar{X}_t\right\|_F^2+\frac{2\lambda^2\eta_{\mathbf{x}}^2}{1-\lambda^2}\left\|V_t-\bar{V}_t\right\|_F^2\end{aligned} \tag{53}$$

Mimic the steps above, we can also get:

$$\left\|Y_{t+1}-\bar{Y}_{t+1}\right\|_F^2 \leq \frac{1+\lambda^2}{2}\left\|Y_t-\bar{Y}_t\right\|_F^2+\frac{2\lambda^2\eta_{\mathbf{y}}^2}{1-\lambda^2}\left\|U_t-\bar{U}_t\right\|_F^2 \tag{54}$$

Lemma 11. For all $t'\in\{0,1,\cdots,T-1\}$ we have

$$\begin{aligned}&\sum_{s=0}^{t'}\mathbb{E}\left\|V_s-\bar{V}_s\right\|_F^2\\ &\leq \frac{2}{1-\lambda^2}\mathbb{E}\left\|V_0-\bar{V}_0\right\|_F^2+\frac{48\lambda^2L^2}{\left(1-\lambda^2\right)^2}\sum_{s=0}^{t'}\left(\mathbb{E}\left\|X_s-\bar{X}_s\right\|_F^2+\mathbb{E}\left\|Y_s-\bar{Y}_s\right\|_F^2\right)\\ &\quad+\frac{24m\lambda^2L^2}{\left(1-\lambda^2\right)^2}\sum_{s=0}^{t'-1}\eta_{\mathbf{y}}^2\mathbb{E}\left\|\bar{\mathbf{u}}_s\right\|^2+\frac{24m\lambda^2L^2}{\left(1-\lambda^2\right)^2}\sum_{s=0}^{t'-1}\eta_{\mathbf{x}}^2\mathbb{E}\left\|\bar{\mathbf{v}}_s\right\|^2+\frac{8m\lambda^2\beta_{\mathbf{x}}^2\sigma^2t'}{1-\lambda^2}\\ &\quad+\frac{8\lambda^2\beta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^2}\sum_{s=0}^{t'-1}\sum_{i=1}^{m}\mathbb{E}\left\|\mathbf{g}_s^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_s^{(i)},\mathbf{y}_s^{(i)}\right)\right\|^2+\frac{8\lambda^2m}{1-\lambda^2}\sum_{s=0}^{t'}\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s}-\mathcal{N}_{\mathbf{x},s-1}\right\|^2\\ &\sum_{s=0}^{t'}\mathbb{E}\left\|U_s-\bar{U}_s\right\|_F^2\\ &\leq \frac{2}{1-\lambda^2}\mathbb{E}\left\|U_0-\bar{U}_0\right\|_F^2+\frac{48\lambda^2L^2}{\left(1-\lambda^2\right)^2}\sum_{s=0}^{t'}\left(\mathbb{E}\left\|X_s-\bar{X}_s\right\|_F^2+\mathbb{E}\left\|Y_s-\bar{Y}_s\right\|_F^2\right)\\ &\quad+\frac{24m\lambda^2L^2}{\left(1-\lambda^2\right)^2}\sum_{s=0}^{t'-1}\eta_{\mathbf{y}}^2\mathbb{E}\left\|\bar{\mathbf{u}}_s\right\|^2+\frac{24m\lambda^2L^2}{\left(1-\lambda^2\right)^2}\sum_{s=0}^{t'-1}\eta_{\mathbf{x}}^2\mathbb{E}\left\|\bar{\mathbf{v}}_s\right\|^2+\frac{8m\lambda^2\beta_{\mathbf{y}}^2\sigma^2t'}{1-\lambda^2}\\ &\quad+\frac{8\lambda^2\beta_{\mathbf{y}}^2}{\left(1-\lambda^2\right)^2}\sum_{s=0}^{t'-1}\sum_{i=1}^{m}\mathbb{E}\left\|\mathbf{h}_s^{(i)}-\nabla_{\mathbf{y}}f_i\left(\mathbf{x}_s^{(i)},\mathbf{y}_s^{(i)}\right)\right\|^2+\frac{8\lambda^2m}{1-\lambda^2}\sum_{s=0}^{t'}\mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s}-\mathcal{N}_{\mathbf{y},s-1}\right\|^2\end{aligned} \tag{55}$$

Proof: Similar as Eq.(51) and (52), by definition of V_t, we obtain:

$$\begin{aligned}&\left\|V_{t+1}-\bar{V}_{t+1}\right\|_F^2\\ &\leq \lambda^2\left\|V_t-\bar{V}_t\right\|_F^2+\lambda^2\left\|G_{t+1}^*-G_t^*\right\|_F^2+2\left\langle\left(V_t-\bar{V}_t\right)(W-J),\left(G_{t+1}^*-G_t^*\right)(W-J)\right\rangle\end{aligned} \tag{56}$$

By the definition of $\mathbf{g}_t^{(i)*}$:

$$\begin{aligned}
&\mathbf{g}_{t+1}^{(i)*} - \mathbf{g}_t^{(i)*} \\
&= \nabla_{\mathbf{x}} F_i\left(\mathbf{x}_{t+1}^{(i)}, \mathbf{y}_{t+1}^{(i)}; \mathbf{z}_{t+1}^{(i)}\right) - \nabla_{\mathbf{x}} F_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}; \mathbf{z}_{t+1}^{(i)}\right) - \beta_{\mathbf{x}}\left(\mathbf{g}_t^{(i)} - \nabla_{\mathbf{x}} f_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}\right)\right) \\
&\quad + \beta_{\mathbf{x}}\left(\nabla_{\mathbf{x}} F_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}; \mathbf{z}_{t+1}^{(i)}\right) - \nabla_{\mathbf{x}} f_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}\right)\right) + n_{\mathbf{x},t+1}^{(i)} - n_{\mathbf{x},t}^{(i)}
\end{aligned} \tag{57}$$

Because of $\mathbb{E}\left[n_{\mathbf{x},t+1}^{(i)} - n_{\mathbf{x},t}^{(i)}\right] = 0 - 0 = 0$, thus

$$\mathbb{E}\left[\mathbf{g}_{t+1}^{(i)^*} - \mathbf{g}_t^{(i)^*}\right] = \nabla_{\mathbf{x}} f_i\left(\mathbf{x}_{t+1}^{(i)}, \mathbf{y}_{t+1}^{(i)}\right) - \nabla_{\mathbf{x}} f_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}\right) - \beta_{\mathbf{x}}\left(\mathbf{g}_t^{(i)} - \nabla_{\mathbf{x}} f_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}\right)\right) \tag{58}$$

Considering about all the agents, we get:

$$\left\|\mathbb{E}\left[G_{t+1}^* - G_t^*\right]\right\|^2 = \sum_{i=1}^{m}\left\|\nabla_{\mathbf{x}} f_i\left(\mathbf{x}_{t+1}^{(i)}, \mathbf{y}_{t+1}^{(i)}\right) - \nabla_{\mathbf{x}} f_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}\right) - \beta_{\mathbf{x}}\left(\mathbf{g}_t^{(i)} - \nabla_{\mathbf{x}} f_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}\right)\right)\right\|^2 \tag{59}$$

Taking expectation on $\mathbf{z}_{t+1}^{(i)}$ the last term of Eq.(56) can be bounded by

$$\begin{aligned}
&\mathbb{E}\left\langle\left(V_t - \bar{V}_t\right)(W - J), \left(G_{t+1}^* - G_t^*\right)(W - J)\right\rangle \\
&= \left\langle\left(V_t - \bar{V}_t\right)(W - J), \mathbb{E}\left[G_{t+1}^* - G_t^*\right](W - J)\right\rangle \le \lambda\left\|V_t - \bar{V}_t\right\|_F \cdot \lambda\left\|\mathbb{E}\left[G_{t+1}^* - G_t^*\right]\right\|_F \\
&\le \frac{1-\lambda^2}{4}\left\|V_t - \bar{V}_t\right\|_F^2 + \frac{\lambda^4}{1-\lambda^2}\left\|\mathbb{E}\left[G_{t+1}^* - G_t^*\right]\right\|_F^2 \\
&\le \frac{1-\lambda^2}{4}\left\|V_t - \bar{V}_t\right\|_F^2 + \frac{2\lambda^4}{1-\lambda^2}\sum_{i=1}^{m}\left\|\nabla_{\mathbf{x}} f_i\left(\mathbf{x}_{t+1}^{(i)}, \mathbf{y}_{t+1}^{(i)}\right) - \nabla_{\mathbf{x}} f_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}\right)\right\|^2 \\
&\quad + \frac{2\lambda^4\beta_{\mathbf{x}}^2}{1-\lambda^2}\sum_{i=1}^{m}\left\|\mathbf{g}_t^{(i)} - \nabla_{\mathbf{x}} f_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}\right)\right\|^2 \\
&\le \frac{1-\lambda^2}{4}\left\|V_t - \bar{V}_t\right\|_F^2 + \frac{2\lambda^4 L^2}{1-\lambda^2}\left(\|X_{t+1} - X_t\|_F^2 + \|Y_{t+1} - Y_t\|_F^2\right) \\
&\quad + \frac{2\lambda^4\beta_{\mathbf{x}}^2}{1-\lambda^2}\sum_{i=1}^{m}\left\|\mathbf{g}_t^{(i)} - \nabla_{\mathbf{x}} f_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}\right)\right\|^2
\end{aligned} \tag{60}$$

Where we use Young's inequality in the second inequality, and then we use Cauchy-Schwartz inequality in the third inequality, and the last inequality is resulted from Assumption 1. Besides, applying Cauchy-Schwartz inequality to Eq.(57) we have:

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{g}_{t+1}^{(i)^*} - \mathbf{g}_t^{(i)^*}\right\|^2 \\
&\le 4\mathbb{E}\left\|\nabla_{\mathbf{x}} F_i\left(\mathbf{x}_{t+1}^{(i)}, \mathbf{y}_{t+1}^{(i)}; \mathbf{z}_{t+1}^{(i)}\right) - \nabla_{\mathbf{x}} F_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}; \mathbf{z}_{t+1}^{(i)}\right)\right\|^2 + 4\beta_{\mathbf{x}}^2\mathbb{E}\left\|\mathbf{g}_t^{(i)} - \nabla_{\mathbf{x}} f_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}\right)\right\|^2 \\
&\quad + 4\beta_{\mathbf{x}}^2\mathbb{E}\left\|\nabla_{\mathbf{x}} F_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}; \mathbf{z}_{t+1}^{(i)}\right) - \nabla_{\mathbf{x}} f_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}\right)\right\|^2 + 4\mathbb{E}\left\|n_{\mathbf{x},t+1}^{(i)} - n_{\mathbf{x},t}^{(i)}\right\|^2 \\
&\le 4L^2\left(\mathbb{E}\left\|\mathbf{x}_{t+1}^{(i)} - \mathbf{x}_t^{(i)}\right\|^2 + \mathbb{E}\left\|\mathbf{y}_{t+1}^{(i)} - \mathbf{y}_t^{(i)}\right\|^2\right) + 4\beta_{\mathbf{x}}^2\mathbb{E}\left\|\mathbf{g}_t^{(i)} - \nabla_{\mathbf{x}} f_i\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}\right)\right\|^2 + 4\beta_{\mathbf{x}}^2\sigma^2 \\
&\quad + 4\mathbb{E}\left\|n_{\mathbf{x},t+1}^{(i)} - n_{\mathbf{x},t}^{(i)}\right\|^2
\end{aligned} \tag{61}$$

where in the last inequality we use Assumption 1 and Assumption 2. Combining Eq.(56) (60), (61) and the definition of $\mathcal{N}_{\mathbf{x},t}$ we can obtain

$$\begin{aligned}
&\mathbb{E}\left\|V_{t+1}-\bar{V}_{t+1}\right\|_F^2 \\
&\leq \frac{1+\lambda^2}{2}\mathbb{E}\left\|V_t-\bar{V}_t\right\|_F^2+\frac{4\lambda^4L^2}{1-\lambda^2}\left(\mathbb{E}\left\|X_{t+1}-X_t\right\|_F^2+\mathbb{E}\left\|Y_{t+1}-Y_t\right\|_F^2\right) \\
&\quad+\frac{4\lambda^4\beta_{\mathbf{x}}^2}{1-\lambda^2}\sum_{i=1}^m\mathbb{E}\left\|\mathbf{g}_t^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right\|^2+4L^2\lambda^2\left(\mathbb{E}\left\|X_{t+1}-X_t\right\|^2+\mathbb{E}\left\|Y_{t+1}-Y_t\right\|^2\right) \\
&\quad+4\beta_{\mathbf{x}}^2\lambda^2\sum_{i=1}^m\mathbb{E}\left\|\mathbf{g}_t^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right\|^2+4m\lambda^2\beta_{\mathbf{x}}^2\sigma^2+4\lambda^2m\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},t+1}-\mathcal{N}_{\mathbf{x},t}\right\|^2 \\
&=\frac{1+\lambda^2}{2}\mathbb{E}\left\|V_t-\bar{V}_t\right\|_F^2+\frac{\left(4\lambda^4L^2+4L^2\lambda^2-4L^2\lambda^4\right)}{1-\lambda^2}\left(\mathbb{E}\left\|X_{t+1}-X_t\right\|_F^2+\mathbb{E}\left\|Y_{t+1}-Y_t\right\|_F^2\right) \\
&\quad+\frac{\left(4\lambda^4\beta_{\mathbf{x}}^2+4\beta_{\mathbf{x}}^2\lambda^2-4\beta_{\mathbf{x}}^2\lambda^4\right)}{1-\lambda^2}\sum_{i=1}^m\mathbb{E}\left\|\mathbf{g}_t^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right\|^2+4m\lambda^2\beta_{\mathbf{x}}^2\sigma^2 \\
&\quad+4\lambda^2m\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},t+1}-\mathcal{N}_{\mathbf{x},t}\right\|^2
\end{aligned}\tag{62}$$

Then, we have the result below.

$$\begin{aligned}
&\mathbb{E}\left\|V_{t+1}-\bar{V}_{t+1}\right\|_F^2 \\
&\leq \frac{1+\lambda^2}{2}\mathbb{E}\left\|V_t-\bar{V}_t\right\|_F^2+\frac{4\lambda^2L^2}{1-\lambda^2}\left(\mathbb{E}\left\|X_{t+1}-X_t\right\|_F^2+\mathbb{E}\left\|Y_{t+1}-Y_t\right\|_F^2\right) \\
&\quad+\frac{4\lambda^2\beta_{\mathbf{x}}^2}{1-\lambda^2}\sum_{i=1}^m\mathbb{E}\left\|\mathbf{g}_t^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right\|^2+4m\lambda^2\beta_{\mathbf{x}}^2\sigma^2+4\lambda^2m\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},t+1}-\mathcal{N}_{\mathbf{x},t}\right\|^2
\end{aligned}\tag{63}$$

Using Eq.(48) for substitution:

$$\begin{aligned}
&\mathbb{E}\left\|V_{t+1}-\bar{V}_{t+1}\right\|_F^2 \\
&\leq\frac{1+\lambda^2}{2}\mathbb{E}\left\|V_t-\bar{V}_t\right\|_F^2+\frac{12\lambda^2L^2}{1-\lambda^2}\left(\mathbb{E}\left\|X_{t+1}-\bar{X}_{t+1}\right\|_F^2+\mathbb{E}\left\|Y_{t+1}-\bar{Y}_{t+1}\right\|_F^2\right) \\
&\quad+\frac{12\lambda^2L^2}{1-\lambda^2}\left(\mathbb{E}\left\|X_t-\bar{X}_t\right\|_F^2+\mathbb{E}\left\|Y_t-\bar{Y}_t\right\|_F^2\right)+\frac{12m\lambda^2L^2\eta_{\mathbf{y}}^2}{1-\lambda^2}\mathbb{E}\left\|\bar{\mathbf{u}}_t\right\|^2+4m\lambda^2\beta_{\mathbf{x}}^2\sigma^2 \\
&\quad+\frac{12m\lambda^2L^2\eta_{\mathbf{x}}^2}{1-\lambda^2}\mathbb{E}\left\|\bar{\mathbf{v}}_t\right\|^2+\frac{4\lambda^2\beta_{\mathbf{x}}^2}{1-\lambda^2}\sum_{i=1}^m\mathbb{E}\left\|\mathbf{g}_t^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right\|^2 \\
&\quad+4\lambda^2m\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},t+1}-\mathcal{N}_{\mathbf{x},t}\right\|^2
\end{aligned}\tag{64}$$

Summing over Eq.(64), we obtain:

$$\begin{aligned}
&\sum_{s=0}^{t'}\mathbb{E}\left\|V_s-\bar{V}_s\right\|_F^2 \\
&\leq \frac{2}{1-\lambda^2}\mathbb{E}\left\|V_0-\bar{V}_0\right\|_F^2+\frac{48\lambda^2L^2}{\left(1-\lambda^2\right)^2}\sum_{s=0}^{t'}\left(\mathbb{E}\left\|X_s-\bar{X}_s\right\|_F^2+\mathbb{E}\left\|Y_s-\bar{Y}_s\right\|_F^2\right) \\
&\quad+\frac{24m\lambda^2L^2}{\left(1-\lambda^2\right)^2}\sum_{s=0}^{t'-1}\eta_{\mathbf{y}}^2\mathbb{E}\left\|\bar{\mathbf{u}}_s\right\|^2+\frac{24m\lambda^2L^2}{\left(1-\lambda^2\right)^2}\sum_{s=0}^{t'-1}\eta_{\mathbf{x}}^2\mathbb{E}\left\|\bar{\mathbf{v}}_s\right\|^2+\frac{8m\lambda^2\beta_{\mathbf{x}}^2\sigma^2t'}{1-\lambda^2} \\
&\quad+\frac{8\lambda^2\beta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^2}\sum_{s=0}^{t'-1}\sum_{i=1}^m\mathbb{E}\left\|\mathbf{g}_s^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_s^{(i)},\mathbf{y}_s^{(i)}\right)\right\|^2+\frac{8\lambda^2m}{1-\lambda^2}\sum_{s=0}^{t'}\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s}-\mathcal{N}_{\mathbf{x},s-1}\right\|^2
\end{aligned}\tag{65}$$

for all $t'\in\{0,1,\cdots,T-1\}$. Here we should notice that term $\mathbb{E}\left\|X_{t+1}-\bar{X}_{t+1}\right\|_F^2$ in Eq.(64) is summed from $\mathbb{E}\left\|X_1-\bar{X}_1\right\|_F^2$ to $\mathbb{E}\left\|X_{t'}-\bar{X}_{t'}\right\|_F^2$, while term $\mathbb{E}\left\|X_t-\bar{X}_t\right\|_F^2$ is summed from $\mathbb{E}\left\|X_0-\bar{X}_0\right\|_F^2$ to $\mathbb{E}\left\|X_{t'-1}-\bar{X}_{t'-1}\right\|_F^2$. As $X_0=\bar{X}_0$, these two terms can be merged together. And it is the same with term $\mathbb{E}\left\|Y_{t+1}-\bar{Y}_{t+1}\right\|_F^2$. Mimic above steps and we can prove the conclusion for $\sum_{s=0}^{t'}\mathbb{E}\left\|U_s-\bar{U}_s\right\|_F^2$ in the similar way.

Lemma 12. We have the local average gradient estimators $\bar{\mathbf{g}}_t$ and $\bar{\mathbf{h}}_t$ satisfy the following conclusion

$$\begin{aligned}
\sum_{s=0}^{t}\sum_{i=1}^{m}\mathbb{E}\left\|\mathbf{g}_s^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_s^{(i)},\mathbf{y}_s^{(i)}\right)\right\|^2 \leq & \frac{m\sigma^2}{\beta_{\mathbf{x}}b_0}+2m\beta_{\mathbf{x}}\sigma^2 t+\frac{12L^2}{\beta_{\mathbf{x}}}\sum_{s=0}^{t}\left(\mathbb{E}\left\|X_s-\bar{X}_s\right\|_F^2\right. \\
& \left.+\mathbb{E}\left\|Y_s-\bar{Y}_s\right\|_F^2\right)+\frac{6mL^2}{\beta_{\mathbf{x}}}\sum_{s=0}^{t-1}\left(\eta_{\mathbf{x}}^2\mathbb{E}\left\|\bar{\mathbf{v}}_s\right\|^2+\eta_{\mathbf{y}}^2\mathbb{E}\left\|\bar{\mathbf{u}}_s\right\|^2\right) \\
\sum_{s=0}^{t}\sum_{i=1}^{m}\mathbb{E}\left\|\mathbf{h}_s^{(i)}-\nabla_{\mathbf{y}}f_i\left(\mathbf{x}_s^{(i)},\mathbf{y}_s^{(i)}\right)\right\|^2 \leq & \frac{m\sigma^2}{\beta_{\mathbf{y}}b_0}+2m\beta_{\mathbf{y}}\sigma^2 t+\frac{12L^2}{\beta_{\mathbf{y}}}\sum_{s=0}^{t}\left(\mathbb{E}\left\|X_s-\bar{X}_s\right\|_F^2\right. \\
& \left.+\mathbb{E}\left\|Y_s-\bar{Y}_s\right\|_F^2\right)+\frac{6mL^2}{\beta_{\mathbf{y}}}\sum_{s=0}^{t-1}\left(\eta_{\mathbf{x}}^2\mathbb{E}\left\|\bar{\mathbf{v}}_s\right\|^2+\eta_{\mathbf{y}}^2\mathbb{E}\left\|\bar{\mathbf{u}}_s\right\|^2\right)
\end{aligned} \tag{66}$$

Proof: According to the definition of $\mathbf{g}_t^{(i)}$, we have:

$$\begin{aligned}
& \mathbf{g}_t^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right) \\
& =(1-\beta_{\mathbf{x}})\left(\mathbf{g}_{t-1}^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)\right)+\beta_{\mathbf{x}}\left(\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_t^{(i)}\right)-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right) \\
& \quad+(1-\beta_{\mathbf{x}})\left(\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_t^{(i)}\right)-\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)};\mathbf{z}_t^{(i)}\right)\right.+\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right) \\
& \quad\left.-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right)
\end{aligned} \tag{67}$$

The last two terms of Eq.(67) will be 0 if we taking expectation of $\mathbf{z}_t^{(i)}$.

$$\begin{aligned}
& \mathbb{E}\left\|(1-\beta_{\mathbf{x}})\left(\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_t^{(i)}\right)-\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)};\mathbf{z}_t^{(i)}\right)\right.+\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)\right. \\
& \quad\left.\left.-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right)\right\|^2 \\
& \leq 2\beta_{\mathbf{x}}^2\mathbb{E}\left\|\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_t^{(i)}\right)-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right\|^2+2(1-\beta_{\mathbf{x}})^2\mathbb{E}\left\|\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_t^{(i)}\right)\right. \\
& \quad\left.-\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)};\mathbf{z}_t^{(i)}\right)\right\|^2
\end{aligned} \tag{68}$$

As Eq.(68) is 0, by using Cauchy-Schwartz we get:

$$\begin{aligned}
& \mathbb{E}\left\|\mathbf{g}_t^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right\|^2 \\
& \leq(1-\beta_{\mathbf{x}})^2\mathbb{E}\left\|\mathbf{g}_{t-1}^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)\right\|^2+2\beta_{\mathbf{x}}^2\mathbb{E}\left\|\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_t^{(i)}\right)\right. \\
& \quad\left.-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right\|^2+2(1-\beta_{\mathbf{x}})^2\mathbb{E}\left\|\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_t^{(i)}\right)-\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)};\mathbf{z}_t^{(i)}\right)\right\|^2 \\
& \leq(1-\beta_{\mathbf{x}})^2\mathbb{E}\left\|\mathbf{g}_{t-1}^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)\right\|^2+2\beta_{\mathbf{x}}^2\sigma^2+2(1-\beta_{\mathbf{x}})^2L^2\left(\mathbb{E}\left\|\mathbf{x}_t^{(i)}-\mathbf{x}_{t-1}^{(i)}\right\|^2\right. \\
& \quad\left.+\mathbb{E}\left\|\mathbf{y}_t^{(i)}-\mathbf{y}_{t-1}^{(i)}\right\|^2\right)
\end{aligned} \tag{69}$$

where we use Assumption 1 and Assumption 2 in the last inequality. Sum above inequality from $i = 1$ to m and we have:

$$\begin{aligned}
&\sum_{i=1}^{m}\mathbb{E}\left\|\mathbf{g}_t^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right\|^2\\
&\le(1-\beta_{\mathbf{x}})^2\sum_{i=1}^{m}\mathbb{E}\left\|\mathbf{g}_{t-1}^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)\right\|^2\\
&\quad+2m\beta_{\mathbf{x}}^2\sigma^2+2\left(1-\beta_{\mathbf{x}}\right)^2L^2\left(\mathbb{E}\left\|X_t-X_{t-1}\right\|^2+\mathbb{E}\left\|Y_t-Y_{t-1}\right\|^2\right)\\
&\le(1-\beta_{\mathbf{x}})^2\sum_{i=1}^{m}\mathbb{E}\left\|\mathbf{g}_{t-1}^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)\right\|^2\\
&\quad+2m\beta_{\mathbf{x}}^2\sigma^2+12\left(1-\beta_{\mathbf{x}}\right)^2L^2\left(\mathbb{E}\left\|X_t-\bar{X}_t\right\|_F^2+\mathbb{E}\left\|Y_t-\bar{Y}_t\right\|_F^2\right)\\
&\quad+6m\left(1-\beta_{\mathbf{x}}\right)^2L^2\left(\eta_{\mathbf{x}}^2\mathbb{E}\left\|\bar{\mathbf{v}}_{t-1}\right\|^2+\eta_{\mathbf{y}}^2\mathbb{E}\left\|\bar{\mathbf{u}}_{t-1}\right\|^2\right)
\end{aligned}\tag{70}$$

Applying Lemma 8 to Eq.(70), we have:

$$\begin{aligned}
&\sum_{s=0}^{t}\sum_{i=1}^{m}\mathbb{E}\left\|\mathbf{g}_s^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_s^{(i)},\mathbf{y}_s^{(i)}\right)\right\|^2\\
&\le\frac{1}{\beta_{\mathbf{x}}}\sum_{i=1}^{m}\mathbb{E}\left\|\mathbf{g}_0^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_0^{(i)},\mathbf{y}_0^{(i)}\right)\right\|^2+\frac{12L^2}{\beta_{\mathbf{x}}}\sum_{s=0}^{t}\left(\mathbb{E}\left\|X_s-\bar{X}_s\right\|_F^2+\mathbb{E}\left\|Y_s-\bar{Y}_s\right\|_F^2\right)\\
&\quad+\frac{6mL^2}{\beta_{\mathbf{x}}}\sum_{s=0}^{t-1}\left(\eta_{\mathbf{x}}^2\mathbb{E}\left\|\bar{\mathbf{v}}_s\right\|^2+\eta_{\mathbf{y}}^2\mathbb{E}\left\|\bar{\mathbf{u}}_s\right\|^2\right)+2m\beta_{\mathbf{x}}\sigma^2t\\
&\le\frac{m\sigma^2}{\beta_{\mathbf{x}}b_0}+2m\beta_{\mathbf{x}}\sigma^2t+\frac{12L^2}{\beta_{\mathbf{x}}}\sum_{s=0}^{t}\left(\mathbb{E}\left\|X_s-\bar{X}_s\right\|_F^2+\mathbb{E}\left\|Y_s-\bar{Y}_s\right\|_F^2\right)\\
&\quad+\frac{6mL^2}{\beta_{\mathbf{x}}}\sum_{s=0}^{t-1}\left(\eta_{\mathbf{x}}^2\mathbb{E}\left\|\bar{\mathbf{v}}_s\right\|^2+\eta_{\mathbf{y}}^2\mathbb{E}\left\|\bar{\mathbf{u}}_s\right\|^2\right)
\end{aligned}\tag{71}$$

for all $t\in\{0,1,\cdots,T-1\}$. Here the last inequality is derived by $\mathbb{E}\left\|\mathbf{g}_0^{(i)}-\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_0^{(i)},\mathbf{y}_0^{(i)}\right)\right\|^2\le\frac{\sigma^2}{b_0}$ due to Lemma 3. The estimation of $\mathbf{h}_t^{(i)}$ can be achieved in the same way as above.

Lemma 13. Let $\eta_{\mathbf{x}}\le\frac{(1-\lambda)^2}{500L}$ and $\eta_{\mathbf{y}}\le\frac{(1-\lambda)^2}{500L}$. The consensus error can be bounded by

$$\begin{aligned}
&\sum_{s=0}^{t}\left(\mathbb{E}\left\|X_s-\bar{X}_s\right\|_F^2+\mathbb{E}\left\|Y_s-\bar{Y}_s\right\|_F^2\right)\\
&\le\frac{16\lambda^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^3}\mathbb{E}\left\|V_0-\bar{V}_0\right\|_F^2+\frac{16\lambda^2\eta_{\mathbf{y}}^2}{\left(1-\lambda^2\right)^3}\mathbb{E}\left\|U_0-\bar{U}_0\right\|_F^2+\frac{576m\lambda^4L^2\left(\eta_{\mathbf{x}}^2+\eta_{\mathbf{y}}^2\right)}{\left(1-\lambda^2\right)^4}\sum_{s=0}^{t-2}\left(\eta_{\mathbf{x}}^2\mathbb{E}\left\|\bar{\mathbf{v}}_s\right\|^2\right.\\
&\quad\left.+\eta_{\mathbf{y}}^2\mathbb{E}\left\|\bar{\mathbf{u}}_s\right\|^2\right)+\frac{64m\lambda^4\left(\beta_{\mathbf{x}}\eta_{\mathbf{x}}^2+\beta_{\mathbf{y}}\eta_{\mathbf{y}}^2\right)\sigma^2}{\left(1-\lambda^2\right)^4b_0}+\frac{192m\lambda^4\left(\beta_{\mathbf{x}}^2\eta_{\mathbf{x}}^2+\beta_{\mathbf{y}}^2\eta_{\mathbf{y}}^2\right)\sigma^2t}{\left(1-\lambda^2\right)^4}\\
&\quad+\frac{64m\lambda^4\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^3}\sum_{s=0}^{t-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s}-\mathcal{N}_{\mathbf{x},s-1}\right\|^2+\frac{64m\lambda^4\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^3}\mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s}-\mathcal{N}_{\mathbf{y},s-1}\right\|^2
\end{aligned}\tag{72}$$

Proof: Combining Lemma 8 and Lemma 10 , for all $t \in \{0, 1, \cdots, T\}$ we have:

$$
\begin{aligned}
&\sum_{s=0}^{t} \left\|X_s - \bar{X}_s\right\|_F^2 \\
&\leq \frac{4\lambda^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^2} \sum_{s=0}^{t-1} \left\|V_s - \bar{V}_s\right\|_F^2 \\
&\leq \frac{8\lambda^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^3} \mathbb{E}\left\|V_0 - \bar{V}_0\right\|_F^2 + \frac{192\lambda^4 L^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^4} \sum_{s=0}^{t-1} \left(\mathbb{E}\left\|X_s - \bar{X}_s\right\|_F^2 + \mathbb{E}\left\|Y_s - \bar{Y}_s\right\|_F^2\right) \\
&\quad + \frac{96m\lambda^4 L^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^4} \sum_{s=0}^{t-2} \eta_{\mathbf{y}}^2 \mathbb{E}\left\|\bar{\mathbf{u}}_s\right\|^2 + \frac{96m\lambda^4 L^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^4} \sum_{s=0}^{t-2} \eta_{\mathbf{x}}^2 \mathbb{E}\left\|\bar{\mathbf{v}}_s\right\|^2 + \frac{32m\lambda^4\beta_{\mathbf{x}}^2\eta_{\mathbf{x}}^2\sigma^2(t-1)}{\left(1-\lambda^2\right)^3} \\
&\quad + \frac{32m\lambda^4\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^3} \sum_{s=0}^{t-1} \mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s} - \mathcal{N}_{\mathbf{x},s-1}\right\|^2 + \frac{32\lambda^4\beta_{\mathbf{x}}^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^4} \sum_{s=0}^{t-2} \sum_{i=1}^{m} \mathbb{E}\left\|\mathbf{g}_s^{(i)} - \nabla_{\mathbf{x}} f_i\left(\mathbf{x}_s^{(i)}, \mathbf{y}_s^{(i)}\right)\right\|^2
\end{aligned}
\tag{73}
$$

Where we use Lemma 11 in the last inequality. Using Lemma 12 to replace the last term in the result.

$$
\begin{aligned}
&\sum_{s=0}^{t} \mathbb{E}\left\|X_s - \bar{X}_s\right\|_F^2 \\
&\leq \frac{8\lambda^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^3} \mathbb{E}\left\|V_0 - \bar{V}_0\right\|_F^2 + \frac{192\lambda^4 L^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^4} \sum_{s=0}^{t-1} \left(\mathbb{E}\left\|X_s - \bar{X}_s\right\|_F^2 + \mathbb{E}\left\|Y_s - \bar{Y}_s\right\|_F^2\right) \\
&\quad + \frac{96m\lambda^4 L^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^4} \sum_{s=0}^{t-2} \left(\eta_{\mathbf{x}}^2 \mathbb{E}\left\|\bar{\mathbf{v}}_s\right\|^2 + \eta_{\mathbf{y}}^2 \mathbb{E}\left\|\bar{\mathbf{u}}_s\right\|^2\right) + \frac{32m\lambda^4\beta_{\mathbf{x}}\eta_{\mathbf{x}}^2\sigma^2}{\left(1-\lambda^2\right)^4 b_0} + \frac{64m\lambda^4\beta_{\mathbf{x}}^3\eta_{\mathbf{x}}^2\sigma^2(t-2)}{\left(1-\lambda^2\right)^4} \\
&\quad + \frac{32m\lambda^4\beta_{\mathbf{x}}^2\eta_{\mathbf{x}}^2\sigma^2(t-1)}{\left(1-\lambda^2\right)^3} + \frac{384\lambda^4\beta_{\mathbf{x}} L^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^4} \sum_{s=0}^{t-2} \left(\mathbb{E}\left\|X_s - \bar{X}_s\right\|_F^2 + \mathbb{E}\left\|Y_s - \bar{Y}_s\right\|_F^2\right) \\
&\quad + \frac{192m\lambda^4\beta_{\mathbf{x}} L^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^4} \sum_{s=0}^{t-3} \left(\eta_{\mathbf{x}}^2 \mathbb{E}\left\|\bar{\mathbf{v}}_s\right\|^2 + \eta_{\mathbf{y}}^2 \mathbb{E}\left\|\bar{\mathbf{u}}_s\right\|^2\right) + \frac{32m\lambda^4\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^3} \sum_{s=0}^{t-1} \mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s} - \mathcal{N}_{\mathbf{x},s-1}\right\|^2 \\
&\leq \frac{8\lambda^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^3} \mathbb{E}\left\|V_0 - \bar{V}_0\right\|_F^2 + \frac{576\lambda^4 L^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^4} \sum_{s=0}^{t-1} \left(\mathbb{E}\left\|X_s - \bar{X}_s\right\|_F^2 + \mathbb{E}\left\|Y_s - \bar{Y}_s\right\|_F^2\right) \\
&\quad + \frac{288m\lambda^4 L^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^4} \sum_{s=0}^{t-2} \left(\eta_{\mathbf{x}}^2 \mathbb{E}\left\|\bar{\mathbf{v}}_s\right\|^2 + \eta_{\mathbf{y}}^2 \mathbb{E}\left\|\bar{\mathbf{u}}_s\right\|^2\right) + \frac{32m\lambda^4\beta_{\mathbf{x}}\eta_{\mathbf{x}}^2\sigma^2}{\left(1-\lambda^2\right)^4 b_0} + \frac{96m\lambda^4\beta_{\mathbf{x}}^2\eta_{\mathbf{x}}^2\sigma^2 t}{\left(1-\lambda^2\right)^4} \\
&\quad + \frac{32m\lambda^4\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^3} \sum_{s=0}^{t-1} \mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s} - \mathcal{N}_{\mathbf{x},s-1}\right\|^2
\end{aligned}
\tag{74}
$$

We use the condition $\beta_{\mathbf{x}} \leq 1$ in the inequality substitutions to simplify the expressions. Similarly, we can get:

$$
\begin{aligned}
&\sum_{s=0}^{t} \mathbb{E}\left\|Y_s - \bar{Y}_s\right\|_F^2 \\
&\leq \frac{8\lambda^2\eta_{\mathbf{y}}^2}{\left(1-\lambda^2\right)^3} \mathbb{E}\left\|U_0 - \bar{U}_0\right\|_F^2 + \frac{576\lambda^4 L^2\eta_{\mathbf{y}}^2}{\left(1-\lambda^2\right)^4} \sum_{s=0}^{t-1} \left(\mathbb{E}\left\|X_s - \bar{X}_s\right\|_F^2 + \mathbb{E}\left\|Y_s - \bar{Y}_s\right\|_F^2\right) \\
&\quad + \frac{288m\lambda^4 L^2\eta_{\mathbf{y}}^2}{\left(1-\lambda^2\right)^4} \sum_{s=0}^{t-2} \left(\eta_{\mathbf{x}}^2 \mathbb{E}\left\|\bar{\mathbf{v}}_s\right\|^2 + \eta_{\mathbf{y}}^2 \mathbb{E}\left\|\bar{\mathbf{u}}_s\right\|^2\right) + \frac{32m\lambda^4\beta_{\mathbf{y}}\eta_{\mathbf{y}}^2\sigma^2}{\left(1-\lambda^2\right)^4 b_0} + \frac{96m\lambda^4\beta_{\mathbf{y}}^2\eta_{\mathbf{y}}^2\sigma^2 t}{\left(1-\lambda^2\right)^4} \\
&\quad + \frac{32m\lambda^4\eta_{\mathbf{y}}^2}{\left(1-\lambda^2\right)^3} \sum_{s=0}^{t-1} \mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s} - \mathcal{N}_{\mathbf{y},s-1}\right\|^2
\end{aligned}
\tag{75}
$$

Add Eq.(74) and (75), we obtain:

$$\begin{aligned}\sum_{s=0}^{t}\left(\mathbb{E}\left\|X_s-\bar{X}_s\right\|_F^2+\mathbb{E}\left\|Y_s-\bar{Y}_s\right\|_F^2\right)\leq&\frac{8\lambda^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^3}\mathbb{E}\left\|V_0-\bar{V}_0\right\|_F^2+\frac{8\lambda^2\eta_{\mathbf{y}}^2}{\left(1-\lambda^2\right)^3}\mathbb{E}\left\|U_0-\bar{U}_0\right\|_F^2\\&+\frac{576\lambda^4L^2\left(\eta_{\mathbf{x}}^2+\eta_{\mathbf{y}}^2\right)}{\left(1-\lambda^2\right)^4}\sum_{s=0}^{t-1}\left(\mathbb{E}\left\|X_s-\bar{X}_s\right\|_F^2+\mathbb{E}\left\|Y_s-\bar{Y}_s\right\|_F^2\right)+\frac{32m\lambda^4\left(\beta_{\mathbf{x}}\eta_{\mathbf{x}}^2+\beta_{\mathbf{y}}\eta_{\mathbf{y}}^2\right)\sigma^2}{\left(1-\lambda^2\right)^4b_0}\\&+\frac{288m\lambda^4L^2\left(\eta_{\mathbf{x}}^2+\eta_{\mathbf{y}}^2\right)}{\left(1-\lambda^2\right)^4}\sum_{s=0}^{t-2}\left(\eta_{\mathbf{x}}^2\mathbb{E}\left\|\bar{\mathbf{v}}_s\right\|^2+\eta_{\mathbf{y}}^2\mathbb{E}\left\|\bar{\mathbf{u}}_s\right\|^2\right)+\frac{96m\lambda^4\left(\beta_{\mathbf{x}}^2\eta_{\mathbf{x}}^2+\beta_{\mathbf{y}}^2\eta_{\mathbf{y}}^2\right)\sigma^2t}{\left(1-\lambda^2\right)^4}\\&+\frac{32m\lambda^4\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^3}\sum_{s=0}^{t-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s}-\mathcal{N}_{\mathbf{x},s-1}\right\|^2+\frac{32m\lambda^4\eta_{\mathbf{y}}^2}{\left(1-\lambda^2\right)^3}\mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s}-\mathcal{N}_{\mathbf{y},s-1}\right\|^2\end{aligned}\tag{76}$$

As $\lambda<1$, when $\eta_{\mathbf{x}}\leq\frac{(1-\lambda)^2}{500L}$ and $\eta_{\mathbf{y}}\leq\frac{(1-\lambda)^2}{500L}$, it holds that $\frac{576\lambda^4L^2\left(\eta_{\mathbf{x}}^2+\eta_{\mathbf{y}}^2\right)}{\left(1-\lambda^2\right)^4}\leq\frac{1}{2}$, thus, we can obtain:

$$\begin{aligned}&\sum_{s=0}^{t}\left(\mathbb{E}\left\|X_s-\bar{X}_s\right\|_F^2+\mathbb{E}\left\|Y_s-\bar{Y}_s\right\|_F^2\right)\\&\leq\frac{16\lambda^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^3}\mathbb{E}\left\|V_0-\bar{V}_0\right\|_F^2+\frac{16\lambda^2\eta_{\mathbf{y}}^2}{\left(1-\lambda^2\right)^3}\mathbb{E}\left\|U_0-\bar{U}_0\right\|_F^2+\frac{576n\lambda^4L^2\left(\eta_{\mathbf{x}}^2+\eta_{\mathbf{y}}^2\right)}{\left(1-\lambda^2\right)^4}\sum_{s=0}^{t-2}\left(\eta_{\mathbf{x}}^2\mathbb{E}\left\|\bar{\mathbf{v}}_s\right\|^2\right.\\&\left.+\eta_{\mathbf{y}}^2\mathbb{E}\left\|\bar{\mathbf{u}}_s\right\|^2\right)+\frac{64m\lambda^4\left(\beta_{\mathbf{x}}\eta_{\mathbf{x}}^2+\beta_{\mathbf{y}}\eta_{\mathbf{y}}^2\right)\sigma^2}{\left(1-\lambda^2\right)^4b_0}+\frac{192m\lambda^4\left(\beta_{\mathbf{x}}^2\eta_{\mathbf{x}}^2+\beta_{\mathbf{y}}^2\eta_{\mathbf{y}}^2\right)\sigma^2t}{\left(1-\lambda^2\right)^4}\\&+\frac{64m\lambda^4\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^3}\sum_{s=0}^{t-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s}-\mathcal{N}_{\mathbf{x},s-1}\right\|^2+\frac{64m\lambda^4\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^3}\mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s}-\mathcal{N}_{\mathbf{y},s-1}\right\|^2\end{aligned}\tag{77}$$

A.2 Proof for main Theorems

Here, we firstly prove the first equation in Theorem 1, we set

$$\beta_{\mathbf{x}}=\frac{\epsilon\min\{1,m\epsilon\}}{20},\quad T=\frac{1500\kappa^3}{(1-\lambda)^2\epsilon\beta_{\mathbf{x}}}\tag{78}$$

Since $\Phi(x)$ is $(\kappa L+L)$-smooth we have:

$$\begin{aligned}\Phi\left(\bar{\mathbf{x}}_t\right)\leq&\Phi\left(\bar{\mathbf{x}}_{t-1}\right)-\eta_{\mathbf{x}}\left\langle\bar{\mathbf{v}}_{t-1},\nabla\Phi\left(\bar{\mathbf{x}}_{t-1}\right)\right\rangle+\eta_{\mathbf{x}}^2\kappa L\left\|\bar{\mathbf{v}}_{t-1}\right\|^2\\=&\Phi\left(\bar{\mathbf{x}}_{t-1}\right)-\frac{\eta_{\mathbf{x}}}{2}\left\|\bar{\mathbf{v}}_{t-1}\right\|^2-\frac{\eta_{\mathbf{x}}}{2}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_{t-1}\right)\right\|^2+\frac{\eta_{\mathbf{x}}}{2}\left\|\bar{\mathbf{v}}_{t-1}-\nabla\Phi\left(\bar{\mathbf{x}}_{t-1}\right)\right\|^2+\eta_{\mathbf{x}}^2\kappa L\left\|\bar{\mathbf{v}}_{t-1}\right\|^2\end{aligned}\tag{79}$$

Then we use Cauchy-Schwartz on above equation, we have:

$$\begin{aligned}\Phi\left(\bar{\mathbf{x}}_t\right)\leq&\Phi\left(\bar{\mathbf{x}}_{t-1}\right)-\frac{\eta_{\mathbf{x}}}{2}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_{t-1}\right)\right\|^2-\left(\frac{\eta_{\mathbf{x}}}{2}-\eta_{\mathbf{x}}^2\kappa L\right)\left\|\bar{\mathbf{v}}_{t-1}\right\|^2+\eta_{\mathbf{x}}\left\|\bar{\mathbf{v}}_{t-1}-\nabla_{\mathbf{x}}f\left(\bar{\mathbf{x}}_{t-1},\bar{\mathbf{y}}_{t-1}\right)\right\|^2\\&+\eta_{\mathbf{x}}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_{t-1}\right)-\nabla_{\mathbf{x}}f\left(\bar{\mathbf{x}}_{t-1},\bar{\mathbf{y}}_{t-1}\right)\right\|^2\end{aligned}\tag{80}$$

Because $\nabla\Phi\left(\bar{\mathbf{x}}_{t-1}\right)=\nabla_{\mathbf{x}}f\left(\bar{\mathbf{x}}_{t-1},\hat{\mathbf{y}}_{t-1}\right)$, according to Assumption 1, the last term satisfies:

$$\left\|\nabla\Phi\left(\bar{\mathbf{x}}_{t-1}\right)-\nabla_{\mathbf{x}}f\left(\bar{\mathbf{x}}_{t-1},\bar{\mathbf{y}}_{t-1}\right)\right\|^2\leq L^2\left\|\hat{\mathbf{y}}_{t-1}-\bar{\mathbf{y}}_{t-1}\right\|^2=L^2\delta_{t-1}\tag{81}$$

Additionally, using Cauchy-Schwartz inequality and Assumption 1 we have:

$$\begin{aligned}&\left\|\bar{\mathbf{v}}_{t-1}-\nabla_{\mathbf{x}}f\left(\bar{\mathbf{x}}_{t-1},\bar{\mathbf{y}}_{t-1}\right)\right\|^2\\&\leq2\left\|\bar{\mathbf{v}}_{t-1}-\frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)\right\|^2+2\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)-\nabla_{\mathbf{x}}f\left(\bar{\mathbf{x}}_{t-1},\bar{\mathbf{y}}_{t-1}\right)\right\|^2\\&\leq2\left\|\bar{\mathbf{v}}_{t-1}-\frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)\right\|^2+2\left\|\sum_{i=1}^{m}\frac{1}{m}\left(\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)-\nabla_{\mathbf{x}}f\left(\bar{\mathbf{x}}_{t-1},\bar{\mathbf{y}}_{t-1}\right)\right)\right\|^2\\&\leq2\left\|\bar{\mathbf{v}}_{t-1}-\frac{1}{m}\sum_{i=1}^{m}\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)\right\|^2+\frac{2L^2}{m}\left(\left\|X_{t-1}-\bar{X}_{t-1}\right\|_F^2+\left\|Y_{t-1}-\bar{Y}_{t-1}\right\|_F^2\right)\end{aligned}\tag{82}$$

Combine Eq.(80) (81) and (82) and we can get the inequality:

$$\begin{aligned}
&\left\|\nabla\Phi\left(\bar{\mathbf{x}}_{t-1}\right)\right\|^2 \\
&\leq \frac{2\left(\Phi\left(\bar{\mathbf{x}}_{t-1}\right)-\Phi\left(\bar{\mathbf{x}}_t\right)\right)}{\eta_{\mathbf{x}}}-\left(1-2\kappa L\eta_{\mathbf{x}}\right)\left\|\bar{\mathbf{v}}_{t-1}\right\|^2+2L^2\delta_{t-1}+\frac{4L^2}{m}\left(\left\|X_{t-1}-\bar{X}_{t-1}\right\|_F^2\right. \\
&\quad\left.+\left\|Y_{t-1}-\bar{Y}_{t-1}\right\|_F^2\right)+4\left\|\bar{\mathbf{v}}_{t-1}-\frac{1}{m}\sum_{i=1}^m\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_{t-1}^{(i)},\mathbf{y}_{t-1}^{(i)}\right)\right\|^2
\end{aligned} \tag{83}$$

Telescoping and taking expectation on Eq.(83) we have:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_t\right)\right\|^2 \\
&\leq \frac{2\left(\Phi\left(\mathbf{x}_0\right)-\mathbb{E}\Phi\left(\bar{\mathbf{x}}_t\right)\right)}{\eta_{\mathbf{x}}T}-\frac{\left(1-2\kappa L\eta_{\mathbf{x}}\right)}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\bar{\mathbf{v}}_t\right\|^2+\frac{2L^2}{T}\sum_{t=0}^{T-1}\mathbb{E}\delta_t \\
&\quad+\frac{4L^2}{mT}\sum_{t=0}^{T-1}\left(\mathbb{E}\left\|X_t-\bar{X}_t\right\|_F^2+\mathbb{E}\left\|Y_t-\bar{Y}_t\right\|_F^2\right)+\frac{4}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\bar{\mathbf{v}}_t-\frac{1}{m}\sum_{i=1}^m\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right\|^2
\end{aligned} \tag{84}$$

Using Lemma 6 to replace

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_t\right)\right\|^2 \\
&\leq \frac{2\left(\Phi\left(\mathbf{x}_0\right)-\Phi^*\right)}{\eta_{\mathbf{x}}T}-\left(1-2\kappa L\eta_{\mathbf{x}}-\frac{40\kappa^4\eta_{\mathbf{x}}^2}{\eta_{\mathbf{y}}^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\bar{\mathbf{v}}_t\right\|^2+\frac{8\kappa L^2\delta_0}{TL\eta_{\mathbf{y}}} \\
&\quad-\frac{28\kappa L\eta_{\mathbf{y}}}{5T}\sum_{t=0}^{T-1}\left(1-\left(1-\frac{\mu\eta_{\mathbf{y}}}{4}\right)^{T-t}\right)\mathbb{E}\left\|\bar{\mathbf{u}}_t\right\|^2+\frac{84\kappa^2L^2}{mT}\sum_{t=0}^{T-1}\left(\mathbb{E}\left\|X_t-\bar{X}_t\right\|_F^2\right. \\
&\quad\left.+\mathbb{E}\left\|Y_t-\bar{Y}_t\right\|_F^2\right)+\frac{4}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\bar{\mathbf{v}}_t-\frac{1}{m}\sum_{i=1}^m\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)}\right)\right\|^2 \\
&\quad+\frac{20\kappa L\eta_{\mathbf{y}}}{T}\sum_{t=1}^{T-1}\left(1-\frac{\mu\eta_{\mathbf{y}}}{4}\right)^{T-t-1}\sum_{s=0}^{t-1}\mathbb{E}\left\|\bar{\mathbf{u}}_t-\frac{1}{m}\sum_{i=1}^m\nabla f_i\left(\mathbf{x}_s^{(i)},\mathbf{y}_s^{(i)}\right)\right\|^2
\end{aligned} \tag{85}$$

And using Lemma 9 to replace the last two terms.

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_t\right)\right\|^2 \\
&\leq \frac{2\left(\Phi\left(\mathbf{x}_0\right)-\Phi^*\right)}{\eta_{\mathbf{x}}T}-\left(1-2\kappa L\eta_{\mathbf{x}}-\frac{40\kappa^4\eta_{\mathbf{x}}^2}{\eta_{\mathbf{y}}^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\bar{\mathbf{v}}_t\right\|^2+\frac{8\kappa L^2\delta_0}{TL\eta_{\mathbf{y}}}+\frac{8\sigma^2}{mb_0T}\left(\frac{1}{\beta_{\mathbf{x}}}+\frac{20\kappa^2}{\beta_{\mathbf{y}}}\right) \\
&\quad+\frac{8\sigma^2}{m}\left(\beta_{\mathbf{x}}+20\kappa^2\beta_{\mathbf{y}}\right)+\frac{4L^2}{mT}\left(21\kappa^2+\frac{12}{m\beta_{\mathbf{x}}}+\frac{240\kappa^2}{m\beta_{\mathbf{y}}}\right)\sum_{t=0}^{T-1}\left(\mathbb{E}\left\|X_t-\bar{X}_t\right\|_F^2+\mathbb{E}\left\|Y_t-\bar{Y}_t\right\|_F^2\right) \\
&\quad+\frac{24L^2}{m\beta_{\mathbf{x}}T}\sum_{t=0}^{T-1}\left(1-\left(1-\beta_{\mathbf{x}}\right)^{T-t}\right)\left(\eta_{\mathbf{x}}^2\mathbb{E}\left\|\bar{\mathbf{v}}_t\right\|^2+\eta_{\mathbf{y}}^2\mathbb{E}\left\|\bar{\mathbf{u}}_t\right\|^2\right)+\frac{480\kappa^2L^2}{m\beta_{\mathbf{y}}T}\sum_{t=0}^{T-1} \\
&\quad\left(1-\left(1-\frac{\mu\eta_{\mathbf{y}}}{4}\right)^{T-t}\right)\left(\eta_{\mathbf{x}}^2\mathbb{E}\left\|\bar{\mathbf{v}}_t\right\|^2+\eta_{\mathbf{y}}^2\mathbb{E}\left\|\bar{\mathbf{u}}_t\right\|^2\right)-\frac{28\kappa L\eta_{\mathbf{y}}}{5T}\sum_{t=0}^{T-1}\left(1-\left(1-\frac{\mu\eta_{\mathbf{y}}}{4}\right)^{T-t}\right)\mathbb{E}\left\|\bar{\mathbf{u}}_t\right\|^2 \\
&\quad+\frac{8}{T}\sum_{s=0}^{T-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s}-\left(1-\beta_{\mathbf{x}}\right)\mathcal{N}_{\mathbf{x},s-1}\right\|^2+\frac{40\kappa L\eta_{\mathbf{y}}}{T}\sum_{t=1}^{T-1}\left(1-\frac{\mu\eta_{\mathbf{y}}}{4}\right)^{T-t-1}\sum_{s=0}^{t-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s}-\left(1-\beta_{\mathbf{y}}\right)\mathcal{N}_{\mathbf{y},s-1}\right\|^2
\end{aligned} \tag{86}$$

For the sum of $1-\beta_{\mathbf{x}}$

$$\frac{1}{\beta_{\mathbf{x}}}\left(1-\left(1-\beta_{\mathbf{x}}\right)^{T-t}\right)=\sum_{s=0}^{T-t-1}\left(1-\beta_{\mathbf{x}}\right)^s \tag{87}$$

we know Eq.(87) is increasing when $\beta_{\mathbf{x}}$ is decreasing.
Hence $\frac{1}{\beta_{\mathbf{x}}}\left(1-(1-\beta_{\mathbf{x}})^{T-t}\right) \leq \frac{300\kappa^2}{(1-\lambda)^2\beta_{\mathbf{x}}}\left(1-(1-\frac{(1-\lambda)^2\beta_{\mathbf{x}}}{300\kappa^2})^{T-t}\right)$. According to the definition of $\beta_{\mathbf{x}}$ and $\eta_{\mathbf{y}}$, we have $\frac{(1-\lambda)^2\beta_{\mathbf{x}}}{300\kappa^2} \leq \frac{\mu\eta_{\mathbf{y}}}{4}$ and $\frac{24L^2}{m\beta_{\mathbf{x}}T}\left(1-(1-\beta_{\mathbf{x}})^{T-t}\right) \leq \frac{7200L^2\kappa^2}{m(1-\lambda)^2\beta_{\mathbf{x}}T}\left(1-\left(1-\frac{\mu\eta_{\mathbf{y}}}{4}\right)^{T-t}\right)$. Therefore, using the definition of $\beta_{\mathbf{x}}, \beta_{\mathbf{y}}$ and $\eta_{\mathbf{y}}$ we obtain:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_t\right)\right\|^2 \\
&\leq \frac{2\left(\Phi\left(\mathbf{x}_0\right)-\Phi^*\right)}{\eta_{\mathbf{x}}T} - \left(1-2\kappa L\eta_{\mathbf{x}}-\frac{40\kappa^4\eta_{\mathbf{x}}^2}{\eta_{\mathbf{y}}^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\bar{\mathbf{v}}_t\right\|^2 + \frac{8\kappa L^2\delta_0}{TL\eta_{\mathbf{y}}} + \frac{8\sigma^2}{mb_0T}\left(\frac{1}{\beta_{\mathbf{x}}}+\frac{20\kappa^2}{\beta_{\mathbf{y}}}\right) \\
&\quad + \frac{8\sigma^2}{m}\left(\beta_{\mathbf{x}}+20\kappa^2\beta_{\mathbf{y}}\right) + \frac{4L^2}{mT}\left(21\kappa^2+\frac{12}{m\beta_{\mathbf{x}}}+\frac{240\kappa^2}{m\beta_{\mathbf{y}}}\right)\sum_{t=0}^{T-1}\left(\mathbb{E}\left\|X_t-\bar{X}_t\right\|_F^2+\mathbb{E}\left\|Y_t-\bar{Y}_t\right\|_F^2\right) \\
&\quad + \left(\frac{24L^2\eta_{\mathbf{x}}^2}{m\beta_{\mathbf{x}}}+\frac{480\kappa^2L^2\eta_{\mathbf{x}}^2}{m\beta_{\mathbf{y}}}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\bar{\mathbf{v}}_t\right\|^2 - \frac{\kappa L\eta_{\mathbf{y}}}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\bar{\mathbf{u}}_t\right\|^2 + \frac{8}{T}\sum_{s=0}^{T-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s}-\left(1-\beta_{\mathbf{x}}\right)\mathcal{N}_{\mathbf{x},s-1}\right\|^2 \\
&\quad + \frac{160\kappa^2}{T}\sum_{s=0}^{T-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s}-\left(1-\beta_{\mathbf{y}}\right)\mathcal{N}_{\mathbf{y},s-1}\right\|^2
\end{aligned}
\tag{88}
$$

Using Lemma 13 to replace $\mathbb{E}\left\|X_s-\bar{X}_s\right\|_F^2+\mathbb{E}\left\|Y_s-\bar{Y}_s\right\|_F^2$

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_t\right)\right\|^2 \\
&\leq \frac{2\left(\Phi\left(\mathbf{x}_0\right)-\Phi^*\right)}{\eta_{\mathbf{x}}T} - \left(1-2\kappa L\eta_{\mathbf{x}}-\frac{40\kappa^4\eta_{\mathbf{x}}^2}{\eta_{\mathbf{y}}^2}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\bar{\mathbf{v}}_t\right\|^2 + \frac{8\kappa L^2\delta_0}{TL\eta_{\mathbf{y}}} + \frac{8\sigma^2}{mb_0T}\left(\frac{1}{\beta_{\mathbf{x}}}+\frac{20\kappa^2}{\beta_{\mathbf{y}}}\right) \\
&\quad + \frac{8\sigma^2}{m}\left(\beta_{\mathbf{x}}+20\kappa^2\beta_{\mathbf{y}}\right) + \frac{4L^2}{mT}\left(21\kappa^2+\frac{12}{m\beta_{\mathbf{x}}}+\frac{240\kappa^2}{m\beta_{\mathbf{y}}}\right)\left(\frac{16\lambda^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^3}\mathbb{E}\left\|V_0-\bar{V}_0\right\|_F^2\right. \\
&\quad \left. + \frac{16\lambda^2\eta_{\mathbf{y}}^2}{\left(1-\lambda^2\right)^3}\mathbb{E}\left\|U_0-\bar{U}_0\right\|_F^2 + \frac{64m\lambda^4\left(\beta_{\mathbf{x}}\eta_{\mathbf{x}}^2+\beta_{\mathbf{y}}\eta_{\mathbf{y}}^2\right)\sigma^2}{\left(1-\lambda^2\right)^4 b_0} + \frac{192m\lambda^4\left(\beta_{\mathbf{x}}^2\eta_{\mathbf{x}}^2+\beta_{\mathbf{y}}^2\eta_{\mathbf{y}}^2\right)\sigma^2T}{\left(1-\lambda^2\right)^4}\right) \\
&\quad + \frac{4L^2}{mT}\left(21\kappa^2+\frac{12}{m\beta_{\mathbf{x}}}+\frac{240\kappa^2}{m\beta_{\mathbf{y}}}\right)\frac{576m\lambda^4L^2\left(\eta_{\mathbf{x}}^2+\eta_{\mathbf{y}}^2\right)}{\left(1-\lambda^2\right)^4}\sum_{t=0}^{T-1}\left(\eta_{\mathbf{x}}^2\mathbb{E}\left\|\bar{\mathbf{v}}_t\right\|^2+\eta_{\mathbf{y}}^2\mathbb{E}\left\|\bar{\mathbf{u}}_t\right\|^2\right) \\
&\quad + \left(\frac{24L^2\eta_{\mathbf{x}}^2}{m\beta_{\mathbf{x}}}+\frac{480\kappa^2L^2\eta_{\mathbf{x}}^2}{m\beta_{\mathbf{y}}}\right)\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\bar{\mathbf{v}}_t\right\|^2 - \frac{\kappa L\eta_{\mathbf{y}}}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\bar{\mathbf{u}}_t\right\|^2 + \frac{8}{T}\sum_{s=0}^{T-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s}-\left(1-\beta_{\mathbf{x}}\right)\mathcal{N}_{\mathbf{x},s-1}\right\|^2 \\
&\quad + \frac{160\kappa^2}{T}\sum_{s=0}^{T-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s}-\left(1-\beta_{\mathbf{y}}\right)\mathcal{N}_{\mathbf{y},s-1}\right\|^2 + \frac{256L^2\lambda^4\eta_{\mathbf{x}}^2}{T\left(1-\lambda^2\right)^3}\left(21\kappa^2+\frac{12}{m\beta_{\mathbf{x}}}+\frac{240\kappa^2}{m\beta_{\mathbf{y}}}\right) \\
&\quad \sum_{s=0}^{T-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s}-\mathcal{N}_{\mathbf{x},s-1}\right\|^2 + \frac{256L^2\lambda^4\eta_{\mathbf{y}}^2}{T\left(1-\lambda^2\right)^3}\left(21\kappa^2+\frac{12}{m\beta_{\mathbf{x}}}+\frac{240\kappa^2}{m\beta_{\mathbf{y}}}\right)\sum_{s=0}^{T-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s}-\mathcal{N}_{\mathbf{y},s-1}\right\|^2
\end{aligned}
\tag{89}
$$

When $\beta_{\mathbf{x}}, \beta_{\mathbf{y}}, \eta_{\mathbf{x}}$ and $\eta_{\mathbf{y}}$ are defined as Theorem 1 , we have

$$
\frac{4L^2}{mT}\left(21\kappa^2+\frac{12}{m\beta_{\mathbf{x}}}+\frac{240\kappa^2}{m\beta_{\mathbf{y}}}\right)\frac{576m\lambda^4L^2\left(\eta_{\mathbf{x}}^2+\eta_{\mathbf{y}}^2\right)}{\left(1-\lambda^2\right)^4}\eta_{\mathbf{y}}^2 \leq \frac{\kappa L\eta_{\mathbf{y}}}{2T}
\tag{90}
$$

and

$$
\begin{aligned}
&1-2\kappa L\eta_{\mathbf{x}}-\frac{40\kappa^4\eta_{\mathbf{x}}^2}{\eta_{\mathbf{y}}^2}-\frac{24L^2\eta_{\mathbf{x}}^2}{m\beta_{\mathbf{x}}}-\frac{480\kappa^2L^2\eta_{\mathbf{x}}^2}{m\beta_{\mathbf{y}}} \\
&-\frac{4L^2}{m}\left(21\kappa^2+\frac{12}{m\beta_{\mathbf{x}}}+\frac{240\kappa^2}{m\beta_{\mathbf{y}}}\right)\frac{576m\lambda^4L^2\left(\eta_{\mathbf{x}}^2+\eta_{\mathbf{y}}^2\right)}{\left(1-\lambda^2\right)^4}\eta_{\mathbf{x}}^2 \geq \frac{2}{5}
\end{aligned}
\tag{91}
$$

Therefore, subtracting the terms containing these two quantities will not affect the validity of the inequality. This simplification is achieved by using this scaling method.

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_t\right)\right\|^2 \\
&\leq \frac{2\left(\Phi\left(\mathbf{x}_0\right)-\Phi^*\right)}{\eta_{\mathbf{x}}T}+\frac{8\kappa L^2\delta_0}{TL\eta_{\mathbf{y}}}+\frac{8\sigma^2}{mb_0T}\left(\frac{1}{\beta_{\mathbf{x}}}+\frac{20\kappa^2}{\beta_{\mathbf{y}}}\right)+\frac{8\sigma^2}{m}\left(\beta_{\mathbf{x}}+20\kappa^2\beta_{\mathbf{y}}\right) \\
&\quad+\frac{4L^2}{mT}\left(21\kappa^2+\frac{12}{m\beta_{\mathbf{x}}}+\frac{240\kappa^2}{m\beta_{\mathbf{y}}}\right)\left(\frac{16\lambda^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^3}\mathbb{E}\left\|V_0-\bar{V}_0\right\|_F^2+\frac{16\lambda^2\eta_{\mathbf{y}}^2}{\left(1-\lambda^2\right)^3}\mathbb{E}\left\|U_0-\bar{U}_0\right\|_F^2\right. \\
&\quad\left.+\frac{64m\lambda^4\left(\beta_{\mathbf{x}}\eta_{\mathbf{x}}^2+\beta_{\mathbf{y}}\eta_{\mathbf{y}}^2\right)\sigma^2}{\left(1-\lambda^2\right)^4b_0}+\frac{192m\lambda^4\left(\beta_{\mathbf{x}}^2\eta_{\mathbf{x}}^2+\beta_{\mathbf{y}}^2\eta_{\mathbf{y}}^2\right)\sigma^2T}{\left(1-\lambda^2\right)^4}\right)+\frac{8}{T}\sum_{s=0}^{T-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s}-\left(1-\beta_{\mathbf{x}}\right)\mathcal{N}_{\mathbf{x},s-1}\right\|^2 \\
&\quad+\frac{160\kappa^2}{T}\sum_{s=0}^{T-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s}-\left(1-\beta_{\mathbf{y}}\right)\mathcal{N}_{\mathbf{y},s-1}\right\|^2+\frac{256L^2\lambda^4\eta_{\mathbf{x}}^2}{T\left(1-\lambda^2\right)^3}\left(21\kappa^2+\frac{12}{m\beta_{\mathbf{x}}}+\frac{240\kappa^2}{m\beta_{\mathbf{y}}}\right)\sum_{s=0}^{T-1} \\
&\quad\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s}-\mathcal{N}_{\mathbf{x},s-1}\right\|^2+\frac{256L^2\lambda^4\eta_{\mathbf{y}}^2}{T\left(1-\lambda^2\right)^3}\left(21\kappa^2+\frac{12}{m\beta_{\mathbf{x}}}+\frac{240\kappa^2}{m\beta_{\mathbf{y}}}\right)\sum_{s=0}^{T-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s}-\mathcal{N}_{\mathbf{y},s-1}\right\|^2
\end{aligned} \tag{92}$$

By Assumption 4 and Cauchy-Schwartz inequality we also have

$$\mathbb{E}\left\|V_0-\bar{V}_0\right\|_F^2=\mathbb{E}\left\|G_0(W-J)\right\|_F^2\leq\lambda^2\mathbb{E}\left\|G_0\right\|_F^2\leq\frac{2m\lambda^2\sigma^2}{b_0}+2\lambda^2\sum_{i=1}^{m}\left\|\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_0,\mathbf{y}_0\right)\right\|^2 \tag{93}$$

Similarly, we have

$$\mathbb{E}\left\|U_0-\bar{U}_0\right\|_F^2\leq\frac{2m\lambda^2\sigma^2}{b_0}+2\lambda^2\sum_{i=1}^{m}\left\|\nabla_{\mathbf{y}}f_i\left(\mathbf{x}_0,\mathbf{y}_0\right)\right\|^2 \tag{94}$$

With the definition of $\beta_{\mathbf{x}}, \beta_{\mathbf{y}}, \eta_{\mathbf{x}}$ and $\eta_{\mathbf{y}}$ are given in Theorem 1, therefore we get:

$$\begin{aligned}
\frac{256L^2\lambda^4\eta_{\mathbf{x}}^2}{T\left(1-\lambda^2\right)^3}\left(21\kappa^2+\frac{12}{n\beta_{\mathbf{x}}}+\frac{240\kappa^2}{n\beta_{\mathbf{y}}}\right)&\leq\frac{\lambda^4(1-\lambda)}{T}\left(m^2\epsilon^2+26\kappa^2\right) \\
\frac{256L^2\lambda^4\eta_{\mathbf{y}}^2}{T\left(1-\lambda^2\right)^3}\left(21\kappa^2+\frac{12}{n\beta_{\mathbf{x}}}+\frac{240\kappa^2}{n\beta_{\mathbf{y}}}\right)&\leq\frac{\lambda^4(1-\lambda)}{T}\left(m^2\epsilon^2+26\kappa^2\right)
\end{aligned} \tag{95}$$

We know that the maximum of $\lambda^4(1-\lambda)$ is $\frac{256}{3075}<1$, meanwhile, by the definition of $\kappa=\frac{L}{\mu}\leq 1$ thus, we have:

$$\begin{aligned}
\frac{256L^2\lambda^4\eta_{\mathbf{x}}^2}{T\left(1-\lambda^2\right)^3}\left(21\kappa^2+\frac{12}{n\beta_{\mathbf{x}}}+\frac{240\kappa^2}{n\beta_{\mathbf{y}}}\right)&\leq\frac{\left(m^2\epsilon^2+3\kappa^2\right)}{T} \\
\frac{256L^2\lambda^4\eta_{\mathbf{y}}^2}{T\left(1-\lambda^2\right)^3\kappa^2}\left(21\kappa^2+\frac{12}{n\beta_{\mathbf{x}}}+\frac{240\kappa^2}{n\beta_{\mathbf{y}}}\right)&\leq\frac{\left(m^2\epsilon^2+3\right)}{T}
\end{aligned} \tag{96}$$

Combine above three inequalities and substitute the parameters with their definitions. We achieve

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_t\right)\right\|^2 \\
&\leq L\left(\Phi\left(\mathbf{x}_0\right)-\Phi^*\right)\epsilon^2+2L^2\delta_0\epsilon^2+2\sigma^2\epsilon^2+\frac{\epsilon^2}{m}\sum_{i=1}^{m}\left\|\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_0,\mathbf{y}_0\right)\right\|^2 \\
&\quad+\frac{\epsilon^2}{m}\sum_{i=1}^{m}\left\|\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_0,\mathbf{y}_0\right)\right\|^2+\frac{\left(m^2\epsilon^2+3\kappa^2\right)}{T}\sum_{s=0}^{T-1}\left(\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s}-\mathcal{N}_{\mathbf{x},s-1}\right\|^2+\mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s}-\mathcal{N}_{\mathbf{y},s-1}\right\|^2\right) \\
&\quad+\frac{8}{T}\sum_{s=0}^{T-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s}-\left(1-\beta_{\mathbf{x}}\right)\mathcal{N}_{\mathbf{x},s-1}\right\|^2+\frac{160\kappa^2}{T}\sum_{s=0}^{T-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s}-\left(1-\beta_{\mathbf{y}}\right)\mathcal{N}_{\mathbf{y},s-1}\right\|^2
\end{aligned} \tag{97}$$

Now, review the definition of $\mathcal{N}_{\mathbf{x},s}$ and $\mathcal{N}_{\mathbf{x},s}$, we can obtain that:

$$\mathcal{N}_{\mathbf{x},t} \sim \mathcal{N}\left(0, \frac{\sigma_{\mathbf{x}}^2}{m} I_{d_1}\right), \mathcal{N}_{y,t} \sim \mathcal{N}\left(0, \frac{\sigma_{\mathbf{y}}^2}{m} I_{d_2}\right) \tag{98}$$

$\mathcal{N}_{\mathbf{x},s}$ and $\mathcal{N}_{\mathbf{x},s-1}$ are independent normally distributed random variables because the noises $n_{\mathbf{x},s}^{(i)}$ and $n_{\mathbf{x},s-1}^{(i)}$ generated at times s and $s-1$ are independent. Therefore, the distribution of $\mathcal{N}_{\mathbf{x},s} - \mathcal{N}_{\mathbf{x},s-1}$ is also normally distributed, with mean 0 and a covariance matrix that is the sum of the covariances of the two independent normal distributions.

$$\mathcal{N}_{\mathbf{x},s} - \mathcal{N}_{\mathbf{x},s-1} \sim \mathcal{N}\left(0, 2\frac{\sigma_{\mathbf{x}}^2}{m} I_{d_1}\right) \tag{99}$$

Therefore:

$$\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s} - \mathcal{N}_{\mathbf{x},s-1}\right\|^2 = \mathrm{Tr}\left(2 \cdot \frac{\sigma_{\mathbf{x}}^2}{m} I_{d_1}\right) = 2\frac{\sigma_{\mathbf{x}}^2}{m} d_1 \tag{100}$$

Sum up from 0 to $T-1$:

$$\sum_{s=0}^{T-1} \mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s} - \mathcal{N}_{\mathbf{x},s-1}\right\|^2 = \sum_{s=0}^{T-1} 2 \cdot \frac{\sigma_{\mathbf{x}}^2}{m} \cdot d_1 = 2\frac{\sigma_{\mathbf{x}}^2}{m} d_1 T \tag{101}$$

Similarly, we can get:

$$\sum_{s=0}^{T-1} \mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s} - \mathcal{N}_{\mathbf{y},s-1}\right\|^2 = 2\frac{\sigma_{\mathbf{y}}^2}{m} d_2 T \tag{102}$$

Mimic the process above, we know that:

$$\mathcal{N}_{\mathbf{x},s} - (1-\beta_{\mathbf{x}})\,\mathcal{N}_{\mathbf{x},s-1} \sim \mathcal{N}\left(0, \frac{\sigma_{\mathbf{x}}^2}{m}\left(I_{d_1} + (1-\beta_{\mathbf{x}})^2 I_{d_1}\right)\right) \tag{103}$$

Therefore, we have:

$$\begin{aligned}
\sum_{s=0}^{T-1} \mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s} - (1-\beta_{\mathbf{x}})\,\mathcal{N}_{\mathbf{x},s-1}\right\|^2 &= T\frac{\sigma_{\mathbf{x}}^2}{m} d_1 \left(1 + (1-\beta_{\mathbf{x}})^2\right) \\
\sum_{s=0}^{T-1} \mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s} - \left(1-\beta_{\mathbf{y}}\right)\mathcal{N}_{\mathbf{y},s-1}\right\|^2 &= T\frac{\sigma_{\mathbf{y}}^2}{m} d_2 \left(1 + \left(1-\beta_{\mathbf{y}}\right)^2\right)
\end{aligned} \tag{104}$$

Therefore, Eq.(97) can be written as:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_t\right)\right\|^2 \leq & L\left(\Phi\left(\mathbf{x}_0\right) - \Phi^*\right)\epsilon^2 + 2L^2\delta_0\epsilon^2 + 2\sigma^2\epsilon^2 + \frac{\epsilon^2}{m}\sum_{i=1}^{m}\left\|\nabla_{\mathbf{x}} f_i\left(\mathbf{x}_0, \mathbf{y}_0\right)\right\|^2 \\
& + \frac{\epsilon^2}{m}\sum_{i=1}^{m}\left\|\nabla_{\mathbf{x}} f_i\left(\mathbf{x}_0, \mathbf{y}_0\right)\right\|^2 + \frac{\left(2m^2\epsilon^2 + 6\kappa^2\right)}{m}\left(\sigma_{\mathbf{x}}^2 d_1 + \sigma_{\mathbf{y}}^2 d_2\right) \\
& + \frac{16\sigma_{\mathbf{x}}^2 d_1}{m} + \frac{320\kappa^2\sigma_{\mathbf{y}}^2 d_2}{m}
\end{aligned} \tag{105}$$

where we use following inequalities for simplification.

$$\begin{aligned}
&\beta_{\mathbf{x}} \geq \beta_{\mathbf{y}}, 4L^2\left(21\kappa^2 + \frac{12}{m\beta_{\mathbf{x}}} + \frac{240\kappa^2}{m\beta_{\mathbf{y}}}\right) \leq 100L^2\kappa^2 + \frac{1000L^2\kappa^2}{m\beta_{\mathbf{y}}} \\
&\frac{L^2\beta_{\mathbf{x}}\eta_{\mathbf{x}}^2}{(1-\lambda)^4 b_0 T} \leq \frac{\epsilon(\min\{1, m\epsilon\})^5\epsilon^2}{20\cdot 400\kappa\cdot 30000\kappa^3\left(15000\kappa^3\right)^2}, \frac{L^2\beta_{\mathbf{y}}\eta_{\mathbf{y}}^2}{(1-\lambda)^4 b_0 T} \leq \frac{\epsilon(\min\{1, m\epsilon\})^5\epsilon^2}{500\cdot 400\kappa\cdot 30000\kappa^3(1500\kappa)^2} \\
&\frac{L^2\beta_{\mathbf{x}}^2\eta_{\mathbf{x}}^2}{(1-\lambda)^4} \leq \frac{\epsilon^2(\min\{1, m\epsilon\})^4}{400\left(15000\kappa^3\right)^2}, \frac{L^2\beta_{\mathbf{y}}^2\eta_{\mathbf{y}}^2}{(1-\lambda)^4} \leq \frac{\epsilon^2(\min\{1, m\epsilon\})^4}{\left(500\kappa^2\right)^2(1500\kappa)^2}
\end{aligned} \tag{106}$$

Therefore, if T is determined by ϵ, we have the first conclusion in Theorem 1:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_t\right)\right\|^2 = O\left(\epsilon^2\right) + O\left(m\epsilon^2\right) + O\left(\sigma_{\mathbf{x}}^2 d_1 + \sigma_{\mathbf{y}}^2 d_2\right) \tag{107}$$

In the above proof, we have established the convergence result when T is determined by ϵ. Next, we analyze the case when T is uncertain, for which we provide the following proof. Before presenting our proof, we first provide some definitions regarding T.

$$T_0 \geq 10m^2, \quad T = \frac{30000\kappa^3 T_0}{(1-\lambda)^2} \tag{108}$$

Similarly, we can obtain:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_t\right)\right\|^2 \\
&\leq \frac{2\left(\Phi\left(\mathbf{x}_0\right)-\Phi^*\right)}{\eta_{\mathbf{x}}T}+\frac{8\kappa L^2\delta_0}{TL\eta_{\mathbf{y}}}+\frac{8\sigma^2}{mb_0T}\left(\frac{1}{\beta_{\mathbf{x}}}+\frac{20\kappa^2}{\beta_{\mathbf{y}}}\right)+\frac{8\sigma^2}{m}\left(\beta_{\mathbf{x}}+20\kappa^2\beta_{\mathbf{y}}\right) \\
&\quad+\frac{4L^2}{mT}\left(21\kappa^2+\frac{12}{m\beta_{\mathbf{x}}}+\frac{240\kappa^2}{m\beta_{\mathbf{y}}}\right)\left(\frac{16\lambda^2\eta_{\mathbf{x}}^2}{\left(1-\lambda^2\right)^3}\mathbb{E}\left\|V_0-\bar{V}_0\right\|_F^2+\frac{16\lambda^2\eta_{\mathbf{y}}^2}{\left(1-\lambda^2\right)^3}\mathbb{E}\left\|U_0-\bar{U}_0\right\|_F^2\right. \\
&\quad\left.+\frac{64m\lambda^4\left(\beta_{\mathbf{x}}\eta_{\mathbf{x}}^2+\beta_{\mathbf{y}}\eta_{\mathbf{y}}^2\right)\sigma^2}{\left(1-\lambda^2\right)^4 b_0}+\frac{192m\lambda^4\left(\beta_{\mathbf{x}}^2\eta_{\mathbf{x}}^2+\beta_{\mathbf{y}}^2\eta_{\mathbf{y}}^2\right)\sigma^2 T}{\left(1-\lambda^2\right)^4}\right)+\frac{8}{T}\sum_{s=0}^{T-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s}-\mathcal{N}_{\mathbf{x},s-1}\right\|^2 \\
&\quad+\frac{160\kappa^2}{T}\sum_{s=0}^{T-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s}-\mathcal{N}_{\mathbf{y},s-1}\right\|^2+\frac{256L^2\lambda^4\eta_{\mathbf{x}}^2}{T\left(1-\lambda^2\right)^3}\left(21\kappa^2+\frac{12}{m\beta_{\mathbf{x}}}+\frac{240\kappa^2}{m\beta_{\mathbf{y}}}\right)\sum_{s=0}^{T-1} \\
&\quad\mathbb{E}\left\|\mathcal{N}_{\mathbf{x},s}-\mathcal{N}_{\mathbf{x},s-1}\right\|^2+\frac{256L^2\lambda^4\eta_{\mathbf{y}}^2}{T\left(1-\lambda^2\right)^3}\left(21\kappa^2+\frac{12}{m\beta_{\mathbf{x}}}+\frac{240\kappa^2}{m\beta_{\mathbf{y}}}\right)\sum_{s=0}^{T-1}\mathbb{E}\left\|\mathcal{N}_{\mathbf{y},s}-\mathcal{N}_{\mathbf{y},s-1}\right\|^2
\end{aligned} \tag{109}$$

Substituting the parameter values given in above Eq.(108) and the relationships between all these parameters in Theorem 1, and using the scaling method to simplify the calculations, we can get:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_t\right)\right\|^2 \leq&\frac{L\left(\Phi\left(\mathbf{x}_0\right)-\Phi^*\right)+2\sigma^2+2L^2\delta_0}{\left(mT_0\right)^{2/3}}+\frac{\frac{1}{m}\sum_{i=1}^m\mathbb{E}\left\|\nabla_{\mathbf{x}}f_i\left(\mathbf{x}_0,\mathbf{y}_0\right)\right\|^2}{T_0} \\
&+\frac{\frac{1}{m}\sum_{i=1}^m\mathbb{E}\left\|\nabla_{\mathbf{y}}f_i\left(\mathbf{x}_0,\mathbf{y}_0\right)\right\|^2}{T_0}+\left(\frac{2m^{2/3}}{T_0^{2/3}}+\frac{3\kappa^2}{m}\right)\left(\sigma_{\mathbf{x}}^2d_1+\sigma_{\mathbf{y}}^2d_2\right) \\
&+\frac{16\sigma_{\mathbf{x}}^2d_1}{m}+\frac{320\kappa^2\sigma_{\mathbf{y}}^2d_2}{m}
\end{aligned} \tag{110}$$

Therefore, if the number of iteration is not fixed, we have the second conclusion in Theorem 1, we have:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi\left(\bar{\mathbf{x}}_t\right)\right\|^2=O\left(\frac{1}{\left(mT_0\right)^{2/3}}\right)+O\left(\frac{1}{T_0}\right)+O\left(\frac{m^{1/3}}{T_0^{2/3}}\right)+O\left(\sigma_{\mathbf{x}}^2d_1+\sigma_{\mathbf{y}}^2d_2\right) \tag{111}$$

B Proof of privacy guarantee

Before we start our privacy analysis, let's introduce moments accountant method [1].

Definition 2. [Privacy Loss [1]] For adjacent datasets D and D', mechanism $\mathcal{M}$, and output $o \in \mathbb{R}$, the privacy loss at o is defined as:

$$c\left(o;\mathcal{M},D,D'\right)=\log\left(\frac{\mathbb{P}[\mathcal{M}(D)=o]}{\mathbb{P}\left[\mathcal{M}\left(D'\right)=o\right]}\right) \tag{112}$$

Definition 3. [Moment [1]] For a mechanism $\mathcal{M}$ and the privacy loss at output o, the λ-th moment is defined as:

$$\alpha_{\mathcal{M}}\left(\lambda;D,D'\right)=\log\left(\mathbb{E}_{o\sim\mathcal{M}(D)}\left[\exp\left(\lambda c\left(o;\mathcal{M},D,D'\right)\right)\right]\right) \tag{113}$$

with the upper bound given by:

$$\alpha_{\mathcal{M}}(\lambda)=\max_{D,D'}\alpha_{\mathcal{M}}\left(\lambda;D,D'\right) \tag{114}$$

Lemma 14. [Composability [1]] Let $\alpha_{\mathcal{M}}(\lambda)$ be defined as above. Suppose $\mathcal{M}$ is composed of several mechanisms $\mathcal{M}_1,\ldots,\mathcal{M}_k$, where $\mathcal{M}_i$ depends on $\mathcal{M}_1,\ldots,\mathcal{M}_{i-1}$. Then, for any λ:

$$\alpha_{\mathcal{M}}(\lambda)\leq\sum_{i=1}^k\alpha_{\mathcal{M}_i}(\lambda) \tag{115}$$

Lemma 15. [Tail Bound [1]] For any $\theta > 0$, mechanism $\mathcal{M}$ is (θ, γ)-differentially private if:

$$\gamma = \min_{\lambda} \exp\left(\alpha_{\mathcal{M}}(\lambda) - \lambda\theta\right) \tag{116}$$

Definition 4. [Rényi Divergence [10]] Let P and Q be probability distributions. For $\rho \in (1, \infty)$, the Rényi Divergence of order ρ between P and Q is defined as:

$$D_\rho(P\|Q) = \frac{1}{\rho - 1} \log\left(\mathbb{E}_{\mathbf{x}\sim P}\left[\left(\frac{P(\mathbf{x})}{Q(\mathbf{x})}\right)^{\rho-1}\right]\right) \tag{117}$$

Lemma 16. For Gaussian distributions $\mathcal{N}(\mu, \sigma^2 I_p)$ and $\mathcal{N}(\nu, \sigma^2 I_p)$, where $\mu, \nu \in \mathbb{R}^p$, $\sigma \in \mathbb{R}$, and $\rho \in (1, \infty)$, the Rényi Divergence of order ρ is given by:

$$D_\rho\left(\mathcal{N}\left(\mu, \sigma^2 I_p\right) \| \mathcal{N}\left(\nu, \sigma^2 I_p\right)\right) = \frac{\rho\|\mu - \nu\|_2^2}{2\sigma^2} \tag{118}$$

Firstly, let's review training process in DP-DM-HSGD:

$$\begin{aligned} \bar{\mathbf{x}}_{t+1} &= \bar{\mathbf{x}}_t - \eta_{\mathbf{x}}\left(\bar{\mathbf{g}}_t + \mathcal{N}_{\mathbf{x},t}\right) \\ \bar{\mathbf{y}}_{t+1} &= \bar{\mathbf{y}}_t + \eta_{\mathbf{y}}\left(\bar{\mathbf{h}}_t + \mathcal{N}_{y,t}\right) \end{aligned} \tag{119}$$

where $\mathcal{N}_{\mathbf{x},t} \sim \mathcal{N}\left(0, \frac{\sigma_{\mathbf{x}}^2}{m} I_{d_1}\right)$, $\mathcal{N}_{\mathbf{y},t} \sim \mathcal{N}\left(0, \frac{\sigma_{\mathbf{y}}^2}{m} I_{d_2}\right)$, and $\sigma_{\mathbf{x}} = c\left(\frac{L_g\sqrt{\left(\frac{8T(T+1)(2T+1)}{3} + 4T\right)\log(1/\gamma)}}{2\theta\sqrt{m}}\right)$

Now, we will present the full proof for Theorem 2. We first analyze parameter $\mathbf{x}$.
When updating $\mathbf{x}$, at iteration t, the randomized mechanism $\mathcal{M}_t$ which may disclose privacy is

$$\begin{aligned} \mathcal{M}_t &= \bar{\mathbf{g}}_t + \mathcal{N}_{\mathbf{x},t} \\ &= \frac{1}{m}\sum_{j=1}^{m}\left(\sum_{k=0}^{t}(1-\beta_{\mathbf{x}})^{t-k}\left[\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(i)}, \mathbf{y}_k^{(i)}; \mathbf{z}_k^{(i)}\right) - \nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(i)}, \mathbf{y}_k^{(i)}; \mathbf{z}_{k+1}^{(i)}\right)\right]\right. \\ &\quad \left. + \nabla_{\mathbf{x}}F_j\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}; \mathbf{z}_{t+1}^{(i)}\right)\right) + \mathcal{N}_{\mathbf{x},t} \end{aligned} \tag{120}$$

We set probability distribution of $\mathcal{M}_t$ over adjacent datasets D, D' as P and Q, respectively, also, we assume the single different data sample is on the m^{th} one, and we obtain:

$$\begin{aligned} P &= \frac{1}{m}\sum_{i=1}^{m-1}\left(\sum_{k=0}^{t}(1-\beta_{\mathbf{x}})^{t-k}\left[\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(i)}, \mathbf{y}_k^{(i)}; \mathbf{z}_k^{(i)}\right) - \nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(i)}, \mathbf{y}_k^{(i)}; \mathbf{z}_{k+1}^{(i)}\right)\right]\right. \\ &\quad \left. + \nabla_{\mathbf{x}}F_j\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}; \mathbf{z}_{t+1}^{(i)}\right)\right) + \frac{1}{m}\left(\sum_{k=0}^{t}(1-\beta_{\mathbf{x}})^{t-k}\left[\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(m)}, \mathbf{y}_k^{(m)}; \mathbf{z}_k^{(m)}\right)\right.\right. \\ &\quad \left.\left. - \nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(m)}, \mathbf{y}_k^{(m)}; \mathbf{z}_{t+1}^{(m)}\right)\right] + \nabla_{\mathbf{x}}F_j\left(\mathbf{x}_t^{(m)}, \mathbf{y}_t^{(m)}; \mathbf{z}_{t+1}^{(m)}\right)\right) + \mathcal{N}_{\mathbf{x},t} \\ Q &= \frac{1}{m}\sum_{i=1}^{m-1}\left(\sum_{k=0}^{t}(1-\beta_{\mathbf{x}})^{t-k}\left[\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(i)}, \mathbf{y}_k^{(i)}; \mathbf{z}_k^{(i)}\right) - \nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(i)}, \mathbf{y}_k^{(i)}; \mathbf{z}_{k+1}^{(i)}\right)\right]\right. \\ &\quad \left. + \nabla_{\mathbf{x}}F_j\left(\mathbf{x}_t^{(i)}, \mathbf{y}_t^{(i)}; \mathbf{z}_{t+1}^{(i)}\right)\right) + \frac{1}{m}\left(\sum_{k=0}^{t}(1-\beta_{\mathbf{x}})^{t-k}\left[\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(m)}, \mathbf{y}_k^{(m)}; \mathbf{z}_k^{(m)\prime}\right)\right.\right. \\ &\quad \left.\left. - \nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(m)}, \mathbf{y}_k^{(m)}; \mathbf{z}_{k+1}^{(m)\prime}\right)\right] + \nabla_{\mathbf{x}}F_j\left(\mathbf{x}_t^{(m)}, \mathbf{y}_t^{(m)}; \mathbf{z}_{t+1}^{(m)\prime}\right)\right) + \mathcal{N}_{\mathbf{x},t} \end{aligned} \tag{121}$$

For the simplicity of the next steps, we set

$$\begin{aligned}
\mathcal{I} &= \frac{1}{m}\sum_{i=1}^{m-1}\left(\sum_{k=0}^{t}(1-\beta_{\mathbf{x}})^{t-k}\left[\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(i)},\mathbf{y}_k^{(i)};\mathbf{z}_k^{(i)}\right)-\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(i)},\mathbf{y}_k^{(i)};\mathbf{z}_{k+1}^{(i)}\right)\right]\right.\\
&\left.+\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_{t+1}^{(i)}\right)\right)+\frac{1}{m}\left(\sum_{k=0}^{t}(1-\beta_{\mathbf{x}})^{t-k}\left[\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(m)},\mathbf{y}_k^{(m)};\mathbf{z}_k^{(m)}\right)\right.\right.\\
&\left.\left.-\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(m)},\mathbf{y}_k^{(m)};\mathbf{z}_{t+1}^{(m)}\right)\right]+\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_t^{(m)},\mathbf{y}_t^{(m)};\mathbf{z}_{t+1}^{(m)}\right)\right)\\
\mathcal{I}' &= \frac{1}{m}\sum_{i=1}^{m-1}\left(\sum_{k=0}^{t}(1-\beta_{\mathbf{x}})^{t-k}\left[\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(i)},\mathbf{y}_k^{(i)};\mathbf{z}_k^{(i)}\right)-\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(i)},\mathbf{y}_k^{(i)};\mathbf{z}_{k+1}^{(i)}\right)\right]\right.\\
&\left.+\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_{t+1}^{(i)}\right)\right)+\frac{1}{m}\left(\sum_{k=0}^{t}(1-\beta_{\mathbf{x}})^{t-k}\left[\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(m)},\mathbf{y}_k^{(m)};\mathbf{z}_k^{(m)\prime}\right)\right.\right.\\
&\left.\left.-\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(m)},\mathbf{y}_k^{(m)};\mathbf{z}_{k+1}^{(m)\prime}\right)\right]+\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_t^{(m)},\mathbf{y}_t^{(m)};\mathbf{z}_{t+1}^{(m)\prime}\right)\right)
\end{aligned}\tag{122}$$

As we define $\mathbf{z}^{(i)}$ as an index sample in local dataset $\mathcal{Z}$, therefore, according to Assumption 7:

$$\left\|\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_t^{(i)}\right)-\nabla_{\mathbf{x}}F_i\left(\mathbf{x}_t^{(i)},\mathbf{y}_t^{(i)};\mathbf{z}_t^{(i)\prime}\right)\right\|_2\leq 2L_g \tag{123}$$

The inequality above stands because $\mathbf{z}_t^{(m)}$ and $\mathbf{z}_t^{(m)\prime}$ are adjacent data samples. Now, we can get:

$$\begin{aligned}
&\left\|\mathcal{I}-\mathcal{I}'\right\|_2^2\\
&=\left\|\frac{1}{m}\left(\sum_{k=0}^{t}(1-\beta_{\mathbf{x}})^{t-k}\left[\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(m)},\mathbf{y}_k^{(m)};\mathbf{z}_k^{(m)}\right)-\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(m)},\mathbf{y}_k^{(m)};\mathbf{z}_{k+1}^{(m)}\right)\right]\right.\right.\\
&\quad\left.+\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_t^{(m)},\mathbf{y}_t^{(m)};\mathbf{z}_{t+1}^{(m)}\right)\right)-\frac{1}{m}\left(\sum_{k=0}^{t}(1-\beta_{\mathbf{x}})^{t-k}\left[\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(m)},\mathbf{y}_k^{(m)};\mathbf{z}_k^{(m)\prime}\right)\right.\right.\\
&\quad\left.\left.\left.-\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(m)},\mathbf{y}_k^{(m)};\mathbf{z}_{k+1}^{(m)\prime}\right)\right]+\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_t^{(m)},\mathbf{y}_t^{(m)};\mathbf{z}_{t+1}^{(m)\prime}\right)\right)\right\|_2^2\\
&\leq\frac{2}{m^2}\left(\sum_{k=0}^{t}(1-\beta_{\mathbf{x}})^{t-k}\left(\left\|\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(m)},\mathbf{y}_k^{(m)};\mathbf{z}_k^{(m)}\right)-\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(m)},\mathbf{y}_k^{(m)};\mathbf{z}_k^{(m)\prime}\right)\right\|_2\right.\right.\\
&\quad\left.\left.-\left\|\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(m)},\mathbf{y}_k^{(m)};\mathbf{z}_{k+1}^{(m)}\right)-\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_k^{(m)},\mathbf{y}_k^{(m)};\mathbf{z}_{k+1}^{(m)\prime}\right)\right\|_2\right)\right)^2\\
&\quad+\frac{2}{m^2}\left(\left\|\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_t^{(m)},\mathbf{y}_t^{(m)};\mathbf{z}_{t+1}^{(m)}\right)-\nabla_{\mathbf{x}}F_j\left(\mathbf{x}_t^{(m)},\mathbf{y}_t^{(m)};\mathbf{z}_{t+1}^{(m)\prime}\right)\right\|_2\right)^2\\
&\leq\frac{2}{m^2}\left(\sum_{k=0}^{t}(1-\beta_{\mathbf{x}})^{t-k}4L_g\right)^2+\frac{8L_g^2}{m^2}\\
&\leq\frac{2L_g^2}{m^2}\left(16\left(\sum_{k=0}^{t}(1-\beta_{\mathbf{x}})^{t-k}\right)^2+4\right)\leq\frac{2L_g^2}{m^2}\left(16t^2+4\right)=\frac{L_g^2}{m^2}\left(32t^2+8\right)
\end{aligned}\tag{124}$$

where we use Young's inequality in the first inequality, and simplify $\sum_{k=0}^{t}(1-\beta_{\mathbf{x}})^{t-k}$ to $\sum_{k=0}^{t}1$ in last inequality as we have a bound $0<\beta_x<1$. Noting that $\mathcal{N}_{\mathbf{x},t}\sim\mathcal{N}\left(0,\frac{\sigma_{\mathbf{x}}^2}{m}I_p\right)$, we have

$$P\sim\mathcal{N}\left(\mathcal{I},\frac{\sigma_{\mathbf{x}}^2}{m}I_p\right),Q\sim\mathcal{N}\left(\mathcal{I}',\frac{\sigma_{\mathbf{x}}^2}{m}I_p\right) \tag{125}$$

With Definition 2 and 3, we obtain:

$$\alpha_{\mathcal{M}_t}\left(\lambda;D,D'\right)=\log\left(\mathbb{E}_{o\sim P}\left[\exp\left(\lambda\log\left(\frac{P}{Q}\right)\right)\right]\right)=\log\left(\mathbb{E}_{o\sim P}\left[\left(\frac{P}{Q}\right)^{\lambda}\right]\right)=\lambda D_{\lambda+1}(P\|Q) \tag{126}$$

Where we use Definition 4 to get the last inequality. From lemma 16, we have:

$$\alpha_{\mathcal{M}_t}\left(\lambda; D, D'\right) = \frac{m\lambda(\lambda+1)\left\|\mathcal{I}-\mathcal{I}'\right\|_2^2}{2\sigma^2} \leq \frac{\left(16t^2+4\right)L_g^2\lambda(\lambda+1)}{m\sigma^2} = \alpha_{\mathcal{M}_t}(\lambda) \tag{127}$$

The inequality holds because $F_i(\cdot,\cdot;\cdot)$ is L-Lipschitz, and the last step holds because of Definition 3. By Lemma 14, there are T iterations, so we have

$$\alpha_{\mathcal{M}}(\lambda) \leq \sum_{t=1}^{T}\alpha_{\mathcal{M}_t}(\lambda) \leq \frac{2\left(16t^2+4\right)L_g^2\lambda^2}{m\sigma^2} \tag{128}$$

where the last inequality holds because $\lambda \in (1,\infty)$. We also assume that the maximum value of $\alpha_{\mathcal{M}_t}(\lambda)$ is no greater than twice the average of its sum.

Taking $\sigma_{\mathbf{x}} = c\frac{L_g\sqrt{\left(\frac{8T(T+1)(2T+1)}{3}+4T\right)\log(1/\gamma)}}{2\theta\sqrt{m}}$, we can guarantee $\alpha_{\mathcal{M}}(\lambda) \leq \lambda\theta/2$, and as the consequence, by Lemma 16, we obtain $\gamma \leq \exp(-\lambda\theta/2)$, and in this way, it leads (θ,γ)-DP to parameter $\bar{\mathbf{x}}$. Similarly, we have the same proof for $\bar{\mathbf{y}}_t$, if $\mathcal{N}_{y,t} \sim \mathcal{N}\left(0, \frac{\sigma_{\mathbf{y}}^2}{m}I_p\right)$ with $\sigma_{\mathbf{y}} = c\frac{L_g\sqrt{\left(\frac{8T(T+1)(2T+1)}{3}+4T\right)\log(1/\gamma)}}{2\theta\sqrt{m}}$ is used when updating $\bar{\mathbf{y}}_t$, then (θ,γ)-DP can be guaranteed. The proof is completed.

C Additional Experiments

C.1 Gradient clipping

We conducted experiments on DPMixSGD and DM-HSGD with a clipping threshold set to clip the top 20% of gradients, which yielded similar results, further confirming the reliability of our original findings. See in Table 3, here σ represents the intensity of the noise added in the DPMixSGD algorithm.

Table 3: AUC Performance of DPMixSGD and DM-HSGD under Different Noise Levels with Gradient Clipping Comparison

Dataset	DM-HSGD	$\sigma = 0.5$		$\sigma = 1$	
		DPMixSGD	DPMixSGD (Clip)	DPMixSGD	DPMixSGD (Clip)
MNIST	0.9937	0.9897	0.9733	0.9796	0.9548
Fashion_MNIST	0.9859	0.9757	0.9493	0.9627	0.9184
ljcnn1	0.9984	0.9962	0.9889	0.9901	0.9711

C.2 Decentralized min-max problem in multilayer perceptron of image classification problem

This experiment focuses on image classification of the Fashion-MNIST [75] dataset using a multilayer perceptron (MLP) model. We introduce corresponding dual variables to formulate a min-max problem. Additionally, we also compare the AUROC performance of the DPMixSGD, DM-HSGD, SGDA, and DP-SGDA algorithms across different scenarios. In this problem, we consider a distributed network composed of m agents. Each agent i possesses its own model parameter $\mathbf{x}_i$ as well as a set of dual variables $y_{a,i}$, $y_{b,i}$, and $y_{w,i}$. These dual variables are typically employed to handle constraints or to model adversarial factors. The optimization objective of the entire MLP system is defined as follows:

$$\min_{\{\mathbf{x}_i\}_{i=1}^m}\max_{\{y_{a,i},y_{b,i},y_{w,i}\}_{i=1}^m}\Phi\left(\{\mathbf{x}_i\},\{y_{a,i},y_{b,i},y_{w,i}\}\right), \tag{129}$$

where Φ is the global objective function, defined as the average of all agents' local objective functions as the following:

$$\Phi\left(\{\mathbf{x}_i\},\{y_{a,i},y_{b,i},y_{w,i}\}\right) = \frac{1}{m}\sum_{i=1}^{m}\phi_i\left(\mathbf{x}_i,y_{a,i},y_{b,i},y_{w,i}\right). \tag{130}$$

Each agent i has a local optimization function ϕ_i defined as:

$$\begin{aligned}&\phi_i\left(\mathbf{x}_i,y_{a,i},y_{b,i},y_{w,i}\right)\\ &= \mathcal{L}\left(\mathbf{x}_i;\mathcal{D}_i\right) + y_{a,i}\cdot f_a\left(\mathbf{x}_i\right) + y_{b,i}\cdot f_b\left(\mathbf{x}_i\right) + y_{w,i}\cdot f_w\left(\mathbf{x}_i\right),\end{aligned} \tag{131}$$

where $\mathcal{L}\left(\mathbf{x}_i;\mathcal{D}_i\right)$ is the primary loss function based on the local dataset $\mathcal{D}_i$ (e.g., cross-entropy loss). $f_a\left(\mathbf{x}_i\right)$, $f_b\left(\mathbf{x}_i\right)$, and $f_w\left(\mathbf{x}_i\right)$ are auxiliary functions associated with the dual variables, introduce to impose additional constraints or model adversarial factors. $y_{a,i}$, $y_{b,i}$, and $y_{w,i}$ are the corresponding dual variables, typically acting as lagrange multipliers to balance the primary loss with the auxiliary terms.

For the image classification algorithms DPMixSGD, DM-HSGD, SGDA, and DP-SGDA, we conduct extensive experimental validations and compare their AUROC metrics. The primary parameters involved in the experiments are as follows: the learning rates for the model

Table 4: AUROC results over epochs for each algorithm during the image classification experiments on Fashion-MNIST dataset.

(a) Impact of total number of agents m.

m	$m = 5$	$m = 10$	$m = 15$	$m = 20$
SGDA	0.7978	0.7227	0.5602	0.5503
DP-SGDA	0.7754	0.7251	0.5367	0.5506
DM-HSGD	0.9352	0.9179	0.9345	0.9087
DPMixSGD	0.9311	0.9310	0.9296	0.9317

(b) Impact of sparsity level p.

p	$p = 0.2$	$p = 0.5$	$p = 0.8$	$p = 1$
SGDA	0.7881	0.7978	0.7978	0.7971
DP-SGDA	0.7816	0.7796	0.7754	0.7769
DM-HSGD	0.9359	0.9357	0.9352	0.9329
DPMixSGD	0.9328	0.9373	0.9311	0.9369

(c) Impact of θ.

θ	$\theta = 0.005$	$\theta = 0.01$	$\theta = 0.05$	$\theta = 0.1$
SGDA	0.7978	0.7978	0.7978	0.7978
DP-SGDA	0.6637	0.5773	0.7066	0.7542
DM-HSGD	0.9351	0.9351	0.9351	0.9351
DPMixSGD	0.9048	0.9213	0.9356	0.9355

(d) Impact of γ.

γ	$\gamma = \frac{1}{60000}$	$\gamma = \frac{1}{30000}$	$\gamma = \frac{1}{5000}$	$\gamma = \frac{1}{1000}$
SGDA	0.7978	0.7978	0.7978	0.7978
DP-SGDA	0.5795	0.5773	0.5732	0.5725
DM-HSGD	0.9351	0.9351	0.9351	0.9351
DPMixSGD	0.9206	0.9213	0.9237	0.9264

parameters $\mathbf{x}$ and their dual variables $\mathbf{y}$ are selected from the set $\{0.01, 0.001, 0.0001\}$. The mini-batch size is set to 64. Specifically, for the DPMixSGD and DM-HSGD algorithms, the initial batch size is set to $b_0 = 64$. The gradient weight adjustment parameters β_x and β_y are chosen from the set $\{0.1, 0.01\}$. Table 4 illustrates the AUROC results over epochs for each algorithm during the image classification experiments. In all compared groups, our proposed method surpasses existing algorithms, because the introduced noise aids in escaping saddle points while expediting the model's training process.