# PDSL: Privacy-Preserved Decentralized Stochastic Learning with Heterogeneous Data Distribution

Lina Wang[a], Yunsheng Yuan[a], Chunxiao Wang[b,c], Feng Li[a,b]

[a] School of Computer Science and Technology, Shandong University, Qingdao, China

[b] Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

[c] Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan, China

Email: linawang425@mail.sdu.edu.cn, ysyuan1028@mail.sdu.edu.cn, wangchx@sdas.org, fli@sdu.edu.cn

## Abstract

In the paradigm of decentralized learning, a group of agents collaborates to learn a global model using distributed datasets without a central server. However, due to the heterogeneity of the local data across the different agents, learning a robust global model is rather challenging. Moreover, the collaboration of the agents relies on their gradient information exchange, which poses a risk of privacy leakage. In this paper, to address these issues, we propose PDSL, a novel privacy-preserved decentralized stochastic learning algorithm with heterogeneous data distribution. On one hand, we innovate in utilizing the notion of Shapley values such that each agent can precisely measure the contributions of its heterogeneous neighbors to the global learning goal; on the other hand, we leverage the notion of differential privacy to prevent each agent from suffering privacy leakage when it contributes gradient information to its neighbors. We conduct both solid theoretical analysis and extensive experiments to demonstrate the efficacy of our PDSL algorithm in terms of privacy preservation and convergence.

## Index Terms

Decentralized learning, data heterogeneity, differential privacy, Shapley values

## I. Introduction

Distributed machine learning refers to a family of algorithms aiming at learning from data distributed among multiple agents [1]–[3]. Based on communication topology, approaches to designing distributed learning algorithms include *centralized learning* and *decentralized learning*. As a popular approach for centralized learning, *Federated Learning* (FL) enables the central server to organize the collaborative training process through the interaction of model parameters while training data can be stored locally at the agents [4], [5]. However, this setup requires constant communication with the central server, potentially creating a bottleneck. To mitigate this concern, some decentralized learning algorithms have been proposed such that agents communicate their updates with their neighbors in a communication network without the assistance of a central server.

One of the fundamental challenges for decentralized learning is data heterogeneity [6]. For example, due to diverse users behaviors, data across different agents are distributed non-independently and identically (Non-IID), while such a Non-IID data distribution may result in a degradation in terms of convergence and accuracy [7]. Although the problem of data heterogeneity has been extensively investigated for centralized learning (e.g., FL) [8]–[11], there are only a very handful of studies on decentralized learning. Recently, decentralized learning algorithms based on cross-gradient information have been proposed in [12], [13]. Each agent aggregates cross-gradient information, i.e., derivatives of its model with respect to its neighbors' datasets, to update its local model. [14] incorporates a recursive gradient correction technique to address data heterogeneity. Unfortunately, these algorithms cannot fairly evaluate the contribution of each agent.

Another significant challenge in decentralized learning is privacy preservation. Decentralized learning entails information exchange among the agents. For example, in [12], [13], each agent calculates cross-gradients for each of its neighbors based on its local dataset, while numerous recent studies have shown that sharing the cross-gradient information can potentially result in the leakage of private local datasets [15]–[17]. To tackle this problem, one popular choice is *Differential Privacy* (DP). For example, [18] incorporates DP into an asynchronous decentralized parallel SGD framework. [19] proposes a differentially private decentralized optimization algorithm that alternates between local gradient descent and gossip averaging. Unfortunately, these state-of-the-art proposals do not take into account the heterogeneous data distribution.

In this paper, we propose PDSL, a novel privacy-preserved decentralized stochastic learning algorithm with heterogeneous data distribution. Specifically, in each round of our PDSL algorithm, each agent first computes its local gradient and cross-gradient, i.e., the derivative of its local model the ones of its neighbors' local models with respect to its local datasets. The agent then adds noise to obtain the perturbed gradients based on DP mechanism. Next, each agent exchanges the perturbed cross-gradient information with its neighbors. Finally, each agent updates its model in a momentum-like manner based on the received perturbed cross-gradient and its perturbed local gradient. To properly weight the perturbed gradients in the aggregation,

each agent utilizes Shapley values to fairly quantify the contributions of its perturbed local gradient and the received perturbed cross-gradient. In summary, our main contributions are as follows:

- We propose PDSL, an innovative and robust decentralized learning algorithm with DP to address heterogeneous data distribution and privacy leakage risks.
- We conduct rigorous theoretical analysis to prove the DP guarantee of our algorithm and verify its convergence under the DP framework.
- We carry out an extensive experimental assessment of PDSL across diverse real-world datasets. The outcomes validate both its practical viability and efficacy.

The remaining of this paper is organized as follows. Related literature and preliminaries are first introduced in Sec. II and Sec. III, respectively. We then present our PDSL algorithm in detail in Sec. IV. Furthermore, we conduct solid theoretical analysis on the privacy protection and convergence of our PDSL algorithm in Sec. V. We also perform extensive experiments to verify the efficacy of our PDSL algorithm in Sec. VI and finally conclude our paper in Sec. VII.

## II. RELATED WORK

The paradigm of decentralized learning has been attracting increasing attention in recent years. [20] integrates *Stochastic Gradient Descent* (SGD) with gossip-averaging as proposed in [21] to develop the *Decentralized Parallel Stochastic Gradient Descent* (D-PSGD) algorithm. [22] considers the distributed optimization of non-smooth convex function and proposes the *Multi-Step Primal-Dual* (MSPD) algorithm for non-smooth decentralized optimization. Moreover, [23] proposes the momentum version of D-PSGD, namely *Decentralized Momentum Stochastic Gradient Descent* (DMSGD) algorithm. [24] presents Soft-DSGD for communication networks featuring lightweight and unreliable communication protocols.

These studies usually assume that data is independently and identically distributed across different agents. However, this assumption may not always hold in real-world applications. Therefore, there exist some state-of-the-art proposals that considering the challenge of data heterogeneity. [25] combines quasi-global momentum with local stochastic gradients to mitigate the difficulties of decentralized learning on heterogeneous data. [26] incorporates *Mutual Knowledge Transfer* (MKT) into decentralized learning to alleviate the negative impact of data heterogeneity. [27] indicates that model inconsistency caused by heterogeneous data and network topology in decentralized FL may result in over-fitting in local models and performance degradation. Therefore, DFedSAM and DFedSAM-MGS are proposed in [27] to enhance the performance of decentralized learning. Specifically, DFedSAM employs *Sharpness Aware Minimization* (SAM) to generate flat model in each client, DFedSAM-MGS further improves DFedSAM by utilizing *Multiple Gossip Steps* (MGS) for better model consistency. Moreover, various tracking approaches, such as gradient tracking [28] and momentum tracking [29], have been introduced to address the issue of heterogeneous data in decentralized learning. [30] introduces *Global Update Tracking* (GUT) algorithm which applies the tracking mechanism with respect to the model updates instead of the gradients. Recently, [12] and [13] adopt the concept of cross-gradients to reduce the impact of heterogeneous data distribution in decentralized learning. Specifically, [12] collects the gradient information of each agent's model on all its neighbors' datasets and uses quadratic programming to project them into a single gradient which is then used to update the model. In [13], each agent aggregates two types of cross-gradients, i.e., the derivatives computed locally using its local data on the neighbors' model parameters, and the gradients of its local model with respect to its neighbors' datasets.

While decentralized learning algorithms enable agents to avoid sharing raw data through local training and exchanging model parameters or gradients, [15], [17], [31] have shown that malicious adversaries can infer private information from them. *Differential privacy* (DP) [32] has been widely used to prevent privacy leakage in machine learning. Despite extensive studies in centralized distributed learning (e.g., FL) [33], [34], only a few works investigates the application of DP in decentralized learning algorithms. [35] applies DP to ADMM-based decentralized updates. [19] proposes Muffliato, a privacy amplification mechanism consisting of local Gaussian noise injection at the agent level followed by a gossiping process for averaging private values. [18] presents a DP-based asynchronous decentralized parallel SGD. However, these proposals do not consider the data heterogeneity across the agents. In this paper, we leverage the notion of Shapley value to design a DP-based decentralized learning algorithm on heterogeneous data distribution.

## III. PRELIMINARIES

We first introduce the basics of decentralized learning and DP in Sec. III-A and Sec. III-B, respectively. We then introduce the concept of Shapley value in Sec. III-C.

## A. Decentralized Stochastic Learning

Consider a set of agents $\mathcal{M} = \{1, 2, \cdots, M\}$. Each agent $i \in \mathcal{M}$ owns a private local dataset $\mathcal{D}_i$. The aim of decentralized stochastic learning is to let the agents collaborate to train a global model with their local datasets. Specifically, the objective can be formulated as follows:

$$\min_{x \in \mathbb{R}^d} \mathcal{F}(x) := \frac{1}{M} \sum_{i=1}^{M} f_i(x) \tag{1}$$

where $x \in \mathbb{R}^d$ denotes model parameter, and

$$f_i(x) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i} \left[ F_i(x; \xi_i) \right] \tag{2}$$

is the expectation of local loss function $F_i$ associated with agent $i$ on its local data distribution $\mathcal{D}_i$.

We assume the agents communicate with each other over a graph $\mathcal{G} = (\mathcal{M}, \mathbf{W})$, where $\mathbf{W} \in [0, 1]^{M \times M}$ is the weighted adjacent matrix of $\mathcal{G}$. We suppose $\mathbf{W}$ is a symmetric doubly stochastic matrix such that $\sum_{j \in \mathcal{M}} \omega_{i,j} = \sum_{i \in \mathcal{M}} \omega_{i,j} = 1$ and $\omega_{i,j} = \omega_{j,i}, \forall i, j \in \mathcal{M}$, where $\omega_{i,j}$ denotes the $(i, j)$-th element of $\mathbf{W}$. Let $\mathcal{M}_i \subseteq \mathcal{M}$ be the set of neighbors of agent $i$ including itself, i.e., $\mathcal{M}_i = \{j \in \mathcal{M} \mid \omega_{i,j} > 0\}$. $\omega_{i,j}$ is known for agent $i$ for any $j \in \mathcal{M}_i$.

In our decentralized learning algorithm, we employ a validation dataset $\mathcal{Q}$ to address the problem of data heterogeneity. Specifically, each agent measures the data relevances (or contributions) of its local model and the ones of its peers to the global training goal. Let $\mathcal{Q}$ denote the global validation dataset, which is loaded in each agent in advance.

## B. Differential Privacy

In this paper, we leverage the notion of *Differential Privacy* (DP) for the purpose of privacy preservation. It is defined in terms of adjacent datasets. Specifically, two datasets are said to be adjacent if they differ in a single entry. A formal definition of DP is given in **Definition** 1.

**Definition 1** (($\epsilon, \delta$)-Differential Privacy). *A randomized mechanism $\mathcal{A} : \mathsf{D} \to \mathsf{R}$ satisfies $(\epsilon, \delta)$-differential privacy, if for any two adjacent datasets $\mathcal{D}, \mathcal{D}' \in \mathsf{D}$ and any subset of outputs $\mathcal{S} \subseteq \mathsf{R}$, it holds that*

$$\Pr\left[\mathcal{A}(\mathcal{D}) \in \mathcal{S}\right] \leq e^\epsilon \Pr\left[\mathcal{A}(\mathcal{D}') \in \mathcal{S}\right] + \delta \tag{3}$$

$\epsilon$ denotes the so-called *privacy budget*, which specifies the privacy loss we can afford. Smaller privacy budget implies stronger privacy preservation is necessitated.

A widely used $(\epsilon, \delta)$-DP mechanism is Gaussian mechanism. Given any query function $q : \mathsf{D} \to \mathsf{R}$, $(\epsilon, \delta)$-DP can be ensurend by adding Gaussian noises to each dimension of $\mathsf{R}$. In particular, for any dataset $\mathcal{D} \in \mathsf{D}$, we have

$$\mathcal{A}(\mathcal{D}) = q(\mathcal{D}) + \mathcal{N}(0, \sigma^2 \mathbf{I}_d) \tag{4}$$

where $\mathcal{N}$ denotes a Gaussian distribution with zero mean, $\mathbf{I}_d$ is a $d$-dimensional identity matrix, and $\sigma^2 \mathbf{I}_d$ denotes the covariance matrix of the Gaussian distribution. We define the sensitivity of $q$ in **Definition** 2, and the noise magnitude $\sigma$ is proportional to the sensitivity of $q$.

**Definition 2** (Sensitivity). *The sensitivity of query function $q$ is the maximum absolute difference derived on the output of any two adjacent datasets $\mathcal{D}, \mathcal{D}' \in \mathsf{D}$, i.e.,*

$$\triangle_2 q = \max_{\mathcal{D}, \mathcal{D}'} \|q(\mathcal{D}) - q(\mathcal{D}')\| \tag{5}$$

## C. Cooperative Game and Shapley Value

**Definition 3** (Cooperative Game). *Let $\mathcal{Z} = \{1, 2, \cdots, Z\}$ denote a set of $Z$ players. A cooperative game is represented by a tuple $(\mathcal{Z}, v)$, where $v : 2^{\mathcal{Z}} \to \mathbb{R}$ is a characteristic function of the game, assigning a real number to each coalition $\mathcal{Z}' \subseteq \mathcal{Z}$ and satisfying $v(\emptyset) = 0$.*

The characteristic function $v(\mathcal{Z}')$ assigns the payoff attainable by coalition $\mathcal{Z}' \subseteq \mathcal{Z}$. The distribution of the payoff of the coalition $\mathcal{Z}'$ among the players is a fundamental issue for cooperative game. As presented in [36], the payoff allocation among the players should meet the following desirable properties, i.e., balance, symmetry, additivity and zero element. Specifically,

- *Balance*. The payoff obtained by each player in $\mathcal{Z}$ must add up to $v(\mathcal{Z})$.
- *Symmetry*. Two players must receive the same payoff if they contribute the same to each coalition.
- *Additivity*. The solution to the sum of two cooperative games must be the sum of what it awards to each of the two games.
- *Zero element*. A player should not receive payoff if it contributes nothing to each coalition.

To design a payoff allocation solution that satisfies the aforementioned properties, [36] proposed to assign the payoff to each player by considering its marginal contributions to all coalitions that the player could possibly participate in. Specifically, the Shapley value of a player is a weighted sum of the player's marginal contributions across all possible coalitions in the game.

**Definition 4** (Shapley Value). *Let $\mathcal{S} \in \Omega(\mathcal{Z})$ denote a permutation of players in $\mathcal{Z}$ and $\mathcal{Z}_i(\mathcal{S})$ denote the coalition made of all predecessors of player $i$ in $\mathcal{S}$. Assuming $\mathcal{S}(i)$ represents the location of $i$ in $\mathcal{S}$, $\mathcal{Z}_i(\mathcal{S})$ can be denoted as*

$$\mathcal{Z}_i(\mathcal{S}) = \{i' \in \mathcal{S} \mid \mathcal{S}(i') < \mathcal{S}(i)\} \tag{6}$$

*Given a cooperative game $(\mathcal{Z}, v)$, the Shapley value for player $i \in \mathcal{Z}$ is defined as follows*

$$\varphi_i(v) = \frac{1}{Z!} \sum_{\mathcal{S} \in \Omega(\mathcal{Z})} \left( v\left(\mathcal{Z}_i(\mathcal{S}) \bigcup \{i\}\right) - v\left(\mathcal{Z}_i(\mathcal{S})\right) \right) \tag{7}$$

In another word, the payoff allocated to player $i$ in a coalition game is the average marginal contribution of $i$ to $\mathcal{Z}_i(\mathcal{S})$ over all $\mathcal{S} \in \Omega(\mathcal{Z})$. It is easy to show that (7) can be re-written as:

$$\varphi(i) = \sum_{\mathcal{Z}' \subseteq \mathcal{Z} \setminus \{i\}} \left( Z\binom{Z-1}{Z'} \right)^{-1} \left( v\left(\mathcal{Z}' \bigcup \{i\}\right) - v\left(\mathcal{Z}'\right) \right) \tag{8}$$

where $Z = |\mathcal{Z}|$ and $Z' = |\mathcal{Z}'|$.

## IV. ALGORITHM DESIGN

In our algorithm, cross-gradient information finely weighted by Shapley values is utilized to address the problem of data heterogeneity. Unfortunately, exchanging the cross-gradient information among the agents may result in privacy leakage. To resolve this issue, we leverage the notion of DP, such that the cross-gradient information can be perturbed through Gaussian mechanism for the purpose of privacy preservation. Specifically, the time span of our algorithm is divided into $T$ rounds. In each round $t$, each agent $i \in \mathcal{M}$ first calculates its local stochastic gradient $g_{i,i}^{[t]}$ according to its local dataset, and perturbs $g_{i,i}^{[t]}$ through the Gaussian mechanism. Then, each agent $i \in \mathcal{M}$ exchanges its current local model $x_i^{[t-1]}$ with its neighbors. Once receiving the local model $x_j^{[t-1]}$ from any neighbor $j \in \mathcal{M}_i$, agent $i$ calculates stochastic cross-gradient $g_{i,j}^{[t]}$ according to its local dataset for agent $j$. $g_{i,j}^{[t]}$ is perturbed by a Gaussian noise $n_{i,j}^{[t]}$, and the perturbed cross-gradient $\hat{g}_{i,j}^{[t]}$ is then sent back to agent $j$. When receiving $\hat{g}_{j,i}^{[t]}$ from each neighbor $j \in \mathcal{M}_i$, agent $i$ calculates an update of its local model $x_i^{[t-1]}$ according to $\hat{g}_{j,i}^{[t]}$. Based on the updates $\left\{ x_{i,j}^{[t]} \right\}_{j \in \mathcal{M}_i}$, agent $i$ calculates Shapley value $\varphi_{i,j}^{[t]}$ for each $j \in \mathcal{M}_i$, and the Shapley values are utilized to compute the weight parameters for the aggregation of the received (perturbed) stochastic gradients. The local model $x_i^{[t]}$ at each agent $i$ is finally updated in a monemtum-like manner.

The pseudo-code of our PDSL algorithm is given in **Algorithm** 1. For each agent $i \in \mathcal{M}$, in each round $t$, agent $i$ first computes its local stochastic gradient using a sample $\xi_{i,t}$ uniformly sampled from its local dataset $\mathcal{D}_i$

$$g_{i,i}^{[t]} = \nabla F_i\left(x_i^{[t-1]}; \xi_{i,t}\right) \tag{9}$$

as shown in Line 2. Then, as presented in Lines $3 \sim 4$, to limit gradient sensitivity and enable a quantifiable privacy assessment, agent $i$ clips its local stochastic gradient $g_{i,i}^{[t]}$

$$\tilde{g}_{i,i}^{[t]} = \frac{g_{i,i}^{[t]}}{\max\left\{1, C^{-1}\left\|g_{i,i}^{[t]}\right\|\right\}} \tag{10}$$

and adds random Gaussian noise $n_{i,i}^{[t]} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ to the clipped gradient such that

$$\hat{g}_{i,i}^{[t]} = \tilde{g}_{i,i}^{[t]} + n_{i,i}^{[t]} \tag{11}$$

where $C$ is the clipping threshold. Next, agent $i$ broadcasts its current local model $x_i^{[t-1]}$ to its neighbors $\mathcal{M}_i \setminus \{i\}$ (see Line 5). As shown in Lines 6~12, once receiving the local model $x_j^{[t-1]}$ from each neighbor $j \in \mathcal{M}_i \setminus \{i\}$, agent $i$ computes the stochastic cross-gradient

$$g_{i,j}^{[t]} = \nabla F_i\left(x_j^{[t-1]}; \xi_{i,t}\right) \tag{12}$$

$g_{i,j}^{[t]}$ is then clipped by

$$\tilde{g}_{i,j}^{[t]} = \frac{g_{i,j}^{[t]}}{\max\left\{1, C^{-1}\left\|g_{i,j}^{[t]}\right\|\right\}} \tag{13}$$

and perturbed by Gaussian noise $n_{i,j}^{[t]} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ as follows

$$\hat{g}_{i,j}^{[t]} = \tilde{g}_{i,j}^{[t]} + n_{i,j}^{[t]} \tag{14}$$

Thereafter, agent $i$ sends the perturbed gradient $\hat{g}_{i,j}^{[t]}$ back to agent $j$.

Based on the gradient $\hat{g}_{j,i}^{[t]}$ returned by each neighbor, agent $i$ updates its local model according to

$$x_{i,j}^{[t]} = x_i^{[t-1]} - \gamma \cdot \hat{g}_{j,i}^{[t]}, \ \forall j \in \mathcal{M}_i \tag{15}$$

where $\gamma$ denotes the learning rate, as illustrated in Lines 13~14. These local model updates are utilized to calculate the Shapley values (see Line 15). Specifically, we define the characteristic function $v : 2^{\mathcal{M}_i} \to \mathbb{R}$ as follows

$$v(\mathcal{M}'; \mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{\xi \in \mathcal{Q}} J\left(\xi; x_{i,\mathcal{M}'}^{[t]}\right), \ \forall \mathcal{M}' \subseteq \mathcal{M}_i \tag{16}$$

where

$$x_{i,\mathcal{M}'}^{[t]} = \frac{1}{|\mathcal{M}'|} \sum_{j \in \mathcal{M}'} x_{i,j}^{[t]} \tag{17}$$

and $J\left(\xi; x_{i,\mathcal{M}'}^{[t]}\right)$ denotes the accuracy of $x_{i,\mathcal{M}'}^{[t]}$ on the data sample $\xi \in \mathcal{Q}$. The Shapley value of any agent $j \in \mathcal{M}_i$ then can be calculated as

$$\varphi_{i,j}^{[t]} = \sum_{\mathcal{M}' \subseteq \mathcal{M}_i \setminus \{j\}} \frac{v(\mathcal{M}' \bigcup \{j\}; \mathcal{Q}) - v(\mathcal{M}'; \mathcal{Q})}{|\mathcal{M}_i| \binom{|\mathcal{M}_i|-1}{|\mathcal{M}'|}} \tag{18}$$

and normalized by

$$\hat{\varphi}_{i,j}^{[t]} = \frac{\varphi_{i,j}^{[t]} - \min_{k \in \mathcal{M}_i} \varphi_{i,k}^{[t]}}{\max_{k \in \mathcal{M}_i} \varphi_{i,k}^{[t]} - \min_{k \in \mathcal{M}_i} \varphi_{i,k}^{[t]}} \tag{19}$$

as illustrated in Line 16. Based on these normalized Shapley values, agent $i$ calculates a weight parameter for $\forall j \in \mathcal{M}_i$

$$\pi_{i,j}^{[t]} = \frac{\hat{\varphi}_{i,j}^{[t]}}{\omega_{i,j} \sum_{k \in \mathcal{M}_i} \hat{\varphi}_{i,k}^{[t]}} \tag{20}$$

as demonstrated in Line 17. Next, as shown in Line 18, agent $i$ calculates a weighted average across the received perturbed stochastic gradients

$$\bar{g}_i^{[t]} = \sum_{j \in \mathcal{M}_i} \pi_{i,j}^{[t]} \hat{g}_{j,i}^{[t]} \tag{21}$$

Finally, the local model $x_i^{[t]}$ is updated in a momentum-like manner. Specifically, agent $i$ first calculates

$$\hat{u}_i^{[t]} = \alpha u_i^{[t-1]} + \bar{g}_i^{[t]} \tag{22}$$

and

$$\hat{x}_i^{[t]} = x_i^{[t-1]} - \gamma \hat{u}_i^{[t]} \tag{23}$$

as shown in Lines 19 and 20, respectively, where $\alpha$ denotes the momentum coefficient. It then broadcasts $\hat{u}_i^{[t]}$ and $\hat{x}_i^{[t]}$ to its neighbors (see Line 21). After receiving $\hat{u}_j^{[t]}$ and $\hat{x}_j^{[t]}$ from $\forall j \in \mathcal{M}_i \setminus \{i\}$, agent $i$ updates its momentum and local model by

$$u_i^{[t]} = \sum_{j \in \mathcal{M}_i} \omega_{i,j} \hat{u}_j^{[t]} \tag{24}$$

and

$$x_i^{[t]} = \sum_{j \in \mathcal{M}_i} \omega_{i,j} \hat{x}_j^{[t]} \tag{25}$$

Based on Sec. III-C, for each agent $i \in \mathcal{M}$, the computation of Shapley values according to (18) primarily depends on the permutations of $\mathcal{M}_i$. This process can lead to high computational complexity, particularly when agent $i$ has a large number of neighbors. To tackle this issue, we employ the Monte Carlo method to compute Shapley values with lower complexity [37], [38]. The pseudo-code of our Monte Carlo-based estimation algorithm (for agent $i$ in round $t$) is presented in **Algorithm** 2. Specifically, we initialize $\varphi_{i,j}^{[t]} = 0$ for each agent $j \in \mathcal{M}_i$. In each iteration $r = 1, 2, \cdots, R$, we randomly generate a

---

**Algorithm 1:** Our PDSL algorithm (at each agent $i$).

---

**Input:** Initial point $x_i^{[0]}$, learning rate $\gamma$, momentum buffer $u_i^{[0]} = 0$, momentum coefficient $\alpha$, doubly stochastic matrix $\mathbf{W}$, the number of iterations $T$, clipping threshold $C$, variance of noise $\sigma^2$, $d$-dimensional identity matrix $\mathbf{I}_d$.

**Output:** Local model $x_i^{[T]}$ of agent $i$

1   **foreach** $t = 1, 2, \cdots, T$ **do**

2     Compute local gradient $g_{i,i}^{[t]}$ according to (9);

3     Clip local stochastic gradient $\tilde{g}_{i,i}^{[t]} = \frac{g_{i,i}^{[t]}}{\max\left\{1, C^{-1}\left\|g_{i,i}^{[t]}\right\|\right\}}$;

4     Randomly generate Gaussian noise $n_{i,i}^{[t]} \in \mathbb{R}^d$ from $\mathcal{N}(0, \sigma^2\mathbf{I}_d)$ and add the noise to local stochastic gradient: $\hat{g}_{i,i}^{[t]} = \tilde{g}_{i,i}^{[t]} + n_{i,i}^{[t]}$;

5     Send $x_i^{[t-1]}$ to each neighbor $j \in \mathcal{M}_i \setminus \{i\}$;

6     **foreach** $j \in \mathcal{M}_i \setminus \{i\}$ **do**

7       Receive $x_j^{[t-1]}$ from $j \in \mathcal{M}_i \setminus \{i\}$;

8       Compute gradient $g_{i,j}^{[t]}$ according to (12);

9       Clip stochastic gradient $\tilde{g}_{i,j}^{[t]} = \frac{g_{i,j}^{[t]}}{\max\left\{1, C^{-1}\left\|g_{i,j}^{[t]}\right\|\right\}}$;

10      Randomly generate Gaussian noise $n_{i,j}^{[t]} \in \mathbb{R}^d$ from $\mathcal{N}(0, \sigma^2\mathbf{I}_d)$ and add the noise to stochastic gradient: $\hat{g}_{i,j}^{[t]} = \tilde{g}_{i,j}^{[t]} + n_{i,j}^{[t]}$;

11      Send $\hat{g}_{i,j}^{[t]}$ back to agent $j$;

12    **end**

13    Receive $\hat{g}_{j,i}^{[t]}$ from each neighbor $j \in \mathcal{M}_i \setminus \{i\}$;

14    Compute $x_{i,j}^{[t]} = x_i^{[t-1]} - \gamma \cdot \hat{g}_{j,i}^{[t]}$ for $\forall j \in \mathcal{M}_i$;

15    Compute Shapley value $\varphi_{i,j}^{[t]}$ for $\forall j \in \mathcal{M}_i$ according to **Algorithm 2**;

16    Compute normalized Shapley value $\hat{\varphi}_{i,j}^{[t]}$ for $\forall j \in \mathcal{M}_i$ according to (19);

17    Compute weight $\pi_{i,j}^{[t]}$ for $\forall j \in \mathcal{M}_i$ based on (20);

18    $\bar{g}_i^{[t]} = \sum_{j \in \mathcal{M}_i} \pi_{i,j}^{[t]} \hat{g}_{j,i}^{[t]}$;

19    $\hat{u}_i^{[t]} = \alpha u_i^{[t-1]} + \bar{g}_i^{[t]}$;

20    $\hat{x}_i^{[t]} = x_i^{[t-1]} - \gamma \hat{u}_i^{[t]}$;

21    Send $\hat{u}_i^{[t]}$ and $\hat{x}_i^{[t]}$ to each neighbor $j \in \mathcal{M}_i \setminus \{i\}$;

22    Receive $\hat{u}_j^{[t]}$ and $\hat{x}_j^{[t]}$ from $\forall j \in \mathcal{M}_i \setminus \{i\}$;

23    $u_i^{[t]} = \sum_{j \in \mathcal{M}_i} \omega_{i,j} \hat{u}_j^{[t]}$;

24    $x_i^{[t]} = \sum_{j \in \mathcal{M}_i} \omega_{i,j} \hat{x}_j^{[t]}$;

25   **end**

---

permutation $\phi_r$ of $\mathcal{M}_i$. Then, as shown in Line 5, we construct $\mathcal{Z}_j(\phi_r)$ for each agent $j \in \mathcal{M}_i$, i.e., the set of predecessors of $j$ in $\phi_r$, and finally compute $v\left(\mathcal{Z}_j\left(\phi_r\right)\bigcup\{j\}; \mathcal{Q}\right)$ and $v\left(\mathcal{Z}_j\left(\phi_r\right); \mathcal{Q}\right)$ using (16), and calculate $\varphi_{i,j}^{[t]}$ by

$$\varphi_{i,j}^{[t]} \leftarrow \varphi_{i,j}^{[t]} + \frac{v\left(\mathcal{Z}_j\left(\phi_r\right)\bigcup\{j\}; \mathcal{Q}\right) - v\left(\mathcal{Z}_j\left(\phi_r\right); \mathcal{Q}\right)}{R} \tag{26}$$

instead of (18) as illustrated in Line 6.

## V. THEORETICAL ANALYSIS

In this section, theoretical analysis is conducted to reveal the performance of our PDSL algorithm in terms of privacy preservation and convergence in Sec. V-A and Sec. V-B, respectively. All proofs of the **Theorems** 1∼2 and **Corollary** 1 are given in the appendices (see Appendix A and Appendix B).

### A. Privacy Preservation

The notion of DP implies that we have better privacy preservation and thus smaller privacy budget if stronger noise (with higher variance of the Gaussian noise in our case) is adopted. **Theorem** 1 reveals a lower bound of $\sigma$ (which determines the minimum magnitude of the Gaussian noise) to ensure $(\epsilon, \delta)$-DP.

**Algorithm 2:** Shapley value estimation (for agent $i$ in round $t$).

---

**Input:** Neighbors $\mathcal{M}_i$, local models $\left\{x_{i,j}^{[t]}\right\}_{j \in \mathcal{M}_i}$ and validation dataset $\mathcal{Q}$, and the maximum number of iterations $R$.

**Output:** Shapley value $\varphi_{i,j}^{[t]}$ of each agent $j \in \mathcal{M}_i$ in round $t$.

1   $\varphi_{i,j}^{[t]} = 0$ for $\forall j \in \mathcal{M}_i$;
2   **foreach** $r = 1, 2, \cdots, R$ **do**
3      Let $\phi_r$ be a random permutation of $\mathcal{M}_i$;
4      **foreach** $j \in \mathcal{M}_i$ **do**
5          $\mathcal{Z}_j(\phi_r) = \left\{j' \in \mathcal{M}_i \mid \phi_r(j') < \phi_r(j)\right\}$;
6          $\varphi_{i,j}^{[t]} \leftarrow \varphi_{i,j}^{[t]} + \frac{v(\mathcal{Z}_j(\phi_r) \bigcup \{j\}; \mathcal{Q}) - v(\mathcal{Z}_j(\phi_r); \mathcal{Q})}{R}$
7      **end**
8   **end**

---

**Theorem 1.** *For any $\epsilon > 0, \delta \in (0, 1)$, **Algorithm 1** can guarantee $(\epsilon, \delta)$-DP in each round, if*

$$\sigma \geq \max_{i \in \mathcal{M}} \left\{ \frac{2C \left( \frac{1}{\omega_{\min}} + \sum_{j \in \mathcal{M}_i} \frac{1}{\omega_{i,j}} \right) \sqrt{2 \ln(1.25/\delta)}}{\hat{\varphi}_{\min} \epsilon \sqrt{\sum_{j \in \mathcal{M}_i} \omega_{i,j}^{-2}}} \right\} \tag{27}$$

*where $C$ is the clipping threshold of any gradient, $\omega_{\min} = \min_{i \in \mathcal{M}, j \in \mathcal{M}_i} \omega_{i,j}$, and $\hat{\varphi}_{\min} = \min_{j \in \mathcal{M}_i, t \in T} \frac{\hat{\varphi}_{i,j}^{[t]}}{\sum_{k \in \mathcal{M}_i} \hat{\varphi}_{i,k}^{[t]}}$.*

### B. Convergence Analysis

Before presenting the main result regarding the convergence of our PDSL algorithm, we first introduce **Assumptions** 1~3 which are commonly employed in the analysis of distributed machine learning in existing state-of-the-art proposals [12], [23]. **Assumption** 1 indicates that the gradient variations of each function $f_i(x)$ are smooth, and $\mathcal{F}(x)$ is also $L$-smooth. **Assumption** 2 bounds the variance of stochastic gradient, and **Assumption** 3 assumes the adjacency matrix of the communication graph is doubly stochastic.

**Assumption 1.** *Each function $f_i(x)$ is L-smooth for any $i \in \mathcal{M}$, i.e., there exists a constant $L > 0$ such that*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^d \tag{28}$$

**Assumption 2.** *There exist $\zeta > 0$, $\kappa > 0$ such that*

$$\mathbb{E}_{\xi \sim \mathcal{D}_i} \left\| \nabla F_i(x; \xi) - \nabla f_i(x) \right\|^2 \leq \zeta^2, \quad \forall i \in \mathcal{M} \tag{29}$$

*and*

$$\left\| \nabla f_i(x) - \nabla \mathcal{F}(x) \right\|^2 \leq \kappa^2, \quad \forall i \in \mathcal{M} \tag{30}$$

**Assumption 3.** *Given the symmetric doubly stochastic matrix $\mathbf{W}$, let $\lambda_i(\mathbf{W})$ denote the $i$-th largest eigenvalue of $\mathbf{W}$. we have $\lambda_1(\mathbf{W}) = 1$ and there exists constant $\rho < 1$ such that $\max\{|\lambda_2(\mathbf{W})|, |\lambda_N(\mathbf{W})|\} \leq \sqrt{\rho}$.*

The main results about the convergence of PDSL is given in **Theorem** 2. In particular, the upper bound of the average gradient magnitude achieved by our PDSL algorithm mainly depends on the initial value of the objective function and the optimal value. In addition, the strength of the Gaussian noise, represented by $\sigma$, also exerts impact on the convergence of our PDSL algorithm. It is a minor polynomial factor and thus cannot disturb the convergence of our algorithm.

**Theorem 2.** *Suppose **Assumptions** 1~3 hold, and learning rate $\gamma$ satisfies the following condition*

$$\frac{(1-\alpha)^2}{\alpha} < \gamma \leq \min \left\{ \frac{(1-\alpha)(1-\sqrt{\rho})}{2\sqrt{26}L}, \alpha(1-\sqrt{\rho})^2 (2\sqrt{13}L)^{-2} \left( -1 + \sqrt{\frac{52L^2(1-\alpha)^2}{\alpha^2(1-\sqrt{\rho})^2} + 1} \right) \right\} \tag{31}$$

*For any $T \geq 1$, we have*

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[ \left\| \nabla \mathcal{F}\left( \bar{x}^{[t-1]} \right) \right\|^2 \right]$$

$$\leq \frac{1}{m_1 T} \left( \mathcal{F}\left( \bar{x}^{[0]} \right) - \mathcal{F}^* \right)$$

$$+ \left( m_2 + m_3 \frac{\gamma^2 \alpha^2}{(1-\alpha)^4} + m_4 \right) \left( \frac{4C^2}{\omega_{\min}^4} + \frac{4\sigma^2 d}{\omega_{\min}^4} + \frac{2\zeta^2}{M} \right)$$
$$+ m_5 \left( \frac{16\gamma^2 \left( C^2 + \sigma^2 d \right)}{\omega_{\min}^4 (1-\alpha)^2 (1-\sqrt{\rho})^2} + \frac{4\gamma^2 \left( 7\zeta^2 + 13\kappa^2 \right)}{(1-\alpha)^2 (1-\sqrt{\rho})^2} \right) \tag{32}$$

for **Algorithm** 1, where $\mathcal{F}^*$ denotes the optimal value of the objective function, and

$$\begin{cases} \bar{x}^{[t]} = \frac{1}{M} \sum_{i=1}^{M} x_i^{[t]} \\ \omega_{\min} = \min_{i \in \mathcal{M}, j \in \mathcal{M}_i} \omega_{i,j} \\ m_1 = \frac{\gamma}{2(1-\alpha)} - \frac{1-\alpha}{2\alpha} \\ m_2 = \frac{1}{m_1} \left( \frac{\alpha L \gamma^2}{2(1-\alpha)^3} + \frac{L\gamma^2}{2(1-\alpha)^2} \right) \\ m_3 = \frac{L(1-\alpha)}{2m_1 \alpha} \\ m_4 = \frac{\alpha \gamma^2}{2m_1 (1-\alpha)^3} \\ m_5 = \frac{L^2 \gamma}{2m_1 (1-\alpha)} \end{cases} \tag{33}$$

Then, **Corollary** 1 provides a more explicit demonstration of the convergence property of our PDSL algorithm.

**Corollary 1.** *Let learning rate $\gamma = O\left( \sqrt{\frac{1}{T}} \right)$. When $T$ is sufficiently large such that*

$$T \geq \max \left\{ \frac{104L^2}{(1-\alpha)^2(1-\sqrt{\rho})^2}, \frac{52^2 L^4}{\left( (1-\sqrt{\rho})\sqrt{52L^2(1-\alpha)^2 + \alpha^2(1-\sqrt{\rho})^2} - \alpha(1-\sqrt{\rho})^2 \right)^2} \right\} \tag{34}$$

*we have*

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla \mathcal{F} \left( \bar{x}^{[t-1]} \right) \right\|^2 \right] \leq K \left( \frac{1}{\sqrt{T}} + \frac{\sigma^2 d}{\sqrt{T}} + \frac{1}{M\sqrt{T}} + \frac{1}{T} + \frac{\sigma^2 d}{T} \right) \tag{35}$$

*for some constant $K > 0$.*

This corollary indicates that, when $T$ is sufficiently large, the convergence rate of our PDSL algorithm mainly depends on $\mathcal{O}\left( \frac{\sigma^2}{\sqrt{T}} \right)$. When 'stronger' Gaussian noise is adopted for higher differential privacy preservation, we have to be content with slower convergence. Furthermore, $\sigma$ is usually a constant and it is revealed that our PDSL algorithm has a linear speedup (with respect to $T$). To the best of our knowledge, this convergence rate is comparable to the well-known results in decentralized SGD algorithms (e.g., [12], [18]).

## VI. EXPERIMENTS

In this section, we evaluate the performance of our PDSL algorithm through extensive experiments. We introduce our experimental setup in Sec. VI-A, and then introduce the state-of-the-art reference algorithms in Sec. VI-B. We finally report and analyze the experiment results in Sec. VI-C.

### A. Datasets and Experiment Setup

We adopt two standard image datasets, i.e., MNIST and CIFAR-10, in our experiments. The MNIST dataset consists of $70,000$ handwritten digits from ten classes, with $60,000$ training images and $10,000$ test images. The CIFAR-10 dataset consists of $60,000$ images in ten classes, such as airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck, $50,000$ images are used for training and the remaining $10,000$ images are used for testing. For both MNIST and CIFAR-10 datasets, we randomly select $2,000$ images ($20\%$) from the original test dataset to serve as the global validation dataset.

We utilize two *Convolutional Neural Network* (CNN) models to assess the efficacy of our PDSL algorithm on MNIST and CIFAR-10 datasets, respectively. The CNN model for MNIST dataset is composed of two $3 \times 3$ convolutional layers, each succeeded by a $2 \times 2$ max pooling layer, then followed by a fully connected layer. The one for CIFAR-10 dataset incorporates two $5 \times 5$ convolutional layers, each succeeded by a $2 \times 2$ max pooling layer, and then two fully connected layers. We adopt ReLU activation functions in the convolutional layers in the above CNN models. The mini-batch size is set to be $250$ for both MNIST and CIFAR-10 datasets. For MNIST dataset, Let momentum coefficient $\alpha = 0.5$ and learning rate $\gamma = 0.001$. For CIFAR-10 dataset, Let $\alpha = 0.7$ and $\gamma = 0.01$.

To obtain heterogeneous data distribution across different agents, we adopt Dirichlet distribution $\text{Dir}(\mu p)$ to compute probability vectors for data labels, where $\mu$ is a parameter that influences the data distribution of agents, $p$ is a $Y$-dimensional

vector with all components are 1, and $Y$ denotes the number of different labels. As $\mu \to 0$, the data distribution across different agents tends to be more diverse. In our experiments, we let $\mu = 0.25$ for both MNIST and CIFAR-10 datasets.

Moreover, we study the performance of our PDSL algorithm with three different communication graph topologies in our experiments, i,e., fully connected graph, ring graph, and bipartite graph. The fully connected graph is a dense topology, while the bipartite and ring graphs are sparser topologies.

### B. Reference Algorithms

In this section, we compare our PDSL algorithm with the following state-of-the-art ones.

- **DP-DPSGD**: In [18], the notion of DP is leveraged to enable privacy preservation in decentralized learning. In the proposed algorithm, each agent uses perturbed gradient information to update its local model. Unfortunately, it does not take into account data heterogeneity.
- **MUFFLIATO**: MUFFLIATO is a privacy amplification mechanism, consisting of local Gaussian noise injection at the individual agents followed by gossiping for averaging the private values [19]. However, MUFFLIATO either does not consider the issue of data heterogeneity.
- **DP-CGA**: *Cross-Gradient Aggregation* (CGA) algorithm [12] is a decentralized learning algorithm with non-IID data distributions. We combine it with DP to ensure both learning efficacy of privacy preservation. Specifically, before sharing the cross-gradient information with its neighbors, each agent perturbs it with Gaussian noise.
- **DP-NET-FLEET**: In the *Decentralized Networked Federated Learning with Recursive Gradient Correction* (NET-FLEET) algorithm [14], a recursive gradient correction technique is adopted to handle heterogeneous data distribution across the agents, such that each agent runs multiple local updates between two consecutive communication rounds with their neighbors. In our experiments, we enable privacy preservation in NET-FLEET algorithm through Gaussian mechanism.

### C. Experiment Results

*1) Experiment Results on MNIST Dataset:* In Fig. 1, we show the average losses of the different algorithms with different numbers of agents $M = 10, 15, 20$ on fully connected communication graphs. We also vary privacy budget $\epsilon = 0.08, 0.1, 0.3$. The results indicate that our PDSL algorithm exhibits a faster convergence rate than the other four reference algorithms across the different privacy budgets. Specifically, our PDSL algorithm can reach its minimum loss value after about 80 rounds. Even with an increased number of agents, the convergence rate of our PDSL algorithm remains consistent, ensuring that 80 rounds are still sufficient for convergence. However, when the number of agents increases (e.g., $M = 15, 20$), the other four reference algorithms require more communication rounds to achieve convergence. Specifically, when $M = 20$, DP-NET-FLEET algorithm reaches convergence within about 100 rounds, while DP-DPSGD, DP-CGA and MUFFLIATO algorithms fail to guarantee convergence even after 180 communication rounds.

Moreover, as demonstrated in Fig. 1, compared to other reference algorithms, our PDSL algorithm achieves a lower average loss across the different values of $\epsilon$. For example, when $M = 10$ and $\epsilon = 0.3$, the average loss of PDSL is about 0.1 when the convergence is reached; it is 1.6 times lower than that of DP-NET-FLEET, 3.5 times lower than that of MUFFLIATO, 3.2 times lower than that of DP-CGA, and 5.8 times lower than that of DP-DPSGD. As the privacy budget $\epsilon$ decreases (e.g., $\epsilon = 0.1, 0.08$) such that more noise is injected, our PDSL algorithm still exhibits a lower average loss compared to other reference algorithms. For instance, when $M = 10, \epsilon = 0.08$, the final average loss of PDSL is 2.5 times lower than that of MUFFLIATO algorithm, 2 times lower than that of DP-CGA algorithm, and 3.5 times lower than that of DP-DPSGD algorithm. Additionally, it is noteworthy that the average loss of PDSL remains nearly unchanged with an increased number of agents (e.g., $M = 15, 20$) when the privacy budget $\epsilon$ remains constant, unlike the other reference algorithms which show a significant increase. For example, when $M = 20, \epsilon = 0.3$, the final average loss of our PDSL algorithm is $6 \sim 15$ times lower than the ones of the reference algorithms.

Fig. 2 and Fig. 3 show the average losses obtained on bipartite and ring graphs, respectively. Even on the sparser communication graphs, our PDSL algorithm has smaller average loss than the other reference algorithms and converges within about $100 \sim 110$ rounds across different values of $\epsilon$. In contrast, the DP-CGA, MUFFLIATO and DP-DPSGD algorithms require much more communication rounds to achieve convergence. Moreover, in comparison to PDSL, the DP-NET-FLEET algorithm demonstrates comparable convergence performance on the ring graph. However, as will be shown later, the test accuracy of DP-NET-FLEET is much smaller.

Table I presents the test accuracy of PDSL along with other reference algorithms across different values of $\epsilon$ on MNIST dataset. The results reveal that PDSL significantly outperforms the other reference algorithms. Take the performance improvement on the fully connected communication graph as an illustration. When $M = 10, \epsilon = 0.08$, the test accuracy of PDSL is 0.938; it is an improvement of 8.5% over DP-DPSGD, 5.7% over DP-CGA, 10% over MUFFLIATO, and 2.6% over DP-NET-FLEET, respectively. Moreover, as the number of agents increases, the test accuracy of PDSL remains relatively stable, in contrast to the significant decline observed in the other reference algorithms. Specifically, when $M = 20, \epsilon = 0.08$, the test accuracy of PDSL is 0.932, demonstrating an improvement of around $3\% \sim 26\%$ over the reference algorithms. Furthermore,
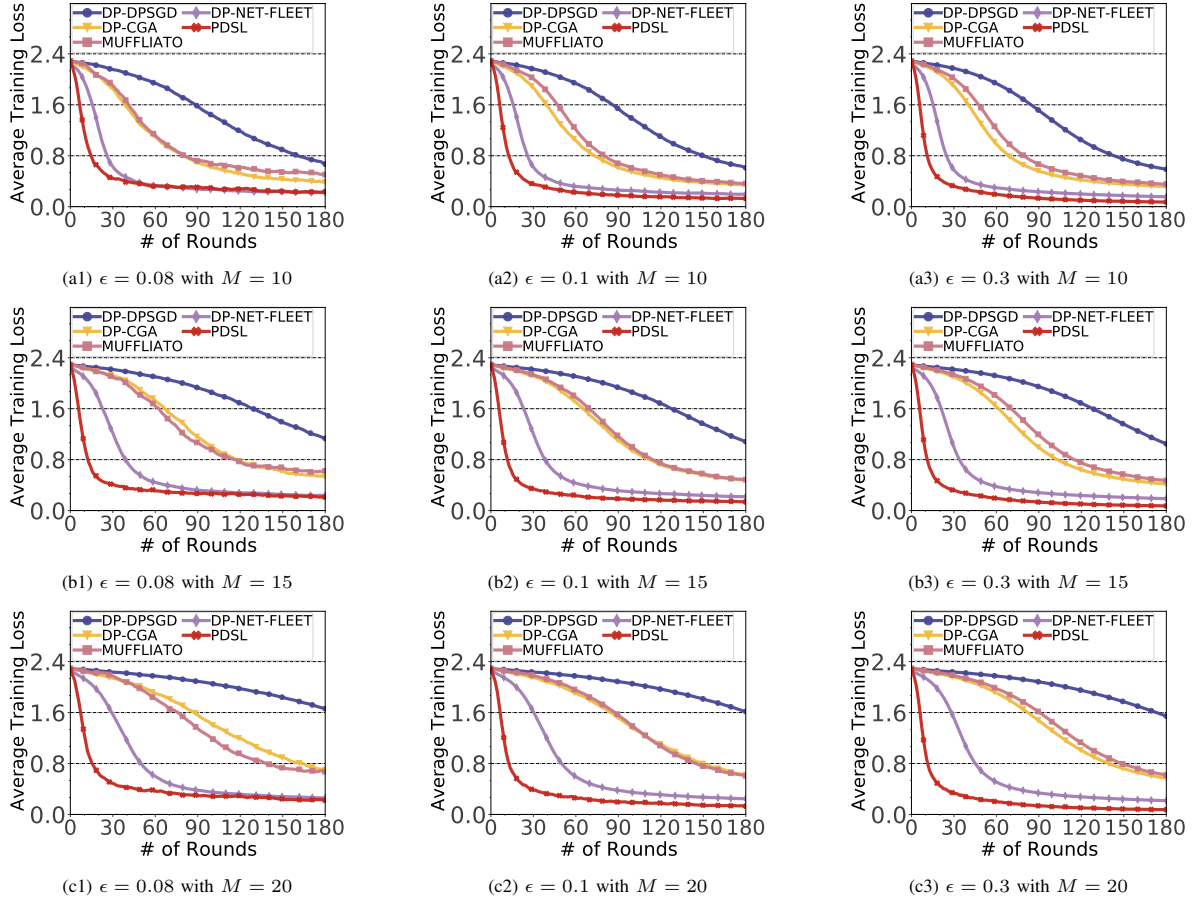
Fig. 1. Comparison results on MNIST dataset over fully connected graphs.

TABLE I
TEST ACCURACY ON MNIST DATASET.

| Privacy budget $\epsilon$ | Methods | Fully-connected | | | Bipartite | | | Ring | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $M=10$ | $M=15$ | $M=20$ | $M=10$ | $M=15$ | $M=20$ | $M=10$ | $M=15$ | $M=20$ |
| | DP-DPSGD | 0.853 | 0.768 | 0.668 | 0.822 | 0.739 | 0.655 | 0.817 | 0.731 | 0.604 |
| | DP-CGA | 0.881 | 0.843 | 0.817 | 0.865 | 0.854 | 0.779 | 0.857 | 0.799 | 0.715 |
| $\epsilon=0.08$ | MUFFLIATO | 0.838 | 0.802 | 0.788 | 0.833 | 0.815 | 0.782 | 0.825 | 0.805 | 0.762 |
| | DP-NET-FLEET | 0.912 | 0.907 | 0.902 | 0.892 | 0.891 | 0.872 | 0.872 | 0.868 | 0.842 |
| | **PDSL** | **0.938** | **0.934** | **0.932** | **0.925** | **0.923** | **0.917** | **0.884** | **0.875** | **0.849** |
| | DP-DPSGD | 0.865 | 0.794 | 0.705 | 0.838 | 0.761 | 0.675 | 0.833 | 0.753 | 0.642 |
| | DP-CGA | 0.902 | 0.869 | 0.851 | 0.887 | 0.869 | 0.813 | 0.881 | 0.829 | 0.769 |
| $\epsilon=0.1$ | MUFFLIATO | 0.887 | 0.879 | 0.846 | 0.885 | 0.872 | 0.841 | 0.887 | 0.868 | 0.835 |
| | DP-NET-FLEET | 0.925 | 0.921 | 0.914 | 0.906 | 0.903 | 0.887 | 0.893 | 0.896 | 0.858 |
| | **PDSL** | **0.957** | **0.952** | **0.955** | **0.946** | **0.949** | **0.943** | **0.913** | **0.907** | **0.871** |
| | DP-DPSGD | 0.875 | 0.817 | 0.723 | 0.851 | 0.781 | 0.693 | 0.847 | 0.775 | 0.653 |
| | DP-CGA | 0.915 | 0.903 | 0.866 | 0.904 | 0.882 | 0.852 | 0.898 | 0.865 | 0.793 |
| $\epsilon=0.3$ | MUFFLIATO | 0.906 | 0.889 | 0.857 | 0.899 | 0.878 | 0.848 | 0.892 | 0.874 | 0.841 |
| | DP-NET-FLEET | 0.938 | 0.929 | 0.922 | 0.927 | 0.917 | 0.903 | 0.912 | 0.905 | 0.875 |
| | **PDSL** | **0.965** | **0.963** | **0.962** | **0.965** | **0.963** | **0.961** | **0.936** | **0.918** | **0.892** |

as privacy budget $\epsilon$ is increased (e.g., $\epsilon = 0.1$, 0.3), the test accuracy of PDSL increases significantly. Table I also reveals that the test accuracy of PDSL is better than the other reference algorithms on the different communication graphs. Especially when $M = 20, \epsilon = 0.3$, the test accuracy of PDSL is 1.38 times higher than DP-DPSGD, 1.12 times higher than DP-CGA, 1.13 times higher than MUFFLIATO, and 1.06 times higher than DP-NET-FLEET on bipartite graphs; On the ring graph, the test accuracy of PDSL is around $1.02 \sim 1.36$ times higher than the those of the reference algorithms.
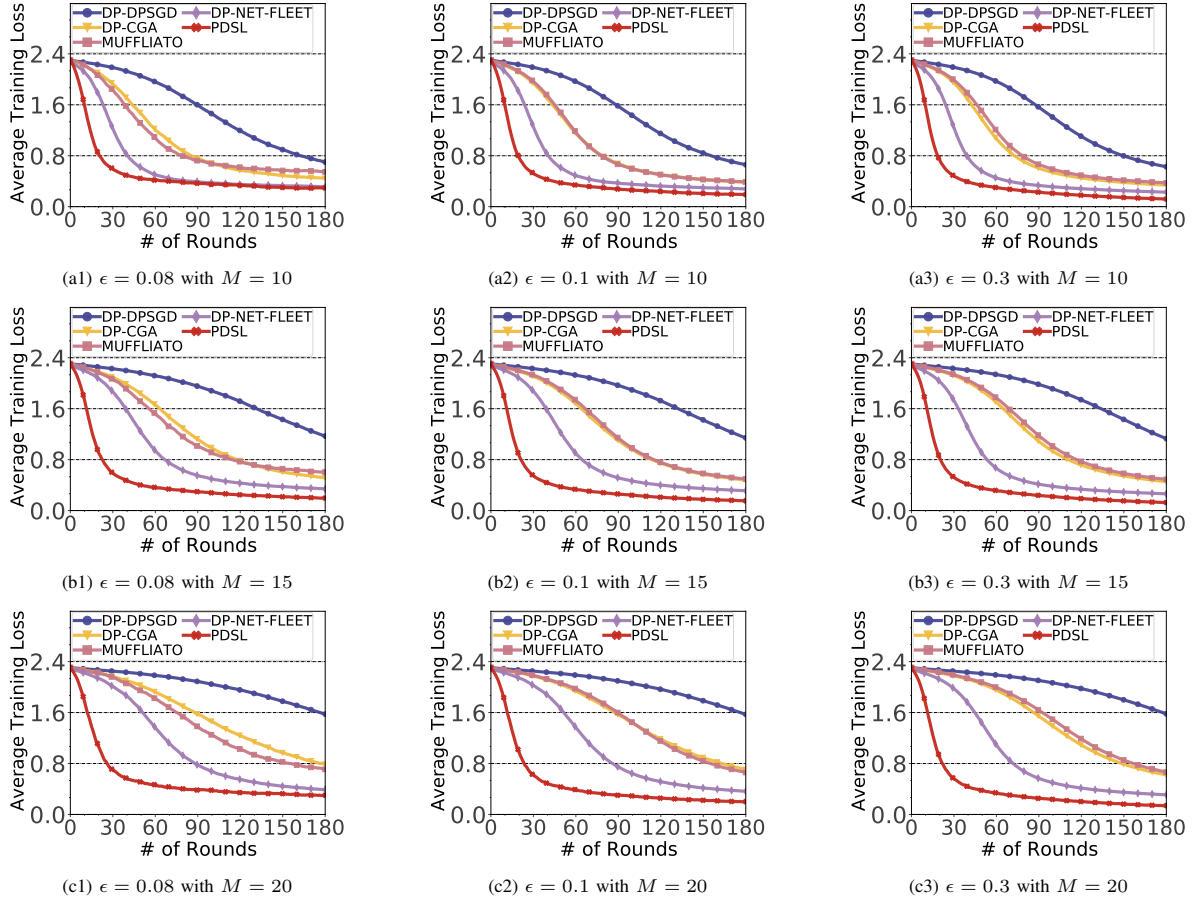
(a1) $\epsilon = 0.08$ with $M = 10$     (a2) $\epsilon = 0.1$ with $M = 10$     (a3) $\epsilon = 0.3$ with $M = 10$

(b1) $\epsilon = 0.08$ with $M = 15$     (b2) $\epsilon = 0.1$ with $M = 15$     (b3) $\epsilon = 0.3$ with $M = 15$

(c1) $\epsilon = 0.08$ with $M = 20$     (c2) $\epsilon = 0.1$ with $M = 20$     (c3) $\epsilon = 0.3$ with $M = 20$

Fig. 2. Comparison results on MNIST dataset over bipartite graphs.

*2) Experiment Results on CIFAR-10 Dataset:* The results of average losses achieved by our PDSL algorithm, as well as the other four reference algorithms across different privacy budget over the three different communication topologies are presented in Fig. 4, Fig. 5 and Fig. 6, respectively. The results indicate that, our PDSL algorithm significantly outperforms DP-DPSGD, DP-CGA and MUFFLIATO algorithms with different privacy budget on both fully connected graphs and bipartite graphs. For example, when $\epsilon = 1, M = 10$, the average loss of PDSL after 200 communication rounds is around 0.65 on the fully connected graph; it is similar to the average loss of DP-NET-FLEET, and 2.2 times smaller than DP-DPSGD, 2.1 times smaller than MUFFLIATO, and 1.67 times smaller than DP-CGA. Furthermore, we can see that PDSL still maintains better average loss than other reference algorithms as the number of agents increases. Specifically, when $\epsilon = 1$ and $M$ is increased to 20, the average loss of PDSL is 2.34 times smaller than DP-DPSGD, 1.74 times smaller than DP-CGA, 2.2 times smaller than MUFFLIATO, while 1.23 times smaller than DP-NET-FLEET. Moreover, compared to PDSL, DP-NET-FLEET exhibits equivalent or even superior convergence properties, as evidenced in Fig. 6. However, the test accuracy for DP-NET-FLEET is lower than that achieved by our PDSL algorithm, a detail that will be further elaborated upon subsequently.

We also compare the test accuracy of the different algorithms with respect to the number of agents and the different values of $\epsilon$ over the three different communication topologies. The results are presented in Table II, from which we can observe that PDSL performs better in test accuracy compared to the other reference algorithms. For instance, when $\epsilon = 1, M = 10$, the test accuracy of our PDSL algorithm on the fully connected graph is 0.758; it is an improvement of 27.3% over DP-DPSGD, 16.2% over DP-CGA, 23% over MUFFLIATO, and 20.7% over DP-NET-FLEET. Additionally, as the privacy budget $\epsilon$ becomes smaller or the number of agents increases, PDSL still maintains superior performance over the other reference algorithms in terms of test accuracy. Specifically, when $\epsilon = 1$ and $M = 20$, our PDSL algorithm has an improvement of around $21\% \sim 30\%$ over the four reference algorithms. When $\epsilon = 0.5, M = 10$, the test accuracy of PDSL reaches 0.691; that is an improvement of $15\% \sim 31\%$ over the reference algorithms. Moreover, our PDSL algorithm still has better test accuracy with sparser communication graphs. For example, when $\epsilon = 1, M = 20$, the test accuracy of PDSL algorithm is around 0.712 on bipartite graphs; it is 1.68 times higher than DP-DPSGD, 1.55 times higher than DP-CGA, 1.52 times higher than
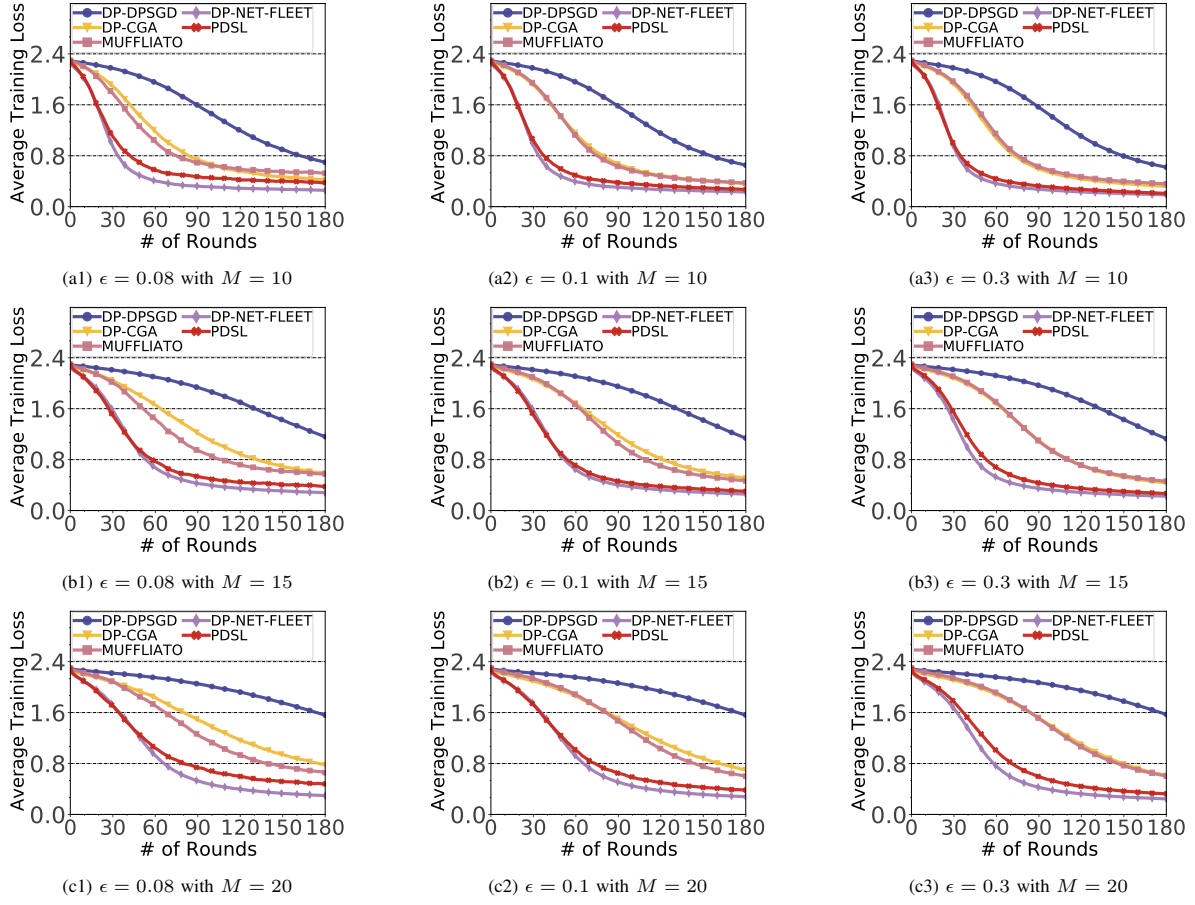
Fig. 3. Comparison results on MNIST dataset over ring graphs.

TABLE II
TEST ACCURACY ON CIFAR-10 DATASET.

| Privacy budget $\epsilon$ | Methods | Fully-connected | | | Bipartite | | | Ring | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $M=10$ | $M=15$ | $M=20$ | $M=10$ | $M=15$ | $M=20$ | $M=10$ | $M=15$ | $M=20$ |
| $\epsilon=0.5$ | DP-DPSGD | 0.453 | 0.451 | 0.423 | 0.441 | 0.443 | 0.412 | 0.374 | 0.388 | 0.326 |
| | DP-CGA | 0.375 | 0.369 | 0.325 | 0.395 | 0.415 | 0.383 | 0.385 | 0.391 | 0.356 |
| | MUFFLIATO | 0.484 | 0.459 | 0.454 | 0.481 | 0.462 | 0.445 | 0.436 | 0.449 | 0.417 |
| | DP-NET-FLEET | 0.541 | 0.524 | 0.502 | 0.455 | 0.458 | 0.408 | 0.483 | 0.471 | 0.439 |
| | **PDSL** | **0.691** | **0.688** | **0.686** | **0.668** | **0.673** | **0.654** | **0.554** | **0.559** | **0.517** |
| $\epsilon=0.7$ | DP-DPSGD | 0.464 | 0.459 | 0.433 | 0.453 | 0.449 | 0.413 | 0.385 | 0.385 | 0.335 |
| | DP-CGA | 0.427 | 0.412 | 0.398 | 0.421 | 0.434 | 0.406 | 0.408 | 0.424 | 0.386 |
| | MUFFLIATO | 0.505 | 0.481 | 0.466 | 0.503 | 0.476 | 0.457 | 0.447 | 0.458 | 0.422 |
| | DP-NET-FLEET | 0.548 | 0.531 | 0.508 | 0.465 | 0.462 | 0.411 | 0.496 | 0.477 | 0.445 |
| | **PDSL** | **0.744** | **0.738** | **0.734** | **0.718** | **0.715** | **0.691** | **0.569** | **0.573** | **0.524** |
| $\epsilon=1$ | DP-DPSGD | 0.485 | 0.463 | 0.447 | 0.461 | 0.455 | 0.422 | 0.403 | 0.383 | 0.353 |
| | DP-CGA | 0.596 | 0.602 | 0.537 | 0.475 | 0.484 | 0.459 | 0.455 | 0.467 | 0.433 |
| | MUFFLIATO | 0.528 | 0.498 | 0.476 | 0.525 | 0.492 | 0.467 | 0.457 | 0.463 | 0.427 |
| | DP-NET-FLEET | 0.551 | 0.539 | 0.512 | 0.473 | 0.466 | 0.416 | 0.501 | 0.483 | 0.452 |
| | **PDSL** | **0.758** | **0.753** | **0.746** | **0.725** | **0.731** | **0.712** | **0.573** | **0.582** | **0.528** |

MUFFLIATO, and 1.71 times higher than DP-NET-FLEET. If ring graphs are adopted, the test accuracy of PDSL algorithm is 1.49 times higher than DP-DPSGD, 1.21 times higher than DP-CGA, 1.23 times higher than MUFFLIATO, and 1.16 times higher than DP-NET-FLEET.
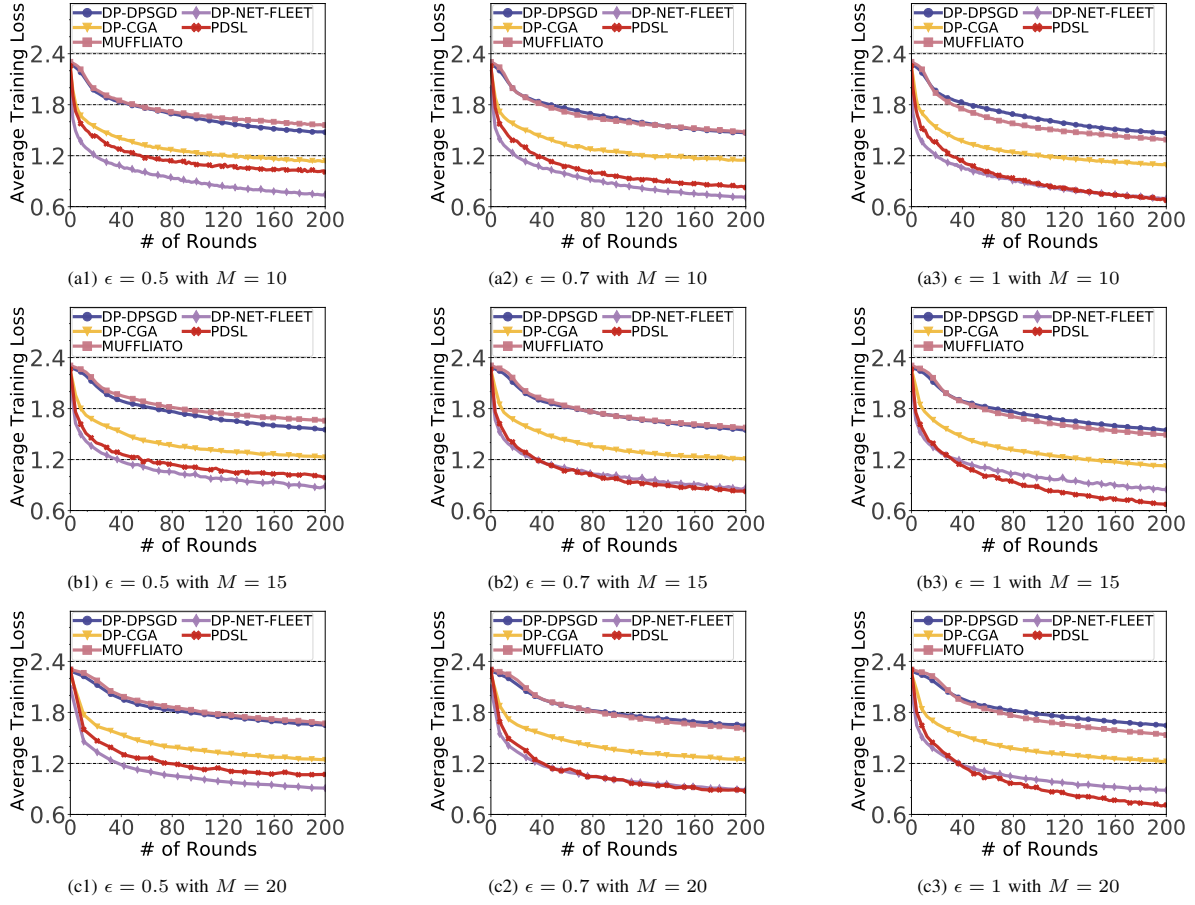
Fig. 4. Comparison results on CIFAR-10 dataset over fully connected graphs.

## VII. CONCLUSION

In this paper, we propose PDSL, a novel privacy-preserved decentralized stochastic learning algorithm with heterogeneous data distribution. In particular, on one hand, we leverage the notion of DP to perturb the cross-gradient information; one the other hand, we utilize the concept of Shapley value to measure the contributions of the perturbed cross-gradient information at each agent for calculating the weighted aggregation of the perturbed gradients. We conduct solid theoretical analysis to reveal the privacy of each agent can be preserved effectively during the decentralized learning process, and the convergence of our algorithm can be ensured with a linear speedup. The efficacy of our algorithm is also verified by our extensive experiments on two real-world datasets.

## REFERENCES

[1] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J. Rellermeyer, "A Survey on Distributed Machine Learning," *ACM Computing Surveys*, vol. 53, no. 2, pp. 30:1–30:33, 2021.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. of the 20th AISTATS*, 2017, pp. 1273–1282.

[3] A. Shamsabadi, A. Gascón, H. Haddadi, and A. Cavallaro, "PrivEdge: From Local to Distributed Private Training and Prediction," *IEEE Trans. on Information Forensics and Security*, vol. 15, pp. 3819–3831, 2020.

[4] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated Machine Learning: Concept and Applications," *ACM Trans. on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–19, 2019.

[5] J. Jin, J. Ren, Y. Zhou, L. Lyu, J. Liu, and D. Dou, "Accelerated Federated Learning with Decoupled Adaptive Optimization," in *Proc. of the 39th ICML*, 2022, pp. 10 298–10 322.

[6] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated Learning on Non-iid Data Silos: An Experimental Study," in *Proc. of the 38th IEEE ICDE*, 2022, pp. 965–978.

[7] T. Li, A. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated Optimization in Heterogeneous Networks," in *Proc. of the MLSys*, 2020, pp. 429–450.

[8] M. Luo, F. Chen, D. Hu, Y. Zhang, J. Liang, and J. Feng, "No Fear of Heterogeneity: Classifier Calibration for Federated Learning with Non-iid Data," in *Proc. of the 35th NeurIPS*, 2021, pp. 5972–5984.

[9] Q. Sun, X. Li, J. Zhang, L. Xiong, W. Liu, J. Liu, Z. Qin, and K. Ren, "Shapleyfl: Robust Federated Learning based on Shapley Value," in *Proc. of the 29th ACM SIGKDD*, 2023, pp. 2096–2108.

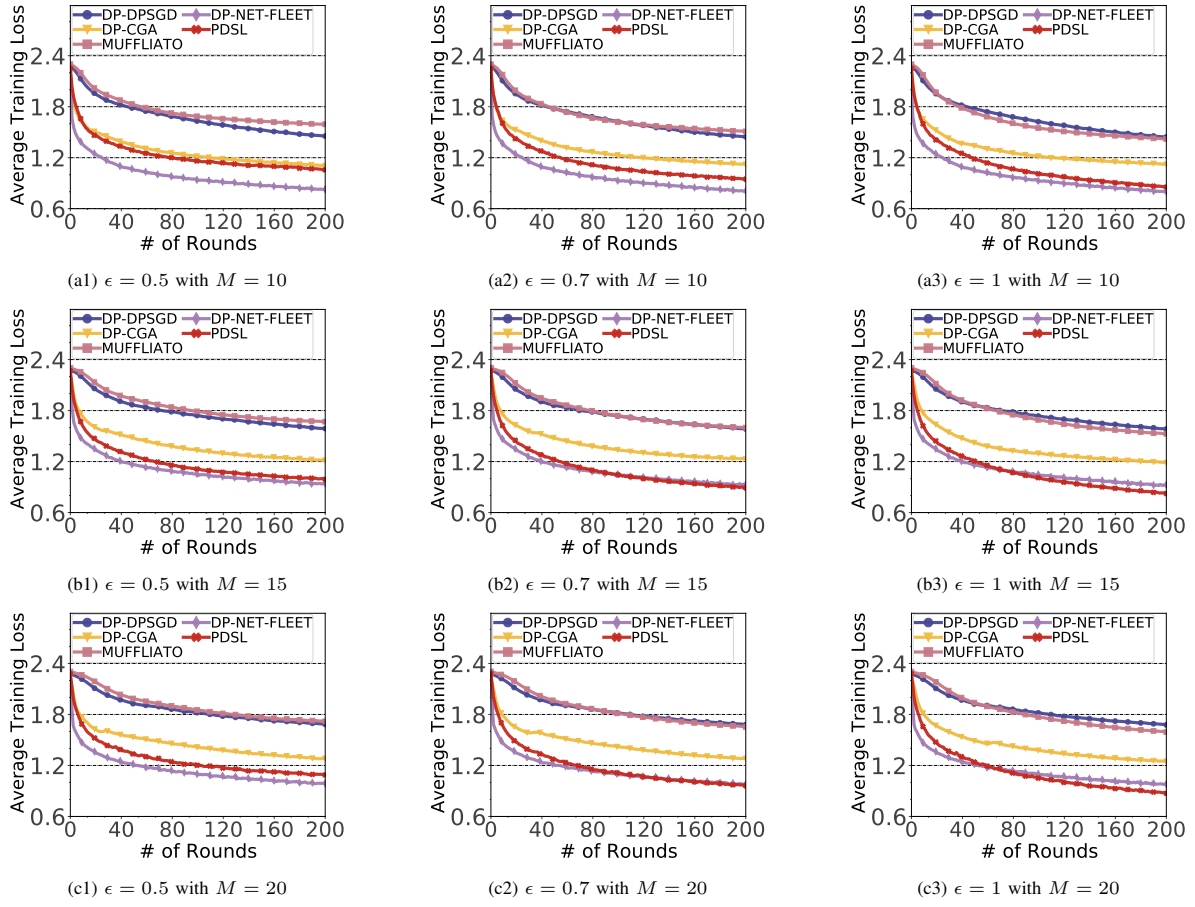Fig. 5. Comparison results on CIFAR-10 dataset over bipartite graphs.

[10] Y. Tan, Y. Liu, G. Long, J. Jiang, Q. Lu, and C. Zhang, "Federated Learning on Non-iid Graphs via Structural Knowledge Sharing," in *Proc. of the 37th AAAI*, 2023, pp. 9953–9961.

[11] S. Chen and B. Li, "Towards Optimal Multi-modal Federated Learning on Non-IID Data with Hierarchical Gradient Blending," in *Proc. of the 41st IEEE INFOCOM*, 2022, pp. 1469–1478.

[12] Y. Esfandiari, S. Tan, Z. Jiang, A. Balu, E. Herron, C. Hegde, and S. Sarkar, "Cross-Gradient Aggregation for Decentralized Learning from Non-IID Data," in *Proc. of the 38th ICML*, 2021, pp. 3036–3046.

[13] S. Aketi, S. Kodge, and K. Roy, "Neighborhood Gradient Mean: An Efficient Decentralized Learning Method for Non-IID Data Distributions," *Trans. on Machine Learning Research*, 2023.

[14] X. Zhang, M. Fang, Z. Liu, H. Yang, J. Liu, and Z. Zhu, "Net-Fleet: Achieving Linear Convergence Speedup for Fully Decentralized Federated Learning with Heterogeneous Data," in *Proc. of the 23rd ACM MobiHoc*, 2022, pp. 71–80.

[15] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in *Proc. of the 38th IEEE S&P*, 2017, pp. 3–18.

[16] M. Fredrikson, S. Jha, and T. Ristenpart, "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures," in *Proc. of the 22nd ACM CCS*, 2015, pp. 1322–1333.

[17] L. Zhu, Z. Liu, and S. Han, "Deep Leakage from Gradients," in *Proc. of the 33rd NeurIPS*, 2019.

[18] J. Xu, W. Zhang, and F. Wang, "A(DP)$^2$SGD: Asynchronous Decentralized Parallel Stochastic Gradient Descent With Differential Privacy," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8036–8047, 2022.

[19] E. Cyffers, M. Even, A. Bellet, and L. Massoulié, "Muffliato: Peer-to-Peer Privacy Amplification for Decentralized Optimization and Averaging," in *Proc. of the 36th NeurIPS*, 2022, pp. 15 889–15 902.

[20] X. Lian, C. Zhang, H. Zhang, C. Hsieh, W. Zhang, and J. Liu, "Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent," in *Proc. of the 31st NeurIPS*, 2017, pp. 5330–534.

[21] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized Gossip Algorithms," *IEEE Trans. on Information Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[22] K. Scaman, F. Bach, S. Bubeck, L. Massoulié, and Y. Lee, "Optimal Algorithms for Non-Smooth Distributed Optimization in Networks," in *Proc. of the 32nd NeurIPS*, 2018.

[23] H. Yu, R. Jin, and S. Yang, "On the Linear Speedup Analysis of Communication Efficient Momentum SGD for Distributed Non-Convex Optimization," in *Proc. of the 36th ICML*, 2019, pp. 7184–7193.

[24] H. Ye, L. Liang, and G. Li, "Decentralized Federated Learning with Unreliable Communications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 3, pp. 487–500, 2022.

[25] T. Lin, S. Karimireddy, S. Stich, and M. Jaggi, "Quasi-Global Momentum: Accelerating Decentralized Deep Learning on Heterogeneous Data," in *Proc. of the 38th ICML*, 2021.
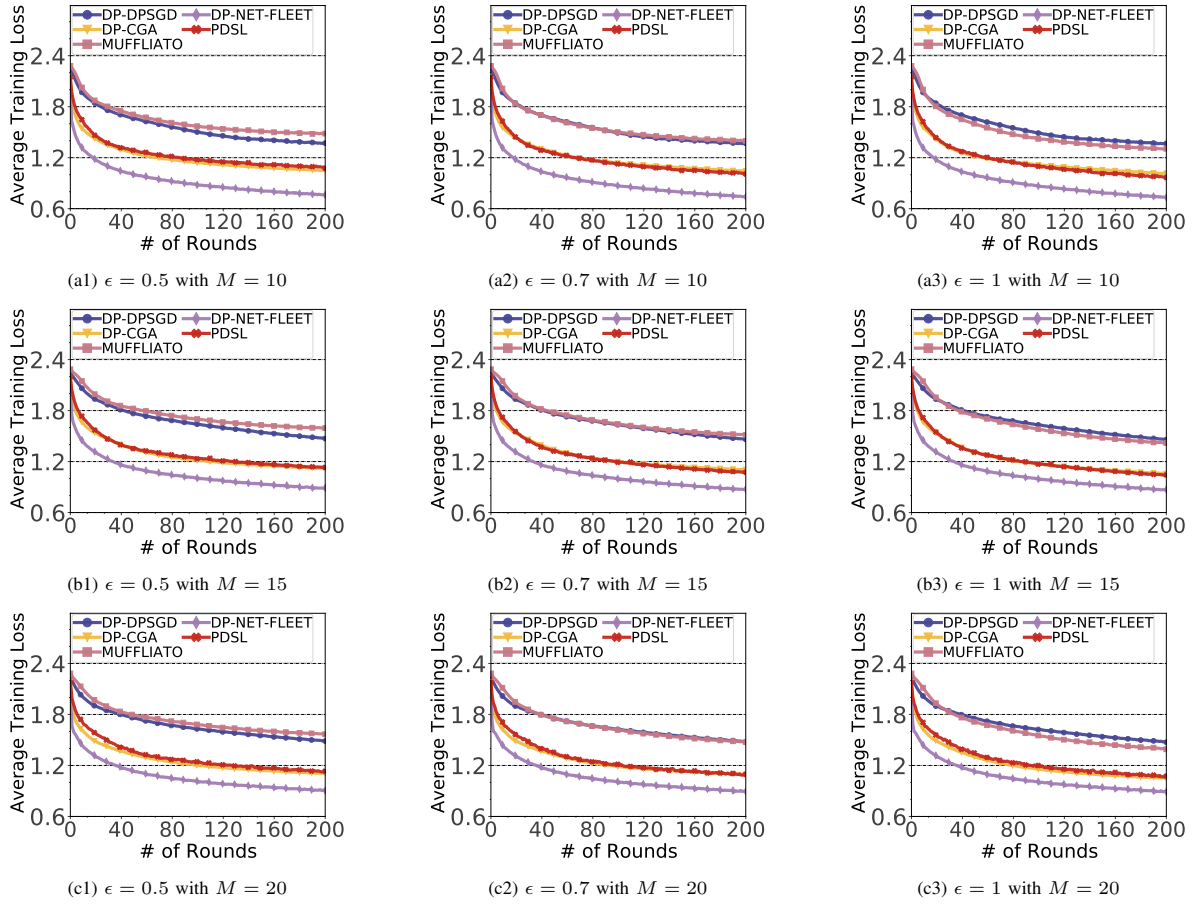
(a1) $\epsilon = 0.5$ with $M = 10$     (a2) $\epsilon = 0.7$ with $M = 10$     (a3) $\epsilon = 1$ with $M = 10$

(b1) $\epsilon = 0.5$ with $M = 15$     (b2) $\epsilon = 0.7$ with $M = 15$     (b3) $\epsilon = 1$ with $M = 15$

(c1) $\epsilon = 0.5$ with $M = 20$     (c2) $\epsilon = 0.7$ with $M = 20$     (c3) $\epsilon = 1$ with $M = 20$

Fig. 6. Comparison results on CIFAR-10 dataset over ring graphs.

[26] C. Li, G. Li, and P. Varshney, "Decentralized Federated Learning via Mutual Knowledge Transfer," *IEEE Internet of Things Journal*, vol. 9, no. 2, pp. 1136–1147, 2021.

[27] Y. Shi, L. Shen, K. Wei, Y. Sun, B. Yuan, X. Wang, and D. Tao, "Improving the Model Consistency of Decentralized Federated Learning," in *Proc. of the 40th ICML*, 2023, pp. 31 269–31 291.

[28] S. Pu and A. Nedić, "Distributed Stochastic Gradient Tracking Methods," *Mathematical Programming*, vol. 187, no. 1, pp. 409–457, 2021.

[29] Y. Takezawa, H. Bao, K. Niwa, R. Sato, and M. Yamada, "Momentum Tracking: Momentum Acceleration for Decentralized Deep Learning on Heterogeneous Data," *Trans. on Machine Learning Research*, vol. 2023, 2023.

[30] S. Aketi, A. Hashemi, and K. Roy, "Global Update Tracking: A Decentralized Learning Algorithm for Heterogeneous Data," in *Proc. of the 37th NeurIPS*, 2024.

[31] L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov, "Exploiting Unintended Feature Leakage in Collaborative Learning," in *Proc. of the 40th IEEE S&P*, 2019, pp. 691–706.

[32] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.

[33] L. Chen, X. Ding, Z. Bao, P. Zhou, and H. Jin, "Differentially Private Federated Learning on Non-iid Data: Convergence Analysis and Adaptive Optimization," *IEEE Trans. on Knowledge and Data Engineering*, vol. 36, no. 9, pp. 4567–4581, 2024.

[34] X. Lin, J. Wu, J. Li, C. Sang, S. Hu, and M. J. Deen, "Heterogeneous Differential-Private Federated Learning: Trading Privacy for Utility Truthfully," *IEEE Trans. on Dependable and Secure Computing*, vol. 20, no. 6, pp. 5113–5129, 2023.

[35] Z. Huang and Y. Gong, "Differentially Private ADMM for Convex Distributed Learning: Improved Accuracy via Multi-step Approximation," *arXiv preprint arXiv:2005.07890*, 2020.

[36] L. Shapley, "Cores of Convex Games," *International journal of game theory*, vol. 1, pp. 11–26, 1971.

[37] J. Castro, D. Gómez, and J. Tejada, "Polynomial Calculation of The Shapley Value Based on Sampling," *Computers & Operations Research*, vol. 36, no. 5, pp. 1726–1730, 2009.

[38] S. Fatima, M. Wooldridge, and N. Jennings, "A Linear Approximation Method for The Shapley Value," *Artificial Intelligence*, vol. 172, no. 14, pp. 1673–1699, 2008.

In **Algorithm** 1, each agent aggregates the perturbed stochastic gradients from its neighbors (see Lines 13~18). In any round $t$, the gradient obtained by agent $i$ is

$$
\begin{aligned}
\bar{g}_i^{[t]} &= \sum_{j\in\mathcal{M}_i} \frac{\hat{\varphi}_{i,j}^{[t]}}{\omega_{i,j}\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}} \left( \frac{g_{j,i}^{[t]}}{\max\left\{1,C^{-1}\left\|g_{j,i}^{[t]}\right\|\right\}} + n_{j,i}^{[t]} \right) \\
&= \sum_{j\in\mathcal{M}_i} \frac{\hat{\varphi}_{i,j}^{[t]}}{\omega_{i,j}\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}} \cdot \frac{g_{j,i}^{[t]}}{\max\left\{1,C^{-1}\left\|g_{j,i}^{[t]}\right\|\right\}} + \sum_{j\in\mathcal{M}_i} \frac{\hat{\varphi}_{i,j}^{[t]}}{\omega_{i,j}\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}} n_{j,i}^{[t]} \\
&\geq \sum_{j\in\mathcal{M}_i} \frac{\hat{\varphi}_{i,j}^{[t]}}{\omega_{i,j}\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}} \cdot \frac{g_{j,i}^{[t]}}{\max\left\{1,C^{-1}\left\|g_{j,i}^{[t]}\right\|\right\}} + \sum_{j\in\mathcal{M}_i} \frac{\hat{\varphi}_{\min}}{\omega_{i,j}} n_{j,i}^{[t]}
\end{aligned} \tag{36}
$$

where $\hat{\varphi}_{\min} = \min_{j\in\mathcal{M}_i,t\in T} \frac{\hat{\varphi}_{i,j}^{[t]}}{\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}}$. For the second term in (36), we have

$$
\mathbb{E}\left[\sum_{j\in\mathcal{M}_i}\frac{\hat{\varphi}_{\min}}{\omega_{i,j}}n_{j,i}^{[t]}\right] = \sum_{j\in\mathcal{M}_i}\frac{\hat{\varphi}_{\min}}{\omega_{i,j}}\mathbb{E}\left[n_{j,i}^{[t]}\right] = 0 \tag{37}
$$

and

$$
\mathrm{Var}\left(\sum_{j\in\mathcal{M}_i}\frac{\hat{\varphi}_{\min}}{\omega_{i,j}}n_{j,i}^{[t]}\right) = (\hat{\varphi}_{\min})^2 \sum_{j\in\mathcal{M}_i}\frac{1}{\omega_{i,j}^2}\mathrm{Var}\left(n_{j,i}^{[t]}\right) = \sum_{j\in\mathcal{M}_i}\frac{\sigma^2 d\,(\hat{\varphi}_{\min})^2}{\omega_{i,j}^2} \tag{38}
$$

Therefore, $\sum_{j\in\mathcal{M}_i}\frac{\hat{\varphi}_{\min}}{\omega_{i,j}}n_{j,i}^{[t]} \sim \mathcal{N}\left(0,\sum_{j\in\mathcal{M}_i}\frac{(\hat{\varphi}_{\min})^2}{\omega_{i,j}^2}\sigma^2\mathbf{I}_d\right)$.

We then bound the $L2$-sensitivity of $q^{[t]} = \sum_{j\in\mathcal{M}_i}\frac{\hat{\varphi}_{i,j}^{[t]}}{\omega_{i,j}\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}}\cdot\frac{g_{j,i}^{[t]}}{\max\left\{1,C^{-1}\left\|g_{j,i}^{[t]}\right\|\right\}}$. For agent $i$, consider any two adjacent datasets $\mathcal{D}_i$ and $\mathcal{D}_i'$. We let the datasets of its neighbors be fixed and thus have

$$
\triangle_2 q^{[t]} = \max_{\mathcal{D}_i,\mathcal{D}_i'}\left\|q^{[t]}(\mathcal{D}_i) - q^{[t]}(\mathcal{D}_i')\right\|
$$

$$
= \max_{\mathcal{D}_i,\mathcal{D}_i'}\left\|\begin{aligned}
&\frac{\hat{\varphi}_{i,i}^{[t]}}{\omega_{i,i}\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}}\cdot\frac{g_{i,i}^{[t]}(\mathcal{D}_i)}{\max\left\{1,C^{-1}\left\|g_{i,i}^{[t]}(\mathcal{D}_i)\right\|\right\}} \\
&+ \sum_{j\in\mathcal{M}_i,j\neq i}\frac{\hat{\varphi}_{i,j}^{[t]}}{\omega_{i,j}\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}}\cdot\frac{g_{j,i}^{[t]}(\mathcal{D}_j)}{\max\left\{1,C^{-1}\left\|g_{j,i}^{[t]}(\mathcal{D}_j)\right\|\right\}} \\
&- \frac{(\hat{\varphi}_{i,i}^{[t]})^-}{\omega_{i,i}\sum_{k\in\mathcal{M}_i}(\hat{\varphi}_{i,k}^{[t]})^-}\cdot\frac{g_{i,i}^{[t]}(\mathcal{D}_i')}{\max\left\{1,C^{-1}\left\|g_{i,i}^{[t]}(\mathcal{D}_i')\right\|\right\}} \\
&- \sum_{j\in\mathcal{M}_i,j\neq i}\frac{(\hat{\varphi}_{i,j}^{[t]})^-}{\omega_{i,j}\sum_{k\in\mathcal{M}_i}(\hat{\varphi}_{i,k}^{[t]})^-}\cdot\frac{g_{j,i}^{[t]}(\mathcal{D}_j)}{\max\left\{1,C^{-1}\left\|g_{j,i}^{[t]}(\mathcal{D}_j)\right\|\right\}}
\end{aligned}\right\|
$$

$$
= \max_{\mathcal{D}_i,\mathcal{D}_i'}\left\|\begin{aligned}
&\frac{\hat{\varphi}_{i,i}^{[t]}}{\omega_{i,i}\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}}\left(\frac{g_{i,i}^{[t]}(\mathcal{D}_i)}{\max\left\{1,C^{-1}\left\|g_{i,i}^{[t]}(\mathcal{D}_i)\right\|\right\}} - \frac{g_{i,i}^{[t]}(\mathcal{D}_i')}{\max\left\{1,C^{-1}\left\|g_{i,i}^{[t]}(\mathcal{D}_i')\right\|\right\}}\right) \\
&+ \left(\frac{\hat{\varphi}_{i,i}^{[t]}}{\omega_{i,i}\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}} - \frac{(\hat{\varphi}_{i,i}^{[t]})^-}{\omega_{i,i}\sum_{k\in\mathcal{M}_i}(\hat{\varphi}_{i,k}^{[t]})^-}\right)\frac{g_{i,i}^{[t]}(\mathcal{D}_i')}{\max\left\{1,C^{-1}\left\|g_{i,i}^{[t]}(\mathcal{D}_i')\right\|\right\}} \\
&+ \sum_{j\in\mathcal{M}_i,j\neq i}\left(\frac{\hat{\varphi}_{i,j}^{[t]}}{\omega_{i,j}\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}} - \frac{(\hat{\varphi}_{i,j}^{[t]})^-}{\omega_{i,j}\sum_{k\in\mathcal{M}_i}(\hat{\varphi}_{i,k}^{[t]})^-}\right)\frac{g_{j,i}^{[t]}(\mathcal{D}_j)}{\max\left\{1,C^{-1}\left\|g_{j,i}^{[t]}(\mathcal{D}_j)\right\|\right\}}
\end{aligned}\right\|
$$

$$
\begin{aligned}
\leq & \max_{\mathcal{D}_i,\mathcal{D}_i'} \left\| \frac{\hat{\varphi}_{i,i}^{[t]}}{\omega_{i,i} \sum_{k\in\mathcal{M}_i} \hat{\varphi}_{i,k}^{[t]}} \left( \frac{g_{i,i}^{[t]}(\mathcal{D}_i)}{\max\left\{1, C^{-1}\left\|g_{i,i}^{[t]}(\mathcal{D}_i)\right\|\right\}} - \frac{g_{i,i}^{[t]}(\mathcal{D}_i')}{\max\left\{1, C^{-1}\left\|g_{i,i}^{[t]}(\mathcal{D}_i')\right\|\right\}} \right) \right\| \\
& + \max_{\mathcal{D}_i,\mathcal{D}_i'} \left\| \left( \frac{\hat{\varphi}_{i,i}^{[t]}}{\omega_{i,i} \sum_{k\in\mathcal{M}_i} \hat{\varphi}_{i,k}^{[t]}} - \frac{(\hat{\varphi}_{i,i}^{[t]})^-}{\omega_{i,i} \sum_{k\in\mathcal{M}_i} (\hat{\varphi}_{i,k}^{[t]})^-} \right) \frac{g_{i,i}^{[t]}(\mathcal{D}_i')}{\max\left\{1, C^{-1}\left\|g_{i,i}^{[t]}(\mathcal{D}_i')\right\|\right\}} \right\| \\
& + \max_{\mathcal{D}_i,\mathcal{D}_i'} \left\| \sum_{j\in\mathcal{M}_i, j\neq i} \left( \frac{\hat{\varphi}_{i,j}^{[t]}}{\omega_{i,j} \sum_{k\in\mathcal{M}_i} \hat{\varphi}_{i,k}^{[t]}} - \frac{(\hat{\varphi}_{i,j}^{[t]})^-}{\omega_{i,j} \sum_{k\in\mathcal{M}_i} (\hat{\varphi}_{i,k}^{[t]})^-} \right) \frac{g_{j,i}^{[t]}(\mathcal{D}_j)}{\max\left\{1, C^{-1}\left\|g_{j,i}^{[t]}(\mathcal{D}_j)\right\|\right\}} \right\| \\
\leq & \frac{2C}{\omega_{i,i}} + \sum_{j\in\mathcal{M}_i} \frac{2C}{\omega_{i,j}}
\end{aligned}
\tag{39}
$$

where $g_{i,i}^{[t]}(\mathcal{D}_i)$ and $g_{i,i}^{[t]}(\mathcal{D}_i')$ represent the gradients corresponding to the adjacent datasets $\mathcal{D}_i$ and $\mathcal{D}_i'$, respectively. $g_{j,i}^{[t]}(\mathcal{D}_j)$ is the gradient of agent $j$'s local dataset $\mathcal{D}_j$ with respect to $i$'s local model. Similarly, for any other agent $j \in \mathcal{M}_i$ such that $j \neq i$, consider any two adjacent datasets $\mathcal{D}_j$ and $\mathcal{D}_j'$, when fixing the datasets of $\forall k \in \mathcal{M}_i \setminus j$, we also have

$$
\triangle_2 q^{[t]} = \max_{\mathcal{D}_j,\mathcal{D}_j'} \left\| q^{[t]}(\mathcal{D}_j) - q^{[t]}(\mathcal{D}_j') \right\| \leq \frac{2C}{\omega_{i,j}} + \sum_{j\in\mathcal{M}_i} \frac{2C}{\omega_{i,j}}
\tag{40}
$$

therefore, based on (39) and (40), we have

$$
\triangle_2 q^{[t]} \leq \frac{2C}{\omega_{\min}} + \sum_{j\in\mathcal{M}_i} \frac{2C}{\omega_{i,j}}
\tag{41}
$$

where $\omega_{\min} = \min_{i\in\mathcal{M}, j\in\mathcal{M}_i} \omega_{i,j}$. Based on **Theorem** 3.22 in [32], the Gaussian mechanism (i.e., (4)) meets $\sigma \geq \frac{\sqrt{2\ln(1.25/\delta)}\triangle_2 q}{\epsilon}$. Therefore, for agent $i$, we have

$$
\sigma \geq \frac{2C \left( \frac{1}{\omega_{\min}} + \sum_{j\in\mathcal{M}_i} \frac{1}{\omega_{i,j}} \right) \sqrt{2\ln(1.25/\delta)}}{\hat{\varphi}_{\min}\epsilon \sqrt{\sum_{j\in\mathcal{M}_i} \omega_{i,j}^{-2}}}
$$

based on which, we have

$$
\sigma \geq \max_{i\in\mathcal{M}} \left\{ \frac{2C \left( \frac{1}{\omega_{\min}} + \sum_{j\in\mathcal{M}_i} \frac{1}{\omega_{i,j}} \right) \sqrt{2\ln(1.25/\delta)}}{\hat{\varphi}_{\min}\epsilon \sqrt{\sum_{j\in\mathcal{M}_i} \omega_{i,j}^{-2}}} \right\}
$$

## APPENDIX B
### PROOF OF **THEOREM** 2

We show the detailed proof of **Theorem** 2. Recall that $\bar{g}_i^{[t]}$ is the average of the perturbed gradients $\left\{\hat{g}_{j,i}^{[t]}\right\}_{j\in\mathcal{M}_i}$ weighted by the Shapley values (see (21)), **Lemma** 1 bounds the (expected) weighted average perturbed gradient.

**Lemma 1.** *When **Assumptions** 1~3 hold, for any $t \geq 1$, we have*

$$
\mathbb{E}\left[ \left\| \frac{1}{M} \sum_{i=1}^M \bar{g}_i^{[t]} \right\|^2 \right] \leq \frac{2C^2}{\omega_{\min}^4} + \frac{2\sigma^2 d}{\omega_{\min}^4}
\tag{42}
$$

*Proof.*

$$
\begin{aligned}
\mathbb{E}\left[ \left\| \frac{1}{M} \sum_{i=1}^M \bar{g}_i^{[t]} \right\|^2 \right] =& \mathbb{E}\left[ \left\| \frac{1}{M} \sum_{i=1}^M \sum_{j\in\mathcal{M}_i} \pi_{i,j}^{[t]} \left( \frac{g_{j,i}^{[t]}}{\max\left\{1, C^{-1}\left\|g_{j,i}^{[t]}\right\|\right\}} + n_{j,i}^{[t]} \right) \right\|^2 \right] \\
\overset{\textcircled{1}}{\leq}& 2\mathbb{E}\left[ \left\| \frac{1}{M} \sum_{i=1}^M \sum_{j\in\mathcal{M}_i} \frac{\hat{\varphi}_{i,j}^{[t]}}{\omega_{i,j} \sum_{k\in\mathcal{M}_i} \hat{\varphi}_{i,k}^{[t]}} \left( \frac{g_{j,i}^{[t]}}{\max\left\{1, C^{-1}\left\|g_{j,i}^{[t]}\right\|\right\}} \right) \right\|^2 \right]
\end{aligned}
$$

$$+ 2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\frac{\hat{\varphi}_{i,j}^{[t]}}{\omega_{i,j}\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}}n_{j,i}^{[t]}\right\|^2\right]$$

$$\overset{\text{②}}{\leq}\frac{2}{M}\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\mathbb{E}\left[\left\|\frac{\hat{\varphi}_{i,j}^{[t]}}{\omega_{i,j}^2\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}}\right\|^2\left\|\frac{g_{j,i}^{[t]}}{\max\left\{1,C^{-1}\left\|g_{j,i}^{[t]}\right\|\right\}}\right\|^2\right]$$

$$+ \frac{2}{M}\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\mathbb{E}\left[\left\|\frac{\hat{\varphi}_{i,j}^{[t]}}{\omega_{i,j}^2\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}}\right\|^2\left\|n_{j,i}^{[t]}\right\|^2\right]$$

$$\overset{\text{③}}{\leq}\frac{2}{M}\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\frac{1}{\omega_{i,j}^4}\mathbb{E}\left[\left\|\frac{g_{j,i}^{[t]}}{\max\left\{1,C^{-1}\left\|g_{j,i}^{[t]}\right\|\right\}}\right\|^2\right]$$

$$+ \frac{2}{M}\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\frac{1}{\omega_{i,j}^4}\mathbb{E}\left[\left\|n_{j,i}^{[t]}\right\|^2\right]$$

$$\overset{\text{④}}{\leq}\frac{2}{M}\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\frac{C^2}{\omega_{i,j}^4}+\frac{2}{M}\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\frac{\sigma^2 d}{\omega_{i,j}^4}$$

$$\overset{\text{⑤}}{\leq}\frac{2C^2}{\omega_{\min}^4}+\frac{2\sigma^2 d}{\omega_{\min}^4}$$

where ① follows from the inequality $\|a+b\|^2\leq 2\|a\|^2+2\|b\|^2$; ② follows from the convexity of $\|\cdot\|^2$ and *Jensen's Inequality*; ③ according to $\frac{\hat{\varphi}_{i,j}^{[t]}}{\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}}\leq 1$; ④ due to the characteristic of clipping process that $\left\|\frac{g_{j,i}^{[t]}}{\max\left\{1,C^{-1}\left\|g_{j,i}^{[t]}\right\|\right\}}\right\|^2\leq C^2$, and the property of Gaussian noises that $\mathbb{E}\left[\left\|n_{j,i}^{[t]}\right\|^2\right]=\sigma^2 d$; and ⑤ follows from $\omega_{\min}=\min_{i\in\mathcal{M},j\in\mathcal{M}_i}\omega_{i,j}$. $\square$

Based on **Lemma** 1, **Lemma** 2 presents that $\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\left(\bar{g}_i^{[t]}-g_{i,i}^{[t]}\right)\right\|^2\right]$ is bounded by $\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right]$.

**Lemma 2.** *When **Assumptions** 1~3 hold, for any $t\geq 1$, we have*

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\left(\bar{g}_i^{[t]}-g_{i,i}^{[t]}\right)\right\|^2\right]$$

$$\leq\frac{4C^2}{\omega_{\min}^4}+\frac{4\sigma^2 d}{\omega_{\min}^4}+\frac{2\zeta^2}{M}+2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right] \tag{43}$$

*Proof.* Recall that $g_{i,i}^{[t]}=\nabla F_i\left(x_i^{[t-1]};\xi_{i,t}\right)$ is mutually independent unbiased stochastic gradient at $x_i^{[t-1]}$ such that $\mathbb{E}\left[g_{i,i}^{[t]}\right]=\nabla f_i\left(x_i^{[t-1]}\right)$ for $\forall i\in\mathcal{M}$. We have

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\left(\bar{g}_i^{[t]}-g_{i,i}^{[t]}\right)\right\|^2\right]\overset{\text{①}}{\leq}2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[t]}\right\|^2\right]+2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}g_{i,i}^{[t]}\right\|^2\right]$$

$$\overset{\text{②}}{\leq}\frac{4C^2}{\omega_{\min}^4}+\frac{4\sigma^2 d}{\omega_{\min}^4}+2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}g_{i,i}^{[t]}\right\|^2\right]$$

$$\overset{\text{③}}{\leq}\frac{4C^2}{\omega_{\min}^4}+\frac{4\sigma^2 d}{\omega_{\min}^4}++2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right]$$

$$+ 2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\left(\nabla F_i\left(x_i^{[t-1]};\xi_{i,t}\right)-\nabla f_i\left(x_i^{[t-1]}\right)\right)\right\|^2\right]$$

$$\overset{\text{\textcircled{4}}}{\leq} \frac{4C^2}{\omega_{\min}^4} + \frac{4\sigma^2 d}{\omega_{\min}^4} + 2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right]$$

$$+ \frac{2}{M^2}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\nabla F_i\left(x_i^{[t-1]};\xi_{i,t}\right) - \nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right]$$

$$\overset{\text{\textcircled{5}}}{\leq} \frac{4C^2}{\omega_{\min}^4} + \frac{4\sigma^2 d}{\omega_{\min}^4} + \frac{2\zeta^2}{M} + 2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right]$$

where ① follows from the basic inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for any vectors $a$ and $b$; ② follows from **Lemma** 1; ③ based on the facts that $\mathbb{E}\left[g_{i,i}^{[t]}\right] = \mathbb{E}\left[\nabla F_i\left(x_i^{[t-1]};\xi_{i,t}\right)\right] = \nabla f_i\left(x_i^{[t-1]}\right)$ and $\mathbb{E}\left[\|a\|^2\right] = \mathbb{E}\left[\|a - \mathbb{E}[a]\|^2\right] + \mathbb{E}\left[\|\mathbb{E}[a]\|^2\right]$ holds for any random vector $a$; ④ follows from the facts that $z_i := \nabla F_i\left(x_i^{[t-1]};\xi_{i,t}\right) - \nabla f_i\left(x_i^{[t-1]}\right)$'s are independent random vectors (with respect to $i$) with zero means and $\mathbb{E}\left[\left\|\sum_{i=1}^{M}z_i\right\|^2\right] = \sum_{i=1}^{M}\mathbb{E}\left[\|z_i\|^2\right]$; ⑤ follows from **Assumption** 2. □

To prove **Theorem** 2, we first present an auxiliary sequence $\left\{\bar{S}^{[t]}\right\}_{t\geq 0}$ such that

$$\bar{S}^{[t]} = \begin{cases} \bar{x}^{[t]} & \text{with } t = 0 \\ \frac{1}{1-\alpha}\bar{x}^{[t]} - \frac{\alpha}{1-\alpha}\bar{x}^{[t-1]} & \text{with } t \geq 1 \end{cases} \tag{44}$$

where $\bar{x}^{[t]} = \frac{1}{M}\sum_{i=1}^{M}x_i^{[t]}$. In the following, the initial value will be set to 0. Based on **Algorithm** 1, we have the following useful **Lemmas**.

**Lemma 3.** *According to **Algorithm** 1, for any $t \geq 1$, we have*

$$\bar{S}^{[t]} - \bar{S}^{[t-1]} = \frac{-\gamma}{M(1-\alpha)}\sum_{i=1}^{M}\bar{g}_i^{[t]} \tag{45}$$

*Proof.* As shown in Line 23 and Line 19 in **Algorithm** 1, we have

$$\bar{u}^{[t]} = \frac{1}{M}\sum_{i=1}^{M}u_i^{[t]} = \frac{1}{M}\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\hat{u}_j^{[t]} = \frac{1}{M}\sum_{i=1}^{M}\hat{u}_i^{[t]}\left(\sum_{j=1}^{M}w_{j,i}\right)$$

$$= \frac{1}{M}\sum_{i=1}^{M}\hat{u}_i^{[t]} = \frac{1}{M}\sum_{i=1}^{M}\left(\alpha u_i^{[t-1]} + \bar{g}_i^{[t]}\right) = \alpha\bar{u}^{[t-1]} + \frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[t]} \tag{46}$$

Moreover, according to Line 24 and Line 20 in **Algorithm** 1, we have

$$\bar{x}^{[t]} = \frac{1}{M}\sum_{i=1}^{M}x_i^{[t]} = \frac{1}{M}\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\hat{x}_j^{[t]} = \frac{1}{M}\sum_{i=1}^{M}\hat{x}_i^{[t]}\left(\sum_{j=1}^{M}w_{j,i}\right)$$

$$= \frac{1}{M}\sum_{i=1}^{M}\hat{x}_i^{[t]} = \frac{1}{M}\sum_{i=1}^{M}\left(x_i^{[t-1]} - \gamma\hat{u}_i^t\right) = \bar{x}^{[t-1]} - \gamma\bar{u}^{[t]} \tag{47}$$

We then prove this lemma by induction. When $t = 1$, we have

$$\bar{S}^{[1]} - \bar{S}^{[0]} = \frac{1}{1-\alpha}\bar{x}^{[1]} - \frac{\alpha}{1-\alpha}\bar{x}^{[0]} - \bar{x}^{[0]}$$

$$= \frac{1}{1-\alpha}\left(\bar{x}^{[1]} - \bar{x}^{[0]}\right) = \frac{-\gamma\bar{u}^{[1]}}{1-\alpha}$$

$$= \frac{-\gamma}{M(1-\alpha)}\sum_{i=1}^{M}\bar{g}_i^{[1]}$$

by considering $\gamma\bar{u}^{[1]} = x^{[0]} - x^{[1]}$ due to (47) and $\bar{u}^{[1]} = \frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[1]}$ due to (46). For any $t > 1$, we have

$$\bar{S}^{[t]} - \bar{S}^{[t-1]} = \left(\frac{1}{1-\alpha}\bar{x}^{[t]} - \frac{\alpha}{1-\alpha}\bar{x}^{[t-1]}\right) - \left(\frac{1}{1-\alpha}\bar{x}^{[t-1]} - \frac{\alpha}{1-\alpha}\bar{x}^{[t-2]}\right)$$

$$= \frac{1}{1-\alpha}\left(\bar{x}^{[t]} - \bar{x}^{[t-1]}\right) - \frac{\alpha}{1-\alpha}\left(\bar{x}^{[t-1]} - \bar{x}^{[t-2]}\right)$$

$$= \frac{1}{1-\alpha}\left(-\gamma\bar{u}^{[t]} - \alpha\left(-\gamma\bar{u}^{[t-1]}\right)\right)$$

$$= \frac{-\gamma}{M(1-\alpha)}\sum_{i=1}^{M}\bar{g}_i^{[t]}$$

which finally completes the proof. □

**Lemma 4.** *For any $T \geq 1$, we have*

$$\sum_{t=1}^{T}\left\|\bar{S}^{[t]} - \bar{x}^{[t]}\right\|^2 \leq \frac{\gamma^2\alpha^2}{(1-\alpha)^4}\sum_{t=1}^{T}\left\|\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[t]}\right\|^2 \tag{48}$$

*Proof.* By applying (46) in **Lemma** 3 recursively, we have

$$\bar{u}^{[t]} = \sum_{\tau=1}^{t}\alpha^{t-\tau}\left(\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[\tau]}\right), \quad \forall t \geq 1 \tag{49}$$

and thus,

$$\bar{S}^{[t]} - \bar{x}^{[t]} = \frac{-\gamma\alpha}{1-\alpha}\bar{u}^{[t]} = \frac{-\gamma\alpha}{1-\alpha}\sum_{\tau=1}^{t}\alpha^{t-\tau}\left(\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[\tau]}\right) \tag{50}$$

Based on (50), we have

$$\left\|\bar{S}^{[t]} - \bar{x}^{[t]}\right\|^2 = \left\|\frac{-\gamma\alpha}{1-\alpha}\sum_{\tau=1}^{t}\alpha^{t-\tau}\left(\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[\tau]}\right)\right\|^2$$

$$\overset{①}{=} \frac{\gamma^2\alpha^2}{(1-\alpha)^2}\phi_t^2\left\|\sum_{\tau=1}^{t}\frac{\alpha^{t-\tau}}{\phi_t}\left(\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[\tau]}\right)\right\|^2$$

$$\overset{②}{\leq} \frac{\gamma^2\alpha^2}{(1-\alpha)^2}\phi_t^2\sum_{\tau=1}^{t}\frac{\alpha^{t-\tau}}{\phi_t}\left\|\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[\tau]}\right\|^2$$

$$\overset{③}{\leq} \frac{\gamma^2\alpha^2}{(1-\alpha)^3}\sum_{\tau=1}^{t}\alpha^{t-\tau}\left\|\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[\tau]}\right\|^2$$

where we have ① by letting $\phi_t = \sum_{\tau=1}^{t}\alpha^{t-\tau} = \frac{1-\alpha^t}{1-\alpha}$; ② follows from the convexity of $\|\cdot\|^2$ and Jensen's Inequality; and ③ by considering $1 - \alpha^t \leq 1$.

Furthermore, by summing $\left\|\bar{S}^{[t]} - \bar{x}^{[t]}\right\|^2$ over $t = 1, 2, \cdots, T$, we obtain

$$\sum_{t=1}^{T}\left\|\bar{S}^{[t]} - \bar{x}^{[t]}\right\|^2 \leq \frac{\gamma^2\alpha^2}{(1-\alpha)^3}\sum_{t=1}^{T}\sum_{\tau=1}^{t}\alpha^{t-\tau}\left\|\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[\tau]}\right\|^2$$

$$\leq \frac{\gamma^2\alpha^2}{(1-\alpha)^3}\sum_{t=1}^{T}\left(\left\|\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[t]}\right\|^2\sum_{\tau=t}^{T}\alpha^{\tau-t}\right)$$

$$\overset{①}{\leq} \frac{\gamma^2\alpha^2}{(1-\alpha)^4}\sum_{t=1}^{T}\left\|\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[t]}\right\|^2$$

where ① follows from $\sum_{\tau=t}^{T}\alpha^{\tau-t} = \frac{1-\alpha^{T-t+1}}{1-\alpha} \leq \frac{1}{1-\alpha}$. □

Before proceeding with our proof, we first show some key notations (see (51)) and facts (i.e., **Fact** 1 and **Fact** 2) which will be helpful in the subsequent discussion. Similar facts are adopted in [12], [20].

$$
\begin{cases}
\bar{\mathbf{G}}^{[t]} & \triangleq \left[\bar{g}_1^{[t]}, \bar{g}_2^{[t]}, \cdots, \bar{g}_M^{[t]}\right] \\
\mathbf{U}^{[t]} & \triangleq \left[u_1^{[t]}, u_2^{[t]}, \cdots, u_M^{[t]}\right] \\
\hat{\mathbf{U}}^{[t]} & \triangleq \left[\hat{u}_1^{[t]}, \hat{u}_2^{[t]}, \cdots, \hat{u}_M^{[t]}\right] \\
\mathbf{X}^{[t]} & \triangleq \left[x_1^{[t]}, x_2^{[t]}, \cdots, x_M^{[t]}\right] \\
\mathbf{G}^{[t]} & \triangleq \left[g_{1,1}^{[t]}, g_{2,2}^{[t]}, \cdots, g_{M,M}^{[t]}\right] \\
\mathbf{H}^{[t]} & \triangleq \left[\nabla f_1\left(x_1^{[t-1]}\right), \cdots, \nabla f_M\left(x_M^{[t-1]}\right)\right] \\
\mathbf{J}^{[t]} & \triangleq \left[\nabla f_1\left(\bar{x}^{[t-1]}\right), \cdots, \nabla f_M\left(\bar{x}^{[t-1]}\right)\right]
\end{cases}
\tag{51}
$$

**Fact 1.** Let $\mathbf{Q} = \frac{1}{M}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ where $\mathbf{1}$ denotes a $M$-dimensional column vector with all entries being $1$. For any doubly stochastic matrix $\mathbf{W}$, we have:

- $\mathbf{Q}\mathbf{W} = \mathbf{W}\mathbf{Q}$;
- $(\mathbf{I} - \mathbf{Q})\mathbf{W} = \mathbf{W}(\mathbf{I} - \mathbf{Q})$;
- For any integer $t \geq 1$, $\|(\mathbf{I} - \mathbf{Q})\mathbf{W}^t\|_{\mathfrak{s}} \leq (\sqrt{\rho})^t$, where $\|\cdot\|_{\mathfrak{s}}$ is the spectrum norm.

**Fact 2.** Let $\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_m$ be $m$ arbitrary real-valued matrices such that $\|\mathbf{B}_i\|_{\mathfrak{F}}^2 = \sum_{k=1}^{M}\|\mathbf{b}_{i,k}\|^2$, where $\|\cdot\|_{\mathfrak{F}}$ is the Frobenius norm, and $\mathbf{b}_{i,k}$ is the $k$-th column of $\mathbf{B}_i$. We have

$$
\left\|\sum_{i=1}^{m}\mathbf{B}_i\right\|_{\mathfrak{F}}^2 \leq \sum_{i=1}^{m}\sum_{j=1}^{m}\|\mathbf{B}_i\|_{\mathfrak{F}}\|\mathbf{B}_j\|_{\mathfrak{F}}
\tag{52}
$$

Next, in **Lemma** 5, we present the bound of $\mathbb{E}\left[\left\|\bar{\mathbf{G}}^{[t]} - \mathbf{G}^{[t]}\right\|_{\mathfrak{F}}^2\right]$, based on which, in **Lemma** 6, we demonstrate that $\sum_{t=1}^{T}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\bar{x}^{[t]} - x_i^{[t]}\right\|^2\right]$ is bounded by $\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right]$.

**Lemma 5.** When all **Assumptions** 1~3 hold, for any $t \geq 1$, we have

$$
\mathbb{E}\left[\left\|\bar{\mathbf{G}}^{[t]} - \mathbf{G}^{[t]}\right\|_{\mathfrak{F}}^2\right]
$$
$$
\leq \frac{4MC^2}{\omega_{\min}^4} + \frac{4M\sigma^2 d}{\omega_{\min}^4} + 2M\zeta^2 + 8M\kappa^2
$$
$$
+ 16L^2\sum_{i=1}^{M}\mathbb{E}\left[\left\|x_i^{[t-1]} - \bar{x}^{[t-1]}\right\|^2\right]
$$
$$
+ 8M\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right]
\tag{53}
$$

*Proof.* According to (51) and **Fact** 2, we have

$$
\mathbb{E}\left[\left\|\bar{\mathbf{G}}^{[t]} - \mathbf{G}^{[t]}\right\|_{\mathfrak{F}}^2\right] = \sum_{i=1}^{M}\mathbb{E}\left[\left\|\bar{g}_i^{[t]} - g_{i,i}^{[t]}\right\|^2\right]
$$
$$
\overset{\textcircled{1}}{\leq} 2\sum_{i=1}^{M}\mathbb{E}\left[\left\|\bar{g}_i^{[t]}\right\|^2\right] + 2\sum_{i=1}^{M}\mathbb{E}\left[\left\|g_{i,i}^{[t]}\right\|^2\right]
$$
$$
\leq 2\sum_{i=1}^{M}\mathbb{E}\left[\left\|\sum_{j\in\mathcal{M}_i}\pi_{i,j}^{[t]}\left(\frac{g_{j,i}^{[t]}}{\max\left\{1, C^{-1}\left\|g_{j,i}^{[t]}\right\|\right\}} + n_{j,i}^{[t]}\right)\right\|^2\right] + 2\sum_{i=1}^{M}\mathbb{E}\left[\left\|g_{i,i}^{[t]}\right\|^2\right]
$$

$$\overset{\text{\textcircled{2}}}{\leq} 4\sum_{i=1}^{M}\mathbb{E}\left[\left\|\sum_{j\in\mathcal{M}_i}\frac{\hat{\varphi}_{i,j}^{[t]}}{\omega_{i,j}\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}}\left(\frac{g_{j,i}^{[t]}}{\max\left\{1,C^{-1}\left\|g_{j,i}^{[t]}\right\|\right\}}\right)\right\|^2\right]$$

$$+4\sum_{i=1}^{M}\mathbb{E}\left[\left\|\sum_{j\in\mathcal{M}_i}\frac{\hat{\varphi}_{i,j}^{[t]}}{\omega_{i,j}\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}}n_{j,i}^{[t]}\right\|^2\right]+2\sum_{i=1}^{M}\mathbb{E}\left[\left\|g_{i,i}^{[t]}\right\|^2\right]$$

$$\overset{\text{\textcircled{3}}}{\leq} 4\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\mathbb{E}\left[\left\|\frac{\hat{\varphi}_{i,j}^{[t]}}{\omega_{i,j}^2\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}}\right\|^2\left\|\frac{g_{j,i}^{[t]}}{\max\left\{1,C^{-1}\left\|g_{j,i}^{[t]}\right\|\right\}}\right\|^2\right]$$

$$+4\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\mathbb{E}\left[\left\|\frac{\hat{\varphi}_{i,j}^{[t]}}{\omega_{i,j}^2\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}}\right\|^2\left\|n_{j,i}^{[t]}\right\|^2\right]+2\sum_{i=1}^{M}\mathbb{E}\left[\left\|g_{i,i}^{[t]}\right\|^2\right]$$

$$\overset{\text{\textcircled{4}}}{\leq} 4\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\frac{C^2}{\omega_{i,j}^4}+4\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\frac{\sigma^2 d}{\omega_{i,j}^4}+2\sum_{i=1}^{M}\mathbb{E}\left[\left\|g_{i,i}^{[t]}\right\|^2\right]$$

$$\overset{\text{\textcircled{5}}}{\leq} 4\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\frac{C^2}{\omega_{i,j}^4}+4\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\frac{\sigma^2 d}{\omega_{i,j}^4}$$

$$+2\sum_{i=1}^{M}\mathbb{E}\left[\left\|\nabla F_i\left(x_i^{[t-1]};\xi_{i,t}\right)-\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right]+2\sum_{i=1}^{M}\mathbb{E}\left[\left\|\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right]$$

$$\leq 4\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\frac{C^2}{\omega_{i,j}^4}+4\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\frac{\sigma^2 d}{\omega_{i,j}^4}$$

$$+2\sum_{i=1}^{M}\mathbb{E}\left[\left\|\nabla F_i\left(x_i^{[t-1]};\xi_{i,t}\right)-\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right]$$

$$+2\sum_{i=1}^{M}\mathbb{E}\left[\left\|\begin{array}{l}\nabla f_i\left(x_i^{[t-1]}\right)-\nabla f_i\left(\bar{x}^{[t-1]}\right)+\nabla f_i\left(\bar{x}^{[t-1]}\right)\\ -\frac{1}{M}\sum_{j=1}^{M}\nabla f_j\left(\bar{x}^{[t-1]}\right)+\frac{1}{M}\sum_{j=1}^{M}\nabla f_j\left(\bar{x}^{[t-1]}\right)\\ -\frac{1}{M}\sum_{j=1}^{M}\nabla f_j\left(x_j^{[t-1]}\right)+\frac{1}{M}\sum_{j=1}^{M}\nabla f_j\left(x_j^{[t-1]}\right)\end{array}\right\|^2\right]$$

$$\overset{\text{\textcircled{6}}}{\leq} 4\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\frac{C^2}{\omega_{i,j}^4}+4\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\frac{\sigma^2 d}{\omega_{i,j}^4}$$

$$+2\sum_{i=1}^{M}\mathbb{E}\left[\left\|\nabla F_i\left(x_i^{[t-1]};\xi_{i,t}\right)-\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right]$$

$$+8\sum_{i=1}^{M}\mathbb{E}\left[\left\|\nabla f_i\left(x_i^{[t-1]}\right)-\nabla f_i\left(\bar{x}^{[t-1]}\right)\right\|^2\right]$$

$$+8\sum_{i=1}^{M}\mathbb{E}\left[\left\|\nabla f_i\left(\bar{x}^{[t-1]}\right)-\frac{1}{M}\sum_{j=1}^{M}\nabla f_j\left(\bar{x}^{[t-1]}\right)\right\|^2\right]$$

$$+8\sum_{i=1}^{M}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{j=1}^{M}\nabla f_j\left(\bar{x}^{[t-1]}\right)-\frac{1}{M}\sum_{j=1}^{M}\nabla f_j\left(x_j^{[t-1]}\right)\right\|^2\right]$$

$$+ 8 \sum_{i=1}^{M} \mathbb{E}\left[\left\|\frac{1}{M}\sum_{j=1}^{M}\nabla f_j\left(x_j^{[t-1]}\right)\right\|^2\right]$$

$$\overset{\textcircled{7}}{\leq} 4\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\frac{C^2}{\omega_{i,j}^4} + 4\sum_{i=1}^{M}\sum_{j\in\mathcal{M}_i}\omega_{i,j}\frac{\sigma^2 d}{\omega_{i,j}^4} + 2M\zeta^2$$

$$+ 8L^2\sum_{i=1}^{M}\mathbb{E}\left[\left\|x_i^{[t-1]}-\bar{x}^{[t-1]}\right\|^2\right] + 8M\kappa^2$$

$$+ 8\sum_{i=1}^{M}\frac{1}{M}\sum_{j=1}^{M}L^2\mathbb{E}\left[\left\|\bar{x}^{[t-1]}-x_j^{[t-1]}\right\|^2\right]$$

$$+ 8M\mathbb{E}\left[\left\|\frac{1}{M}\sum_{j=1}^{M}\nabla f_j\left(x_j^{[t-1]}\right)\right\|^2\right]$$

$$\overset{\textcircled{8}}{\leq} \frac{4MC^2}{\omega_{\min}^4} + \frac{4M\sigma^2 d}{\omega_{\min}^4} + 2M\zeta^2 + 8M\kappa^2$$

$$+ 16L^2\sum_{i=1}^{M}\mathbb{E}\left[\left\|x_i^{[t-1]}-\bar{x}^{[t-1]}\right\|^2\right]$$

$$+ 8M\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right]$$

where $\textcircled{1}$ and $\textcircled{2}$ follow from $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$; $\textcircled{3}$ follows from the convexity of $\|\cdot\|^2$ and *Jensen's Inequality*; $\textcircled{4}$ follows from $\frac{\hat{\varphi}_{i,j}^{[t]}}{\sum_{k\in\mathcal{M}_i}\hat{\varphi}_{i,k}^{[t]}} \leq 1$, and the characteristic of clipping process that $\left\|\frac{g_{j,i}^{[t]}}{\max\left\{1,C^{-1}\left\|g_{j,i}^{[t]}\right\|\right\}}\right\|^2 \leq C^2$, and the property of Gaussian noises that $\mathbb{E}\left[\left\|n_{j,i}^{[t]}\right\|^2\right] = \sigma^2 d$; $\textcircled{5}$ follows from the facts that $\mathbb{E}\left[g_{i,i}^{[t]}\right] = \mathbb{E}\left[\nabla F_i\left(x_i^{[t-1]};\xi_{i,t}\right)\right] = \nabla f_i\left(x_i^{[t-1]}\right)$ and $\mathbb{E}\left[\|a\|^2\right] = \mathbb{E}\left[\|a-\mathbb{E}[a]\|^2\right] + \mathbb{E}\left[\|\mathbb{E}[a]\|^2\right]$ holds for any random vector $a$; $\textcircled{6}$ follows from $\|a+b+c+d\|^2 \leq 4\|a\|^2 + 4\|b\|^2 + 4\|c\|^2 + 4\|d\|^2$ for any vectors $a,b,c,d$; $\textcircled{7}$ follows from **Assumption** 2, and $\left\|\nabla f_i\left(x_i^{[t-1]}\right) - \nabla f_i\left(\bar{x}^{[t-1]}\right)\right\|^2 \leq L^2\left\|x_i^{[t-1]}-\bar{x}^{[t-1]}\right\|^2$ by the smoothness of $f_i$ in **Assumption** 1, and $\left\|\frac{1}{M}\sum_{j=1}^{M}\nabla f_j(\bar{x}^{[t-1]}) - \frac{1}{M}\sum_{j=1}^{M}\nabla f_j(x_j^{[t-1]})\right\|^2 \leq \frac{1}{M}\sum_{j=1}^{M}\left\|\nabla f_j(\bar{x}^{[t-1]}) - \nabla f_j(x_j^{[t-1]})\right\|^2 \leq L^2\frac{1}{M}\sum_{j=1}^{M}\left\|\bar{x}^{[t-1]}-x_j^{[t-1]}\right\|^2$ where the first inequality holds due to the convexity of $\|\cdot\|^2$ and *Jensen's Inequality*; $\textcircled{8}$ follows from $\omega_{\min} = \min_{i\in\mathcal{M},j\in\mathcal{M}_i}\omega_{i,j}$. $\square$

**Lemma 6.** *Let* **Assumptions** *1∼3 hold and* $\left\{\bar{x}^{[t]}\right\}_{t\geq 1}$ *be the sequence obtained by the iterations of* **Algorithm** *1. If momentum coefficient* $\alpha\in[0,1)$ *and learning rate*

$$\gamma \leq \frac{(1-\alpha)(1-\sqrt{\rho})}{2\sqrt{26}L} \tag{54}$$

*then for any* $T\geq 1$, *we have*

$$\sum_{t=1}^{T}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\bar{x}^{[t]}-x_i^{[t]}\right\|^2\right]$$

$$\leq T\frac{\frac{16\gamma^2}{\omega_{\min}^4}\left(C^2+\sigma^2 d\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2} + T\frac{4\gamma^2\left(7\zeta^2+13\kappa^2\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2}$$

$$+ \frac{52\gamma^2}{(1-\alpha)^2(1-\sqrt{\rho})^2}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right] \tag{55}$$

*Proof.* Recall the definition of $\mathbf{U}^{[t]}$ and $\hat{\mathbf{U}}^{[t]}$, according to Line 23 and Line 19 in **Algorithm** 1, for all $t\geq 1$, we have

$$\mathbf{U}^{[t]} = \hat{\mathbf{U}}^{[t]}\mathbf{W} = \left(\alpha\mathbf{U}^{[t-1]}+\bar{\mathbf{G}}^{[t]}\right)\mathbf{W} \tag{56}$$

By applying (56) recursively, we obtain

$$\mathbf{U}^{[t]} = \alpha^t \mathbf{U}^{[0]} \mathbf{W}^t + \sum_{\tau=1}^{t} \bar{\mathbf{G}}^{[\tau]} \alpha^{t-\tau} \mathbf{W}^{t-\tau+1} \overset{\mathbf{U}^{[0]}=0}{=} \sum_{\tau=1}^{t} \bar{\mathbf{G}}^{[\tau]} \alpha^{t-\tau} \mathbf{W}^{t-\tau+1} \tag{57}$$

Recalling the definition of $\mathbf{X}^{[t]}$, according to Line 24 and Line 20 in **Algorithm** 1, for all $t \geq 1$, we have

$$\mathbf{X}^{[t]} = \left( \mathbf{X}^{[t-1]} - \gamma \hat{\mathbf{U}}^{[t]} \right) \mathbf{W} = \mathbf{X}^{[t-1]} \mathbf{W} - \gamma \hat{\mathbf{U}}^{[t]} \mathbf{W} = \mathbf{X}^{[t-1]} \mathbf{W} - \gamma \mathbf{U}^{[t]} \tag{58}$$

By applying (58) recursively, we can obtain

$$\mathbf{X}^{[t]} = \mathbf{X}^{[0]} \mathbf{W}^t - \gamma \sum_{\tau=1}^{t} \mathbf{U}^{[\tau]} \mathbf{W}^{t-\tau} \tag{59}$$

Multiplying both sides by $(\mathbf{I} - \mathbf{Q})$ with $\mathbf{Q}$ defined in **Fact** 1, we have

$$\mathbf{X}^{[t]}(\mathbf{I} - \mathbf{Q}) = \mathbf{X}^{[0]} \mathbf{W}^t (\mathbf{I} - \mathbf{Q}) - \gamma \sum_{\tau=1}^{t} \mathbf{U}^{[\tau]} \mathbf{W}^{t-\tau} (\mathbf{I} - \mathbf{Q})$$

$$\overset{\textcircled{1}}{=} \mathbf{X}^{[0]}(\mathbf{I} - \mathbf{Q}) \mathbf{W}^t - \gamma \sum_{\tau=1}^{t} \mathbf{U}^{[\tau]} (\mathbf{I} - \mathbf{Q}) \mathbf{W}^{t-\tau} \tag{60}$$

where $\textcircled{1}$ follows from $\mathbf{W}(\mathbf{I} - \mathbf{Q}) = (\mathbf{I} - \mathbf{Q})\mathbf{W}$ in **Fact** 1.

By considering $\mathbf{X}^{[0]}(\mathbf{I} - \mathbf{Q}) = 0$ due to all columns of $\mathbf{X}^{[0]}$ are identical, we have

$$\mathbb{E} \left[ \left\| \mathbf{X}^{[t]}(\mathbf{I} - \mathbf{Q}) \right\|_{\mathfrak{F}}^2 \right]$$

$$= \mathbb{E} \left[ \left\| -\gamma \sum_{\tau=1}^{t} \mathbf{U}^{[\tau]}(\mathbf{I} - \mathbf{Q}) \mathbf{W}^{t-\tau} \right\|_{\mathfrak{F}}^2 \right]$$

$$= \mathbb{E} \left[ \left\| -\gamma \sum_{\tau=1}^{t} \sum_{\tau'=1}^{\tau} \bar{\mathbf{G}}^{[\tau']} \alpha^{\tau-\tau'} \mathbf{W}^{\tau-\tau'+1}(\mathbf{I} - \mathbf{Q}) \mathbf{W}^{t-\tau} \right\|_{\mathfrak{F}}^2 \right]$$

$$= \mathbb{E} \left[ \left\| -\gamma \sum_{\tau=1}^{t} \sum_{\tau'=1}^{\tau} \bar{\mathbf{G}}^{[\tau']} \alpha^{\tau-\tau'} \mathbf{W}^{t-\tau'+1}(\mathbf{I} - \mathbf{Q}) \right\|_{\mathfrak{F}}^2 \right]$$

$$= \mathbb{E} \left[ \left\| -\gamma \sum_{\tau=1}^{t} \bar{\mathbf{G}}^{[\tau]} \left( \sum_{\tau'=\tau}^{t} \alpha^{\tau'-\tau} \mathbf{W}^{t-\tau+1} \right)(\mathbf{I} - \mathbf{Q}) \right\|_{\mathfrak{F}}^2 \right]$$

$$= \mathbb{E} \left[ \left\| -\gamma \sum_{\tau=1}^{t} \bar{\mathbf{G}}^{[\tau]} \left( \sum_{\tau'=\tau}^{t} \alpha^{\tau'-\tau} \right) \mathbf{W}^{t-\tau+1}(\mathbf{I} - \mathbf{Q}) \right\|_{\mathfrak{F}}^2 \right]$$

$$= \mathbb{E} \left[ \left\| -\gamma \sum_{\tau=1}^{t} \frac{1 - \alpha^{t-\tau}}{1 - \alpha} \bar{\mathbf{G}}^{[\tau]}(\mathbf{I} - \mathbf{Q}) \mathbf{W}^{t-\tau+1} \right\|_{\mathfrak{F}}^2 \right]$$

$$= \gamma^2 \mathbb{E} \left[ \left\| \sum_{\tau=1}^{t} \frac{1 - \alpha^{t-\tau}}{1 - \alpha} \left( (\bar{\mathbf{G}}^{[\tau]} - \mathbf{G}^{[\tau]})(\mathbf{I} - \mathbf{Q}) \mathbf{W}^{t-\tau+1} + \mathbf{G}^{[\tau]}(\mathbf{I} - \mathbf{Q}) \mathbf{W}^{t-\tau+1} \right) \right\|_{\mathfrak{F}}^2 \right]$$

$$\overset{\textcircled{1}}{\leq} 2\gamma^2 \underbrace{\mathbb{E} \left[ \left\| \sum_{\tau=1}^{t} \frac{1 - \alpha^{t-\tau}}{1 - \alpha} \left( \bar{\mathbf{G}}^{[\tau]} - \mathbf{G}^{[\tau]} \right)(\mathbf{I} - \mathbf{Q}) \mathbf{W}^{t-\tau+1} \right\|_{\mathfrak{F}}^2 \right]}_{\text{I}}$$

$$+ 2\gamma^2 \underbrace{\mathbb{E} \left[ \left\| \sum_{\tau=1}^{t} \frac{1 - \alpha^{t-\tau}}{1 - \alpha} \mathbf{G}^{[\tau]}(\mathbf{I} - \mathbf{Q}) \mathbf{W}^{t-\tau+1} \right\|_{\mathfrak{F}}^2 \right]}_{\text{II}} \tag{61}$$

where we have ① by considering the inequality $\|\mathbf{A} + \mathbf{B}\|_{\mathfrak{F}}^2 \le 2\|\mathbf{A}\|_{\mathfrak{F}}^2 + 2\|\mathbf{B}\|_{\mathfrak{F}}^2$.

Next, we first give the upper bound of term I

$$\mathbb{E}\left[\left\|\sum_{\tau=1}^{t} \frac{1-\alpha^{t-\tau}}{1-\alpha}\left(\bar{\mathbf{G}}^{[\tau]} - \mathbf{G}^{[\tau]}\right)(\mathbf{I}-\mathbf{Q})\mathbf{W}^{t-\tau+1}\right\|_{\mathfrak{F}}^2\right]$$

$$\overset{①}{\le} \sum_{\tau=1}^{t}\sum_{\tau'=1}^{t}\mathbb{E}\left[\left\|\frac{1-\alpha^{t-\tau}}{1-\alpha}(\bar{\mathbf{G}}^{[\tau]}-\mathbf{G}^{[\tau]})(\mathbf{I}-\mathbf{Q})\mathbf{W}^{t-\tau+1}\right\|_{\mathfrak{F}}\left\|\frac{1-\alpha^{t-\tau'}}{1-\alpha}(\bar{\mathbf{G}}^{[\tau']}-\mathbf{G}^{[\tau']})(\mathbf{I}-\mathbf{Q})\mathbf{W}^{t-\tau'+1}\right\|_{\mathfrak{F}}\right]$$

$$\overset{②}{\le} \frac{1}{(1-\alpha)^2}\sum_{\tau=1}^{t}\sum_{\tau'=1}^{t}\mathbb{E}\left[\left\|\bar{\mathbf{G}}^{[\tau]}-\mathbf{G}^{[\tau]}\right\|_{\mathfrak{F}}\left\|(\mathbf{I}-\mathbf{Q})\mathbf{W}^{t-\tau+1}\right\|_{\mathfrak{s}}\left\|\bar{\mathbf{G}}^{[\tau']}-\mathbf{G}^{[\tau']}\right\|_{\mathfrak{F}}\left\|(\mathbf{I}-\mathbf{Q})\mathbf{W}^{t-\tau'+1}\right\|_{\mathfrak{s}}\right]$$

$$\overset{③}{\le} \frac{1}{(1-\alpha)^2}\sum_{\tau=1}^{t}\sum_{\tau'=1}^{t}\rho^{(t+1-\frac{\tau+\tau'}{2})}\mathbb{E}\left[\left\|\bar{\mathbf{G}}^{[\tau]}-\mathbf{G}^{[\tau]}\right\|_{\mathfrak{F}}\left\|\bar{\mathbf{G}}^{[\tau']}-\mathbf{G}^{[\tau']}\right\|_{\mathfrak{F}}\right]$$

$$\overset{④}{\le} \frac{1}{(1-\alpha)^2}\sum_{\tau=1}^{t}\sum_{\tau'=1}^{t}\rho^{(t+1-\frac{\tau+\tau'}{2})}\left(\frac{1}{2}\mathbb{E}\left[\left\|\bar{\mathbf{G}}^{[\tau]}-\mathbf{G}^{[\tau]}\right\|_{\mathfrak{F}}^2\right]+\frac{1}{2}\mathbb{E}\left[\left\|\bar{\mathbf{G}}^{[\tau']}-\mathbf{G}^{[\tau']}\right\|_{\mathfrak{F}}^2\right]\right)$$

$$\overset{⑤}{\le} \frac{1}{(1-\alpha)^2(1-\sqrt{\rho})}\sum_{\tau=1}^{t}\rho^{(\frac{t+1-\tau}{2})}\mathbb{E}\left[\left\|\bar{\mathbf{G}}^{[\tau]}-\mathbf{G}^{[\tau]}\right\|_{\mathfrak{F}}^2\right]$$

$$\overset{⑥}{\le} \frac{1}{(1-\alpha)^2(1-\sqrt{\rho})}\sum_{\tau=1}^{t}\rho^{(\frac{t+1-\tau}{2})}\begin{pmatrix}\frac{4MC^2}{\omega_{\min}^4}+\frac{4M\sigma^2 d}{\omega_{\min}^4}+2M\zeta^2+8M\kappa^2\\[2mm]+16L^2\sum_{i=1}^{M}\mathbb{E}\left[\left\|x_i^{[\tau-1]}-\bar{x}^{[\tau-1]}\right\|^2\right]\\[2mm]+8M\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[\tau-1]}\right)\right\|^2\right]\end{pmatrix}$$

$$\le \frac{\frac{4M}{\omega_{\min}^4}\left(C^2+\sigma^2 d\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2}+\frac{2M\left(\zeta^2+4\kappa^2\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2}$$

$$+\frac{16L^2}{(1-\alpha)^2(1-\sqrt{\rho})}\sum_{\tau=1}^{t}\rho^{(\frac{t+1-\tau}{2})}\sum_{i=1}^{M}\mathbb{E}\left[\left\|x_i^{[\tau-1]}-\bar{x}^{[\tau-1]}\right\|^2\right]$$

$$+\frac{8M}{(1-\alpha)^2(1-\sqrt{\rho})}\sum_{\tau=1}^{t}\rho^{(\frac{t+1-\tau}{2})}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[\tau-1]}\right)\right\|^2\right]\tag{62}$$

where ① follows from **Fact** 2; ② follows from the inequality $\left|\frac{1-\alpha^{t-\tau}}{1-\alpha}\right| \le \frac{1}{1-\alpha}$ and the inequality $\|\mathbf{AB}\|_{\mathfrak{F}} \le \|\mathbf{B}\|_{\mathfrak{s}}\|\mathbf{A}\|_{\mathfrak{F}}$ if matrix $\mathbf{B}$ is symmetric; ③ follows from **Fact** 1; ④ follows from the inequality $\|\mathbf{A}\|_{\mathfrak{F}}\|\mathbf{B}\|_{\mathfrak{F}} \le \frac{1}{2}(\|\mathbf{A}\|_{\mathfrak{F}}^2 + \|\mathbf{B}\|_{\mathfrak{F}}^2)$; ⑤ follows from the inequality $\sum_{\tau'=1}^{t}\rho^{\left(t+1-\frac{\tau+\tau'}{2}\right)} = \rho^{\frac{t+1-\tau}{2}}(1-\sqrt{\rho})^{-1}$; ⑥ follows from **Lemma** 5.

We then bound term II as follows

$$\mathbb{E}\left[\left\|\sum_{\tau=1}^{t}\frac{1-\alpha^{t-\tau}}{1-\alpha}\left(\mathbf{G}^{[\tau]}(\mathbf{I}-\mathbf{Q})\mathbf{W}^{t-\tau+1}\right)\right\|_{\mathfrak{F}}^2\right]$$

$$\le \sum_{\tau=1}^{t}\sum_{\tau'=1}^{t}\mathbb{E}\left[\left\|\frac{1-\alpha^{t-\tau}}{1-\alpha}\left(\mathbf{G}^{[\tau]}(\mathbf{I}-\mathbf{Q})\mathbf{W}^{t-\tau+1}\right)\right\|_{\mathfrak{F}}\left\|\frac{1-\alpha^{t-\tau'}}{1-\alpha}\left(\mathbf{G}^{[\tau']}(\mathbf{I}-\mathbf{Q})\mathbf{W}^{t-\tau'+1}\right)\right\|_{\mathfrak{F}}\right]$$

$$\le \frac{1}{(1-\alpha)^2}\sum_{\tau=1}^{t}\sum_{\tau'=1}^{t}\rho^{(t+1-\frac{\tau+\tau'}{2})}\mathbb{E}\left[\left\|\mathbf{G}^{[\tau]}\right\|_{\mathfrak{F}}\left\|\mathbf{G}^{[\tau']}\right\|_{\mathfrak{F}}\right]$$

$$\le \frac{1}{(1-\alpha)^2}\sum_{\tau=1}^{t}\sum_{\tau'=1}^{t}\rho^{(t+1-\frac{\tau+\tau'}{2})}\left(\frac{1}{2}\mathbb{E}\left[\left\|\mathbf{G}^{[\tau]}\right\|_{\mathfrak{F}}^2\right]+\frac{1}{2}\mathbb{E}\left[\left\|\mathbf{G}^{[\tau']}\right\|_{\mathfrak{F}}^2\right]\right)$$

$$\leq \frac{1}{(1-\alpha)^2} \sum_{\tau=1}^{t} \sum_{\tau'=1}^{t} \rho^{(t+1-\frac{\tau+\tau'}{2})} \mathbb{E}\left[\left\|\mathbf{G}^{[\tau]}\right\|_{\mathfrak{F}}^2\right]$$

$$\leq \frac{1}{(1-\alpha)^2(1-\sqrt{\rho})} \sum_{\tau=1}^{t} \rho^{(\frac{t+1-\tau}{2})} \mathbb{E}\left[\left\|\mathbf{G}^{[\tau]}\right\|_{\mathfrak{F}}^2\right] \tag{63}$$

It is apparent that the bound of II is dependent on $\mathbb{E}\left[\left\|\mathbf{G}^{[\tau]}\right\|_{\mathfrak{F}}^2\right]$. Therefore, we have

$$\mathbb{E}\left[\left\|\mathbf{G}^{[\tau]}\right\|_{\mathfrak{F}}^2\right] = \mathbb{E}\left[\left\|\mathbf{G}^{[\tau]} - \mathbf{H}^{[\tau]} + \mathbf{H}^{[\tau]} - \mathbf{J}^{[\tau]} + \mathbf{J}^{[\tau]} - \mathbf{J}^{[\tau]}\mathbf{Q} + \mathbf{J}^{[\tau]}\mathbf{Q} - \mathbf{H}^{[\tau]}\mathbf{Q} + \mathbf{H}^{[\tau]}\mathbf{Q}\right\|_{\mathfrak{F}}^2\right]$$

$$\leq 5\mathbb{E}\left[\left\|\mathbf{G}^{[\tau]} - \mathbf{H}^{[\tau]}\right\|_{\mathfrak{F}}^2\right] + 5\mathbb{E}\left[\left\|\mathbf{H}^{[\tau]} - \mathbf{J}^{[\tau]}\right\|_{\mathfrak{F}}^2\right] + 5\mathbb{E}\left[\left\|\mathbf{J}^{[\tau]} - \mathbf{J}^{[\tau]}\mathbf{Q}\right\|_{\mathfrak{F}}^2\right]$$

$$+ 5\mathbb{E}\left[\left\|\mathbf{J}^{[\tau]}\mathbf{Q} - \mathbf{H}^{[\tau]}\mathbf{Q}\right\|_{\mathfrak{F}}^2\right] + 5\mathbb{E}\left[\left\|\mathbf{H}^{[\tau]}\mathbf{Q}\right\|_{\mathfrak{F}}^2\right]$$

$$\leq 5\mathbb{E}\left[\sum_{i=1}^{M}\left\|\nabla F_i\left(x_i^{[\tau-1]};\xi_{i,\tau}\right) - \nabla f_i\left(x_i^{[\tau-1]}\right)\right\|^2\right]$$

$$+ 5\mathbb{E}\left[\sum_{i=1}^{M}\left\|\nabla f_i\left(x_i^{[\tau-1]}\right) - \nabla f_i\left(\bar{x}^{[\tau-1]}\right)\right\|^2\right]$$

$$+ 5\mathbb{E}\left[\sum_{i=1}^{M}\left\|\nabla f_i\left(\bar{x}^{[\tau-1]}\right) - \frac{1}{M}\sum_{j=1}^{M}\nabla f_j\left(\bar{x}^{[\tau-1]}\right)\right\|^2\right]$$

$$+ 5\mathbb{E}\left[\sum_{i=1}^{M}\left\|\frac{1}{M}\sum_{j=1}^{M}\nabla f_j\left(\bar{x}^{[\tau-1]}\right) - \frac{1}{M}\sum_{j=1}^{M}\nabla f_j\left(x_j^{[\tau-1]}\right)\right\|^2\right]$$

$$+ 5\mathbb{E}\left[\sum_{i=1}^{M}\left\|\frac{1}{M}\sum_{j=1}^{M}\nabla f_j\left(x_j^{[\tau-1]}\right)\right\|^2\right]$$

$$\overset{①}{\leq} 5M\zeta^2 + 5M\kappa^2 + 10L^2\sum_{i=1}^{M}\mathbb{E}\left[\left\|x_i^{[\tau-1]} - \bar{x}^{[\tau-1]}\right\|^2\right]$$

$$+ 5M\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[\tau-1]}\right)\right\|^2\right] \tag{64}$$

where ① follows from **Assumption** 2, and

$$\sum_{i=1}^{M}\left\|\nabla f_i(x_i^{[\tau-1]}) - \nabla f_i(\bar{x}^{[\tau-1]})\right\|^2 \leq L^2\sum_{i=1}^{M}\left\|x_i^{[\tau-1]} - \bar{x}^{[\tau-1]}\right\|^2$$

due to the smoothness of each $f_i$ by **Assumption** 1, and

$$\sum_{i=1}^{M}\left\|\frac{1}{M}\sum_{j=1}^{M}\nabla f_j(\bar{x}^{[\tau-1]}) - \frac{1}{M}\sum_{j=1}^{M}\nabla f_j(x_j^{[\tau-1]})\right\|^2$$

$$\leq \sum_{i=1}^{M}\frac{1}{M}\sum_{j=1}^{M}\left\|\nabla f_j(\bar{x}^{[\tau-1]}) - \nabla f_j(x_j^{[\tau-1]})\right\|^2 \leq L^2\sum_{j=1}^{M}\left\|\bar{x}^{[\tau-1]} - x_j^{[\tau-1]}\right\|^2$$

where the first inequality holds due to the the convexity of $\|\cdot\|^2$ and *Jensen's Inequality*. By substituting (64) into (63), we continue to derive the bound of **II** as follows,

$$\mathbb{E}\left[\left\|\sum_{\tau=1}^{t}\frac{1-\alpha^{t-\tau}}{1-\alpha}\left(\mathbf{G}^{[\tau]}(\mathbf{I}-\mathbf{Q})\mathbf{W}^{t-\tau+1}\right)\right\|_{\mathfrak{F}}^2\right]$$

$$\leq \frac{1}{(1-\alpha)^2(1-\sqrt{\rho})} \sum_{\tau=1}^{t} \rho^{(\frac{t+1-\tau}{2})} \left( \begin{array}{c} 5M\zeta^2 + 5M\kappa^2 + 10L^2 \sum_{i=1}^{M} \mathbb{E}\left[\left\|x_i^{[\tau-1]} - \bar{x}^{[\tau-1]}\right\|^2\right] \\ + 5M\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M} \nabla f_i\left(x_i^{[\tau-1]}\right)\right\|^2\right] \end{array} \right)$$

$$\leq \frac{5M\left(\zeta^2+\kappa^2\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2} + \frac{10L^2}{(1-\alpha)^2(1-\sqrt{\rho})} \sum_{\tau=1}^{t} \rho^{(\frac{t+1-\tau}{2})} \sum_{i=1}^{M} \mathbb{E}\left[\left\|x_i^{[\tau-1]} - \bar{x}^{[\tau-1]}\right\|^2\right]$$

$$+ \frac{5M}{(1-\alpha)^2(1-\sqrt{\rho})} \sum_{\tau=1}^{t} \rho^{(\frac{t+1-\tau}{2})} \mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M} \nabla f_i\left(x_i^{[\tau-1]}\right)\right\|^2\right] \tag{65}$$

By substituting (62) and (65) into (61), we have

$$\mathbb{E}\left[\left\|\mathbf{X}^{[t]}(\mathbf{I}-\mathbf{Q})\right\|_{\mathfrak{F}}^2\right] \leq \frac{\frac{8\gamma^2 M}{\omega_{\min}^4}\left(C^2+\sigma^2 d\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2} + \frac{4\gamma^2 M\left(\zeta^2+4\kappa^2\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2}$$

$$+ \frac{32\gamma^2 L^2}{(1-\alpha)^2(1-\sqrt{\rho})} \sum_{\tau=1}^{t} \rho^{(\frac{t+1-\tau}{2})} \sum_{i=1}^{M} \mathbb{E}\left[\left\|x_i^{[\tau-1]} - \bar{x}^{[\tau-1]}\right\|^2\right]$$

$$+ \frac{16\gamma^2 M}{(1-\alpha)^2(1-\sqrt{\rho})} \sum_{\tau=1}^{t} \rho^{(\frac{t+1-\tau}{2})} \mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M} \nabla f_i\left(x_i^{[\tau-1]}\right)\right\|^2\right]$$

$$+ \frac{10\gamma^2 M\left(\zeta^2+\kappa^2\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2} + \frac{20\gamma^2 L^2}{(1-\alpha)^2(1-\sqrt{\rho})} \sum_{\tau=1}^{t} \rho^{(\frac{t+1-\tau}{2})} \sum_{i=1}^{M} \mathbb{E}\left[\left\|x_i^{[\tau-1]} - \bar{x}^{[\tau-1]}\right\|^2\right]$$

$$+ \frac{10\gamma^2 M}{(1-\alpha)^2(1-\sqrt{\rho})} \sum_{\tau=1}^{t} \rho^{(\frac{t+1-\tau}{2})} \mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M} \nabla f_i\left(x_i^{[\tau-1]}\right)\right\|^2\right]$$

$$\leq \frac{\frac{8\gamma^2 M}{\omega_{\min}^4}\left(C^2+\sigma^2 d\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2} + \frac{\gamma^2 M\left(14\zeta^2+26\kappa^2\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2}$$

$$+ \frac{52\gamma^2 L^2}{(1-\alpha)^2(1-\sqrt{\rho})} \sum_{\tau=1}^{t} \rho^{(\frac{t+1-\tau}{2})} \sum_{i=1}^{M} \mathbb{E}\left[\left\|x_i^{[\tau-1]} - \bar{x}^{[\tau-1]}\right\|^2\right]$$

$$+ \frac{26\gamma^2 M}{(1-\alpha)^2(1-\sqrt{\rho})} \sum_{\tau=1}^{t} \rho^{(\frac{t+1-\tau}{2})} \mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M} \nabla f_i\left(x_i^{[\tau-1]}\right)\right\|^2\right] \tag{66}$$

Summing over $t=1,2,\cdots,T$ and noting that $\mathbb{E}\left[\left\|\mathbf{X}^{[0]}(\mathbf{I}-\mathbf{Q})\right\|_{\mathfrak{F}}^2\right]=0$, we have

$$\sum_{t=1}^{T} \mathbb{E}\left[\left\|\mathbf{X}^{[t]}(\mathbf{I}-\mathbf{Q})\right\|_{\mathfrak{F}}^2\right] \leq T\frac{\frac{8\gamma^2 M}{\omega_{\min}^4}\left(C^2+\sigma^2 d\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2} + T\frac{\gamma^2 M\left(14\zeta^2+26\kappa^2\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2}$$

$$+ \frac{52\gamma^2 L^2}{(1-\alpha)^2(1-\sqrt{\rho})} \sum_{t=1}^{T}\sum_{\tau=1}^{t} \rho^{(\frac{t+1-\tau}{2})} \sum_{i=1}^{M} \mathbb{E}\left[\left\|x_i^{[\tau-1]} - \bar{x}^{[\tau-1]}\right\|^2\right]$$

$$+ \frac{26\gamma^2 M}{(1-\alpha)^2(1-\sqrt{\rho})} \sum_{t=1}^{T}\sum_{\tau=1}^{t} \rho^{(\frac{t+1-\tau}{2})} \mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M} \nabla f_i\left(x_i^{[\tau-1]}\right)\right\|^2\right]$$

$$\leq T\frac{\frac{8\gamma^2 M}{\omega_{\min}^4}\left(C^2+\sigma^2 d\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2} + T\frac{\gamma^2 M\left(14\zeta^2+26\kappa^2\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2}$$

$$+ \frac{52\gamma^2 L^2}{(1-\alpha)^2(1-\sqrt{\rho})^2} \sum_{t=1}^{T}\sum_{i=1}^{M} \mathbb{E}\left[\left\|x_i^{[t]} - \bar{x}^{[t]}\right\|^2\right]$$

$$+\frac{26\gamma^2 M}{(1-\alpha)^2(1-\sqrt{\rho})^2}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right] \tag{67}$$

Dividing the both sides of (67) by $M$, we have

$$\sum_{t=1}^{T}\frac{1}{M}\mathbb{E}\left[\left\|\mathbf{X}^{[t]}(\mathbf{I}-\mathbf{Q})\right\|_{\mathfrak{F}}^2\right]\leq T\frac{\frac{8\gamma^2}{\omega_{\min}^4}\left(C^2+\sigma^2 d\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2}+T\frac{\gamma^2\left(14\zeta^2+26\kappa^2\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2}$$
$$+\frac{52\gamma^2 L^2}{(1-\alpha)^2(1-\sqrt{\rho})^2}\sum_{t=1}^{T}\frac{1}{M}\mathbb{E}\left[\left\|\mathbf{X}^{[t]}(\mathbf{I}-\mathbf{Q})\right\|_{\mathfrak{F}}^2\right]$$
$$+\frac{26\gamma^2}{(1-\alpha)^2(1-\sqrt{\rho})^2}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i(x_i^{[t-1]})\right\|^2\right] \tag{68}$$

Rearranging terms and dividing the both sides of (68) by $1-\frac{52\gamma^2 L^2}{(1-\alpha)^2(1-\sqrt{\rho})^2}$, we complete the proof as follows

$$\sum_{t=1}^{T}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\bar{x}^{[t]}-x_i^{[t]}\right\|^2\right]$$
$$\leq\frac{1}{1-\frac{52\gamma^2 L^2}{(1-\alpha)^2(1-\sqrt{\rho})^2}}\left(T\frac{\frac{8\gamma^2}{\omega_{\min}^4}\left(C^2+\sigma^2 d\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2}+T\frac{\gamma^2\left(14\zeta^2+26\kappa^2\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2}\right)$$
$$+\frac{1}{1-\frac{52\gamma^2 L^2}{(1-\alpha)^2(1-\sqrt{\rho})^2}}\left(\frac{26\gamma^2}{(1-\alpha)^2(1-\sqrt{\rho})^2}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right]\right)$$
$$\overset{\text{\textcircled{1}}}{\leq}T\frac{\frac{16\gamma^2}{\omega_{\min}^4}\left(C^2+\sigma^2 d\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2}+T\frac{4\gamma^2\left(7\zeta^2+13\kappa^2\right)}{(1-\alpha)^2(1-\sqrt{\rho})^2}+\frac{52\gamma^2}{(1-\alpha)^2(1-\sqrt{\rho})^2}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right] \tag{69}$$

where $\text{\textcircled{1}}$ follows because $\gamma\leq\frac{(1-\alpha)(1-\sqrt{\rho})}{2\sqrt{26}L}$ is chosen to ensure $\left(1-\frac{52\gamma^2 L^2}{(1-\alpha)^2(1-\sqrt{\rho})^2}\right)^{-1}\leq 2$. $\qquad\square$

According to the above **Lemmas**, we now give the main proof of **Theorem** 2. Based on the $L$-smoothness of $\mathcal{F}$, we have

$$\mathbb{E}\left[\mathcal{F}\left(\bar{S}^{[t]}\right)\right]\leq\mathbb{E}\left[\mathcal{F}\left(\bar{S}^{[t-1]}\right)\right]+\mathbb{E}\left[\left\langle\nabla\mathcal{F}\left(\bar{S}^{[t-1]}\right),\bar{S}^{[t]}-\bar{S}^{[t-1]}\right\rangle\right]+\frac{L}{2}\mathbb{E}\left[\left\|\bar{S}^{[t]}-\bar{S}^{[t-1]}\right\|^2\right] \tag{70}$$

Based on **Lemma** 3, we have

$$\mathbb{E}\left[\left\langle\nabla\mathcal{F}\left(\bar{S}^{[t-1]}\right),\bar{S}^{[t]}-\bar{S}^{[t-1]}\right\rangle\right]=\frac{-\gamma}{1-\alpha}\mathbb{E}\left[\left\langle\nabla\mathcal{F}\left(\bar{S}^{[t-1]}\right),\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[t]}\right\rangle\right]$$
$$=\frac{-\gamma}{1-\alpha}\mathbb{E}\left[\left\langle\nabla\mathcal{F}\left(\bar{S}^{[t-1]}\right)-\nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right),\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[t]}\right\rangle\right]$$
$$+\frac{-\gamma}{1-\alpha}\mathbb{E}\left[\left\langle\nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right),\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[t]}\right\rangle\right]$$
$$=\underbrace{\frac{-\gamma}{1-\alpha}\mathbb{E}\left[\left\langle\nabla\mathcal{F}\left(\bar{S}^{[t-1]}\right)-\nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right),\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[t]}\right\rangle\right]}_{\text{III}}$$
$$+\underbrace{\frac{-\gamma}{1-\alpha}\mathbb{E}\left[\left\langle\nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right),\frac{1}{M}\sum_{i=1}^{M}\left(\bar{g}_i^{[t]}-g_{i,i}^{[t]}\right)\right\rangle\right]}_{\text{IV}}$$
$$+\underbrace{\frac{-\gamma}{1-\alpha}\mathbb{E}\left[\left\langle\nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right),\frac{1}{M}\sum_{i=1}^{M}g_{i,i}^{[t]}\right\rangle\right]}_{\text{V}} \tag{71}$$

which can be bounded by bounding III, IV and V. For III, we have

$$\frac{-\gamma}{1-\alpha}\mathbb{E}\left[\left\langle \nabla\mathcal{F}\left(\bar{S}^{[t-1]}\right) - \nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right), \frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[t]}\right\rangle\right]$$

$$\overset{\textcircled{1}}{\leq}\frac{1-\alpha}{2\alpha L}\mathbb{E}\left[\left\|\nabla\mathcal{F}\left(\bar{S}^{[t-1]}\right) - \nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right)\right\|^2\right] + \frac{\alpha L\gamma^2}{2(1-\alpha)^3}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[t]}\right\|^2\right]$$

$$\overset{\textcircled{2}}{\leq}\frac{(1-\alpha)L}{2\alpha}\mathbb{E}\left[\left\|\bar{S}^{[t-1]} - \bar{x}^{[t-1]}\right\|^2\right] + \frac{\alpha L\gamma^2}{2(1-\alpha)^3}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[t]}\right\|^2\right] \tag{72}$$

where $\textcircled{1}$ follows from the basic inequality $\langle a,b\rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$ with $a = \frac{\sqrt{1-\alpha}}{\sqrt{\alpha L}}\left(\nabla\mathcal{F}\left(\bar{S}^{[t-1]}\right) - \nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right)\right)$ and $b = \frac{-\gamma\sqrt{\alpha L}}{(1-\alpha)^{\frac{3}{2}}}\left(\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[t]}\right)$; and $\textcircled{2}$ based on the smoothness of function $\mathcal{F}(\cdot)$. For IV, we have

$$\frac{-\gamma}{1-\alpha}\mathbb{E}\left[\left\langle \nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right), \frac{1}{M}\sum_{i=1}^{M}\left(\bar{g}_i^{[t]} - g_{i,i}^{[t]}\right)\right\rangle\right]$$

$$\leq\frac{1-\alpha}{2\alpha}\mathbb{E}\left[\left\|\nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right)\right\|^2\right] + \frac{\alpha\gamma^2}{2(1-\alpha)^3}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\left(\bar{g}_i^{[t]} - g_{i,i}^{[t]}\right)\right\|^2\right] \tag{73}$$

by applying $\langle a,b\rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$ with $a = \frac{-\sqrt{(1-\alpha)}}{\sqrt{\alpha}}\nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right)$ and $b = \frac{\gamma\sqrt{\alpha}}{(1-\alpha)^{\frac{3}{2}}}\left(\frac{1}{M}\sum_{i=1}^{M}\left(\bar{g}_i^{[t]} - g_{i,i}^{[t]}\right)\right)$. Then, for V, we have

$$\mathbb{E}\left[\left\langle \nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right), \frac{1}{M}\sum_{i=1}^{M}g_{i,i}^{[t]}\right\rangle\right] = \mathbb{E}\left[\left\langle \nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right), \frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\rangle\right] \tag{74}$$

Therein,

$$\left\langle \nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right), \frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\rangle$$

$$\overset{\textcircled{1}}{=}\frac{1}{2}\left(\left\|\nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right)\right\|^2 + \left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2 - \left\|\nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right) - \frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right)$$

$$\overset{\textcircled{2}}{\geq}\frac{1}{2}\left(\left\|\nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right)\right\|^2 + \left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2 - \frac{L^2}{M}\sum_{i=1}^{M}\left\|\bar{x}^{[t-1]} - x_i^{[t-1]}\right\|^2\right) \tag{75}$$

where we have $\textcircled{1}$ according to $\langle a,b\rangle = \frac{1}{2}\left(\|a\|^2 + \|b\|^2 - \|a-b\|^2\right)$ with $a = \nabla\mathcal{F}(\bar{x}^{[t-1]})$ and $b = \frac{1}{M}\sum_{i=1}^{M}\nabla f_i(x_i^{[t-1]})$; $\textcircled{2}$ since

$$\left\|\nabla\mathcal{F}(\bar{x}^{[t-1]}) - \frac{1}{M}\sum_{i=1}^{M}\nabla f_i(x_i^{[t-1]})\right\|^2 = \left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i(\bar{x}^{[t-1]}) - \frac{1}{M}\sum_{i=1}^{M}\nabla f_i(x_i^{[t-1]})\right\|^2$$

$$\overset{\textcircled{1}}{\leq}\frac{1}{M}\sum_{i=1}^{M}\left\|\nabla f_i(\bar{x}^{[t-1]}) - \nabla f_i(x_i^{[t-1]})\right\|^2$$

$$\overset{\textcircled{2}}{\leq}\frac{1}{M}\sum_{i=1}^{M}L^2\left\|\bar{x}^{[t-1]} - x_i^{[t-1]}\right\|^2$$

where $\textcircled{1}$ due to the convexity of $\|\cdot\|^2$ and *Jensen's Inequality*; and $\textcircled{2}$ according to the smoothness of $f_i(\cdot)$. By substituting (75) into (74), we obtain the bound of V as follows

$$\frac{-\gamma}{1-\alpha}\mathbb{E}\left[\left\langle \nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right), \frac{1}{M}\sum_{i=1}^{M}g_{i,i}^{[t]}\right\rangle\right] \leq \frac{-\gamma}{2(1-\alpha)}\mathbb{E}\left[\left\|\nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right)\right\|^2\right]$$

$$+ \frac{-\gamma}{2(1-\alpha)} \mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M} \nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right]$$

$$- \frac{-\gamma L^2}{2M(1-\alpha)}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\bar{x}^{[t-1]} - x_i^{[t-1]}\right\|^2\right] \tag{76}$$

By combining the bounds of III, IV, and V (see (72), (73) and (76), respectively), we obtain the bound of (71), then we continue to bound $\mathbb{E}\left[\mathcal{F}\left(\bar{S}^{[t]}\right)\right]$ based on (70) as follows

$$\mathbb{E}\left[\mathcal{F}\left(\bar{S}^{[t]}\right)\right] \le \mathbb{E}\left[\mathcal{F}\left(\bar{S}^{[t-1]}\right)\right] + \frac{(1-\alpha)L}{2\alpha}\mathbb{E}\left[\left\|\bar{S}^{[t-1]} - \bar{x}^{[t-1]}\right\|^2\right]$$

$$+ \left(\frac{\alpha L\gamma^2}{2(1-\alpha)^3} + \frac{L\gamma^2}{2(1-\alpha)^2}\right)\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[t]}\right\|^2\right]$$

$$+ \left(\frac{1-\alpha}{2\alpha} - \frac{\gamma}{2(1-\alpha)}\right)\mathbb{E}\left[\left\|\nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right)\right\|^2\right]$$

$$+ \frac{\alpha\gamma^2}{2(1-\alpha)^3}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\left(\bar{g}_i^{[t]} - g_{i,i}^{[t]}\right)\right\|^2\right]$$

$$- \frac{\gamma}{2(1-\alpha)}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right]$$

$$+ \frac{\gamma L^2}{2(1-\alpha)}\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\bar{x}^{[t-1]} - x_i^{[t-1]}\right\|^2\right] \tag{77}$$

By rearranging (77), we have

$$\mathbb{E}\left[\left\|\nabla\mathcal{F}\left(\bar{x}^{[t-1]}\right)\right\|^2\right] \le \frac{1}{m_1}\left(\mathbb{E}\left[\mathcal{F}\left(\bar{S}^{[t-1]}\right)\right] - \mathbb{E}\left[\mathcal{F}\left(\bar{S}^{[t]}\right)\right]\right)$$

$$+ m_2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[t]}\right\|^2\right] + m_3\mathbb{E}\left[\left\|\bar{S}^{[t-1]} - \bar{x}^{[t-1]}\right\|^2\right]$$

$$+ m_4\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\left(\bar{g}_i^{[t]} - g_{i,i}^{[t]}\right)\right\|^2\right]$$

$$+ \frac{m_5}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\bar{x}^{[t-1]} - x_i^{[t-1]}\right\|^2\right]$$

$$- m_6\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\nabla f_i\left(x_i^{[t-1]}\right)\right\|^2\right] \tag{78}$$

where $m_1 = \frac{\gamma}{2(1-\alpha)} - \frac{1-\alpha}{2\alpha}$, $m_2 = \frac{1}{m_1}\left(\frac{\alpha L\gamma^2}{2(1-\alpha)^3} + \frac{L\gamma^2}{2(1-\alpha)^2}\right)$, $m_3 = \frac{L(1-\alpha)}{2m_1\alpha}$, $m_4 = \frac{\alpha\gamma^2}{2m_1(1-\alpha)^3}$, $m_5 = \frac{L^2\gamma}{2m_1(1-\alpha)}$, $m_6 = \frac{\gamma}{2m_1(1-\alpha)}$. Summing over $t = 1, \cdots, T$, we have

$$\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\mathcal{F}(\bar{x}^{[t-1]})\right\|^2\right] \le \frac{1}{m_1}\left(\mathbb{E}\left[\mathcal{F}\left(\bar{S}^{[0]}\right)\right] - \mathbb{E}\left[\mathcal{F}(\bar{S}^{[T]})\right]\right)$$

$$+ m_2\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\bar{g}_i^{[t]}\right\|^2\right]$$

$$+ m_3\sum_{t=1}^{T}\mathbb{E}\left[\left\|\bar{S}^{[t-1]} - \bar{x}^{[t-1]}\right\|^2\right]$$

$$+ m_4 \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \frac{1}{M} \sum_{i=1}^{M} \left( \bar{g}_i^{[t]} - g_{i,i}^{[t]} \right) \right\|^2 \right]$$

$$+ m_5 \sum_{t=1}^{T} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E} \left[ \left\| \bar{x}^{[t-1]} - x_i^{[t-1]} \right\|^2 \right]$$

$$- m_6 \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \frac{1}{M} \sum_{i=1}^{M} \nabla f_i \left( x_i^{[t-1]} \right) \right\|^2 \right] \tag{79}$$

By combining **Lemma** 1, **Lemma** 2, **Lemma** 4 and **Lemma** 6 in (79), we have,

$$\sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla \mathcal{F}(\bar{x}^{[t-1]}) \right\|^2 \right]$$

$$\leq \frac{1}{m_1} \left( \mathbb{E} \left[ \mathcal{F} \left( \bar{S}^{[0]} \right) \right] - \mathbb{E} \left[ \mathcal{F} \left( \bar{S}^{[T]} \right) \right] \right)$$

$$+ T \left( m_2 + m_3 \frac{\gamma^2 \alpha^2}{(1-\alpha)^4} + m_4 \right) \left( \frac{4C^2}{\omega_{\min}^4} + \frac{4\sigma^2 d}{\omega_{\min}^4} + \frac{2\zeta^2}{M} \right)$$

$$+ T m_5 \left( \frac{\frac{16\gamma^2}{\omega_{\min}^4} \left( C^2 + \sigma^2 d \right)}{(1-\alpha)^2 (1-\sqrt{\rho})^2} + \frac{4\gamma^2 \left( 7\zeta^2 + 13\kappa^2 \right)}{(1-\alpha)^2 (1-\sqrt{\rho})^2} \right)$$

$$- \left( m_6 - 2m_4 - m_5 \frac{52\gamma^2}{(1-\alpha)^2 (1-\sqrt{\rho})^2} \right) \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \frac{1}{M} \sum_{i=1}^{M} \nabla f_i \left( x_i^{[t-1]} \right) \right\|^2 \right] \tag{80}$$

Dividing both sides of (80) by $T$ and employing the fact that $\bar{S}^{[0]} = \bar{x}^{[0]}$, we have

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla \mathcal{F} \left( \bar{x}^{[t-1]} \right) \right\|^2 \right]$$

$$\leq \frac{1}{m_1 T} \left( \mathcal{F} \left( \bar{x}^{[0]} \right) - \mathcal{F}^* \right)$$

$$+ \left( m_2 + m_3 \frac{\gamma^2 \alpha^2}{(1-\alpha)^4} + m_4 \right) \left( \frac{4C^2}{\omega_{\min}^4} + \frac{4\sigma^2 d}{\omega_{\min}^4} + \frac{2\zeta^2}{M} \right)$$

$$+ m_5 \left( \frac{\frac{16\gamma^2}{\omega_{\min}^4} \left( C^2 + \sigma^2 d \right)}{(1-\alpha)^2 (1-\sqrt{\rho})^2} + \frac{4\gamma^2 \left( 7\zeta^2 + 13\kappa^2 \right)}{(1-\alpha)^2 (1-\sqrt{\rho})^2} \right) \tag{81}$$

when

$$m_6 - 2m_4 - m_5 \frac{52\gamma^2}{(1-\alpha)^2 (1-\sqrt{\rho})^2} \geq 0 \tag{82}$$

To ensure the inequality (82), we have

$$m_1 = \frac{\gamma}{2(1-\alpha)} - \frac{1-\alpha}{2\alpha} > 0 \tag{83}$$

and

$$1 - \frac{2\alpha\gamma}{(1-\alpha)^2} - \frac{52L^2\gamma^2}{(1-\alpha)^2 (1-\sqrt{\rho})^2} \geq 0 \tag{84}$$

therefore, the step size $\gamma$ can be defined as

$$\frac{(1-\alpha)^2}{\alpha} < \gamma \leq \min \left\{ \frac{(1-\alpha)(1-\sqrt{\rho})}{2\sqrt{26}L}, \frac{(1-\sqrt{\rho})\sqrt{52L^2(1-\alpha)^2 + \alpha^2(1-\sqrt{\rho})^2}}{52L^2} - \frac{\alpha(1-\sqrt{\rho})^2}{52L^2} \right\} \tag{85}$$

We continue to present the proof of **Corollary** 1. If letting $\gamma = \mathcal{O} \left( \sqrt{\frac{1}{T}} \right)$, we have the constants in (33) represented by

$$m_1 = \mathcal{O} \left( \sqrt{\frac{1}{T}} \right), \ m_2 = \mathcal{O} \left( \sqrt{\frac{1}{T}} \right), \ m_3 = \mathcal{O} \left( \sqrt{T} \right), \ m_4 = \mathcal{O} \left( \sqrt{\frac{1}{T}} \right), \ m_5 = \mathcal{O} \left( 1 \right), \ \text{and} \ m_6 = \mathcal{O} \left( 1 \right) \tag{86}$$

Then, for each term in the right hand side of (81), we have

$$\frac{1}{m_1 T} \left( \mathcal{F} \left( \bar{x}^{[0]} \right) - \mathcal{F}^* \right) = \mathcal{O} \left( \frac{1}{\sqrt{T}} \right), \tag{87}$$

$$\left( m_2 + \frac{m_3 \gamma^2 \alpha^2}{(1-\alpha)^4} + m_4 \right) \left( \frac{4C^2}{\omega_{\min}^4} + \frac{4\sigma^2 d}{\omega_{\min}^4} + \frac{2\zeta^2}{M} \right) = \mathcal{O} \left( \frac{1}{\sqrt{T}} \right) + \mathcal{O} \left( \frac{\sigma^2 d}{\sqrt{T}} \right) + \mathcal{O} \left( \frac{1}{M\sqrt{T}} \right), \tag{88}$$

and

$$m_5 \left( \frac{\frac{16\gamma^2}{\omega_{\min}^4} \left( C^2 + \sigma^2 d \right)}{(1-\alpha)^2 (1 - \sqrt{\rho})^2} + \frac{4\gamma^2 \left( 7\zeta^2 + 13\kappa^2 \right)}{(1-\alpha)^2 (1 - \sqrt{\rho})^2} \right) = \mathcal{O} \left( \frac{\sigma^2 d}{T} \right) + \mathcal{O} \left( \frac{1}{T} \right) \tag{89}$$

by combining which, we can conclude that there exists constant $K$ such that

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla \mathcal{F} \left( \bar{x}^{[t-1]} \right) \right\|^2 \right] \leq K \left( \frac{1}{\sqrt{T}} + \frac{\sigma^2 d}{\sqrt{T}} + \frac{1}{M\sqrt{T}} + \frac{1}{T} + \frac{\sigma^2 d}{T} \right) \tag{90}$$

If we fix $M$ and $\sigma$, let $T$ be sufficiently large, the convergence rate of **Algorithm** 1 can be characterized by $\mathcal{O} \left( \frac{1}{\sqrt{T}} \right)$.