每个local model training通过SAM寻找更平坦的求解平面，再内嵌多轮gossip round，即与邻居节点多次整合，来取得更好的一致性，类似GNN的传递平均导致过平滑。

Improving the Model Consistency of Decentralized Federated Learning

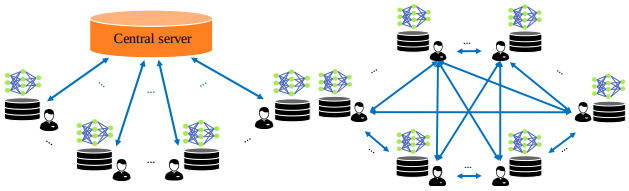
Yifan Shi [1] **Li Shen** [2] **Kang Wei** [3] **Yan Sun** [4] **Bo Yuan** [1] **Xueqian Wang** [1] **Dacheng Tao** [4]

Abstract

To mitigate the privacy leakages and communication burdens of Federated Learning (FL), decentralized FL (DFL) discards the central server and each client only communicates with its neighbors in a decentralized communication network. However, existing DFL suffers from high inconsistency among local clients, which results in severe distribution shift and inferior performance compared with centralized FL (CFL), especially on heterogeneous data or sparse communication topologies. To alleviate this issue, we propose two DFL algorithms named DFedSAM and DFedSAM-MGS to improve the performance of DFL. Specifically, DFedSAM leverages gradient perturbation to generate local flat models via Sharpness Aware Minimization (SAM), which searches for models with uniformly low loss values. DFedSAM-MGS further boosts DFedSAM by adopting Multiple Gossip Steps (MGS) for better model consistency, which accelerates the aggregation of local flat models and better balances communication complexity and generalization. Theoretically, we present improved convergence rates $\mathcal{O}(\frac{1}{\sqrt{KT}}+\frac{1}{T}+\frac{1}{K^{1/2}T^{3/2}(1-\lambda)^2})$ and $\mathcal{O}(\frac{1}{\sqrt{KT}}+\frac{1}{T}+\frac{\lambda^Q+1}{K^{1/2}T^{3/2}(1-\lambda^Q)^2})$ in non-convex setting for DFedSAM and DFedSAM-MGS, respectively, where $1-\lambda$ is the spectral gap of gossip matrix and Q is the number of MGS. Empirically, our methods can achieve competitive performance compared with CFL methods and outperform existing DFL methods.

[1]Tsinghua University, Shenzhen, China [2]JD Explore Academy, Beijing, China [3]Hong Kong Polytechnic University, Hong Kong, China [4]The University of Sydney, Australia. Correspondence to: Li Shen <mathshenli@gmail.com>, Xueqian Wang <wang.xq@sz.tsinghua.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.



(a) CFL framework (b) DFL framework

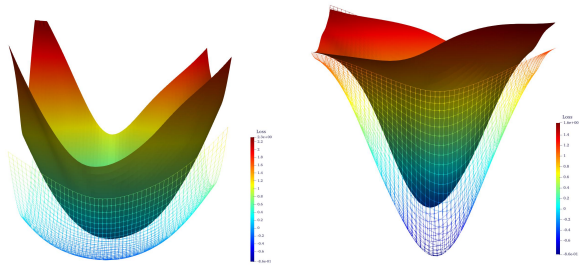
Figure 1. Illustrations of CFL (a) and DFL (b). For DFL, the various communication topologies are shown in **Appendix** A.

1. Introduction

Federated learning (FL) (Mcmahan et al., 2017; Li et al., 2020b) allows distributed clients to collaboratively train a shared model under the orchestration of the cloud without transmitting local data. However, almost all FL paradigms employ a central server to communicate with clients, which faces several critical challenges, such as computational resources limitation, high communication bandwidth cost, and privacy leakage (Kairouz et al., 2021). Compared to the centralized FL (CFL, Figure 1(a)), decentralized FL (DFL, Figure 1(b)), where the clients only communicate with their neighbors without a central server, offers communication advantage and further preserves the data privacy (Kairouz et al., 2021; Wang et al., 2021; Sun et al., 2022).

Due to the participants with different hardware and network capabilities in the real federated system, DFL is a promising field of research that has been frequently considered as a challenge in several review articles in recent years (Beltrán et al., 2022; Kairouz et al., 2021). In practice, the most promising DFL application scenarios include healthcare (Nguyen et al., 2022), industry 4.0 such as the blockchain system (Kang et al., 2022; Li et al., 2022), mobile services in the internet-of-things (IoT) (Wang et al., 2022b), the robust networks for Unmanned Aerial Vehicles (UAVs) (Wang et al., 2020) and internet-of-vehicles (Yu et al., 2020).

However, DFL suffers from severe inconsistency among local models due to heterogeneous data distribution and model aggregation locality caused by intrinsic network connectivity. This inconsistency may result in severe over-fitting in local models and performance degradation (Sun et al., 2022). To explore the mechanism behind this phenomenon, we present the structure of the loss landscapes (Li et al.,



(a) Fashion-MNIST (b) CIFAR-10

Figure 2. Loss landscapes comparison between CFL and DFL on Fashion-MNIST (Xiao et al., 2017) and CIFAR-10 (Krizhevsky et al., 2009). DFedAvg (surface plot) features a sharper landscape than FedAvg (mesh plot) with poorer generalization ability.

2018) of FedAvg (Mcmahan et al., 2017) and decentralized FedAvg (DFedAvg, Sun et al. (2022)) on Fashion-MNIST (Xiao et al., 2017) and CIFAR-10 (Krizhevsky et al., 2009) with the same setting in Figure 2. It is clear that DFL has a sharper landscape than CFL.

Motivation. Most FL algorithms face the over-fitting issue of local models on heterogeneous data. And many solutions (Sahu et al., 2018; Li et al., 2020c; Karimireddy et al., 2020; Yang et al., 2021; Acar et al., 2021; Wang et al., 2022a) have been proposed for CFL. In DFL, this issue is further exacerbated due to the sharp loss landscape caused by the inconsistency of local models (see Figure 2). Therefore, the performance of decentralized schemes is generally worse than that of centralized schemes with the same setting (Sun et al., 2022). Consequently, an important question is: *can we design DFL algorithms that can mitigate the inconsistency among local models and achieve similar performance to its centralized counterpart?*

To answer this question, we propose two DFL algorithms: DFedSAM and DFedSAM-MGS. Specifically, DFedSAM overcomes the local over-fitting issue via gradient perturbation with SAM (Foret et al., 2021) in each client to generate local flat models. Since each client aggregates the flat models from its neighbors, a potential flat aggregated model can be generated, which results in good generalization ability. To further boost the performance of DFedSAM, DFedSAM-MGS integrates multiple gossip steps (MGS) (Ye et al., 2020; Ye & Zhang, 2021; Li et al., 2020a) to accelerate the aggregation of local flatness models by increasing the number of gossip steps of local communications. It achieves a better trade-off between communication complexity and learning performance by bridging the gap between CFL and DFL since DFL can be roughly regarded as CFL with a sufficient number of gossip steps (**Section** 5.4).

Theoretically, we present the convergence rates for our algorithms in the stochastic non-convex setting. We show that the bound can be looser when the connectivity of the communication topology λ is sufficiently sparse, or the data homogeneity β is sufficiently large. Meanwhile, as the number of consensus/gossip steps Q in MGS increases, the bound tends to be tighter as the impact of communication topology can be alleviated (**Section** 4). The theoretical results explain why the adoption of SAM and MGS in DFL can ensure better performance with various types of communication network topology. Empirically, we conduct extensive experiments on CIFAR-10 and CIFAR-100 datasets in both identical data distribution (IID) and non-IID settings. The experimental results confirm that our algorithms can achieve competitive performance compared to CFL baselines and outperform existing DFL baselines (**Section** 5.2).

Contribution. Our main contributions are summarized as:

- We propose two effective DFL schemes: DFedSAM and DFedSAM-MGS. DFedSAM reduces the inconsistency of local models with local flat models, and DFedSAM-MGS further improves the consistency via MGS acceleration and features a better trade-off between communication and generalization.
- We present improved convergence rates $\mathcal{O}(\frac{1}{\sqrt{KT}} + \frac{1}{T} + \frac{1}{K^{1/2}T^{3/2}(1-\lambda)^2})$ and $\mathcal{O}(\frac{1}{\sqrt{KT}} + \frac{1}{T} + \frac{\lambda^Q+1}{K^{1/2}T^{3/2}(1-\lambda^Q)^2})$ for DFedSAM and DFedSAM-MGS in the non-convex settings, respectively, which theoretically verify the effectiveness of our approaches.
- We conduct extensive experiments to demonstrate the efficacy of DFedSAM and DFedSAM-MGS, which can achieve competitive performance compared with both CFL and DFL baselines.

2. Related Work

Decentralized Federated Learning (DFL). In DFL, clients only communicate with their neighbors in various communication networks without a central server, which offers communication advantage and better preserves data privacy in comparison to CFL. Lalitha et al. (2018; 2019) take a Bayesian-like approach to introduce a belief over the model parameter space of the clients in a fully DFL framework. Roy et al. (2019) propose the first server-less, peer-to-peer FL approach BrainTorrent and apply it to medical applications in a highly dynamic peer-to-peer FL environment. Sun et al. (2022) apply the multiple local iterations with SGD and quantization method to reduce the communication cost and provide the convergence results in various convex settings. Dai et al. (2022) develop a decentralized sparse training technique to further lower the communication and computation cost. Our work focuses on DFL (Lalitha et al., 2018; 2019; Roy et al., 2019; Sun et al., 2022; Dai et al.,

2022) rather than decentralized training (Lian et al., 2017; Ye et al., 2020; Li et al., 2020a; Chen et al., 2021; Yuan et al., 2021; Ye & Zhang, 2021; Warnat-Herresthal et al., 2021; Koloskova et al., 2020; Hashemi et al., 2022; Zhang et al., 2022) [1].

Sharpness Aware Minimization (SAM). SAM (Foret et al., 2021) is an effective optimizer for training deep learning models, which leverages the flat geometry of the loss landscape to improve model generalization ability. Recently, Andriushchenko & Flammarion (2022) study the properties of SAM and provide convergence results of SAM for nonconvex objectives. As an effective optimizer, SAM and its variants have been applied to various machine learning (ML) tasks (Zhao et al., 2022; Kwon et al., 2021; Du et al., 2021; Liu et al., 2022; Abbas et al., 2022; Shi et al., 2023; Zhong et al., 2022; Mi et al., 2022). For instance, Qu et al. (2022) and Caldarola et al. (2022) adopt SAM to improve generalization, and thus mitigate the distribution shift problem and achieve SOTA performance for CFL (Sun et al., 2023; Qu et al., 2022). However, to the best of our knowledge, few if any efforts have been devoted to the empirical performance and theoretical analysis of SAM in the DFL setting.

Multiple Gossip Steps (MGS). The advantage of increasing the frequency of local communications within a network topology is investigated in FastMix (Ye et al., 2020), in which the optimal computational complexity and near-optimal communication complexity are established. DeEPCA (Ye & Zhang, 2021) integrates FastMix into a decentralized PCA algorithm to accelerate the training process. DeLi-CoCo (Hashemi et al., 2022) performs multiple compression gossip steps in each iteration for fast convergence with arbitrary communication compression. Network-DANE (Li et al., 2020a) uses multiple gossip steps and generalizes DANE to decentralized scenarios. In general, by increasing the number of gossip steps, local clients can reach a better consensus model to improve the performance of CFL. However, MGS has yet to be explored to mitigate the model inconsistency in the DFL setting.

The work most related to this paper is DFedAvg and DFedAvg with momentum (DFedAvgM) in Sun et al. (2022), which leverage multiple local iterations with the SGD optimizer and significantly improve the performance of classic decentralized parallel SGD method D-PSGD (Lian et al., 2017). However, DFL may still suffer from inferior performance due to the severe model inconsistency among clients. Another related work is FedSAM (Qu et al., 2022), which integrates SAM into CFL to enhance the flatness of local model and achieves new SOTA performance for CFL. On top of the aforementioned studies, we extend the SAM optimizer to the DFL setting and simultaneously provide its convergence guarantee in the nonconvex setting. Furthermore, we bride the gap between CFL and DFL via adopting MGS in DFedSAM-MGS, which largely mitigates the model inconsistency in DFL.

[1] In this work, DFL refers to local training with multiple local iterates, whereas decentralized learning/training focuses on one-step local training. For instance, D-PSGD (Lian et al., 2017) is a decentralized training algorithm, which uses the one-step SGD to train local models in each communication round.

3. Methodology

In this section, we introduce the problem setting in DFL and present the details of the proposed algorithms.

3.1. Problem Setting

In this work, we are interested in solving the following finite-sum stochastic non-convex minimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) := \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), \quad f_i(\mathbf{x}) = \mathbb{E}_{\xi\sim\mathcal{D}_i} F_i(\mathbf{x};\xi), \tag{1}$$

where $\mathcal{D}_i$ denotes the data distribution in the i-th client, which is heterogeneous across clients; m is the number of clients, and $F_i(\mathbf{x};\xi)$ is the local objective function associated with data samples ξ. Equation (1) is known as the empirical risk minimization (ERM) with many applications in ML. In Figure 1(b), the communication network in the decentralized network topology among clients is modeled as an undirected connected graph $\mathcal{G} = (\mathcal{N}, \mathcal{V}, \boldsymbol{W})$, where $\mathcal{N} = \{1, 2, \ldots, m\}$ refers to the set of clients, and $\mathcal{V} \subseteq \mathcal{N} \times \mathcal{N}$ refers to the set of communication channels, each connecting two distinct clients. Furthermore, there is no central server in the decentralized setting and all clients only communicate with their neighbors via the communication channels $\mathcal{V}$. In addition, we assume that Equation (1) is well-defined and denote f^* as the minimal value of f: $f(x) \geq f(x^*) = f^*$ for all $x \in \mathbb{R}^d$.

The Challenges in DFL. With the absence of the central server, communication connections become an important factor for decentralized optimization. Furthermore, communication is more careful in classical FL scenarios than computation (Mcmahan et al., 2017; Li et al., 2020b; Kairouz et al., 2021; Qu et al., 2022), so that the client adopts multi-step local iterations by default due to the large communication overhead in classic FL methods such as FedAvg (Mcmahan et al., 2017). Consequently, the major technical difficulties in DFL are summarized as follows:

- *Various communication topologies.* The topology is measured by the spectral gap $1 - \lambda \in (0, 1]$ of $\mathbf{W}$, and the value of λ increases as the connectivity is more sparse. It has a significant negative impact on model training (convergence rate and generalization ability), especially on heterogeneous data or in face of sparse connectivity of communication networks, such as Ring

topology and Grid topology where $\lambda \approx 16\pi^2/(3m^2)$ and $\lambda = \mathcal{O}(1/(m\log_2(m)))$, with m being the size of the clients, respectively (Sun et al., 2022; Zhu et al., 2022).

- *Multi-step local iterations.* After multiple local iterations, the gradient estimation may tend to be biased. The implication is that the corresponding theoretical analysis may be more difficult and the empirical efficacy may also suffer compared to the one-step local iteration.

Algorithm 1 DFedSAM and DFedSAM-MGS

Input : Total number of clients m, total number of communication rounds T, the number of consensus steps per gradient iteration Q, learning rate η, and total number of the local iterates are K.
Output : The consensus model $\mathbf{x}^T$ after the final communication of all clients.

Initialization: Randomly initialize each model $\mathbf{x}^0(i)$.
for $t = 0$ **to** $T-1$ **do**
  for *node* i *in parallel* **do**
    for $k = 0$ **to** $K-1$ **do**
      Set $\mathbf{y}^{t,0}(i) \leftarrow \mathbf{x}^t(i)$, $\mathbf{y}^{t,-1}(i) = \mathbf{y}^{t,0}(i)$
      Sample a batch of local data ξ_i and calculate local gradient $\mathbf{g}^{t,k}(i) = \nabla F_i(\mathbf{y}^{t,k};\xi_i)$
      $\tilde{\mathbf{g}}^{t,k}(i) = \nabla F_i(\mathbf{y}^{t,k} + \delta(\mathbf{y}^{t,k});\xi_i)$ with $\delta(\mathbf{y}^{t,k}) = \rho\mathbf{g}^{t,k}(i)/\left\|\mathbf{g}^{t,k}(i)\right\|_2$
      $\mathbf{y}^{t,k+1}(i) = \mathbf{y}^{t,k}(i) - \eta\tilde{\mathbf{g}}^{t,k}(i)$
    end
    $\mathbf{z}^t(i) \leftarrow \mathbf{y}^{t,K}(i)$
    Receive neighbors' models $\mathbf{z}^t(l)$ from neighborhood set $\mathcal{S}_{k,t}$ with adjacency matrix $\boldsymbol{W}$.
    $\mathbf{x}^{t+1}(i) = \sum_{l\in\mathcal{N}(i)} w_{i,l}\mathbf{z}^t(l)$
    for $q = 0$ **to** $Q-1$ **do**
      $\mathbf{x}^{t,q+1}(i) = \sum_{l\in\mathcal{N}(i)} \boldsymbol{w}_{i,l}\mathbf{z}^{t,q}(l)$ $(\mathbf{z}^{t,0}(i) = \mathbf{z}^t(i))$
      $\mathbf{z}^{t,q+1}(i) = \mathbf{x}^{t,q+1}(i)$
    end
    $\mathbf{x}^{t+1}(i) = \mathbf{x}^{t,Q}(i)$
  end
end

SAM，寻找更平坦的解平面，通过对模型添加扰动，优化这个扰动后的模型

3.2. DFedSAM and DFedSAM-MG

Instead of searching for a solution via SGD (Bottou, 2010; Bottou et al., 2018), SAM (Foret et al., 2021) aims to seek a solution in a flat region by adding a small perturbation to the models, i.e., $x + \delta$ with more robust performance. As shown in Figure 2, the decentralized scheme has a sharper landscape with poorer generalization ability compared with the centralized scheme. In this paper, we incorporate the SAM optimizer into DFL to explore this feature, and the local loss function of the proposed DFedSAM is defined as:

$$f_i(\mathbf{x}) = \mathbb{E}_{\xi\sim\mathcal{D}_i} \max_{\|\delta_i\|_2\le\rho} F_i(\mathbf{y}^{t,k}(i) + \delta_i;\xi_i), \quad i \in \mathcal{N} \tag{2}$$

添加的扰动不超过一定半径

where $\mathbf{y}^{t,k}(i) + \delta_i$ is the perturbed model, and ρ is a predefined constant controlling the radius of the perturbation and $\|\cdot\|_2$ is a l_2-norm, which is simplified to $\|\cdot\|$ in the rest. Similar to CFL methods, in DFedSAM, clients can update local model parameters with multiple local iterates before conducting communication. Specifically, for each client $i \in \{1, 2, ..., m\}$, in each local iteration $k \in \{0, 1, ..., K-1\}$ and each communication round $t \in \{0, 1, ..., T-1\}$, the k-th inner iteration in client i is performed as:

$$\mathbf{y}^{t,k+1}(i) = \mathbf{y}^{t,k}(i) - \eta\tilde{\mathbf{g}}^{t,k}(i), \tag{3}$$

where $\tilde{\mathbf{g}}^{t,k}(i) = \nabla F_i(\mathbf{y}^{t,k} + \delta(\mathbf{y}^{t,k});\xi)$, $\delta(\mathbf{y}^{t,k}) = \rho\mathbf{g}^{t,k}/\left\|\mathbf{g}^{t,k}\right\|_2$, with the first order Taylor expansion around $\mathbf{y}^{t,k}$ for a small value of ρ (Foret et al., 2021). After K inner iterations in each client, parameters are updated as $\mathbf{z}^t(i) \leftarrow \mathbf{y}^{t,K}(i)$ and sent to its neighbors $l \in \mathcal{N}(i)$ after local updates. Then, each client averages its parameters with the information of its neighbors (including itself):

$$\mathbf{x}^{t+1}(i) = \sum_{l\in\mathcal{N}(i)} w_{i,l}\mathbf{z}^t(l). \tag{4}$$

Furthermore, we employ the multiple gossip steps (MGS) technique (Ye et al., 2020; Ye & Zhang, 2021; Hashemi et al., 2022) to achieve better consistency among local models, named DFedSAM-MGS, to ensure a balance between the communication cost and generalization ability in DFL. Specifically, the generation of MGS at the q-th step $(q \in \{0, 1, ..., Q-1\})$ can be viewed as two steps in terms of exchanging messages and local gossip update as follows:

类似多轮comm round，整合邻居节点的模型参数

$$\mathbf{x}^{t,q+1}(i) = \sum_{l\in\mathcal{N}(i)} \boldsymbol{w}_{i,l}\mathbf{z}^{t,q}(l), \text{ and } \mathbf{z}^{t,q+1}(i) = \mathbf{x}^{t,q+1}(i). \tag{5}$$

At the end of MGS, it holds that $\mathbf{x}^{t+1}(i) = \mathbf{x}^{t,Q}(i)$. Our proposed algorithms are summarized in Algorithm 1. It is clear that, in DFedSAM, there is a trade-off between the local computation complexity and communication overhead via multiple local iterations, whereas the local communication is only performed at once. By contrast, DFedSAM-MGS performs multiple local communications with a larger Q to make all local clients synchronized. Therefore, DFedSAM-MGS can be viewed as a compromise between DFL and CFL.

Compared with existing SOTA DFL methods: DFedAvg and DFedAvgM (Sun et al., 2022), the benefits of DFedSAM and DFedSAM-MGS lie in three-fold: (i) SAM is introduced to alleviate the local over-fitting issue caused by the inconsistency among local models via seeking a flat model at each client in DFL, and can also help make the consensus model flat; (ii) In DFedSAM-MGS, MGS is used to accelerate the aggregation of local flatness models for better consistency among local models based on DFedSAM

and properly balance the communication complexity and learning performance; (iii) Furthermore, we also present the corresponding theory unifying the impact of the gradient perturbation ρ in SAM, the number of local communications Q in MGS, and the network typology λ, along with data homogeneity β upon the convergence rate in **Section** 4.

4. Convergence Analysis

In this section, we give the convergence analysis of DFedSAM and DFedSAM-MGS for the general non-convex FL setting, and the detailed proof is presented in **Appendix** D. Below, we introduce the necessary notations and assumptions.

Definition 4.1. (The gossip/mixing matrix). [Definition 1, (Sun et al., 2022)] The gossip matrix $\mathbf{W} = [w_{i,j}] \in [0,1]^{m\times m}$ is assumed to have these properties: (i) (Graph) If $i \neq j$ and $(i,j) \notin \mathcal{V}$, then $w_{i,j} = 0$, otherwise, $w_{i,j} > 0$; (ii) (Symmetry) $\mathbf{W} = \mathbf{W}^\top$; (iii) (Null space property) $\text{null}\{\mathbf{I} - \mathbf{W}\} = \text{span}\{\mathbf{1}\}$; (iv) (Spectral property) $\mathbf{I} \succeq \mathbf{W} \succ -\mathbf{I}$. With these properties, the eigenvalues of W can be shown to satisfy $1 = |\lambda_1(\mathbf{W}))| > |\lambda_2(\mathbf{W}))| \geq \cdots \geq |\lambda_m(\mathbf{W}))|$. Furthermore, $\lambda := \max\{|\lambda_2(\mathbf{W})|, |\lambda_m(\mathbf{W}))|\}$ and $1-\lambda \in (0,1]$ is the denoted as the spectral gap of $\mathbf{W}$.

Definition 4.2. (Homogeneity parameter). [Definition 2, (Li et al., 2020a)] For any $i \in \{1,2,\ldots,m\}$ and the parameter $\mathbf{x} \in \mathbb{R}^d$, the homogeneity parameter β can be defined as:

$$\beta := \max_{1\leq i\leq m} \beta_i, \quad with\, \beta_i := \sup_{\mathbf{x}\in\mathbb{R}^d} \|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|.$$

Assumption 4.3. (Lipschitz smoothness). The function f_i is differentiable and ∇f_i is L-Lipschitz continuous, $\forall i \in \{1,2,\ldots,m\}$, i.e., $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|$, for all $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$.

Assumption 4.4. (Bounded variance). The gradient of the function f_i have σ_l-bounded variance:

$$\mathbb{E}_{\xi_i} \|\nabla F_i(\mathbf{y};\xi_i) - \nabla f_i(\mathbf{x})\|^2 \leq \sigma_l^2, \forall i \in \{1,2,\ldots,m\},$$

and the global variance is also bounded, i.e., $\frac{1}{m}\sum_{i=1}^m \|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq \sigma_g^2$ for all $\mathbf{x} \in \mathbb{R}^d$. It is not hard to verify that the σ_g is smaller than the homogeneity parameter β, i.e., $\sigma_g^2 \leq \beta^2$.

Note that above mentioned assumptions are mild and commonly used in characterizing the convergence rate of FL (Sun et al., 2022; Ghadimi & Lan, 2013; Yang et al., 2021; Bottou et al., 2018; Yu et al., 2019; Reddi et al., 2021). Compared with classic decentralized parallel SGD methods such as D-PSGD (Lian et al., 2017), the difficulty is that $\mathbf{z}^t(i) - \mathbf{x}^t(i)$ may fail to be an unbiased gradient estimation $\nabla f_i(\mathbf{x}^t(i))$ after multiple local iterates, thereby merging the multiple local iterations is non-trivial. Furthermore, the various topologies of communication networks in DFL are quite different with SAM in CFL (Qu et al., 2022). Below, we adopt the averaged parameter $\overline{\mathbf{x}^t} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}^t(i)$ of all clients as the approximated solution of Problem (1).

Theorem 4.5. *Assume Assumptions 4.3 and 4.4 hold, and the parameters $\{\mathbf{x}^t(i)\}_{t\geq 0}$ are generated via Algorithm 1. Meanwhile, assume the learning rate of SAM in each client satisfies $0 < \eta \leq \frac{1}{10KL}$. Let $\overline{\mathbf{x}^t} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}^t(i)$ and denote $\mathbf{\Phi}(\lambda, m, Q)$ as the metric related with three parameters in terms of the number of spectral gap, the clients and multiple gossip steps:*

$$\mathbf{\Phi}(\lambda, m, Q) = \frac{\lambda^Q + 1}{(1-\lambda)^2 m^{2(Q-1)}} + \frac{\lambda^Q + 1}{(1-\lambda^Q)^2}, \tag{6}$$

Then, we have the gradient estimation of DFedSAM or DFedSAM-MGS for solving Problem (1):

$$\min_{1\leq t\leq T} \mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 \leq \frac{2[f(\overline{\mathbf{x}^1}) - f^*]}{T(\eta K - 32\eta^3K^2L^2 - 6\eta^2KL)} + \alpha(K,\rho,\eta) + \mathbf{\Phi}(\lambda,m,Q)\beta(K,\rho,\eta), \tag{7}$$

where T is the number of communication rounds and the constants are given as:

$$\alpha(K,\rho,\eta) = \frac{\eta KL}{2(\eta K - 32\eta^3K^2L^2 - 6\eta^2KL)}\Big(2KL(\frac{4K^3L^2\eta^2\rho^4}{2K-1} + 8K\eta^2(L^2\rho^2 + \sigma_g^2 + \sigma_l^2) + \frac{\rho^2}{2K-1}) + \eta(L^2\rho^2 + \sigma_l^2)\Big),$$

$$\beta(K,\rho,\eta) = \frac{\eta^4KL^3(16\eta KL + 3)}{\eta K - 32\eta^3K^2L^2 - 6\eta^2KL}\Big(2K(\frac{4K^3L^2\rho^4}{2K-1} + 8K(L^2\rho^2 + \sigma_g^2 + \sigma_l^2)) + \frac{2K\rho^2}{\eta^2(2K-1)}\Big).$$

With Theorem 4.5, we state the following convergence rates for DFedSAM and DFedSAM-MGS.

Corollary 4.6. *Let the local adaptive learning rate satisfy $\eta = \mathcal{O}(1/L\sqrt{KT})$. With the similar assumptions required in **Theorem** 4.6, and setting the perturbation parameter $\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$. The convergence rate for DFedSAM satisfies:*

$$\min_{1\leq t\leq T} \mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 = \mathcal{O}\Big(\frac{(f(\overline{\mathbf{x}^1}) - f^*) + \sigma_l^2}{\sqrt{KT}} + \frac{K(\beta^2 + \sigma_l^2)}{T} + \frac{L^2}{K^{1/2}T^{3/2}} + \frac{\beta^2 + \sigma_l^2}{K^{1/2}T^{3/2}(1-\lambda)^2}\Big).$$

Remark 4.7. Due to the effect of the radius of the perturbation ρ by using SAM, DFedSAM can achieve a better bound than the state-of-the-art (SOTA) bounds such as $\mathcal{O}\Big(\frac{1}{\sqrt{T}} + \frac{\sigma_g^2}{\sqrt{T}} + \frac{\sigma_g^2 + B^2}{(1-\lambda)^2T^{3/2}}\Big)$ in (Sun et al., 2022), where B is the upper bound of the gradient. And the reason is that the diverse impact of various communication topologies λ on convergence rate can be alleviated by the effect of the radius of the perturbation ρ, i.e., $\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$ (we set $\rho = 0.01$

Table 1. The performance (%) of all algorithms on two datasets in both IID and non-IID settings.

Task	Algorithm	Dirichlet 0.3			Dirichlet 0.6			IID		
		Train	Validation	Generalization error	Train	Validation	Generalization error	Train	Validation	Generalization error
CIFAR-10	FedAvg	99.99	82.39	17.60	99.99	84.17	15.82	99.99	84.70	15.29
	FedSAM	99.75	82.49	16.26	99.89	85.04	14.85	99.98	84.98	15.00
	D-PSGD	98.59	68.23	30.36	99.09	70.58	28.51	99.75	73.23	26.52
	DFedAvg	99.75	73.55	26.20	99.93	74.67	25.26	99.95	75.55	24.40
	DFedAvgM	99.93	79.96	19.97	99.95	81.56	17.39	99.95	82.07	17.88
	DisPFL	99.90	72.19	27.71	99.93	74.43	25.50	99.95	76.18	23.77
	DFedSAM	99.41	82.04	17.37	99.44	84.38	15.06	99.44	85.30	14.14
	DFedSAM-MGS	99.53	84.26	**15.27**	99.65	85.14	**14.51**	99.69	86.47	**13.22**
CIFAR-100	FedAvg	99.99	48.36	51.63	99.99	53.06	46.93	99.99	54.16	45.83
	FedSAM	99.99	**52.98**	**47.01**	99.99	**55.88**	**44.11**	99.99	**59.60**	**40.39**
	D-PSGD	90.72	27.98	62.74	90.15	30.62	59.53	92.19	33.64	59.55
	DFedAvg	99.56	27.62	61.94	99.56	32.82	66.74	99.68	36.77	632.91
	DFedAvgM	99.56	45.11	54.45	99.60	45.50	54.10	99.78	47.98	51.80
	DisPFL	97.20	30.15	67.05	99.48	32.44	67.04	99.69	35.98	63.71
	DFedSAM	99.87	48.66	51.21	99.85	52.70	47.15	99.97	53.12	46.85
	DFedSAM-MGS	99.92	52.37	47.55	99.95	54.91	45.04	99.97	56.15	43.82

when $T = 1000$). Note that the bound can be tighter as λ decreases, which is dominated by $\frac{\beta^2+\sigma_l^2}{K^{1/2}T^{3/2}(1-\lambda)^2}$ as $\lambda \geq 1-\frac{1}{\sqrt{T}}$, whereas as β increases, it can be degraded. Furthermore, when the smoothness is not good, which means that L is large, and the additional term $\mathcal{O}(\frac{L^2}{K^{1/2}T^{3/2}})$ can be neglected compared to other terms, which comes from the additional SGD step for smoothness via SAM local optimizer.

Corollary 4.8. *Assume $Q > 1$ and T is large enough and $\eta = \mathcal{O}(1/L\sqrt{KT})$. With the similar assumptions required in **Theorem** 4.6 and perturbation amplitude $\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$, the convergence rate for DFedSAM-MGS satisfies:*

$$\min_{1\leq t\leq T} \mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 = \mathcal{O}\Big(\frac{(f(\overline{\mathbf{x}^1}) - f^*) + \sigma_l^2}{\sqrt{KT}} + \frac{K(\beta^2+\sigma_l^2)}{T} + \frac{L^2}{K^{1/2}T^{3/2}} + \mathbf{\Phi}(\lambda, m, Q)\frac{\beta^2+\sigma_l^2}{K^{1/2}T^{3/2}}\Big).$$

Remark 4.9. The impact of the network topology $(1-\lambda)$ can be alleviated as Q increases and the number of clients m is large enough, meanwhile, the term $\frac{\lambda^Q+1}{(1-\lambda)^2 m^{2(Q-1)}}$ of $\mathbf{\Phi}(\lambda, m, Q)$ can be neglected and the term $\frac{\lambda^Q+1}{(1-\lambda^Q)^2}$ is close to 1. That means by using the proposed Q-step gossip procedure, model consistency among clients can be improved, and thus DFL with various communication topologies can be roughly viewed as CFL. Thus, the negative effect of the gradient variances σ_l^2 and β^2 can be alleviated especially on sparse network topology where λ is close to 1. In practice, a suitable step number $Q > 1$ can possibly achieve a communication-accuracy trade-off in DFL.

Remark 4.10. Note that $\mathbf{\Phi}(\lambda, m, Q) = \frac{2(\lambda+1)}{(1-\lambda)^2} \leq \frac{4}{(1-\lambda)^2}$ for $Q = 1$ and the result is the same as **Corollary** 4.6.

5. Experiments

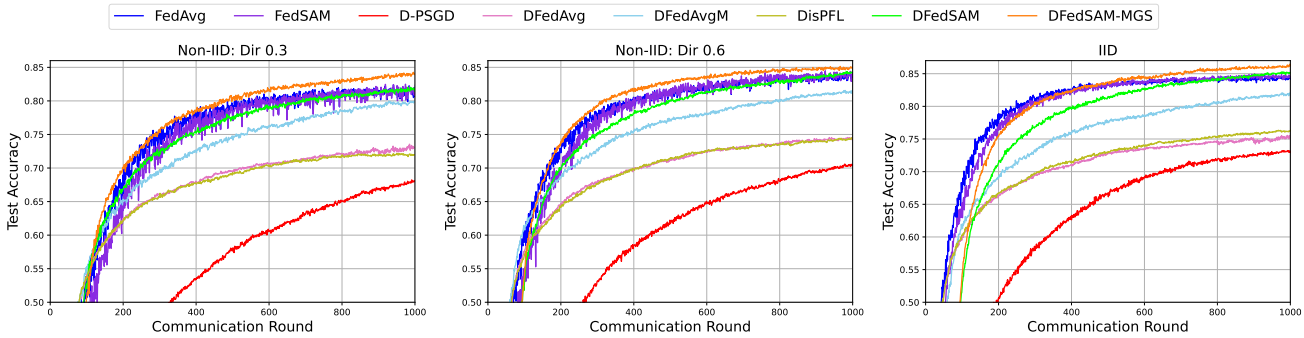
In this section, we evaluate the efficacy of our algorithms compared with six baselines from CFL and DFL settings. In addition, we conduct several experiments to verify the impact of the communication network topology in **Section** 4. Furthermore, several ablation studies are conducted.

5.1. Experiment Setup

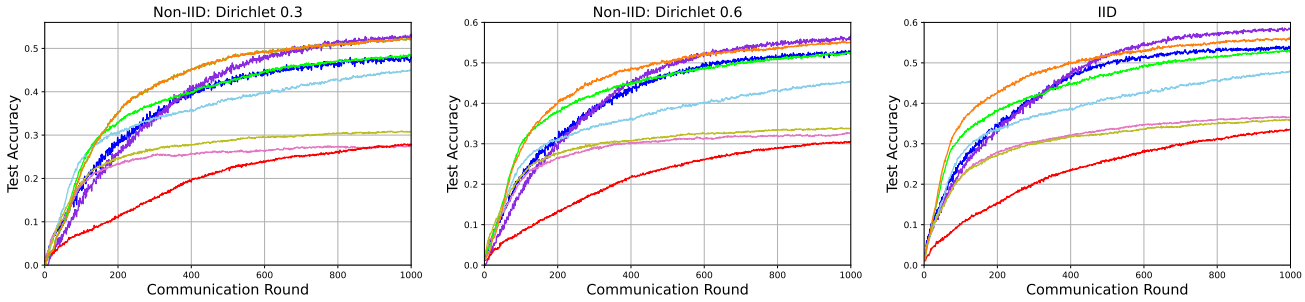
Dataset and Data Partition. The efficacy of the proposed DFedSAM and DFedSAM-MGS is evaluated on CIFAR-10 and CIFAR-100 datasets (Krizhevsky et al., 2009) in both IID and non-IID settings. Specifically, Dirichlet Partition (Hsu et al., 2019) and Pathological Partition are used for simulating non-IID across federated clients, where the former partitions the local data of each client by splitting the total dataset through sampling the label ratios from the Dirichlet distribution Dir(α) with parameters $\alpha = 0.3$ and $\alpha = 0.6$. And the Pathological Partition is placed in **Appendix** C due to limited space.

Baselines. The compared baselines cover several SOTA methods in both the CFL and DFL settings. Specifically, centralized baselines include FedAvg (Mcmahan et al., 2017) and FedSAM (Qu et al., 2022). For decentralized setting, D-PSGD (Lian et al., 2017), DFedAvg and DFedAvgM (Sun et al., 2022), along with DisPFL (Dai et al., 2022), are used for comparison.

Implementation Details. The total number of clients is set to 100, among which 10% clients participates in communication. Specifically, all clients perform the local iteration step for decentralized methods and only participated clients can perform local update for centralized methods. We ini-



(a) CIFAR-10



(b) CIFAR-100

Figure 3. Test accuracy of all baselines from both CFL and DFL with (a) CIFAR-10 and (b) CIFAR-100 in both IID and non-IID settings.

tialize the local learning rate to 0.1 with a decay rate 0.998 per communication round for all experiments. For CIFAR-10 and CIFAR-100 datasets, VGG-11 (He et al., 2016) and ResNet-18 (Simonyan & Zisserman, 2014) are adopted as the backbones in each client, respectively. The number of communication rounds is 1000 in the experiments for comparing all baselines and investigating the topology-aware performance. In addition, all ablation studies are conducted on the CIFAR-10 dataset and the number of communication rounds is set to 500.

Communication Configurations. For a fair comparison between decentralized and centralized settings, we apply a dynamic time-varying connection topology for decentralized methods to ensure that, in each round, the number of connections are no more than that in the central server. Note that the number of clients communicating with their neighbors can be controlled to keep the communication volume consistent with centralized methods. Following earlier works, the communication complexity is measured by the times of local communications. More details of the experiments are presented in **Appendix** B due to limited space.

5.2. Performance Evaluation

Performance comparison with baselines. We evaluate DFedSAM and DFedSAM-MGS ($Q = 4$) with $\rho = 0.01$ on CIFAR-10 and CIFAR-100 datasets in both settings compared with all baselines from CFL and DFL. From the results in Table 1 and Figure 3, it is clearly seen that our proposed algorithms outperform other decentralized methods on both datasets, and DFedSAM-MGS outperforms and achieves similar performance as the SOTA centralized baseline FedSAM on CIFAR-10 and CIFAR-100, respectively. Specifically, the training accuracy and testing accuracy are presented in Table 1 to show the generalization performance. We can see that the performance improvement is more obvious than all other baselines on CIFAR-10 with the same communication round. For instance, the differences between training accuracy and test accuracy (aka. generalization error) on CIFAR-10 in IID setting are 14.14% in DFedSAM, 13.22% in DFedSAM-MGS, 15.29% in FedAvg and 15% in FedSAM. That means our algorithms achieve a generalization level comparable to centralized baselines. In general, the upper-performance limitation of DFedSAM is consistent with FedSAM, as they apply different optimizers to the same optimization problem.

Impact of non-IID level (β). In Table 1, we can see that our algorithms are robust to different participation cases. The heterogeneous data distribution of local clients is set to various participation levels including IID, Dirichlet 0.6, and Dirichlet 0.3, which makes the training of the global/consensus model more difficult. On CIFAR-10, as the non-IID level increases, DFedSAM-MGS achieves better generalization than FedSAM as the generalization error in DFedSAM-MGS $\{15.27\%, 14.51\%, 13.22\%\}$ are lower than those in Fed-

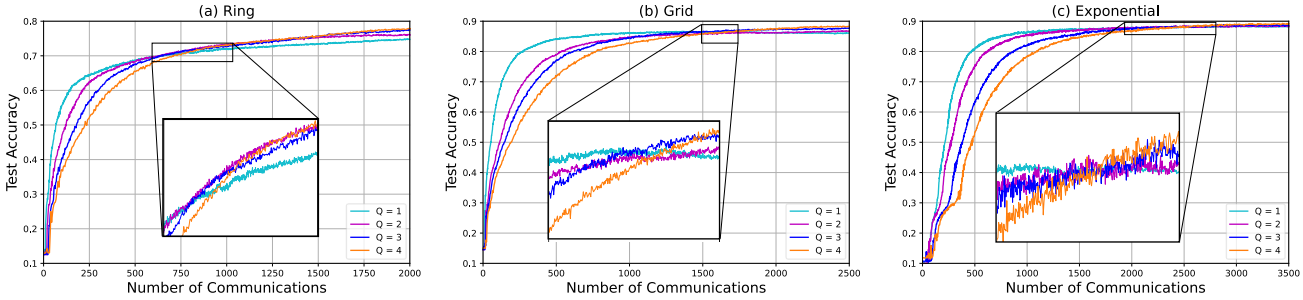


Figure 4. Test accuracy with the number of local communications in various values of Q.

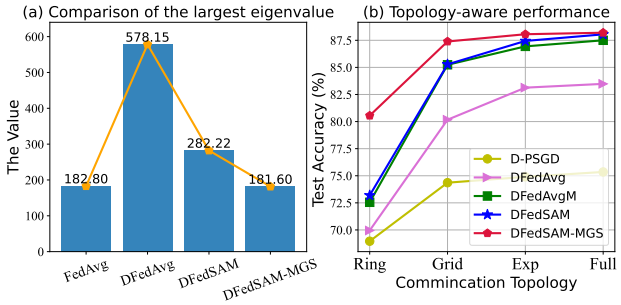


Figure 5. Comparison of the largest eigenvalue in the Hessian matrix in Figure 5(a), a known measure for flatness of loss landscape (Yao et al., 2020), and topology-aware performance of DFL methods in Figure 5(b).

SAM $\{17.26\%, 14.85\%, 15\%\}$. Similarly, the generalization error in DFedSAM $\{17.37\%, 15.06\%, 14.10\%\}$ are lower than those in FedAvg $\{17.60\%, 15.82\%, 15.27\%\}$. These observations confirm that our algorithms are more robust than baselines on various heterogeneous degrees.

Measuring on the flatness of loss landscapes. To evaluate the flatness of loss landscapes produced by our algorithms compared to FedAvg and DFedAvg (Figure 2(a) and 2(b)), we conduct some experiments on partitioned CIFAR-10 dataset with Dirichlet distribution ($\alpha = 0.6$) and VGG-11 model after all models are converged and present the largest eigenvalue of the Hessian matrix (Yao et al., 2020) in Figure 5. Note that the smaller the largest eigenvalue, the flatter the loss landscape. It is clear that the resulting models are found in flatter minima as expected and DFedSAM-MGS is close to FedAvg in terms of the largest eigenvalue.

5.3. Topology-aware Performance

We verify the influence of various communication topologies and gossip averaging steps in DFedSAM and DFedSAM-MGS. Different from the comparison of CFL and DFL in **Section** 5.2, we only need to verify the key properties for the DFL methods in this section. Thus, the communication type is set to "Complete", so that each client can communicate with its neighbors in the same communication round.

Table 2. Testing accuracy (%) in various network topologies compared with decentralized algorithms on CIFAR-10.

Algorithm	Ring	Grid	Exp	Full
D-PSGD	68.96	74.36	74.90	75.35
DFedAvg	69.95	80.17	83.13	83.48
DFedAvgM	72.55	85.24	86.94	87.50
DFedSAM	73.19 ↑	85.28↑	87.44 ↑	88.05 ↑
DFedSAM-MGS	80.55 ↑	87.39↑	88.06 ↑	88.20 ↑

The degree of sparse connectivity λ is Ring $>$ Grid $>$ Exponential (abbreviated as "Exp") $>$ Full-connected (abbreviated as "Full") in DFL. From Table 2 and Figure 5(b), our algorithms are superior to all decentralized baselines in various communication networks, which is coincided with our theoretical findings. Specifically, compared with DFedAvgM, DFedSAM, and DFedSAM-MGS can significantly improve the performance in the ring topology by 0.64% and 8.0%, respectively. Meanwhile, the performance of DFedSAM-MGS in various topologies is always better than that of other methods. This observation confirms that multiple gossip steps can alleviate the impact of network topology with a smaller $Q = 4$. Therefore, our algorithms can achieve better generalization and model consistency with various communication topologies.

5.4. Ablation Study

We verify the influence of each component and hyperparameter in DFedSAM with $Q = 1$. All the ablation studies are conducted with the "exponential" topology except the study of the impact of Q with three topologies, and the communication type is "Complete" same as **Section** 5.3.

Consensus/gossip steps Q. In Figure 4, we investigate the balance between learning performance and communication complexity with three network topologies. In general, larger Q values can better solve the local consistency problem, but may also introduce additional communication cost. In the decentralized training setting, it is easier to obtain an optimal Q. Thus, we treat Q as a hyperparameter in our experiments and investigate the different balance points for

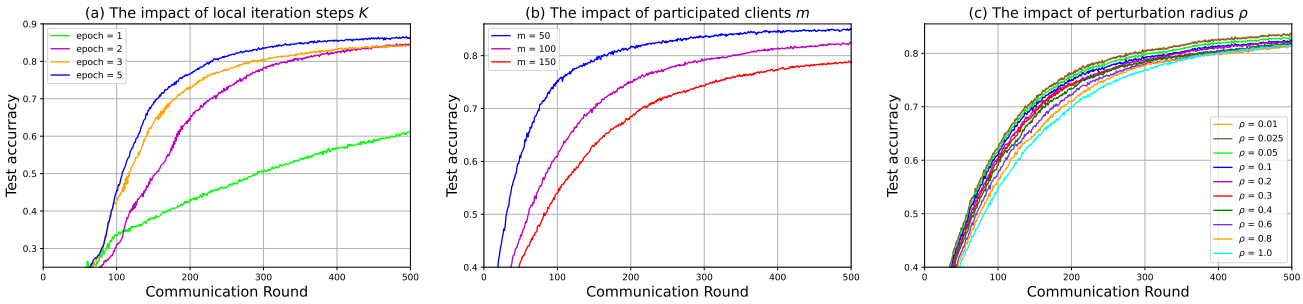


Figure 6. Impact of hyper-paramerters: local iteration steps K, participated clients m, perturbation radius ρ.

different values of $1 \leq Q \leq 4$ under various communication topologies in Figure 4 (a), (b) and (c), such as ring, grid, and exponential. As the number of local communications increases, model performance is also improved but the communication complexity increases too. It is clear that the balance points are different but with the same tendency with different topologies. Also, a relatively larger Q can bring better performance for a given communication complexity. Therefore, we choose $Q = 4$ in DFedSAM-MGS under 1000 communication rounds for a better balance.

Local iteration steps K. Large local iteration steps K can help the convergence as shown in previous DFL work (Sun et al., 2022) with theoretical guarantees. To investigate the acceleration on T by adopting a larger local iteration steps K, we fix the total batchsize and change local training epochs. As shown in Figure 6 (a), our algorithms can accelerate the convergence in theoretical results (see **Section** 4.5) as a larger K value is adopted.

Number of participated clients m. In Figure 6 (b), we compare the performance with different numbers of client participation $m = \{50, 100, 150\}$ with the same hyperparameters. Compared with $m = 150$, the m value (50 or 100) can achieve better convergence and test accuracy as the number of local data increases. We can find that a small client size tends to produce better training performance due to more samples in each client and the relationship between training performance and client size is approximately linear.

Perturbation radius ρ. The impact of the perturbation radius ρ comes from the fact that the added perturbation is accumulated when the communication round T increases. It is a trade-off between test accuracy and the generalization. To select a proper value for our algorithms, we conduct experiments with various perturbation radius values from the set $\{0.01, 0.025, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0\}$ in Figure 6 (c). With $\rho = 0.01$, we achieve a satisfactory trade-off between convergence and performance. Meanwhile, $\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$ can make a linear speedup on convergence (see **Section** 4.5).

The effectiveness of SAM and MGS. To validate the effectiveness of SAM and MGS, we can compare these algorithms including DFedAvg, DFedSAM, and FedSAM-MGS in Table 1, DFedSAM can achieve performance improvement and better generalization compared with DFedAvg as the SAM optimizer is adopted. DFedSAM-MGS can further boost the performance compared with FedSAM as MGS can also make models consistent among clients and accelerate the convergence rates.

6. Conclusions and Future Work

In this paper, we focus on the challenge of model inconsistency caused by heterogeneous data and network topology in DFL from the perspective of model generalization. We propose two DFL algorithms: DFedSAM and DFedSAM-MGS with better model consistency among clients. DFedSAM adopts SAM to produce the flat model in each client, thereby improving the generalization by generating a consensus/global flat model. DFedSAM-MGS further improves the model consistency based on DFedSAM by accelerating the aggregation of local flat models and reaching a better trade-off between learning performance and communication complexity. For theoretical findings, we unify the impacts of gradient perturbation in SAM, local communications in MGS, and network topology, along with data homogeneity upon the convergence rate in DFL. Furthermore, empirical results also verify the superiority of our approaches. For future work, we aim to obtain an in-depth understanding of the effect of SAM and MGS and achieve better generalization in DFL.

Ackonwledgements

This work is supported by the Science and Technology Innovation 2030 – "Brain Science and Brain-like Research" key Project (No. 2021ZD0201405), the National Key R&D Program of China (2022YFB4701400/4701402), and the National Natural Science Foundation of China (No. U21B6002, U1813216, 52265002).

References

Abbas, M., Xiao, Q., Chen, L., Chen, P., and Chen, T. Sharpmaml: Sharpness-aware model-agnostic meta learning. In *International Conference on Machine Learning, ICML*, pp. 10–32, 2022.

Acar, D. A. E., Zhao, Y., Navarro, R. M., Mattina, M., Whatmough, P. N., and Saligrama, V. Federated learning based on dynamic regularization. In *9th International Conference on Learning Representations, ICLR*, 2021.

Andriushchenko, M. and Flammarion, N. Towards understanding sharpness-aware minimization. In *International Conference on Machine Learning, ICML*, Proceedings of Machine Learning Research, pp. 639–668. PMLR, 2022.

Beltrán, E. T. M., Pérez, M. Q., Sánchez, P. M. S., Bernal, S. L., Bovet, G., Pérez, M. G., Pérez, G. M., and Celdrán, A. H. Decentralized federated learning: Fundamentals, state-of-the-art, frameworks, trends, and challenges. *arXiv preprint arXiv:2211.08413*, 2022.

Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer, 2010.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Caldarola, D., Caputo, B., and Ciccone, M. Improving generalization in federated learning by seeking flat minima. *CoRR*, abs/2203.11834, 2022.

Chen, C., Zhang, J., Shen, L., Zhao, P., and Luo, Z. Communication efficient primal-dual algorithm for nonconvex nonsmooth distributed optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1594–1602. PMLR, 2021.

Chen, H.-Y. and Chao, W.-L. On bridging generic and personalized federated learning for image classification. *arXiv preprint arXiv:2107.00778*, 2021.

Dai, R., Shen, L., He, F., Tian, X., and Tao, D. Dispfl: Towards communication-efficient personalized federated learning via decentralized sparse training. In *International Conference on Machine Learning, ICML*, Proceedings of Machine Learning Research, pp. 4587–4604. PMLR, 2022.

Deng, Y., Kamani, M. M., and Mahdavi, M. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Du, J., Yan, H., Feng, J., Zhou, J. T., Zhen, L., Goh, R. S. M., and Tan, V. Efficient sharpness-aware minimization for improved training of neural networks. In *International Conference on Learning Representations*, 2021.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Hashemi, A., Acharya, A., Das, R., Vikalo, H., Sanghavi, S., and Dhillon, I. On the benefits of multiple gossip steps in communication-constrained decentralized federated learning. *IEEE Transactions on Parallel and Distributed Systems, TPDS*, pp. 2727–2739, 2022.

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

Hsu, T.-M. H., Qi, H., and Brown, M. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Huang, Y., Chu, L., Zhou, Z., Wang, L., Liu, J., Pei, J., and Zhang, Y. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7865–7873, 2021.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, pp. 1–210, 2021.

Kang, J., Ye, D., Nie, J., Xiao, J., Deng, X., Wang, S., Xiong, Z., Yu, R., and Niyato, D. Blockchain-based federated learning for industrial metaverses: Incentive scheme with optimal aoi. In *2022 IEEE International Conference on Blockchain (Blockchain)*, pp. 71–78. IEEE, 2022.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pp. 5381–5393. PMLR, 2020.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kwon, J., Kim, J., Park, H., and Choi, I. K. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pp. 5905–5914. PMLR, 2021.

Lalitha, A., Shekhar, S., Javidi, T., and Koushanfar, F. Fully decentralized federated learning. In *Third workshop on Bayesian Deep Learning (NeurIPS)*, 2018.

Lalitha, A., Kilinc, O. C., Javidi, T., and Koushanfar, F. Peer-to-peer federated learning on graphs. *arXiv preprint arXiv:1901.11173*, 2019.

Li, B., Cen, S., Chen, Y., and Chi, Y. Communication-efficient distributed optimization in networks with gradient tracking and variance reduction. *Journal of Machine Learning Research, JMLR*, pp. 180:1–180:51, 2020a.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Li, J., Shao, Y., Wei, K., Ding, M., Ma, C., Shi, L., Han, Z., and Poor, H. V. Blockchain assisted decentralized federated learning (blade-fl): Performance analysis and resource allocation. *IEEE Transactions on Parallel and Distributed Systems*, 33(10):2401–2415, 2022. doi: 10.1109/TPDS.2021.3138848.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, pp. 50–60, 2020b.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, pp. 429–450, 2020c.

Li, T., Hu, S., Beirami, A., and Smith, V. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 5330–5340, 2017.

Liu, Y., Mai, S., Chen, X., Hsieh, C.-J., and You, Y. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12360–12370, 2022.

Mcmahan, H. B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. Y. Communication-efficient learning of deep networks from decentralized data. pp. 1273–1282, 2017.

Mi, P., Shen, L., Ren, T., Zhou, Y., Sun, X., Ji, R., and Tao, D. Make sharpness-aware minimization stronger: A sparsified perturbation approach. *arXiv preprint arXiv:2210.05177*, 2022.

Nedic, A. and Ozdaglar, A. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

Nguyen, T., Dakka, M., Diakiw, S., VerMilyea, M., Perugini, M., Hall, J., and Perugini, D. A novel decentralized federated learning approach to train on globally distributed, poor quality, and protected private medical data. *Scientific Reports*, 12(1):8888, 2022.

Qu, Z., Li, X., Duan, R., Liu, Y., Tang, B., and Lu, Z. Generalized federated learning via sharpness aware minimization. In *International Conference on Machine Learning, ICML*, pp. 18250–18280, 2022.

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=LkFG3lB13U5.

Roy, A. G., Siddiqui, S., Pölsterl, S., Navab, N., and Wachinger, C. Braintorrent: A peer-to-peer environment for decentralized federated learning. *arXiv preprint arXiv:1905.06731*, 2019.

Sahu, A. K., Li, T., Sanjabi, M., Zaheer, M., Talwalkar, A., and Smith, V. On the convergence of federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, pp. 3, 2018.

Shi, Y., Liu, Y., Wei, K., Shen, L., Wang, X., and Tao, D. Make landscape flatter in differentially private federated learning. *arXiv preprint arXiv:2303.11242*, 2023.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sun, T., Li, D., and Wang, B. Decentralized federated averaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Sun, Y., Shen, L., Huang, T., Ding, L., and Tao, D. Fedspeed: Larger local interval, less communication round, and higher generalization accuracy. *arXiv preprint arXiv:2302.10429*, 2023.

Wang, H., Marella, S., and Anderson, J. Fedadmm: A federated primal-dual algorithm allowing partial participation. *arXiv preprint arXiv:2203.15104*, 2022a.

Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., Al-Shedivat, M., Andrew, G., Avestimehr, S., Daly, K., Data, D., et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

Wang, L., Xu, Y., Xu, H., Chen, M., and Huang, L. Accelerating decentralized federated learning in heterogeneous edge computing. *IEEE Transactions on Mobile Computing*, 2022b.

Wang, Y., Su, Z., Zhang, N., and Benslimane, A. Learning in the air: Secure federated learning for uav-assisted crowdsensing. *IEEE Transactions on network science and engineering*, 8(2):1055–1069, 2020.

Warnat-Herresthal, S., Schultze, H., Shastry, K. L., Manamohan, S., Mukherjee, S., Garg, V., Sarveswara, R., Händler, K., Pickkers, P., Aziz, N. A., et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature*, pp. 265–270, 2021.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yang, H., Fang, M., and Liu, J. Achieving linear speedup with partial worker participation in non-IID federated learning. In *International Conference on Learning Representations*, 2021.

Yao, Z., Gholami, A., Keutzer, K., and Mahoney, M. W. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE international conference on big data (Big data)*, pp. 581–590. IEEE, 2020.

Ye, H. and Zhang, T. Deepca: Decentralized exact pca with linear convergence rate. *J. Mach. Learn. Res.*, 22(238): 1–27, 2021.

Ye, H., Zhou, Z., Luo, L., and Zhang, T. Decentralized accelerated proximal gradient descent. *Advances in Neural Information Processing Systems*, 33:18308–18317, 2020.

Yu, H., Yang, S., and Zhu, S. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5693–5700, 2019.

Yu, Z., Hu, J., Min, G., Xu, H., and Mills, J. Proactive content caching for internet-of-vehicles based on peer-to-peer federated learning. In *2020 IEEE 26th International Conference on Parallel and Distributed Systems (ICPADS)*, pp. 601–608. IEEE, 2020.

Yuan, Y., Chen, R., Sun, C., Wang, M., Hua, F., Yi, X., Yang, T., and Liu, J. Defed: A principled decentralized and privacy-preserving federated learning algorithm. *arXiv preprint arXiv:2107.07171*, 2021.

Zhang, M., Sapra, K., Fidler, S., Yeung, S., and Alvarez, J. M. Personalized federated learning with first order model optimization. *arXiv preprint arXiv:2012.08565*, 2020.

Zhang, X., Fang, M., Liu, Z., Yang, H., Liu, J., and Zhu, Z. Net-fleet: Achieving linear convergence speedup for fully decentralized federated learning with heterogeneous data. *arXiv preprint arXiv:2208.08490*, 2022.

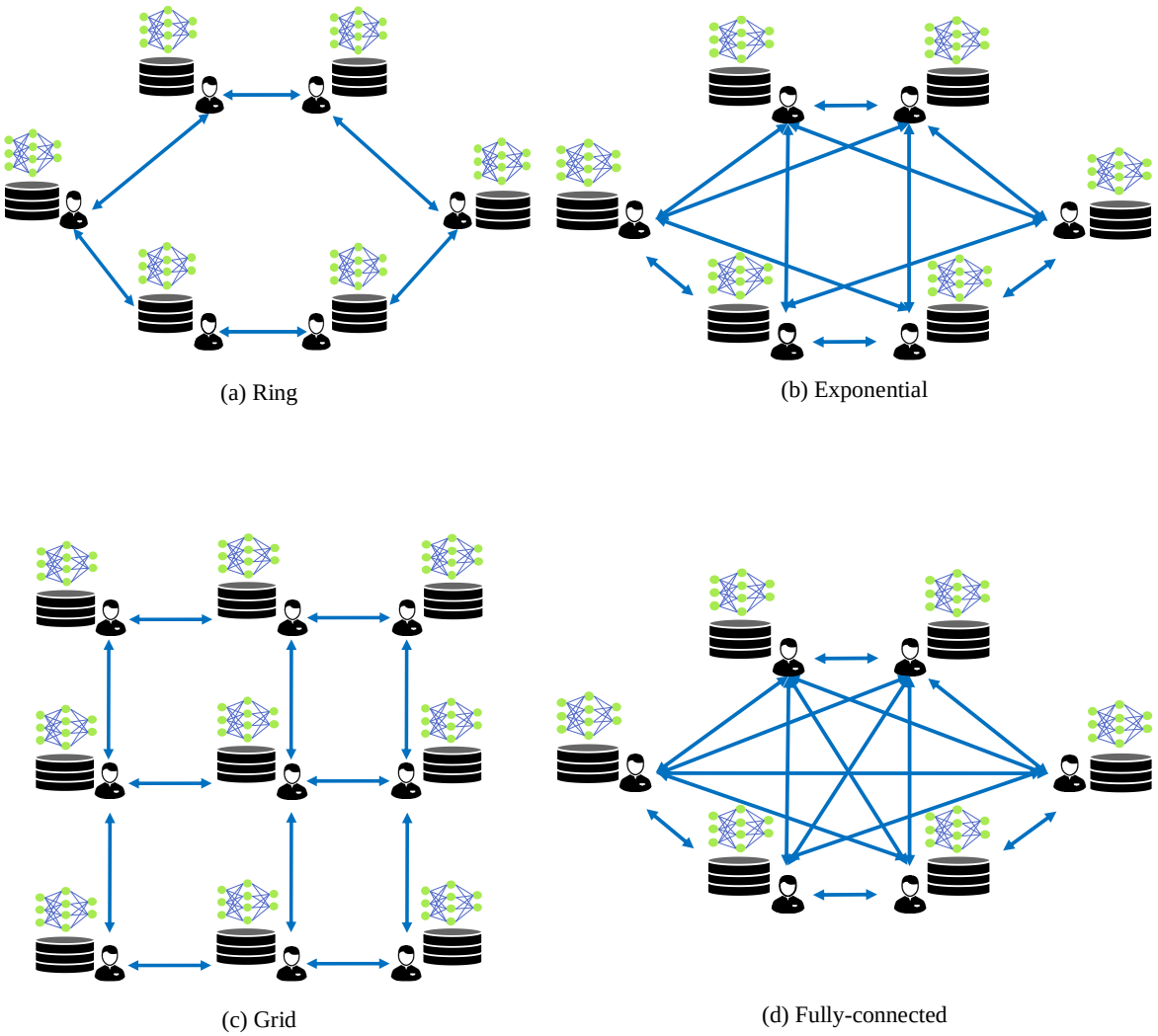
Zhao, Y., Zhang, H., and Hu, X. Penalizing gradient norm for efficiently improving generalization in deep learning. In *International Conference on Machine Learning, ICML*, pp. 26982–26992. PMLR, 2022.

Zhong, Q., Ding, L., Shen, L., Mi, P., Liu, J., Du, B., and Tao, D. Improving sharpness-aware minimization with fisher mask for better generalization on language models. *arXiv preprint arXiv:2210.05497*, 2022.

Zhu, T., He, F., Zhang, L., Niu, Z., Song, M., and Tao, D. Topology-aware generalization of decentralized sgd. In *International Conference on Machine Learning, ICML*, pp. 27479–27503. PMLR, 2022.

Supplementary Material for “ Improving the Model Consistency of Decentralized Federated Learning ”

A. Communication Network Topologies



(a) Ring (b) Exponential

(c) Grid (d) Fully-connected

Figure 7. An overview of the various communication network topologies in decentralized setting.

B. More Details on Algorithm Implementation

B.1. Datasets and backbones.

CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) are labeled subsets of the 80 million images dataset. They both share the same $60,000$ input images. CIFAR-100 has a finer labeling, with 100 unique labels, in comparison to CIFAR-10, having 10 unique labels. The VGG-11 as the backbone is used for CIFAR-10, and the ResNet is chose for CIFAR-100, where the batch-norm layers are replaced by group-norm layers due to a detrimental effect of batch-norm.

B.2. More details about baselines.

FedAvg is the classic FL method via the vanilla weighted averaging to parallel train a global model with a central server. FedSAM applies SAM to be the local optimizer for improving the model generalization performance. For decentralized schemes, D-PSGD is a classic decentralized parallel SGD method to reach a consensus model [2], DFedAvg is the decentralized FedAvg, and DFedAvgM uses SGD with momentum based on DFedAvg to train models on each client and performs multiple local training steps before each communication. Furthermore, DisPFL is a novel personalized FL framework in a decentralized communication protocol, which uses a decentralized sparse training technique, thus for a fair comparison, we report the global accuracy in DisPFL.

B.3. Hyperparameters.

The total client number is set to 100, and the number of connection s in each client is restrict at most 10 neighbors in decentralized setting. For centralized setting, the sample ratio of client is set to 0.1. The local learning rate is set to 0.1 decayed with 0.998 after each communication round for all experiments, and the global learning rate is set to 1.0 for centralized methods. The batch size is fixed to 128 for all the experiments. We run 1000 global communication rounds for CIFAR-10 and CIFAR-100. SGD optimizer is used with weighted decayed parameter 0.0005 for all baselines except FedSAM. Other optimizer hyper-parameters $\rho = 0.01$ for our algorithms (DFedSAM and DFedSAM-MGS with $Q = 1$) via grid search on the set $\{0.01, 0.025, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0\}$ and the value of ρ in FedSAM is followed by (Qu et al., 2022), respectively. And following by (Sun et al., 2022), the local optimization with momentum 0.9 for DFedAvgM. For local iterations K, the training epoch in D-PSGD is set to 1, that for all other methods is set to 5.

B.4. Communication configurations.

Specifically, such as (Dai et al., 2022), the decentralized methods actually generate far more communication volume than centralized methods because each client in the network topology needs to transmit the local information to their neighbors. However, only the partly sampled clients can upload their parameter updates with a central server in centralized setting. Therefore, for a fair comparison, we use a dynamic time-varying connection topology for decentralized methods in **Section** 5.2, we restrict each client can communicate with at most 10 neighbors which are random sampled without replacement from all clients, and only 10 clients who are neighbors to each other can perform one gossip step to exchange their local information in DFedSAM. In DFedSAM-MGS, the gossip step is performed Q times, $10 \times Q$ clients are sampled without replacement can perform one gossip step to exchange their local information.

C. More Details on Experiments under the Pathological Partition

Pathological Partition. To make the experiment more convincing, we also conduct some experiments on CIFAR-10 and CIFAR-100 with VGG-11 and ResNet-18 models, respectively, under the pathological data partition setup (Zhang et al., 2020) after 500 communication rounds. Where the sorted data is divided into 200 partitions with 100 clients and each client is randomly assigned 2 partitions from 2 classes. It is clear that training performance is more difficult in this partition setup compared with Dirichlet distribution partition setup. That means larger heterogeneous levels which is closer to real-world scenario.

Comparison with all baselines. From Table 3, we can obviously see that the final performance is worse than that in Dir(α) due to larger heterogeneous levels especially on more complex dataset such as CIFAR-100. It is seen that in this partition way, the ill-impact of the data heterogeneity is more severe on performance. However, our methods still have benefited from SAM compared with other methods, thereby verifying the effective of our algorithms. Furthermore, it also exposes the all-fit-one model may be not suitable for more serious statistical heterogeneity setting. For this issue, existing many works have adopted some technique to alleviate it, such as model personalization in FL (Zhang et al., 2020; Li et al., 2021; Chen & Chao, 2021; Deng et al., 2020; Huang et al., 2021).

[2] In this work, we focus on decentralized FL which refers to the local training with multiple local iterates, whereas decentralized learning/training focuses on one-step local training. For instance, D-PSGD (Lian et al., 2017) is a decentralized training algorithm, which uses the one-step SGD to train local models in each communication round.

Table 3. The performance on CIFAR-10 and CIFAR-100 under pathological data partition.

Task	Test accuracy (%) for all algorithms sampling only 2 classes from the whole data classes in each client							
	FedAvg	FedSAM	D-PSGD	DisPFL	DFedAvg	DFedAvgM	DFedSAM	DFedSAM-MGS
CIFAR-10	63.23	65.61	39.53	47.61	45.67	51.64	60.58	64.45
CIFAR-100	9.66	13.07	5.60	5.69	5.35	7.65	9.91	12.07

D. Convergence Analysis for DFedSAM and DFedSAM-MGS

In the following, we present the proof of convergence results for DFedSAM and DFedSAM-MGS, respectively. Note that the proof of **Theorem** 4.5 is thoroughly introduced in two sections D.2 and D.3 as follows, where $Q = 1$ and $Q > 1$, respectively.

D.1. Preliminary Lemmas

Lemma D.1 (Lemma 4, (Lian et al., 2017)). *For any $t \in \mathbb{Z}^+$, the mixing matrix $\mathbf{W} \in \mathbb{R}^m$ satisfies $\|\mathbf{W}^t - \mathbf{P}\|_{\text{op}} \leq \lambda^t$, where $\lambda := \max\{|\lambda_2|, |\lambda_m(W)|\}$ and for a matrix $\mathbf{A}$, we denote its spectral norm as $\|\mathbf{A}\|_{\text{op}}$. Furthermore, $\mathbf{1} := [1, 1, \ldots, 1]^\top \in \mathbb{R}^m$ and*

$$\mathbf{P} := \frac{\mathbf{1}\mathbf{1}^\top}{m} \in \mathbb{R}^{m \times m}.$$

In [Proposition 1, (Nedic & Ozdaglar, 2009)], the author also proved that $\|W^t - \mathbf{P}\|_{\text{op}} \leq C\lambda^t$ for some $C > 0$ that depends on the matrix.

Lemma D.2 (Lemma A.5, (Qu et al., 2022)). *(Bounded global variance of $\|\nabla f_i(\mathbf{x} + \delta_i) - \nabla f(\mathbf{x} + \delta)\|^2$.) An immediate implication of Assumptions 4.3 and 4.4, the variance of local and global gradients with perturbation can be bounded as follows:*

$$\|\nabla f_i(\mathbf{x} + \delta_i) - \nabla f(\mathbf{x} + \delta)\|^2 \leq 3\sigma_g^2 + 6L^2\rho^2.$$

Lemma D.3 (Lemma B.1, (Qu et al., 2022)). *(Bounded $\mathcal{E}_\delta$ of DFedSAM). the updates of DFedSAM for any learning rate satisfying $\eta \leq \frac{1}{4KL}$ have the drift due to $\delta_{i,k} - \delta$:*

$$\mathcal{E}_\delta = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[\|\delta_{i,k} - \delta\|^2] \leq 2K^2\beta^2\eta^2\rho^2.$$

where $\delta = \rho\frac{\nabla F(\mathbf{x})}{\|\nabla F(\mathbf{x})\|}$, $\quad \delta_{i,k} = \rho\frac{\nabla F_i(\mathbf{y}^{t,k}, \xi)}{\|\nabla F_i(\mathbf{y}^{t,k}, \xi)\|}$.

Lemma D.4. *Assume that Assumptions 4.3 and 4.4 hold, and $(\mathbf{y}^{t,k}(i) + \delta_{i,k})_{t\geq 0}$, $(\mathbf{x}^{t,k}(i))_{t\geq 0}$ are generated by DFedSAM for all $i \in \{1, 2, ..., m\}$. If the client update of DFedSAM for any learning rate $\eta \leq \frac{1}{10KL}$, it then follows:*

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left\|(\mathbf{y}^{t,k}(i) + \delta_{i,k}) - \mathbf{x}^t(i)\right\|^2 \leq 2K\Big(&\frac{4K^3L^2\eta^2\rho^4}{2K-1} + 8K\eta^2(L^2\rho^2 + \sigma_g^2 + \sigma_l^2) \\ &+ \frac{8K\eta^2}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2\Big) + \frac{2K\rho^2}{2K-1},\end{aligned} \tag{8}$$

where $0 \leq k \leq K-1$.

Proof.

For any local iteration $k \in \{0, 1, ..., K-1\}$ in any node i, it holds

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left\|(\mathbf{y}^{t,k}(i) + \delta_{i,k}) - \mathbf{x}^t(i)\right\|^2 &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\mathbf{y}^{t,k-1}(i) + \delta_{i,k} - \eta\nabla F_i(\mathbf{y}^{t,k-1}(i) + \delta_{i,k-1}) - \mathbf{x}^t(i)\right\|^2 \\ &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|\mathbf{y}^{t,k-1}(i) + \delta_{i,k-1} - \mathbf{x}^t(i) + \delta_{i,k} - \delta_{i,k-1} - \eta\Big(\nabla F_i(\mathbf{y}^{t,k-1}(i) + \delta_{i,k-1}) - \nabla F_i(\mathbf{y}^{t,k-1}) \\ &+ \nabla F_i(\mathbf{y}^{t,k-1}) - \nabla f_i(\mathbf{x}^t) + \nabla f_i(\mathbf{x}^t) - \nabla f(\mathbf{x}^t) + \nabla f(\mathbf{x}^t)\Big)\|^2 \\ &\leq \mathrm{I} + \mathrm{II},\end{aligned}$$

where $\mathrm{I} = (1+\frac{1}{2K-1})\frac{1}{m}\sum_{i=1}^{m}\Big(\mathbb{E}\left\|\mathbf{y}^{t,k-1}(i)+\delta_{i,k-1}-\mathbf{x}^{t}(i)\right\|^2+\mathbb{E}\|\delta_{i,k}-\delta_{i,k-1}\|^2\Big)$ *and*

$$\mathrm{II} = \frac{2K}{m}\sum_{i=1}^{m}\mathbb{E}\|-\eta\Big(\nabla F_i(\mathbf{y}^{t,k-1}(i)+\delta_{i,k-1})-\nabla F_i(\mathbf{y}^{t,k-1}) + \nabla F_i(\mathbf{y}^{t,k-1})-\nabla f_i(\mathbf{x}^t)+\nabla f_i(\mathbf{x}^t)-\nabla f(\mathbf{x}^t)+\nabla f(\mathbf{x}^t)\Big)\|^2,$$

With ***Lemma*** *D.3 and Assumptions, the bounds are*

$$\mathrm{I} = (1+\frac{1}{2K-1})\frac{1}{m}\sum_{i=1}^{m}\Big(\mathbb{E}\left\|\mathbf{y}^{t,k-1}(i)+\delta_{i,k-1}-\mathbf{x}^{t}(i)\right\|^2+2K^2L^2\eta^2\rho^4\Big),$$

and

$$\mathrm{II} = \frac{8K\eta^2}{m}\sum_{i=1}^{m}\Big(L^2\rho^2+\sigma_l^2+\sigma_g^2+\mathbb{E}\left\|\nabla f(\mathbf{x}^t)\right\|^2\Big),$$

where $\mathbb{E}\left\|\delta_{i,k-1}\right\|^2\le\rho^2$.
Thus, we can obtain

$$\mathbb{E}\left\|(\mathbf{y}^{t,k}(i)+\delta_{i,k})-\mathbf{x}^t(i)\right\|^2 \le (1+\frac{1}{2K-1})\mathbb{E}\left\|(\mathbf{y}^{t,k-1}(i)+\delta_{i,k-1})-\mathbf{x}^t(i)\right\|^2 + \frac{4K^3L^2\eta^2\rho^4}{2K-1}+8K\eta^2(L^2\rho^2+\sigma_g^2+\sigma_l^2)+\frac{8K\eta^2}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2,$$

where $\mathbb{E}\left\|\nabla f(\mathbf{x}^t)\right\|^2=\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2$, $f(\mathbf{x}):=\frac{1}{m}\sum_{i=1}^{m}f_i(\mathbf{x})$, *and* $\nabla f_i(\mathbf{x}^t):=\nabla f(\mathbf{x}^t(i))$.
The recursion from $\tau=0$ *to* k *yields*

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left\|(\mathbf{y}^{t,k}(i)+\delta_{i,k})-\mathbf{x}^t(i)\right\|^2 &\le \frac{1}{m}\sum_{i=1}^{m}\sum_{\tau=1}^{K-1}(1+\frac{1}{2K-1})^{\tau}\Big(\frac{4K^3L^2\eta^2\rho^4}{2K-1}+8K\eta^2(L^2\rho^2+\sigma_g^2+\sigma_l^2)\\ &\quad+\frac{8K\eta^2}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2\Big)+(1+\frac{1}{2K-1})\rho^2\\ &\le 2K(\frac{4K^3L^2\eta^2\rho^4}{2K-1}+8K\eta^2(L^2\rho^2+\sigma_g^2+\sigma_l^2)\\ &\quad+\frac{8K\eta^2}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2)+\frac{2K\rho^2}{2K-1}.\end{aligned}$$

This completes the proof.

Lemma D.5. *Assume that the number of local iterations K is large enough. Let* $\{\mathbf{x}^t(i)\}_{t\ge0}$ *be generated by DFedSAM for all* $i\in\{1,2,...,m\}$ *and any learning rate* $\eta>0$*, we have following bound:*

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\|\mathbf{x}^{t,k}(i)-\overline{\mathbf{x}^t}\|^2]\le\frac{C_2\eta^2}{(1-\lambda)^2},$$

where $C_2=2K(\frac{4K^3L^2\rho^4}{2K-1}+8K(L^2\rho^2+\sigma_g^2+\sigma_l^2)+\frac{8K}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2)+\frac{2K\rho^2}{\eta^2(2K-1)}$.
Proof.
Following [Lemma 4, (Sun et al., 2022)], we denote $\mathbf{Z}^t:=\left[\mathbf{z}^t(1),\mathbf{z}^t(2),\ldots,\mathbf{z}^t(m)\right]^\top\in\mathbb{R}^{m\times d}$. *With these notation, we have*

$$\mathbf{X}^{t+1}=\mathbf{W}\mathbf{Z}^t=\mathbf{W}\mathbf{X}^t-\zeta^t, \tag{9}$$

where $\zeta^t := \mathbf{W}\mathbf{X}^t - \mathbf{W}\mathbf{Z}^t$. *The iteration equation (9) can be rewritten as the following expression*

$$\mathbf{X}^t = \mathbf{W}^t\mathbf{X}^0 - \sum_{j=0}^{t-1}\mathbf{W}^{t-1-j}\zeta^j. \tag{10}$$

Obviously, it follows

$$\mathbf{W}\mathbf{P} = \mathbf{P}\mathbf{W} = \mathbf{P}. \tag{11}$$

According to Lemma D.1, it holds

$$\|\mathbf{W}^t - \mathbf{P}\| \leq \lambda^t.$$

Multiplying both sides of equation (10) with $\mathbf{P}$ *and using equation (11), we then get*

$$\mathbf{P}\mathbf{X}^t = \mathbf{P}\mathbf{X}^0 - \sum_{j=0}^{t-1}\mathbf{P}\zeta^j = -\sum_{j=0}^{t-1}\mathbf{P}\zeta^j, \tag{12}$$

where we used initialization $\mathbf{X}^0 = \boldsymbol{0}$. *Then, we are led to*

$$\begin{aligned}\|\mathbf{X}^t - \mathbf{P}\mathbf{X}^t\| &= \|\sum_{j=0}^{t-1}(\mathbf{P} - \mathbf{W}^{t-1-j})\zeta^j\| \\ &\leq \sum_{j=0}^{t-1}\|\mathbf{P} - \mathbf{W}^{t-1-j}\|_{op}\|\zeta^j\| \leq \sum_{j=0}^{t-1}\lambda^{t-1-j}\|\zeta^j\|.\end{aligned} \tag{13}$$

With Cauchy inequality,

$$\begin{aligned}\mathbb{E}\|\mathbf{X}^t - \mathbf{P}\mathbf{X}^t\|^2 &\leq \mathbb{E}(\sum_{j=0}^{t-1}\lambda^{\frac{t-1-j}{2}}\cdot\lambda^{\frac{t-1-j}{2}}\|\zeta^j\|)^2 \\ &\leq (\sum_{j=0}^{t-1}\lambda^{t-1-j})(\sum_{j=0}^{t-1}\lambda^{t-1-j}\mathbb{E}\|\zeta^j\|^2)\end{aligned}$$

Direct calculation gives us

$$\mathbb{E}\|\zeta^j\|^2 \leq \|\mathbf{W}\|^2\cdot\mathbb{E}\|\mathbf{X}^j - \mathbf{Z}^j\|^2 \leq \mathbb{E}\|\mathbf{X}^j - \mathbf{Z}^j\|^2.$$

With ***Lemma*** *D.4 and Assumption 3, for any* j,

$$\begin{aligned}&\mathbb{E}\|\mathbf{X}^j - \mathbf{Z}^j\|^2 \\ &\leq m\Big(2K(\frac{4K^3L^2\rho^4}{2K-1} + 8K(L^2\rho^2 + \sigma_g^2 + \sigma_l^2) + \frac{8K}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2) + \frac{2K\rho^2}{\eta^2(2K-1)}\Big)\eta^2.\end{aligned}$$

Thus, we get

$$\begin{aligned}&\mathbb{E}\|\mathbf{X}^t - \mathbf{P}\mathbf{X}^t\|^2 \\ &\leq \frac{m\Big(2K(\frac{4K^3L^2\rho^4}{2K-1} + 8K(L^2\rho^2 + \sigma_g^2 + \sigma_l^2) + \frac{8K}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2) + \frac{2K\rho^2}{\eta^2(2K-1)}\Big)\eta^2}{(1-\lambda)^2}.\end{aligned}$$

The fact that $\mathbf{X}^t - \mathbf{P}\mathbf{X}^t = \begin{pmatrix}\mathbf{x}^t(1) - \overline{\mathbf{x}^t} \\ \mathbf{x}^t(2) - \overline{\mathbf{x}^t} \\ \vdots \\ \mathbf{x}^t(m) - \overline{\mathbf{x}^t}\end{pmatrix}$ *then proves the result.*

Lemma D.6. *Assume that the number of local iteration K is large enough. Let $\{\mathbf{x}^t(i)\}_{t\geq 0}$ be generated by DFedSAM-MGS for all $i \in \{1, 2, ..., m\}$ and any learning rate $\eta > 0$, we have following bound:*

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\|\mathbf{x}^{t,k}(i) - \overline{\mathbf{x}^t}\|^2] \leq C_2\eta^2\Big(\frac{\lambda^Q+1}{(1-\lambda)^2 m^{2(Q-1)}} + \frac{\lambda^Q+1}{(1-\lambda^Q)^2}\Big),$$

where $C_2 = 2K(\frac{4K^3L^2\rho^4}{2K-1} + 8K(L^2\rho^2+\sigma_g^2+\sigma_l^2) + \frac{8K}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2) + \frac{2K\rho^2}{\eta^2(2K-1)}$.

Proof.

Following [Lemma 4, (Sun et al., 2022)] and ***Lemma*** *D.5, we denote* $\mathbf{Z}^t := \left[\mathbf{z}^t(1), \mathbf{z}^t(2), \ldots, \mathbf{z}^t(m)\right]^\top \in \mathbb{R}^{m\times d}$. *With these notation, we have*

$$\mathbf{X}^{t+1} = \mathbf{W}^Q\mathbf{Z}^t = \mathbf{W}^Q\mathbf{X}^t - \zeta^t, \tag{14}$$

where $\zeta^t := \mathbf{W}^Q\mathbf{X}^t - \mathbf{W}^Q\mathbf{Z}^t$. *The iteration equation (14) can be rewritten as the following expression*

$$\mathbf{X}^t = (\mathbf{W}^t)^Q\mathbf{X}^0 - \sum_{j=0}^{t-1}\mathbf{W}^{(t-1-j)Q}\zeta^j. \tag{15}$$

Obviously, it follows

$$\mathbf{W}^Q\mathbf{P} = \mathbf{P}\mathbf{W}^Q = \mathbf{P}. \tag{16}$$

According to Lemma D.1, it holds

$$\|\mathbf{W}^t - \mathbf{P}\| \leq \lambda^t.$$

Multiplying both sides of equation (15) with $\mathbf{P}$ *and using equation (16), we then get*

$$\mathbf{P}\mathbf{X}^t = \mathbf{P}\mathbf{X}^0 - \sum_{j=0}^{t-1}\mathbf{P}\zeta^j = -\sum_{j=0}^{t-1}\mathbf{P}\zeta^j,$$

where we used initialization $\mathbf{X}^0 = \boldsymbol{0}$. *Then, we are led to*

$$\begin{aligned}
\|\mathbf{X}^t - \mathbf{P}\mathbf{X}^t\| &= \|\sum_{j=0}^{t-1}(\mathbf{P} - \mathbf{W}^{Q(t-1-j)})\zeta^j\| \\
&\leq \sum_{j=0}^{t-1}\|\mathbf{P} - \mathbf{W}^{Q(t-1-j)}\|_{op}\|\zeta^j\| \\
&\leq \sum_{j=0}^{t-1}\lambda^{t-1-j}\|\mathbf{W}^{(t-1-j)(Q-1)}\|\|\zeta^j\| \\
&\leq \sum_{j=0}^{t-1}\lambda^{t-1-j}\|\mathbf{W}^{t-1-j} - \mathbf{P} + \mathbf{P}\|^{Q-1}\|\zeta^j\|.
\end{aligned}$$

With Cauchy inequality,

$$\begin{aligned}
\mathbb{E}\|\mathbf{X}^t - \mathbf{P}\mathbf{X}^t\|^2 &\leq (\sum_{j=0}^{t-1}\lambda^{t-1-j}(\lambda^{(Q-1)(t-1-j)} + \frac{1}{m^{Q-1}})\sum_{j=0}^{t-1}\lambda^{t-1-j}(\lambda^{(Q-1)(t-1-j)} + \frac{1}{m^{Q-1}})\mathbb{E}\|\zeta^j\|^2) \\
&\leq (\sum_{j=0}^{t-1}(\lambda^{Q(t-1-j)} + \frac{\lambda^{t-1-j}}{m^{Q-1}})\sum_{j=0}^{t-1}(\lambda^{Q(t-1-j)} + \frac{\lambda^{t-1-j}}{m^{Q-1}})\mathbb{E}\|\zeta^j\|^2) \\
&\leq \mathbb{E}\|\zeta^j\|^2\Big(\frac{1}{(1-\lambda)^2 m^{2(Q-1)}} + \frac{1}{(1-\lambda^Q)^2}\Big).
\end{aligned}$$

Direct calculation gives us

$$\begin{aligned}\mathbb{E}\|\zeta^j\|^2 &\leq \|\mathbf{W}^Q\|^2 \cdot \mathbb{E}\|\mathbf{X}^j - \mathbf{Z}^j\|^2 \\ &\leq \|\mathbf{W}-\mathbf{P}+\mathbf{P}\|^{2Q}\|\mathbf{X}^j - \mathbf{Z}^j\|^2 \\ &\leq (\|\mathbf{W}-\mathbf{P}\|^{2Q} + \|\mathbf{P}\|^{2Q})\mathbb{E}\|\mathbf{X}^j - \mathbf{Z}^j\|^2 \\ &\leq (\lambda^Q+1)\mathbb{E}\|\mathbf{X}^j - \mathbf{Z}^j\|^2.\end{aligned}$$

*With **Lemma** D.4 and Assumption 3, for any j,*

$$\begin{aligned}&\mathbb{E}\|\mathbf{X}^j - \mathbf{Z}^j\|^2 \\ &\leq m\Big(2K(\frac{4K^3L^2\rho^4}{2K-1} + 8K(L^2\rho^2+\sigma_g^2+\sigma_l^2) + \frac{8K}{m}\sum_{i=1}^m \mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2) + \frac{2K\rho^2}{\eta^2(2K-1)}\Big)\eta^2.\end{aligned}$$

Thus, we get

$$\mathbb{E}\|\mathbf{X}^t - \mathbf{P}\mathbf{X}^t\|^2 \leq mC_2\eta^2\Big(\frac{\lambda^Q+1}{(1-\lambda)^2m^{2(Q-1)}} + \frac{\lambda^Q+1}{(1-\lambda^Q)^2}\Big),$$

where $C_2 = 2K(\frac{4K^3L^2\rho^4}{2K-1} + 8K(L^2\rho^2+\sigma_g^2+\sigma_l^2) + \frac{8K}{m}\sum_{i=1}^m \mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2) + \frac{2K\rho^2}{\eta^2(2K-1)}$.

The fact that $\mathbf{X}^t - \mathbf{P}\mathbf{X}^t = \begin{pmatrix} \mathbf{x}^t(1) - \overline{\mathbf{x}^t} \\ \mathbf{x}^t(2) - \overline{\mathbf{x}^t} \\ \vdots \\ \mathbf{x}^t(m) - \overline{\mathbf{x}^t} \end{pmatrix}$ *then proves the result.*

D.2. Proof of convergence results for DFedSAM.

Noting that $\mathbf{P}\mathbf{X}^{t+1} = \mathbf{P}\mathbf{W}\mathbf{Z}^t = \mathbf{P}\mathbf{Z}^t$, that is also

$$\overline{\mathbf{x}^{t+1}} = \overline{\mathbf{z}^t},$$

where $\mathbf{X} := \big[\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(m)\big]^\top \in \mathbb{R}^{m\times d}$ and $\mathbf{Z} := \big[\mathbf{z}(1), \mathbf{z}(2), \ldots, \mathbf{z}(m)\big]^\top \in \mathbb{R}^{m\times d}$.
Thus we have

$$\overline{\mathbf{x}^{t+1}} - \overline{\mathbf{x}^t} = \overline{\mathbf{x}^{t+1}} - \overline{\mathbf{z}^t} + \overline{\mathbf{z}^t} - \overline{\mathbf{x}^t} = \overline{\mathbf{z}^t} - \overline{\mathbf{x}^t}, \tag{17}$$

where $\overline{\mathbf{z}^t} := \frac{\sum_{i=1}^m \mathbf{z}^t(i)}{m}$ and $\overline{\mathbf{x}^t} := \frac{\sum_{i=1}^m \mathbf{x}^t(i)}{m}$. In each node,

$$\begin{aligned}\overline{\mathbf{z}^t} - \overline{\mathbf{x}^t} &= \frac{\sum_{i=1}^m(\sum_{k=0}^{K-1}\mathbf{y}^{t,k+1}(i) - \mathbf{y}^{t,k}(i))}{m} \\ &= \frac{\sum_{i=1}^m\sum_{k=0}^{K-1}(-\eta\tilde{\mathbf{g}}^{t,k}(i))}{m} \\ &= \frac{\sum_{i=1}^m\sum_{k=0}^{K-1}(-\eta\nabla F_i(\mathbf{y}^{t,k} + \rho\nabla F_i(\mathbf{y}^{t,k};\xi)/\left\|\nabla F_i(\mathbf{y}^{t,k};\xi)\right\|_2);\xi)}{m}.\end{aligned} \tag{18}$$

The Lipschitz continuity of ∇f:

$$\mathbb{E}f(\overline{\mathbf{x}^{t+1}}) \leq \mathbb{E}f(\overline{\mathbf{x}^t}) + \mathbb{E}\langle\nabla f(\overline{\mathbf{x}^t}), \overline{\mathbf{z}^t} - \overline{\mathbf{x}^t}\rangle + \frac{L}{2}\mathbb{E}\|\overline{\mathbf{x}^{t+1}} - \overline{\mathbf{x}^t}\|^2, \tag{19}$$

where we used (17).
And (18) is used:

$$
\begin{aligned}
&\mathbb{E}\langle K\nabla f(\overline{\mathbf{x}^t}), (\overline{\mathbf{z}^t}-\overline{\mathbf{x}^t})/K\rangle = \mathbb{E}\langle K\nabla f(\overline{\mathbf{x}^t}), -\eta\nabla f(\overline{\mathbf{x}^t})+\eta\nabla f(\overline{\mathbf{x}^t})+(\overline{\mathbf{z}^t}-\overline{\mathbf{x}^t})/K\rangle\\
&=-\eta K\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2+\mathbb{E}\langle K\nabla f(\overline{\mathbf{x}^t}), \eta\nabla f(\overline{\mathbf{x}^t})+(\overline{\mathbf{z}^t}-\overline{\mathbf{x}^t})/K\rangle\\
&\overset{a)}{=}-\eta K\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2+\mathbb{E}\langle K\nabla f(\overline{\mathbf{x}^t}), \frac{\eta}{mK}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\Big(\nabla F_i(\frac{1}{m}\sum_{i=1}^{m}\mathbf{x}^t(i))-\nabla F_i(\mathbf{y}^{t,k}+\delta_{i,k};\xi)\Big)\rangle\\
&\overset{b)}{\leq}-\eta K\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2+\eta\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|\cdot\mathbb{E}\left\|\frac{L}{m^2}\sum_{i=1}^{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}(\mathbf{x}^t(i)-\mathbf{y}^{t,k}-\delta_{i,k})\right\|\\
&\overset{c)}{\leq}\frac{-\eta K}{2}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2+\frac{\eta L^2K^2}{2K}\Big(2K(\frac{4K^3L^2\eta^2\rho^4}{2K-1}+8K\eta^2(L^2\rho^2+\sigma_g^2+\sigma_l^2)\\
&+\frac{8K\eta^2}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2)+\frac{2K\rho^2}{2K-1}\Big),
\end{aligned}
\tag{20}
$$

where $a)$ uses $\nabla f_i=\mathbb{E}\nabla F_i$ and $\nabla f=\frac{1}{m}\sum_{i=1}^{m}\nabla f_i$, $b)$ uses the Lipschitz continuity, and $c)$ uses **Lemma** D.4. Meanwhile, we can get

$$
\begin{aligned}
\frac{L}{2}\mathbb{E}\left\|\overline{\mathbf{x}^{t+1}}-\overline{\mathbf{x}^t}\right\|^2=\frac{L}{2}\mathbb{E}\left\|\overline{\mathbf{z}^t}-\overline{\mathbf{x}^t}\right\|^2&\leq\frac{L}{2}\frac{1}{m}\sum_{i=1}^{m}\left\|\mathbf{y}^{t,K}(i)-\mathbf{x}^t(i)\right\|^2\\
&\leq\frac{L}{2}\mathbb{E}\left\|\frac{-\eta\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(\mathbf{y}^{t,k}+\delta_{i,k};\xi)}{m}\right\|^2\\
&\leq\frac{\eta^2L}{2m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla F_i(\mathbf{y}^{t,k}+\delta_{i,k};\xi)-\nabla F_i(\mathbf{y}^{t,k};\xi)+\nabla F_i(\mathbf{y}^{t,k};\xi)\\
&\quad-\nabla f_i(\mathbf{x}^t)+\nabla f_i(\mathbf{x}^t)\|^2\\
&\overset{a)}{\leq}\frac{3\eta^2KL}{2}\left(L^2\rho^2+\sigma_l^2+\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|\nabla f(\mathbf{x}^t(i)\|^2\right),
\end{aligned}
\tag{21}
$$

where $a)$ uses **Assumptions** 4.3 and 4.4.
Thus, (19) can be represented as

$$
\begin{aligned}
\mathbb{E}f(\overline{\mathbf{x}^{t+1}})\leq&\mathbb{E}f(\overline{\mathbf{x}^t})-\frac{\eta K}{2}\mathbb{E}\left\|\nabla\mathbb{E}f(\overline{\mathbf{x}^t})\right\|^2+\frac{\eta L^2K}{2}C_1\\
&+\frac{8\eta^3K^2L^2}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2+\frac{3\eta^2KL}{2}\left(L^2\rho^2+\sigma_l^2+\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|\nabla f(\mathbf{x}^t(i)\|^2\right),
\end{aligned}
\tag{22}
$$

where $C_1=2K(\frac{4K^3L^2\eta^2\rho^4}{2K-1}+8K\eta^2(L^2\rho^2+\sigma_g^2+\sigma_l^2))+\frac{2K\rho^2}{2K-1}$.
Furthermore, with **Lemma** D.5, we can get

$$
\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2&\leq 2L^2\frac{\sum_{i=1}^{m}\left\|\mathbf{x}^t(i)-\overline{\mathbf{x}^t}\right\|^2}{m}+2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2\\
&\overset{a)}{\leq}2L^2\frac{C_2\eta^2}{(1-\lambda)^2}+2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2,
\end{aligned}
\tag{23}
$$

where a) uses **Lemma** D.5 and $C_2=2K(\frac{4K^3L^2\rho^4}{2K-1}+8K(L^2\rho^2+\sigma_g^2+\sigma_l^2)+\frac{8K}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2)+\frac{2K\rho^2}{\eta^2(2K-1)}$.
Therefore, we have

$$
\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2\leq\frac{2L^2C_3\eta^2+2(1-\lambda)^2)\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2}{(1-\lambda)^2-32L^2\eta^2K^2},
\tag{24}
$$

where $C_3 = 2K(\frac{4K^3L^2\rho^4}{2K-1} + 8K(L^2\rho^2+\sigma_g^2+\sigma_l^2)) + \frac{2K\rho^2}{\eta^2(2K-1)}$.
And then, (19) can be represented as

$$\begin{aligned}\mathbb{E}f(\overline{\mathbf{x}^{t+1}}) &\leq \mathbb{E}f(\overline{\mathbf{x}^t}) - \frac{\eta K}{2}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 + \frac{\eta L^2 K C_1}{2} + 8\eta^3K^2L^2(\frac{2L^2C_3\eta^2 + 2(1-\lambda)^2)\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2}{(1-\lambda)^2 - 32L^2\eta^2K^2}) \\ &+ \frac{3\eta^2KL}{2}\left(L^2\rho^2 + \sigma_l^2 + \frac{1}{m}\sum_{i=1}^m\mathbb{E}\|\nabla f(\mathbf{x}^t(i)\|^2\right) \\ &\overset{a)}{\leq} \mathbb{E}f(\overline{\mathbf{x}^t}) + (16\eta^3K^2L^2 - \frac{\eta K}{2} + 3\eta^2KL)\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 + \frac{3\eta^2KL}{2}\left(L^2\rho^2+\sigma_l^2\right) \\ &+ \frac{\eta KL^2C_1}{2} + \frac{16\eta^5K^2L^4C_3 + 3\eta^4KL^3C_3}{(1-\lambda)^2}.\end{aligned} \tag{25}$$

Where a) uses (24) and η is a very small value, summing the inequality (25) from $t=1$ to T, and then we can get the proved result as below:

$$\min_{1\leq t\leq T}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 \leq \frac{2f(\overline{\mathbf{x}^1}) - 2f^*}{T(\eta K - 32\eta^3K^2L^2 - 6\eta^2KL)} + \frac{\frac{\eta L^2KC_1}{2} + \frac{\eta^4KL^3C_3(16\eta KL+3)}{(1-\lambda)^2} + \frac{3\eta^2KL}{2}\left(L^2\rho^2+\sigma_l^2\right)}{\eta K - 32\eta^3K^2L^2 - 6\eta^2KL}.$$

If we choose the learning rate $\eta = \mathcal{O}(1/L\sqrt{KT})$ and $\eta \leq \frac{1}{10KL}$, the number of communication round T is large enough, we have

$$\min_{1\leq t\leq T}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 = \mathcal{O}\Big(\frac{(f(\overline{\mathbf{x}^1}) - f^*) + L^2\rho^2 + \sigma_l^2}{\sqrt{KT}} + \frac{K(L^2\rho^2+\sigma_g^2+\sigma_l^2)}{T} + \frac{K^{3/2}L^2\rho^4}{T^{3/2}(1-\lambda)^2} + \frac{L^2\rho^2+\sigma_g^2+\sigma_l^2}{K^{1/2}T^{3/2}(1-\lambda)^2}\Big).$$

When perturbation amplitude ρ proportional to the learning rate, e.g., $\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$, and then we have:

$$\min_{1\leq t\leq T}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 = \mathcal{O}\Big(\frac{(f(\overline{\mathbf{x}^1}) - f^*) + \sigma_l^2}{\sqrt{KT}} + \frac{K(\sigma_g^2+\sigma_l^2)}{T} + \frac{L^2}{K^{1/2}T^{3/2}} + \frac{\sigma_g^2+\sigma_l^2}{K^{1/2}T^{3/2}(1-\lambda)^2}\Big).$$

Under **Definition** 4.2, we can get

$$\min_{1\leq t\leq T}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 = \mathcal{O}\Big(\frac{(f(\overline{\mathbf{x}^1}) - f^*) + \sigma_l^2}{\sqrt{KT}} + \frac{K(\beta^2+\sigma_l^2)}{T} + \frac{L^2}{K^{1/2}T^{3/2}} + \frac{\beta^2+\sigma_l^2}{K^{1/2}T^{3/2}(1-\lambda)^2}\Big).$$

This completes the proof.

D.3. Proof of convergence results for DFedSAM-MGS

With multiple gossiping steps, $\mathbf{x}^0$ and $\mathbf{z}^0$ are represented as $\mathbf{x}$ and $\mathbf{z}$, respectively. Meanwhile, $\mathbf{Z}^{t,Q} = \mathbf{Z}^{t,0}\cdot\mathbf{W}^Q = \mathbf{Z}^t\cdot\mathbf{W}^Q$. Noting that $\mathbf{P}\mathbf{X}^{t+1} = \mathbf{P}\mathbf{W}^Q\mathbf{Z}^t = \mathbf{P}\mathbf{Z}^t (Q>1)$, that is also

$$\overline{\mathbf{x}^{t+1}} = \overline{\mathbf{z}^t},$$

where $\mathbf{X} := \left[\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(m)\right]^\top \in \mathbb{R}^{m\times d}$ and $\mathbf{Z} := \left[\mathbf{z}(1), \mathbf{z}(2), \ldots, \mathbf{z}(m)\right]^\top \in \mathbb{R}^{m\times d}$.
Thus we have

$$\overline{\mathbf{x}^{t+1}} - \overline{\mathbf{x}^t} = \overline{\mathbf{x}^{t+1}} - \overline{\mathbf{z}^t} + \overline{\mathbf{z}^t} - \overline{\mathbf{x}^t} = \overline{\mathbf{z}^t} - \overline{\mathbf{x}^t}, \tag{26}$$

where $\overline{\mathbf{z}^t} := \frac{\sum_{i=1}^m \mathbf{z}^t(i)}{m}$ and $\overline{\mathbf{x}^t} := \frac{\sum_{i=1}^m \mathbf{x}^t(i)}{m}$. In each node,

$$\begin{aligned}\overline{\mathbf{z}^t} - \overline{\mathbf{x}^t} &= \frac{\sum_{i=1}^m(\sum_{k=0}^{K-1}\mathbf{y}^{t,k+1}(i) - \mathbf{y}^{t,k}(i))}{m} \\ &= \frac{\sum_{i=1}^m\sum_{k=0}^{K-1}(-\eta\tilde{\mathbf{g}}^{t,k}(i))}{m} \\ &= \frac{\sum_{i=1}^m\sum_{k=0}^{K-1}(-\eta\nabla F_i(\mathbf{y}^{t,k} + \rho\nabla F_i(\mathbf{y}^{t,k};\xi)/\left\|\nabla F_i(\mathbf{y}^{t,k};\xi)\right\|_2);\xi)}{m}.\end{aligned} \tag{27}$$

The Lipschitz continuity of ∇f:

$$\mathbb{E}f(\overline{\mathbf{x}^{t+1}}) \leq \mathbb{E}f(\overline{\mathbf{x}^t}) + \mathbb{E}\langle \nabla f(\overline{\mathbf{x}^t}), \overline{\mathbf{z}^t} - \overline{\mathbf{x}^t}\rangle + \frac{L}{2}\mathbb{E}\|\overline{\mathbf{x}^{t+1}} - \overline{\mathbf{x}^t}\|^2, \tag{28}$$

where we used (26).
And (27) is used:

$$\begin{aligned}
&\mathbb{E}\langle K\nabla f(\overline{\mathbf{x}^t}), (\overline{\mathbf{z}^t} - \overline{\mathbf{x}^t})/K\rangle = \mathbb{E}\langle K\nabla f(\overline{\mathbf{x}^t}), -\eta\nabla f(\overline{\mathbf{x}^t}) + \eta\nabla f(\overline{\mathbf{x}^t}) + (\overline{\mathbf{z}^t} - \overline{\mathbf{x}^t})/K\rangle \\
&= -\eta K\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 + \mathbb{E}\langle K\nabla f(\overline{\mathbf{x}^t}), \eta\nabla f(\overline{\mathbf{x}^t}) + (\overline{\mathbf{z}^t} - \overline{\mathbf{x}^t})/K\rangle \\
&\overset{a)}{=} -\eta K\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 + \mathbb{E}\langle K\nabla f(\overline{\mathbf{x}^t}), \frac{\eta}{mK}\sum_{i=1}^m\sum_{k=0}^{K-1}\Big(\nabla F_i(\frac{1}{m}\sum_{i=1}^m \mathbf{x}^t(i)) - \nabla F_i(\mathbf{y}^{t,k} + \delta_{i,k}; \xi)\Big)\rangle \\
&\overset{b)}{\leq} -\eta K\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 + \eta\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\| \cdot \mathbb{E}\left\|\frac{L}{m^2}\sum_{i=1}^m\sum_{i=1}^m\sum_{k=0}^{K-1}(\mathbf{x}^t(i) - \mathbf{y}^{t,k} - \delta_{i,k})\right\| \\
&\overset{c)}{\leq} \frac{-\eta K}{2}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 + \frac{\eta L^2K^2}{2K}\Big(2K(\frac{4K^3L^2\eta^2\rho^4}{2K-1} + 8K\eta^2(L^2\rho^2 + \sigma_g^2 + \sigma_l^2) \\
&+ \frac{8K\eta^2}{m}\sum_{i=1}^m\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2) + \frac{2K\rho^2}{2K-1}\Big),
\end{aligned} \tag{29}$$

where $a)$ uses $\nabla f_i = \mathbb{E}\nabla F_i$ and $\nabla f = \frac{1}{m}\sum_{i=1}^m \nabla f_i$, $b)$ uses the Lipschitz continuity, and $c)$ uses **Lemma** D.4. Meanwhile, we can get

$$\begin{aligned}
\frac{L}{2}\mathbb{E}\left\|\overline{\mathbf{x}^{t+1}} - \overline{\mathbf{x}^t}\right\|^2 = \frac{L}{2}\mathbb{E}\left\|\overline{\mathbf{z}^t} - \overline{\mathbf{x}^t}\right\|^2 &\leq \frac{L}{2}\frac{1}{m}\sum_{i=1}^m\left\|\mathbf{y}^{t,K}(i) - \mathbf{x}^t(i)\right\|^2 \\
&\leq \frac{L}{2}\mathbb{E}\left\|\frac{-\eta\sum_{i=1}^m\sum_{k=0}^{K-1}\nabla F_i(\mathbf{y}^{t,k} + \delta_{i,k}; \xi)}{m}\right\|^2 \\
&\leq \frac{\eta^2 L}{2m}\sum_{i=1}^m\sum_{k=0}^{K-1}\mathbb{E}\|\nabla F_i(\mathbf{y}^{t,k} + \delta_{i,k}; \xi) - \nabla F_i(\mathbf{y}^{t,k}; \xi) + \nabla F_i(\mathbf{y}^{t,k}; \xi) \\
&\quad - \nabla f_i(\mathbf{x}^t) + \nabla f_i(\mathbf{x}^t)\|^2 \\
&\overset{a)}{\leq} \frac{3\eta^2 KL}{2}\left(L^2\rho^2 + \sigma_l^2 + \frac{1}{m}\sum_{i=1}^m\mathbb{E}\|\nabla f(\mathbf{x}^t(i)\|^2\right),
\end{aligned} \tag{30}$$

where $a)$ uses **Assumptions** 4.3 and 4.4.
Thus, (28) can be represented as

$$\begin{aligned}
\mathbb{E}f(\overline{\mathbf{x}^{t+1}}) \leq &\mathbb{E}f(\overline{\mathbf{x}^t}) - \frac{\eta K}{2}\mathbb{E}\left\|\nabla\mathbb{E}f(\overline{\mathbf{x}^t})\right\|^2 + \frac{\eta L^2K}{2}C_1 \\
&+ \frac{8\eta^3K^2L^2}{m}\sum_{i=1}^m\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2 + \frac{3\eta^2KL}{2}\left(L^2\rho^2 + \sigma_l^2 + \frac{1}{m}\sum_{i=1}^m\mathbb{E}\|\nabla f(\mathbf{x}^t(i)\|^2\right),
\end{aligned} \tag{31}$$

where $C_1 = 2K(\frac{4K^3L^2\eta^2\rho^4}{2K-1} + 8K\eta^2(L^2\rho^2 + \sigma_g^2 + \sigma_l^2) + \frac{2K\rho^2}{2K-1}$.
Furthermore, with **Lemma** D.6, we can get

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^m\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2 &\leq 2L^2\frac{\sum_{i=1}^m\left\|\mathbf{x}^t(i) - \overline{\mathbf{x}^t}\right\|^2}{m} + 2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 \\
&\overset{a)}{\leq} 2L^2C_2\eta^2\Big(\frac{\lambda^Q + 1}{(1-\lambda)^2m^{2(Q-1)}} + \frac{\lambda^Q + 1}{(1-\lambda^Q)^2}\Big) + 2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2,
\end{aligned} \tag{32}$$

where a) uses **Lemma** D.6 and $C_2 = 2K(\frac{4K^3L^2\rho^4}{2K-1} + 8K(L^2\rho^2+\sigma_g^2+\sigma_l^2) + \frac{8K}{m}\sum_{i=1}^m \mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2) + \frac{2K\rho^2}{\eta^2(2K-1)}$. Moreover, we have

$$\frac{1}{m}\sum_{i=1}^m \mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2 \leq \frac{2L^2C_3\eta^2\Big(\frac{\lambda^Q+1}{(1-\lambda)^2m^{2(Q-1)}} + \frac{\lambda^Q+1}{(1-\lambda^Q)^2}\Big) + 2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2}{1-32L^2\eta^2K^2\Big(\frac{\lambda^Q+1}{(1-\lambda)^2m^{2(Q-1)}} + \frac{\lambda^Q+1}{(1-\lambda^Q)^2}\Big)}, \tag{33}$$

where $C_3 = 2K(\frac{4K^3L^2\rho^4}{2K-1} + 8K(L^2\rho^2+\sigma_g^2+\sigma_l^2)) + \frac{2K\rho^2}{\eta^2(2K-1)}$.
Therefore, (28) is

$$\begin{aligned}
\mathbb{E}f(\overline{\mathbf{x}^{t+1}}) \leq{}& \mathbb{E}f(\overline{\mathbf{x}^t}) - \frac{\eta K}{2}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 + \frac{\eta L^2KC_1}{2} + 8\eta^3K^2L^2(2L^2C_3\eta^2\Big(\frac{\lambda^Q+1}{(1-\lambda)^2m^{2(Q-1)}} \\
&+ \frac{\lambda^Q+1}{(1-\lambda^Q)^2}\Big) + 2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2) + \frac{3\eta^2KL}{2}\left(L^2\rho^2+\sigma_l^2 + \frac{1}{m}\sum_{i=1}^m\mathbb{E}\|\nabla f(\mathbf{x}^t(i)\|^2\right) \\
\overset{a)}{\leq}{}& \mathbb{E}f(\overline{\mathbf{x}^t}) + (16\eta^3K^2L^2 - \frac{\eta K}{2} + 3\eta^2KL)\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 + \frac{3\eta^2KL}{2}\left(L^2\rho^2+\sigma_l^2\right) \\
&+ \frac{\eta KL^2C_1}{2} + (16\eta^5K^2L^4C_3 + 3\eta^4KL^3C_3)\Big(\frac{\lambda^Q+1}{(1-\lambda)^2m^{2(Q-1)}} + \frac{\lambda^Q+1}{(1-\lambda^Q)^2}\Big).
\end{aligned} \tag{34}$$

Where a) uses (33) and η is a very small value, summing the inequality (34) from $t=1$ to T, and then we can get the proved result as below:

$$\begin{aligned}
\min_{1\leq t\leq T}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 \leq{}& \frac{2f(\overline{\mathbf{x}^1}) - 2f^*}{T(\eta K - 32\eta^3K^2L^2 - 6\eta^2KL)} \\
&+ \frac{\frac{\eta L^2KC_1}{2} + \eta^4KL^3C_3(16\eta KL+3)\Big(\frac{\lambda^Q+1}{(1-\lambda)^2m^{2(Q-1)}} + \frac{\lambda^Q+1}{(1-\lambda^Q)^2}\Big) + \frac{3\eta^2KL}{2}\left(L^2\rho^2+\sigma_l^2\right)}{\eta K - 32\eta^3K^2L^2 - 6\eta^2KL}.
\end{aligned}$$

Summing the inequality (34) from $t=1$ to T, and then we can get the proved result as below:

$$\min_{1\leq t\leq T}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 \leq \frac{2f(\overline{\mathbf{x}^1}) - 2f^*}{T(\eta K - 32\eta^3K^2L^2)} + \frac{\frac{\eta L^2KC_1}{2} + 16C_2\eta^5K^2L^4\Big(\frac{\lambda^Q+1}{(1-\lambda)^2m^{2(Q-1)}} + \frac{\lambda^Q+1}{(1-\lambda^Q)^2}\Big)}{\eta K - 32\eta^3K^2L^2}.$$

If we choose the learning rate $\eta = \mathcal{O}(1/L\sqrt{KT})$ and $\eta \leq \frac{1}{10KL}$, the number of communication round T is large enough with **Definition** 4.2 and $\mathbf{\Phi}(\lambda,m,Q) = \frac{\lambda^Q+1}{(1-\lambda)^2m^{2(Q-1)}} + \frac{\lambda^Q+1}{(1-\lambda^Q)^2}$ is the key parameter to the convergence bound with the number of spectral gap, the clients and multiple gossiping steps. Thus we have

$$\min_{1\leq t\leq T}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 = \mathcal{O}\Big(\frac{(f(\overline{\mathbf{x}^1}) - f^*) + L^2\rho^2 + \sigma_l^2}{\sqrt{KT}} + \frac{K(L^2\rho^2+\sigma_g^2+\sigma_l^2)}{T} + \mathbf{\Phi}(\lambda,m,Q)\frac{K^2L^2\rho^4 + L^2\rho^2 + \sigma_g^2 + \sigma_l^2}{K^{1/2}T^{3/2}}\Big).$$

When perturbation amplitude ρ proportional to the learning rate, e.g., $\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$, and then we have:

$$\min_{1\leq t\leq T}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 = \mathcal{O}\Big(\frac{(f(\overline{\mathbf{x}^1}) - f^*) + \sigma_l^2}{\sqrt{KT}} + \frac{K(\beta^2+\sigma_l^2)}{T} + \frac{L^2}{K^{1/2}T^{3/2}} + \mathbf{\Phi}(\lambda,m,Q)\frac{\beta^2+\sigma_l^2}{K^{1/2}T^{3/2}}\Big).$$

This completes the proof.