Improving Model Consistency of Decentralized Federated Learning via Sharpness Aware Minimization and Multiple Gossip Approaches

Anonymous authors
Paper under double-blind review

Abstract

To mitigate the privacy leakages and reduce the communication burden of Federated Learning (FL), decentralized FL (DFL) discards the central server and each client only communicates with its neighbors in the decentralized communication network. However, existing DFL algorithms tend to feature high inconsistency among local models, which results in severe distribution shifts across clients and inferior performance compared with centralized FL (CFL), especially on heterogeneous data or with sparse connectivity of communication topology. To alleviate this challenge, we propose two DFL algorithms named DFedSAM and DFedSAM-MGS to improve the performance. Specifically, DFedSAM leverages gradient perturbation to generate local flatness models via Sharpness Aware Minimization (SAM), which searches for model parameters with uniformly low loss function values. In addition, DFedSAM-MGS further boosts DFedSAM by adopting the technique of Multiple Gossip Steps (MGS) for a better model consistency, which accelerates the aggregation of local flatness models and better balances the communication complexity and learning performance. In the theoretical perspective, we present the improved convergence rates $\mathcal{O}\big(\frac{1}{T} + \frac{1}{T^2(1-\lambda)^2}\big)$ and $\mathcal{O}\big(\frac{1}{T} + \frac{\lambda^Q+1}{T^2(1-\lambda^Q)^2}\big)$ in the stochastic non-convex setting for DFedSAM and DFedSAM-MGS, respectively, where $1-\lambda$ is the spectral gap of the gossip matrix W and Q is the gossip steps in MGS. Meanwhile, we empirically confirm that our methods can achieve competitive performance compared with CFL baselines and outperform existing DFL baselines.

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) framework, decentralized FL (DFL, see Figure 1), in which 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).

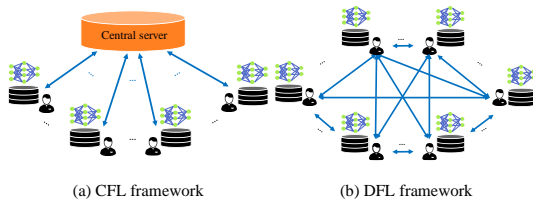

(a) CFL framework   (b) DFL framework

Figure 1: Illustrations of the types of communication CFL (a) and DFL (b) framework. For decentralized setting, the various communication network topologies are illustrated in **Appendix** A.

However, DFL suffers from bottlenecks such as severe inconsistency of local models due to heterogeneous data and model aggregation locality caused by the network connectivity of communication topology. This inconsistency results in severe over-fitting in local models and model performance degradation. Therefore, the global/consensus model may bring inferior performance compared with CFL, especially on heterogeneous data or in face of the sparse connectivity of communication net-

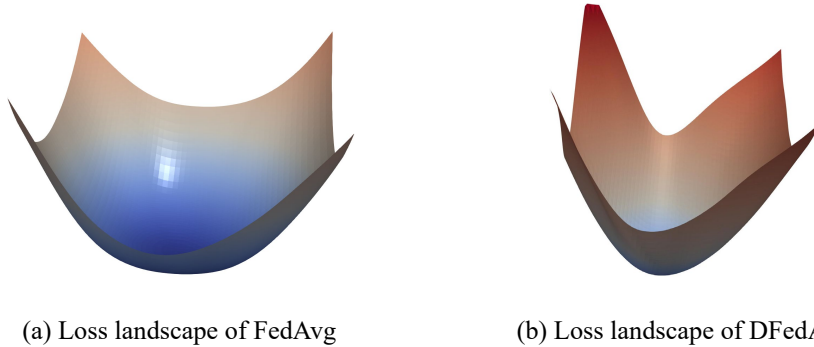(a) Loss landscape of FedAvg　　　　　　　(b) Loss landscape of DFedAvg

Figure 2: Loss landscapes comparison between CFL and DFL with the same setting. FedAvg has a more flat landscape, whereas DFedAvg has a sharper landscape than FedAvg with poorer generalization ability.

works. Similar performance pattern of DFL has also been demonstrated by Sun et al. (2022). To explore the reasons behind this phenomenon, we present the structure of the loss landscapes (Li et al., 2018) for FedAvg (Mcmahan et al., 2017) and decentralized FedAvg (DFedAvg, Sun et al. (2022)) on Fashion-MNIST (Xiao et al., 2017) with the same setting in Figure 2 (a) and (b). It is clearly seen that DFL method has a sharper landscape than CFL method.

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

To address this question, we propose two DFL algorithms: DFedSAM and DFedSAM-MGS. Specifically, DFedSAM overcomes local model over-fitting issue via gradient perturbation with SAM (Foret et al., 2021) in each client to generate local flatness models. Since each client aggregates the flatness models from its neighbors, a potential flat aggregated model can be generated, which results in high 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 realizes 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 sufficiently large number of gossip steps (see **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, while as the consensus/gossip steps Q in MGS increase, it is tighter as the impact of communication topology can be alleviated (see **Section** 4). The theoretical results directly explain why the application 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 the identical data distribution (IID) and non-IID settings. The experimental results confirm that our algorithms achieve competitive performance compared to CFL baselines and outperform DFL baselines (see **Section** 5.2).

Contribution. Our main contributions can be summarized as three-fold:

- We propose two DFL algorithms DFedSAM and DFedSAM-MGS. DFedSAM alleviates the inconsistency of local models through getting local flatness models, while DFedSAM-MGS achieves a better consistency based on DFedSAM via the aggregation acceleration and has a better trade-off between communication and generalization.

- We present the convergence rates $\mathcal{O}\left(\frac{1}{T} + \frac{1}{T^2(1-\lambda)^2}\right)$ and $\mathcal{O}\left(\frac{1}{T} + \frac{\lambda^Q+1}{T^2(1-\lambda^Q)^2}\right)$ for DFedSAM and DFedSAM-MGS in the non-convex settings, respectively, and show that our algorithms can achieve the linear speedup for convergence.

- We conduct extensive experiments to verity the efficacy of our proposed 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 in comparison to CFL, which offers communication advantage and preserves the data privacy. Lalitha et al. (2018; 2019) take a Bayesian-like approach by introducing 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 approach Brain-Torrent to FL and apply it on medical application in a highly dynamic peer-to-peer FL environment. Sun et al. (2022) apply the multiple local iteration with SGD and quantization method to further reduce the communication cost, and provide the convergence results in various convexity setting. Dai et al. (2022) develop a decentralized sparse training technique to further save the communication and computation cost.

Sharpness Aware Minimization (SAM). SAM (Foret et al., 2021) is an effective optimizer for training deep learning models, which leverages the flatness 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 non-convex objectives. As a powerful 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). Specifically, Qu et al. (2022) and Caldarola et al. (2022) integrate SAM to improve the generalization, and thus mitigate the distribution shift problem and achieve a new SOTA performance for CFL. However, to the best of our knowledge, no 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 times of local communications within a network topology is investigated in Ye et al. (2020), in which FastMix is proposed with multi-consensus and gradient tracking and they establish the optimal computational complexity and a near optimal communication complexity. DeEPCA (Ye & Zhang, 2021) integrates FastMix into a decebtralized 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 approach to a better consensus model towards the performance in CFL. Thus, the use of MGS can also potentially 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 leverages 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 suffers from inferior performance due to the severe model inconsistency issue among the clients. Another related work is FedSAM (Qu et al., 2022), which integrates SAM optimizer into CFL to enhance the flatness of local model and achieves new SOTA performance for CFL. On top of the aforementioned studies, we are the first to extend the SAM optimizer to the DFL setting and simultaneously provide its convergence guarantee in the nonconvex setting. Furthermore, we bride the gap of CFL and DFL via adopting MGS in DFedSAM-MGS, which largely mitigates the model inconsistency in DFL.

3 METHODOLOGY

In this section, we try to solve this issue in the DFL setting. Below, we first initialize the problem setup in DFL and then describe the proposed DFedSAM and DFedSAM-MGS in detail.

3.1 PROBLEM SETUP

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

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

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 the training data samples ξ. Problem (1) is known as empirical risk minimization (ERM) and models many applications in ML. As shown in Figure 1(b), we model the communication network in the decentralized network topology between clients as an undirected connected graph $\mathcal{G} = (\mathcal{N}, \mathcal{V}, \boldsymbol{W})$, where $\mathcal{N} := \{1, 2, \ldots, m\}$ represents the set of clients, and $\mathcal{V} \subseteq \mathcal{N} \times \mathcal{N}$ represents the set of communication channels, each connecting two distinct clients. Furthermore, we emphasis that there is no central server in the decentralized setting and all clients only communicate with their neighbors with respect to the communication channels $\mathcal{V}$. In addition, we assume Problem (1) is well-defined and denote f^* as the minimal value of f, i.e., $f(x) \geq f(x^*) = f^*$ for all $x \in \mathbb{R}^d$.

3.2 DFedSAM and DFedSAM-MG Algorithms

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 flatness region through adding a small perturbation to the models, i.e., $x + \delta$ with more robust performance. As shown in Figure 2, decentralized schemes has a sharper landscape with poorer generalization ability than centralized schemes. However, the study focus on this issue remains unexplored. In this paper, we extend to SAM optimizer into DFL for investigating this issue, dubbed DFedSAM, whose local loss function is defined as:

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

where $\mathbf{y}^{t,k}(i) + \delta_i$ is viewed as the perturbed model, ρ is a predefined constant controlling the radius of the perturbation and $\| \cdot \|_2^2$ is a l_2-norm, which can be simplified to $\| \cdot \|_2$ in the rest. Similar with CFL methods, in DFL, DFedSAM allows that clients can update the local model parameters with multiple local iterates before communication are performed. Specifically, for each client $i \in \{1, 2, ..., m\}$, each local iteration $k \in \{0, 1, ..., K - 1\}$ in 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 $\tilde{\mathbf{g}}^{t,k}(i) = \nabla F_i(\mathbf{y}^{t,k} + \delta(\mathbf{y}^{t,k}); \xi)$ and $\delta(\mathbf{y}^{t,k}) = \rho \mathbf{g}^{t,k} / \|\mathbf{g}^{t,k}\|_2$. Following by (Foret et al., 2021), using first order Taylor expansion around $\mathbf{y}^{t,k}$ for a small value of ρ. 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:

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

On the other hand, we use multiple gossip steps (MGS) technique (Ye et al., 2020; Ye & Zhang, 2021; Hashemi et al., 2022) to achieve a better consistency among local models based on DFedSAM, dubbed DFedSAM-MGS, thereby further boosting the performance. DFedSAM-MGS provides a balance between the communication cost and generalization ability in DFL setting. Specifically, the produce 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:

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

At the end of MGS, $\mathbf{x}^{t+1}(i) = \mathbf{x}^{t,Q}(i)$. Both DFedSAM and DFedSAM-MGS are summarized in Algorithm 1 (see **Appendix** C). It is clearly seen that DFedSAM may generate the trade-off between the local computation complexity and communication overhead via multiple local iterations, whereas the local communication is only performed at once. While DFedSAM-MGS performs multiple local communications with a larger Q to make all local clients synchronized. Therefore, DFedSAM-MGS can be viewed as compromising 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 first alleviate

local over-fitting issue caused by the inconsistency of local models via seeking a flatness model at each client in DFL, and also contribute to make consensus model flat; (ii) MGS in DFedSAM-MGS is used to further accelerate the aggregation of local flatness models for a better consistency among local models based on DFedSAM and properly balances the communication complexity and learning performance; (iii) Furthermore, we also present the theories unifying the impacts of gradient perturbation ρ in SAM, several times of local communications Q in MGS, and various network typologies λ, along with data homogeneity β upon the convergence rate in **Section** 4.

4 CONVERGENCE ANALYSIS

In this section, we show the convergence results of DFedSAM and DFedSAM-MGS for general non-convex FL setting, and the detailed proof is presented in **Appendix** E. Below, we first give several useful and necessary notations and assumptions.

Definition 1 *(The gossip/mixing matrix). (Sun et al., 2022, Definition 1) 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**) $\mathrm{null}\{\mathbf{I} - \mathbf{W}\} = \mathrm{span}\{\mathbf{1}\}$; (iv) (**Spectral property**) $\mathbf{I} \succeq \mathbf{W} \succ -\mathbf{I}$. Under these properties, eigenvalues of W can be shown satisfying $1 = |\lambda_1(\mathbf{W}))| > |\lambda_2(\mathbf{W}))| \geq \cdots \geq |\lambda_m(\mathbf{W}))|$. Furthermore, $\lambda := \max\{|\lambda_2(\mathbf{W})|, |\lambda_m(\mathbf{W}))|\}$ and $1 - \lambda \in (0,1]$ is the denoted as spectral gap of $\mathbf{W}$.*

Definition 2 *(Homogeneity parameter). (Li et al., 2020a, Definition 2) 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 1 *(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 2 *(Bounded variance). The gradient of the function f_i have σ_l-bounded variance, i.e.,*

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

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., $\sigma_g^2 \leq \beta^2$.

Assumption 3 *(Bounded gradient). For any $i \in \{1,2,\ldots,m\}$ and $\mathbf{x} \in \mathbb{R}^d$, we have $\|\nabla f_i(\mathbf{x})\| \leq B$.*

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). Difference with classic decentralized parallel SGD methods such as D-PSGD (Lian et al., 2017), the technical difficulty is that $\mathbf{z}^t(i) - \mathbf{x}^t(i)$ fails 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 network 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 to be the approximated solution of Problem (1).

Theorem 4.1 *Let Assumptions 1, 2 and 3 hold, and the parameters $\{\mathbf{x}^t(i)\}_{t \geq 0}$ is generated via Algorithm 1. Meanwhile, assume the learning rate of SAM in each client satisfy $0 < \eta \leq \frac{1}{10KL}$. 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):

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

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

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

$$\beta(K, \rho, \eta, \lambda) = \frac{64\eta^5 K^3 L^4}{\eta K - 32\eta^3 K^2 L^2} \Big(\frac{4K^3 L^2 \rho^4}{2K - 1} + 8K(L^2 \rho^2 + \sigma_g^2 + \sigma_l^2) + 8KB^2 + \frac{\rho^2}{\eta^2 (2K - 1)} \Big).$$

With Theorem 4.1, we state following convergence rates for DFedSAM and DFedSAM-MGS.

Corollary 4.1.1 *Let the local adaptive learning rate satisfy $\eta = \mathcal{O}(1/L\sqrt{KT})$. With the similar assumptions required in **Theorem** 4.1.1, and setting the perturbation parameter $\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$. Then, the convergence rate for DFedSAM satisfies:*

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

Remark 1 DFedSAM can achieve a linear speedup on the general non-convex setting as long as $T \geq K$, which is significantly better than the state-of-the-art (SOTA) bounds such as $\mathcal{O}\Big(\frac{1}{\sqrt{T}} + \frac{\sigma_g^2}{\sqrt{T}} + \frac{\sigma_g^2 + B^2}{(1 - \lambda)^2 T^{3/2}} \Big)$ in (Sun et al., 2022). Note that the bound can be tighter as λ decreases, which is dominated by $\frac{K(\beta^2 + \sigma_l^2)}{T^2 (1 - \lambda)^2}$ terms as $\lambda \leq 1 - \frac{K^{1/4}}{T^{3/2}}$, whereas as β increases, it can be degraded.

Corollary 4.1.2 *Let $Q > 1$, T be large enough and $\eta = \mathcal{O}(1/L\sqrt{KT})$. With the similar assumptions required in **Theorem** 4.1.1 and perturbation amplitude ρ being $\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$, Then, the convergence rate for DFedSAM-MGS satisfies:*

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

Remark 2 The impact of the network topology $(1 - \lambda)$ can be alleviated as Q increases and the number of clients m is large enough, the term $\frac{\lambda^Q + 1}{(1 - \lambda)^2 m^{2(Q-1)}}$ of $\Phi(\lambda, m, Q)$ 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 in the 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 degraded especially on sparse network topology where λ is close to 1. In practice, a suitable steps $Q > 1$ is possible to achieve a communication-accuracy trade-off in DFL setting.

5 EXPERIMENTS

In this section, we evaluate the efficacy of our algorithms compared to 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, Dir Partition (Hsu et al., 2019) is used for simulating non-IID across federated clients, where the local data of each client is partitioned by splitting the total dataset through sampling the label ratios from the Dirichlet distribution Dir(α) with parameters $\alpha = 0.3$ and $\alpha = 0.6$.

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 as 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 initialize the local learning rate as 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 set as 1000 in the experiments for comparing with all baselines and studying on topology-aware performance. In addition, all the ablation studies are conducted on CIFAR-10 dataset and the number of communication rounds is set as 500.

Communication Configurations. For a fair comparison between decentralized and centralized setting, 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. In specific, 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. The more experiments setup are presented in **Appendix** B due to the limited space.

Table 1: Train accuracy (%) and test accuracy (%) on two data in both IID and non-IID settings.

| Algorithm	CIFAR-10						CIFAR-100					
	Dirichlet 0.3		Dirichlet 0.6		IID		Dirichlet 0.3		Dirichlet 0.6		IID	
	Train	Validation	Train	Validation	Train	Validation	Train	Validation	Train	Validation	Train	Validation
FedAvg	99.99	82.39	99.99	84.17	99.99	84.70	99.99	48.36	99.99	53.06	99.99	54.16
FedSAM	99.75	82.49	99.89	85.04	99.98	84.98	99.99	51.98	99.99	54.88	99.99	**59.60**
D-PSGD	98.59	68.23	99.09	70.58	99.75	73.23	90.72	27.98	90.15	30.62	92.19	33.64
DFedAvg	99.75	73.55	99.93	74.67	99.95	75.55	99.56	27.62	99.56	32.82	99.68	36.77
DFedAvgM	99.93	79.96	99.95	81.56	99.95	82.07	99.56	45.11	99.60	45.50	99.78	47.98
DisPFL	99.90	72.19	99.93	74.43	99.95	85.30	97.20	30.15	99.48	32.44	99.69	35.98
DFedSAM	99.41	82.04	99.44	84.38	99.44	**85.30**	99.87	48.66	99.85	50.70	99.97	53.12
DFedSAM-MGS	99.53	**84.26**	99.65	**85.14**	99.69	**86.47**	99.92	**52.37**	99.95	**54.91**	99.97	56.15

(a) CIFAR-10
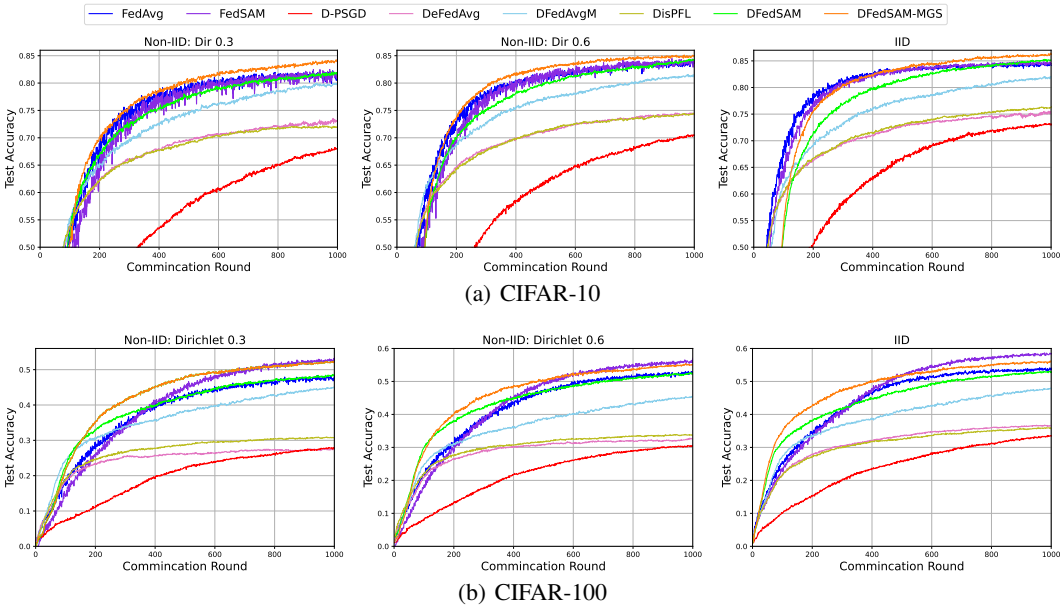
(b) CIFAR-100

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

5.2 PERFORMANCE EVALUATION

Performance with compared baselines. In Table 1 and Figure 3, 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 these results, it is clearly seen that our proposed algorithms outperform other decentralized methods on this two datasets, and DFedSAM-MGS outperforms and roughly achieves the performance of 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 difference between training accuracy and test accuracy on CIFAR-10 in IID setting is 14.14% in DFedSAM, 13.22% in DFedSAM-MGS, 15.29% in FedAvg and 15% in Fed-SAM. That means our algorithms achieve a comparable generalization than centralized baselines.

Impact of non-IID levels (β). In Table 1, we can see our algorithms are robust to different participation cases. Heterogeneous data distribution of local client is set to various participation levels from IID, Dirichlet 0.6 and Dirichlet 0.3, which makes the training of global/consensus model is more difficult. For instance, on CIFAR-10, as non-IID levels increases, DFedSAM-MGS achieves better generalization than FedSAM as the difference between training accuracy and test accuracy in DFedSAM-MGS $\{15.27\%, 14.51\%, 13.22\%\}$ are lower than that in FedSAM $\{17.26\%, 14.85\%, 15\%\}$. Similarly, the difference in DFedSAM $\{17.37\%, 15.06\%, 14.10\%\}$ are lower than that in FedAvg $\{17.60\%, 15.82\%, 15.27\%\}$. These observations confirm that our algorithms are more robust than baselines in various degrees of heterogeneous data.

5.3 TOPOLOGY-AWARE PERFORMANCE

We verify the influence of various communication topologies and gossip averaging steps in DFedSAM and DFedSAM-MGS. Different with 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 as "Complete", i.e., each client can communicate with its neighbors in the same communication round.

The degree of sparse connectivity λ is: ring > grid > exponential > full-connected in DFL. From Table 2, our algorithms are obviously superior to all decentralized baselines in various communication networks, which is coincided with our theoretical findings. Specifically, compared with DFedAvgM, DFedSAM and

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

Algorithm	Ring	Grid	Exponential	Full-connected
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 ↑

DFedSAM-MGS can significantly improve the performance in the ring topology with 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 Q can alleviate the impact of network topology with a smaller $Q = 4$. Therefore, our algorithms can generate the better generalization and model consistency in various communication topologies.

5.4 ABLATION STUDY

Below, we verify the influence of each component and hyper-parameter in DFedSAM where $Q = 1$. All the ablation studies are conducted in "exponential" topology except the study for Q in three topologies, and the communication type is the same as **Section** 5.3: "Complete".
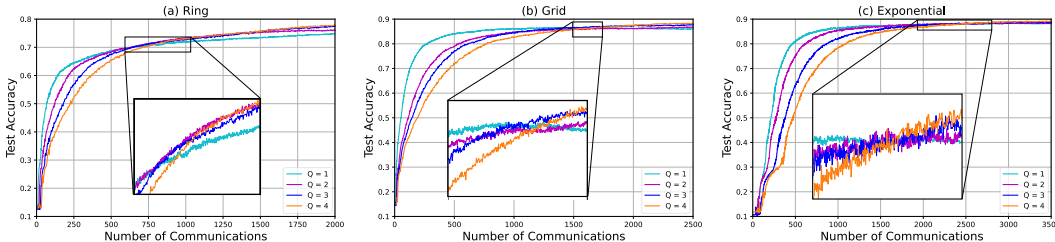

Figure 4: Test accuracy with the number of local communications in various values of Q.

Consensus/gossip steps Q. In Figure 4, we investigate where the balance between learning performance and communication complexity in three network topologies is. We choose four multiple steps $Q = \{1, 2, 3, 4\}$ and study the different balance points under the different steps in three net-

work topologies in Figure 4 (a), (b) and (c). As the number of local communications increases, model performance is also improved but the communication complexity increases too. It is clearly seen that the balance point is different but with the same tendency in different topologies. And a relatively larger Q can bring better performance for a given communication complexity. Therefore, we select $Q = 4$ in DFedSAM-MGS under 1000 communication rounds for a better balance.
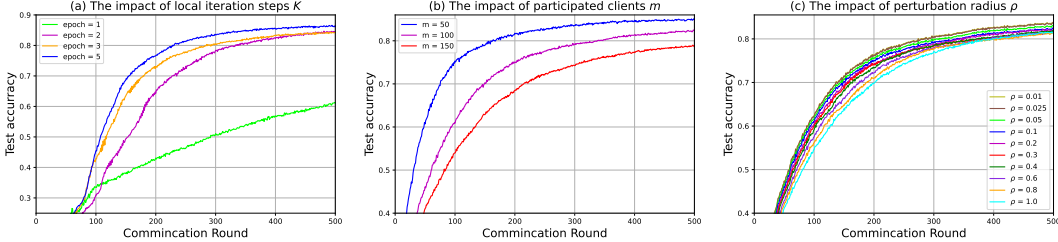

Figure 5: Impact of hyper-paramerters: local iteration steps K, participated clients m, perturbation radius ρ.

Local iteration steps K. Large local iteration steps K can help the convergence in previous DFL work (Sun et al., 2022) with the 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 5 (a), our algorithms can accelerate the convergence in theoretical results (see **Section** 4.1) as a larger local iteration steps K is adopted.

Number of participated clients m. As shown in Figure 5 (b), we compare the performance between different number of client participation $m = \{50, 100, 150\}$ with the same hyper-parameters. Compared with larger $m = 150$, the smaller $m = \{50, 100\}$ can achieve better convergence and test accuracy as the number of local data increases, which makes indirectly local model more generalization, thereby improving the performance of consensus model.

Perturbation radius ρ. Perturbation radius ρ has the impact on performance as the adding 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 some experiments on various perturbation radius from the set $\{0.01, 0.025, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0\}$ in Figure 5 (c). As $\rho = 0.01$, we achieve a better convergence and performance. Meanwhile, $\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$ can make a linear speedup on convergence (see **Section** 4.1).

The effectiveness of SAM and MGS. To validate the effectiveness of SAM and MGS, respectively, we compare four algorithms including DFedAvg, DFedSAM and FedSAM-MGS with the same setting. From Table 3, DFedSAM can achieve the

Table 3: Test accuracy of DFedAvg and DFedSAM along with DFedSAM-MGS.

Algorithm	Train (%)	Validation (%)	Differential value (%)
DFedAvg	98.86	82.10	16.76
DFedSAM	92.26	84.70 ↑	7.56 ↓
DFedSAM-MGS	92.55	85.56 ↑	6.99 ↓

performance improvement and better generalization compared with DFedAvg as 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 model inconsistency challenge caused by heterogeneous data and network connectivity of communication topology in DFL and overcome this challenge from the perspectives of model generalization. We propose two DFL frameworks: DFedSAM and DFedSAM-MGS with better model consistency among clients. DFedSAM adopts SAM to achieve the flatness model in each client, thereby improving the generalization by generating a consensus/global flatness model. Meanwhile, 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 confirm a linear speedup and unify the impacts of gradient perturbation in SAM, local communications in MGS, and network typology, along with data homogeneity upon the convergence rate in DFL. Furthermore, empirical results also verify the superiority of our approaches. For future work, we will continue towards understanding of the effect of SAM and MGS for more desira ble generalization in DFL.

REFERENCES

Momin Abbas, Quan Xiao, Lisha Chen, Pin-Yu Chen, and Tianyi Chen. Sharp-maml: Sharpness-aware model-agnostic meta learning. In *International Conference on Machine Learning, ICML*, pp. 10–32, 2022.

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N. Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *9th International Conference on Learning Representations, ICLR*, 2021.

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

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

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

Debora Caldarola, Barbara Caputo, and Marco Ciccone. Improving generalization in federated learning by seeking flat minima. *CoRR*, abs/2203.11834, 2022.

Rong Dai, Li Shen, Fengxiang He, Xinmei Tian, and Dacheng Tao. 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.

Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, and Vincent Tan. Efficient sharpness-aware minimization for improved training of neural networks. In *International Conference on Learning Representations*, 2021.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

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

Abolfazl Hashemi, Anish Acharya, Rudrajit Das, Haris Vikalo, Sujay Sanghavi, and Inderjit Dhillon. 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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, pp. 1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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

Anusha Lalitha, Shubhanshu Shekhar, Tara Javidi, and Farinaz Koushanfar. Fully decentralized federated learning. In *Third workshop on Bayesian Deep Learning (NeurIPS)*, 2018.

Anusha Lalitha, Osman Cihan Kilinc, Tara Javidi, and Farinaz Koushanfar. Peer-to-peer federated learning on graphs. *arXiv preprint arXiv:1901.11173*, 2019.

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

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, pp. 50–60, 2020b.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, pp. 429–450, 2020c.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. 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.

Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12360–12370, 2022.

H Brendan Mcmahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera Y Arcas. Communication-efficient learning of deep networks from decentralized data. pp. 1273–1282, 2017.

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

Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *International Conference on Machine Learning, ICML*, pp. 18250–18280, 2022.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=LkFG3lB13U5.

Abhijit Guha Roy, Shayan Siddiqui, Sebastian Pölsterl, Nassir Navab, and Christian Wachinger. Braintorrent: A peer-to-peer environment for decentralized federated learning. *arXiv preprint arXiv:1905.06731*, 2019.

Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. On the convergence of federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, pp. 3, 2018.

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

Tao Sun, Dongsheng Li, and Bao Wang. Decentralized federated averaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Han Wang, Siddartha Marella, and James Anderson. Fedadmm: A federated primal-dual algorithm allowing partial participation. *arXiv preprint arXiv:2203.15104*, 2022.

Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

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

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

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

Haishan Ye, Ziang Zhou, Luo Luo, and Tong Zhang. Decentralized accelerated proximal gradient descent. *Advances in Neural Information Processing Systems*, 33:18308–18317, 2020.

Hao Yu, Sen Yang, and Shenghuo Zhu. 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.

Yang Zhao, Hao Zhang, and Xiuyuan Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. In *International Conference on Machine Learning, ICML*, pp. 26982–26992. PMLR, 2022.
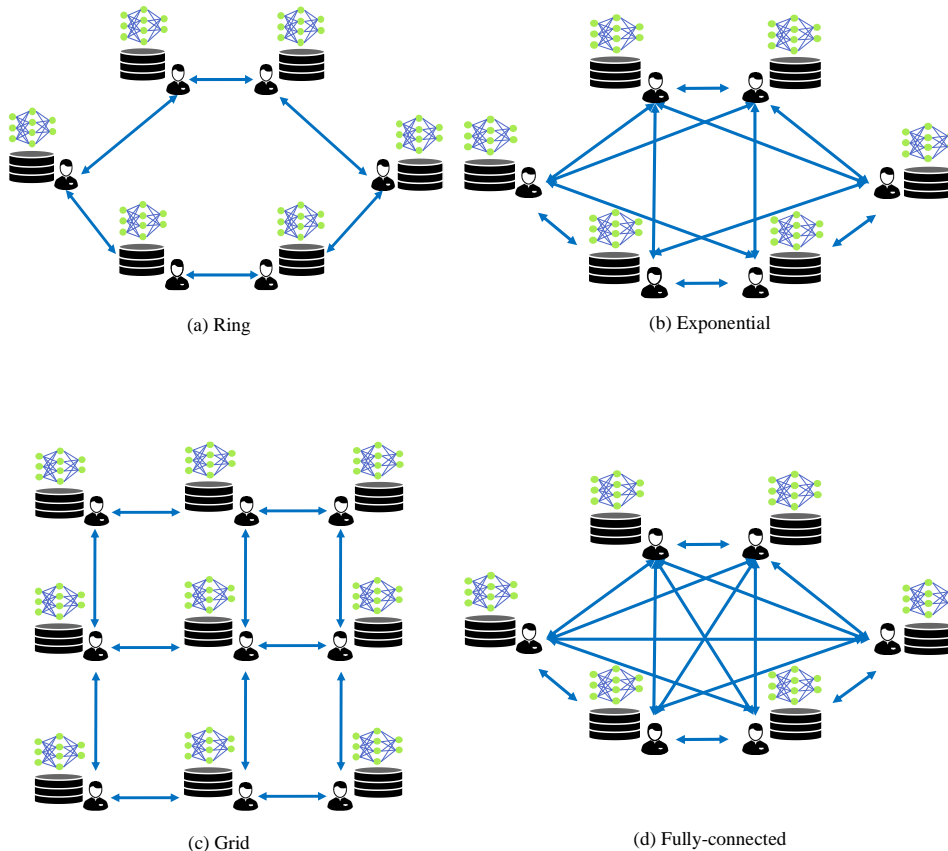
A   COMMUNICATION NETWORK TOPOLOGIES

(a) Ring

(b) Exponential

(c) Grid

(d) Fully-connected

Figure 6: 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 [1], 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

---

[1] 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.

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 Fed-SAM. 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 ALGORITHMS

D 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 $\|\mathbf{W}^t - \mathbf{P}\|_{\mathrm{op}} \leq \lambda^t$, where $\lambda := \max\{|\lambda_2|, |\lambda_m(W)|\}$ and for a matrix $\mathbf{A}$, we denote its spectral norm as $\|\mathbf{A}\|_{\mathrm{op}}$. Furthermore, $\mathbf{1} := [1, 1, \ldots, 1]^\top \in \mathbb{R}^m$ and*

$$\mathbf{P} := \frac{\mathbf{11}^\top}{m} \in \mathbb{R}^{m \times m}.$$

In [Proposition 1, (Nedic & Ozdaglar, 2009)], the author also proved that $\|W^t - \mathbf{P}\|_{\mathrm{op}} \leq C\lambda^t$ 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 1 and 2, 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 $\delta = \rho \frac{\nabla F(\mathbf{x})}{\|\nabla F(\mathbf{x})\|}$, $\delta_{i,k} = \rho \frac{\nabla F_i(\mathbf{y}^{t,k}, \xi)}{\|\nabla F_i(\mathbf{y}^{t,k}, \xi)\|}$.

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: Generate consensus model $\mathbf{x}^T$ after the final communication of all clients with their neighbors.

1 **Initialization:** Randomly initialize each client's model $\mathbf{x}^0(i)$.

2 **for** $t = 0$ **to** $T - 1$ **do**

3   **for** *node i in parallel* **do**

4   **for** $k = 0$ **to** $K - 1$ **do**

5   Set $\mathbf{y}^{t,0}(i) \leftarrow \mathbf{x}^t(i)$, $\mathbf{y}^{t,-1}(i) = \mathbf{y}^{t,0}(i)$

6   Sample a batch of local data ξ_i and calculate local gradient $\mathbf{g}^{t,k}(i) = \nabla F_i(\mathbf{y}^{t,k}; \xi_i)$

7        $\tilde{\mathbf{g}}^{t,k}(i) = \nabla F_i(\mathbf{y}^{t,k} + \delta(\mathbf{y}^{t,k}); \xi_i)$ with $\delta(\mathbf{y}^{t,k}) = \rho \mathbf{g}^{t,k} / \left\| \mathbf{g}^{t,k} \right\|_2$

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

9   **end**

10      $\mathbf{z}^t(i) \leftarrow \mathbf{y}^{t,K}(i)$

11   Receive neighbors' models $\mathbf{z}^t(l)$ from neighborhood set $\mathcal{S}_{k,t}$ with adjacency matrix $\mathbf{W}$.

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

13   **for** $q = 0$ **to** $Q - 1$ **do**

14        $\mathbf{x}^{t,q+1}(i) = \sum_{l \in \mathcal{N}(i)} \boldsymbol{w}_{i,l} \mathbf{z}^{t,q}(l)$ $(\mathbf{z}^{t,0}(i) = \mathbf{z}^t(i))$ (Exchanging messages)

15        $\mathbf{z}^{t,q+1}(i) = \mathbf{x}^{t,q+1}(i)$ (Local gossip update)

16   **end**

17      $\mathbf{x}^{t+1}(i) = \mathbf{x}^{t,Q}(i)$

18   **end**

19 **end**

E   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** 4.1 is thoroughly introduced in two sections E.2 and E.3 as follows, where $Q = 1$ and $Q > 1$, respectively.

E.1   PRELIMINARY LEMMAS

Lemma E.1 *Assume that Assumptions 1 and 2 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:*

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

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

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

and

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

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

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

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

$$\le 2K(\frac{4K^3L^2\eta^2\rho^4}{2K-1} + 8K\eta^2(L^2\rho^2 + \sigma_g^2 + \sigma_l^2)$$

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

This completes the proof.

Lemma E.2 Assume that Assumption 3 holds and the number of local iteration K is large enough. Let $\{\mathbf{x}^t(i)\}_{t\ge 0}$ 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 $C_2 = 2K(\frac{4K^3L^2\rho^4}{2K-1} + 8K(L^2\rho^2 + \sigma_g^2 + \sigma_l^2) + 8KB^2) + \frac{2K\rho^2}{\eta^2(2K-1)}$.
Proof.
Following [Lemma 4, (Sun et al., 2022)], we denote $\mathbf{Z}^t := \left[\mathbf{z}^t(1), \mathbf{z}^t(2), \dots, \mathbf{z}^t(m)\right]^\top \in \mathbb{R}^{m \times d}$. 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

$$\mathbf{WP} = \mathbf{PW} = \mathbf{P}. \tag{11}$$

According to Lemma D.1, it holds

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

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

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

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

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

With Cauchy inequality,

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

Direct calculation gives us

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

*With **Lemma E.1** and Assumption 3, for any j,*

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{X}^j - \mathbf{Z}^j\|^2 \\
&\le m\Big(2K(\frac{4K^3L^2\rho^4}{2K-1} + 8K(L^2\rho^2 + \sigma_g^2 + \sigma_l^2) + 8KB^2) + \frac{2K\rho^2}{\eta^2(2K-1)}\Big)\eta^2.
\end{aligned}
$$

Thus, we get

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{X}^t - \mathbf{PX}^t\|^2 \\
&\le \frac{m\Big(2K(\frac{4K^3L^2\rho^4}{2K-1} + 8K(L^2\rho^2 + \sigma_g^2 + \sigma_l^2) + 8KB^2) + \frac{2K\rho^2}{\eta^2(2K-1)}\Big)\eta^2}{(1-\lambda)^2}.
\end{aligned}
$$

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

Lemma E.3 *Assume that Assumption 3 holds and the number of local iteration K is large enough. Let $\{\mathbf{x}^t(i)\}_{t \ge 0}$ be generated by DFedSAM-MGS for all $i \in \{1, 2, ..., m\}$ and any learning rate $\eta > 0$, we have following bound:*

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

where $C_2 = 2K(\frac{4K^3L^2\rho^4}{2K-1} + 8K(L^2\rho^2 + \sigma_g^2 + \sigma_l^2) + 8KB^2) + \frac{2K\rho^2}{\eta^2(2K-1)}$.
Proof.

Following [Lemma 4, (Sun et al., 2022)] and **Lemma** E.2, we denote $\mathbf{Z}^t := \left[\mathbf{z}^t(1), \mathbf{z}^t(2), \dots, \mathbf{z}^t(m)\right]^\top \in \mathbb{R}^{m \times d}$. 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 $\mathbf{X}^0 = \mathbf{0}$. 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,

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

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 E.1** and Assumption 3, for any j,

$$\begin{aligned}
&\mathbb{E}\|\mathbf{X}^j - \mathbf{Z}^j\|^2 \\
&\leq m\Big(2K(\frac{4K^3 L^2 \rho^4}{2K-1} + 8K(L^2 \rho^2 + \sigma_g^2 + \sigma_l^2) + 8KB^2) + \frac{2K\rho^2}{\eta^2(2K-1)}\Big)\eta^2.
\end{aligned}$$

Thus, we get

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

where $C_2 = 2K(\frac{4K^3 L^2 \rho^4}{2K-1} + 8K(L^2\rho^2 + \sigma_g^2 + \sigma_l^2) + 8KB^2) + \frac{2K\rho^2}{\eta^2(2K-1)}$.

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.*

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

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

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:

$$\mathbb{E}\langle K\nabla f(\overline{\mathbf{x}^t}), (\overline{\mathbf{z}^t} - \overline{\mathbf{x}^t})/K \rangle = \mathbb{E}\langle K\nabla f(\overline{\mathbf{x}^t}), -\eta\nabla f(\overline{\mathbf{x}^t}) + \eta\nabla f(\overline{\mathbf{x}^t}) + (\overline{\mathbf{z}^t} - \overline{\mathbf{x}^t})/K \rangle$$

$$= -\eta K\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 + \mathbb{E}\langle K\nabla f(\overline{\mathbf{x}^t}), \frac{\eta}{mK}\sum_{i=1}^m\sum_{k=0}^{K-1}\Big(\nabla f(\mathbf{x}^t(i)) - \nabla F_i(\mathbf{y}^{t,k} + \delta_{i,k}; \xi)\Big)\rangle$$

$$\overset{a)}{\leq} \eta\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\| \cdot \left\|\frac{L}{m}\sum_{i=1}^m\sum_{k=0}^{K-1}(\mathbf{x}^t(i) - \mathbf{y}^{t,k} - \delta_{i,k})\right\| \tag{20}$$

$$\overset{b)}{\leq} \frac{\eta K}{2}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 + \frac{\eta L^2 K^2}{2K}\Big(2K(\frac{4K^3 L^2 \eta^2 \rho^4}{2K-1} + 8K\eta^2(L^2\rho^2 + \sigma_g^2 + \sigma_l^2)$$

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

where $a)$ uses the Lipschitz continuity, $b)$ uses **Lemma** E.1. Meanwhile, we can get

$$\frac{L}{2}\mathbb{E}\left\|\overline{\mathbf{x}^{t+1}} - \overline{\mathbf{x}^t}\right\|^2 = \frac{L}{2}\mathbb{E}\left\|\overline{\mathbf{z}^t} - \overline{\mathbf{x}^t}\right\|^2 \leq \frac{L}{2}\frac{1}{m}\sum_{i=1}^m\left\|\mathbf{y}^{t,K}(i) - \mathbf{x}^t(i)\right\|^2$$

$$\leq \frac{L}{2}\mathbb{E}\left\|\frac{-\eta\sum_{i=1}^m\sum_{k=0}^{K-1}\nabla F_i(\mathbf{y}^{t,k} + \delta_{i,k}; \xi)}{m}\right\|^2 \tag{21}$$

$$\overset{a)}{\leq} \frac{L}{2}\eta^2 K^2 B^2,$$

where $a)$ uses **Assumption** 3.

Furthermore, (19) can be represented as

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

where $C_1 = 2K(\frac{4K^3 L^2 \eta^2 \rho^4}{2K-1} + 8K\eta^2(L^2\rho^2 + \sigma_g^2 + \sigma_l^2) + \frac{2K\rho^2}{2K-1}$.

Thus, with **Lemma** E.2, 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** E.2 and $C_2 = 2K(\frac{4K^3 L^2 \rho^4}{2K-1} + 8K(L^2\rho^2 + \sigma_g^2 + \sigma_l^2) + 8KB^2) + \frac{2K\rho^2}{\eta^2(2K-1)}$.

Therefore, (19) is

$$
\begin{aligned}
\mathbb{E}f(\overline{\mathbf{x}^{t+1}}) \leq & \mathbb{E}f(\overline{\mathbf{x}^t}) - \frac{\eta K}{2}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 + \frac{\eta L^2 K C_1}{2} \\
& + 8\eta^3 K^2 L^2(2L^2\frac{C_2\eta^2}{(1-\lambda)^2} + 2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2) \\
\leq & \mathbb{E}f(\overline{\mathbf{x}^t}) + (16\eta^3 K^2 L^2 - \frac{\eta K}{2})\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 \\
& + \frac{\eta L^2 K C_1}{2} + \frac{16C_2\eta^5 K^2 L^4}{(1-\lambda)^2}.
\end{aligned}
\tag{24}
$$

Summing the inequality (24) from $t = 1$ to T, and then we can get the proved result as below:

$$
\min_{1\leq t\leq T}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 \leq \frac{2f(\overline{\mathbf{x}^1}) - 2f^*}{T(\eta K - 32\eta^3 K^2 L^2)} + \frac{\frac{\eta L^2 K C_1}{2} + \frac{16C_2\eta^5 K^2 L^4}{(1-\lambda)^2}}{\eta K - 32\eta^3 K^2 L^2}.
$$

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, we have

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

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

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

Under **Definition** 2, we can get

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

This completes the proof.

E.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 $\mathbf{PX}^{t+1} = \mathbf{PW}^Q\mathbf{Z}^t = \mathbf{PZ}^t(Q > 1)$, that is also

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

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

$$\overline{\mathbf{x}^{t+1}} - \overline{\mathbf{x}^t} = \overline{\mathbf{x}^{t+1}} - \overline{\mathbf{z}^t} + \overline{\mathbf{z}^t} - \overline{\mathbf{x}^t} = \overline{\mathbf{z}^t} - \overline{\mathbf{x}^t}, \tag{25}$$

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,

$$\begin{aligned}
\overline{\mathbf{z}^t} - \overline{\mathbf{x}^t} &= \frac{\sum_{i=1}^m (\sum_{k=0}^{K-1} \mathbf{y}^{t,k+1}(i) - \mathbf{y}^{t,k}(i))}{m} \\
&= \frac{\sum_{i=1}^m \sum_{k=0}^{K-1}(-\eta \tilde{\mathbf{g}}^{t,k}(i))}{m} \\
&= \frac{\sum_{i=1}^m \sum_{k=0}^{K-1}(-\eta \nabla F_i(\mathbf{y}^{t,k} + \rho \nabla F_i(\mathbf{y}^{t,k};\xi)/\left\|\nabla F_i(\mathbf{y}^{t,k};\xi)\right\|_2);\xi)}{m}.
\end{aligned} \tag{26}$$

The Lipschitz continuity of ∇f:

$$\mathbb{E}f(\overline{\mathbf{x}^{t+1}}) \leq \mathbb{E}f(\overline{\mathbf{x}^t}) + \mathbb{E}\langle \nabla f(\overline{\mathbf{x}^t}), \overline{\mathbf{z}^t} - \overline{\mathbf{x}^t}\rangle + \frac{L}{2}\mathbb{E}\|\overline{\mathbf{x}^{t+1}} - \overline{\mathbf{x}^t}\|^2, \tag{27}$$

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

$$\begin{aligned}
&\mathbb{E}\langle K\nabla f(\overline{\mathbf{x}^t}), (\overline{\mathbf{z}^t} - \overline{\mathbf{x}^t})/K\rangle = \mathbb{E}\langle K\nabla f(\overline{\mathbf{x}^t}), -\eta\nabla f(\overline{\mathbf{x}^t}) + \eta\nabla f(\overline{\mathbf{x}^t}) + (\overline{\mathbf{z}^t} - \overline{\mathbf{x}^t})/K\rangle \\
&= -\eta K\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 + \mathbb{E}\langle K\nabla f(\overline{\mathbf{x}^t}), \frac{\eta}{mK}\sum_{i=1}^m \sum_{k=0}^{K-1}\left(\nabla f(\mathbf{x}^t(i)) - \nabla F_i(\mathbf{y}^{t,k} + \delta_{i,k};\xi)\right)\rangle \\
&\overset{a)}{\leq} \eta\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\| \cdot \left\|\frac{L}{m}\sum_{i=1}^m \sum_{k=0}^{K-1}(\mathbf{x}^t(i) - \mathbf{y}^{t,k} - \delta_{i,k})\right\| \\
&\overset{b)}{\leq} \frac{\eta K}{2}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 + \frac{\eta L^2 K^2}{2K}\left(2K(\frac{4K^3 L^2 \eta^2 \rho^4}{2K-1} + 8K\eta^2(L^2\rho^2 + \sigma_g^2 + \sigma_l^2)\right. \\
&\left. + \frac{8K\eta^2}{m}\sum_{i=1}^m \mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2) + \frac{2K\rho^2}{2K-1}\right),
\end{aligned} \tag{28}$$

where $a)$ uses the Lipschitz continuity, $b)$ uses **Lemma** E.1. Meanwhile, we can get

$$\begin{aligned}
\frac{L}{2}\mathbb{E}\left\|\overline{\mathbf{x}^{t+1}} - \overline{\mathbf{x}^t}\right\|^2 = \frac{L}{2}\mathbb{E}\left\|\overline{\mathbf{z}^t} - \overline{\mathbf{x}^t}\right\|^2 &\leq \frac{L}{2}\frac{1}{m}\sum_{i=1}^m \left\|\mathbf{y}^{t,K}(i) - \mathbf{x}^t(i)\right\|^2 \\
&\leq \frac{L}{2}\mathbb{E}\left\|\frac{-\eta\sum_{i=1}^m \sum_{k=0}^{K-1}\nabla F_i(\mathbf{y}^{t,k} + \delta_{i,k};\xi)}{m}\right\|^2 \\
&\overset{a)}{\leq} \frac{L}{2}\eta^2 K^2 B^2,
\end{aligned} \tag{29}$$

where $a)$ uses **Assumption** 3.
Furthermore, (19) can be represented as

$$\begin{aligned}
\mathbb{E}f(\overline{\mathbf{x}^{t+1}}) \leq &\mathbb{E}f(\overline{\mathbf{x}^t}) - \frac{\eta K}{2}\mathbb{E}\left\|\nabla\mathbb{E}f(\overline{\mathbf{x}^t})\right\|^2 + \frac{\eta L^2 K}{2}C_1 \\
&+ \frac{8\eta^3 K^2 L^2}{m}\sum_{i=1}^m \mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2 + \frac{L}{2}\eta^2 K^2 B^2,
\end{aligned} \tag{30}$$

where $C_1 = 2K(\frac{4K^3L^2\eta^2\rho^4}{2K-1} + 8K\eta^2(L^2\rho^2 + \sigma_g^2 + \sigma_l^2) + \frac{2K\rho^2}{2K-1}$.

Thus, with **Lemma** E.3, we can get

$$
\frac{1}{m}\sum_{i=1}^m \mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2 \le 2L^2 \frac{\sum_{i=1}^m \left\|\mathbf{x}^t(i) - \overline{\mathbf{x}^t}\right\|^2}{m} + 2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2
$$

$$
\overset{a)}{\le} 2L^2 C_2 \eta^2 \Big(\frac{\lambda^Q + 1}{(1-\lambda)^2 m^{2(Q-1)}} + \frac{\lambda^Q + 1}{(1-\lambda^Q)^2}\Big) + 2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2,
$$

(31)

where a) uses **Lemma** E.3 and $C_2 = 2K(\frac{4K^3L^2\rho^4}{2K-1} + 8K(L^2\rho^2 + \sigma_g^2 + \sigma_l^2) + 8KB^2) + \frac{2K\rho^2}{\eta^2(2K-1)}$.

Therefore, (19) is

$$
\mathbb{E}f(\overline{\mathbf{x}^{t+1}}) \le \mathbb{E}f(\overline{\mathbf{x}^t}) - \frac{\eta K}{2}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 + \frac{\eta L^2 K C_1}{2}
$$

$$
+ 8\eta^3 K^2 L^2 (2L^2 C_2 \eta^2 \Big(\frac{\lambda^Q + 1}{(1-\lambda)^2 m^{2(Q-1)}} + \frac{\lambda^Q + 1}{(1-\lambda^Q)^2}\Big) + 2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2)
$$

$$
\le \mathbb{E}f(\overline{\mathbf{x}^t}) + (16\eta^3 K^2 L^2 - \frac{\eta K}{2})\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2
$$

$$
+ \frac{\eta L^2 K C_1}{2} + 16 C_2 \eta^5 K^2 L^4 \Big(\frac{\lambda^Q + 1}{(1-\lambda)^2 m^{2(Q-1)}} + \frac{\lambda^Q + 1}{(1-\lambda^Q)^2}\Big).
$$

(32)

Summing the inequality (32) from $t = 1$ to T, and then we can get the proved result as below:

$$
\min_{1 \le t \le T} \mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 \le \frac{2f(\overline{\mathbf{x}^1}) - 2f^*}{T(\eta K - 32\eta^3 K^2 L^2)} + \frac{\frac{\eta L^2 K C_1}{2} + 16 C_2 \eta^5 K^2 L^4 \Big(\frac{\lambda^Q+1}{(1-\lambda)^2 m^{2(Q-1)}} + \frac{\lambda^Q+1}{(1-\lambda^Q)^2}\Big)}{\eta K - 32\eta^3 K^2 L^2}.
$$

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

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

$$
\Phi(\lambda, m, Q)\Big(\frac{L^2\rho^2}{T} + \frac{K^2L^2\rho^4}{T^2} + \frac{K(L^2\rho^2 + \beta^2 + \sigma_l^2 + B^2)}{T^2}\Big)\Big).
$$

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

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

$$
+ \Phi(\lambda, m, Q)\frac{L^2 + K(\beta^2 + \sigma_l^2 + B^2)}{T^2}\Big).
$$

This completes the proof.