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]

[1] Tsinghua University, Shenzhen, China; [2] JD Explore Academy, Beijing, China;
[3] Hong Kong Polytechnic University, Hong Kong, China; [4] University of Sydney, Sydney, Australia.

shiyf21@mails.tsinghua.edu.cn; mathshenli@gmail.com; wang.xq@sz.tsinghua.edu.cn

Background

General Federated Learning (FL) with central server

- A learning paradigm allows distributed clients to collaboratively train a shared model without sharing data under the coordination of the central server.

- Challenges: privacy leakages and communication burdens.

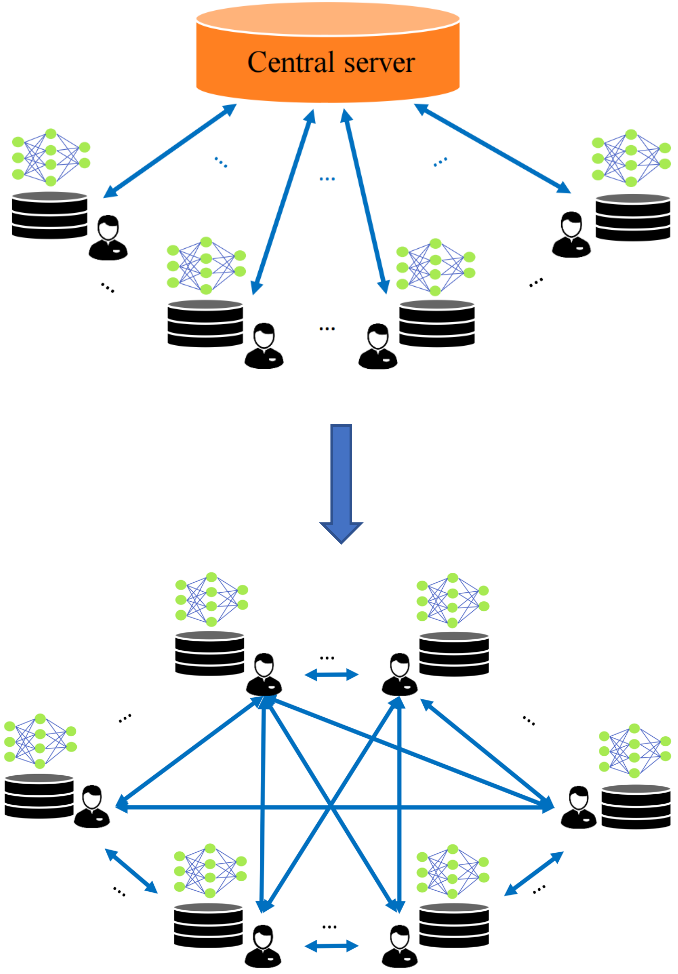
A solution

Decentralized Federated Learning (DFL)

- It discards the central server and each client only communicates with its neighbors in a decentralized communication network.

- But it may suffer from **high inconsistency** among local clients, which results in
  1. severe distribution shift
  2. inferior performance

  } Compared with centralized FL (CFL)



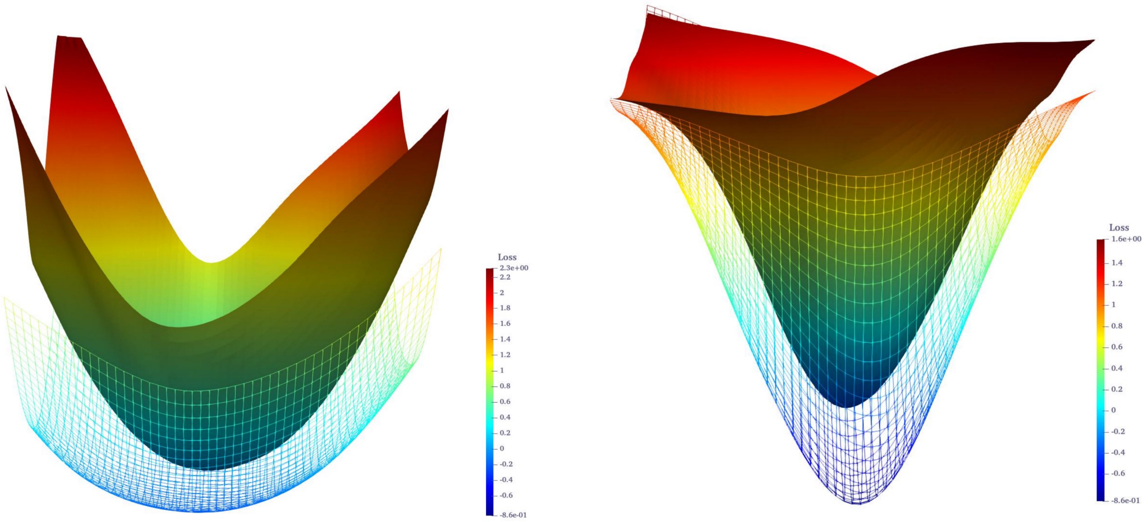
(b) DFL framework

Motivation

Observation:

Compare the structure of loss landscapes for **FedAvg (mesh plot)** ***v.s.*** **Decentralized FedAvg (surface plot)** on partitioned Fashion-MNIST and CIFAR-10 datasets with the same setting.

Sharper loss landscape means poor generalization ability



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

Research Question:

Can we design DFL algorithms that can mitigate the inconsistency among local models and achieve similar performance to its centralized counterpart?

Tsinghua University

ICML 40 Years International Conference On Machine Learning

Problem Setting and Challenges in DFL

Problem Setting:

The 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: $f(x) \ge f(x^*) = f^*$ for all $x \in \mathbb{R}^d$.

Challenges in DFL:

- *Various communication topologies*. A significant negative impact on model training (convergence rate and generalization ability).

- *Multi-step local iterations*. The corresponding theoretical analysis may be more difficult and the empirical efficacy may also suffer compared to the one-step local iteration.

The Details of Methodology

Algorithm

- Local loss function is defined as:

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

- The k-th inner iteration in client i is performed as:

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

$$\tilde{\mathbf{g}}^{t,k}(i) = \nabla F_i(\mathbf{y}^{t,k} + \delta(\mathbf{y}^{t,k}); \xi)$$

$$\delta(\mathbf{y}^{t,k}) = \rho\mathbf{g}^{t,k} / \left\|\mathbf{g}^{t,k}\right\|_2$$

- Each client averages its parameters with the information of its neighbors (including itself):

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

- MGS at the q-th step (q ∈ {0, 1, ..., Q − 1}):

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

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 $\delta(\mathbf{y}^{t,k}) = \rho\mathbf{g}^{t,k} / \left\|\mathbf{g}^{t,k}\right\|_2$
      $\mathbf{y}^{t,k+1}(i) = \mathbf{y}^{t,k}(i) - \eta\tilde{\mathbf{g}}^{t,k}(i)$
    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

Theoretical Analysis

Problem Setting

$$\min_{\mathbf{x}\in R^d} f(\mathbf{x}) := \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), f_i(\mathbf{x}) = \mathrm{E}_{\xi\sim D_i} F_i(\mathbf{x};\xi)$$

Assumption

- **Homogeneity parameter**

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

- **Lipschitz smoothness**

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \le L\|\mathbf{x}-\mathbf{y}\|. \quad \forall i \in \{1,2,\ldots,m\}$$

- **Bounded variance**

Local : $\mathrm{E}_{\xi_i}\left\|\nabla F_i(\mathbf{y};\xi_i) - \nabla f_i(\mathbf{x})\right\|^2 \le \sigma_l^2, \forall i \in \{1,2,\ldots,m\}$

Global : $\frac{1}{m}\sum_{i=1}^{m}\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \le \sigma_g^2$

- ## Convergence Analysis

  - **DFedSAM**

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

  - **DFedSAM-MGS**

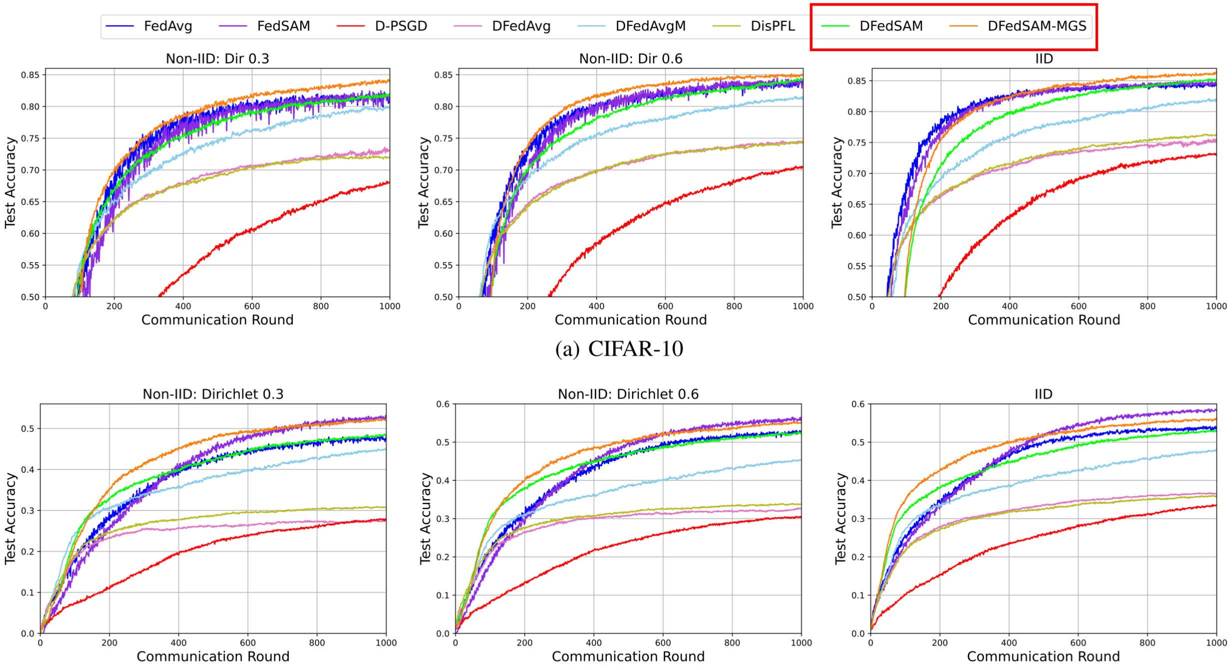
$$\min_{1\le t\le T} \mathrm{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2 = O\Big(\frac{(f(\overline{\mathbf{x}^1}) - f^*) + \sigma_l^2}{\sqrt{KT}} + \frac{K(\beta^2+\sigma_l^2)}{T} + \frac{L^2}{K^{1/2}T^{3/2}} + \mathbf{\Phi}(\lambda,m,Q)\frac{\beta^2+\sigma_l^2}{K^{1/2}T^{3/2}}\Big)$$

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

$1-\lambda$ and m is the spectral gap of gossip matrix and the total numbers of clients, and Q is the number of MGS.

Experiments

Performance with compared baselines.



(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.

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

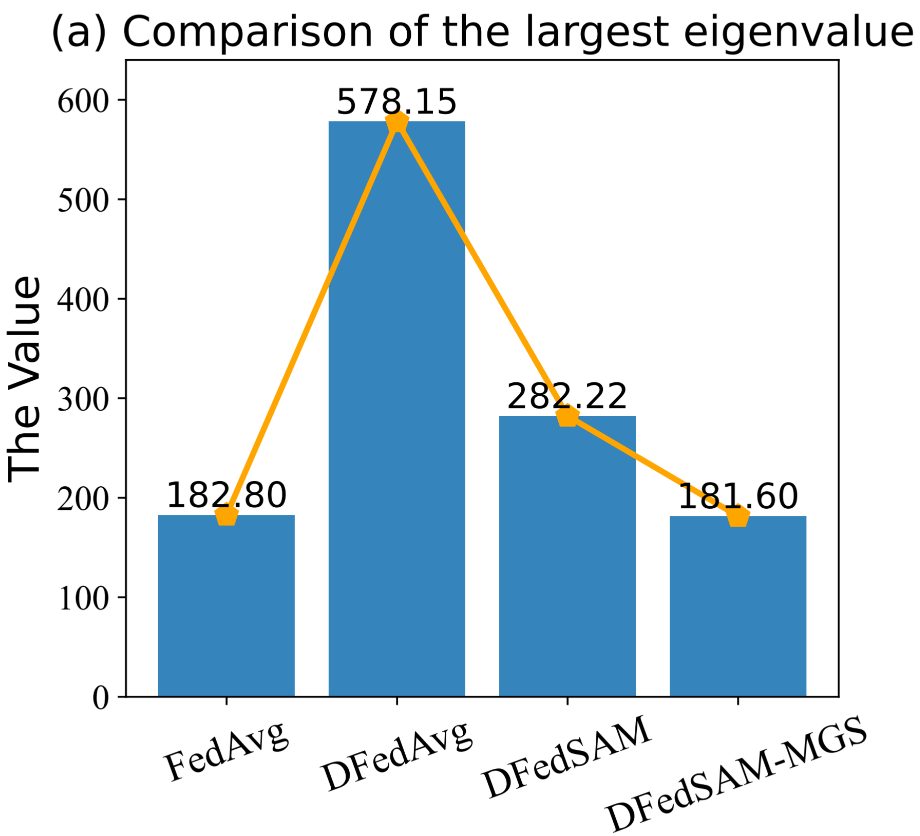
- Outperform other baselines on both accuracy and generalization perspectives.
- More robust than baselines in various degrees of heterogeneous data.

Tsinghua University

ICML 40 Years International Conference On Machine Learning

Experiments

- **Measuring on the Flatness of Loss Landscape**



The smaller the largest eigenvalue, the flatter the loss landscape.

- **Topology-aware Performance**

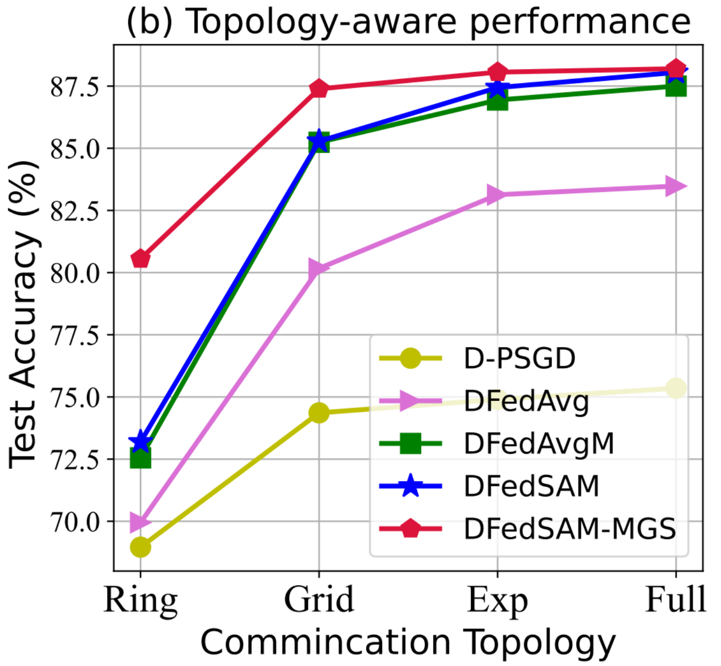


Table 2. Testing accuracy (%) in various network topologies compared with decentralized algorithms on CIFAR-10.

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 ↑

Our algorithms can achieve better generalization and model consistency with various communication topologies.

Thank You!

For any questions, you can find us at

shiyf21@mails.tsinghua.edu.cn;
kang.wei@njust.edu.cn;
mathshenli@gmail.com

Scan for paper!

Overview

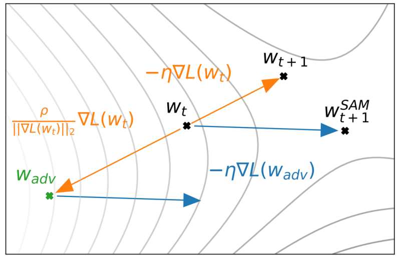
1. Background:

Decentralized Federated Learning (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).

2. Our goal:

We aim to improve the model consistency of DFL via leveraging *Sharpness-Aware Minimization (SAM)* optimizer and Multiple Gossip Steps (MGS).

SAM optimizer



3. Our contribution:

- 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***, 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.