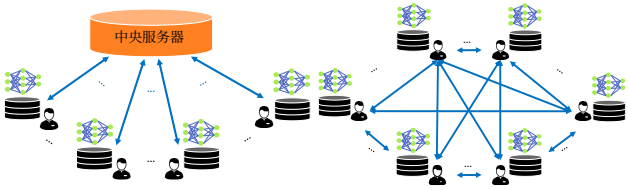
arXiv:2302.04083v2 [cs.LG] 9 Jun 2023

改进 去中心化联邦学习的模型一致性

史一帆 [1] 沈立 [2] 魏康 [3] 孙岩 [4] 袁博 [1] 王学谦 [1] 陶大程 [4]

摘要

为减轻联邦学习（FL）中的隐私泄露和通信负担，去中心化联邦学习（DFL）摒弃中央服务器，每个客户端仅在去中心化通信网络中与其邻居通信。然而，现有 DFL 中本地客户端之间的一致性较差，导致严重的分布偏移；尤其是在异构数据或通信拓扑稀疏的情况下，其性能不如集中式联邦学习（CFL）。为缓解这一问题，我们提出了两种 DFL 算法，分别称为 DFedSAM 和 DFedSAM-MGS，以提升 DFL 的性能。具体而言，DFedSAM 利用梯度扰动，通过尖锐度感知最小化（SAM）生成本地平坦模型；该方法旨在寻找损失值均匀较低的模型。DFedSAM-MGS 则通过采用多次 gossip 步骤（MGS）进一步增强 DFedSAM，以提升模型一致性；这能加速本地平坦模型的聚合，并更好地权衡通信复杂度与泛化能力。从理论上看，我们分别证明了 DFedSAM 和 DFedSAM-MGS 在非凸情形下具有改进的收敛速率 $\mathcal{O}(\frac{1}{\sqrt{KT}}+\frac{1}{T}+\frac{1}{K^{1/2}T^{3/2}(1-\lambda)^2})$ 和 $\mathcal{O}(\frac{1}{\sqrt{KT}}+\frac{1}{T}+\frac{\lambda^Q+1}{K^{1/2}T^{3/2}(1-\lambda^Q)^2})$，其中 $1-\lambda$ 为 gossip 矩阵的谱间隙，Q 为 MGS 的次数。从实验结果来看，与集中式联邦学习方法相比，我们的方法具有有竞争力的性能，并优于现有的 DFL 方法。



(a) 集中式联邦学习框架 (b) 去中心化联邦学习框架

图 *1*。CFL（a）和 DFL（b）的示意图。DFL 的各种通信拓扑见**附录** A。

1. 引言

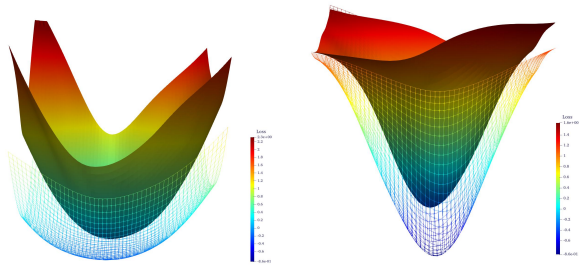
联邦学习（FL）（Mcmahan 等人，2017年；Li 等人，2020b）使分布式客户端能够在云端的协调下协同训练共享模型，而无需传输本地数据。然而，几乎所有联邦学习范式都采用中央服务器与客户端通信，这面临若干关键挑战，例如计算资源有限、通信带宽成本高以及隐私泄露（Kairouz 等人，2021年）。与集中式联邦学习（CFL，图 1（a））相比，去中心化联邦学习（DFL，图 1（b））中，客户端仅与邻居通信，无需中央服务器，因此具有通信优势，并能进一步保护数据隐私（Kairouz 等人，2021年；Wang 等人，2021年；Sun 等人，2022年）。

由于实际联邦系统中的参与者具备不同的硬件和网络能力，去中心化联邦学习是一个很有前景的研究领域，近年来在多篇综述文章中经常被视为一项挑战（Beltr´an 等人。，2022年；Kairouz 等人，2021年）。在实践中，最具前景的去中心化联邦学习应用场景包括医疗保健（Nguyen 等人，2022年）、工业 4.0，例如区块链系统（Kang 等人，2022年；Lietal。，2022年）、物联网（IoT）中的移动服务（Wang 等人，2022b）、无人机（UAVs）的鲁棒网络（Wang 等人，2020年）以及车联网（Yu 等人，2020年）。

然而，由于异构数据分布以及内在网络连接性导致的模型聚合局部性，DFL 中的本地模型存在严重的不一致性。这种不一致性可能导致本地模型严重过拟合和性能下降（Sun 等人，2022年）。为探究这一现象背后的机制，我们展示了损失景观的结构（Li 等人，

[1]中国深圳清华大学 [2]中国北京京东探索研究院 [3]中国香港香港理工大学 [4]澳大利亚悉尼大学。通讯作者：沈立 <mathshenli@gmail.com>，王学谦<wang.xq@sz.tsinghua.edu.cn>。

论文集： 40th 国际机器学习会议，美国夏威夷檀香山。PMLR 202，2023。版权归作者所有，2023年。



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

图2. Fashion-MNIST（Xiao 等人，2017年）和 CIFAR-10（Krizhevsky 等人，2009年）上中心化联邦学习与去中心化联邦学习的损失景观对比。与 FedAvg（网格图）相比，DFedAvg（曲面图）的损失景观更陡峭，泛化能力更差。

2018年），并在与图Mcmahan 等人，2017年）相同的设置下，在 Fashion-MNIST（Xiao 等人，2017年）和 CIFAR-10（Krizhevsky 等人，2009年）上比较 FedAvg（）与去中心化 FedAvg（DFedAvg，Sun 等人（2022年））。如图2所示，DFL 的损失景观明显比 CFL 更陡峭。

动机。 大多数联邦学习算法都面临异构数据下本地模型过拟合的问题。针对中心化联邦学习，已有许多解决方案（Sahu 等人，2018年；Li 等人，2020c；Karimireddy 等人，2020年；Yang 等人，2021年；Acar 等人，2021年；Wang 等人，2022a）。在去中心化联邦学习中，由于本地模型不一致导致损失景观陡峭，这一问题进一步加剧（见图2）。因此，在相同设置下，去中心化方案的性能通常不如集中式方案（Sun 等人，2022年）。因此，一个重要问题是：我们能否设计去中心化联邦学习算法，以减轻本地模型之间的不一致性，并达到与其中心化对应方案相近的性能？

为回答这一问题，我们提出了两种去中心化联邦学习算法：DFedSAM 和 DFedSAM-MGS。具体而言，DFedSAM 通过在每个客户端中采用 SAM（Foret 等人，2021年）进行梯度扰动，克服本地过拟合问题，从而生成本地平坦模型。由于每个客户端都会聚合来自其邻居的平坦模型，因此可以生成潜在的平坦聚合模型，从而获得良好的泛化能力。为进一步提升 DFedSAM 的性能，DFedSAM-MGS 融合了多步 Gossip（MGS）（Ye 等人，2020年；Ye 和 Zhang，2021年；Li 等人，2020a），通过增加本地通信的 gossip 步骤数量，加速本地平坦模型的聚合。由于去中心化联邦学习可以粗略地视为具有足够数量 gossip 步骤的中心化联邦学习（第5.4节），该方法通过弥合中心化联邦学习与去中心化联邦学习之间的差距，实现了通信复杂度与学习性能之间更优的权衡。

从理论上讲，我们给出了算法在随机非凸设置下的收敛率。我们表明，当通信拓扑的连通性 λ 足够稀疏，或数据同质性 β 足够高时，界可能会更宽松。同时，随着 MGS 中共识/八卦步数 Q 增加，界趋于更紧，因为通信拓扑的影响可以得到缓解（第4节）。理论结果解释了为何在去中心化联邦学习中采用 SAM 和 MGS 能够确保在各种通信网络拓扑下取得更好的性能。通过实证研究，我们在相同数据分布（IID）和非独立同分布设置下，对 CIFAR-10 和 CIFAR-100 数据集进行了大量实验。实验结果证实，与集中式联邦学习基线相比，我们的算法能够取得有竞争力的性能，并优于现有的去中心化联邦学习基线（第5.2节）。

贡献。 我们的主要贡献总结如下：

- 我们提出了两种有效的去中心化联邦学习方案：DFedSAM 和 DFedSAM-MGS。DFedSAM 通过本地平坦模型降低本地模型的不一致性，而 DFedSAM-MGS 则借助 MGS加速进一步提升一致性，并在通信与泛化之间实现了更好的权衡。

- 我们分别为非凸情形下的 DFedSAM 和 DFedSAM-MGS 提出了改进的收敛速率 $\mathcal{O}(\frac{1}{\sqrt{KT}} + \frac{1}{T} + \frac{1}{K^{1/2}T^{3/2}(1-\lambda)^2})$ 和 $\mathcal{O}(\frac{1}{\sqrt{KT}} + \frac{1}{T} + \frac{\lambda^Q+1}{K^{1/2}T^{3/2}(1-\lambda^Q)^2})$，从理论上验证了我们方法的有效性。

- 我们进行了大量实验，以展示 DFedSAM 和 DFedSAM-MGS 的有效性。与中心化联邦学习和去中心化联邦学习基线相比，这两种方法均能取得有竞争力的性能。

2. 相关工作

去中心化联邦学习（DFL）。在 DFL 中，客户端仅在没有中央服务器的各种通信网络中与其邻居进行通信；与 CFL 相比，这具有通信优势，并能更好地保护数据隐私。Lalitha 等人（2018年；2019年）在完整的 DFL 框架中采用类贝叶斯方法，在客户端的模型参数空间上引入信念。Roy等人（2019年）提出了首个无服务器的点对点联邦学习方法 BrainTorrent，并将其应用于高度动态的点对点联邦学习环境中的医疗应用。Sun 等人（2022年）将多次本地迭代与随机梯度下降和量化方法相结合，以降低通信成本，并给出了各种凸情形下的收敛结果。Dai等人（2022年）开发了一种去中心化稀疏训练技术，以进一步降低通信成本和计算成本。我们的工作聚焦于 DFL（Lalitha 等人，2018年；2019年；Roy等人，2019年；Sun 等人，2022年；Dai等人，

2022年），而非去中心化训练（Lian 等人，2017年；Ye 等人，2020年；Li 等人，2020a；Chen 等人，2021年；Yuan 等人，2021年；Ye 和 Zhang，2021年；Warnat-Herresthal 等人，2021年；Koloskova 等人，2020年；Hashemi 等人，2022年；Zhang 等人，2022年）[1]。

尖锐度感知最小化（SAM）。SAM（Foret 等人，2021）是一种用于训练深度学习模型的有效优化器，它利用损失景观的平坦几何特性来提升模型泛化能力。近期，Andriushchenko 和 Flammarion（2022）研究了 SAM 的性质，并给出了 SAM 在非凸目标上的收敛结果。作为一种有效的优化器，SAM 及其变体已应用于各种机器学习（ML）任务（Zhao 等人，2022；Kwon 等人，2021；Du 等人，2021；Liu 等人，2022；Abbas 等人，2022；Shi 等人，2023；Zhong 等人，2022；Mi 等人，2022）。例如，Qu 等人（2022）和 Caldarola 等人（2022）采用 SAM 来提升泛化能力，从而缓解分布偏移问题，并在中心化联邦学习（CFL）中达到最先进的性能（Sun 等人，2023；Qu 等人，2022）。然而，据我们所知，几乎没有研究致力于探究 SAM 在去中心化联邦学习（DFL）场景中的实证表现和理论分析。

多步 Gossip（MGS）。FastMix（Ye 等人，2020年）研究了在网络拓扑中提高本地通信频率的优势，并实现了最优计算复杂度和近乎最优的通信复杂度。DeEPCA（Ye 和 Zhang，2021年）将 FastMix 集成到去中心化主成分分析算法中，以加速训练过程。DeLi-CoCo（Hashemi 等人，2022年）在每次迭代中执行多次压缩 gossip 步骤，从而在任意通信压缩条件下实现快速收敛。Network-DANE（Li 等人，2020a）采用多次 gossip 步骤，并将 DANE 推广到去中心化场景。一般而言，增加 gossip 步骤的数量可以使本地客户端达成更好的共识模型，从而提升中心化联邦学习的性能。然而，在去中心化联邦学习场景中，MGS 尚未被用于缓解模型不一致性。

与本文最相关的工作是 Sun 等人Sun 等人(2022年)提出的 DFedAvg 和带动量的 DFedAvg（DFedAvgM）。这两种方法利用随机梯度下降优化器进行多次本地迭代，显著提升了经典去中心化并行随机梯度下降方法 D-PSGD 的性能（Lian 等人，2017年）。然而，由于客户端之间存在严重的模型不一致性，DFL 的性能仍可能不佳。另一项相关工作是 FedSAM （ Qu 等人，2022 年），该方法将 SAM 融入 CFL，以提升本地模型的平坦性，并为 CFL 实现了新的最先进性能。在上述研究的基础上，我们将 SAM 优化器扩展到 DFL 设定，并同时为其提供了非凸设定下的收敛保证。此外，我们通过在 DFedSAM-MGS 中采用 MGS，弥合了 CFL 与 DFL 之间的差距，从而大幅缓解了 DFL 中的模型不一致性。

[1]在本文中，DFL 指采用多个本地迭代点进行本地训练，而去中心化学习/训练则侧重于单步本地训练。例如，D-PSGD（Lian 等人，2017年）是一种去中心化训练算法，在每个通信轮次中使用单步随机梯度下降来训练本地模型。

3. 方法

本节介绍去中心化联邦学习中的问题设定，并详细阐述所提出的算法。

3.1. 问题设定

在本文中，我们关注如何求解以下有限和随机非凸最小化问题：

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

其中， $\mathcal{D}_i$ 表示第 i个客户端的数据分布，该分布在不同客户端之间具有异质性； m 表示客户端数量； $F_i(\mathbf{x};\xi)$是与数据样本 ξ相关联的局部目标函数。公式（1）称为经验风险最小化（ERM），在机器学习中有许多应用。如图 1(b) 所示，客户端之间去中心化网络拓扑中的通信网络建模为无向连通图 $\mathcal{G} = (\mathcal{N}, \mathcal{V}, \boldsymbol{W})$，其中 $\mathcal{N} = \{1, 2, \ldots, m\}$ 表示客户端集合，$\mathcal{V} \subseteq \mathcal{N} \times \mathcal{N}$ 表示通信信道集合，每条信道连接两个不同的客户端。此外，在去中心化设置中不存在中央服务器，所有客户端仅通过通信信道 $\mathcal{V}$与其邻居通信。另外，我们假设公式（1）定义良好，并将 f^*记为 f:$f(x) \geq f(x^*) = f^*$ 对所有 $x \in \mathbb{R}^d$的最小值。

去中心化联邦学习中的挑战。由于缺少中央服务器，通信连接成为去中心化优化中的一个重要因素。此外，在经典联邦学习场景中，通信比计算更受关注（Mcmahan 等人，2017年；Li 等人，2020b；Kairouz 等人，2021年；Qu 等人，2022年），因此，由于 FedAvg 等经典联邦学习方法的通信开销较大（Mcmahan 等人，2017年），客户端默认采用多步本地迭代。因此，去中心化联邦学习中的主要技术难点总结如下：

- 各种通信拓扑。拓扑结构通过谱间隙 $1-\lambda \in (0, 1]$ $\mathbf{W}$ 的谱间隙)来衡量， λ 的值会随着连通性变得更加稀疏而增大。这会对模型训练（收敛率和泛化能力）产生显著的负面影响，尤其是在异构数据情况下，或面对通信网络连通性稀疏的问题时，例如环形

拓扑结构和网格拓扑，其中 $\lambda \approx 16\pi^2/(3m^2)$和 $\lambda = \mathcal{O}(1/(m对数_2(m)))$，其中 m 分别为客户端的数量（Sun 等人，2022年；Zhu 等人，2022年）。

• 多步局部迭代。经过多次局部迭代后，梯度估计可能会趋于有偏。这意味着，与单步本地迭代相比，相应的理论分析可能更加困难，经验效果也可能受到影响。

算法 1 DFed edS

输入： 客户端总数 m，通信轮数总计 T，每次梯度迭代的共识步骤数 Q，学习率 η，以及局部迭代总数迭代为 K。

输出： 最终一次通信后，$\mathbf{x}^T$ 共识模型所有客户端的。

初始化： 随机初始化每个模型$\mathbf{x}^0(i)$。
对于 $t=0$ 到 $T-1$，执行
  对于节点 i 并行执行
    对于 $k=0$ 至 $K-1$ 执行
      设置$\mathbf{y}^{t,0}(i) \leftarrow \mathbf{x}^t(i)$、$\mathbf{y}^{t,-1}(i) = \mathbf{y}^{t,0}(i)$
      采样一个本地数据批次 ξ_i 并计算本地梯度 $\mathbf{g}^{t,k}(i) = \nabla F_i(\mathbf{y}^{t,k}; \xi_i)$
      $\tilde{\mathbf{g}}^{t,k}(i) = \nabla F_i(\mathbf{y}^{t,k} + \delta(\mathbf{y}^{t,k}); \xi_i)$，其中 $\delta(\mathbf{y}^{t,k}) = \rho \mathbf{g}^{t,k}(i) / \left\|\mathbf{g}^{t,k}(i)\right\|_2$
      $\mathbf{y}^{t,k+1}(i) = \mathbf{y}^{t,k}(i) - \eta \tilde{\mathbf{g}}^{t,k}(i)$
    end
    $\mathbf{z}^t(i) \leftarrow \mathbf{y}^{t,K}(i)$
    从邻域集合接收邻居模型 $\mathbf{z}^t(l)$
    $\mathcal{S}_{k,t}$ 邻接矩阵为 $\boldsymbol{W}$。
    $\mathbf{x}^{t+1}(i) = \sum_{l \in \mathcal{N}(i)} w_{i,l} \mathbf{z}^t(l)$
    对于 $q=0$ 到 $Q-1$，执行
      $\mathbf{x}^{t,q+1}(i) = \sum_{l \in \mathcal{N}(i)} \boldsymbol{w}_{i,l} \mathbf{z}^{t,q}(l)$ $(\mathbf{z}^{t,0}(i) = \mathbf{z}^t(i))$
      $\mathbf{z}^{t,q+1}(i) = \mathbf{x}^{t,q+1}(i)$
    end
    $\mathbf{x}^{t+1}(i) = \mathbf{x}^{t,Q}(i)$
  end
结束

3.2. DFedSAM 和 DFedSAM-MG

不同于通过随机梯度下降（Bottou，2010年；Bottou 等人，2018年）搜索解，SAM（Foret 等人，2021年）旨在通过向模型添加小扰动，在平坦区域中寻找解，即 $x+\delta$ 从而获得更稳健的性能。如图2所示，与中心化方案相比，去中心化方案具有更尖锐的损失景观和更差的泛化能力。本文将SAM优化器引入DFL，以探索这一特性，并将所提出的DFedSAM的局部损失函数定义为：

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

其中，$\mathbf{y}^{t,k}(i) + \delta_i$ 是经过扰动的模型，ρ 是一个预先定义的常数，用于控制扰动半径，$\|\cdot\|_2$ 是 l_2-范数，在下文中简化为 $\|\cdot\|$。与集中式联邦学习方法类似，在DFedSAM中，客户端可以在进行通信之前通过多次局部迭代更新本地模型参数。具体而言，对于每个客户端 $i \in \{1,2,...,m\}$，在每次局部迭代$k \in \{0,1,...,K-1\}$和每个通信轮次 $t \in \{0,1,...,T-1\}$中，客户端 i的第 k次内部迭代执行如下操作：

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

其中，$\tilde{\mathbf{g}}^{t,k}(i) = \nabla F_i(\mathbf{y}^{t,k} + \delta(\mathbf{y}^{t,k}); \xi)$，$\delta(\mathbf{y}^{t,k}) = \rho \mathbf{g}^{t,k} / \left\|\mathbf{g}^{t,k}\right\|_2$，对于较小的 ρ，在 $\mathbf{y}^{t,k}$ 处进行一阶泰勒展开（Foret 等人，2021）。在每个客户端进行 K 次内层迭代后，参数更新为 $\mathbf{z}^t(i) \leftarrow \mathbf{y}^{t,K}(i)$，并在本地更新后发送给其邻居 $l \in \mathcal{N}(i)$。随后，每个客户端将自身参数与邻居（包括自身）的信息进行平均：

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

此外，我们采用多步 Gossip（MGS）技术（Ye 等人，2020年；Ye 和 Zhang，2021年；Hashemi 等人，2022年），以改善本地模型之间的一致性，该方法称为 DFedSAM-MGS，旨在确保去中心化联邦学习中的通信成本与泛化能力之间取得平衡。具体而言，第 q步（$q \in \{0,1,...,Q-1\}$）中 MGS 的生成可分为消息交换和本地 gossip 更新两个步骤，如下所示：

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

在 MGS 结束时，有 $\mathbf{x}^{t+1}(i) = \mathbf{x}^{t,Q}(i)$。我们提出的算法总结于算法1。显然，在 DFedSAM 中，通过多次本地迭代，可以在本地计算复杂度和通信开销之间进行权衡，而本地通信仅执行一次。相比之下，DFedSAM-MGS 会进行多次本地通信，并采用更大的 Q 以使所有本地客户端保持同步。因此，DFedSAM-MGS 可视为 DFL 与 CFL 之间的一种折中方案。

与现有最先进的去中心化联邦学习方法：DFedAvg 和 DFedAvgM（Sun 等人，2022年）相比，DFed-SAM 和 DFedSAM-MGS 的优势主要体现在三个方面：（i）在去中心化联邦学习中，通过在每个客户端寻找平坦模型，引入 SAM 来缓解由本地模型不一致导致的本地过拟合问题，同时也有助于使共识模型保持平坦；（ii）在 DFedSAM-MGS 中，基于 DFedSAM 使用 MGS 加速本地平坦模型的聚合，以提升本地模型之间的一致性

并适当平衡通信复杂度与学习性能；（iii）此外，我们还在 ρ 中提出相应理论，统一阐释梯度扰动Q 、MGS 中的本地通信次数 λ、网络拓扑 β 以及数据同质性对**第**4节中收敛率的影响。

4. 收敛性分析

本节针对一般的非凸联邦学习设定，分析 DFed-SAM 和 DFedSAM-MGS 的收敛性，详细证明见 **附录** D。下面介绍所需的记号和假设。

定义 4.1.（gossip/混合矩阵）。 [定义 1，（Sun 等人，2022）] 假设 gossip 矩阵 $\mathrm{W} = [w_{i,j}] \in [0,1]^{m\times m}$ 具有以下性质：(i)（图）若 $i \neq j$ 且 $(i,j) \notin \mathcal{V}$，则 $w_{i,j} = 0$；否则，$w_{i,j} > 0$；(ii)（对称性）$\mathrm{W} = \mathrm{W}^\top$；(iii)（零空间性质）$\mathrm{null}\{\mathrm{I} - \mathrm{W}\} = \mathrm{span}\{\mathbf{1}\}$；(iv)（谱性质）$\mathrm{I} \succeq \mathrm{W} \succ -\mathrm{I}$。根据这些性质，可以证明 W 的特征值满足 $1 = |\lambda_1(\mathrm{W})| > |\lambda_2(\mathrm{W})| \geq \cdots \geq |\lambda_m(\mathrm{W})|$。此外，$\lambda := \max\{|\lambda_2(\mathrm{W})|, |\lambda_m(\mathrm{W})|\}$，且 $1-\lambda \in (0,1]$ 表示 W 的谱间隙。

定义 4.2。（同质性参数）。 [定义 2，（Li 等人，2020a）] 对于任意 $i \in \{1,2,\ldots,m\}$ 以及参数$\mathbf{x} \in \mathbb{R}^d$，同质性参数 β 可定义为：

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

假设 4.3。（利普希茨光滑性）。函数 f_i可微，且 ∇f_i 是 L-利普希茨连续的，$\forall i \in \{1,2,\ldots,m\}$，即 $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|$，对所有$\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$均成立。

假设 4.4。（有界方差）。函数 f_i 的梯度具有 σ_l-有界方差：

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

且全局方差也有界，即$\frac{1}{m}\sum_{i=1}^m \|\nabla f_i(\mathbf{x}) - \nabla f(\overline{\mathbf{x}})\|^2 \leq \sigma_g^2$对所有$\mathbf{x} \in \mathbb{R}^d$均成立。不难验证，$\sigma_g$ 小于同质性参数 β，即 $\sigma_g^2 \leq \beta^2$。

需要注意的是，上述假设较为温和，且通常用于刻画联邦学习的收敛率（Sun 等人，2022；Ghadimi 和 Lan，2013；Yang 等人，2021；Bottou 等人，2018；Yu 等人，2019；Reddi 等人，2021）。与 D-PSGD（Lian 等人，2017）等经典去中心化并行随机梯度下降方法相比，难点在于，经过多次本地迭代后，$\mathbf{z}^t(i) - \mathbf{x}^t(i)$ 可能不再是$\nabla f_i(\mathbf{x}^t(i))$的无偏梯度估计，因此合并多次本地迭代并非易事。此外，去中心化联邦学习中的各种通信网络拓扑与中心化联邦学习中的 SAM 截然不同（Qu 等人，2022年）。下文中，我们采用 $\overline{\mathbf{x}^t} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}^t(i)$ 所有客户端的平均参数，作为问题（1）的近似解。

定理 4.5。 假设假设*4.3* 和 *4.4* 成立，且参数 $\{\mathbf{x}^t(i)\}_{t\geq 0}$ 由算法*1*生成。同时，假设每个客户端中 *SAM* 的学习率满足 $0 < \eta \leq \frac{1}{10KL}$。令 $\overline{\mathbf{x}^t} = \frac{1}{m}\sum_{i=1}^m \mathbf{x}^t(i)$ 并将 $\mathbf{\Phi}(\lambda, m, Q)$ 记为与谱间隙数量、客户端数量和多次 *gossip* 步骤这三个参数相关的度量：

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

那么，对于求解问题（*1*）的 *DFedSAM* 或 *DFedSAM-MGS*，其梯度估计为：

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

其中，T 表示通信轮次，常数定义如下：

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

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

根据定理 4.5，我们给出 DFedSAM 和 DFedSAM-MGS 的以下收敛速率。

推论 4.6。 设本地自适应学习率满足$\eta = \mathcal{O}(1/L\sqrt{KT})$。在满足与 定理 *4.6*相似的假设下，并令扰动参数$\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$。*DFedSAM* 的收敛率满足：

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

备注 4.7。 由于使用 SAM 时扰动半径 ρ 的影响，DFedSAM 能够获得比最先进水平（SOTA）的界更好的界，例如$\mathcal{O}\Big(\frac{1}{\sqrt{T}} + \frac{\sigma_g}{\sqrt{T}} + \frac{\sigma_g + B^2}{1-\lambda}\Big)$在（Sun 等人，2022年）中给出的界，其中 B是梯度的上界。其原因是，不同通信拓扑 λ 对收敛率的不同影响可以通过扰动半径 ρ的作用得到缓解，即 $\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$ (我们设 $\rho = 0.01$

表 *1.* 所有算法在两个数据集的 IID 和非独立同分布设置下的性能（%）。

Task	算法	Dirichlet 0.3			Dirichlet 0.6			IID		
		训练	验证	泛化误差	训练	验证	泛化误差	训练	验证	泛化误差
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

当 $T = 1000$ 时）。请注意，随着 λ 减小，界会变得更紧，这主要取决于 $\frac{\beta^2+\sigma_l^2}{K^{1/2}T^{3/2}}\frac{}{(1-\lambda)^2}$，当 $\lambda \geq 1-\frac{1}{\sqrt{T}}$ 时如此；而随着 β 增大，界可能会变差。此外，当光滑性较差时，即 L 较大时，与其他项相比，额外项 $\mathcal{O}(\frac{L^2}{K^{1/2}T^{3/2}})$ 可以忽略不计。该项源于通过 SAM 局部优化器进行光滑性处理时所增加的随机梯度下降步骤。

推论 4.8。假设 $Q > 1$ 和 T 足够大，并且 $\eta = \mathcal{O}(1/L\sqrt{KT})$。在定理 *4.6* 所要求的类似假设以及扰动幅度 $\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$ 下，*DFedSAM-MGS* 的收敛率满足：

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

备注 4.9. 随着 Q 增大且客户端数量 m 足够多，网络拓扑（$1-\lambda$）的影响可以得到缓解。同时，$\frac{\lambda^Q+1}{(1-\lambda)^2 m^{2(Q-1)}}$ 关于 $\mathbf{\Phi}(\lambda,\ m,\ Q)$ 的项可以忽略，而项 $\frac{\lambda^Q+1}{(1-\lambda^Q)^2}$ 接近于 1。这意味着，采用所提出的 Q 步 gossip 过程可以提高客户端之间的模型一致性，因此，具有各种通信拓扑的 DFL 大致可以视为 CFL。由此，梯度方差 σ_l^2 和 β^2 的负面影响可以得到缓解，尤其是在 λ 接近于 1 的稀疏网络拓扑中。在实践中，合适的步数 $Q > 1$ 可能实现 DFL 中的通信-精度权衡。

注 4.10。注意 $\mathbf{\Phi}(\lambda,\ m,\ Q) = \frac{2(\lambda+1)}{(1-\lambda)^2} \leq \frac{4}{(1-\lambda)^2}$ 对于 $Q = 1$，结果与 **推论** 4.6 相同。

5. 实验

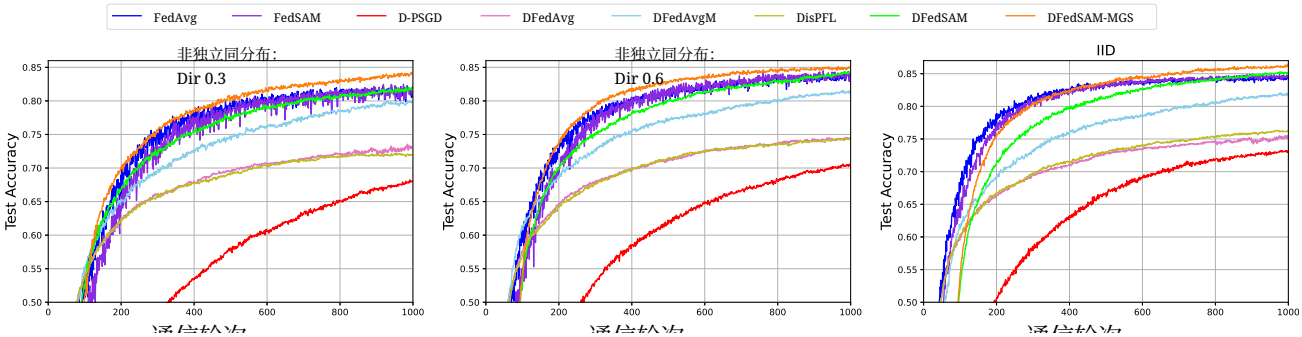
本节将我们的算法与中心化联邦学习和去中心化联邦学习场景下的六种基线方法进行比较，以评估其有效性。此外，我们还开展了若干实验，以验证通信网络拓扑的影响，相关内容见 **第**4节。此外，还开展了若干消融实验。

5.1. 实验设置

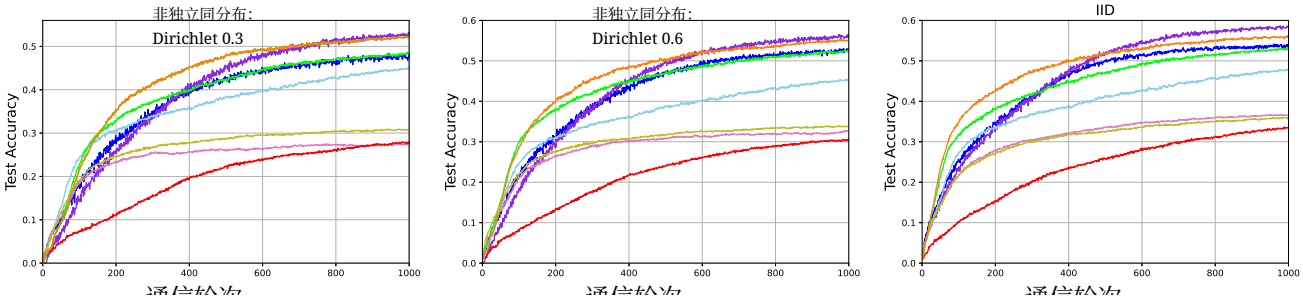
数据集与数据划分。在 IID 和非 IID 设置下，使用 CIFAR-10 和 CIFAR-100 数据集对所提出的 DFedSAM 和 DFedSAM-MGS 的有效性进行评估（Krizhevsky 等人，2009年）。具体而言，采用狄利克雷划分（Hsu 等人，2019年）和病态划分来模拟联邦客户端之间的非 IID，其中前者通过从狄利克雷分布 Dir（α）中采样标签比例，并以参数 $\alpha = 0.3$ 和 $\alpha = 0.6$对总数据集进行划分，从而划分每个客户端的本地数据。由于篇幅有限，病态划分的内容列于**附录**C。

基线方法。对比基线涵盖了集中化联邦学习和去中心化联邦学习设置下的多种最先进方法。具体而言，集中式基线包括 FedAvg（Mcmahan 等人，2017年）和 FedSAM（Qu 等人，2022年）。在去中心化设置下，采用 D-PSGD（Lian 等人，2017年）、DFedAvg 和 DFedAvgM（Sun 等人，2022年），以及 DisPFL（Dai 等人，2022年）进行比较。

实现细节。客户端总数设为 100，其中 10% 的客户端参与通信。具体而言，对于去中心化方法，所有客户端都执行本地迭代步；对于中心化方法，仅参与通信的客户端可以执行本地更新。我们将本地学习率初始设为



(a) CIFAR-10



(b) CIFAR-100

图 *3.* 所有 在独立同分布和非独立同分布设置下，中心化联邦学习与去中心化联邦学习的所有基线方法在 (a) CIFAR-10 和 (b) CIFAR-100 上的测试准确率 s.

0。1，所有实验中每个通信轮次的衰减率为 0.998。对于 CIFAR-10 和 CIFAR-100 数据集，每个客户端分别采用 VGG-11（He 等人，2016）和 ResNet-18（Simonyan 和 Zisserman，2014）作为骨干网络。在用于比较所有基线方法和研究拓扑感知性能的实验中，通信轮次为1000 。此外，所有消融实验均在 CIFAR-10 数据集上进行，通信轮次设为500。

通信配置。 为公平比较去中心化设置与集中式设置，我们为去中心化方法采用动态时变连接拓扑，以确保每轮的连接数不超过中央服务器的连接数。请注意，可以控制与邻居通信的客户端数量，使通信量与集中式方法保持一致。按照先前的研究，通信复杂度以本地通信次数衡量。受篇幅限制，实验的更多细节见 **附录**B 。

5.2. 性能评估

与基线方法的性能比较。我们在两种设置下，针对 CIFAR-10 和 CIFAR-100 数据集，将 DFedSAM 和 DFedSAM-MGS（$Q = 4$）与 $\rho = 0.01$ 进行评估，并与来自中心化联邦学习和去中心化联邦学习的所有基线方法进行比较。根据结果（见表1 和图3），可以清楚地看到，我们提出的算法在两个数据集上均优于其他去中心化方法；DFedSAM-MGS 在 CIFAR-10 和 CIFAR-100 上的表现分别优于和接近最先进水平的中心化基线 FedSAM。具体而言，表1 列出了训练准确率和测试准确率，以展示泛化性能。可以看到，在 CIFAR-10 上，相同通信轮次下，我们的性能提升比所有其他基线方法都更加显著。例如，在 IID 设置下，CIFAR-10 上训练准确率与测试准确率之差（即泛化误差）分别为：DFedSAM 的14.14% 、DFedSAM-MGS 的13.22% 、FedAvg 的15.29% 和 FedSAM 的15%。这意味着我们的算法实现了与中心化基线相当的泛化水平。总体而言，DFedSAM 的性能上限与 FedSAM 一致，因为二者针对同一优化问题采用了不同的优化器。

非独立同分布程度的影响（β）。从表1中可以看出，我们的算法在不同参与程度下均具有鲁棒性。将本地客户端的异构数据分布设置为不同参与程度，包括独立同分布、Dirichlet 0.6 和 Dirichlet 0.3，这使得全局/共识模型的训练更加困难。在 CIFAR-10 上，随着非独立同分布程度增加，DFedSAM-MGS 的泛化能力优于 FedSAM，因为 DFedSAM-MGS 的泛化误差 {15.27%14。51% 13。22%} 低于 Fed-

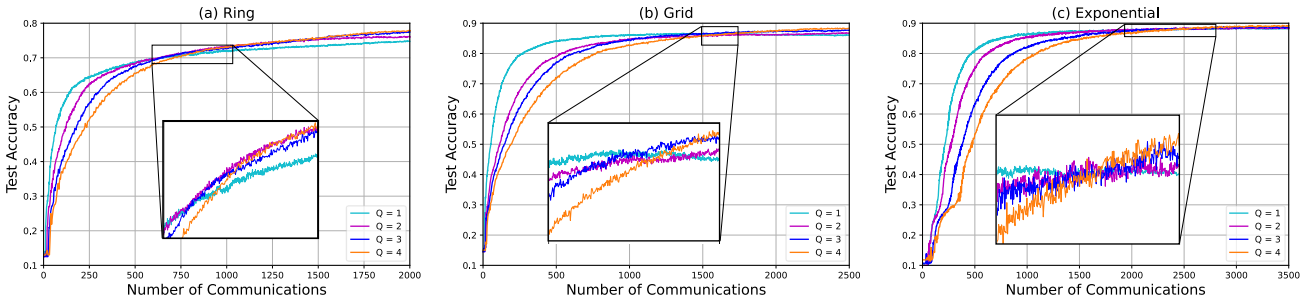


图 *4*。 不同 Q取值下，测试准确率随本地通信次数的变化。

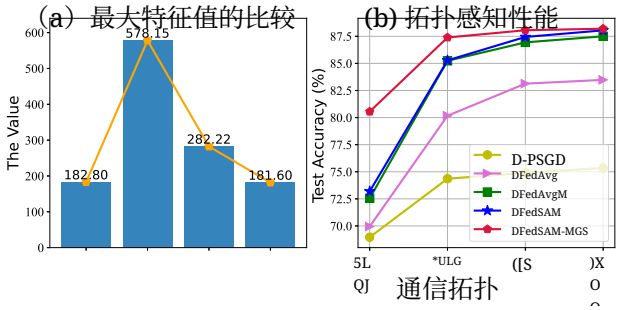


图 *5*。 图 5（a）中海森矩阵最大特征值（Yao 等人，2020 年），这是衡量损失景观平坦度的一个已知指标；以及图 5（b）中去中心化联邦学习方法的拓扑感知性能。

SAM {17.26% 14。85%，15%}。同样，DFedSAM {17.37 % 15。06% 14。10%} 的泛化误差低于 FedAvg {17.60% 15。82% 15。27%}。这些观察结果证实，我们的算法在不同异质程度下比基线方法更具鲁棒性。

损失景观平坦度的测量。 为了评估与 FedAvg 和 DFedAvg（图 2（a） 和 2（b））相比，我们的算法所产生的损失景观平坦度，我们在具有狄利克雷分布（$\alpha = 0.6$）的划分的CIFAR-10数据集上使用 VGG-11模型进行实验，待所有模型收敛后，给出海森矩阵的最大特征值（Yao 等人，2020年），如图5所示。请注意，最大特征值越小，损失景观越平坦。显然，所得模型如预期一样位于更平坦的极小值处，并且就最大特征值而言，DFedSAM-MGS接近 FedAvg。

5.3. 拓扑感知性能

我们验证了不同通信拓扑结构和 gossip 平均步骤对 DFedSAM 和 DFedSAM-MGS 的影响。不同于 **第** 5.2 节中对 CFL 和 DFL 的比较，本节只需验证 DFL 方法的关键性质。因此，通信类型设置为“完全连接”，使每个客户端都能在同一通信轮次中与其邻居进行通信。

表2. 在 CIFAR-10 上，不同网络拓扑下去中心化算法的测试准确率（%）。

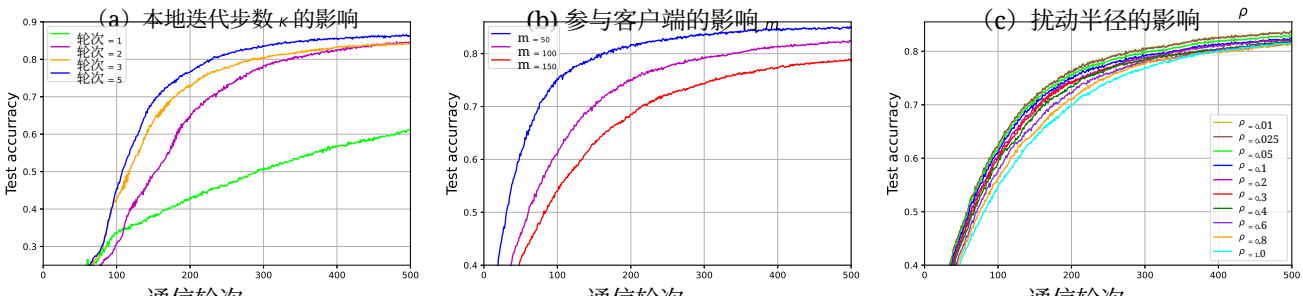
算法	Ring	Grid	Exp	Full
D-PSGD	68.96	74.36	74.90	75.35
DFedAvg	69.95	80.17	83.13	83.48
DFedAvgM	72.55	85.24	86.94	87.50
DFedSAM	73.19 ↑	85.28↑	87.44 ↑	88.05 ↑
DFedSAM-MGS	80.55 ↑	87.39↑	88.06 ↑	88.20 ↑

稀疏连通性的程度 λ 为环形 $>$ 网格 $>$ 指数型（缩写为“Exp”） $>$ 和全连接（缩写为“Full”）DFL。根据表 2和图 5(b)，我们的算法在各种通信网络中均优于所有去中心化基线方法，这与我们的理论发现一致。具体而言，与 DFedAvgM 相比，DFedSAM 和 DFedSAM-MGS 在环形拓扑中的性能分别显著提升了 0.64% 和 8.0%。同时，DFedSAM-MGS 在各种拓扑结构中的性能始终优于其他方法。这一观察结果证实，多次 gossip 步骤可以在更小的 $Q = 4$ 下减轻网络拓扑的影响。因此，我们的算法能够在各种通信拓扑结构下实现更好的泛化和模型一致性。

5.4. 消融研究

我们通过 $Q = 1$ 验证 DFedSAM 中各个组件和超参数的影响。除研究 Q 在三种拓扑结构下的影响外，所有消融实验均采用“指数型”拓扑结构，通信类型为“完全连接”，与 **第** 5.3 节相同。

共识/八卦步数 Q。在图4中，我们研究了三种网络拓扑下学习性能与通信复杂度之间的平衡。总的来说，较大的Q 值可以更好地解决局部一致性问题，但也可能带来额外的通信成本。在去中心化训练设置中，更容易获得最优的 Q。因此，在实验中，我们将 Q 视为超参数，并研究其在不同



*Fi*图 6. 超参数的影响：本地迭代步数 K、参与客户端数量 m、扰动半径 ρ.

值下 $1 \leq Q \leq 4$图4 （a）、（b）和（c）展示了不同通信拓扑下的情况，例如环形、网格和指数型拓扑。随着本地通信次数增加，模型性能也有所提升，但通信复杂度也随之增加。显然，不同拓扑下的平衡点各不相同，但趋势一致。此外，在给定通信复杂度的情况下，较大的 Q 可以带来更好的性能。因此，为了取得更好的平衡，我们在 DFedSAM-MGS 中选择了 $Q = 4$，通信轮次为 1000。

本地迭代步数 K。较大的本地迭代步数 K如先前去中心化联邦学习研究（Sun 等人，2022年）在理论保证下所示，有助于收敛。为了研究采用较大的本地迭代步数 T 对 K加速效果的影响，我们固定总批量大小并改变本地训练轮数。如图 6 (a) 所示，随着 **第** 4.5节中理论结果所示，更大的 K 取值能够加快算法收敛。

参与客户端数量 m。在图 6 (b) 中，我们比较了在相同超参数下不同客户端参与数量 $m = \{50, 100, 150\}$ 时的性能。与 $m = 150$相比， m 取值为 50 或 100 时，随着本地数据量增加，能够实现更好的收敛效果和测试准确率。我们发现，较小的客户端规模往往能带来更好的训练性能，这是因为每个客户端拥有更多样本，而且训练性能与客户端规模之间近似呈线性关系。

扰动半径 ρ。扰动半径的影响 ρ 源于通信轮次增加时，所添加的扰动会不断累积。 T 这需要在测试准确率与泛化之间进行权衡。为了为我们的算法选择合适的取值，我们针对集合中的多个扰动半径取值进行了实验 $\{0.01, 0.025, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0\}$，如图6（c）所示。取值为 $\rho = 0.01$时，我们在收敛与性能之间取得了令人满意的平衡。同时，$\rho = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ 可以使收敛实现线性加速（参见**第**4.5节）。

SAM 和 MGS 的有效性。为了验证其有效性，我们可以比较表格中的这些算法，包括 DFedAvg、DFedSAM 和 FedSAM-MGS 1。采用 SAM 优化器后，与 DFedAvg 相比，DFedSAM 能够提升性能并实现更好的泛化。与 FedSAM 相比，DFedSAM-MGS 能进一步提升性能，因为 MGS 还能使各客户端的模型保持一致，并加快收敛速率。

6. 结论与未来工作

本文从模型泛化的角度，聚焦于去中心化联邦学习中由异构数据和网络拓扑结构导致的模型不一致这一挑战。我们提出了两种去中心化联邦学习算法：DFedSAM 和 DFedSAM-MGS，它们能够提升客户端之间的模型一致性。DFedSAM 在每个客户端上采用 SAM 来生成平坦模型，进而通过生成共识/全局平坦模型来改善泛化能力。DFedSAM-MGS 则在 DFedSAM 的基础上进一步提升模型一致性：它通过加速本地平坦模型的聚合，在学习性能与通信复杂度之间实现更好的权衡。在理论研究方面，我们统一分析了 SAM 中的梯度扰动、MGS 中的本地通信、网络拓扑结构以及数据同质性对去中心化联邦学习收敛率的影响。此外，实验结果也验证了我们方法的优越性。未来，我们计划深入理解 SAM 和 MGS 的作用，并提升去中心化联邦学习的泛化能力。

致谢

本研究得到科技创新2030——“脑科学与类脑研究”重大项目（编号：2021ZD0201405）、国家重点研发计划（2022YFB4701400/4701402）以及国家自然科学基金（编号：U21B6002、U1813216、52265002）的资助。

参考文献

Abbas, M., Xiao, Q., Chen, L., Chen, P., and Chen, T. Sharp-maml：感知锐度的模型无关元学习。发表于国际机器学习会议，*ICML*，第10–32页，2022年。

Acar, D. A. E.、Zhao, Y.、Navarro, R. M.、Mattina, M.、Whatmough, P. N. 和 Saligrama, V. 基于动态正则化的联邦学习。载于 第九届国际学习表征会议，*ICLR*，2021年。

Andriushchenko, M. 和 Flammarion, N. 迈向锐度感知最小化。载于 国际机器学习会议，*ICML*，《机器学习研究会议录》，第639–668页。PMLR，2022年。

Beltr´an, E. T. M.、P´erez, M. Q.、S´anchez, P. M. S.、Bernal, S. L.、Bovet, G.、P´erez, M. G.、P´erez, G. M. 和 Celdr´an, A. H. 去中心化联邦学习：基础、最先进的技术、框架、趋势与挑战。*arXiv* 预印本 *arXiv:2211.08413*，2022年。

Bottou, L. 使用随机梯度下降进行大规模机器学习。见 *COMPSTAT'2010*论文集，第 177–186 页。施普林格，2010年。

Bottou, L., Curtis, F. E., and Nocedal, J. 大规模机器学习的优化方法。*SIAM*评论，60(2):223–311，2018年。

Caldarola, D., Caputo, B., and Ciccone, M. 通过寻求平坦极小值改进联邦学习中的泛化能力。*CoRR*，abs/2203.11834，2022年。

Chen, C., Zhang, J., Shen, L., Zhao, P., and Luo, Z. 面向非凸非光滑分布式优化的高通信效率原始-对偶算法。见国际人工智能与统计会议，第 1594–1602 页。PMLR，2021年。

Chen, H.-Y. and Chao, W.-L. 关于如何弥合图像分类中通用联邦学习与个性化联邦学习之间的差距。*arXiv*预印本 *arXiv:2107.00778*，2021年。

Dai, R., Shen, L., He, F., Tian, X. 和 Tao, D. DisPFL：通过去中心化稀疏训练实现高通信效率的个性化联邦学习。载于 国际机器学习会议，*ICML*，《机器学习研究论文集》，第4587–4604页。PMLR，2022年。

Deng, Y., Kamani, M. M. 和 Mahdavi, M. 自适应个性化联邦学习。*arXiv* 预印本 *arXiv:2003.13461*，2020年。

Du, J., Yan, H., Feng, J., Zhou, J. T., Zhen, L., Goh, R. S. M. 和 Tan, V. 提出高效的锐度感知最小化方法，以改善神经网络的训练。发表于国际学习表征会议，2021年。

Foret, P., Kleiner, A., Mobahi, H，等d Neyshabur, B 。锐度感知最小化：高效改进泛化能力。载于 国际学习表征会议，2021年。Ghadimi, S. 和 Lan, G. 非凸随机规划的随机一阶与零阶方法。《*SIAM* 优化杂志》，23(4):2341–2368，2013年。

Hashemi, A., Acharya, A., Das, R., Vikalo, H., Sanghavi, S. 和 Dhillon, I. 通信受限的去中心化联邦学习中多次 gossip 步骤的益处。*IEEE*并行与分布式系统汇刊，*TPDS*，第2727–2739页，2022年。

He, K., Zhang, X., Ren, S. 和 Sun, J. 用于图像识别的深度残差学习。载于*IEEE* 计算机视觉与模式识别会议论文集，第 770–778 页，2016年。

Hsu, T.-M. H., Qi, H. 和 Brown, M. 衡量非独立同分布数据对联邦视觉分类的影响。*arXiv* 预印本 *arXiv:1909.06335*，2019年。

Huang, Y., Chu, L., Zhou, Z., Wang, L., Liu, J., Pei, J. 和 Zhang, Y. 非独立同分布数据上的个性化跨机构联邦学习。载于*AAAI* 人工智能会议论文集，第 35 卷，第 7865–7873 页，2021年。

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R. 等。联邦学习的进展与开放问题。《机器学习基础与趋势》，第 1–210 页，2021年。

Kang, J., Ye, D., Nie, J., Xiao, J., Deng, X., Wang, S., Xiong, Z., Yu, R., 和 Niyato, D. 面向工业元宇宙的基于区块链的联邦学习：具有最优信息新鲜度的激励机制。载于*2022*年 *IEEE* 国际区块链会议（*Blockchain*），第 71–78 页。IEEE，2022年。

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., 和 Suresh, A. T. SCAFFOLD：用于联邦学习的随机控制平均。在国际机器学习会议，第 5132–5143 页。PMLR，2020年。

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., 和 Stich, S. 关于具有变化拓扑和本地更新的去中心化随机梯度下降的统一理论。载于国际机器学习会议，第 5381–5393 页。PMLR，2020年。

Krizhevsky、A.、Hinton、G. 等。学习多层特征：从小型图像中学习。2009年。

Kwon, J., Kim, J., Park, H. 和 Choi, I. K. Asam：用于深度神经网络尺度不变学习的自适应尖锐度感知最小化。载于国际机器学习会议，第 5905–5914 页。PMLR，2021年。

Lalitha, A., Shekhar, S., Javidi, T. 和 Koushanfar, F. 完全去中心化联邦学习。载于第三届贝叶斯深度学习研讨会（*NeurIPS*），2018年。

Lalitha, A., Kilinc, O. C., Javidi, T., and Koushanfar, F. 图上的点对点联邦学习。*arXiv*预印本 *arXiv:1901.11173*，2019年。

Li, B., Cen, S., Chen, Y., and Chi, Y. 采用梯度跟踪和方差缩减的网络通信高效分布式优化。机器学习研究期刊，*JMLR*，第180:1–180:51页，2020a。

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. 神经网络损失景观的可视化。神经信息处理系统进展，第31卷，2018年。

Li, J., Shao, Y., Wei, K., Ding, M., Ma, C., Shi, L., Han, Z., and Poor, H. V. 区块链辅助的去中心化联邦学习（BLADE-FL）：性能分析与资源分配。*IEEE*并行与分布式系统汇刊，33(10):2401–2415，2022年。doi: 10. 1109/TPDS.2021.3138848。

Li, T., Sahu, A. K., Talwalkar, A., 和 Smith, V. 联邦学习：挑战、方法与未来方向。*IEEE*信号处理杂志，第 50–60 页，2020b。

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., 和 Smith, V. 异构网络中的联邦优化。机器学习与系统会议录，第 429–450 页，2020c。

Li, T., Hu, S., Beirami, A., 和 Smith, V. Ditto：通过个性化实现公平且稳健的联邦学习。发表于国际机器学习会议，第 6357–6368 页。PMLR，2021年。

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., 和 Liu, J. 去中心化算法能否胜过集中式算法？去中心化并行随机梯度下降的一个案例研究。发表于神经信息处理系统进展，第 5330–5340 页，2017年。

Liu, Y., Mai, S., Chen, X., Hsieh, C.-J., 和 You, Y. 迈向高效且可扩展的锐度感知最小化。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 12360–12370 页，2022年。

Mcmahan, H. B., Moore, E., Ramage, D., Hampson, S., 和 Arcas, B. A. Y. 基于去中心化数据的深度网络通信高效学习。第 1273–1282 页，2017年。

Mi, P., Shen, L., Ren, T., Zhou, Y., Sun, X., Ji, R., and Tao, D. 增强锐度感知最小化：一种稀疏化扰动方法。*arXiv* 预印本 *arXiv:2210.05177*，2022年。

Nedic, A. 和 Ozdaglar, A. 多智能体优化的分布式次梯度方法。*IEEE*自动控制汇刊，54(1):48–61，2009年。Nguyen, T., Dakka, M., Diakiw, S., VerMilyea, M., Perugini, M., Hall, J., 和 Perugini, D. 一种新颖的去中心化联邦学习方法，可利用全球分布的低质量受保护的私人医疗数据进行训练。《科学报告》，12(1):8888，2022年。Qu, Z., Li, X., Duan, R., Liu, Y., Tang, B., 和 Lu, Z. 通过锐度感知最小化实现广义联邦学习。载于国际机器学习会议，*ICML*，第18250–18280页，2022年。Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Kone˘cn´y, J., Kumar, S., 和 McMahan, H. B. 自适应联邦优化。载于国际学习表征会议，2021年。网址：https://openreview.net/forum?id=LkFG3lB13U5。

Roy, A. G., Siddiqui, S., P¨olsterl, S., Navab, N., 和 Wachinger, C. BrainTorrent：一种用于去中心化联邦学习的点对点环境。*arXiv*预印本 *arXiv:1905.06731*，2019年。

Sahu, A. K., Li, T., Sanjabi, M., Zaheer, M., Talwalkar, A., 和 Smith, V. 关于异构网络中联邦优化的收敛。*arXiv*预印本 *arXiv:1812.06127*，第 3 页，2018年。

Shi, Y., Liu, Y., Wei, K., Shen, L., Wang, X., and Tao, D. 使差分隐私联邦学习中的损失景观更平坦。*arXiv*预印本 *arXiv:2303.11242*,2023。

Simonyan, K. and Zisserman, A. 用于大规模图像识别的超深卷积网络。*arXiv* 预印本 *arXiv:1409.1556*, 2014。

Sun, T., Li, D., and Wang, B. 去中心化联邦平均。*IEEE*模式分析与机器智能汇刊,2022年。

Sun, Y., Shen, L., Huang, T., Ding, L., and Tao, D. Fedspeed：更大的本地间隔、更少的通信轮次和更高的泛化准确率。*arXiv*预印本 *arXiv:2302.10429*, 2023。

Wang, H., Marella, S., and Anderson, J. Fedadmm：一种允许部分参与的联邦原始-对偶算法。*arXiv* 预印本 *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., 等. 联邦优化指南。*arXiv* 预印本 *arXiv:2107.06917*, 2021年。

Wang, L., Xu, Y., Xu, H., Chen, M., 和Huang, L. 加速异构边缘计算中的去中心化联邦学习。*IEEE* 移动计算汇刊，2022年b。

Wang, Y., Su, Z., Zhang, N., 和 Benslimane, A. 空中学习：面向无人机辅助群智感知的安全联邦学习。*IEEE*网络科学与工程汇刊，8(2):1055–1069，2020年。

Warnat-Herresthal, S., Schultze, H., Shastry, K. L., Manamohan, S., Mukherjee, S., Garg, V., Sarveswara, R., H¨andler, K., Pickkers, P., Aziz, N. A., 等。面向去中心化机密临床机器学习的群体学习。《自然》，第265–270页，2021年。

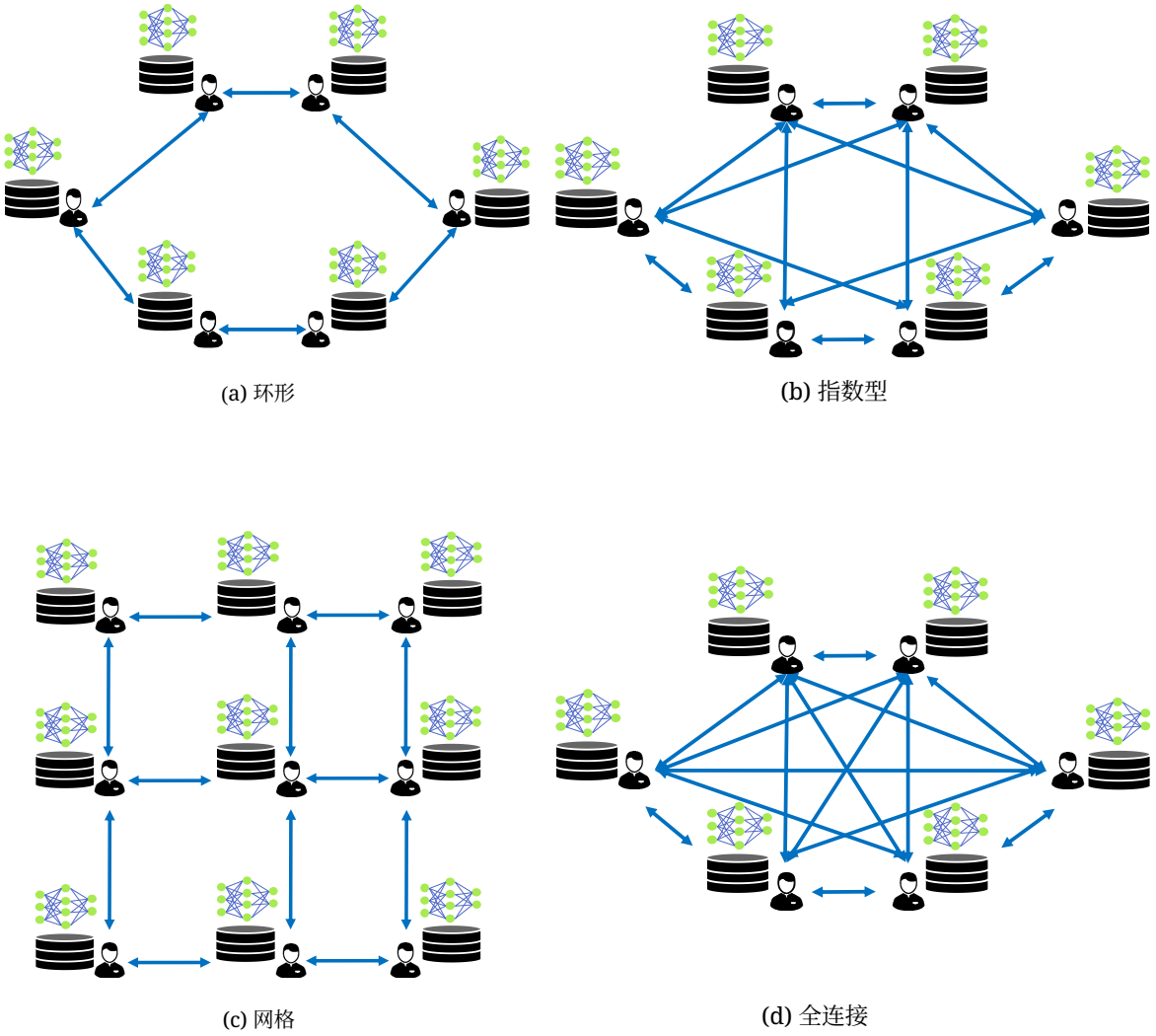
Xiao, H., Rasul, K., 和 Vollgraf, R. Fashion-MNIST：一种用于评测机器学习算法的新型图像数据集。*arXiv*预印本 *arXiv:1708.07747*，2017年。

Yang, H., Fang, M., 和 Liu, J. 在非独立同分布联邦学习中通过部分工作节点参与实现线性加速。载于国际学习表征会议，2021年。Yao, Z., Gholami, A., Keutzer, K., 和 Mahoney, M. W. Py-hessian：从海森矩阵的视角看神经网络。载于*2020 IEEE*国际大数据会议（*Big Data*），第581–590页。IEEE，2020年。Ye, H. 和 Zhang, T. Deepca：具有线性收敛率的去中心化精确主成分分析。《style id='5'>机器学习研究杂志，22(238)：1–27，2021年。Ye, H., Zhou, Z., Luo, L., 和 Zhang, T. 去中心化加速近端梯度下降。神经信息处理系统进展，33：18308–18317，2020年。Yu, H., Yang, S., 和 Zhu, S. 并行重启随机梯度下降：实现更快收敛并减少通信——揭示模型平均为何适用于深度学习。载于*AAAI*人工智能会议论文集，第5693–5700页，2019年。Yu, Z., Hu, J., Min, G., Xu, H., 和 Mills, J. 基于点对点联邦学习的车联网主动内容缓存。载于*2020 IEEE*第*26*届并行与分布式系统国际会议（*ICPADS*），第601–608页。IEEE，2020年。Yuan, Y., Chen, R., Sun, C., Wang, M., Hua, F., Yi, X., Yang, T., 和 Liu, J. DeFed：一种原理严谨的去中心化隐私保护联邦学习算法。*arXiv*预印本 *arXiv:2107.07171*，2021年。Zhang, M., Sapra, K., Fidler, S., Yeung, S., 和 Alvarez, J. M. 采用一阶模型优化的个性化联邦学习。*arXiv*预印本 *arXiv:2012.08565*，2020年。

Zhang, X., Fang, M., Liu, Z., Yang, H., Liu, J., and Zhu, Z. Net-fleet：利用异构数据实现完全去中心化联邦学习的线性收敛加速。*arXiv*预印本 *arXiv:2208.08490*，2022年。Zhao, Y., Zhang, H., and Hu, X. 通过惩罚梯度范数，高效提升深度学习的泛化能力。收录于国际机器学习会议，*ICML*，第26982–26992 页。PMLR，2022年。Zhong, Q., Ding, L., Shen, L., Mi, P., Liu, J., Du, B., and Tao, D. 利用 Fisher 掩码改进锐度感知最小化，以提升语言模型的泛化能力。*arXiv*预印本 *arXiv:2210.05497*，2022年。Zhu, T., He, F., Zhang, L., Niu, Z., Song, M., and Tao, D. 去中心化随机梯度下降的拓扑感知泛化。收录于国际机器学习会议，*ICML*，第 27479–27503 页。PMLR，2022年。

《改进去中心化联邦学习的模型一致性》补充材料

A. 通信网络拓扑



(a) 环形 (b) 指数型

(c) 网格 (d) 全连接

图 7. 去中心化设置中各种通信网络拓扑的概览。

B. 算法实现的更多细节

B.1. 数据集和骨干网络。

CIFAR-10 和 CIFAR-100（Krizhevsky 等人，2009年）是从8000万张图像数据集中标注而成的子集。两者共享相同的6万张输入图像。CIFAR-100 的标注更精细，包含 100 个唯一标签；相比之下，CIFAR-10 包含 10 个唯一标签。CIFAR-10 使用 VGG-11 作为骨干网络，CIFAR-100 则选用 ResNet；由于批归一化层会带来不利影响，因此将其替换为组归一化层。

B.2. 基线方法的更多详情。

FedAvg 是经典的联邦学习方法，通过常规加权平均，在中央服务器的支持下并行训练全局模型。FedSAM 使用 SAM 作为局部优化器，以提升模型泛化性能。对于去中心化方案，D-PSGD 是一种经典的去中心化并行随机梯度下降方法，用于达成模型共识[2]，DFedAvg 是去中心化版本的 FedAvg，而 DFedAvgM 则基于 DFedAvg 使用带动量的随机梯度下降在各个客户端上训练模型，并在每次通信前执行多个本地训练步骤。此外，DisPFL 是一种采用去中心化通信协议的新型个性化联邦学习框架，它使用去中心化稀疏训练技术。因此，为了公平比较，我们报告 DisPFL 的全局准确率。

B.3. 超参数。

客户端总数设为 100；在去中心化设置中，每个客户端的连接数最多限制为 10 个邻居。在集中式设置中，客户端的样本比例设为 0.1。所有实验的本地学习率均设为 0.1，并在每个通信轮次后乘以 0.998 进行衰减；集中式方法的全局学习率设为 1.0。所有实验的批量大小固定为 128。我们在 CIFAR-10 和 CIFAR-100 上运行 1000 个全局通信轮次。除 FedSAM 外，所有基线方法均使用带有 0.0005 权重衰减参数的随机梯度下降优化器。我们算法（DFedSAM 和使用 $Q = 1$ 的 DFedSAM-MGS）的其他优化器超参数 $\rho = 0.01$ 通过在集合 $\{0.01, 0.025, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0\}$ 上进行网格搜索确定；FedSAM 中 ρ 的取值则分别遵循（Qu 等人，2022）。此外，按照（Sun 等人，2022），DFedAvgM 的本地优化使用 0.9 的动量。对于本地迭代 K，D-PSGD 的训练周期设为 1，所有其他方法则设为 5。

B.4. 通信配置。

具体而言，例如（Dai等人，2022年），与集中式方法相比，去中心化方法实际上会产生多得多的通信量，因为在网络拓扑中，每个客户端都需要将本地信息传输给其邻居。然而，在集中式设置下，只有部分采样的客户端能够通过中央服务器上传其参数更新。因此，为了进行公平比较，我们在**第**5.2节中为去中心化方法使用动态时变连接拓扑，限制每个客户端最多与10个邻居通信，这些邻居从所有客户端中随机无放回地采样；并且只有彼此互为邻居的10个客户端才能执行一次 gossip 步骤，以交换其本地信息。在 DFedSAM-MGS 中，gossip 步骤执行 Q 次，$10 \times Q$ 个无放回采样的客户端可以执行一次 gossip 步骤，以交换其本地信息。

C. 病态划分下实验的更多细节

病态划分。 为了使实验更具说服力，我们还分别使用 VGG-11 和 ResNet-18 模型，在病态数据划分设置下对 CIFAR-10 和 CIFAR-100 进行了实验（Zhang 等人，2020年），经过500个通信轮次后进行评估。具体而言，排序后的数据被划分为200个分区，并分配给100个客户端；每个客户端从2个类别中随机分配2个分区。显然，与狄利克雷分布划分设置相比，在这种划分设置下训练更加困难。这意味着异质性程度更高，也更接近真实场景。

与所有基线方法的比较。从表 3中，我们可以清楚地看到，由于更高的异质性程度，最终性能比 Dir(α) 中的性能更差，尤其是在 CIFAR-100 等更复杂的数据集上。在这种划分方式下，可以看出数据异质性对性能的不利影响更加严重。然而，与其他方法相比，我们的方法仍然受益于 SAM，从而验证了我们的算法的有效性。此外，这也表明，全局统一模型可能不适用于统计异质性更严重的场景。针对这一问题，现有许多工作采用了一些技术来缓解其影响，例如联邦学习中的模型个性化（Zhang 等人，2020年；Li 等人，2021年；Chen &Chao，2021年；Deng 等人，2020年；Huang 等人，2021年）。

[2]在本文中，我们关注去中心化联邦学习，即包含多个本地迭代点的本地训练；而去中心化学习/训练则关注单步本地训练。例如，D-PSGD（Lian 等人，2017年）是一种去中心化训练算法，在每个通信轮次中使用单步随机梯度下降来训练本地模型。

表 3. 病态数据划分下在 CIFAR-10 和 CIFAR-100 上的性能。

Task	每个客户端仅从全部数据类别中抽取 2 个类别时，各算法的测试准确率（%）							
	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. DFedSAM 和 DFedSAM-MGS 的收敛性分析

下面，我们分别给出 DFedSAM 和 DFedSAM-MGS 的收敛结果证明。请注意，**定理** 4.5 的证明在以下两个章节中进行了详细介绍：D.2 和 D.3 ，其中 $Q = 1$ 和 $Q > 1$分别表示相应内容。

D.1. 预备引理

引理 *D.1*（引理 *4*，*Lian* 等人，*2017*）。对于任意 $t \in \mathbb{Z}^+$，混合矩阵 $W \in \mathbb{R}^m$ 满足 $\|W^t - P\|_{\text{op}} \leq \lambda^t$，其中$\lambda := max\{|\lambda_2|, |\lambda_m(W)|\}$，对于矩阵 A，我们将其谱范数记为 $\|A\|_{\text{op}}$。此外，$1 := [1, 1, \ldots, 1]^\top \in \mathbb{R}^m$ 且

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

在 [命题 1，（Nedic 和 Ozdaglar，2009年）], 作者还证明了 $\|W^t - \mathbf{P}\|_{\text{op}} \leq C\lambda^t$ 对于某个 $C > 0$ 取决于该矩阵的量。

引理 D.2（引理 A.5，（Qu 等人，2022年））。（全局梯度的有界方差，$\|\nabla f_i(\mathbf{x} + \delta_i) - \nabla f(\mathbf{x} + \delta)\|^2$。）假设*4.3* 和*4.4*的直接推论是，带扰动的局部梯度和全局梯度的方差可界定如下：

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

引理 D.3 (引理 B.1, (Qu 等人，2022年))。*(DFedSAM* 的有界性*)*。对于满足 $\eta \leq \frac{1}{4KL}$ 的任意学习率，*DFedSAM* 的更新因 $\delta_{i,\ k} - \delta$ 而产生的漂移为 $\mathcal{E}_\delta$：

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

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

引理 D.4。假设假设 *4.3* 和 *4.4* 成立，且 $(\mathbf{y}^{t,\ k}(i) + \delta_{i,\ k})_{t \geq 0}$、$(\mathbf{x}^{t,\ k}(i))_{t \geq 0}$ 均由 *DFedSAM* 在所有 $i \in \{1, 2, ..., m\}$ 下生成。若 *DFedSAM* 在任意学习率 $\eta \leq \frac{1}{10KL}$ 下的客户端更新，则可得：

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

其中 $0 \leq k \leq K - 1$。

证明。
对于任意节点 i 的任意局部迭代 $k \in \{0, 1, ..., K-1\}$，均有

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

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

d

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

根据引理*D.3* 和假设，界为

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

and

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

其中 $\mathbb{E}\left\|\delta_{i,\ k-1}\right\|^2\le\rho^2$。

因此，我们可以得到

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

其中 $\mathbb{E}\left\|\nabla f(\mathbf{x}^t)\right\|^2=\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2$, $f(\mathbf{x})\colon=\frac{1}{m}\sum_{i=1}^{m}f_i(\mathbf{x})$，且 $\nabla f_i(\mathbf{x}^t)\colon=\nabla f(\mathbf{x}^t(i))$。

从 $\tau=0$到 k 的递推得到

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

至此，证明完成。

引理 D.5. 假设本地迭代次数 K 足够大。令 $\{\mathbf{x}^t(i)\}_{t\ge0}$ 为 *DFedSAM* 在所有 $i\in\{1,2,...,m\}$ 和任意学习率 $\eta>0$下生成的结果，我们有以下界:

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

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

根据 *[*引理 *4*，*(Sun* 等人，*2022*年*)]*, 记 $\mathbf{Z}^t:=\left[\mathbf{z}^t(1),\ \mathbf{z}^t(2)\ldots,\ \mathbf{z}^t(m)\right]^\top\in\mathbb{R}^{m\times d}$。采用这些记号，可得

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

where $\zeta^t := \mathbf{W}\mathbf{X}^t - \mathbf{W}\mathbf{Z}^t$。迭代方程 *(9)* 可重写为以下表达式

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

显然，可得

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

根据引理*D.1*，有

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

将方程（*10*）两边同乘$\mathbf{P}$ 并利用方程（*11*），可得

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

其中使用了初始化$\mathbf{X}^0 = \boldsymbol{0}$。因此，可得

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

根据柯西不等式，

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

直接计算可得

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

根据引理 *D.4* 和假设 *3*，对任意 j,

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

因此，我们得到

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

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.*

引理 D.6。 假设局部迭代次数 K 足够大。对于所有 $i \in \{1, 2, ..., m\}$ 以及任意学习率 $\eta > 0$，由 *DFedSAM-MGS* 生成 $\{\mathbf{x}^t(i)\}_{t\geq 0}$，则有以下界：

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

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

根据 *[引理 4，（Sun 等人，2022年）]* 和 引理 *D.5*，记 $\mathbf{Z}^t := \left[\mathbf{z}^t(1),\ \mathbf{z}^t(2)\ldots,\ \mathbf{z}^t(m)\right]^\top \in \mathbb{R}^{m\times d}$。根据这些记号，我们有

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

其中 $\zeta^t := \mathbf{W}^Q\mathbf{X}^t - \mathbf{W}^Q\mathbf{Z}^t$。迭代方程（*14*）可改写为以下表达式

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

显然，可得

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

根据引理*D.1*，可得

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

将等式（*15*）两边同乘$\mathbf{P}$，并利用等式（*16*），可得

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

其中我们使用了初始化$\mathbf{X}^0 = \boldsymbol{0}$。于是可得

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

根据柯西不等式，

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

直接计算可得

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

根据 引理 *D.4* 和假设 *3*，对于任意 j,

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

因此，我们得到

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

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

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. DFedSAM 收敛结果的证明。

注意到$\mathbf{P}\mathbf{X}^{t+1} = \mathbf{P}\mathbf{W}\mathbf{Z}^t = \mathbf{P}\mathbf{Z}^t$，这也意味着

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

其中 $\mathbf{X} := \left[\mathbf{x}(1),\ \mathbf{x}(2).....,\ \mathbf{x}(m)\right]^\top \in \mathbb{R}^{m\times d}$ 以及 $\mathbf{Z} := \left[\mathbf{z}(1),\ \mathbf{z}(2).....,\ \mathbf{z}(m)\right]^\top \in \mathbb{R}^{m\times d}$。因此我们有

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

其中 $\mathbf{z}^t := \frac{\sum_{i=1}^m \mathbf{z}^t(i)}{m}$ 以及 $\mathbf{x}^t := \frac{\sum_{i=1}^m \mathbf{x}^t(i)}{m}$ 在每个节点中,

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

∇f的利普希茨连续性:

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

其中我们使用了（17）。
并且（18）可得：

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

其中 $a)$ 使用 $\nabla f_i=\mathbb{E}\nabla F_i$ 和 $\nabla f=\frac{1}{m}\sum_{i=1}^{m}\nabla f_i$, $b)$ 使用利普希茨连续性，而 $c)$ 使用 **引理** D.4。同时，我们可以得到

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

其中 $a)$ 使用 **假设** 4.3 和 4.4。因此，
（19）可表示为

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

其中 $C_1=2K(\frac{4K^3L^2\eta^2\rho^4}{2K-1}+8K\eta^2(L^2\rho^2+\sigma_g^2+\sigma_l^2))+\frac{2K\rho^2}{2K-1}$。
此外，根据**引理** D.5，我们可以得到

$$
\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2&\le2L^2\frac{\sum_{i=1}^{m}\left\|\mathbf{x}^t(i)-\overline{\mathbf{x}^t}\right\|^2}{m}+2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2\\
&\overset{a)}{\le}2L^2\frac{C_2\eta^2}{(1-\lambda)^2}+2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}^t})\right\|^2,
\end{aligned}
\tag{23}
$$

其中 a) 使用 **引理** D.5和 $C_2=2K(\frac{4K^3L^2\rho^4}{2K-1}+8K(L^2\rho^2+\sigma_g^2+\sigma_l^2)+\frac{8K}{m}\sum_{i=1}^{m}\mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2)+\frac{2K\rho^2}{\eta^2\,2K-1}$。
因此，我们有

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

其中 $C_3 = 2K(\frac{4K^3L^2\rho^4}{2K-1} + 8K(L^2\rho^2+\sigma_g^2+\sigma_l^2)) + \frac{2K\rho^2}{\eta^2\ 2K-1}$
()。然后，(19) 可以表示为

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

其中 a) 使用（24）且 η 是一个非常小的值，将不等式（25）从 $t=1$ 到 T求和，即可得到如下已证明的结果:

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

如果我们选择学习率 $\eta = \mathcal{O}(1/L\sqrt{KT})$ 和 $\eta \le \frac{1}{10KL}$，且通信轮次 T足够大，则有

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

当 p　扰动幅度 ρ 与学习率成正比，例如 $\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$，则有:　—

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

根据**定义** 4.2，可得

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

证明完毕。

D.3. DFedSAM-MGS 收敛结果的证明

经过多个 Gossip 步数后，$\mathbf{x}^0$和$\mathbf{z}^0$分别表示为$\mathbf{x}$和$\mathbf{z}$。同时，$\mathbf{Z}^{t,Q} = \mathbf{Z}^{t,0}\cdot\mathbf{W}^Q = \mathbf{Z}^t\cdot\mathbf{W}^Q$。注意到$\mathbf{P}\mathbf{X}^{t+1} = \mathbf{P}\mathbf{W}^Q$ $\mathbf{Z}^t = \mathbf{P}\mathbf{Z}^t(Q>1)$，也就是说

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

其中$\mathbf{X} := \left[\mathbf{x}(1), \mathbf{x}(2)\ldots, \mathbf{x}(m)\right]^\top \in \mathbb{R}^{m\times d}$ 以及$\mathbf{Z} := \left[\mathbf{z}(1), \mathbf{z}(2)\ldots, \mathbf{z}(m)\right]^\top \in \mathbb{R}^{m\times d}$。因此，我们有

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

其中 $\mathbf{z}^t := \frac{\sum_{i=1}^m \mathbf{z}^t(i)\ m}{}$和 $\mathbf{x}^t := \frac{\sum_{i=1}^m \mathbf{x}^t(i)\ m}{}$ 在每个节点中,

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

∇f 的利普希茨连续性：

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

其中我们使用了（26）。
并且使用（27）：

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

其中 $a)$ 使用 $\nabla f_i = \mathbb{E}\nabla F_i$ 和 $\nabla f = \frac{1}{m}\sum_{i=1}^m \nabla f_i$，$b)$ 使用利普希茨连续性，而 $c)$ 使用 **引理** D.4。与此同时，我们可以得到

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

其中 $a)$ 使用 **假设** 4.3 和 4.4。
因此，（28）可表示为

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

其中 $C_1 = 2K\ (\frac{4K^3L^2\eta^2\rho^4}{2K-1} + 8K\eta^2\ (L^2\rho^2 + \sigma_g^2 + \sigma_l^2)\ + \frac{2K\rho^2}{2K-1}$。
此外，根据**引理** D.6，我们可以得到

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

其中 a) 使用了**引理** D.6and $C_2 = 2K(\frac{4K^3L^2\rho^4}{2K-1} + 8K(L^2\rho^2+\sigma_g^2+\sigma_l^2) + \frac{8K}{m}\sum_{i=1}^m \mathbb{E}\left\|\nabla f(\mathbf{x}^t(i))\right\|^2) + \frac{2K\rho^2}{\eta^2(2K-1)}$()。此外，我们有

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

其中 $C_3 = 2K(\frac{4K^3L^2\rho^4}{2K-1}+8K(L^2\rho^2+\sigma_g^2+\sigma_l^2)) + \frac{2K\rho^2}{\eta^2(2K-1)}$()。
因此，(28)为

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

其中，a) 使用了 (33)，且 η 是一个非常小的值，将不等式 (34) 从 $t=1$ 到 T求和，即可得到如下证明结果：

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

将不等式 (34) 从 $t=1$ 到 T求和，即可得到如下证明结果：

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

如果我们选择学习率 $\eta = \mathcal{O}(1/L\sqrt{KT})$ 和 $\eta \le \frac{1}{10KL}$，且通信轮次 T足够大，根据**定义** 4.2 以及 $\mathbf{\Phi}(\lambda, m, Q) = \frac{\lambda^Q+1}{(1-\lambda)^2m^{2(Q-1)}} + \frac{\lambda^Q+1}{1-\lambda^Q}$()~~是收敛界的关键参数，~~与谱间隙、客户端数量和多次 Gossip 步数有关。
因此我们有

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

当 p 扰动幅度 ρ 与学习率成正比，例如 $\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$，于是我们有：

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

至此，证明完成。