arXiv:2305.00873v2 [cs.LG] 2 May 2023

迈向客户端级差分隐私下联邦学习中更平坦的损失景观与更好的泛化

Yifan Shi、Kang Wei，*IEEE* 会员、Li Shen、Yingqi Liu、Xueqian Wang，*IEEE* 会员、Bo Yuan，*IEEE* 高级会员和 Dacheng Tao，*IEEE* 会士

摘要——为防御推理攻击并缓解联邦学习（FL）中的敏感信息泄露，客户端级差分隐私联邦学习（DPFL）通过裁剪本地更新并添加随机噪声，已成为隐私保护的事实标准。然而，现有 DPFL 方法往往会形成尖锐的损失景观，且权重扰动鲁棒性较差，从而导致严重的性能下降。为缓解这些问题，我们提出一种名为 DP-FedSAM 的新型 DPFL 算法，利用梯度扰动来减轻差分隐私带来的负面影响。具体而言，DP-FedSAM 将尖锐度感知最小化（SAM）优化器纳入其中，以生成稳定性和权重扰动鲁棒性更高的局部平坦模型，从而使本地更新的范数较小，并增强对差分隐私噪声的鲁棒性，进而提升性能。为进一步降低随机噪声的幅度并取得更好的性能，我们采用本地更新稀疏化技术，提出 DP-FedSAM-top$_k$ 。从理论角度出发，我们给出收敛分析，研究所提算法如何缓解差分隐私导致的性能下降。同时，我们基于 R´enyi 差分隐私给出严格的隐私保证，并对本地更新进行敏感度分析和泛化分析。最后，实证结果表明，与现有 DPFL 最先进水平的基线方法相比，我们的算法取得了最先进水平的性能。

索引词—联邦学习、客户端级差分隐私、尖锐度感知最小化、本地更新稀疏化

1 引言

联邦学习（FL） [2] 使分布式客户端能够协作训练共享模型，而无需共享数据。然而，联邦学习面临着隐私泄露的严重困境 [3]。近期研究表明，好奇的服务器可以借助精心设计的生成模型和/或影子模型，推断客户端的隐私信息，例如成员关系和数据特征 [4], [5], [6], [7], [8]。为解决这一问题，联邦学习中引入了差分隐私（DP） [9] ，它可以保护任意客户端数据中的每个样本（实例级差分隐私 [10], [11], [12], [13]），也可以保护客户端之间的信息（客户端级差分隐私 [14], [15], [16], [17], [18], [19]）。一般而言，由于能够获得更好的模型性能，客户端级差分隐私更适用于现实场景。例如，谷歌已将采用客户端级差分隐私 [20], [14]的语言预测模型应用于移动设备。通常，基于高斯噪声扰动的方法是常用的为确保强客户端级差分隐私而采用。然而，该方法包含两项操作：将本地更新的范数裁剪至敏感度阈值 C ，并添加与模型大小成正比的随机噪声，其标准差（STD）也由 C 决定。这些步骤可能导致严重的性能下降 [21], [22]，尤其是在大型复杂模型 [23], 上，例如 ResNet-18 [24]，或在异构数据场景下。

这一问题背后的原因有两方面：（i）由于裁剪操作，尤其是在 C 值较小时，本地更新中包含的有用信息会被丢弃；（ii）添加随机噪声会严重破坏本地更新，并导致本地模型之间的方差增大，从而加剧本地模型之间的不一致性，尤其是在 C 值较大时 [21]。现有研究尝试通过限制本地更新的范数 [21] ，并采用本地更新稀疏化技术 [21], [22] ，来减轻裁剪和添加随机噪声带来的不利影响。然而，与不考虑隐私的方法（如 FedAvg[25]）相比，模型性能下降仍然显著。

1.1 研究动机

为进一步探究这一现象，我们在图 1 (a) 和 (b) 中分别比较了 FedAvg[25] 与 DP-FedAvg [14], [15] 在采用狄利克雷分布（$\alpha = 0.6$）进行划分的 CIFAR-10 数据集 [27] 以及以 ResNet-18 为骨干网络 [24] 时的损失景观结构和曲面等高线 [26] 。请注意，经过较长的通信

• 本文的早期版本部分内容曾发表于 *2023*年 *IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*） *[1]*。

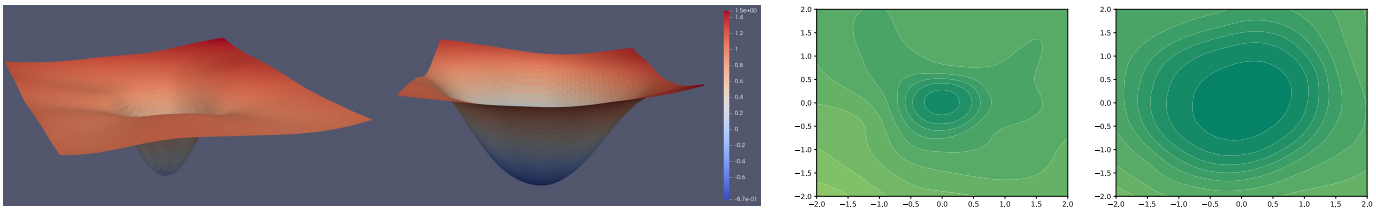
• *Yifan Shi* 和 *Xueqian Wang* 任职于中国深圳 *518055* 清华大学深圳国际研究生院（电子邮箱：*shiyf21@mails.tsinghua.edu.cn*；*wang.xq@sz.tsinghua.edu.cn*）。

• *Kang Wei* 和 *Yingqi Liu* 任职于中国南京 *210094* 南京理工大学电气与光学工程学院（电子邮箱：{*kang.wei, lyq*}*@njust.edu.cn*）。

• *Li Shen* 任职于中国北京京东探索研究院（电子邮箱：*mathshenli@gmail.com*）。

• *Bo Yuan* 任职于中国深圳 *518055* 深圳市智慧与策略科技有限公司（电子邮箱：*boyuan@ieee.org*）。

• *Dacheng Tao* 任职于澳大利亚新南威尔士州 *2006* 悉尼大学。（电子邮箱：*dacheng.tao@gmail.com*）。



(a) 损失景观。 (b) 损失曲面等高线。

图1. 损失景观 (a) 和曲面等高线 (b) 对比：DP-FedAvg（左）与 FedAvg（右）。

轮次后，DP-FedAvg 的损失值仍然更高，因此其收敛效果不如 FedAvg。此外，FedAvg 的损失景观更平坦，而 DP-FedAvg 的损失景观更陡峭，这分别导致其泛化能力较差（极小值更尖锐，见图 1 (a)）和权重扰动鲁棒性较弱（见图 1 (b)）；其原因分别是裁剪后的本地更新信息和加剧的模型不一致性。基于这些观察，一个值得探讨的研究问题是：我们能否通过使损失景观更平坦、泛化能力更强，进一步缓解性能下降？

为回答这一问题，我们提出了采用梯度扰动的 DP-FedSAM，以提升模型性能。具体而言，每个客户端通过 SAM 优化器 [28] 生成一个局部平坦模型，从而提高稳定性。随后，通过聚合多个平坦的本地模型，可以生成一个潜在的全局平坦模型，从而提升泛化能力并增强对差分隐私噪声的鲁棒性，显著改善性能，并在性能与隐私之间实现更好的权衡。为进一步减小随机噪声的幅度，同时获得更好的性能，我们在 DP-FedSAM 的基础上采用本地更新稀疏化技术 [22], [21]，提出了 DP-FedSAM-top$_k$ 。理论上，我们在随机非凸情形下给出了更紧的界

$$\mathcal{O}\left(\frac{L^2\sqrt{T\sigma^2C^2pd}}{m^2\sqrt{K}}+\frac{T^2}{\sqrt{KT}+\sum_{t=1}^{T}(\alpha^t\sigma_g^2+\tilde{\alpha}^tL^2)}\right)+(\cdots)$$

，其中 $\frac{1}{T}\sum_{t=1}^{T}\alpha^t$ 和 $\frac{1}{T}\sum_{t=1}^{T}\tilde{\alpha}^t$ 均为有界常数。p、K 和 T 分别表示稀疏率、本地迭代步数和通信轮次。具体而言，在裁剪和添加噪声操作之前，本地更新的平均范数 α^t 以及客户端之间的本地更新一致性 $\tilde{\alpha}^t$ 表示为

$$\bar{\alpha}^t := \frac{1}{M}\sum_{i=1}^{M}\alpha_i^t \quad \text{and} \quad \tilde{\alpha}^t := \frac{1}{M}\sum_{i=1}^{M}|\alpha_i^t-\bar{\alpha}_i^t|,$$

其中，$\alpha_i^t = \min(1, \frac{C}{\eta\|\sum_{k=0}^{K-1}\tilde{\mathbf{g}}^{t,k}(i)\|_2})$ 衡量了客户端 i 在第 t 个通信轮次中本地更新裁剪带来的负面影响。请注意，η 是学习率，$\tilde{\mathbf{g}}^{t,k}(i)$ 是第 k 个本地迭代中使用本地 SAM 优化器计算的梯度。接下来，我们介绍 SAM 如何减轻差分隐私的不利影响。对于裁剪操作，DP-FedSAM 降低了 l_2 范数，并减轻了本地更新之间的不一致性对收敛造成的负面影响。对于添加噪声操作，我们获得了更强的权重扰动鲁棒性，从而减少随机噪声造成的性能损失，使其在联邦学习训练过程中更能抵御差分隐私噪声。同时，我们对算法进行了敏感度、隐私、敏感度和泛化分析。在独立同分布（IID）设置下，我们在 EMNIST、CIFAR-10 和 CIFAR-100 数据集上进行了大量实验。以及非独立同分布设置。此外，我们研究损失景观、曲面等高线以及本地更新的范数分布，以探索SAM与差分隐私结合后的内在影响；这些研究与理论分析共同证实了 DP-FedSAM的效果。

1.2 贡献

我们的工作主要有四方面贡献。

- 我们提出了两种新颖的方案 DP-FedSAM 和 DP- FedSAM-top$_k$ ，用于缓解差分隐私联邦学习中的性能下降问题（见第4节）。
- 我们在随机非凸情形下建立了更优的收敛速度，其界比传统界更紧。 [21], [22]此外，我们还提供了严格的隐私保证、敏感度和泛化分析（见第5节）。
- 我们首次深入分析了本地更新的平均范数 α^t 以及客户端之间的本地更新一致性 $\tilde{\alpha}^t$ 对收敛的作用（见第5.3节的定理3）。同时，我们通过实验验证了本地更新范数在减轻不利影响方面的理论结果（见第6.3节）。
- 我们进行了大量实验以验证所提算法的效果。与若干强大的差分隐私联邦学习基线方法相比，我们的算法达到了最先进水平（SOTA）的性能（见第6节）。

1.3 组织结构

第2节回顾客户端级差分隐私联邦学习、SAM 优化器以及差分隐私联邦学习中的模型压缩相关工作。第3节介绍联邦学习、差分隐私和本地更新稀疏化的背景。第4节阐述所提出的 DP-FedSAM 和 DP-FedSAM-顶部$_k$ 。此外，我们在第5节对所提方法的敏感度、收敛性和泛化能力进行了理论分析。第6节给出了全面的实验评估。本文最后在第7节总结全文，并提出未来工作的方向。

2 相关工作

客户端级差分隐私联邦学习。客户端级差分隐私联邦学习是保护每个客户端数据的事实标准解决方案。DP-FedAvg [29] 是这一设定下的首次尝试，它在移动键盘中训练语言预测模型，并通过采用高斯机制和组合隐私保证来确保客户端级差分隐私保证。随后，[30], [31] 中的工作提出了一个全面的端到端系统，该系统对数据进行适当的离散化，并在执行安全聚合之前添加离散高斯噪声。同时，AE-DPFL [32]利用基于数据标签的投票机制，而非对梯度取平均，从而避免依赖维度并显著降低通信成本。

Fed-SMP [22]使用稀疏化模型扰动（SMP）来减轻隐私保护对模型准确率的影响。不同于上述方法，近期一项研究 [21] 重新审视了这一问题，并分别利用有界本地更新正则化（BLUR）和本地更新稀疏化（LUS），在执行差分隐私保证操作前限制本地更新的范数并减小噪声规模。然而，性能下降的问题仍然存在。

尖锐度感知最小化（SAM）。SAM [28]是一种有效的优化器，可用于训练深度学习（DL）模型。它利用损失景观的平坦性几何来提升模型泛化能力。近期，[33] 中的研究探讨了 SAM 的性质，并给出了其在非凸目标上的收敛结果。作为一种强大的优化器，SAM 及其变体已应用于各种机器学习（ML）任务 [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44]。具体而言， [45], [43], 和 [46] 中的研究将 SAM 融入模型，以提升泛化能力，进而缓解联邦学习中的分布偏移问题，并取得了新的最先进水平。然而，据我们所知，目前针对差分隐私联邦学习中 SAM 的实证表现和理论分析所做的工作仍然有限 [1]。本文在 [1]的基础上，进一步对结合本地更新稀疏化的 SAM 本地优化器进行更详细、系统的评估。

差分隐私联邦学习中的稀疏化。为了更好地权衡性能与隐私保护，许多研究利用隐私保护中的稀疏化技术，引入大量随机噪声 [22], [21], [11],[47], [48]。该方法以 k/d 的稀疏率（d 表示权重尺度）仅保留本地模型每一层中相对较大的权重，其余权重则置为零。其优势在于能够减少随机噪声的数量（无需在稀疏权重位置添加噪声），并提高模型性能。在差分隐私联邦学习中，稀疏化技术可分为两种策略：随机稀疏化和基于权重的稀疏化。例如，Fed-SPA [11] 将随机稀疏化与梯度扰动相结合，以实现更好的效用-隐私权衡，并降低实例级差分隐私联邦学习的通信成本。Fed-SMP [22] 采用结合这两种策略的稀疏模型扰动（SMP），以减轻隐私保护对模型准确率的影响。文献 [21] 则利用基于权重的稀疏化，在执行保证差分隐私的操作之前减少噪声规模。

与本文最相关的工作是 DP-FedAvg[14], Fed-SMP [22], 以及采用 LUS 和 BLUR 的 DP-FedAvg[21]。然而，由于随机噪声加剧了客户端之间的模型不一致性，这些工作性能较差。与现有工作不同，我们尝试通过使损失景观更加平坦、提升权重扰动能力的稳健性来缓解这一问题。此外，另一项相关工作是 FedSAM [45]，它整合 SAM 优化器以增强本地模型的平坦性，并在联邦学习中取得了新的最先进水平。在上述研究的基础上，我们首次将 SAM 优化器扩展至差分隐私联邦学习，以有效缓解性能下降问题。同时，我们还将在非凸情形下，对敏感度、隐私、收敛和泛化进行了理论分析。最后，我们通过实验验证了理论结果，并证实了该方法相较于差分隐私联邦学习中现有最先进方法的性能优势。

3 预备知识

本节首先介绍联邦学习的问题设定，然后介绍差分隐私中的关键术语，最后介绍稀疏化技术。

3.1 联邦学习

考虑一个由 M 个客户端组成的通用联邦学习系统，每个客户端都拥有自己的本地数据集。分别令 S_i 表示客户端 i 持有的训练样本集，其中 $i \in \mathcal{U} = \{1, 2, \ldots, M\}$。形式上，联邦学习任务可表示为：

$$\mathbf{w}^\star = \arg\min_{\mathbf{w}} \sum_{i \in \mathcal{U}} p_i f_i(\mathbf{w}, S_i), \tag{1}$$

其中 $p_i = |S_i|/|S| \geq 0$，且 $\sum_{i \in \mathcal{U}} p_i = 1$；$f_i(\cdot)$ 是局部损失函数，其中 $f_i(\mathbf{w}) = F_i(\mathbf{w}; \xi_i)$，$\xi_i$ 是客户端 i 的一批次样本数据。我们假设 $S = \{(x_1, y_1) \ldots, (x_N, y_N) | x_i \in \mathcal{X} \subset \mathbb{R}^{d_X}, y_i \in \mathcal{Y} \subset \mathbb{R}^{d_Y}, i = 1, \ldots, N\}$ 是所有客户端持有的完整训练数据集，其中 x_i 是第 i 个特征，y_i 是对应的标签；d_X 和 d_Y 分别是特征和标签的维度。同时，我们定义 $z_i = (x_i, y_i)$，并假设 z_i 满足数据分布 $\mathcal{D}$。$|S_i|$ 是训练数据集 S_i 的大小，$|S| = \sum_{i \in \mathcal{U}} |S_i|$ 是训练数据集的总大小。对于第 i 个客户端，通过以下方式在其私有训练数据 S_i 上学习本地模型：

$$\mathbf{w}_i = \mathbf{w}_i^t - \eta \nabla f_i(\mathbf{w}_i, S_i). \tag{2}$$

通常，各客户端的局部损失函数 $f_i(\cdot)$ 具有相同的表达形式。随后， M 相关客户端基于异构训练数据 $S_i, \forall i \in \mathcal{U}$，协同学习全局模型 $\mathbf{w}$ 。

3.2 差分隐私

差分隐私（DP） [9] 是一种用于衡量隐私风险的严格隐私概念。本文考虑其一种放宽形式：R´ényi差分隐私（RDP） [49] ，并在定理 2 中用于隐私计算。此外，我们采用最大散度来给出定理 5 中的泛化界。

定义 1 (Rényi´差分隐私， [49])。给定实数 $\alpha \in (1, \infty)$和隐私参数 $\rho \geq 0$，若随机机制 $\mathcal{M}$满足以下条件，则称其满足(α, ρ)-*Rényi*差分隐私：对于任意两个相邻数据集 U, U'，若它们仅在一个样本数据上不同，则其间的 *Rényi* α-散度，即 $\mathcal{M}(U)$与 $\mathcal{M}(U')$之间的散度，满足：

$$D_\alpha \left[\mathcal{M}(U) \| \mathcal{M}\left(U'\right)\right] := \frac{1}{\alpha - 1} \log \mathbb{E}\left[\left(\frac{\mathcal{M}(U)}{\mathcal{M}\left(U'\right)}\right)^\alpha\right] \leq \rho, \tag{3}$$

其中，期望取自 $\mathcal{M}\left(U'\right)$ 的输出。

Rényi 差分隐私是一种有用的分析工具，可用于衡量隐私并准确表征隐私损失尾部的保证，其严格强于 $\delta > 0$ 的 (ϵ, δ)-差分隐私。我们基于该工具对每位用户的隐私损失进行隐私分析。此外，我们

通过证明联邦学习训练过程在通信轮次中满足(ϵ, δ)-差分隐私，且最大散度近似误差为δ，从而给出高概率泛化界，其中通信轮次遵循 [50]。

定义 2 (δ-近似最大散度， [9])。 对于任意随机变量 $X \in U$ 和 $Y \in U$，其中 U 是一个数据集，δ-近似散度定义为 X 与 Y 之间的

$$D_{\infty}^{\delta}(X \parallel Y)=\max_{U \subseteq \operatorname{Supp}(X): \mathbb{P}(Y \in U) \geq \delta}\left[\log \frac{\mathbb{P}(Y \in U)-\delta}{\mathbb{P}(Y \in U)}\right]. \quad (4)$$

定义 3 (l_2 敏感度， [21])。 设 $\mathcal{F}$ 为一个函数，则 $\mathcal{F}$的L_2-敏感度定义为 $\mathcal{S}$ =最大值$_{U \simeq U'} \|\mathcal{F}(U)-\mathcal{F}(U')\|_2$，其中最大化是在所有相邻数据集对 U 和 U'上进行的。

函数 $\mathcal{F}$ 的敏感度刻画了在最坏情况下个人数据对函数 $\mathcal{F}$ 造成的变化幅度。因此，它在确定确保差分隐私所需噪声的大小方面起着关键作用。

定义 4 （客户端级差分隐私， [51]）。如果对于任意两个相邻数据集 U、U'（通过添加或移除任意客户端的全部数据构造），随机化算法$\mathcal{M}$ 对每个可能的输出子集 O 都满足以下不等式，则称其为（ϵ， δ）-差分隐私：

$$\Pr[\mathcal{M}(U) \in O] \leq e^{\epsilon} \Pr\left[\mathcal{M}\left(U'\right) \in O\right]+\delta. \quad (5)$$

在客户端级差分隐私中，我们旨在确保任何客户端的参与信息都受到保护。因此，无论某个客户端是否参与，我们都需要使本地更新保持相似。

3.3 本地更新稀疏化

为减少随机噪声量，并在性能与隐私保护之间取得更好的权衡，现有工作通常在客户端级差分隐私联邦学习中采用本地更新稀疏化技术，也称为 top-k 稀疏器，该方法基于每层权重的幅值进行稀疏化。

定义 5 （本地更新稀疏化， [21], [22]）。对于 $1 \leq k \leq d$和本地更新向量$\mathbf{x} \in \mathbb{R}^d$，*top-$k$* 稀疏器$\mathrm{top}_k$： $\mathbb{R}^d \rightarrow \mathbb{R}^d$ 定义如下

$$[\operatorname{top}_k(\mathbf{x})]_j := \begin{cases}[\mathbf{x}]_{\pi(j)}, & \textit{if } j \leq k \\ 0, & \textit{otherwise}\end{cases}, \quad (6)$$

其中 $\Omega_k = \binom{[d]}{k}$ 表示 $[d]$中所有 k元素子集构成的集合，且 π 是 $[d]$ 的一个排列，使得 $|[\mathbf{x}]_{\pi(j)}| \geq |[\mathbf{x}]_{\pi(j+1)}|$，对于 $j \in [1, d-1]$。注意，执行本地更新稀疏化后的稀疏率为 $p = k/d$ 。

因此，在对本地模型更新进行裁剪并在差分隐私中添加随机噪声后，top-k 稀疏化器会丢弃不太可能重要的本地更新参数，只将幅值最大的 k 参数/坐标传输到服务器。

4 方法论

为重新审视差分隐私联邦学习中的性能下降问题，我们在图 1(a) 和 (b) 中分别考察了 FedAvg 和 DP-FedAvg 的损失景观与曲面等高线。我们发现，与联邦学习方法相比，差分隐私联邦学习方法产生的损失景观更加尖锐，泛化能力和权重扰动鲁棒性也更差。这意味着差分隐私联邦学习方法可能导致平坦性较差，并使模型对噪声扰动敏感。本文计划从优化器视角入手，在每个客户端采用 SAM 优化器，称为 DP-FedSAM，其局部损失函数定义如下：

$$f_i(\mathbf{w}) = \mathbb{E}_{\xi \sim \mathcal{D}_i} \max_{\|\delta_i\|_2 \leq \rho} F_i(\mathbf{w}^{t,k}(i) + \delta_i; \xi_i), \quad i \in \mathcal{N}, \quad (7)$$

其中，$\mathbf{w}^{t,k}(i) + \delta_i$ 表示扰动模型， ρ 是一个预先定义的常数，用于控制扰动半径，而 $\|\cdot\|_2$ 表示 l_2-范数。与通过随机梯度下降（SGD）寻找解不同， [52], [53], SAM [28]旨在通过添加一个小扰动来寻找平坦区域中的解，即 $w + \delta$，从而获得更稳健的性能。具体而言，对于每个客户端$i \in \{1, 2, ..., M\}$，以及每个通信轮次 $t \in \{0, 1, ..., T-1\}$中的每次本地迭代 $k \in \{0, 1, ..., K-1\}$，客户端 i 中的第 k次内部迭代执行如下：

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

$$\tilde{\mathbf{g}}^{t,k}(i) = \nabla F_i(\mathbf{w}^{t,k} + \delta(\mathbf{w}^{t,k}); \xi), \delta(\mathbf{w}^{t,k}) = \frac{\rho \mathbf{g}^{t,k}}{\|\mathbf{g}^{t,k}\|_2}, \quad (9)$$

其中， $\delta(\mathbf{w}^{t,k})$ 是通过围绕$\mathbf{w}^{t,k}$ [28]进行一阶泰勒展开计算得到的。随后，我们采用抽样机制、进行梯度裁剪并添加高斯噪声，以确保客户端级差分隐私。请注意，这种抽样方法可以增强隐私保证，因为它降低了泄露特定个体信息的可能性。每个通信轮次以概率 $q = m/M$ 抽取m 个客户端，这对于衡量隐私损失十分重要。我们对这些抽样客户端的本地更新进行裁剪，具体如下：

$$\tilde{\Delta}_i^t = \Delta_i^t \cdot \min\left(1, \frac{C}{\|\Delta_i^t\|_2}\right). \quad (10)$$

裁剪后，我们向本地更新添加高斯噪声，以确保客户端级差分隐私，具体如下：

$$\breve{\Delta}_i^t = \tilde{\Delta}_i^t + \mathcal{N}(0, \sigma^2 C^2 \cdot \mathbf{I}_d/m). \quad (11)$$

给定噪声方差 σ^2，可基于抽样高斯机制 [54]计算累积隐私预算 ϵ 。

随后，我们根据定义5采用稀疏化技术，以减小所添加随机噪声的幅度，称为 DP-FedSAM-top$_k$。其中，本地模型更新根据稀疏率 p保留所有参数中幅度最大的部分。形式上，根据定义5，该操作定义如下：

$$\hat{\Delta}_i^t = \breve{\Delta}_i^t \odot \boldsymbol{m}^t, \quad (12)$$

其中，运算符 $\odot$ 表示逐元素乘法，掩码向量 $\boldsymbol{m}^t$ 的第 j 个坐标若被 top-k 稀疏化器选中，则等于 1，DP-FedSAM-top$_k$ 在 $p = 1.0$ 时与 DP-FedSAM 相同。我们在算法 1 中总结了训练过程。

与现有的差分隐私联邦学习方法相比， [14], [22], [21], 我们算法的优势体现在以下五个方面：(i) 我们将 SAM 引入差分隐私联邦学习，通过在每个客户端中寻找平坦模型来缓解严重的性能下降；这种性能下降是由本地模型间不一致性加剧造成的。具体而言，DP-FedSAM 能够带来更好的泛化

算法 1： DP-FedSAM 和 DP-FedSAM-top$_k$

输入： 客户端总数 M，采样比例 q，通信轮次总数 T，裁剪阈值 C，本地学习率 η，以及本地总数迭代次数为 K。

输出： 全局模型 $\mathbf{w}^T$。

初始化： 随机初始化全局模型 $\mathbf{w}^0$。

对于 $t=0$ **到** $T-1$，**执行**

  随机抽取一组 $m=qM$ 个客户端，不进行放回抽样，记为 $\mathcal{W}^t$。

  并行执行 客户端 $i=1$ **到** m

    执行 $k=0$ **到** $K-1$

      将全局参数更新为本地参数 参数$\mathbf{w}^t(i) \leftarrow \mathbf{w}^t$。

      采样一批本地数据 ξ_i 并计算本地梯度 $\mathbf{g}^{t,k}(i) = \nabla F_i(\mathbf{w}^{t,k}(i);\xi_i)$。

      根据公式 (9) 进行梯度扰动。

      根据公式 (8) 执行本地迭代更新。

    end

    $\Delta_i^t = \mathbf{w}^{t,K}(i) - \mathbf{w}^{t,0}(i)$.

    根据公式 (11) 进行裁剪并添加噪声以实现差分隐私。生成本地更新 $\hat{\Delta}_i^t$，以便上传至服务器端，通过 选项 I 或II。

    返回 $\hat{\Delta}_i^t$.

  end

  $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t + \frac{1}{m}\sum_{i\in\mathcal{W}^t} \hat{\Delta}_i^t$。

结束

选项 I: （DP-FedSAM）

获取 $\hat{\Delta}_i^t$，不进行稀疏化：$\hat{\Delta}_i^t = \breve{\Delta}_i^t$。

选项 II: （DP-FedSAM-上$_k$）

获取 $\hat{\Delta}_i^t$ 根据公式进行本地更新稀疏化 (12)。

能力，并通过使全局模型的损失景观更加平坦，增强对差分隐私噪声的鲁棒性；(ii) 我们详细分析了 DP-FedSAM 如何减轻差分隐私带来的负面影响。我们从本地更新的平均范数 α^t 和客户端间的本地更新一致性 $\tilde{\alpha}^t$两方面对收敛性进行理论分析，并通过观察本地更新的范数分布和平均范数对这些结果进行实证验证（见第 6.3 节）；(iii) 我们对所提出的算法进行了敏感度、隐私和泛化分析（见第 5 节）；(iv) 我们还提出了统一理论，分析 SAM 中梯度扰动 ρ 、本地更新的平均范数 α^t、裁剪操作中的客户端间本地更新一致性 $\tilde{\alpha}^t$ 以及随机噪声的方差 σ^2C^2/m对收敛速度的影响（见第 5.3 节）；(v) 我们探索了结合 SAM 本地优化器的本地更新稀疏化，以进一步提升性能。

5 理论分析

本节将对我们的算法进行严格分析，包括敏感度、隐私和收敛速度。详细证明见**附录A**。下面，我们首先给出几个关键假设。

假设1 (Lipschitz 光滑性)。函数 F_i可微，且 ∇F_i 是 L-*Lipschitz* 连续的，$\forall i \in\{1,2,\ldots,M\}$，即 $\|\nabla F_i(\mathbf{w}) - \nabla F_i(\mathbf{w}')\| \le L\|\mathbf{w}-\mathbf{w}'\|$，对于$\mathbf{w}$，$\mathbf{w}' \in \mathbb{R}^d$。

假设2 (有界方差)。函数 f_i 的梯度具有 σ_l-有界方差，即 $\mathbb{E}_{\xi_i}\left\|\nabla F_i(\mathbf{w}^k(i);\xi_i) - \nabla F_i(\mathbf{w}(i))\right\|^2 \le \sigma_l^2$，$\forall i \in \{1,2,\ldots,M\}$，$k\in\{1,...,K-1\}$，并且全局方差也有界，即 $\frac{1}{M}\sum_{i=1}^M \|\nabla \underline{f_i}(\mathbf{w}) - \nabla f(\mathbf{w})\|^2 \le \sigma_g^2$对于 $\mathbf{w}\in\mathbb{R}^d$。不难验证，$\sigma_g$小于齐次性参数 β，即 $\sigma_g^2 \le \beta^2$。

假设 3 (有界梯度)。对于任意 $i\in\{1,2,\ldots,M\}$和$\mathbf{w}\in\mathbb{R}^d$，我们有 $\|\nabla f_i(\mathbf{w})\| \le B$。

请注意，上述假设较为宽松，且常用于刻画联邦学习[55], [56], [57], [53], [58], [45], [21], [22]的收敛速度。此外，深度学习中常使用梯度裁剪来防止梯度爆炸现象，从而使梯度有界。我们的算法面临的技术挑战在于：（i）SAM 如何减轻差分隐私的影响；（ii）如何详细分析裁剪操作导致的客户端间一致性和本地更新平均范数的影响。

5.1 敏感度分析

首先，我们研究第 t 个通信轮次裁剪前客户端 $i\in\{1,2,...,M\}$ 的本地更新 Δ_i^t 的敏感度。该敏感度上界可粗略衡量隐私保护程度。根据定义 3，客户端 i 在第 t 个通信轮次的敏感度可表示为 $\mathcal{S}_{\Delta_i^t}$ 。

定理 *1*（敏感度分析）。令 $\Delta_i^t(x)$ 和 $\Delta_i^t(y)$ 表示第 t 个通信轮次的本地更新，模型 *x(i)* 和 *y(i)* 分别基于两个仅相差一个样本的数据集。假设初始模型参数 $w^t(i) = x^{t,0}(i) = y^{t,0}(i)$，则本地更新的期望平方敏感度 $\mathcal{S}^2_{\Delta_i^t}$ 的上界为

$$\mathbb{E}\mathcal{S}^2_{\Delta_i^t} \le \frac{6\eta^2\rho^2KL^2(12K^2L^2\eta^2+10)}{1-2\eta^2L^2K} \tag{13}$$

当局部自适应学习率满足 $\eta = \mathcal{O}(1/L\sqrt{KT})$，且扰动幅度 ρ 与学习率成正比时，例如 $\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$，我们有

$$\mathbb{E}\mathcal{S}^2_{\Delta_i^t} \le \mathcal{O}\left(\frac{1}{T^2}\right). \tag{14}$$

在差分隐私联邦学习中，使用随机梯度下降的本地更新的期望平方敏感度为 $\mathbb{E}\mathcal{S}^2_{\Delta_i^t\ SGD} \le \frac{6\eta^2\sigma_l^2K}{1-3\eta^2KL^2}\mathbb{E}\mathcal{S}^2_{\Delta_i^t\ SGD} \le$。。因此，(

$\mathcal{O}(\frac{\sigma_l^2}{KL^2T})$当 $\eta=\mathcal{O}(1/L\sqrt{KT})$时。

注 1. 显然，$\mathbb{E}\mathcal{S}^2_{\Delta_i^t,\ SAM}$ 的上界比$\mathbb{E}\mathcal{S}^2_{\Delta_i^t,\ SGD}$的上界更紧。对于隐私保护而言，这意味着 *DP-FedSAM* 的隐私保证优于 *DP-FedAvg*。同时，在本地迭代中，我们的算法兼具更好的客户端间模型一致性和训练稳定性。

5.2 隐私分析

为实现客户端级隐私保护，我们分析对本地更新进行裁剪后的聚合过程的敏感度。下面给出隐私分析。

引理 1. *DP-FedSAM* 中客户端级差分隐私的敏感度可表示为 C/m。

证明。给定两个相邻批次 $\mathcal{W}^t$ 和 $\mathcal{W}^{t,\text{ adj}}$，其中$\mathcal{W}^{t,\text{ adj}}$ 多出或少一个客户端，我们有

$$\left\|\frac{1}{m}\sum_{i\in\mathcal{W}^t}\Delta_i^t-\frac{1}{m}\sum_{j\in\mathcal{W}^{t,\text{adj}}}\Delta_j^t\right\|_2=\frac{1}{m}\left\|\Delta_{j'}^t\right\|_2\leq\frac{C}{m},\tag{15}$$

其中，$\Delta_{j'}^t$ 是两个批次存在差异处客户端的本地更新。□

备注 2。 敏感度的取值可以决定添加随机噪声时的方差大小。

已有研究 [22] 表明，随机梯度下降和稀疏化满足 Rényi 差分隐私，而 SAM 优化器仅在随机梯度下降的基础上添加扰动，并在训练过程中影响模型。由于 SAM 和稀疏化都在差分隐私处理之前执行，因此二者均满足 Rényi 差分隐私。因此，添加高斯噪声后，我们使用 Rényi 差分隐私计算训练过程中的累积隐私预算 [54]，如下所示：´。

定理 2 (隐私计算)。经过 T 个通信轮次后，累积隐私预算计算如下：

$$\epsilon=\bar{\epsilon}+\frac{(\alpha-1)\log(1-\frac{1}{\alpha})-\log(\alpha)-\log(\delta)}{\alpha-1},\tag{16}$$

其中

$$\bar{\epsilon}=\frac{T}{\alpha-1}\ln\mathbb{E}_{z\sim\mu_0(z)}\left[\left(1-q+\frac{q\mu_1(z)}{\mu_0(z)}\right)^{\alpha}\right],\tag{17}$$

且 q 是客户端选择的采样率；$\mu_0(z)$ 和$\mu_1(z)$ 分别表示 $\mathcal{N}(\mathbf{0},\sigma)$ 的高斯概率密度函数（*PDF*）以及两个高斯分布$q\mathcal{N}(\mathbf{1},\sigma)$的混合分布 $+(1-q)\mathcal{N}(\mathbf{0},\sigma)$；$\sigma$ 是式（11）中的噪声标准差；α 是一个可调变量。

注 3. 较低的采样率 q 可以通过降低隐私预算来增强隐私保证，但由于每个通信轮次中参与的客户端数量减少，也可能降低训练性能。

5.3 收敛分析

下面，我们分析 DP-FedSAM 如何减轻差分隐私带来的负面影响。我们的主要贡献是联合考察本地更新的平均范数 α^t 和客户端间的本地更新一致性 $\tilde{\alpha}^t$ 对收敛速率的影响。此外，我们还在第 6.3 节通过实验验证了这些结果。

定理 3 (收敛界)。在假设 *1-4* 下，本地学习率满足 $\eta=\mathcal{O}(1/L\sqrt{KT})$，令 f^*表示 f 的最小值，即 $f(x)\geq f(x^*)=f^*$ 对所有 $x\in\mathbb{R}^d$ 均成立。给定稀疏率 p，扰动幅度 ρ 与学习率成正比，即 $\rho=\mathcal{O}(1/\sqrt{T})$，且由算法 *1* 生成的输出序列 $\{\mathbf{w}^t\}$，我们有：

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\overline{\alpha}^t\left\|\nabla f\left(\mathbf{w}^t\right)\right\|^2\right]\leq\underbrace{\mathcal{O}\left(\frac{2L(f(\mathbf{w}^1)-f^*)}{\sqrt{KT}}+\frac{L^2\sigma_l^2}{KT^2}\right)}_{\textit{From FedSAM}}+\underbrace{\underbrace{\mathcal{O}\left(\frac{\sum_{t=1}^{T}(\overline{\alpha}^t\sigma_g^2+\tilde{\alpha}^tL^2)}{T^2}\right)}_{\textit{Clipping}}+\underbrace{\mathcal{O}\left(\frac{L^2\sqrt{T}\sigma^2C^2pd}{m^2\sqrt{K}}\right)}_{\textit{Adding noise}}}_{\textit{From operations for DP}}$$

其中

$$\overline{\alpha}^t:=\frac{1}{M}\sum_{i=1}^{M}\alpha_i^t\quad\textit{and}\quad\tilde{\alpha}^t:=\frac{1}{M}\sum_{i=1}^{M}|\alpha_i^t-\overline{\alpha_i^t}|,\tag{18}$$

其中 $\alpha_i^t=\min(1,\frac{C}{\eta\|\sum_{k=0}^{K-1}\tilde{\mathbf{g}}_i^{t,k}(i)\|_2})$。请注意，$\alpha^t$和 $\tilde{\alpha}^t$ 分别衡量了在 *DP-FedSAM* 中进行裁剪和添加噪声操作之前本地更新的平均范数以及客户端间本地更新一致性。

注 4. 与先前的工作 $\mathcal{O}\left(\frac{1}{\sqrt{KT}}+\frac{6K\sigma_g^2+\sigma_l^2}{T}+\frac{B^2\sum_{t=1}^{T}(\alpha^t+\tilde{\alpha}^t}{T}+\frac{L^2\sqrt{T}\sigma^2C^2pd}{m^2\sqrt{K}}\right)$相比，所提出的算法在一般非凸情形下能够获得更紧的界：在 *[21]* 中以及在 *[22]*, 中，我们的界分别降低了本地和全局方差σ_l^2、σ_g^2的影响。此外，我们首次从理论上分析了本地更新的平均范数 α^t 和客户端间本地更新不一致性 $\tilde{\alpha}^t$ 对收敛的影响。与先前的工作 *[21]* 相比，由于采用了本地 *SAM* 优化器，α^t 和 $\tilde{\alpha}^t$ 在收敛时带来的负面影响也显著减轻。这意味着我们能够有效缓解差分隐私中裁剪操作造成的性能下降，并在对称噪声下实现更好的性能。我们还在多个真实世界数据集上通过实验证实了这一理论结果（见第 *6.2* 节和第 *6.3* 节）。

5.4 泛化分析

为了基于先前的理论工作构建我们算法的泛化上界 [50]，我们首先证明每个通信轮次中的联邦学习训练过程都满足具有最大散度的差分隐私。请注意，这一差分隐私保证不同于联邦学习中的客户端级差分隐私分析（见定理 2），因为我们遵循 [50]将迭代式联邦学习方法视为迭代式机器学习算法。下面，我们给出算法 1 在迭代通信轮次中的$(\tilde{\varepsilon},\frac{m}{N}\delta)$-差分隐私。

定理 *4*（迭代通信轮次中的（$\tilde{\varepsilon},\frac{m}{N}\delta$）-差分隐私）。假设某种联邦学习方法有 T 个通信轮次：$\{w^t(S)\}_{t=1}^T$，其中 w^t 为全局模型。令 $v=\frac{1}{\|S_{\mathcal{I}}\|}\sum_{z\in S_{\mathcal{I}}}g(z,w^t)-\frac{1}{\|S'_{\mathcal{I}}\|}\sum_{z\in S'_{\mathcal{I}}}g(z,w^t)$，其中 $g(z,w^t)$ 是数据 z 处的梯度值，且 $S_{\mathcal{I}},S'_{\mathcal{I}}(\mathcal{I}$ 是一个小批量数据

表1 所有对比方法在 IID 和非独立同分布设置下两种数据的平均训练准确率（%）和测试准确率（%）。

Task	算法	狄利克雷分布0.3		狄利克雷分布0.6		IID	
		训练	验证	训练	验证	训练	验证
EMNIST	DP-FedAvg	99.28±0.02	73.10±0.16	99.55±0.02	82.20±0.35	99.66±0.40	81.90±0.86
	Fed-SMP-随机$_k$	99.24±0.02	73.72±0.53	99.71±0.01	82.18±0.73	99.71±0.61	84.16±0.83
	Fed-SMP-顶部$_k$	99.31±0.04	75.75±0.35	99.72±0.02	83.41±0.91	99.73±0.40	83.32±0.52
	DP-FedAvg-模糊	99.12±0.02	73.71±0.02	99.66±0.00	83.20±0.01	99.67±0.03	82.92±0.49
	DP-FedAvg-模糊处理	99.63±0.08	76.25±0.35	99.72±0.02	83.41±0.91	99.74±0.45	82.92±0.49
	DP-FedSAM	96.28±0.64	76.81±0.81	95.07±0.45	84.32±0.19	95.61±0.94	85.90±0.72
	DP-FedSAM-顶部$_k$	94.77±0.11	**77.27±0.67**	95.87±1.52	**84.80±0.60**	96.12±0.85	**87.70±0.83**
CIFAR-10	DP-FedAvg	93.65±0.47	47.98±0.24	93.65±0.42	50.05±0.47	93.65±0.15	50.90±0.86
	Fed-SMP-随机$_k$	95.46±0.43	48.14±0.12	95.36±0.06	51.33±0.36	95.36±0.06	50.61±0.20
	Fed-SMP-顶端$_k$	95.49±0.14	49.93±2.29	95.49±0.09	54.11±0.83	95.49±0.10	53.30±0.45
	DP-FedAvg-模糊	95.47±0.12	47.66±0.01	99.66±0.42	51.05±0.01	94.50±0.05	52.56±0.47
	DP-FedAvg-模糊了	96.79±0.51	51.23±0.66	99.72±0.09	54.11±0.83	96.45±0.30	53.48±0.76
	DP-FedSAM	90.38±0.90	53.92±0.55	90.83±0.15	54.14±0.60	90.83±0.16	55.58±0.50
	DP-FedSAM-顶部$_k$	93.25±0.60	**54.85±0.86**	92.60±0.65	**57.00±0.69**	91.52±0.11	**58.82±0.51**
CIFAR-100	DP-FedAvg	91.14±0.16	16.10±0.71	92.33±0.08	15.92±0.39	94.01±0.10	17.47±0.47
	Fed-SMP-随机$_k$	90.70±0.01	17.25±0.16	92.28±0.32	17.50±0.19	94.31±0.02	17.68±0.44
	Fed-SMP-顶部$_k$	92.58±0.24	18.58±0.25	93.51±0.11	18.07±0.09	95.06±0.05	19.09±0.56
	DP-FedAvg-模糊	91.27±0.01	17.03±0.09	92.33±0.03	17.92±0.01	94.01±0.04	18.47±0.02
	DP-FedAvg-模糊	92.98±0.24	18.98±0.25	94.01±0.11	18.27±0.19	95.46±0.05	19.59±0.06
	DP-FedSAM	82.19±0.01	18.88±0.31	85.47±0.13	19.09±0.15	87.12±0.37	20.64±0.48
	DP-FedSAM-顶部$_k$	84.49±0.24	**20.85±0.63**	88.23±0.23	**21.24±0.69**	89.86±0.21	**22.30±0.05**

是两个相邻的样本集。对于任意第 t 个通信轮次，我们有 （$\tilde{\varepsilon}$，δ）-差分隐私保证，其中

$$\tilde{\varepsilon} = \log\left(\frac{N-m}{N} + \frac{m}{N}\exp\left(\frac{L\rho\sqrt{2\log\frac{1}{\delta}}}{\sigma C d} + \left(\frac{\sqrt{2}L\rho}{\sigma C d}\right)^2\right)\right),$$
$$\delta = \min_{1\le t\le T}\exp\left(-\frac{\sqrt{2}t\|\mathbf{v}\|\sigma C d}{m}\sqrt{\log\frac{1}{\tilde{\delta}}}\right)\mathbb{E}(e^{t\langle \mathbf{v}, W'\rangle}). \tag{19}$$

在上述内容中，$\|\mathbf{v}\| \le \frac{2L\rho}{m}$ 和 $0 < \tilde{\delta} \le 1$, σ^2 表示高斯噪声方差，而 C 和 d 分别表示本地更新的裁剪阈值以及$\mathbf{w}^{t,k}$的维度。

注释 5. $\tilde{\varepsilon}$ 和 δ 主要由参与客户端的数量 m、裁剪阈值 C、本地更新的维度 d、差分隐私噪声的标准差 σ、L-光滑系数 L以及 *SAM* 优化器中的扰动半径 ρ 决定。一般而言，$\tilde{\varepsilon}$ 和 δ的值越小，隐私保护能力 *[9]*越强。

设迭代式机器学习算法 $\mathcal{A}$ （算法 1）在遵循数据分布 $\mathcal{D}$ 的训练样本集 S 上学习假设 $\mathcal{A}(S)$。算法 1 的期望风险 $\hat{}\mathcal{R}_{\mathcal{D}}(\mathcal{A}(S))$ 和经验风险 $\mathcal{R}_{\mathcal{S}}(\mathcal{A}(S))$ 定义如下:

$$\mathcal{R}_{\mathcal{D}}(\mathcal{A}(S)) = \mathbb{E}_{z\sim\mathcal{D}}\ell(\mathcal{A}(S), z) = \max_{\|\delta\|_2\le\rho}\mathbb{E}_{z\sim\mathcal{D}}\ell(\mathbf{w}+\delta_z; z),$$
$$\hat{\mathcal{R}}_S(\mathcal{A}(S)) = \frac{1}{N}\sum_{i=1}^{N}\ell(\mathcal{A}(S), z_i) = \max_{\|\delta\|_2\le\rho}\frac{1}{N}\sum_{i=1}^{N}\ell(\mathbf{w}+\delta_{z_i}; z_i).$$

其中，ℓ： $\mathcal{H}\times\mathcal{Z}\to\mathbb{R}^+$ 是损失函数，$\delta = \rho\ ^{g\ (z,\ \mathbf{w})\|g\ (z,\ \mathbf{w})\|_2}$，且 N 是训练样本量。下面给出泛化界。

定理 5 (泛化界**)**。 在定理 *4* 和 ˜的条件下，假设损失函数 $\|\mathcal{L}\|_\infty \le 1$, $0 < \delta \le 1$ 是任意的正实数常数。于是，对于数据空间 $\mathcal{Z}$上的任意数据分布 $\mathcal{D}$，我们有以下不等式:

$$\mathbb{P}\left[\left|\hat{\mathcal{R}}_S(\mathcal{A}(S)) - \mathcal{R}_{\mathcal{D}}(\mathcal{A}(S))\right| < 4\varepsilon'\right] > 1 - \frac{2e^{-1.7\varepsilon'}\delta'}{\varepsilon'}\ln\left(\frac{2}{\varepsilon'}\right), \tag{20}$$

where $\varepsilon' = \sqrt{2T\log\left(\frac{1}{\tilde{\delta}}\right)\tilde{\varepsilon}^2 + T\tilde{\varepsilon}\frac{e^{\tilde{\varepsilon}}-1}{e^{\tilde{\varepsilon}}+1}}$ *and*

$$\delta' = e^{-\frac{\varepsilon'+T\tilde{\varepsilon}}{2}}\left(\frac{1}{1+e^{\tilde{\varepsilon}}}\left(\frac{2T\tilde{\varepsilon}}{T\tilde{\varepsilon}-\varepsilon'}\right)\right)^T\left(\frac{T\tilde{\varepsilon}+\varepsilon'}{T\tilde{\varepsilon}-\varepsilon'}\right)^{-\frac{\varepsilon'+T\tilde{\varepsilon}}{2\tilde{\varepsilon}}}$$
$$+ 2 - \left(1 - e^{\tilde{\varepsilon}}\frac{m\delta}{N(1+e^{\tilde{\varepsilon}})}\right)^{\left\lceil\frac{\varepsilon'}{\tilde{\varepsilon}}\right\rceil}\left(1 - \frac{m\delta}{N(1+e^{\tilde{\varepsilon}})}\right)^{T-\left\lceil\frac{\varepsilon'}{\tilde{\varepsilon}}\right\rceil}$$
$$- \left(1 - \frac{m\delta}{N(1+e^{\tilde{\varepsilon}})}\right)^T,$$

注意，这个概率不等式可以视为一个泛化界 *[50]*，，其中 T 是全局训练迭代次数（通信轮次），m 是全局批量大小（每轮参与的客户端数），而$N \ge \frac{2}{\varepsilon'^2}\ln\left(\frac{16}{e^{-\varepsilon'}\delta'}\right)$ 是每次本地迭代中的训练样本量。

备注6. 上述界表明，定理 *4* 得到的 $\tilde{\varepsilon}$ 和 δ 越小，泛化性能越好。因此，当差分隐私噪声的标准差 σ、本地更新的维度 d 和裁剪阈值 C 越大，且参与客户端的数量m、L-光滑系数 L 以及 *SAM* 优化器中的扰动半径ρ 越小时，泛化界越小。这意味着训练误差与测试误差之间的差距更小，从而实现更好的泛化性能。

6 实验

在本节中，我们开展了大量实验，以验证 DP-FedSAM 和 DP-FedSAM-top$_k$ 的有效性。

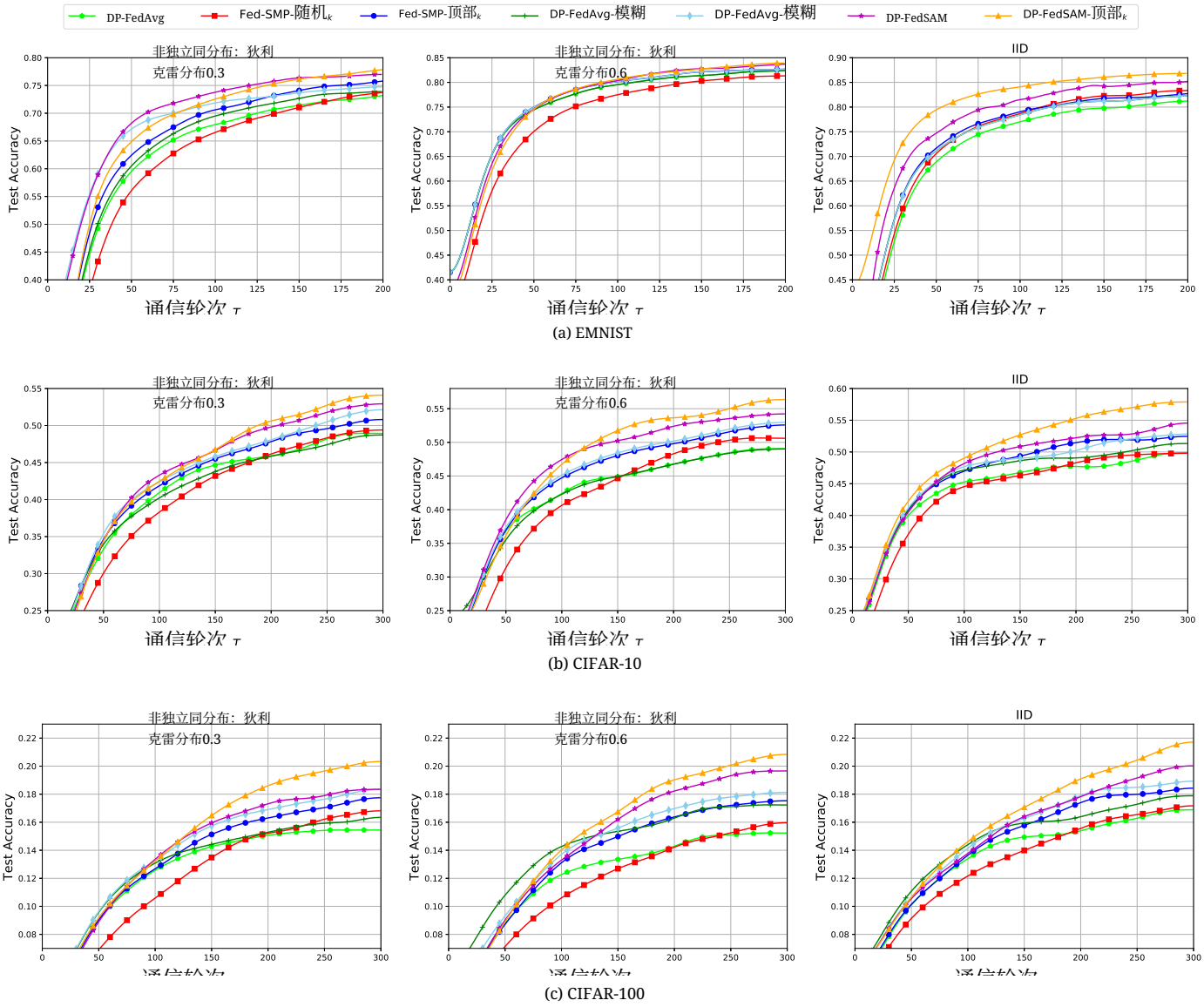


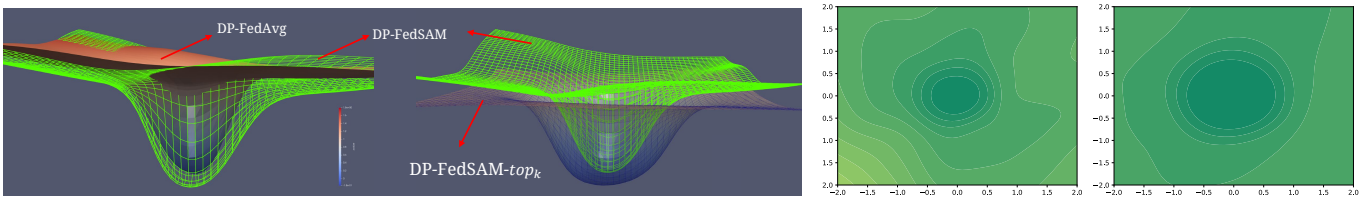
图 2. 所有对比方法在对称噪声下于*EMNIST*、*CIFAR-10* 和*CIFAR-100* 上的平均测试准确率。

6.1 实验设置

数据集与数据划分。 我们在三个数据集上评估 DP-FedSAM 的效果，包括**EMNIST** [59],**CIFAR-10**和 **CIFAR-100** [27],，涵盖独立同分布和非独立同分布设置。EMNIST [59]是一个包含 62 个类别的图像分类数据集，我们使用其中 20% 的数据，共计 88,800 个训练样本和 14,800 个验证样本。CIFAR-10 和 CIFAR-100 [27] 均包含 60,000 张图像，分为 50,000 个训练样本和 10,000 个验证样本。与包含 10 个唯一标签的 CIFAR-10 相比，CIFAR-100 的标签划分更细，包含 100 个唯一标签。此外，默认情况下，我们通过狄利克雷分配将这些数据集分配给 500 个客户端。此外，我们使用 Dir 划分 [60] 来模拟联邦客户端之间的非独立同分布设置：根据基于参数为 $\alpha = 0.3$ 和 $\alpha = 0.6$的狄利克雷分布 Dir(α) 所确定的标签比例，从原始数据集中采样，为每个客户端创建本地数据。

基线方法。 我们关注能够确保客户端-级差分隐私的差分隐私联邦学习方法。因此，我们考虑以下差分隐私联邦学习基线方法：**DP-FedAvg** [14] 通过直接对本地更新采用高斯机制来确保客户端级差分隐私保证；**DP-FedAvg-blur** [21] 在 DP-FedAvg 的基础上加入正则化方法（BLUR）；**DP-FedAvg-blurs** [21]采用本地更新稀疏化（LUS）和 BLUR 来提升 DP-FedAvg 的性能；**Fed-SMP-**rand$_k$ 和**Fed-SMP-**top$_k$ [22] 分别利用随机稀疏化和 top$_k$稀疏化技术，以降低差分隐私噪声对模型准确率的影响。

配置。 对于 EMNIST，我们使用一个简单的卷积神经网络模型，并训练 200 个通信轮次。对于 CIFAR-10 和 CIFAR-100 数据集，我们使用 ResNet-18 [24] 骨干网络，并训练 300 个通信轮次。在所有实验中，我们将客户端数量 M 设为 500。对于 EMNIST 数据集，我们将小批量大小设为 32，并使用一个简单的卷积神经网络模型进行训练，该模型包含两个带有 5×5 卷积核的卷积层、最大池化层，以及一个 512 单元的全连接层。对于 CIFAR-10 和 CIFAR-100 数据集，我们将



(a) 损失景观：左侧为 DP-FedAvg 与 DP-FedSAM 的对比，右侧为 DP-FedSAM 与 DP-FedSAM-top$_k$ 的对比。(b) DP-FedSAM 和 DP-FedSAM- top$_k$ 的损失曲面等高线分别位于左侧和右侧。

图3. 损失景观（a）和曲面等高线（b）的比较。与图1（a）左侧相同设置下的 DP-FedAvg 相比，DP-FedSAM 的损失景观更平坦，泛化能力更强（平坦极小值，见图3（a）左侧），权重扰动鲁棒性也更高（见图3（b）左侧）。与此同时，与图3（a）和（b）右侧的 DP-FedSAM 相比，DP-FedSAM-top$_k$ 也具有类似优势。

小批量大小设为 50，并使用 ResNet-18 [24]架构进行训练。对于每种算法和每个数据集，学习率通过在 $\{10^{-0.5}, 10^{-1}, 10^{-1.5}, 10^{-2}\}$范围内进行网格搜索来确定。权重扰动比率 ρ 通过在$\{0.01, 0.1, 0.3, 0.5, 0.7, 1.0\}$范围内进行网格搜索来确定。对于所有采用稀疏化技术的方法，稀疏率设为 $p = 0.4$。客户端的默认样本比例 q 为 0.1。本地学习率 η 设为 0.1，衰减率为 0.0005 ，动量为0.5，训练轮数为 30。隐私参数方面，噪声乘数 σ 设为 0.95 ，隐私失败概率为 $\delta = \frac{1}{M}$。裁剪阈值C通过在 $\{0.1, 0.2, 0.4, 0.6, 0.8\}$范围内进行网格搜索来选择；我们发现，当 $C \geq 0.6$ 时，EMNIST 上可能出现梯度爆炸现象，而 $C = 0.2$ 在三个数据集上的表现最佳。权重扰动比率设为 $\rho = 0.5$。每项实验进行 3 次试验，并报告每项实验中平均测试准确率的最佳值。

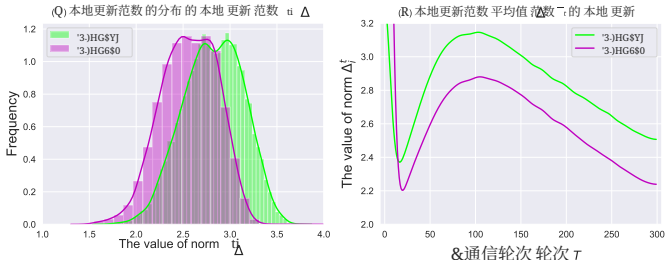


图 4. 本地更新的范数分布及平均范数。

6.2 实验评估

总体性能比较。在表1和图2中，我们在两种设置下，将 DP-FedSAM 和 DP-FedSAM-top$_k$ 应用于 EM-NIST、CIFAR-10 和 CIFAR-100，并与所有基线方法进行了比较。显然，在对称噪声下，就准确率和泛化而言，我们提出的算法始终优于其他基线方法。这表明，我们显著提升了性能，并在差分隐私联邦学习中实现了性能与隐私之间更好的权衡。例如，在 IID 设置下，DP-FedSAM 和85.90% 以及87.70% DP-FedSAM-top$_k$ 在 EMNIST 上的平均测试准确率分别为上述数值，均优于其他基线方法。同时，训练准确率与测试准确率之间的差异为 9。

71% 在 DP-FedSAM 中，以及8.40% 在 DP-FedSAM-top$_k$ 中；相比之下，DP-FedAvg 中为17.74% ，Fed-SMP-top$_k$ 中为16.41% 。因此，这表明我们的算法显著缓解了差分隐私导致的性能下降问题。

非独立同分布程度的影响。在表1所示的不同参与情况下进行实验，我们进一步展示了所提出算法在泛化方面的鲁棒性。我们将本地客户端的异构数据分布设置为不同程度，包括独立同分布、狄利克雷分布0.6和狄利克雷分布0.3，使全局模型的训练更具挑战性。在 EMNIST 上，随着非独立同分布程度降低，DP-FedSAM 的泛化性能优于 DP-FedAvg，且 DP-FedSAM 的训练准确率与测试准确率之间的差异(19.47%10.75%9.71%) 较小。DP-FedAvg 中的差异(26.18%17.35%17.74%)。同样，DP-FedSAM-top$_k$(17.50%11.07%8.40%) 中的差异也小于 Fed- SMP-top$_k$ (23.56%16.31%16.41%)中的差异。这些观察结果证实，与基线方法相比，我们的算法在不同程度的异构数据下具有更强的鲁棒性。

6.3 联邦学习中结合 SAM 的差分隐私讨论

在本小节中，我们从本地更新范数以及损失景观和等高线的可视化两个方面，讨论 SAM 如何减轻差分隐私带来的负面影响。同时，我们还研究了在不同隐私预算 ϵ 下结合 SAM 的训练性能，并与所有基线方法进行比较。这些实验在 CIFAR-10 上使用 ResNet-18 [24] 和狄利克雷 $\alpha = 0.6$进行。

损失景观与等高线。为了可视化极小值的结构，并研究与 DP-FedAvg 相比，DP-FedSAM 和 DP-FedSAM-top$_k$ ($p = 0.4$) 对差分隐私噪声的鲁棒性，我们在图3中展示了损失景观和曲面等高线 [26] 。显然，在左侧的图1(a)中，DP-FedSAM 的极小值更平坦，对差分隐私噪声的鲁棒性也优于 DP-FedAvg。此外，在右侧的图1(a)和(b)中，DP-FedSAM-top$_k$ 的极小值也比 DP-FedSAM 更平坦，对差分隐私噪声的鲁棒性也更强。这表明，在差分隐私联邦学习中使用 SAM 本地优化器和本地更新稀疏化技术，可以获得更平坦的损失景观和更强的泛化能力。此外，这也表明我们提出的算法具有更好的泛化能力，并使训练过程更适用于差分隐私联邦学习场景。

本地更新的范数。为验证关于缓解本地更新范数负面影响的理论结果，我们在裁剪阈值为 $C = 0.2$ 的情况下，对 DP-FedSAM 和 DP-FedAvg 进行了实验，如图 4 所示。我们展示了范数 Δ_i^t 分布和平均范数 Δ^t —

表2 不同隐私预算下的性能比较 ϵ 在 *CIFAR-10* 和 *CIFAR-100* 上。

Task	算法	不同隐私预算下的平均测试准确率（%） ϵ			
		$\epsilon = 4$	$\epsilon = 6$	$\epsilon = 8$	$\epsilon = 10$
CIFAR-10	DP-FedAvg	38.23 ± 0.15	43.87 ± 0.62	46.74 ± 0.03	49.06 ± 0.49
	Fed-SMP-随机$_k$	33.78 ± 0.92	42.21 ± 0.21	48.20 ± 0.05	50.62 ± 0.14
	Fed-SMP-顶端$_k$	38.99 ±0.50	46.24 ± 0.80	49.78 ± 0.78	52.51 ± 0.83
	DP-FedAvg-模糊	38.23 ± 0.70	43.93 ± 0.48	46.74 ± 0.92	49.06 ± 0.13
	DP-FedAvg-模糊	39.39 ±0.43	46.64 ± 0.36	50.18 ± 0.27	52.91 ± 0.57
	DP-FedSAM	**39.89± 0.17**	47.92 ± 0.23	51.30 ± 0.95	53.18 ± 0.40
	DP-FedSAM-顶部$_k$	38.96 ± 0.61	**49.17 ± 0.15**	**53.64 ± 0.12**	**56.36 ± 0.36**
CIFAR-100	DP-FedAvg	9.65 ± 0.34	12.81 ± 0.29	14.30 ± 0.05	15.23 ± 0.24
	Fed-SMP-随机$_k$	7.95 ± 0.91	11.51 ± 0.18	14.20 ± 0.90	15.92 ± 1.03
	Fed-SMP-顶端$_k$	9.90 ± 0.89	14.22 ± 0.82	16.57 ± 0.75	17.71 ± 0.36
	DP-FedAvg-模糊	9.90 ± 0.89	14.22 ± 0.82	16.57 ± 0.75	17.71 ± 0.36
	DP-FedAvg-模糊	9.65 ± 0.34	12.81 ± 0.29	14.30 ± 0.05	15.23 ± 0.24
	DP-FedSAM	10.03 ± 0.63	14.46 ± 1.21	18.20 ± 1.34	19.65 ± 0.80
	差分隐私-FedSAM-顶部$_k$	**10.08 ± 0.68**	**15.26 ± 0.78**	**18.99 ± 1.38**	**20.79 ± 1.07**

表 3 不同稀疏率下的性能比较 p。

Task	性能	不同稀疏率 p					
		$p = 1.0$	$p = 0.1$	$p = 0.2$	$p = 0.4$	$p = 0.6$	$p = 0.8$
CIFAR-10	训练 (%)	90.83 ± 0.15	91.35 ± 0.66	92.83 ± 0.09	92.60 ± 0.65	92.43 ± 0.37	91.89 ± 0.17
	验证 (%)	54.14 ± 0.60	55.05 ± 0.74	56.12 ± 0.41	**57.00 ± 0.69**	56.42 ± 0.13	56.14 ± 0.32
	增益（%）比较 与 $p = 1.0$	0.00	0.91 ± 0.14	1.98 ± 0.19	**2.86 ± 0.09**	2.28 ± 0.47	2.00 ± 0.28
CIFAR-100	训练（%）	85.47 ± 0.13	86.29 ± 0.26	87.49 ± 0.30	88.23 ± 0.23	86.41 ± 0.37	85.66 ± 0.14
	验证 (%)	19.09 ± 0.15	20.38 ± 0.24	20.62 ± 0.85	**21.24 ± 0.69**	20.55 ± 0.84	19.79 ± 0.28
	增益（%）对比 与 $p = 1.0$	0.00	1.29 ± 0.09	1.53 ± 0.80	**2.15 ± 0.54**	1.46 ± 0.69	0.70 ± 0.13

通信轮次中裁剪前的本地更新范数。与 DP-FedAvg 相比，根据图 4 (a)，我们的方案中大多数范数分布在较小的数值范围内，这意味着裁剪操作丢弃的信息更少。同时，如图 4 (b) 所示，平均范数 Δ^t小于 DP-FedAvg。这些观察结果也与第5节中的理论结果一致。

6.4 不同隐私预算下的性能

表 2 展示了在不同隐私预算 ϵ 下，CIFAR-10 和 CIFAR-100 数据集上的测试准确率。具体而言，在 CIFAR-10 上，DP-FedSAM 和 DP-FedSAM-top$_k$的表现显著优于 DP-FedAvg 和 Fed-SMP-top$_k$，在相同的 ϵ下分别高出1% $\sim$ 4% 和 3% $\sim$ 4%。在复杂度更高的 CIFAR-100 数据集上，我们的算法也取得了显著的性能提升。也就是说，DP-FedSAM 和 DP-FedSAM-top$_k$ 显著提升了 DP-FedAvg 和 Fed-SMP-top$_k$ 的准确率，在相同的 ϵ下分别高出1% $\sim$ 4%和1% $\sim$ 3%。此外，随着隐私预算 ϵ 增加，测试准确率往往也会提高，这表明需要在训练性能和隐私之间保持适当的平衡。

6.5 本地更新稀疏化讨论

在表3中，我们研究了不同稀疏率p 对 CIFAR-10 和 CIFAR-100 数据集上模型性能提升的影响。具体而言，我们可以看到，当 p 的值从0.1 增至0.4时，验证准确率（测试准确率）有所提升；随后，当 p 的值从0.4 增至1.0时，准确率则有所下降。这表明稀疏率 p存在一个最优值。原因在于稀疏化操作引入的信息误差与差分隐私操作导致的随机噪声幅度之间需要取得平衡。当 p 的值足够小时，信息误差较大；由于大量添加了噪声的参数被丢弃，因此相较于差分隐私中的随机噪声，这种误差对性能造成的不利影响更为显著。与此同时，当 p 的值足够大时，随机噪声相较于信息误差，对性能造成的不利影响更为显著。因此，当 p 的值为0.4时，信息误差与随机噪声幅度之间达到了恰当的权衡。请注意，当 $p = 1.0$时，DP-FedSAM 与 DP-FedSAM-top$_k$ 相同。

6.6 消融研究

在本小节中，我们验证 DP-FedSAM 中各个组件和超参数的作用。所有消融研究均在 EMNIST 数据集上使用狄利克雷分布0.6进行。

扰动权重 ρ。随着通信轮次 T 增加，添加的扰动会不断累积，因此扰动权重 ρ 会影响性能。

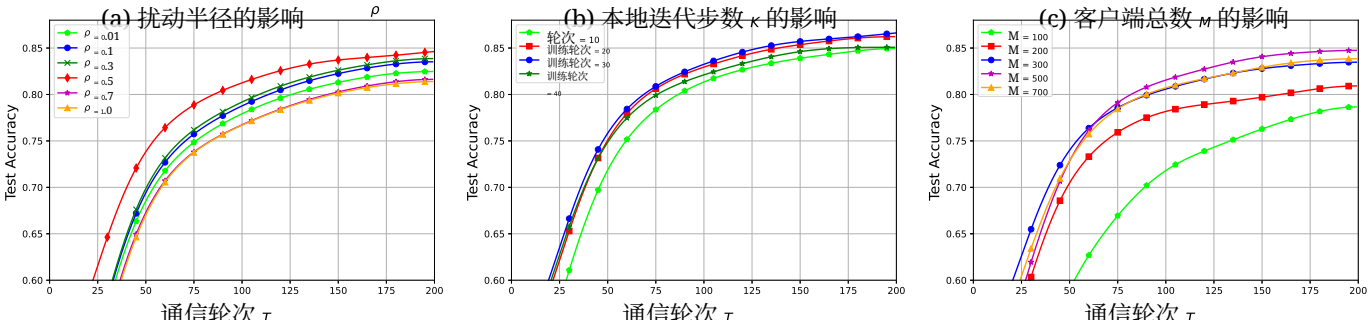


图5. 超参数的影响：扰动半径 ρ、本地迭代步数 K、客户端总数 M。

表4 平均训练准确率和测试准确率。

算法	训练（%）	验证 (%)	差异值（%）
DP-FedAvg	99.55±0.02	82.20± 0.35	17.35±0.32
DP-FedSAM	95.07±0.45	84.32±0.19 ↑	10.75±0.26 ↓
Fed-SMP-顶端$_k$	99.72±0.02	83.41 ± 0.91	16.31± 0.89
差分隐私-FedSAM-顶部$_k$	95.87±0.52	84.80±0.60 ↑	11.07±0.08 ↓

为了为我们的算法选择合适的值，我们在图 5 (a) 中针对不同的扰动半径进行了实验，取值为$\{0.01, 0.1, 0.3, 0.5, 0.7, 1.0\}$ 时，$\rho = 0.5$，我们获得了更好的收敛性和性能。

本地迭代步数 K。较大的本地迭代步数有助于加快此前差分隐私联邦学习工作中的收敛 [21]，并具有理论保证。为探究采用更大的 K 对T 的加速效果，我们固定总批量大小并改变本地训练轮数。在图 5 (b) 中，采用更大的 K（即使用更大的轮数值）时，我们的算法可以加快定理 3 中的收敛。然而，随着 K 过大（例如，轮数为 $= 40$），裁剪对训练的不利影响会增大。

客户端数量 M。 我们在图 5 (c) 中比较了不同客户端参与数量$m = \{100, 200, 300, 500, 700\}$ 下的性能。总体而言，在相同设置下，较小的 m 值往往会导致性能变差，这是因为差分隐私噪声的方差较大 σ^2C^2/m 。同时，当 m 过大（例如 $M = 700$）时，随着本地数据量减少，性能可能会下降。

SAM 的效果。如表4所示，采用 SAM 优化器时，与 DP-FedAvg 和 Fed-SMP-top$_k$ 相比，DP-FedSAM 和 DP-FedSAM-top$_k$ 能够显著提升性能并改善泛化。

7 结论

本文聚焦于因模型信息丢失和模型不一致性加剧而导致的严重性能下降这一难题。我们的主要贡献在于，首次从优化器视角缓解这一问题，并提出了两种新颖有效的框架 DP-FedSAM 和 DP-FedSAM-top$_k$，以获得更平坦的损失景观和更好的泛化。同时，我们详细分析了 SAM 如何减轻差分隐私带来的不利影响并实现更紧的收敛界，还给出了敏感度、隐私、以及泛化分析。此外，我们研究了 SAM 本地优化器与稀疏化策略的结合，该结合兼具更平坦的损失景观和更好泛化的优势。我们还首次分析了本地更新的平均范数与客户端间本地更新一致性对训练的综合影响，并进行了相应的实验评估。最后，实证结果也验证了我们的方法在多个真实世界数据集上优于先进的基线方法。

参考文献

[1] Y. Shi、Y. Liu、K. Wei、L. Shen、X. Wang 和 D. Tao，“让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年。[2] T. Li、A. K. Sahu、A. Talwalkar 和 V. Smith，“联邦学习：挑战、方法与未来方向”，*IEEE*信号处理杂志，第 50–60 页，2020年。[3] P. Kairouz、H. B. McMahan、B. Avent、A. Bellet、M. Bennis、A. N. Bhagoji、K. Bonawitz、Z. Charles、G. Cormode、R. Cummings、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。[4] M. Fredrikson、S. Jha 和 T. Ristenpart，“利用置信度信息的模型反演攻击及基本应对措施”，载于*ACM SIGSAC*计算机与通信安全会议（*CCS*）论文集，第 1322–1333 页，2015年。[5] R. Shokri、M. Stronati、C. Song 和 V. Shmatikov，“针对机器学习模型的成员推断攻击”，载于*IEEE*安全与隐私研讨会（*SP*）论文集，第 3–18 页，2017年。[6] L. Melis、C. Song、E. De Cristofaro 和 V. Shmatikov，“利用协作学习中的非预期特征泄露”，载于*IEEE*安全与隐私研讨会（*SP*）论文集，第 691–706 页，2019年。[7] M. Nasr、R. Shokri 和 A. Houmansadr，“深度学习的全面隐私分析：针对集中式和联邦学习的被动与主动白盒推理攻击”，载于*IEEE*安全与隐私研讨会（*SP*）论文集，第 739–753 页，2019年。[8] L. Zhang、L. Shen、L. Ding、D. Tao 和 L.-Y. Duan，“通过无数据知识蒸馏微调全局模型，以应对非独立同分布联邦学习”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 10174–10183 页，2022年。[9] C. Dwork 和 A. Roth，“差分隐私的算法基础”，理论计算机科学基础与趋势，第 211–407 页，2014年8月。[10] N. Agarwal、A. T. Suresh、F. X. Yu、S. Kumar 和 B. McMahan，“cpSGD：通信高效且具差分隐私的分布式随机梯度下降”，载于神经信息处理系统年会（*NeurIPS*）论文集，第 7575–7586 页，2018年12月。[11] R. Hu、Y. Gong 和 Y. Guo，“采用稀疏化增强隐私和自适应优化的联邦学习”，载于第三十届国际人工智能联合会议（*IJCAI*）论文集，第 1463–1469 页，2021年8月。[12] L. Sun 和 L. Lyu，“采用无噪声差分隐私的联邦模型蒸馏”，载于第三十届国际人工智能联合会议（*IJCAI*）论文集，第 1563–1570 页，2021年8月。

[13] L. Sun、J. Qian 和 X. Chen，“LDP-FL：采用本地差分隐私的联邦学习实用私有聚合”，载于第三十届国际人工智能联合会议（IJCAI）论文集，第 1571–1578 页，2021年8月。[14] H. B. McMahan、D. Ramage、K. Talwar 和 L. Zhang，“学习差分隐私循环语言模型”，载于国际学习表征会议（ICLR）论文集，2018年4月。[15] R. C. Geyer、T. Klein 和 M. Nabi，“差分隐私联邦学习：客户端层面的视角”，Co RR，2017年8月。[16] P. Kairouz、Z. Liu 和 T. Steinke，“用于安全聚合联邦学习的分布式离散高斯机制”，载于国际机器学习会议（ICML）论文集，第 5201–5212 页，2021年7月。[17] K. Wei、J. Li、M. Ding、C. Ma、H. Su、B. Zhang 和 H. V. Poor，“用户级隐私保护联邦学习：分析与性能优化”，《IEEE移动计算汇刊》，第 21 卷，第 9 期，第 3388–3401 页，2022年。[18] R. Hu、Y. Gong 和 Y. Guo，“采用稀疏化模型扰动的联邦学习：提升客户端级差分隐私下的准确率”，CoRR，2022年。[19] A. Cheng、P. Wang、X. S. Zhang 和 J. Cheng，“采用局部正则化和稀疏化的差分隐私联邦学习”，CoRR，2022年。[20] R. C. Geyer、T. Klein 和 M. Nabi，“差分隐私联邦学习：客户端层面的视角”，arXiv预印本 arXiv:1712.07557，2017年。[21] A. Cheng、P. Wang、X. S. Zhang 和 J. Cheng，“采用局部正则化和稀疏化的差分隐私联邦学习”，载于IEEE/CVF计算机视觉与模式识别会议论文集，2022年。[22] R. Hu、Y. Gong 和 Y. Guo，“采用稀疏化模型扰动的联邦学习：提升客户端级差分隐私下的准确率”，arXiv预印本 arXiv:2202.07178，2022年。[23] K. Simonyan 和 A. Zisserman，“用于大规模图像识别的超深卷积网络”，arXiv预印本 arXiv:1409.1556，2014年。[24] K. He、X. Zhang、S. Ren 和 J. Sun，“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。[25] B. McMahan、E. Moore、D. Ramage、S. Hampson 和 B. A. y Arcas，“利用去中心化数据进行深度网络的通信高效学习”，载于国际人工智能与统计会议（AISTATS）论文集，第 54 卷，第 1273–1282 页，2017年4月。[26] H. Li、Z. Xu、G. Taylor、C. Studer 和 T. Goldstein，“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。[27] A. Krizhevsky、G. Hinton、等，“从微型图像中学习多层特征”，2009年。[28] P. Foret、A. Kleiner、H. Mobahi 和 B. Neyshabur，“尖锐度感知最小化：高效提升泛化能力”，载于国际学习表征会议，2021年。[29] H. B. McMahan、D. Ramage、K. Talwar 和 L. Zhang，“学习差分隐私循环语言模型”，载于 ICLR，2018年。[30] P. Kairouz、Z. Liu 和 T. Steinke，“用于安全聚合联邦学习的分布式离散高斯机制”，ArXiv，第 abs/2102.06387 卷，2021年。[31] O. Thakkar、G. Andrew 和 H. B. McMahan，“采用自适应裁剪的差分隐私学习”，ArXiv，第 abs/1905 03871 卷，2019年。。[32] Y. Zhu、X. Yu、Y.-H. Tsai、F. Pittaluga、M. Faraki、M. Chandraker 和 Y.-X. Wang，“用于差分隐私联邦学习的基于投票的方法”，ArXiv，第 abs/2010.04851 卷，2020年。[33] M. Andriushchenko 和 N. Flammarion，“迈向对尖锐度感知最小化的理解”，载于国际机器学习会议（ICML），《机器学习研究论文集》，第 639–668 页，PMLR，2022年。[34] Y. Zhao、H. Zhang 和 X. Hu，“惩罚梯度范数以高效提升深度学习的泛化能力”，载于国际机器学习会议（ICML），第 26982–26992 页，PMLR，2022年。[35] J. Kwon、J. Kim、H. Park 和 I. K. Choi，“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。[36] J. Du、H. Yan、J. Feng、J. T. Zhou、L. Zhen、R. S. M. Goh 和 V. Tan，“用于提升神经网络训练”，载于《国际学习表征会议》，2021年。[37] Y. Liu、S. Mai、X. Chen、C.-J. Hsieh 和 Y. You，“迈向高效且可扩展的尖锐度感知最小化”，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。[38] M. Abbas 、 Q. Xiao 、 L. Chen 、 P. Chen 和 T. Chen，“Sharp-maml：尖锐度感知的模型无关元学习”，载于《国际机器学习会议》（ICML），第 10–32 页，2022年。[39] P. Mi、L. Shen、T. Ren、Y. Zhou、X. Sun、R. Ji 和 D. Tao，“增强尖锐度感知最小化：一种稀疏化扰动方法”，载于《神经信息处理系统进展》（A. H. Oh、A. Agarwal、D. Belgrave 和 K. Cho 编），2022年。[40] Y. Shi、L. Shen、K. Wei、Y. Sun、B. Yuan、X. Wang 和 D. Tao，“提升去中心化联邦学习的模型一致性”，arXiv 预印本 arXiv:2302.04083，2023年。[41] Q. Zhong、L. Ding、L. Shen、P. Mi、J. Liu、B. Du 和 D. Tao，“利用费舍尔掩码改进尖锐度感知最小化，以提升语言模型的泛化能力”，arXiv 预印本 arXiv:2210.05497，2022年。[42] Z. Huang、X. Xia、L. Shen、J. Yu、C. Gong、B. Han 和 T. Liu，“通过最坏情况尖锐度最小化实现抵御损坏的稳健泛化”，[43] Y. Sun、L. Shen、T. Huang、L. Ding 和 D. Tao，“Fedspeed：更大的本地间隔、更少的通信轮次和更高的泛化准确率”，载于《国际学习表征会议》。[44] H. Sun、L. Shen、Q. Zhong、L. Ding、S. Chen、J. Sun、J. Li、G. Sun 和 D. Tao，“Adasam：通过自适应学习率和动量增强尖锐度感知最小化，以训练深度神经网络”，arXiv 预印本 arXiv:2303.00565，2023年。[45] Z. Qu、X. Li、R. Duan、Y. Liu、B. Tang 和 Z. Lu，“通过尖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。[46] D. Caldarola、B. Caputo 和 M. Ciccone，“通过寻找平坦极小值提升联邦学习的泛化能力”，CoRR，第 abs/2203.11834 卷，2022年。[47] J. Zhu 和 M. B. Blaschko，“通过随机稀疏化梯度改进差分隐私随机梯度下降”，arXiv 电子印本，第 arXiv–2112 页，2021年。[48] R. Liu、Y. Cao、M. Yoshikawa 和 H. Chen，“Fedsel：采用前k维选择的本地差分隐私联邦随机梯度下降”，载于《高级应用数据库系统：第25届国际会议，DASFAA 2020，济州，韩国，2020年9月24–27日，论文集，第一部分25》，第 485–501 页，Springer，2020年。[49] I. Mironov，“Rényi 差分隐私”，载于《IEEE计算机安全基础研讨会》（CSF）论文集，第 263–275 页，2017年。[50] F. He、B. Wang 和 D. Tao，“迭代式差分隐私学习算法的更紧泛化界”，载于《人工智能中的不确定性》，第 802–812 页，PMLR，2021年。[51] H. B. McMahan、D. Ramage、K. Talwar 和 L. Zhang，“学习差分隐私循环语言模型”，arXiv 预印本 arXiv:1710.06963，2017年。[52] L. Bottou，“利用随机梯度下降进行大规模机器学习”，载于《COMPSTAT’2010 论文集》，第 177–186 页，Springer，2010年。[53] L. Bottou、F. E. Curtis 和 J. Nocedal，“大规模机器学习的优化方法”，《SIAM评论》，第 60 卷，第 2 期，第 223–311 页，2018年。[54] A. Yousefpour、I. Shilov、A. Sablayrolles、D. Testuggine、K. Prasad、M. Malek、J. Nguyen、S. Gosh、A. Bharadwaj、J. Zhao、G. Cormode 和 I. Mironov，“Opacus：PyTorch 中易于使用的差分隐私库”，载于 NeurIPS《机器学习中的隐私》（PriML）研讨会论文集（线上），2021年12月。[55] T. Sun、D. Li 和 B. Wang，“去中心化联邦平均”，《IEEE模式分析与机器智能汇刊》，2022年。[56] S. Ghadimi 和 G. Lan，“非凸随机规划的随机一阶和零阶方法”，《SIAM优化杂志》，第 23 卷，第 4 期，第 2341–2368 页，2013年。[57] H. Yang、M. Fang 和 J. Liu，“ 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。[58] S. J. Reddi、Z. Charles、M. Zaheer、Z. Garrett、K. Rush、J. Konečný、S. Kumar 和 H. B. McMahan，“自适应联邦优化”，载于《国际学习表征会议》，2021年。

[59] G. Cohen、S. Afshar、J. Tapson 和 A. Van Schaik，“EMNIST：将 MNIST 扩展至手写字母”，发表于 *2017*年国际联合神经网络会议（*IJCNN*），第 2921–2926 页，IEEE，2017年。 [60] T.-M. H. Hsu、H. Qi 和 M. Brown，“衡量非同一数据分布对联邦视觉分类的影响”，*arXiv* 预印本 *arXiv:1909.06335*，2019年。 [61] B. Balle、G. Barthe 和 M. Gaboardi，“ 通过子采样实现隐私放大：基于耦合与散度的紧致分析 ”，发表于 *NeurIPS*，2018 年。

《客户端级差分隐私下联邦学习的更平坦损失景观与更好的泛化》补充材料

附录 A 主要证明

A.1 符号与预备知识

$S = \{(x_1, y_1) \ldots, (x_N, y_N) | x_i \in \mathcal{X} \subset \mathbb{R}^{d_X}, y_i \in \mathcal{Y} \subset \mathbb{R}^{d_Y}, i = 1, \ldots, N\}$是一个训练样本集，其中，$x_i$是第 i个特征，y_i是相应的标签，d_X 和 d_Y 分别是特征和标签的维度。为简洁起见，我们定义 $z_i = (x_i, y_i)$。我们还定义随机变量 $Z = (X, Y)$，使得所有 $z_i = (x_i, y_i)$都是变量 $Z = (X, Y) \in \mathcal{Z}$，$Z \sim \mathcal{D}$的独立同分布（i.i.d.）观测，其中 $\mathcal{D}$ 是数据分布。对于机器学习算法 $\mathcal{A}$，它学习一个假设 $\mathcal{A}(S)$，$\mathcal{A}(S) \in \mathcal{H} \subset \mathcal{Y}^{\mathcal{X}} = \{f : \mathcal{X} \to \mathcal{Y}\}$。

算法 $\mathcal{R}_D$（$\mathcal{A}$（S））的期望风险和经验风险 ˆ $\mathcal{R}_{\mathcal{S}}$（$\mathcal{A}$（$S$））定义如下：

$$\mathcal{R}_{\mathcal{D}}(\mathcal{A}(S)) = \mathbb{E}_{z\sim\mathcal{D}}\ell(\mathcal{A}(S), z) = \max_{\|\delta\|_2\leq\rho} \mathbb{E}_{z\sim\mathcal{D}}\ell(\mathbf{w} + \delta_z; z),$$

$$\hat{\mathcal{R}}_S(\mathcal{A}(S)) = \frac{1}{N}\sum_{i=1}^{N}\ell(\mathcal{A}(S), z_i) = \max_{\|\delta\|_2\leq\rho} \frac{1}{N}\sum_{i=1}^{N}\ell(\mathbf{w} + \delta_{z_i}; z_i),$$

其中 ℓ： $\mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$ 是 损失函数，$\delta = \rho \frac{g\ (z,\ \mathbf{w})}{\|g\ (z,\ \mathbf{w})\|_2}$（$g$（$z$，$\mathbf{w}$）是梯度），$N$ 是训练样本量。

定义 6（泛化界，[50]）。泛化误差定义为期望风险与经验风险之差，

$$\mathrm{Gen}_{S,\mathcal{A}(S)} \triangleq \mathcal{R}_{\mathcal{D}}(\mathcal{A}(S)) - \hat{\mathcal{R}}_S(\mathcal{A}(S)),$$

其上界称为泛化界。

定义 7（差分隐私，[9]）。*hhh*随机算法 $\mathcal{A}$ 称为 (ε, δ)-差分隐私的，如果对于任意假设子集$\mathcal{H}_0 \subset \mathcal{H}$ 以及任意一对相邻样本集 S 和 S'，二者仅相差一个样本（称为 S 和 S' 相邻），均有

$$\log\left[\frac{\mathbb{P}_{\mathcal{A}(S)}(\mathcal{A}(S) \in \mathcal{H}_0) - \delta}{\mathbb{P}_{\mathcal{A}(S')}(\mathcal{A}(S') \in \mathcal{H}_0)}\right] \leq \varepsilon.$$

如果算法 $\mathcal{A}$ 满足（ε，0）-差分隐私，也称其满足 ε-差分隐私。

定义 8（多样本集学习算法，[50]）。假设将大小为 kN 的训练样本集 S 划分为 k 个子样本集 S_1，.....，S_k，每个子样本集的大小为 N。换言之，S 由 k 个子样本集构成，如下所示：

$$S = (S_1, \ldots, S_k).$$

由多样本集算法 $\mathcal{B}$ 在数据集 S 上学习得到的假设 $\mathcal{B}$（S）定义如下：

$$\mathcal{B} : \mathcal{Z}^{k\times N} \mapsto \mathcal{H} \times \{1, \ldots, k\},\ \mathcal{B}(S) = \left(h_{\mathcal{B}(S)}, i_{\mathcal{B}(S)}\right).$$

A.2 预备引理

引理 2（引理 B.1, [45]）。在假设 *1-2* 下，对于满足 $\eta \leq \frac{1}{4KL}$ 的任意学习率，更新由 $\delta_{i,\ k} - \delta$ 引起的漂移为：

$$\frac{1}{M}\sum_i \mathbb{E}[\|\delta_{i,k} - \delta\|^2] \leq 2K^2L^2\eta^2\rho^2.$$

其中

$$\delta = \rho\frac{\nabla f(\mathbf{w}^t)}{\|\nabla f(\mathbf{w}^t)\|},\quad \delta_{i,k} = \rho\frac{\nabla F_i(\mathbf{w}^{t,k}, \xi_i)}{\|\nabla F_i(\mathbf{w}^{t,k}, \xi_i)\|}.$$

引理 3 (引理 B.2, [45])。 在上述假设下，对于满足 $\eta_l \leq \frac{1}{10KL}$ 的任意学习率，更新因$\mathbf{w}^{t,k}(i)$ $-\mathbf{w}^t$而产生的漂移为：

$$\frac{1}{M}\sum_i \mathbb{E}[\|\mathbf{w}^{t,k}(i) - \mathbf{w}^t\|^2] \leq 5K\eta^2\Big(2L^2\rho^2\sigma_l^2 + 6K(3\sigma_g^2 + 6L^2\rho^2) + 6K\|\nabla f(\mathbf{w}^t)\|^2\Big) + 24K^3\eta^4L^4\rho^2.$$

引理 4 (定理 9, [61])。 该引理给出了无放回均匀子采样后差分隐私参数的界。设 $\mathcal{M}^o: \mathcal{Z}^m \mapsto \Delta\mathcal{H}$ 是任意一个保持(ε, δ)差分隐私的机制。设 $\mathcal{M}^{wo}: \mathcal{Z}^N \mapsto \Delta\mathcal{Z}^m$ 是无放回均匀子采样机制。则机制 $\mathcal{M}^o \circ \mathcal{M}^{wo}$满足$(\log(1+(m/N)(e^\varepsilon-1))$ $m\delta/N)$差分隐私。

引理 5 (定理 4, [50])。 该引理给出了单步隐私保护方法与迭代式机器学习方法之间的关系。假设一个迭代式机器学习算法 $\mathcal{A}$ 有 T步: $\{W_i(S)\}_{i=1}^T$。具体而言，我们将第 i 个迭代器定义如下，

$$\mathcal{M}_i : (W_{i-1}(S), S) \mapsto W_i(S).$$

假设 W_0 是初始假设（不依赖于 S）。如果对于任意固定的 W_{i-1}, $\mathcal{M}_i(W_{i-1}, S)$ 满足 *(ε_i, δ)*-差分隐私，那么 $\{W_i\}_{i=0}^T$ 是 *(ε', δ')*-差分隐私的，且

$$\varepsilon' = \sqrt{2\log\left(\frac{1}{\tilde{\delta}}\right)\left(\sum_{i=1}^T \varepsilon_i^2\right)} + \sum_{i=1}^T \varepsilon_i \frac{e^{\varepsilon_i}-1}{e^{\varepsilon_i}+1},$$

$$\begin{aligned}\delta' =& e^{-\frac{\varepsilon'+T\varepsilon}{2}}\left(\frac{1}{1+e^\varepsilon}\left(\frac{2T\varepsilon}{T\varepsilon-\varepsilon'}\right)\right)^T\left(\frac{T\varepsilon+\varepsilon'}{T\varepsilon-\varepsilon'}\right)^{-\frac{\varepsilon'+T\varepsilon}{2\varepsilon}} - \left(1-\frac{\delta}{1+e^\varepsilon}\right)^T \\ &+ 2 - \left(1-e^\varepsilon\frac{\delta}{1+e^\varepsilon}\right)^{\left\lceil\frac{\varepsilon'}{\varepsilon}\right\rceil}\left(1-\frac{\delta}{1+e^\varepsilon}\right)^{T-\left\lceil\frac{\varepsilon'}{\varepsilon}\right\rceil}.\end{aligned}$$

Not ˜明 $0 < \delta \leq 1$ 是任意正实数常量，这与定义 *2* 中的 δ 相同 .

引理 6 (定理 1, [50])。 该引理为任意 *(ε, δ)*-差分隐私机器学习算法给出了高概率泛化界。假设算法 $\mathcal{A}$满足 *(ε, δ)*-差分隐私，训练样本量 N $\geq$最大值 $\left\{\frac{c_1}{\varepsilon^2}\ln\left(\frac{c_2}{e^{-c_3\varepsilon}\delta}\right), c_5^{c_4}(1-c_6e^{-\varepsilon})\ln c_7e^{-\varepsilon}\delta\right\}$，且损失函数 $\|l\|_\infty \leq 1$。那么，对于数据空间 $\mathcal{Z}$ 上的任意数据分布 $\mathcal{D}$，有以下不等式，

$$\mathbb{P}\left[\left|\hat{\mathcal{R}}_S(\mathcal{A}(S)) - \mathcal{R}(\mathcal{A}(S))\right| < 4\varepsilon\right] > 1 - \frac{2e^{-1.7\varepsilon}\delta}{\varepsilon}\ln\left(\frac{2}{\varepsilon}\right).$$

其中， $c_1, ..., c_7$ 均为正常数。

引理 7. 在一次通信轮次中，两个相邻数据集仅相差客户端 i 的一个样本，其对应的两个模型参数在通信轮次 t ，

$$\sum_{k=0}^{K-1}\|\mathbf{y}^{t,k}(i) - \mathbf{x}^{t,k}(i)\|_2^2 \leq 2K\max\|\Delta_i^t(\mathbf{y}) - \Delta_i^t(\mathbf{x})\|_2^2.$$

证明。 回顾客户端 i 的本地更新为 $\sum_{k=0}^{K-1}\mathbf{w}^{t,k}(i) = \sum_{k=0}^{K-1}\mathbf{w}^{t,k-1}(i) + \Delta_i^t$, （假设初始值为$\mathbf{w}^{t,-1} = \mathbf{w}^{t,0} = \mathbf{w}^t$）。那么，

$$\begin{aligned}\sum_{k=0}^{K-1}\|\mathbf{y}^{t,k}(i) - \mathbf{x}^{t,k}(i)\|_2^2 \leq& 2\sum_{k=0}^{K-1}\|\mathbf{y}^{t,k-1}(i) - \mathbf{x}^{t,k-1}(i)\|_2^2 \\ &+ 2\|\Delta_i^t(\mathbf{y}) - \Delta_i^t(\mathbf{x})\|_2^2.\end{aligned}$$

从 $\tau = 0$到 k 的递推关系可得

$$\sum_{k=0}^{K-1}\|\mathbf{y}^{t,k}(i) - \mathbf{x}^{t,k}(i)\|_2^2 \overset{a)}{\leq} 2K\max\|\Delta_i^t(\mathbf{y}) - \Delta_i^t(\mathbf{x})\|_2^2.$$

其中，a）使用初始值$\mathbf{w}^t(i) = \mathbf{x}^{t,0}(i) = \mathbf{y}^{t,0}(i)$ 以及 $0 < k \leq K$。 □

引理 8. 在 假设*1*和*3*下，选定客户端经过裁剪操作后的本地更新平均值为

$$\mathbb{E}\|\frac{1}{m}\sum_{i\in\mathcal{W}^t}\tilde{\Delta}_i^t\|^2 \leq 3K\eta^2(L^2\rho^2 + B^2)$$

证明。

$$\begin{aligned}\mathbb{E}\|\frac{1}{m}\sum_{i\in\mathcal{W}^t}\tilde{\Delta}_i^t\|^2 \leq& \mathbb{E}\|\frac{1}{m}\sum_{i\in\mathcal{W}^t}\sum_{i=0}^{K-1}\eta\tilde{\mathbf{g}}^{t,k}(i)\cdot\alpha_i^t\|^2 \leq \frac{\eta^2}{m}\sum_{i\in\mathcal{W}^t}\sum_{i=0}^{K-1}\mathbb{E}\|\nabla F_i(\mathbf{w}^{t,k}(i)+\delta;\xi_i) - \nabla F_i(\mathbf{w}^{t,K}(i);\xi_i) \\ &+ \nabla F_i(\mathbf{w}^{t,k}(i);\xi_i) - \nabla F_i(\mathbf{w}^t(i)) + \nabla F_i(\mathbf{w}^t(i))\|^2 \\ &\overset{a)}{\leq} 3K\eta^2(L^2\rho^2 + B^2),\end{aligned}$$

其中，a）使用假设1和假设3，且

$$\alpha_i^t := \min\Big(1, \frac{C}{\eta\|\sum_{k=0}^{K-1}\tilde{\mathbf{g}}^{t,k}(i)\|_2}\Big).$$

□

A.3 敏感度分析证明

定理 *1* 的证明。 回顾一下，在客户端 i上，裁剪并添加噪声之前的本地更新为 $\Delta_i^t = \mathbf{w}^{t,K}(i) - \mathbf{w}^{t,0}(i)$。于是,

$$\begin{aligned}
\mathbb{E}\mathcal{S}^2_{\Delta_i^t} &= \max \mathbb{E}\|\Delta_i^t(\mathbf{x}) - \Delta_i^t(\mathbf{y})\|_2^2 \\
&= \mathbb{E}\|\mathbf{x}^{t,K}(i) - \mathbf{x}^{t,0}(i) - (\mathbf{y}^{t,K}(i) - \mathbf{y}^{t,0}(i))\|_2^2 \\
&= \eta^2\mathbb{E}\sum_{k=0}^{K-1}\|\nabla F_i(\mathbf{x}^{t,k}(i)+\delta_x;\xi_i) - \nabla F_i(\mathbf{y}^{t,k}(i)+\delta_y;\xi_i^{'})\|_2^2 \\
&= \eta^2L^2\mathbb{E}\sum_{k=0}^{K-1}\|\mathbf{y}^{t,k}(i) - \mathbf{x}^{t,k}(i) + (\delta_y - \delta_x)\|_2^2 \\
&\overset{a)}{\leq} 2\eta^2L^2K\max\|\Delta_i^t(\mathbf{y}) - \Delta_i^t(\mathbf{x})\|_2^2 + 2\eta^2L^2\rho^2\mathbb{E}\sum_{k=0}^{K-1}\Big\|\frac{\nabla F_i(\mathbf{y}^{t,k}(i)+\delta_y;\xi_i^{'})}{\|\nabla F_i(\mathbf{y}^{t,k}(i)+\delta_y;\xi_i^{'})\|_2} - \frac{\nabla F_i(\mathbf{y}^{t,k}(i);\xi_i^{'})}{\|\nabla F_i(\mathbf{y}^{t,k}(i);\xi_i^{'})\|_2} \\
&+ \Big(\frac{\nabla F_i(\mathbf{x}^{t,k}(i);\xi_i)}{\|\nabla F_i(\mathbf{x}^{t}(i,k);\xi_i)\|_2} - \frac{\nabla F_i(\mathbf{x}^{t,k}(i)+\delta_x;\xi_i)}{\|\nabla F_i(\mathbf{x}^{t,k}(i)+\delta_x;\xi_i)\|_2}\Big) + \frac{\nabla F_i(\mathbf{y}^{t,k}(i);\xi_i^{'}))}{\|\nabla F_i(\mathbf{y}^{t,k}(i);\xi_i^{'}))\|_2} - \frac{\nabla F_i(\mathbf{x}^{t,k}(i);\xi_i)}{\|\nabla F_i(\mathbf{x}^{t,k}(i);\xi_i)\|_2}\Big\|_2^2 \\
&\leq 2\eta^2L^2K\max\|\Delta_i^t(\mathbf{y}) - \Delta_i^t(\mathbf{x})\|_2^2 + 6\eta^2\rho^2L^2\mathbb{E}\sum_{k=0}^{K-1}\Big(4 + \frac{1}{\rho^2}\Big\|\rho\frac{\nabla F_i(\mathbf{y}^{t,k}(i);\xi_i^{'}))}{\|\nabla F_i(\mathbf{y}^{t,k}(i);\xi_i^{'}))\|_2} - \rho\frac{\nabla f(\mathbf{y}^t)}{\|\nabla f(\mathbf{y}^t)\|_2} \\
&+ \Big(\rho\frac{\nabla f(\mathbf{x}^t)}{\|\nabla f(\mathbf{x}^t)\|_2} - \rho\frac{\nabla F_i(\mathbf{x}^{t,k}(i);\xi_i)}{\|\nabla F_i(\mathbf{x}^{t,k}(i);\xi_i)\|_2}\Big) + \rho\frac{\nabla f(\mathbf{y}^t)}{\|\nabla f(\mathbf{y}^t)\|_2} - \rho\frac{\nabla f(\mathbf{x}^t)}{\|\nabla f(\mathbf{x}^t)\|_2}\Big\|_2^2\Big) \\
&\overset{b)}{\leq} 2\eta^2L^2K\mathcal{S}^2_{\Delta_i^t} + 6\eta^2\rho^2KL^2(4 + 12K^2L^2\eta^2 + 6) \\
&\leq \frac{6\eta^2\rho^2KL^2(12K^2L^2\eta^2+10)}{1-2\eta^2L^2K}
\end{aligned} \tag{21}$$

其中，a) 和 b) 分别使用引理 7 和引理 2。

当局部自适应学习率满足 $\eta = \mathcal{O}(1/L\sqrt{KT})$ 且扰动幅度 ρ 与学习率成正比时，例如 $\rho = \mathcal{O}(\frac{1}{\sqrt{T}})$，我们有

$$\mathbb{E}\mathcal{S}^2_{\Delta_i^t} \leq \mathcal{O}\left(\frac{1}{T^2}\right). \tag{22}$$

□

作为比较，我们还给出了差分隐私联邦学习中采用随机梯度下降的本地更新的期望平方敏感度，如下所示。显然，$\mathbb{E}\mathcal{S}^2_{\Delta_i^t, SAM}$ 中的上界比 $\mathbb{E}\mathcal{S}^2_{\Delta_i^t, SGD}$中的更紧。

联邦学习中随机梯度下降的敏感度证明。

$$\begin{aligned}
\mathbb{E}\mathcal{S}^2_{\Delta_i^t, SGD} &= \max \mathbb{E}\|\Delta_i^t(\mathbf{x}) - \Delta_i^t(\mathbf{y})\|_2^2 = \eta^2\mathbb{E}\sum_{i=0}^{K-1}\|\nabla F_i(\mathbf{x}^{t,k}(i);\xi_i) - \nabla F_i(\mathbf{y}^{t,k}(i);\xi_i^{'})\|_2^2 \\
&= \eta^2\mathbb{E}\sum_{i=0}^{K-1}\|\nabla F_i(\mathbf{x}^{t,k}(i);\xi_i) - \nabla F_i(\mathbf{x}^t(i)) + \nabla F_i(\mathbf{x}^t(i)) - \nabla F_i(\mathbf{y}^t(i)) + \nabla F_i(\mathbf{y}^t(i)) - \nabla F_i(\mathbf{y}^{t,k}(i);\xi_i^{'})\|_2^2 \\
&\overset{a)}{\leq} 3\eta^2\mathbb{E}\sum_{i=0}^{K-1}(2\sigma_l^2 + L^2\|y^{t,k}(i) - x^{t,k}(i)\|_2^2) \\
&\overset{b)}{\leq} 6\eta^2K\sigma_l^2 + 3\eta^2L^2K\max\mathbb{E}\|\Delta_i^t(\mathbf{x}) - \Delta_i^t(\mathbf{y})\|_2^2 \\
&\leq \frac{6\eta^2\sigma_l^2K}{1-3\eta^2KL^2}.
\end{aligned} \tag{23}$$

其中 a) 和 b) 使用了假设 分别为 1-2 和引理 7。因此 $\mathbb{E}\mathcal{S}^2_{\Delta_i^t,\ SGD} \leq \mathcal{O}(\frac{\sigma_l^2}{KL^2T})$ 当 $\eta = \mathcal{O}(1/L\sqrt{KT})$时。 □

A.4 收敛性分析证明

定理 *3* 的证明。为方便起见，我们定义如下记号：

$$\tilde{\Delta}_i^t = -\eta \sum_{k=0}^{K-1} \tilde{\mathbf{g}}^{t,k}(i) \cdot \alpha_i^t;$$
$$\overline{\Delta_i^t} = -\eta \sum_{k=0}^{K-1} \tilde{\mathbf{g}}^{t,k}(i) \cdot \overline{\alpha}^t,$$

其中

$$\alpha_i^t := \min\Big(1, \frac{C}{\eta \| \sum_{k=0}^{K-1} \tilde{\mathbf{g}}^{t,k}(i) \|}\Big),$$
$$\overline{\alpha}^t := \frac{1}{M}\sum_{i=1}^{M} \alpha_i^t,$$
$$\tilde{\alpha}^t := \frac{1}{M}\sum_{i=1}^{M} |\alpha_i^t - \overline{\alpha}^t|.$$

∇f的 Lipschitz 连续性：

$$\begin{aligned}
&\mathbb{E} f(\mathbf{w}^{t+1}) \\
&\le \mathbb{E} f(\mathbf{w}^t) + \mathbb{E}\Big\langle \nabla f(\mathbf{w}^t), \mathbf{w}^{t+1} - \mathbf{w}^t \Big\rangle + \mathbb{E}\frac{L}{2}\|\mathbf{w}^{t+1} - \mathbf{w}^t\|^2 \\
&= \mathbb{E} f(\mathbf{w}^t) + \mathbb{E}\Big\langle \nabla f(\mathbf{w}^t), \frac{1}{m}\sum_{i\in\mathcal{W}^t} \tilde{\Delta}_i^t + z_i^t \Big\rangle + \frac{L}{2}\mathbb{E}\Big\| \frac{1}{m}\sum_{i\in\mathcal{W}^t} \tilde{\Delta}_i^t + z_i^t \Big\|^2 \\
&= \mathbb{E} f(\mathbf{w}^t) + \underbrace{\Big\langle \nabla f(\mathbf{w}^t), \mathbb{E}\frac{1}{m}\sum_{i\in\mathcal{W}^t} \tilde{\Delta}_i^t \Big\rangle}_{\text{I}} + \frac{L}{2}\underbrace{\mathbb{E}\Big\langle \| \frac{1}{m}\sum_{i\in\mathcal{W}^t} \tilde{\Delta}_i^t \|^2 \Big\rangle}_{\text{II}} + \frac{L\sigma^2 C^2 pd}{2m^2},
\end{aligned} \tag{24}$$

其中， d 表示$\mathbf{w}_i^{t\,k}$,的维度， p 为稀疏率，噪声 z_i^t 的均值为零。接下来，我们分别分析 I 和 II。

对于 I，有

$$\Big\langle \nabla f(\mathbf{w}^t), \mathbb{E}\frac{1}{m}\sum_{i\in\mathcal{W}^t} \tilde{\Delta}_i^t \Big\rangle = \Big\langle \nabla f(\mathbf{w}^t), \mathbb{E}\frac{1}{M}\sum_{i=1}^{M} \tilde{\Delta}_i^t - \overline{\Delta}_i^t \Big\rangle + \Big\langle \nabla f(\mathbf{w}^t), \mathbb{E}\frac{1}{M}\sum_{i=1}^{M} \overline{\Delta}_i^t \Big\rangle. \tag{25}$$

Th接下来，我们分别对上述等式中的两项进行界定。对于第一项，有

$$\begin{aligned}
&\mathbb{E}\Big\langle \nabla f(\mathbf{w}^t), \mathbb{E}\frac{1}{M}\sum_{i=1}^{M} \tilde{\Delta}_i^t - \overline{\Delta}_i^t \Big\rangle \\
&\le \mathbb{E}\Big\langle \nabla f(\mathbf{w}^t), \mathbb{E}\frac{1}{M}\sum_{i=1}^{M}\sum_{k=0}^{K-1} \eta |\alpha_i^t - \overline{\alpha}^t| \tilde{\mathbf{g}}^{t,k}(i) \Big\rangle \\
&\le \frac{\eta K}{M}\sum_{i=1}^{M} \mathbb{E}|\alpha_i^t - \overline{\alpha}^t| \Big\langle \nabla F_i(\mathbf{w}^t), \tilde{\mathbf{g}}^{t,k}(i) \Big\rangle \\
&\overset{a)}{\le} \frac{\eta K}{M}\sum_{i=1}^{M} \mathbb{E}|\alpha_i^t - \overline{\alpha}^t| \Big(-\frac{1}{2}(\|\nabla F_i(\mathbf{w}^{t,k})\|^2 + \|F_i(\mathbf{w}^{t,k} + \delta; \xi_i)\|^2) + \frac{1}{2}\|\nabla F_i(\mathbf{w}^{t,k} + \delta; \xi_i) - \nabla F_i(\mathbf{w}^{t,k}; \xi_i)\|^2 \Big) \\
&\overset{b)}{\le} \eta\tilde{\alpha}^t K(\frac{1}{2}L^2\rho^2 - B^2),
\end{aligned} \tag{26}$$

其中， $\tilde{\alpha}^t = \frac{1}{M}\sum_{i=1}^{M} |\alpha_i^t - \overline{\alpha}^t|$， a) 使用 $\langle a, b\rangle = -\frac{1}{2}\|a\|^2 = \frac{1}{2}\|b\|^2 + \frac{1}{2}\|a-b\|^2$， b) 基于假设1,3。对于第二项，我们有

$$\begin{aligned}
&\Big\langle \nabla f(\mathbf{w}^t), \mathbb{E}\frac{1}{M}\sum_{i=1}^{M} \overline{\Delta}_i^t \Big\rangle \\
&\overset{a)}{\le} \frac{-\overline{\alpha}^t \eta K}{2}\|\nabla f(\mathbf{w}^t)\|^2 - \frac{\overline{\alpha}^t}{2K}\mathbb{E}\Big\| \frac{1}{\overline{\alpha}^t M}\sum_{i=1}^{M} \overline{\Delta}_i^t \Big\|^2 + \frac{\overline{\alpha}^t}{2}\underbrace{\mathbb{E}\Big\| \sqrt{K}\nabla f(\mathbf{w}^t) - \frac{1}{\overline{\alpha}^t M\sqrt{K}}\sum_{i=1}^{M} \overline{\Delta}_i^t \Big\|^2}_{\text{III}},
\end{aligned} \tag{27}$$

其中 a) 使用 $\langle a,b\rangle=-\frac{1}{2}\|a\|^2-\frac{1}{2}\|b\|^2+\frac{1}{2}\|a-b\|^2$ 和 $0<\eta<1$。接下来，我们对 III 进行如下界定：

$$\begin{aligned}\text{III}&=K\mathbb{E}\Big\|\nabla f(\mathbf{w}^t)+\frac{1}{MK}\sum_{i=1}^{M}\sum_{k=0}^{K-1}\nabla\eta F_i(\mathbf{w}^{t,k}+\delta;\xi_i)\Big\|^2\\&\le\frac{1}{M}\sum_{i=1}^{M}\sum_{k=0}^{K-1}\mathbb{E}\Big\|\eta(F_i(\mathbf{w}^{t,k}+\delta;\xi_i)-\nabla F_i(\mathbf{w}^{t,k};\xi_i))+\eta(\nabla F_i(\mathbf{w}^{t,k};\xi_i)-\nabla F_i(\mathbf{w}^t))+(1+\eta)\nabla F_i(\mathbf{w}^t)\Big\|^2\\&\overset{a)}{\le}3K\eta^2L^2\Big(\rho^2+\mathbb{E}\|\mathbf{w}^{t,k}-\mathbf{w}^t\|^2+2B^2\Big)\\&\overset{b)}{\le}3K\eta^2L^2\Big[\rho^2+5K\eta^2\Big(2L^2\rho^2\sigma_l^2+6K(3\sigma_g^2+6L^2\rho^2)+6K\|\nabla f(\mathbf{w}^t)\|^2\Big)+24K^3\eta^4L^4\rho^2+B^2\Big],\end{aligned}\tag{28}$$

其中 $0<\eta<1$，a) 和 b 分别使用假设1、3 和引理 3。对于 II，我们使用
引理 8。然后，结合公式 12-16，可得

$$\begin{aligned}\mathbb{E}f(\mathbf{w}^{t+1})\le\mathbb{E}f(\mathbf{w}^t)+\eta\tilde{\alpha}_tK(\frac{1}{2}L^2\rho^2-B^2)-\frac{\overline{\alpha}^t\eta K}{2}\|\nabla f(\mathbf{w}^t)\|^2-\frac{\eta\overline{\alpha}^t}{2K}\mathbb{E}\Big\|\frac{1}{\eta\overline{\alpha}^tM}\sum_{i=1}^{M}\overline{\Delta}_i^t\Big\|^2\\+\frac{3\overline{\alpha}^t\eta^2L^2K}{2}\Big[\rho^2+5K\eta^2\Big(2L^2\rho^2\sigma_l^2+6K(3\sigma_g^2+6L^2\rho^2)+6K\|\nabla f(\mathbf{w}^t)\|^2\Big)\\+24K^3\eta^4L^4\rho^2+B^2\Big]+\frac{3\eta^2KL(L^2\rho^2+B^2)}{2}+\frac{L\sigma^2C^2pd}{2m^2}.\end{aligned}\tag{29}$$

当 $\eta\le\frac{1}{3\sqrt{KL}}$ 时，不等式为

$$\begin{aligned}\mathbb{E}f(\mathbf{w}^{t+1})\le\mathbb{E}f(\mathbf{w}^t)-\frac{\overline{\alpha}^t\eta K}{2}\mathbb{E}\|\nabla f(\mathbf{w}^t)\|^2+\frac{\tilde{\alpha}^t\eta KL^2\rho^2}{2}+\frac{3\overline{\alpha}^t\eta^2KL^2\rho^2}{2}-\tilde{\alpha}^t\eta KB^2\\+\frac{15\overline{\alpha}^tK\eta^4L^2}{2}\Big(2L^2\rho^2\sigma_l^2+6K(3\sigma_g^2+6L^2\rho^2)+6K\|\nabla f(\mathbf{w}^t)\|^2\Big)+36\eta^6K^4L^6\rho^2\\+\frac{3\eta^2KL(L^2\rho^2+B^2)}{2}+\frac{L\sigma^2C^2pd}{2m^2}.\end{aligned}\tag{30}$$

对 t 从 1 到 T 求和，可得

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\overline{\alpha}^t\|f(\mathbf{w}^t)\|^2\Big]\le\frac{2L(f(\mathbf{w}^1)-f^*)}{\sqrt{KT}}+\frac{1}{T}\sum_{t=1}^{T}\tilde{\alpha}^tL^2\rho^2-2\tilde{\alpha}^tB^2+30\eta^2L^2\frac{1}{T}\sum_{t=1}^{T}\overline{\alpha}^t\Big(2L^2\rho^2\sigma_l^2+6K(3\sigma_g^2+6L^2\rho^2)\Big)\\+72\eta^4K^3L^6\rho^2+3\eta L(L^2\rho^2+B^2)+\frac{L\sigma^2C^2pd}{\eta m^2K}\end{aligned}\tag{31}$$

假设局部自适应学习率满足 $\eta=\mathcal{O}(1/L\sqrt{KT})$，$\frac{1}{T}\sum_{t=1}^{T}\tilde{\alpha}^t$ 和 $\frac{1}{T}\sum_{t=1}^{T}\alpha^t$ 都是衡量裁剪影响的两个重要参数。同时，$\frac{1}{T}\sum_{t=1}^{T}\tilde{\alpha}^t$ 和 $\frac{1}{T}\sum_{t=1}^{T}\alpha^t$ 也都以1为界。于是，我们得到

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\overline{\alpha}^t\left\|\nabla f\left(\mathbf{w}^t\right)\right\|^2\right]\le\underbrace{\mathcal{O}\left(\frac{2L(f(\mathbf{w}^1)-f^*)}{\sqrt{KT}}+\frac{\sigma_l^2L^2\rho}{KT}\right)}_{\text{From FedSAM}}+\underbrace{\underbrace{\mathcal{O}\left(\sum_{t=1}^{T}(\frac{\overline{\alpha}^t\sigma_g^2}{T^2}+\frac{\tilde{\alpha}^tL^2\rho^2}{T})\right)}_{\text{Clipping}}+\underbrace{\mathcal{O}\left(\frac{L^2\sqrt{T}\sigma^2C^2pd}{m^2\sqrt{K}}\right)}_{\text{Adding noise}}}_{\text{From operations for DP}}.\tag{32}$$

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

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\overline{\alpha}^t\left\|\nabla f\left(\mathbf{w}^t\right)\right\|^2\right]\le\underbrace{\mathcal{O}\left(\frac{2L(f(\mathbf{w}^1)-f^*)}{\sqrt{KT}}+\frac{L^2\sigma_l^2}{KT^2}\right)}_{\text{From FedSAM}}+\underbrace{\underbrace{\mathcal{O}\left(\sum_{t=1}^{T}(\frac{\overline{\alpha}^t\sigma_g^2}{T^2}+\frac{\tilde{\alpha}^tL^2}{T^2})\right)}_{\text{Clipping}}+\underbrace{\mathcal{O}\left(\frac{L^2\sqrt{T}\sigma^2C^2pd}{m^2\sqrt{K}}\right)}_{\text{Adding noise}}}_{\text{From operations for DP}}.\tag{33}$$

□

A.5 泛化分析证明

证明。 主要证明思路是从差分隐私（差分隐私）的角度推导泛化界。证明框架可分为三个阶段：(1) 首先，取所提出算法的全局模型，并将其归类为迭代式机器学习算法。(2) 然后，计算算法中每一步/每一轮的差分隐私。(3) 借助 [50]第5节提供的桥梁，将其拓展到差分隐私联邦学习。

首先，在整个训练过程中始终存在一个全局模型 $\mathbf{w}^t$ 。为简单起见，我们将迭代 t 时的全局模型定义为 $\mathbf{w}^t$ ，该迭代也对应于通信轮次。于是，FL 范式可以看作利用 m 参与客户端上的梯度信息，对 $\mathbf{w}^t$ 进行迭代优化。我们还将 $\mathcal{N}(0, \sigma^2 C^2 \cdot \mathrm{I}_d/m)$ 表示为差分隐私中添加的高斯噪声，其中 σ^2 为高斯噪声方差。我们将 τ 定义为 m 参与客户端，并将总体迭代步数定义为 T。在我们的算法中，通过使用 SAM本地优化器，本地更新模型空间的直径定义为 $D \triangleq \max_{\mathbf{w},z,z'} \|\nabla \ell(z, \mathbf{w}) - \nabla \ell(z', \mathbf{w})\| = \max_{\mathbf{w},z,z'} \| \max_{\|\delta\|_2 \le \rho} (\nabla \ell(\mathbf{w} + \delta_z;\ z) - \nabla \ell(\mathbf{w} + \delta_{z'}; z'))\|$。为简洁起见，我们还将 $G_{\mathcal{B}}(\mathbf{w}) \triangleq \frac{1}{\|\mathcal{B}\|} \sum_{z \in \mathcal{B}} g(z, \mathbf{w})$ 表示为 g 在 $\mathcal{B}$ 上的均值。我们还使用 $\boldsymbol{p}$ 表示概率密度，其中 $\boldsymbol{p}^V$ 表示在任意随机变量 V 条件下的概率密度。

随后，由于本地更新经过了裁剪，我们将本地更新模型空间的直径定义为 [50]

$$\begin{aligned} D &= \max_{\mathbf{w},z,z'} \| \max_{\|\delta\|_2 \le \rho} (\nabla \ell(\mathbf{w} + \delta_z; z) - \nabla \ell(\mathbf{w} + \delta_{z'}; z')) \| \\ &\overset{a}{=} 2L\rho, \end{aligned}$$

其中，a）使用了假设1，且 $\delta = \rho \ {}^{g(z,\mathbf{w})}/_{\|g(z,\mathbf{w})\|_2}$。接下来，我们计算每个步骤/轮次的差分隐私。回顾算法 1，第 5–10 行表示本地训练过程，即梯度信息。算法 1 中的第 3 行等价于从大小为 $[N]$ 的索引集中 $\mathcal{I}_t$ 无放回地均匀抽取一个大小为 τ 的小批量，并令 $\mathcal{B}_t = S_{\mathcal{I}_t}$。此外，对于固定的 $\mathbf{w}^{t-1}$、$\mathcal{I}$，以及任意两个相邻的样本集 S 和 S'，我们有

$$\begin{aligned} \frac{\boldsymbol{p}^{S,\mathcal{I}_t}(\mathbf{w}^t = \mathbf{w} | \mathbf{w}^{t-1})}{\boldsymbol{p}^{S',\mathcal{I}_t}(\mathbf{w}^t = \mathbf{w} | \mathbf{w}^{t-1})} &= \frac{\boldsymbol{p}^{S,\mathcal{I}_t}(\eta_t(G_{S_{\mathcal{I}}}(\mathbf{w}^{t-1}) + \mathcal{N}(0, \sigma^2 C^2 \cdot \mathbf{I}_d/m)) = \mathbf{w} - \mathbf{w}^{t-1})}{\boldsymbol{p}^{S',\mathcal{I}_t}(\eta_t(G_{S'_{\mathcal{I}}}(\mathbf{w}^{t-1}) + \mathcal{N}(0, \sigma^2 C^2 \cdot \mathbf{I}_d/m)) = \mathbf{w} - \mathbf{w}^{t-1})} \\ &= \frac{\boldsymbol{p}^{\mathcal{I}_t, \mathbf{w}^{t-1}}(\mathcal{N}(0, \sigma^2 C^2 \cdot \mathbf{I}_d/m) = \mathbf{w}')}{\boldsymbol{p}^{S,S',\mathcal{I}_t,\mathbf{w}^{t-1}}(G_{S'_{\mathcal{I}}}(\mathbf{w}^{t-1}) - G_{S_{\mathcal{I}}}(\mathbf{w}^{t-1}) + \mathcal{N}(0, \sigma^2 C^2 \cdot \mathbf{I}_d/m) = \mathbf{w}')}, \end{aligned} \tag{34}$$

其中 $\eta_t \mathbf{w}' = \mathbf{w} - \mathbf{w}^{t-1} - \eta_t G_{S_{\mathcal{I}}}(\mathbf{w}^{t-1})$。因此，考虑加性高斯噪声时，如果 $\mathbf{w} \sim \mathbf{w}^{t-1} + \eta_t(G_{S_{\mathcal{I}}}(\mathbf{w}^{t-1}) + \mathcal{N}(0, \sigma^2 C^2 \cdot \mathrm{I}_d/m))$，那么 $\mathbf{w}' \sim G_{S_{\mathcal{I}}}(\mathbf{w}^{t-1}) + \mathcal{N}(0, \sigma^2 C^2 \cdot \mathrm{I}_d/m)$。

为简单起见，根据差分隐私的定义，我们定义

$$D_p^{S,S',\mathcal{I}_t,\mathbf{w}^{t-1}}(\mathbf{w}') = \log \frac{\boldsymbol{p}^{\mathcal{I}_t, \mathbf{w}^{t-1}}(\mathcal{N}(0, \sigma^2 C^2 \cdot \mathbf{I}_d/m) = \mathbf{w}')}{\boldsymbol{p}^{S,S',\mathcal{I}_t,\mathbf{w}^{t-1}}(G_{S'_{\mathcal{I}}}(\mathbf{w}^{t-1}) - G_{S_{\mathcal{I}}}(\mathbf{w}^{t-1}) + \mathcal{N}(0, \sigma^2 C^2 \cdot \mathbf{I}_d/m) = \mathbf{w}')}, \tag{35}$$

根据高斯分布的定义，进一步可得

$$\begin{aligned} D_p(\mathbf{w}') &= -\frac{\|\mathbf{w}'\|^2}{2\sigma^2 C^2 d^2/m^2} + \frac{\|\mathbf{w}' - G_{S'_{\mathcal{I}}}(\mathbf{w}^{t-1}) + G_{S_{\mathcal{I}}}(\mathbf{w}^{t-1})\|^2}{2\sigma^2 C^2 d^2/m^2} \\ &= \frac{2\langle \mathbf{w}', -G_{S'_{\mathcal{I}}}(\mathbf{w}^{t-1}) + G_{S_{\mathcal{I}}}(\mathbf{w}^{t-1})\rangle + \|G_{S'_{\mathcal{I}}}(\mathbf{w}^{t-1}) - G_{S_{\mathcal{I}}}(\mathbf{w}^{t-1})\|^2}{2\sigma^2 C^2 d^2/m^2}. \end{aligned} \tag{36}$$

记 $-G_{S'_{\mathcal{I}}}(\mathbf{w}^{t-1}) + G_{S_{\mathcal{I}}}(\mathbf{w}^{t-1})$ 为$\mathbf{v}$。根据 $D = 2L\rho$ （梯度空间的直径）的定义，可得

$$\|\mathbf{v}\| < \frac{1}{m} D < \frac{2L\rho}{m}. \tag{37}$$

另一方面，由于 $\langle \mathbf{v}, \mathbf{w}' \rangle \sim \mathcal{N}(0, \|\mathbf{v}\|^2 \sigma^2 C^2 d^2/m^2)$，根据切尔诺夫界技巧，可得

$$\begin{aligned} \mathbb{P}\left(\langle \mathbf{v}, \mathbf{w}' \rangle \ge \frac{2\sqrt{2} L\rho\sigma C d}{m^2} \sqrt{\log \frac{1}{\tilde{\delta}}} \right) &\le \mathbb{P}\left(\langle \mathbf{v}, \mathbf{w}' \rangle \ge \frac{\sqrt{2}\|\mathbf{v}\|\sigma C d}{m} \sqrt{\log \frac{1}{\tilde{\delta}}} \right) \\ &\le \min_{1 \le t \le T} \exp\left(-\frac{\sqrt{2} t \|\mathbf{v}\| \sigma C d}{m} \sqrt{\log \frac{1}{\tilde{\delta}}} \right) \mathbb{E}(e^{t\langle \mathbf{v}, \mathbf{w}' \rangle}). \end{aligned} \tag{38}$$

其中 $0 < \delta \le 1$ 是任意正实常数，然后我们定义

$$\delta = \min_{1 \le t \le T} \exp\left(-\frac{\sqrt{2} t \|\mathbf{v}\| \sigma C d}{m} \sqrt{\log \frac{1}{\tilde{\delta}}} \right) \mathbb{E}(e^{t\langle \mathbf{v}, \mathbf{w}' \rangle}). \tag{39}$$

因此，以至少 $1-\delta$ 的概率（相对于 $\mathbf{w}'$），我们有

$$\begin{aligned} D_p(\mathbf{w}') &\leq \frac{\sqrt{2}L\rho\sigma Cd\sqrt{\log\frac{1}{\delta}}+2L^2\rho^2}{\sigma^2C^2d^2} \\ &\leq \frac{L\rho\sqrt{2\log\frac{1}{\delta}}}{\sigma Cd}+\left(\frac{\sqrt{2}L\rho}{\sigma Cd}\right)^2. \end{aligned} \tag{40}$$

结合引理 4，可得算法 1 中的每一步均满足 $(\tilde{\varepsilon}, \frac{m}{N}\delta)$-差分隐私，其中 $\tilde{\varepsilon}$ 定义为

$$\tilde{\varepsilon}=\log\left(\frac{N-m}{N}+\frac{m}{N}\exp\left(\frac{L\rho\sqrt{2\log\frac{1}{\delta}}}{\sigma Cd}+\left(\frac{\sqrt{2}L\rho}{\sigma Cd}\right)^2\right)\right). \tag{41}$$

应用引理 5，可得迭代步骤满足 (ε', δ')-差分隐私保证。最后，结合引理 6 与 (ε', δ') 即可完成证明。

□