arXiv:2602.01039v1 [cs.LG] 1 Feb 2026

通过分布外检测实现联邦学习的自适应双重加权框架

Zhiwei Ling、Hailiang Zhao，*IEEE* 会员、Chao Zhang、Xiang Ao、Ziqi Wang、Cheng Zhang、Zhen Qin、Xinkui Zhao、Kingsum Chow、Yuanqing Wu，*IEEE* 高级会员，以及 MengChu Zhou，*IEEE* 会士

摘要—联邦学习（FL）能够在保护数据隐私的同时，实现大规模分布式服务节点之间的协同模型训练，因此已成为边缘云环境中智能服务系统的基石。然而，在面向实际服务的部署中，由异构用户、设备和应用场景产生的数据天然具有非独立同分布特性。这种严重的数据异质性会严重削弱全局模型的收敛稳定性和泛化能力，并最终影响其提供的服务质量。为应对这一挑战，我们提出了 FLOOD，一种受分布外（OOD）检测启发的新型联邦学习框架。FLOOD 通过联合控制本地训练和全局聚合的双重加权机制，动态抵消数据异质性带来的不利影响。在客户端层级，该方法通过提高伪分布外样本的权重，自适应地重新加权监督损失，从而促进对分布不一致或具有挑战性数据的鲁棒学习。在服务器层面，该方法依据客户端的 OOD 置信度分数对客户端贡献进行加权，进一步优化模型聚合，优先采用具有更高分布内一致性的客户端更新，从而增强全局模型的鲁棒性和收敛稳定性。在多种非独立同分布设置下的多个基准数据集上进行的大量实验表明，FLOOD 在准确率和泛化能力方面均持续优于最先进的联邦学习方法。此外，FLOOD 还是一个正交插件模块：它能够无缝集成到现有联邦学习算法中，在不修改其核心优化逻辑的情况下，提升其在数据异质性环境下的性能。这些特性使 FLOOD 成为在现实联邦环境中部署可靠智能服务的实用且可扩展解决方案。

索引词—联邦学习、分布外检测、数据异质性、移动边缘计算、自适应样本加权、动态聚合校正。

一、引言

联邦学习（FL）已成为一项变革性面向服务的分布式机器学习范式，使大规模分散式客户端能够协同进行模型训练，同时保护数据隐私 [1]–[3]。通过在不集中原始数据的前提下，促进智能模型在异构服务节点上持续优化，联邦学习为下一代数据驱动型服务构建了基础设施。其有效性已在多个服务计算领域得到广泛验证，包括自动驾驶 [4], [5],智能金融系统 [6], [7],和智慧医疗平台 [8], [9]。

Zhiwei Ling、Hailiang Zhao、Chao Zhang、Ziqi Wang、Zhen Qin、Xinkui Zhao、Kingsum Chow和Hailiang Zhao就职于浙江大学软件技术学院及浙江省数智服务技术重点实验室（电子邮箱：{zwling, hliangzhao, 22351272, wangziqi0312, zhenqin, zhaoxinkui, kingsum.chow}@zju.edu.cn。）。Cheng Zhang就职于中国浙江财经大学信息技术与人工智能学院（电子邮箱：zhangcheng@zufe.edu.cn）。Xiang Ao就职于浙江大学计算机科学与技术学院（电子邮箱：aoxiangyx@zju.edu.cn）。Yuanqing Wu就职于中山大学智能系统工程学院，中国深圳518107（电子邮箱：yqwuzju@163.com）。MengChu Zhou就职于美国新泽西理工学院电气与计算机工程系，纽瓦克，NJ 07102（电子邮箱：zhou@njit.edu）。

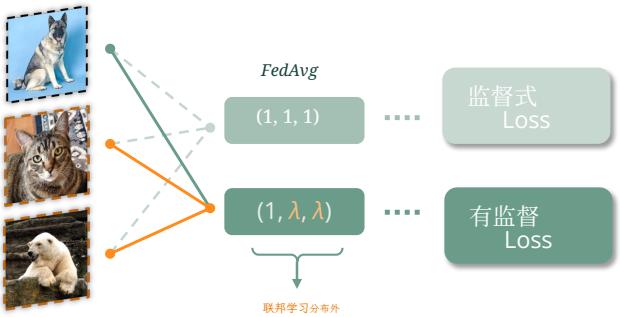


图1：FLOOD与FEDAVG在客户端的关键差异。橙色边框突出显示伪分布外样本，而 λ 表示在监督损失中分配给这些样本的重新分配权重。

在典型的联邦学习赋能服务工作流中，每个客户端从编排服务器接收全局服务模型，在私有服务数据上执行本地训练，并上传加密模型更新。服务器随后聚合这些更新，以改进全局模型，供下一通信轮次使用。这种去中心化设计确保敏感用户数据保留在本地设备上，从而满足严格的隐私要求和监管合规标准 [10]–[12]。因此，联邦学习为大规模部署智能服务提供了可扩展且保护隐私的基础。

尽管前景广阔，但 FL 在现实服务系统中的实际部署受到一项根本性挑战的阻碍：数据异质性。不同用户、设备和运行环境所生成的服务数据，天然呈现出非独立同分布（非 IID）特征 [13], [14]。这类分布差异往往会导致本地训练模型产生显著偏差。当这些存在偏差的更新被简单聚合（例如采用统一平均）时，得到的全局模型会出现收敛不稳定、准确率下降以及服务可靠性受损等问题。这一问题构成了实现可扩展、稳健且高质量的基于联邦学习的智能服务的关键瓶颈[15]。为缓解这一问题，已有大量方法被提出，主要分为

两类：服务器端模型调整方法 [16]–[18] 以及客户端训练校正技术 [19]–[21]。尽管这些方法已展现出显著的改进效果，但它们主要作用于模型参数或梯度层面。因此，它们往往忽视了导致性能退化的一个更为根本的原因：客户端数据分布之间的内在不匹配。

在本研究中，我们通过分布外（OOD）检测的视角重新框定数据异质性，从而提出一种新颖的认识。分布外检测是指深度神经网络（DNN）识别显著偏离训练分布 [22]–[24] 的输入的能力。尽管分布外检测已在集中式学习场景中被广泛探索以进行不确定性量化，但我们表明，分布外置信度分数同样可作为丰富、可解释且切实可操作的信号，用于刻画联邦环境中的分布差异。具体而言，即便客户端的本地数据集在内部看似一致，某些样本在当前全局模型下仍常常给出较低的预测置信度。这些低置信度实例可被解释为伪分布外 样本，它们定量反映本地数据偏离全局数据流形的程度。重要的是，这一解释将数据异质性从一个不可避免的障碍转化为一种可度量且可利用的性质，从而能够同时指导本地训练动态和全局聚合策略。

基于这一洞见，我们提出 FLOOD，这是一种新型联邦学习框架，通过受分布外信号指导的双重加权策略，明确且系统地应对数据异质性。如图 1 所示，与 FEDAVG 相比，FLOOD 通过为伪分布外样本分配更高的损失权重来增强本地训练，从而鼓励本地模型更有效地关注具有挑战性且分布失配的数据。在服务器端，FLOOD 根据客户端的总体 OOD 置信度分数动态调整客户端聚合权重，优先纳入本地数据与不断演变的全局分布更为一致的客户端所提交的更新。本文的主要贡献总结如下：

- 我们通过分布外检测框架，对联邦学习中的数据异质性进行了有原则的重新诠释，揭示了可解释的信号，从而促进跨客户端泛化能力的提升。
- 我们设计了 FLOOD，这是一种联邦学习框架，在本地训练期间纳入分布外感知的样本加权，并在全局聚合期间采用分布外引导的客户端重新加权，从而有效缓解非独立同分布数据带来的不利影响。
- 我们在多个基准数据集和多种非独立同分布配置上进行了广泛实验，结果表明，FLOOD 的表现始终显著优于现有先进联邦学习方法。

本文其余部分安排如下。第II 节回顾联邦学习和分布外检测的相关工作。第III 节介绍必要的背景概念和问题形式化。第IV 节详细介绍 FLOOD 的设计，随后进行全面的实验评估，见第 V节。最后，第 VI 节总结全文并强调主要发现。

II. 相关工作

A. 联邦学习

联邦学习已成为分布式机器学习的一种开创性范式 [25]–[27], ，但非独立同分布数据带来的挑战往往会影响其有效性。为应对数据异质性，现有方法大致可分为两个主要方向：客户端侧局部训练校正和服务器端全局模型调整。

与基础联邦学习框架（FEDAVG [13]）相比，客户端侧局部训练校正方法旨在通过重新定义局部优化目标，减轻非独立同分布数据带来的固有偏差。例如，FEDPROX [19] 引入近端正则化项，以限制本地模型与全局模型之间的差异，从而促进更稳定的聚合。同样，MOON [21]通过引入对比学习目标来对齐本地表征与全局表征，进一步提升本地训练一致性。为解决数据异质性导致的维度坍塌问题，FEDDECORR[14]引入去相关项，以促进特征表征的多样性。FEDDC [28] 将本地模型与全局模型之间的差异显式建模为额外的损失项，从而弥合参数差距并提升训练稳定性。此外，FEDBR [29] 平衡本地模型输出，以减少学习偏差，同时保留本地数据的内在特征。

尽管取得了这些进展，大多数现有方法通常依赖于附加在标准监督损失上的辅助正则化项，而损失函数本身保持不变。相比之下，我们提出的 FLOOD 通过直接修改监督损失，引入了另一种思路。FLOOD 根据单个样本的分布外置信度分数为其分配自适应权重，从而在训练过程中优先处理难分类数据点。这不仅解决了性能退化的根本原因，即数据异质性，还能与基于正则化的现有策略无缝集成，形成互补。据我们所知，FLOOD 是首个在客户端层级纳入分布外引导的自适应损失重加权的框架。

全局模型调整方法旨在通过引入微调、知识蒸馏和动态聚合权重调整等技术来增强全局模型，为传统的直接聚合策略提供更有效的替代方案。例如，FEDAVGM [16]引入服务器动量，以平滑全局更新并减少由本地模型不一致导致的波动。针对客户端数据贡献不均的问题，FEDNOVA [30] 对聚合权重进行归一化，以平衡本地更新的整合，从而确保全局过程的公平性和稳定性。FEDDISCO[17] 进一步通过综合考虑客户端差异度量和数据量来改进聚合，在严格的理论保证下实现更紧的误差界。除聚合策略外，FEDFTG [18]还采用无数据知识蒸馏机制对全局模型进行微调，无需原始数据即可实现从本地模型中迁移知识。此外，FEDMR [31] 引入了一种逐层打乱策略

在更新过程中对全局模型参数进行处理，从而促进模型收敛至更平坦的最优解，并提升鲁棒性。

尽管这些方法提升了全局模型聚合和泛化性能，但它们主要关注聚合后的增强，却忽视了数据异质性和有偏的本地训练所引发的上游挑战。相比之下，FLOOD 通过双重加权策略，整合客户端侧损失重新加权与服务器端聚合调整，明确针对这些根本原因。这使 FLOOD 能够全面应对非独立同分布数据导致的性能下降，并弥合联邦学习中本地训练校正与全局模型调整之间的差距。

B. 分布外检测

深度神经网络模型是现代数据驱动服务系统的核心智能引擎，通常在训练数据和分布内（ID）测试样本上都能取得出色的性能 [32]。然而，在面向服务的现实部署中，这些模型会持续接触范围广泛的分布外输入，这些输入源于动态用户行为、多样化的服务环境以及不断变化的数据源 [33]。由于模型在识别此类输入方面存在固有局限，误分类分布外样本可能会引发严重的决策错误，并在高风险服务场景中造成难以控制的风险。因此，分布外检测的首要目标是赋予模型准确识别并拒绝分布外输入的能力，以避免其对服务准确性、系统可靠性和整体服务质量造成不利影响。

在现有方法中，事后评分方法因其简洁性、计算效率以及与预训练模型的良好兼容性而尤为引人关注。这些方法无需修改模型架构，而是利用模型输出（如逻辑值或中间特征）进行分布外分类。因此，它们能够轻松集成到现有流程中，且只需增加极少的计算开销。沿着这一思路，研究者提出了多种事后分布外评分方法来量化预测置信度，大致可分为基于逻辑值和基于特征 的方法。作为最早且最具代表性的基于逻辑值的方法之一，Hendrycks等人 [22]采用逻辑值的最大 Softmax 概率作为分布外评分函数。沿着这一思路，Energy [24] 进一步基于逻辑值推导出的能量函数构造 OOD 分数，从而提供更具表达力的置信度度量。对于基于特征的方法，ReAct [34] 通过截断超过预定义阈值的特征值，减轻异常值激活的影响，从而抑制隐藏表示中的异常响应。ASH [35]进一步提出了一种简单的特征抑制策略，即移除大量样本特征，同时有选择地调整其余特征。

III. 预备知识

A. 联邦学习问题设置

联邦学习涉及一组 N客户端，记为 $\mathcal{C}_1$，$\mathcal{C}_2$，…，$\mathcal{C}_N$，其中每个客户端 $\mathcal{C}_k$ 都持有一个本地数据集 $\mathcal{D}_k$。其目标是联邦学习旨在协同优化一个全局模型 θ，同时保护客户端数据的隐私。该目标通过最小化所有客户端的总体损失来实现，其形式化表示为

$$\arg\min_{\theta} \left\{ \mathcal{F}(\theta) = \sum_{k=1}^{N} p_k \mathcal{L}_k(\theta) \right\}, \tag{1}$$

其中，$\mathcal{L}_k(\theta)$ 表示客户端 $\mathcal{C}_k$ 的局部经验损失，定义如下：

$$\mathcal{L}_k(\theta) := \frac{1}{n_k} \sum_{i=1}^{n_k} \Phi(\theta; x_i, y_i). \tag{2}$$

这里，(x_i, y_i) 表示客户端 $\mathcal{C}_k$ 上的第 i 个训练样本，n_k 是该客户端上的样本数，$\Phi(\theta; x, y)$ 表示逐样本损失函数。聚合权重 p_k 与客户端数据集的大小成正比，定义如下：

$$p_k = \frac{n_k}{\sum_{i=1}^{N} n_i}. \tag{3}$$

在第 t 个通信轮次中，选择一部分客户端 $\mathcal{C}^t \subseteq \mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N$ 参与训练过程。每个被选中的客户端 $\mathcal{C}_k \in \mathcal{C}^t$ 从服务器接收当前全局模型 θ^t，在其私有数据集 $\mathcal{D}_k$ 上执行若干步局部优化，并将更新后的本地模型 θ_k^t 上传至服务器。随后，服务器聚合收到的本地模型，以更新全局模型：

$$\theta^{t+1} = \sum_{k:\mathcal{C}_k \in \mathcal{C}^t} p_k \theta_k^t. \tag{4}$$

这一过程在多个通信轮次中重复进行，直到全局模型 θ 收敛。

B. 事后 *OOD* 分数

设 $\mathcal{D}_{\text{id}}$ 和 $\mathcal{D}_{\text{ood}}$ 分别表示 ID 数据集和 OOD 数据集。分布外检测的目标是判断给定样本（x，y）属于 $\mathcal{D}_{\text{id}}$ 还是 $\mathcal{D}_{\text{ood}}$。对于给定的 DNN 模型 θ，事后 OOD 检测技术会采用启发式评分函数 $\mathrm{S}(\cdot;\cdot)$，为每个输入样本（x，y）分配一个 OOD 置信度分数。MSP[22] 和 Energy [24] 等代表性方法采用不同的评分策略。具体而言，MSP 会计算模型输出的最大 softmax 概率，而 Energy 则会评估预测概率分布的“能量”值。这些分数可有效指示测试样本是否符合训练分布。在实际应用中，样本会根据以下判定标准被分类为 ID 或 OOD：

$$(x, y) \sim \begin{cases} \mathcal{D}_{\text{ood}}, & \text{if } \mathrm{S}(\theta; x) \leq \tau, \\ \mathcal{D}_{\text{id}}, & \text{if } \mathrm{S}(\theta; x) > \tau, \end{cases} \tag{5}$$

其中 τ 是一个预定义阈值，用于区分 $\mathcal{D}$分布内与 $\mathcal{D}$分布外。如式（5）所述，置信度分数高于 τ的样本被判定为分布内，而分数低于 τ的样本被识别为分布外。这种基于阈值的判定规则计算高效，且引入的部署开销可忽略不计，使事后分布外检测方法特别适用于资源受限场景，或作为真实世界系统中预训练模型的轻量级插件模块。

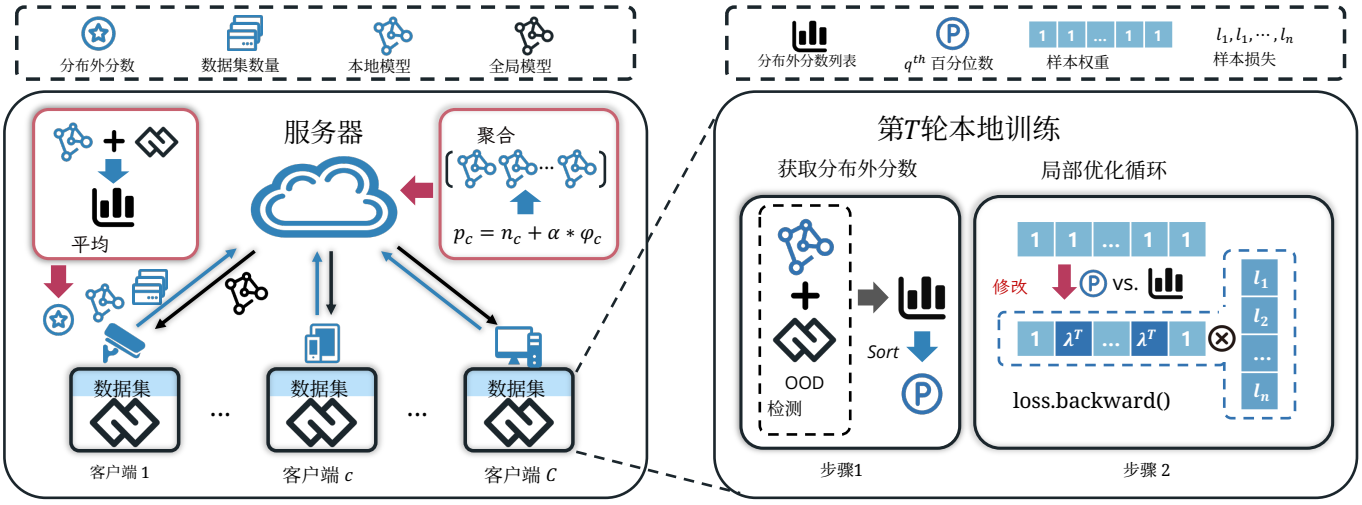


图 2：FLOOD 的框架与工作流程。右侧展示了客户端训练过程，其中使用分布外检测来调整监督损失中的样本权重。左侧展示了服务器端流程，服务器根据客户端提供的信息动态重新分配聚合权重。

IV. FLOOD 框架

在本节中，我们提出FLOOD（通过分布外检测实现的联邦学习），这是一种新颖的框架，通过将分布外检测融入客户端训练与服务器端聚合，以增强非独立同分布数据设置下的联邦学习。FLOOD的核心思想是，依据其与不断演化的全局模型之间估计得到的分布对齐程度，动态调整单个样本在局部优化过程中的重要性，以及不同客户端在全局聚合过程中的可靠性。通过将分布外感知作为自适应控制信号，FLOOD建立了一种原则性机制，以减轻数据异质性的不利影响。

FLood 的架构和工作流程如图 2所示。该框架引入了两个紧密耦合的组件，即自适应样本加权（第IV-B节）和动态聚合校正（第IV-C节），二者共同构成双重加权策略。具体而言，前者在客户端优化过程中突出本地低置信度样本，而后者则根据模型在本地数据集上的置信度，在服务器端调整聚合权重。

A. 概述

图2 展示了 FLOOD 的框架与工作流程。不同于传统联邦学习方法，FLOOD 在客户端优化过程中，根据单个样本的分布外置信度分数分配动态权重，从而自适应地调节训练过程。同时，它根据各客户端报告的分布外信息调整聚合权重，改进服务器端聚合。该双重加权策略共同增强了全局模型的鲁棒性和泛化能力。FLOOD 的每个训练轮次 t 包括以下步骤：

1) 客户端选择与模型广播：在第 t轮训练开始时，服务器随机选择一组客户端 $\mathcal{C}^t$，并将当前全局模型分发给 θ^t 每个选定的客户端。此步骤遵循传统联邦学习框架的标准协议。

2) 分布外感知局部训练： 收到全局模型 θ^t后，每个客户端都会对其本地数据集进行分布外检测，以评估样本级置信度分数，并根据预定义阈值将数据划分为真实分布内样本和伪分布外（异常）子集。在局部训练阶段，FLOOD 会为伪分布外样本分配更高的损失权重，从而优先学习这些难分类数据点。请注意，随着模型不断演进，样本的分类可能会在训练迭代过程中动态变化。

3) 分布外置信度聚合与上传： 完成本地训练后，每个客户端会将其本地数据集的分布外置信度分数聚合为单个标量 ϕ_c。聚合后的置信度分数 ϕ_c 将与更新后的本地模型 θ_c^t 和数据集大小 n_c一同上传至服务器。纳入 ϕ_c 后，服务器便能据此对全局聚合权重进行合理调整。

4) 置信度引导的全局聚合： 服务器根据收到的本地模型、数据集规模和聚合后的分布外置信度分数，计算用于更新全局模型的新聚合权重。置信度分数较高的客户端 ϕ_c 会获得更大的聚合权重，因此对全局更新 θ^{t+1}贡献更大，从而提升模型在异质数据分布下的鲁棒性和泛化能力。

算法 1 概述了服务器端逻辑，其中蓝色模块表示相较于标准联邦学习框架（如 Food 、FedAvg [13]所引入的联邦学习。第 1 行初始化全局模型 θ^0，第 3-9 行描述了每个通信轮次中的核心步骤。每一轮中，服务器会随机选择 M 个客户端组成活跃集 $\mathcal{C}^t$，如第 3 行所述；随后按照第 4行所示，将当前全局模型 θ^t 广播给这些客户端。第 5-7 行（对应并行执行模块）展示了客户端

算法 1 FLOOD（服务器端）

输入：R：通信轮次数量；N：客户端总数；M：每轮选取的客户端数量。
输出：θ^T：经过 T 轮后的最终全局模型。

1：初始化全局模型：θ^0
2：**对于** $t = 0$ 到 $R-1$ **执行**
3：　随机选择 M 个客户端：$\mathcal{C}^t \subseteq \mathcal{C}$
4：　将 θ^t 广播给 $\mathcal{C}^t$ 中的所有客户端
5：　**并行执行**每个客户端 $c \in \mathcal{C}^t$ **的操作**
6：　　$\theta_c^t,\ n_c,\ \phi_c \leftarrow$ 局部更新$(\theta^t,\ t)$
7：　**结束循环**
8：　通过式（11）计算客户端权重：$p_c = f(\phi_c)$
9：　聚合：$\theta^{t+1} \leftarrow \sum_{c \in \mathcal{C}^t} p_c \theta_c^t$
10：**结束循环**
11：**返回** θ^T

算法 2 FLOOD 的局部更新（在客户端 c 上）

输入：θ^t：第 t 轮的全局模型；E：本地训练轮数；$\mathcal{D}_c$：客户端的本地数据集；η：学习率；T：权重增长的停止轮次。
输出：θ_c^t：更新后的本地模型；n_c：样本数量；ϕ_c：聚合的分布外（OOD）分数。

1：初始化本地模型：$\theta_c^t \leftarrow \theta^t$
2：**对于** $e = 0$ 至 $E-1$ **执行**
3：　**对于** 小批次 $(X_b,\ Y_b) \sim \mathcal{D}_c$ **执行**
4：　　计算基础损失：Loss $\leftarrow \Phi(\theta_c^t;\ X_b,\ Y_b)$
5：　　通过公式（6）和（7）计算分数，τ
6：　　通过公式（8）计算掩码
7：　　加权损失：损失 $\leftarrow$ MEAN(损失 $\circ$ 掩码)
8：　　更新模型：$\theta_c^t \leftarrow \theta_c^t - \eta \nabla_{\theta^t}\ {}_c\texttt{Loss}$
9：　**结束循环**
10：**结束循环**
11：汇总分布外分数：$\phi_c \leftarrow \frac{1}{|\mathcal{D}_c|} \sum_{x \in \mathcal{D}_c}$ 分数$_x$
12：**返回** $\theta_c^t, |\mathcal{D}_c|, \phi_c$

端的更新过程，其中每个活跃客户端都会获取其本地模型θ_c^t、数据集大小 n_c以及聚合后的分布外置信度分数 ϕ_c（第 6行）。第 8行中，服务器根据 p_c，使用（ ϕ_c 11）计算每个客户端的聚合权重。第 9 行将所有本地模型聚合，以生成更新后的全局模型 θ^{t+1}。经过 R轮后，第 11 行输出最终的全局模型 。

B. 自适应样本加权

为解决客户端内数据异质性问题，FLOOD 利用事后 OOD 检测技术，找出偏离预期数据分布的样本。具体而言，FLOOD 会为每个样本计算一个 OOD 置信度分数，该分数反映模型对其主要任务预测结果的信心。分数越高，表示预测确定性越强；分数越低，则意味着不确定性越高。通过应用预定义的阈值，可将本地数据集划分为两类：真实分布内（真实 *ID*）样本和伪分布外（伪分布外）样本。直观来说，不确定性较高的伪分布外样本需要得到针对性学习侧重，而预测可靠的真实 ID 样本则不需要太多干预。

算法2概述了每个客户端执行的 LOCALUPDATE 过程，其中蓝色标注的行表示集成了我们的自适应加权机制。第1 行使用从服务器接收的全局模型 θ_c^t 初始化客户端模型。第2–10 行描述了进行E 个训练轮次的本地优化过程。在每个训练轮次中，客户端按照第3行的规定，遍历从本地数据集 $\mathcal{D}_c$中抽取的小批次。第4 行计算当前小批次中每个样本的标准监督损失，记为 Loss $= [l_1, l_2, \ldots]$。在模型更新之前，FLOOD 会为小批次中的每个输入样本 x 计算置信度分数（$X_b,\ Y_b$）：

$$\text{SCORE}_x := \text{S}(\theta_c^t, x), \tag{6}$$

其中，S(·; ·) 表示所选的分布外（OOD）评分函数，例如基于能量的评分 [24] 或最大软最大概率（MSP） [22]。批次中各样本的分数向量记为 Score $=$ [SCORE$_1$，SCORE$_2$，$\ldots$]。

随后，将分数向量的第 q 个百分位数确定为动态阈值 τ：

$$\tau = [\texttt{Score}]_{q\%}. \tag{7}$$

该阈值能够识别伪分布外样本，随后通过第6行的掩码操作为其分配更高的损失权重。重加权损失在第7行计算，本地模型据此在第8行相应更新。最后，在完成全部本地训练轮数后，客户端在其整个数据集上聚合 OOD 分数以得到 ϕ_c，如第11行所示，并在第12行返回更新后的模型、数据集大小和 OOD 置信度分数。

利用该阈值，我们将每个样本分类为真实 ID（$\text{SCORE}_x \geq \tau$）或伪分布外（$\text{SCORE}_x < \tau$）。为优先学习伪分布外样本，FLOOD 采用自适应加权机制。每个样本 (x, y) 被赋予权重掩码，定义如下：

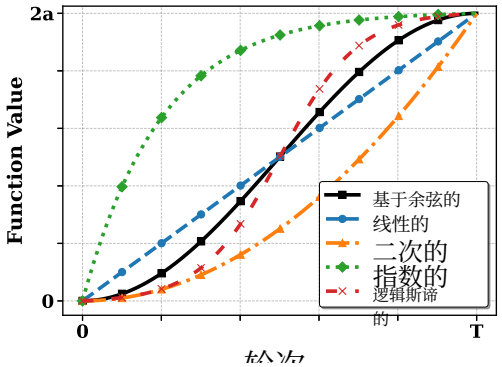
$$\texttt{Mask}_x^t = \lambda^t \cdot \mathbb{I}(\text{SCORE}_x < \tau) + \mathbb{I}(\text{SCORE}_x \geq \tau), \tag{8}$$

其中 $\mathbb{I}(\cdot)$为指示函数。对于置信度分数大于阈值 τ 的样本，其在掩码中的权重设为 1；否则，其权重设为 λ^t。

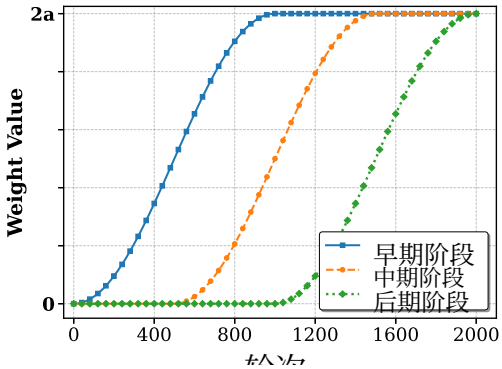
为了更好地促进模型对这些伪分布外样本的学习，我们在训练过程中采用逐步递增的权重调度策略。随着模型能力增强，权重增长会在第 T 轮停止，以避免对真实分布内样本的学习产生不利影响。此外，由于模型在训练早期的分布外检测能力并不可靠，权重增长过快可能会导致训练不稳定性。随着权重增长逐渐接近停止点，其增长速率也应随之降低。为实现这一效果，FLOOD 使用余弦函数来调节权重增长，具体公式如下：

$$\lambda^t = a \cdot \left(1 - \cos\frac{\pi t}{T}\right), \tag{9}$$

其中 a控制最大放大因子，T 表示 λ^t 达到其平台期的轮次。该调度



(a) 调度函数。



(b) 递增阶段。

图 3：FLOOD 中权重调整曲线示意图。(a) 不同权重调度函数的比较，包括基于余弦、线性、二次、指数和逻辑斯谛的调度方案。(b) 不同起始阶段下的权重增长。

策略缓解了由不成熟的分布外估计引起的早期不稳定性，同时随着模型随时间变得越来越可靠，逐步增强伪分布外样本的影响。

除余弦调度方案外，我们还进一步研究了几种替代加权函数和增长周期，以考察不同动态变化对其有效性的影响。候选函数定义如下：

- **Linear:** $\lambda^t = 2a \cdot \frac{t}{T}$
- **Quadratic:** $\lambda^t = 2a \cdot (\frac{t}{T})^2$
- **Exponential:** $\lambda^t = 2a \cdot \frac{1-\exp(-kt)}{1-\exp(-kT)}$
- **Logistic:** $\lambda^t = 2a \cdot \frac{\sigma(\alpha(t-T/2))-\sigma(-\alpha T/2)}{\sigma(\alpha T/2)-\sigma(-\alpha T/2)}$, where $\sigma(\cdot)$ denotes the logistic sigmoid function

图 3a 展示了这些调度函数的增长模式，说明了它们在初始增长速度和收敛行为方面的差异。为进一步分析时间特征，我们还改变了权重增长的起始点，并考虑了三种配置：早期开始、中期开始和晚期开始，如图 3b所示。

C. 动态聚合校正

借鉴 FEDDISCO [17] 和 FEDNOVA[30], 的思路，FLOOD 引入了一种动态聚合机制，根据客户端的分布外置信度分数调整客户端权重。这一设计通过自适应地反映本地更新的质量，增强了全局模型的鲁棒性。相比之下，FEDDISCO 依赖静态数据差异度量，而 FEDNOVA 根据有效迭代次数而非模型可靠性调整聚合权重。

完成本地训练后，每个客户端 c 在其本地数据集 $\mathcal{D}_c$上执行分布外检测。OOD 分数的计算开销较低，因为不涉及模型更新，从而最大限度地减少了计算开销。对于客户端数据集中的每个样本（x， y），使用函数 $\mathrm{S}(\theta_c^t; x)$获取 OOD 置信度分数，其中 θ_c^t 表示第t轮次的本地模型。随后计算客户端 c的聚合的分布外（OOD）分数 ϕ_c：

$$\phi_c = \frac{1}{n_c} \sum_{x \in \mathcal{D}_c} \mathrm{S}(\theta_c^t; x), \tag{10}$$

其中， n_c 表示本地数据集 $\mathcal{D}_c$的大小。标量值 ϕ_c 与更新后的本地模型 θ_c^t 和数据集大小 n_c一同传输至服务器，几乎不会增加通信开销。此外，这种设计能够保护数据隐私，因为 ϕ_c 不会泄露有关数据集类别分布或敏感特征的具体信息。

为将分布外分数 ϕ_c 纳入聚合过程，FLOOD 扩展了传统的基于数据量的加权方案。具体而言，客户端k 的调整后聚合权重定义如下：

$$p_c = \mathrm{NORM}\Big(\mathrm{NORM}(n_c) + \alpha \cdot \mathrm{NORM}(\phi_c)\Big), \tag{11}$$

其中，NORM(·)表示归一化操作，用于将各分量缩放至相近的范围； α 是一个可调参数，用于控制分布外分数对最终聚合权重的相对影响。该公式确保分布外分数较高的客户端（这表明它们对本地数据分布的理解更有把握）能够对全局模型更新作出更大贡献。

通过将数据量和分布外置信度同时纳入聚合权重，FLOOD 在利用大型数据集与优先采用高质量、高置信度本地模型更新之间实现了平衡。因此，全局模型能够更好地抵御数据异质性，有效捕捉不同客户端的多样化模式，同时降低对任何特定数据子集过拟合的风险。此外，这种动态加权方案使 FLOOD 能够根据不断变化的数据分布和模型能力，自适应地调整各客户端的贡献。随着全局模型不断改进，各客户端更新的相对重要性也会自然变化，从而形成更具韧性和泛化能力的联邦模型。

D. 收敛性分析

现在，我们在联邦优化的标准假设下，对 FLOOD 进行收敛性分析。设 $F(\theta) \sum_{c=1}^{N} \frac{n_c}{n} \mathcal{L}_c\ \theta$ $x\ y$ $\mathcal{L}_c\ \theta$ $= (\ )$ 表示全局目标函数，其中 $(\ )\ \mathbb{E}_{(x,y)\sim\mathcal{D}_c}[\Phi(\theta; \sum_{c=1}^{N} n_c$ ，$)$] 是客户端上的真实本地风险，而 。

在 FLOOD 中，每个客户端 c在轮次 t优化加权代理损失：

$$\widetilde{\mathcal{L}}_c^t(\theta) := \frac{1}{n_c} \sum_{i=1}^{n_c} w_{c,i}^t\, \Phi(\theta; x_i, y_i), \tag{12}$$

其中，样本权重$w_{c,i}^t$ = 掩码$_{x_i}^t > 0$ 由当前全局模型的分布外置信度分数决定$\tilde{\theta}^t$（见（8））。注意，$\mathcal{L}_c^t$ 会随时间变化，因为它取决于θ^t。本地训练后，服务器使用由报告的OOD置信度分数计算得到的动态权重 p_c^t 进行更新聚合 ϕ_c^t（见（11））。全局更新为：

$$\theta^{t+1} = \sum_{c \in \mathcal{C}^t} p_c^t \, \theta_c^t, \tag{13}$$

其中，θ_c^t 是客户端 c返回的本地模型。

为分析收敛性，我们作出以下假设：(A1) 每个 $\mathcal{L}_c$ 都是 L-光滑的，即对于所有 θ、θ'，$\|\nabla\mathcal{L}_c(\theta) - \nabla\mathcal{L}_c(\theta')\| \le L\|\theta - \theta'\|$。(A2) 对于所有和，随机梯度满足 $\mathbb{E}_{(x,y)}[\|\nabla\Phi(\theta;\, x, y)$ $\nabla\mathcal{L}_c\ \theta\ \|^2] \le \sigma^2$ c θ ()。(A3) 存在常数 $W_{\min}$、$W_{\max} > 0$，使得对于所有 c、i、t，$W_{\min} \le w_{c,i}^t \le W_{\max}$。这是按设计成立的，因为 $\lambda^t \le 2a$ (第 IV-B 节)。（A4）将本地梯度偏差定义为 b_c^t：$= \nabla\mathcal{L}_c^t(\theta^t)$ $-\nabla\mathcal{L}_c(\theta^t)$。我们假设对于某个 $\delta \ge 0$ 及所有 c、t，有 $\|b_c^t\| \le \delta$。(A5) 令 $q_c = n_c/n$ 为理想的数据比例权重。实际聚合权重 p_c^t 满足

$$\left\| \sum_{c \in \mathcal{C}^t} p_c^t \nabla\mathcal{L}_c(\theta^t) - \nabla F(\theta^t) \right\| \le \epsilon, \tag{14}$$

对于某些 $\epsilon \ge 0$。

假设 A4 和 A5 刻画了 FLOOD 的 OOD感知机制所引入的偏差。在实践中，经过良好校准的 OOD评分和适度的加权调度（例如余弦退火）可确保 δ 和 ϵ 保持较小。在这些假设下，我们得到如下收敛保证。

定理 1（联邦学习OOD的收敛性）。假设学习率设置为 $\eta = \frac{1}{\sqrt{T}}$。则由 联邦学习OOD 生成的迭代序列 $\{\theta^t\}_{t=0}^{T-1}$满足

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla F(\theta^t)\right\|^2\right] \le \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left((\epsilon+\delta)^2\right), \tag{15}$$

其中 ϵ和 δ 分别定义于假设*A4*和*A5*中。

证明见补充材料。

V. 实验

在本节中，我们使用 PyTorch [36] 实现 FLOOD，以评估其性能。[1] 所有实验均在一台 Ubuntu 工作站上进行，该工作站配备 512 GB 内存、一颗 AMD EPYC 7Y43 CPU 和两块 NVIDIA GeForce RTX 4090 GPU。为系统评估所提出框架的有效性和通用性，我们围绕四个核心研究问题（RQs）组织实验研究：

RQ1（优越性）：在不同数据异质性水平下，FLOOD 是否优于最先进的联邦学习方法?

RQ2（可扩展性）：在实际应用中，FLOOD 在不同联邦学习设置下的扩展表现如何?

RQ3（兼容性）：FLOOD 能否有效集成到现有的客户端训练校正方法中?

RQ4（消融）：FLOOD 的所有关键组件是否都能提升其整体性能？该框架对超参数有多敏感?

表 I：不同联邦学习基线方法的超参数设置。

基线	超参数	值
FEDAVG	None	–
FEDPROX	μ：正则化系数	$\mu = 0.1$
SCAFFOLD	None	–
MOON	μ：正则化系数 τ：温度	$\mu = 1.0$, $\tau = 0.5$
FEDDECORR	β：正则化系数	$\beta = 0.1$
FEDNP	λ：正则化系数	$\lambda = 0.01$
FEDAVGM	ρ：服务器动量	$\rho = 0.1$
FEDNOVA	None	–
FEDDISCO	a：平衡系数；调整系数 b	$a = 0.5$, $b = 1$
FEDMR	None	–
FLOOD	q：分布外检测阈值 a：放大系数 T：停止轮次 α：聚合比率	$q = 0.7$ $a = 200$ $T = 1000$ $\alpha = 0.5$

A. 设置

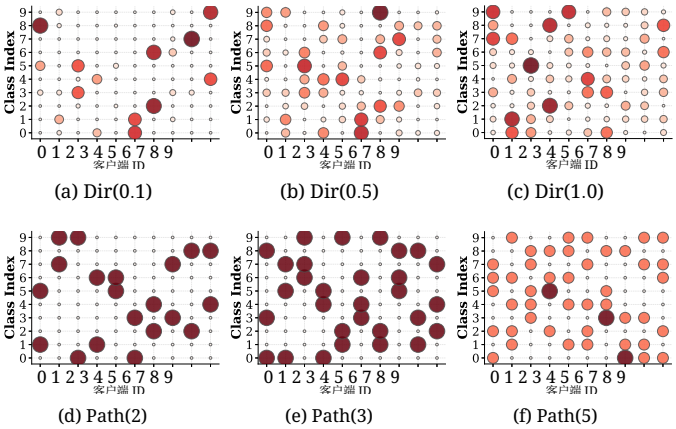
a) 基线方法：我们将 FLOOD 与十种最先进的联邦学习方法进行比较，包括经典基线 FEDAVG [13]，、五种客户端本地训练校正方法（FEDPROX [19]，、SCAFFOLD [20]，、MOON [21]，、FEDDECORR [14] 和 FEDNP [37]，），以及四种服务器端全局模型调整方法（FEDAVGM [16]，、FEDNOVA[30]，、FEDDISCO [17]，和 FEDMR [31]）。对于 FLOOD，我们采用 ENERGY分数 [24] 作为默认的分布外检测方法。在客户端，分布外检测阈值 q 设为 70%，放大系数 a 设为 200，权重稳定轮次 T 设为 1000。在服务器端，聚合系数 α 固定为 0.5。所有基线方法的详细超参数配置汇总于表 I。

b) 数据集和模型：我们在三个广泛使用的基准数据集上评估我们的方法：CIFAR-10 [38]，、CIFAR-100 [38]，和 SVHN [39]。为模拟联邦学习中的非独立同分布（*i.i.d.*）数据分布，我们遵循两种广泛采用的实验方案 [13], [40]，，即狄利克雷分布 Dir(β) 和病态分布 Path(r)，对数据进行划分。对于狄利克雷分布，我们在实验中考虑 $\beta \in \{0.1, 0.5, 1.0\}$。对于病态分布，参数 r 设置为 $\{2, 3, 5\}$。β和 r 的值越小，表示数据异质性程度越高。图4展示了两种划分方案下数据分布差异的可视化结果。此外，为考察 FLOOD 在不同模型架构下的鲁棒性，我们使用两种具有代表性的神经网络模型 ResNet-8 [41] 和 MobileNet-V2 [42]开展实验。

[1]我们的代码可在以下地址获取：https://github.com/LingAndZero/FLood

表 II：在使用 ResNet-8 的狄利克雷异质设置上的分类准确率结果（均值 ± 标准差）。所有数值均以百分比表示，最佳和次佳结果分别以**粗体** 和下划线标出。

基线	CIFAR-10			CIFAR-100			SVHN		
	Dir(0.1)	*Dir*(0.5)	*Dir*(1.0)	*Dir*(0。1)	*Dir*(0.5)	*Dir*(1。0)	*Dir*(0。1)	*Dir*(0.5)	*Dir*(1.0)
FEDAVG	$72.79_{(5.95)}$	$90.18_{(0.25)}$	$90.82_{(0.11)}$	$65.51_{(0.24)}$	$67.26_{(0.13)}$	$67.59_{(0.11)}$	$86.49_{(7.44)}$	$94.62_{(0.13)}$	$94.94_{(0.03)}$
FEDPROX	$71.48_{(6.42)}$	$90.48_{(0.29)}$	$90.66_{(0.09)}$	$65.55_{(0.25)}$	$67.31_{(0.16)}$	$67.43_{(0.15)}$	$84.94_{(7.95)}$	$94.70_{(0.13)}$	$94.82_{(0.05)}$
SCAFFOLD	$70.61_{(3.55)}$	$90.73_{(0.23)}$	$\underline{91.89}_{(0.20)}$	$67.59_{(0.28)}$	$70.38_{(0.16)}$	$\underline{70.77}_{(0.11)}$	$91.68_{(1.90)}$	$95.06_{(0.25)}$	$95.33_{(0.11)}$
MOON	$73.72_{(4.55)}$	$90.48_{(0.26)}$	$90.78_{(0.11)}$	$65.90_{(0.23)}$	$67.65_{(0.14)}$	$67.00_{(0.14)}$	$85.83_{(7.25)}$	$94.79_{(0.13)}$	$94.89_{(0.04)}$
FEDDECORR	$70.53_{(6.69)}$	$90.34_{(0.29)}$	$90.57_{(0.09)}$	$64.21_{(0.26)}$	$66.11_{(0.14)}$	$65.70_{(0.16)}$	$85.04_{(8.78)}$	$94.60_{(0.13)}$	$94.65_{(0.06)}$
FEDNP	$66.60_{(8.28)}$	$89.57_{(0.48)}$	$90.41_{(0.17)}$	$64.78_{(0.31)}$	$67.18_{(0.14)}$	$67.39_{(0.14)}$	$82.44_{(10.49)}$	$94.17_{(0.30)}$	$94.56_{(0.10)}$
FEDAVGM	$70.29_{(7.52)}$	$90.51_{(0.26)}$	$90.95_{(0.11)}$	$65.53_{(0.22)}$	$67.59_{(0.15)}$	$67.22_{(0.11)}$	$85.42_{(7.57)}$	$94.85_{(0.13)}$	$94.93_{(0.05)}$
FEDNOVA	$71.16_{(6.56)}$	$90.26_{(0.30)}$	$91.09_{(0.13)}$	$66.09_{(0.23)}$	$67.60_{(0.10)}$	$67.48_{(0.13)}$	$84.93_{(5.66)}$	$94.79_{(0.14)}$	$94.89_{(0.05)}$
FEDDISCO	$\underline{75.75}_{(4.66)}$	$90.32_{(0.34)}$	$90.73_{(0.14)}$	$65.75_{(0.23)}$	$67.35_{(0.13)}$	$67.85_{(0.14)}$	$89.25_{(3.82)}$	$94.72_{(0.11)}$	$95.03_{(0.06)}$
FEDMR	$36.03_{(6.19)}$	$\underline{90.74}_{(0.24)}$	$91.69_{(0.13)}$	$\underline{67.68}_{(0.20)}$	$\underline{70.56}_{(0.10)}$	$69.88_{(0.12)}$	$\underline{92.91}_{(1.37)}$	$\underline{95.60}_{(0.05)}$	$\underline{95.76}_{(0.04)}$
FLOOD（我们的方法）	$\mathbf{84.43}_{(1.66)}$	$\mathbf{92.92}_{(0.09)}$	$\mathbf{93.28}_{(0.08)}$	$\mathbf{70.98}_{(0.13)}$	$\mathbf{71.30}_{(0.17)}$	$\mathbf{71.44}_{(0.12)}$	$\mathbf{94.50}_{(0.34)}$	$\mathbf{95.67}_{(0.05)}$	$\mathbf{95.85}_{(0.04)}$



(a) Dir(0.1) (b) Dir(0.5) (c) Dir(1.0)

(d) Path(2) (e) Path(3) (f) Path(5)

图 4：狄利克雷和病态非独立同分布划分协议下各客户端的数据分布。

c) 联邦学习实现：我们考虑一个包含 100 个客户端的联邦学习设置。为更好地反映通信资源有限的实际联邦学习场景，每个通信轮次仅随机选择 10% 的客户端，即 10 个客户端参与。服务器执行 2000 个通信轮次，每个被选中的客户端在每轮进行 5 个本地训练轮次，批量大小为 50。所有客户端均采用随机梯度下降优化器进行局部优化，学习率为 0.01，动量为 0.9，权重衰减为 0.0005。此外，在每轮结束时应用 0.998 的学习率衰减因子，以促进训练过程中的稳定收敛。

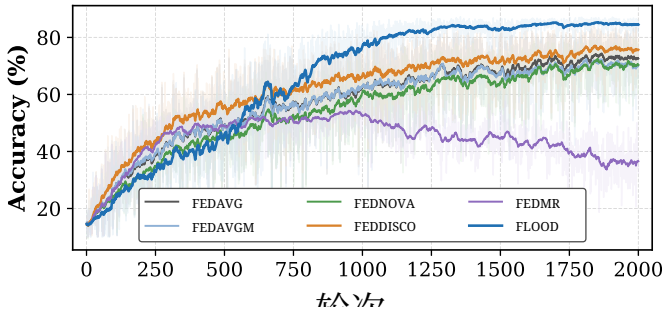
B. 与最先进基线方法的比较（*RQ1*）

表II 和III对比了 FLOOD 与具有代表性的联邦学习基线方法在三种基准数据集、两种异质性协议下的综合性能。为确保评估公平且稳定，我们报告了最后 50 个通信轮次中的平均分类准确率和标准差。从表II和III可以明显看出 FLOOD 在所有异质性设置和数据集上始终取得最佳或接近最佳的性能。在狄利克雷分布下，FLOOD 相较于客户端校正方法和服务器端聚合方法均表现出明显且持续的提升。例如，在 Dir(0.1) 设置下，FLOOD 在 CIFAR-10 数据集上的准确率比次优基线高出 8.68%，彰显了其缓解严重数据异质性的强大能力。在 CIFAR-100 和 SVHN 上也观察到类似的性能差距，表明 FLOOD 能够适应复杂程度各异的数据集，具有较强的鲁棒性。值得注意的是，与 Dir(1.0) 或 Path(5) 等较温和的场景相比，在 Dir(0.1) 和 Path(2) 等高度异质的设置下，FLOOD 的性能提升更加显著。这一趋势表明，当数据异质性对联邦优化构成重大挑战时，FLOOD 尤其有效。一种可能的解释是，在数据异质性较低的场景中，真实 ID 样本与伪分布外样本之间的差异不那么明显，因此分布外感知重加权带来的潜在收益也会降低。相反，在高度异质的条件下，FLOOD 能够更准确地识别分布不匹配样本和不可靠的客户端更新，从而带来显著更大的性能提升。我们还观察到，FEDMR 在少数设置下表现出具有竞争力或略优的性能。不过，在这些情况下，它与 FLOOD 之间的性能差距相对较小。相比之下，在大多数评估场景中，尤其是在数据异质性严重的情况下，FLOOD 始终明显优于 FEDMR。

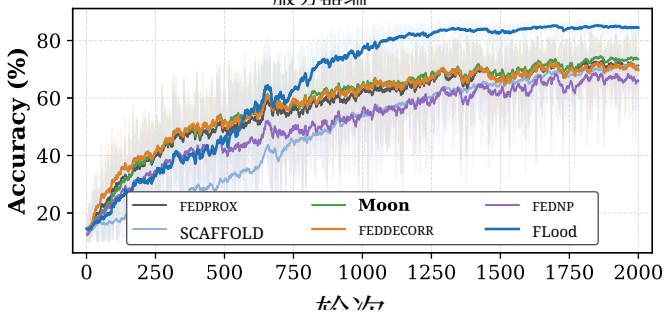
我们在图5中进一步展示了不同联邦学习方法的训练动态。如两组对比所示，准确率曲线揭示了 FLOOD 特有的学习轨迹。在训练初期，FLOOD 的表现起初落后于大多数基线方法。这一现象符合预期，因为此阶段的全局模型判别能力有限，其分布外检测信号也尚不成熟。因此，真实 ID 样本与伪分布外样本之间的区分不够可靠，可能导致本地更新期间的权重分配欠佳。随着训练进入中期，

表 III：在病理性异构设置下使用 ResNet-8 的分类准确率结果（均值 ± 标准差）。所有数值均以百分比表示，最佳和次佳结果分别以**粗体**和下划线标出。

基线	CIFAR-10			CIFAR-100			SVHN		
	Path(2)	Path(3)	Path(5)	Path(2)	Path(3)	Path(5)	Path(2)	Path(3)	Path(5)
FEDAVG	$43.53_{(4.53)}$	$71.33_{(5.55)}$	$89.72_{(0.51)}$	$8.64_{(2.20)}$	$26.22_{(3.45)}$	$53.62_{(1.66)}$	$80.00_{(7.59)}$	$91.35_{(3.67)}$	$94.76_{(0.23)}$
FEDPROX	$44.68_{(4.50)}$	$72.55_{(5.43)}$	$89.80_{(0.57)}$	$9.24_{(2.01)}$	$27.58_{(3.08)}$	$53.44_{(1.67)}$	$79.39_{(7.84)}$	$91.75_{(3.48)}$	$94.73_{(0.24)}$
SCAFFOLD	$50.16_{(5.17)}$	$73.67_{(4.51)}$	$\underline{90.27}_{(0.57)}$	$9.04_{(2.20)}$	$29.49_{(2.78)}$	$54.56_{(1.35)}$	$79.50_{(7.08)}$	$91.39_{(4.23)}$	$95.26_{(0.10)}$
MOON	$45.05_{(5.07)}$	$72.71_{(5.40)}$	$89.83_{(0.46)}$	$8.49_{(2.37)}$	$26.86_{(3.17)}$	$54.09_{(1.59)}$	$78.39_{(7.64)}$	$91.41_{(3.46)}$	$94.71_{(0.30)}$
FEDDECORR	$\underline{61.21}_{(5.15)}$	$\underline{78.42}_{(3.94)}$	$89.73_{(0.42)}$	$\underline{12.02}_{(2.69)}$	$\underline{30.11}_{(3.28)}$	$\underline{55.32}_{(0.87)}$	$83.13_{(6.42)}$	$92.50_{(2.34)}$	$94.53_{(0.37)}$
FEDNP	$43.74_{(5.54)}$	$70.16_{(6.09)}$	$88.66_{(0.99)}$	$8.50_{(2.90)}$	$26.69_{(3.14)}$	$51.42_{(1.83)}$	$78.30_{(8.86)}$	$90.33_{(4.58)}$	$94.18_{(0.36)}$
FEDAVGM	$42.29_{(4.07)}$	$71.60_{(5.71)}$	$90.08_{(0.59)}$	$7.47_{(1.95)}$	$24.63_{(3.31)}$	$53.64_{(1.68)}$	$77.82_{(8.07)}$	$90.97_{(3.71)}$	$94.72_{(0.45)}$
FEDNOVA	$44.70_{(4.26)}$	$72.29_{(5.60)}$	$89.73_{(0.59)}$	$9.00_{(2.19)}$	$25.89_{(3.30)}$	$52.94_{(1.63)}$	$76.93_{(7.76)}$	$90.27_{(4.14)}$	$94.86_{(0.20)}$
FEDDISCO	$43.05_{(4.18)}$	$72.62_{(5.14)}$	$89.40_{(0.47)}$	$8.62_{(2.19)}$	$28.14_{(3.30)}$	$54.21_{(1.61)}$	$79.64_{(7.45)}$	$91.65_{(3.57)}$	$94.75_{(0.17)}$
FEDMR	$40.08_{(3.79)}$	$76.72_{(1.32)}$	$88.59_{(0.53)}$	$2.88_{(0.52)}$	$4.21_{(0.71)}$	$32.51_{(1.71)}$	$\mathbf{87.99}_{(5.08)}$	$\mathbf{94.43}_{(0.89)}$	$\mathbf{95.89}_{(0.05)}$
FLOOD（本文）	$\mathbf{62.37}_{(4.30)}$	$\mathbf{86.16}_{(2.96)}$	$\mathbf{93.34}_{(0.09)}$	$\mathbf{15.31}_{(2.43)}$	$\mathbf{35.52}_{(2.39)}$	$\mathbf{61.05}_{(0.69)}$	$\underline{86.55}_{(6.48)}$	$\underline{93.67}_{(2.73)}$	$\underline{95.78}_{(0.04)}$



(a) 对比 服务器端 方法。

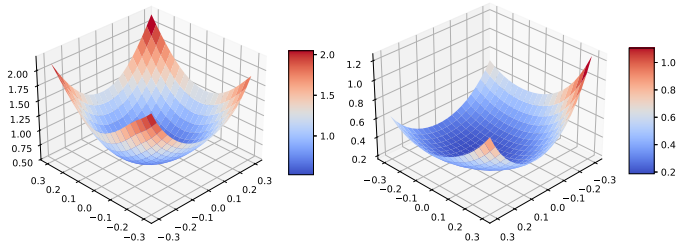


（b）与客户端方法的对比。

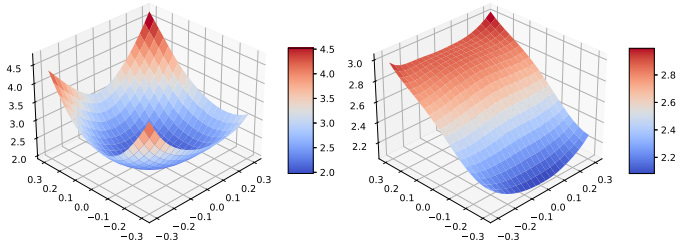
图5：在 Dir(0.1) 场景下，多种联邦学习方法在 CIFAR-10 数据集上的准确率曲线。上方将 FLOOD 与 FEDAVG 及服务器端全局调整方法进行对比，下方则将其与客户端侧局部训练校正方法进行对比。

模型与潜在数据分布的契合度逐渐提高，从而使 FLOOD 能够更准确地识别伪分布外样本，并相应地调整其权重。因此，其性能逐步追平并最终超越现有方法。在后期训练轮次中，FLOOD 展现出明显且持续的优势。不断
可靠的分布外置信度估计能够更有针对性地关注困难样本。总体趋势证实，

FLOOD 得益于其渐进式学习机制：一旦分布外信号趋于稳定，该方法便能有效利用这些信号，从而实现更优的收敛行为和最终性能。



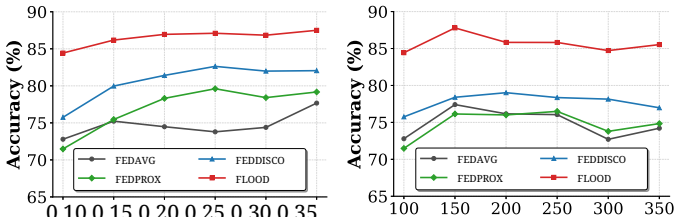
（a）FEDAVG-Dir(0.1)　（b）FLOOD-Dir(0.1)



（c）FEDAVG-Path(2)　（d）FLOOD-Path(2)

图6：FEDAVG 和 FLOOD 的损失景观。

图 6 展示了在不同数据异质性水平下，使用 FEDAVG 和 FLOOD 训练的模型的损失景观。在所有设置下，FEDAVG 都一致地收敛到尖锐极小值，并且沿多个方向呈现陡峭的曲率，这表明其相关 Hessian 谱包含相对较大的特征值。相比之下，FLOOD 产生了明显更平坦的极小值，其特征是更宽的谷底区域和显著降低的曲率。从 [43], 的角度来看，平坦极小值意味着模型参数的微小扰动只会导致损失略微增加，这被广泛认为是泛化能力提升的有力指标。因此，这些景观可视化结果提供了明确的几何证据，表明在非独立同分布（*i.i.d.*）数据分布下，FLOOD 能够获得比 FEDAVG 更稳定、更具鲁棒性且泛化能力更好的解。



(a) 参与比例。 (b) 客户端数量。

图 7：参与比例和客户端数量的影响。

方法	Path(2)	↑	Path(3)	↑	Path(5)	↑
FEDPROX	62.45	**17.77**	84.80	**12.25**	92.83	**3.03**
SCAFFOLD	58.20	**8.04**	81.52	**7.85**	92.98	**2.71**
FEDDECORR	69.31	**8.10**	87.57	**9.15**	92.99	**3.26**
FEDNP	59.35	**15.61**	86.09	**15.93**	92.19	**3.53**

表IV：在病态异质性条件下，针对 CIFAR-10 数据集，将 FLOOD 与客户端侧本地训练校正方法进行对比的正交实验。

C. 可扩展性分析（*RQ2*）

图 7 评估了 FLOOD 相对于两个关键系统级参数的可扩展性，即客户端参与比例和客户端总数。在图7a中，我们改变每轮通信中参与客户端的比例，同时固定其他所有设置。结果表明，在所有参与比例下，FLOOD 都 consistently 优于 FEDAVG、FEDPROX 和 FEDDISCO，展现出对部分客户端可用性的强大鲁棒性。此外，随着参与率提高，FLOOD 的性能仍保持稳定，表明它能够有效利用额外的客户端更新，同时不会引入聚合不稳定性或性能下降。图7b 进一步通过增加系统中的客户端总数来考察可扩展性。随着客户端数量增长，FLOOD 仍能持续实现稳健且稳定的准确率，而基线方法则表现出更大的性能波动。这一行为表明，FLOOD 中具备 OOD 感知能力的聚合机制能够有效地选择性增强可靠的客户端更新，从而在系统规模和客户端多样性增加的情况下维持稳定的学习动态。总体而言，这些结果表明，FLOOD 在参与率和客户端总数方面都具有良好的可扩展性，凸显了其对系统级变化的鲁棒性，以及其在大规模联邦学习部署中的实用性。

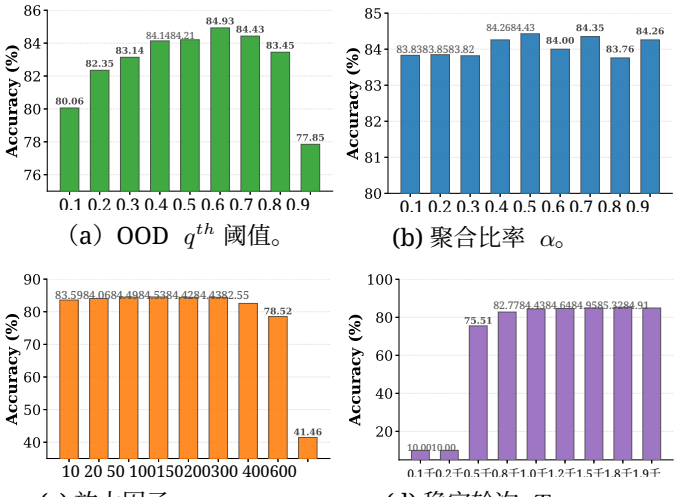
D. 与客户端侧联邦学习方法的兼容性（*RQ3*）

通过直接修改局部监督损失，FLOOD 可以无缝集成为正交插件模块，以增强现有的客户端侧局部训练校正方法。表 IV 展示了在不同病理异质性设置下进行正交兼容性实验的结果。对于每种基线方法，该表报告了集成 FLOOD 后的绝对分类准确率，以及在各异质性水平下相较于原始基线方法所取得的性能提升。结果清楚地表明，FLOOD 能够持续提升所有受评估客户端方法的性能，证实了其良好的兼容性和互补性。值得注意的是，当与 F 结合时

ED-PROX、FLOOD 在 Path(2) 中将准确率大幅提升了最多 17.77%，凸显了分布外感知样本重新加权在校正有偏本地更新方面的有效性。此外，FLOOD 与 FEDDECORR 集成时表现尤为出色，表明它能够进一步提升先进本地训练校正策略在各种异质性设置下的效果。

E. 消融研究（*RQ4*）

a) 超参数的影响：我们分析了 FLOOD 对四个关键超参数（即 q、a、T 和 α）的敏感性。图 8a 展示了本地分布外阈值 q 对 FLOOD 性能的影响。如图所示，当 q 处于 $[0.2, 0.8]$范围内时，FLOOD 的性能稳定且出色，表明其对适度的阈值变化具有鲁棒性。当 q 过小时（例如 $q = 0.1$），性能会急剧下降，因为过多样本被划分为伪分布外样本，从而过度抑制真实分布内样本，妨碍有效的监督学习。相反，阈值过大也会导致性能略有下降，因为对伪分布外样本的关注不足。



（a）OOD q^{th} 阈值。 (b) 聚合比率 α。

(c) 放大因子 a。 (d) 稳定轮次 T。

图8：四个关键超参数的影响。

为考察服务器端聚合的影响，图8b对聚合比率 α进行了网格搜索。结果表明，当 α =0.5时，FLOOD取得最优性能，而偏离该值会导致轻微的性能下降。当 αis 过小时，分布外置信度的影响有限，聚合在很大程度上由数据量主导。相反，α 过大可能会过度强调一部分客户端，从而导致有偏差且次优的聚合。

图8c探讨了放大因子 a 对 FLOOD 行为的影响。结果表明，随着 a从较小值增大，性能有所提升，这说明放大伪分布外样本是有益的。然而，一旦 a超过某个阈值，准确率便开始下降。这表明，过度激进的放大可能会主导

学习动态，导致模型过度关注伪分布外样本，并在监督训练中失去平衡。

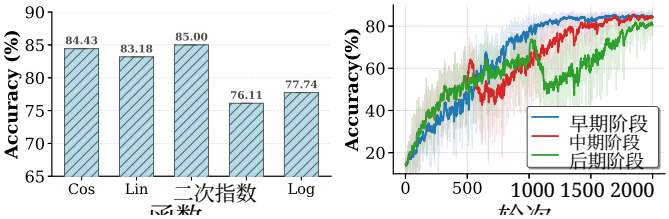
我们在图8d中研究了权重调度的时间特性。当 T 较小时，会过早引入强重新加权，从而导致不稳定训练和较差的性能。增大 T 会逐步提升性能，反映出将基于分布外数据的重新加权推迟到模型足够可靠之后的重要性。超过某一阈值后，进一步增大 T 带来的收益有限，表明延迟放大的收益最终会趋于饱和。

综合来看，这些结果表明，FLOOD的有效性取决于对其核心超参数的协同调优，这些超参数共同平衡异质数据分布下的训练稳定性与鲁棒性。

表V：两个组件在两种模型上的影响。

	FEDAVG	+DAC	+ASW	FLOOD
ResNet	$72.79_{(5.95)}$	$74.46_{(4.62)}$	$82.71_{(2.10)}$	**84.43**$_{(1.66)}$
MobileNet	$55.01_{(7.36)}$	$56.99_{(5.95)}$	$63.40_{(6.80)}$	**66.57**$_{(3.92)}$

b) 各组件的影响：表 V 展示了 FLOOD 的消融实验结果，通过比较其解耦组件在两种模型架构上的表现，其中 ASW 表示自适应样本加权，DAC 表示动态聚合校正。结果表明，这两个组件均有助于提升性能，优于FEDAVG基线。具体而言，仅使用 ASW 就能带来显著提升，说明在本地训练过程中着重处理伪分布外样本，能够有效缓解有偏本地更新。同样，加入 DAC 后，通过基于客户端可靠性优化服务器端聚合，也能提升性能。更重要的是，ASW 与 DAC 结合后，在 ResNet 和 MobileNet 两种模型架构上始终取得最佳性能，显著优于单独使用任一组件。



(a) 调度函数。 （b）递增阶段。

图9：调度策略对 FLOOD 的影响。

c) 权重调度策略的影响：图9a 比较了不同的调度函数，结果表明，余弦和二次等平滑、渐进的方案始终优于指数调度等更激进的方案，这说明渐进式权重放大对于稳定优化至关重要。图 9b进一步分析了不同的增长阶段设计，结果显示，在适当阶段启动权重增长有助于实现更好的收敛；启动过早会受到不可靠的分布外估计影响，而启动过晚则会限制强调困难样本所带来的益处。总体而言，这些结果表明，函数形式和权重增长的时机对于在 FLOOD 中有效利用分布外信号都起着关键作用。

表VI：不同 OOD方法 对 FLOOD 的影响。

MSP	MAXLOGIT	ENERGY	GEN	ASH-S	ASH-B	ASH-P
Dir(0.1) 84.45	84.81	84.43	85.86	10.00	74.79	84.50
Dir(0.5) 93.44	93.40	92.92	93.11	93.13	92.63	93.02
Dir(1.0) 93.37	93.39	93.28	93.42	93.48	93.54	93.51
Path(2) 61。55	56.50	62.37	55.43	10.00	56.97	61.05
Path(3) 86。35	85.96	86.16	84.42	85.01	84.29	85.12
Path(5) 92。94	93.00	93.34	93.01	93.15	92.56	93.07

*d) OOD*方法的影响：我们进一步研究了不同分布外检测方法对 FLOOD 性能的影响，结果汇总于表 VI。具体而言，我们评估了一组具有代表性的 OOD评分函数，包括 MSP [22], MAXLOGIT [44], EN- ERGY [24], GEN [45], 以及 ASH变体（ASH-S、ASH-B 和 ASH-P） [35]。总体而言，在两种异质性设置下，大多数分布外检测方法相较于FEDAVG基线均带来了明显的性能提升，这表明在实际场景中，FLOOD 基本不受具体 OOD评分函数选择的影响。值得注意的是，某些 OOD方法（如 ASH-S）在严重异质性条件下（例如 Dir(0。1) 和 Path(2)）表现不佳，导致性能崩溃。这表明，过于激进或不稳定的 OOD评分可能会在重新加权过程中引入噪声。相比之下，ENERGY、MSP 和 MAXLOGIT 等稳定的评分函数在所有设置下都能取得持续强劲且可靠的结果。

VI. 结论

联邦学习中数据异质性这一关键挑战源于客户端数据固有的非*i.i.d.*特性，仍是开发稳健可靠的全局模型的一大障碍。本文提出了 FLOOD，这是一种新颖的联邦学习框架，通过利用分布外检测动态缓解数据异质性带来的不利影响，从而应对这一挑战。FLOOD采用双重加权策略，自适应地调整客户端的样本权重和服务器端的聚合权重，从而有效抵消非*i.i.d.* 数据所造成的不一致性。FLOOD在本地训练期间着重处理伪分布外样本，并提高更可靠的本地模型的贡献权重，从而增强全局模型的泛化能力，同时保护客户端隐私。在多个基准数据集上开展的大量实验表明，FLOOD始终优于最先进的联邦学习方法，在异构数据分布下取得了显著提升。重要的是，FLOOD与现有联邦学习框架互补且相互正交，因此可广泛应用于各种联邦学习场景。

参考文献

[1] Z. Ling, Z. Yue, J. Xia, T. Wang, M. Chen, and X. Lian, 《Fedentropy: 基于最大熵判定的客户端选择，实现非独立同分布场景下的高效联邦学习》，载于国际并行与分布式处理及应用研讨会，2023年，第56–63页。

[2]T. Liu, J. Xia, Z. Ling, X. Fu, S. Yu, and M. Chen, 《面向人工智能物联网应用的高效联邦学习：基于知识蒸馏》，IEEE物联网期刊，第10卷，第8期，4月15日，第7229–7243页，2023年。[3] L. Yu, X. Li 等,《基于联邦学习数据敏感度的差分隐私参数动态优化方法》，高级计算系统期刊，第5卷，第6期，第1–13页，2025年。[4] Y. Fu, X. Tang, C. Li, F. R. Yu, and N. Cheng, 《面向自动驾驶的安全个性化联邦学习算法》,IEEE智能交通系统汇刊，第25卷，第12期，第20 378–20 389页，2024年。[5] J. Ma, C. Tian, L. Li, and C. Xu, 《Fedmg：面向自动驾驶控制的联邦多全局优化框架》，载于国际服务质量研讨会，2024年，第1–10页。[6] A. Karampasi, P. I. Radoglou-Grammatikis, M. Pawlicki, R. Choras, R. M. de Pozuelo, P. G. Sarigiannidis, D. Puchalski, A. Pawlicka, R. Kozik, and M. Choras, 《迈向金融系统中的透明人工智能网络安全：CaixaBank试点项目中的联邦学习与可解释人工智能部署》,载于IEEE国际数据挖掘会议研讨会，2024年，第270–277页。[7] S. J ain and K. R. Jerripothula, 《商业图像来源的联邦学习》，载于IEEE/CVF冬季计算机视觉应用会议，2023年，第6523–6532页。[8] J. Liu, Y. Li, M. Zhao, L. Liu, and N. Kumar, 《EPFFL：面向分布式电子医疗数据共享服务的联邦学习隐私与公平性增强》，IEEE可靠与安全计算汇刊，第22卷，第2期，第1239–1252页，2025年。[9] C. Huang, Y. He, X. Han, W. Jiao, Z. Jin, and L. Wang, 《Unitrans：用于增强边缘医疗协作的统一纵向联邦知识迁移框架》，IEEE移动计算汇刊，第24卷，第12期，第13 408–13 425页，2025年。[10] H. Zhou, H. Dai, G. Yang, and Y. Xiang, 《面向边缘计算中隐私保护与效率的鲁棒联邦学习》，IEEE服务计算汇刊，第18卷，第3期，第1739–1752页，2025年。[11] M. P. Uddin, Y. Xiang, Y. Zhao, M. Ali, Y. Zhang, and L. Gao, 《联邦学习中的自适应正则化与弹性估计》，IEEE服务计算汇刊，第17卷，第4期，第1369–1381页，2024年。[12] Y. Xu, Y. Zhu, Z. Wang, H. Xu, and Y. Liao, 《通过针对非独立同分布数据进行逐层聚合来增强联邦学习》，IEEE服务计算汇刊，第18卷，第2期，第798–811页，2025年。[13] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, 《从去中心化数据中高效通信地学习深度网络》，载于国际人工智能与统计会议，2017年，第1273–1282页。[14] Y. Shi, J. Liang, W. Zhang, V. Y. F. Tan, and S. Bai, 《迈向对异构联邦学习中维度坍塌的理解与缓解》，载于国际学习表征会议，2023年。[15] S. Deng, H. Zhao, Z. Wang, G. Cheng, P. Chen, W. Qian, Z. Ling, J. Yin, A. Y. Zomaya, and S. Dustdar, 《智能体服务计算》，arXiv预印本arXiv:2509.24380，2025年。[16] T. H. Hsu, H. Qi, and M. Brown, 《衡量非相同数据分布对联邦视觉分类的影响》，CoRR，第abs/1909.06335卷，2019年。[17] R. Ye, M. Xu, J. Wang, C. Xu, S. Chen, and Y. Wang, 《FEDDISCO：具有差异感知协作的联邦学习》，载于国际机器学习会议，第202卷，2023年，第39 879–39 902页。[18] L. Zhang, L. Shen, L. Ding, D. Tao, and L. Duan, 《通过无数据知识蒸馏微调全局模型，以实现非独立同分布联邦学习》，载于计算机视觉与模式识别，2022年，第10 164–10 173页。[19] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, 《异构网络中的联邦优化》，载于机器学习与系统会议，2020年，第429–450页。[20] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, 《SCAFFOLD：联邦学习的随机控制平均》，载于国际机器学习会议，2020年，第5132–5143页。[21] Q. Li, B. He, and D. Song, 《模型对比式联邦学习》，载于计算机视觉与模式识别，2021年，第10 713–10 722页。[22] D. Hendrycks and K. Gimpel, 《一种用于检测神经网络中误分类样本和分布外样本的基线方法》，载于国际学习表征会议，2017年。

[23] S. Liang、Y. Li 和 R. Srikant，“增强神经网络中分布外图像检测的可靠性”，载于国际学习表征会议，2018年。[24] W. Liu、X. Wang、J. D. Owens 和 Y. Li，“基于能量的分布外检测”，载于神经信息处理系统会议，2020年。[25] M. Hu、P. Zhou、Z. Yue、Z. Ling、Y. Huang、A. Li、Y. Liu、X. Lian 和 M. Chen，“Fedcross：通过多模型交叉聚合实现精准联邦学习”，载于IEEE国际数据工程会议，2024年，第2137–2150页。[26] S. Liang、Q. Cui、X. Huang、B. Zhao、Y. Hou 和 X. Tao，“面向异构移动边缘的高效分层联邦服务”，IEEE服务计算汇刊，第18卷，第1期，第140–155页，2025年。[27] L. Luo、C. Zhang、H. Yu、G. Sun、S. Luo 和 S. Dustdar，“针对异构客户端-边缘-云网络的通信高效联邦学习与自适应聚合”，IEEE服务计算汇刊，第17卷，第6期，第3241–3255页，2024年。[28] L. Gao、H. Fu、L. Li、Y. Chen、M. Xu 和 C. Xu，“通过局部漂移解耦与校正实现非独立同分布数据上的联邦学习”，载于计算机视觉与模式识别，2022年，第10 102–10 111页。[29] Y. Guo、X. Tang 和 T. Lin，“通过降低局部学习偏差，改善异构数据上的联邦学习”，载于国际机器学习会议，第202卷，2023年，第12 034–12 054页。[30] J. Wang、Q. Liu、H. Liang、G. Joshi 和 H. V. Poor，“解决异构联邦优化中的目标不一致问题”，载于神经信息处理系统会议，2020年。[31] M. Hu、Z. Yue、X. Xie、C. Chen、Y. Huang、X. Wei、X. Lian、Y. Liu 和 M. Chen，“聚合是唯一选择吗？通过逐层模型重组实现联邦学习”，载于ACM SIGKDD知识发现与数据挖掘会议，2024年，第1096–1107页。[32] H. Yuan、J. Bi、Z. Wang、J. Zhang、M. Zhou 和 R. Buyya，“边缘计算中的多视角节能深度学习”，IEEE物联网期刊，2025年。[33] Z. Ling、Y. Chang、H. Zhao、X. Zhao、K. Chow 和 S. Deng，“Cadref：通过类别感知解耦相对特征利用实现稳健的分布外检测”，载于IEEE/CVF计算机视觉与模式识别会议，2025年，第4968–4977页。[34] Y. Sun、C. Guo 和 Y. Li，“ReAct：利用整流激活进行分布外检测”，载于神经信息处理系统会议，2021年，第144–157页。[35] A. Djurisic、N. Bozanic、A. Ashok 和 R. Liu，“用于分布外检测的极简激活整形”，载于国际学习表征会议，2023年。[36] A. Paszke、S. Gross、F. Massa、A. Lerer、J. Bradbury、G. Chanan、T. Killeen、Z. Lin、N. Gimelshein、L. Antiga、A. Desmaison、A. K¨opf、E. Z. Yang、Z. DeVito、M. Raison、A. Tejani、S. Chilamkurthy、B. Steiner、L. Fang、J. Bai 和 S. Chintala，“PyTorch：一种命令式风格的高性能深度学习库”，载于神经信息处理系统进展论文集，2019年，第8024–8035页。[37] X. Wu、H. Huang、Y. Ding、H. Wang、Y. Wang 和 Q. Xu，“Fednp：通过联邦神经传播实现非独立同分布场景下的联邦学习”，载于AAAI人工智能会议，2023年，第10 399–10 407页。[38] A. Krizhevsky、G. Hinton 等，“从微小图像中学习多层特征”，2009年。[39] Y. Netzer、T. Wang、A. Coates、A. Bissacco、B. Wu、A. Y. Ng 等，“利用无监督特征学习读取自然图像中的数字”，载于NIPS深度学习与无监督特征学习研讨会，第2011卷，第2期，2011年，第4页。[40] H. Wang、Z. Kaplan、D. Niu 和 B. Li，“利用强化学习优化非独立同分布数据上的联邦学习”，载于IEEE计算机通信会议，2020年，第1698–1707页。[41] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，载于计算机视觉与模式识别，2016年，第770–778页。[42] M. Sandler、A. G. Howard、M. Zhu、A. Zhmoginov 和 L. Chen，“Mobilenetv2：倒置残差与线性瓶颈”，载于计算机视觉与模式识别，2018年，第4510–4520页。[43] M. Hardt、B. Recht 和 Y. Singer，“训练更快，泛化更好：随机梯度下降的稳定性”，载于国际机器学习会议，第48卷，2016年，第1225–1234页。[44] D. Hendrycks、S. Basart、M. Mazeika、A. Zou、J. Kwon、M. Mostajabi、J. Steinhardt 和 D. Song，“面向现实场景的可扩展分布外检测”，载于国际机器学习会议，第162卷，2022年，第8759–8773页。[45] X. Liu、Y. Lochman 和 C. Zach，“GEN：突破基于softmax的分布外检测极限”，载于计算机视觉与模式识别

，2023，第 23 页，946–23955。

附录

A. 定理1的证明

为简单起见，我们分析每个被选中的客户端执行单个局部梯度步（即 $E = 1$），且迷你批次大小为 B 的情况。该分析可借助 [20] 中的技术扩展至 $E > 1$。

令 θ^t 表示第 t 轮开始时的全局模型。每个参与客户端 $c \in \mathcal{C}^t$ 根据其重新加权后的损失计算随机梯度：

$$g_c^t := \frac{1}{B} \sum_{(x,y)\in\xi_c^t} w_{c,(x,y)}^t \nabla \Phi(\theta^t; x, y),$$

其中，ξ_c^t 是从 $\mathcal{D}_c$ 中抽取的大小为 B 的随机小批量，并且根据假设 A3，$w_{c,(x,y)}^t \in [W_{\min}, W_{\max}]$。随后，局部更新为 $\theta_c^t = \theta^t - \eta g_c^t$，服务器进行如下聚合：

$$\theta^{t+1} = \sum_{c\in\mathcal{C}^t} p_c^t \theta_c^t = \theta^t - \eta \sum_{c\in\mathcal{C}^t} p_c^t g_c^t.$$

根据 F 的 L-光滑性（由假设 A1 得出），我们有标准的下降不等式：

$$F(\theta^{t+1}) \le F(\theta^t) + \langle \nabla F(\theta^t), \theta^{t+1} - \theta^t \rangle + \frac{L}{2}\|\theta^{t+1} - \theta^t\|^2. \tag{16}$$

代入更新规则可得

$$F(\theta^{t+1}) \le F(\theta^t) - \eta \left\langle \nabla F(\theta^t), \sum_c p_c^t g_c^t \right\rangle + \frac{L\eta^2}{2} \left\| \sum_c p_c^t g_c^t \right\|^2. \tag{17}$$

对小批次采样和客户端选择中的随机性取期望（以 θ^t 为条件），并注意到 $\mathbb{E}[g_c^t \mid \theta^t] = \nabla \mathcal{L}_c^t(\theta^t)$，可得

$$\mathbb{E}[F(\theta^{t+1}) \mid \theta^t] \le F(\theta^t) - \eta \left\langle \nabla F(\theta^t), \sum_c p_c^t \nabla \widetilde{\mathcal{L}}_c^t(\theta^t) \right\rangle + \frac{L\eta^2}{2} \mathbb{E}\left[\left\| \sum_c p_c^t g_c^t \right\|^2 \middle| \theta^t \right]. \tag{18}$$

现在我们对内积项进行分解。根据偏差 $b_c^t = \nabla \mathcal{L}_c^t(\theta^t) - \nabla \mathcal{L}_c(\theta^t)$和理想全局梯度$\nabla F(\theta^t) = \sum_c q_c \nabla \mathcal{L}_c(\theta^t)$的定义，其中 $q_c = n_c/n$，可得

$$\sum_c p_c^t \nabla \widetilde{\mathcal{L}}_c^t(\theta^t) = \sum_c p_c^t \left(\nabla \mathcal{L}_c(\theta^t) + b_c^t \right) = \underbrace{\sum_c p_c^t \nabla \mathcal{L}_c(\theta^t)}_{=:h^t} + \sum_c p_c^t b_c^t. \tag{19}$$

根据假设 (A5)，$\|h^t - \nabla F(\theta^t)\| \le \epsilon$；根据假设 A4，$\|\sum_c p_c^t b_c^t\| \le \sum_c p_c^t \|b_c^t\| \le \delta$。因此，

$$\begin{aligned} &\left\langle \nabla F(\theta^t), \sum_c p_c^t \nabla \widetilde{\mathcal{L}}_c^t(\theta^t) \right\rangle \\ &= \langle \nabla F(\theta^t), h^t \rangle + \left\langle \nabla F(\theta^t), \sum_c p_c^t b_c^t \right\rangle \\ &\ge \|\nabla F(\theta^t)\|^2 - \|\nabla F(\theta^t)\| \cdot \|h^t - \nabla F(\theta^t)\| \\ &\quad - \|\nabla F(\theta^t)\| \cdot \left\| \sum_c p_c^t b_c^t \right\| \\ &\ge \|\nabla F(\theta^t)\|^2 - (\epsilon + \delta)\|\nabla F(\theta^t)\|. \end{aligned} \tag{20}$$

接下来，我们界定二阶矩项。利用 Jensen 不等式以及客户端更新的独立性，

$$\mathbb{E}\left[\left\| \sum_c p_c^t g_c^t \right\|^2 \middle| \theta^t \right] \le \sum_c p_c^t \, \mathbb{E}\left[\|g_c^t\|^2 \,\middle|\, \theta^t \right]. \tag{21}$$

现在将 $g_c^t = \nabla \mathcal{L}_c^t(\theta^t) + \xi_c^t$ 分解，其中 $\xi_c^t = g_c^t - \mathbb{E}[g_c^t]$ 满足 $\mathbb{E}[\xi_c^t] = 0$，且根据假设 A2 和 A3，$\mathbb{E}[\|\xi_c^t\|^2] \le \frac{W_{\max}^2}{B}\sigma^2$。于是

$$\begin{aligned} \mathbb{E}[\|g_c^t\|^2 \mid \theta^t] &= \|\nabla \widetilde{\mathcal{L}}_c^t(\theta^t)\|^2 + \mathbb{E}[\|\xi_c^t\|^2] \\ &\le \left(\|\nabla \mathcal{L}_c(\theta^t)\| + \|b_c^t\| \right)^2 + \frac{W_{\max}^2 \sigma^2}{B} \\ &\le (G + \delta)^2 + \frac{W_{\max}^2 \sigma^2}{B}, \end{aligned} \tag{22}$$

这里我们对所有 c, θ 使用了标准的有界梯度假设$\|\nabla \mathcal{L}_c(\theta)\| \le G$。

为简单起见，我们将 δ 并入常数，并采用较宽松但更简洁的界 $\|\widetilde{\nabla}(\theta^t)\| \le W_{\max} G$，这是成立的，因为样本权重有界于 $W_{\max}$，且梯度经过了平均。因此，

$$\mathbb{E}\left[\left\| \sum_c p_c^t g_c^t \right\|^2 \middle| \theta^t \right] \le W_{\max}^2 G^2 + \frac{W_{\max}^2 \sigma^2}{B}. \tag{23}$$

综上所述，我们得到

$$\mathbb{E}[F(\theta^{t+1}) \mid \theta^t] \le F(\theta^t) - \eta \|\nabla F(\theta^t)\|^2 + \eta(\epsilon + \delta)\|\nabla F(\theta^t)\| + \frac{L\eta^2}{2}\left(W_{\max}^2 G^2 + \frac{W_{\max}^2 \sigma^2}{B} \right). \tag{24}$$

应用杨氏不等式，即 $(\epsilon + \delta)\|\nabla F(\theta^t)\| \le \frac{1}{2}\|\nabla F(\theta^t)\|^2 + \frac{1}{2}(\epsilon + \delta)^2$，可得

$$\mathbb{E}[F(\theta^{t+1}) \mid \theta^t] \le F(\theta^t) - \frac{\eta}{2}\|\nabla F(\theta^t)\|^2 + \frac{\eta}{2}(\epsilon + \delta)^2 + \frac{L\eta^2}{2}\left(W_{\max}^2 G^2 + \frac{W_{\max}^2 \sigma^2}{B} \right). \tag{25}$$

对截至轮次 t 的所有随机性取全期望，并从 $t=0$ 到 $T-1$进行望远镜求和，可得

$$\begin{aligned}\frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(\theta^t)\|^2\right] &\le F(\theta^0)-\mathbb{E}[F(\theta^T)]+\frac{\eta T}{2}(\epsilon+\delta)^2 \\ &\quad+\frac{L\eta^2 T}{2}\left(W_{\max}^2G^2+\frac{W_{\max}^2\sigma^2}{B}\right) \\ &\le F(\theta^0)-F^*+\frac{\eta T}{2}(\epsilon+\delta)^2 \\ &\quad+\frac{L\eta^2 T}{2}\left(W_{\max}^2G^2+\frac{W_{\max}^2\sigma^2}{B}\right),\end{aligned} \tag{26}$$

其中 $F^*=\inf_\theta F(\theta)$。

将两边同时除以 $\frac{\eta T}{2}$，得到

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(\theta^t)\|^2\right] &\le \frac{2(F(\theta^0)-F^*)}{\eta T}+(\epsilon+\delta)^2 \\ &\quad+L\eta\left(W_{\max}^2G^2+\frac{W_{\max}^2\sigma^2}{B}\right).\end{aligned} \tag{27}$$

最后，令学习率为 $\eta=\frac{1}{\sqrt{T}}$，可得

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\theta^t)\right\|^2\right] &\le \frac{2(F(\theta^0)-F^*)}{\sqrt{T}}+(\epsilon+\delta)^2 \\ &\quad+\frac{L}{\sqrt{T}}\left(W_{\max}^2G^2+\frac{W_{\max}^2\sigma^2}{B}\right),\end{aligned} \tag{28}$$

由此得到结果

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla F(\theta^t)\right\|^2\right]\le\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)+\mathcal{O}\left((\epsilon+\delta)^2\right). \tag{29}$$