面向噪声鲁棒训练的混合联邦学习

Yongjun Kim，IEEE研究生会员；Hyeongjun Park，IEEE研究生会员；Hwanjin Kim，IEEE会员；Junil Choi，IEEE高级会员

arXiv:2601.04483v1 [cs.LG] 8 Jan 2026

摘要——联邦学习（FL）和联邦蒸馏（FD）是增强隐私保护、用于训练用户设备模型的分布式学习范式，在噪声鲁棒性和学习速度之间各有不同的权衡。为弥补各自的不足，我们提出一种混合联邦学习（HFL）框架，其中每个用户设备（UE）传输梯度或逻辑值，基站（BS）则选择联邦学习更新和联邦蒸馏更新在每轮中的权重。我们推导了混合联邦学习框架的收敛性，并提出两种利用混合联邦学习自由度（DoF）的方法：（i）通过Jenks优化进行自适应用户设备聚类；（ii）通过阻尼牛顿法进行自适应权重选择。数值结果表明，同时利用这两种自由度时，混合联邦学习在低信噪比（SNR）下可取得更高的测试准确率。

索引词——联邦学习（FL）、联邦蒸馏（FD）、混合联邦学习（HFL）、噪声鲁棒训练

I. 引言

下一代无线网络预计需要以低延迟处理来自海量设备的数据。然而，由于隐私法规严格且回传容量有限，集中式学习往往难以实施 [1]。联邦学习（FL） [2] 通过将训练数据保留在本地、仅交换模型参数或梯度来应对这些挑战。这种去中心化范式带来三项优势：（i）避免原始数据暴露，从而增强隐私保护；（ii）通过轻量级传输降低通信开销；（iii）降低自动驾驶等对时延要求严格的5G/6G应用的延迟 [3]。

尽管具备这些优势，联邦学习在无线链路上仍面临挑战。高维梯度会产生大量上行链路和下行链路流量，并给低功耗边缘设备带来计算和能源负担。为缓解这一问题，联邦蒸馏（FD） [4], [5] 交换逻辑值而非模型参数，从而缩短信号、降低通信和能源成本、在监管要求更严格的环境中提供更强的隐私保护，并支持无需在每个客户端提供标签的分布式学习。

尽管联邦蒸馏已成为一种颇具前景的智能边缘服务方案，但它以牺牲峰值准确率为代价，换取较低的通信开销和对上行链路噪声更强的鲁棒性 [6], [7]。在无噪声的上行链路条件下，仅交换低维逻辑值通常会比联邦学习取得更低的准确率 [8], ，因为当上行链路噪声污染传输时联邦学习会遭受显著的性能下降，因为高维梯度会被直接扭曲。联邦蒸馏通过传输对噪声不那么敏感的低维逻辑值，对受功率限制的边缘设备而言更具鲁棒性 [9]。

为解决联邦学习和联邦蒸馏各自的局限性，我们提出一种混合联邦学习框架，其中每个用户设备在每轮通信中自适应地传输梯度或逻辑值；在信道条件良好时保留联邦学习的准确性，在条件不利时则利用联邦蒸馏的噪声鲁棒性。本文的主要贡献概述如下：

1）我们设计了一种新颖的混合联邦学习框架，联合利用梯度和逻辑值传输，并对其进行收敛性分析。2）我们提出了通过自适应用户设备聚类和权重选择来优化传输自由度的方法。

本文其余部分安排如下。第 II 节介绍系统模型，包括信号传输与接收。第 III 节详述 HFL 方案、其收敛性分析，以及自适应用户设备聚类和权重选择方法。第 IV 节给出仿真结果，验证 HFL 在低信噪比下的抗噪测试准确率。最后，第 V 节总结全文。

II. 系统模型

我们考虑一个协作学习系统，其中配备 N 根天线的基站 (BS) 为一组 K 个单天线用户设备 $\mathcal{K}$ 提供服务，即 $|\mathcal{K}|= K$。每个用户设备 $k \in \mathcal{K}$ 持有一个本地训练数据集 $\mathcal{D}_k$。在通信轮次 t，基站向用户设备

k $\mathcal{K}_1$ =发送指示 $\{k_1 \quad k_1 \in \mathcal{K}\ \mathcal{I}_{k_1}^{(t)} = 0\}$ $|\mathcal{K}_1|= K_1$ 符 $\mathcal{I}_k^{(t)} \in \{0,1\}$，以指定其上行传输数据类型。定义联邦学习集合: , 其中 和联邦蒸馏集合 $\mathcal{K}_2 = \{k_2 \colon k_2 \in \mathcal{K}, \mathcal{I}_{k_2}^{(t)} = 1\}$，其中 $|\mathcal{K}_2|= K - K_1$。集合 $\mathcal{K}_1$ 中的用户设备 k_1 传输梯度 $\mathbf{g}_{k_1}^{(t)} \in \mathbb{R}^P$，其中 P 为向量化梯度的长度。集合 $\mathcal{K}_2$ 中的用户设备 k_2 在包含 P_{pub} 个输入的公共数据集 $\mathcal{D}_{\text{pub}}$ 上计算逻辑值，为分类任务中的C 个类别生成 $\mathbf{z}_{k_2}^{(t)} \in \mathbb{R}^{CP_{\text{pub}}}$。

对于无线传输，两种上行传输数据类型均通过以下处理映射为复信号。令 $\mathbf{u}_k^{(t)}$ 表示 $\mathbf{g}_{k_1}^{(t)}$ 或 $\mathbf{z}_{k_2}^{(t)}$。首先，将实数元素两两配对，构成复向量 $\tilde{\mathbf{u}}_k^{(t)}[m] = \mathbf{u}_k^{(t)}[2m-1] + j\,\mathbf{u}_k^{(t)}[2m]$，其中 $1 \le m \le \lceil |\mathbf{u}_k^{(t)}|/2 \rceil$，$\lceil \cdot \rceil$ 表示向上取整函数，$[m]$ 表示第 m 个元素。为确保缩放一致，对其进行标准化处理：

Y. Kim、H. Park 和 J. Choi 任职于韩国大田 34141 的 KAIST 电气工程学院（电子邮箱：yongjunkim、mika0303、junil@kaist.ac.kr）。H. Kim 任职于韩国大邱 41566 的庆北国立大学电子工程学院（电子邮箱：hwanjin@knu.ac.kr）

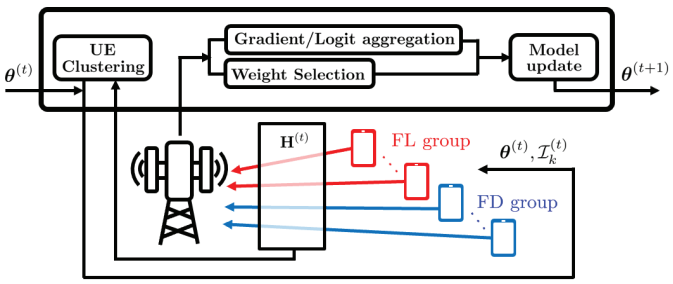


图 1：第 t 轮的分层联邦学习框架。

$\tilde{\mathbf{u}}_k[m] = (\tilde{\mathbf{u}}_k[m] - \mu_{\tilde{\mathbf{u}}_k})/\sigma_{\tilde{\mathbf{u}}_k}$，其中 $\mu_{\tilde{\mathbf{u}}_k}$ 和 $\sigma_{\tilde{\mathbf{u}}_k}$ 分别是 $\tilde{\mathbf{u}}_k$ 的均值和标准差，随后进行 $\tilde{\mathbf{u}}_k[m]$ 归一化 $\tilde{\mathbf{u}}_k[m] = \tilde{\mathbf{u}}_k[m]/\|\tilde{\mathbf{u}}_k\|_\infty$。我们将 $L = \max_k |\tilde{\mathbf{u}}_k|$ 设为每轮的时隙数。如果 $\tilde{\mathbf{u}}_k$ 的长度小于 L，则进行零填充以统一信号长度，从而得到用户设备 k 的最终发射信号 $\mathbf{x}_k^{(t)} \in \mathbb{C}^{1\times L}$。将所有用户设备的信号堆叠后得到 $\mathbf{X}^{(t)} \in \mathbb{C}^{K\times L}$，并通过无线信道传输。为聚焦于 $\mathbf{X}^{(t)}$ 中的噪声影响，我们假设 $\mu_{\tilde{\mathbf{u}}_k}$, $\sigma_{\tilde{\mathbf{u}}_k}$, $\|\tilde{\mathbf{u}}_k\|_\infty$ 的传输以及所有下行信号均无差错。

传输 $\mathbf{X}^{(t)} = [\mathbf{x}_1^{(t)}\ \mathbf{x}_2^{(t)}\ \cdots\ \mathbf{x}_K^{(t)}] \in \mathbb{C}^{N\times K}$ 时，考虑一个瑞利衰落信道 $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_N]$。我们假设每个通信轮次内信道保持恒定。每个通信轮次 t 包含 L 个同步时隙，索引为 $m = 1, 2, \cdots, L$，在此期间传输 $\mathbf{X}^{(t)}$ 的每一列。第 m 个时隙中基站接收到的信号为

$$\mathbf{y}^{(t)}[m] = \sqrt{\rho}\,\mathbf{H}^{(t)}\mathbf{x}^{(t)}[m] + \mathbf{n}^{(t)}[m], \tag{1}$$

其中，$\mathbf{x}^{(t)}[m] \in \mathbb{C}^{K\times 1}$ 是 $\mathbf{X}^{(t)}$ 的第 m 个列向量，$\mathbf{n}^{(t)} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$ 是加性白复高斯噪声，$\mathbf{I}_N$ 是 $N\times N$ 单位矩阵，ρ 是信噪比。对于 $N \geq K$，基站对 $\mathbf{y}^{(t)}[m]$ 应用迫零（ZF）解码，检测到的信号 $\hat{\mathbf{x}}^{(t)}[m] \in \mathbb{C}^{K\times 1}$ 表示为

$$\frac{(\mathbf{H}^{(t)\mathrm{H}}\mathbf{H}^{(t)})^{-1}\mathbf{H}^{(t)\mathrm{H}}}{\sqrt{\rho}}\,\mathbf{y}^{(t)}[m] = \mathbf{x}^{(t)}[m] + \tilde{\mathbf{n}}^{(t)}[m], \tag{2}$$

其中 $\tilde{\mathbf{n}}^{(t)}[m] \sim \mathcal{CN}\left(\mathbf{0}, (\rho\,\mathbf{H}^{(t)\mathrm{H}}\mathbf{H}^{(t)})^{-1}\right)$。对所有 $m \in [1, L]$ 检测后，$\hat{\mathbf{x}}_k^{(t)} \in \mathbb{C}^{1\times L}$ 被映射回 $\hat{\mathbf{g}}_{k_1}^{(t)}$ 或 $\hat{\mathbf{z}}_{k_2}^{(t)}$，分别对应每个 $k_1 \in \mathcal{K}_1$ 和 $k_2 \in \mathcal{K}_2$，即恢复出的梯度和逻辑值。

III. 混合联邦学习

我们提出了一种如图 1 所示的混合联邦学习框架，使每个用户设备能够根据信道状况传输其梯度或逻辑值。我们首先介绍单轮操作并推导其收敛性，表明所得界推广了联邦学习和联邦蒸馏的界。随后，我们介绍动态用户设备聚类和权重选择方法，以提供额外的自由度。

A. 混合联邦学习框架

在通信轮次 t 开始时，全局模型参数 $\boldsymbol{\theta}^{(t)}$ 会广播给所有用户设备，$\boldsymbol{\theta}_k^{(t)} = \boldsymbol{\theta}^{(t)}$ 以便。回想一下，若满足以下条件，则每个用户设备都属于联邦学习组：$I_k^{(t)} = 0$；若满足 $I_k^{(t)} = 1$，则属于联邦蒸馏组。在用户设备端，每个联邦学习组用户设备 k_1 都会在本地训练，并得到梯度 $\mathbf{g}_{k_1}^{(t)} = \nabla F_{k_1}(D_{k_1}^{(t)}; \boldsymbol{\theta}_{k_1}^{(t)})$，其中 F_{k_1} 是交叉熵（CE）损失，$D_{k_1}^{(t)}$ 是用户设备 k_1 在轮次 t 采样、用于随机梯度下降（SGD）的本地数据集。类似地，联邦蒸馏组用户设备 k_2 首先使用其私有数据集 $D_{k_2}^{(t)}$ 执行本地随机梯度下降，以更新 $\boldsymbol{\theta}_{k_2}^{(t)}$，然后生成逻辑值 $\mathbf{z}_{k_2}^{(t)} = f_{k_2}(D^{(t)}\text{pub}; \boldsymbol{\theta}_{k_2}^{(t)})$，其中 f_{k_2} 是用户设备 k_2 的神经网络（NN），其参数为 $\boldsymbol{\theta}_{k_2}^{(t)}$；$D^{(t)}\text{pub}$ 是所有用户设备和基站共享的公共数据集，在轮次 t 采样，用于随机梯度下降。联邦学习组和联邦蒸馏组用户设备将 $\mathbf{g}_{k_1}^{(t)}$ 和 $\mathbf{z}_{k_2}^{(t)}$ 转换为传输信号 $\mathbf{X}^{(t)}$，通过上行链路发送，并按照第 II 节所述重构 $\hat{\mathbf{g}}_{k_1}^{(t)}$ 和 $\hat{\mathbf{z}}_{k_2}^{(t)}$。回想一下，$\hat{\mathbf{g}}_{k_1}^{(t)} = \mathbf{g}_{k_1}^{(t)} + \mathbf{e}_{k_1}^{(t)}$，且 $\hat{\mathbf{z}}_{k_2}^{(t)} = \mathbf{z}_{k_2}^{(t)} + \mathbf{e}_{k_2}^{(t)}$，其中 $\mathbf{e}_{k_1}^{(t)}$ 和 $\mathbf{e}_{k_2}^{(t)}$ 是有效噪声项，其取值由信噪比、信道和预处理决定，并作用于每个用户设备的梯度和逻辑值。

随后，在基站处，对重构的梯度和逻辑值进行聚合，得到

$$\hat{\mathbf{g}}^{(t)} = \sum_{k_1=1}^{K_1} \frac{|D_{k_1}^{(t)}|}{\sum_{\ell=1}^{K_1}|D_\ell^{(t)}|}\hat{\mathbf{g}}_{k_1}^{(t)}, \tag{3}$$

$$\hat{\mathbf{z}}^{(t)} = \sum_{k_2=K_1+1}^{K} \frac{|D_{k_2}^{(t)}|}{\sum_{\ell=K_1+1}^{K}|D_\ell^{(t)}|}\hat{\mathbf{z}}_{k_2}^{(t)}, \tag{4}$$

可等价地写作 $\hat{\mathbf{g}}^{(t)} = \mathbf{g}^{(t)} + \mathbf{e}_g^{(t)}$ 和 $\hat{\mathbf{z}}^{(t)} = \mathbf{z}^{(t)} + \mathbf{e}_z^{(t)}$。随后，按照 FL 和 FD 更新方向的凸组合更新全局模型：

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \alpha\eta_1\hat{\mathbf{g}}^{(t)} - (1-\alpha)\eta_2\nabla Q(D_{\text{pub}}^{(t)}; \boldsymbol{\theta}^{(t)}, \hat{\mathbf{z}}^{(t)}), \tag{5}$$

其中，$Q(D^{(t)}\text{pub}; \boldsymbol{\theta}^{(t)}, \hat{\mathbf{z}}^{(t)}) = \mathrm{KL}(\hat{\mathbf{z}}^{(t)}/\tau \,\|\, f(D^{(t)}\text{pub}; \boldsymbol{\theta}^{(t)})/\tau)$，$\mathrm{KL}(\cdot\|\cdot)$ 表示 Kullback–Leibler 散度（KLD）。这里，η_1 和 η_2 是学习率，f 是基站的 NN，τ 是控制分布平滑度的温度参数。在无噪声上行链路条件下，即 $\mathbf{e}_g^{(t)} = \mathbf{e}_z^{(t)} = \mathbf{0}$ 时，FL 和 FD 是 HFL 的特殊情况。具体而言，当 $\alpha = 1$ 和 $K = K_1$ 时，HFL 退化为 FL [2]；而当 $\alpha = 0$ 和 $K_1 = 0$ 时，则退化为 FD [10]。

B. 收敛性分析

我们首先陈述用于证明所提出的分层联邦学习框架收敛性的假设。

假设 III.1。（凸性）：令 $F(D; \boldsymbol{\theta}) = \sum_k \lambda_k F_k(D_k; \boldsymbol{\theta})$，其中 D 为总数据集，$0 < \lambda_k < 1$ 满足 $\sum_k \lambda_k = 1$。函数 $F(D; \boldsymbol{\theta})$ 对所有 $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ 均为 μ_1-强凸，即

$$F(D;\boldsymbol{\theta}_1) \geq F(D;\boldsymbol{\theta}_2) + \langle \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \nabla F(D;\boldsymbol{\theta}_2)\rangle + \frac{\mu_1}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2, \tag{6}$$

其中，$\langle\cdot,\cdot\rangle$ 表示内积。类似地，在式 (5) 后定义的 $Q(D; \boldsymbol{\theta}, \mathbf{z})$ 是 μ_2-强凸的，具体如下：

$$Q(D;\boldsymbol{\theta}_1, \mathbf{z}) \geq Q(D;\boldsymbol{\theta}_2, \mathbf{z}) + \langle \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \nabla Q(D;\boldsymbol{\theta}_2, \mathbf{z})\rangle + \frac{\mu_2}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2. \tag{7}$$

假设 III.2。（有界方差）：令 $D^{(t)}$ 表示在轮次 t采样的数据集。对于 F 和 Q，带小批量的随机梯度方差分别由某个 ψ_1^2 和ψ_2^2 上界。

$$\mathbb{E}\Big[\|\nabla F(D^{(t)};\boldsymbol{\theta})-\nabla F(D;\boldsymbol{\theta})\|^2\Big]\le\psi_1^2, \tag{8}$$

$$\mathbb{E}\Big[\|\nabla Q(D^{(t)};\boldsymbol{\theta},\mathbf{z})-\nabla Q(D;\boldsymbol{\theta},\mathbf{z})\|^2\Big]\le\psi_2^2, \tag{9}$$

其中 $D^{(t)}\subset D$。

假设 III.3.（有界梯度）：梯度范数分别被 G_1^2 和 G_2^2 上界限制，其中 F 和 Q。

$$\mathbb{E}\big[\|\nabla F(D;\boldsymbol{\theta})\|^2\big]\le G_1^2, \tag{10}$$

$$\mathbb{E}\big[\|\nabla Q(D;\boldsymbol{\theta},\mathbf{z})\|^2\big]\le G_2^2. \tag{11}$$

这些假设常用于联邦学习收敛性分析 [11]。此外，为处理逻辑值噪声，我们还作出以下假设。

假设 III.4.（logit 上的 Lipschitz 连续性）： Q 的梯度关于 z 满足 Lipschitz 连续性。

$$\|\nabla Q(D;\boldsymbol{\theta},\mathbf{z}_1)-\nabla Q(D;\boldsymbol{\theta},\mathbf{z}_2)\|\le L\|\mathbf{z}_1-\mathbf{z}_2\|. \tag{12}$$

基于这些假设，我们证明了所提出的分层联邦学习框架的收敛性。

命题 III.1. 当 $t\to\infty$时，假设 $\mathbf{e}_z^{(t)}$和 $\mathbb{E}[\mathbf{e}_g^{(t)}]=\mathbb{E}[\mathbf{e}_z^{(t)}]=\mathbf{0}$ 较小，则通过混合联邦学习训练得到的序列 $\{\boldsymbol{\theta}^{(t)}\}$ 满足 $\mathbb{E}\Big[\|\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}^*\|^2\Big]\le A/\bar{\mu}$，其中 $\bar{\mu}$ 和 A 为常数，且 $0<\bar{\mu}<1$。

证明。定义 σ_g、 σ_z、 $\boldsymbol{\theta}^*$ 和 $\boldsymbol{\delta}_q^{(t)}$如下。令 $\sigma_g=\sup_t\mathbb{E}[\|\mathbf{e}_g^{(t)}\|^2]$ 和 $\sigma_z=\sup_t\mathbb{E}[\|\mathbf{e}_z^{(t)}\|^2]$。假设最优参数 $\boldsymbol{\theta}^*$ 满足 $F(D^{(t)};\boldsymbol{\theta}^*)\le F(D^{(t)};\boldsymbol{\theta}^{(t)})$，且对所有 $\boldsymbol{\theta}^{(t)}$、 $D^{(t)}$ 和 $\mathbf{z}^{(t)}$，均有 $Q(D^{(t)};\boldsymbol{\theta}^*,\mathbf{z}^{(t)})\le Q(D^{(t)};\boldsymbol{\theta}^{(t)},\mathbf{z}^{(t)})$。最后， $\boldsymbol{\delta}_Q^{(t)}$ 定义为

$$\boldsymbol{\delta}_Q^{(t)}=\nabla Q(D_{\text{pub}}^{(t)};\boldsymbol{\theta}^{(t)},\hat{\mathbf{z}}^{(t)})-\nabla Q(D_{\text{pub}}^{(t)};\boldsymbol{\theta}^{(t)},\mathbf{z}^{(t)}). \tag{13}$$

当 $\mathbf{e}_z^{(t)}$较小时，泰勒展开的一阶形式给出 $\boldsymbol{\delta}_Q^{(t)}\approx\mathbf{e}_z^{(t)}\frac{\partial}{\partial\mathbf{z}}\nabla Q(D^{(t)}\text{pub};\boldsymbol{\theta}^{(t)},\mathbf{z})|_{\mathbf{z}=\mathbf{z}^{(t)}}$，从而可得 $\mathbb{E}[\boldsymbol{\delta}_Q^{(t)}]=\mathbf{0}$。根据假设 III.4， $\|\boldsymbol{\delta}_Q^{(t)}\|$ 有上界 $L\|\mathbf{e}_z^{(t)}\|$。

因此，更新规则 (5) 可改写为

$$\begin{aligned}\boldsymbol{\theta}^{(t+1)}=&\boldsymbol{\theta}^{(t)}-\alpha\eta_1(\nabla F(D^{(t)};\boldsymbol{\theta}^{(t)})+\mathbf{e}_g^{(t)})\\&-(1-\alpha)\eta_2(\nabla Q(D_{\text{pub}}^{(t)};\boldsymbol{\theta}^{(t)},\mathbf{z}^{(t)})+\boldsymbol{\delta}_Q^{(t)}).\end{aligned} \tag{14}$$

为推导 $\mathbb{E}[\|\boldsymbol{\theta}^{(t+1)}-\boldsymbol{\theta}^*\|^2]$的上界，推导过程如下。首先，根据假设 III.1 和 $\boldsymbol{\theta}^*$ 的定义，

$$\begin{aligned}&-2\alpha\eta_1\mathbb{E}[\langle\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}^*,\nabla F(D^{(t)};\boldsymbol{\theta}^{(t)})\rangle]\\&-2(1-\alpha)\eta_2\mathbb{E}[\langle\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}^*,\nabla Q(D^{(t)};\boldsymbol{\theta}^{(t)},\mathbf{z}^{(t)})\rangle]\\&\le-\alpha\eta_1\mu_1\mathbb{E}[\|\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}^*\|^2]-(1-\alpha)\eta_2\mu_2\mathbb{E}[\|\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}^*\|^2].\end{aligned} \tag{15}$$

其次，根据假设 III.2 和 III.3，

$$\begin{aligned}&\mathbb{E}[\|\nabla F(D^{(t)};\boldsymbol{\theta}^{(t)})-\nabla F(D;\boldsymbol{\theta}^{(t)})+\nabla F(D;\boldsymbol{\theta}^{(t)})\|^2\\&\le\psi_1^2+G_1^2.\end{aligned} \tag{16}$$

这些假设还可得出 $\mathbb{E}[\|\nabla Q(D^{(t)};\boldsymbol{\theta}^{(t)},\mathbf{z}^{(t)})\|^2]\le\psi_2^2+G_2^2$。最后，根据柯西-施瓦茨不等式，两个梯度的平均内积可由$\sqrt{(\psi_1^2+G_1^2)(\psi_2^2+G_2^2)}$界定。综合上述结果，我们得到

$$\begin{aligned}&\mathbb{E}[\|\boldsymbol{\theta}^{(t+1)}-\boldsymbol{\theta}^*\|^2]\\&\le(1-\alpha\eta_1\mu_1-(1-\alpha)\eta_2\mu_2)\mathbb{E}[\|\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}^*\|^2]\\&+\alpha^2\eta_1^2(\psi_1^2+G_1^2+\sigma_g)+(1-\alpha)^2\eta_2^2(\psi_2^2+G_2^2+L^2\sigma_z)\\&+2\alpha(1-\alpha)\eta_1\eta_2\sqrt{(\psi_1^2+G_1^2)(\psi_2^2+G_2^2)}\\&=(1-\bar{\mu})\mathbb{E}[\|\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}^*\|^2]+A.\end{aligned} \tag{17}$$

当通过调整 η_1、 η_2 和 α 来改变 $0<\bar{\mu}<1$ 时，可推导出上界 $A/\bar{\mu}$，即 $t\to\infty$。 □

同样，该上界也涵盖了 FL 和 FD 这两种特殊情况，其推导见 [6]。当 $\sigma_g=\sigma_z=0$ 和 $\alpha=1$时，可得 FL 的上界；否则，当 $\sigma_g=\sigma_z=0$ 和 $\alpha=0$时，可得 FD 的上界。尽管所提出的分层联邦学习框架的收敛性分析是在逻辑值误差较小的假设下推导的，因此并不严格适用于低信噪比场景，但第 IV 节的实证结果表明，通过恰当利用下一小节介绍的 HFL 自由度，该算法在低信噪比下仍能收敛，其表现与逻辑值误差较小的情形一致。

C. 自由度：自适应用户设备聚类与权重选择

在混合联邦学习的每个通信轮次中，基站会确定两项配置：将用户设备分配到联邦学习组或联邦蒸馏组，即选择传输梯度或逻辑值；以及用于融合由此产生的联邦学习和联邦蒸馏全局更新的权重。这些决策利用瞬时信道状态信息（信道状态信息）以及公共数据上的损失反馈，并假设通过轻量级计算实施这些决策所需的时间不超过信道相干时间。

1) 自适应用户设备聚类：对于每个用户设备 k，基站计算均衡后噪声增强因子 $q_k=1/(\rho[\mathbf{H}^{(t)\text{H}}\mathbf{H}^{(t)}]_{kk})$。 q_k 越小，表示梯度传输越可靠；而 q_k 越大，则表明由于逻辑值具有噪声鲁棒性，传输逻辑值更为可取。为了将用户设备划分到联邦学习组和联邦蒸馏组，我们对集合 $\{q_k\}_{k=1}^K$ 应用 Jenks自然断点优化 [12] ，并设定 $S=2$ 个类别。Jenks优化旨在最小化类内方差，并确定阈值 $q^\star$ ，以区分低噪声增强因子和高噪声增强因子。这种方法等价于一维 k-均值聚类 [13]。随后，基站在满足 $q_k\le q^\star$ 时将用户设备 k 分配到联邦学习组 $\mathcal{I}_k^{(t)}=0$ ，在满足 $q_k>q^\star$ 时将其分配到联邦蒸馏组 $\mathcal{I}_k^{(t)}=1$ 。

2) 自适应权重选择：计算出联邦学习和联邦蒸馏更新后，将它们各自的全局梯度下降方向定义为$\mathbf{d}_{\text{fd}}=-\eta_2\nabla Q(D^{(t)},\boldsymbol{\theta}^{(t)},\mathbf{z}^{(t)})$和 $\mathbf{d}(\text{pub},\hat{\mathbf{z}})$。为合并这两个方向，我们对凸权重 $(\sigma(s^{(w)}),1-\sigma(s^{(w)}))$ 进行优化，其中 $\sigma(\cdot)$ 是 Sigmoid函数，而 $s^{(w)}\in\mathbb{R}$ 是

算法1 混合联邦学习

对每个 $t \in \{1, ..., T\}$ 执行
  [BS]
  通过 Jenks优化方法，使用以下内容决定 $\mathcal{I}_k^{(t)}$ 的 $\forall k$ 获取的信道状态信息 $\mathrm{H}_{t\circ}$
  广播 $\boldsymbol{\theta}^{(t)}$ ，$\mathcal{I}_k^{(t)}$ 至所有用户设备。
  [UE]
  训练用户设备以获得梯度 $\mathrm{g}_k^{(t)}$.
  对于联邦学习组用户设备，将 g 转换为 $\mathbf{x}_{k_1}^{(t)}$ as $\mathbf{x}_{k_1\circ}^{(t)}$ 对于联邦蒸馏组用户设备，使用训练好的 $\boldsymbol{\theta}_{k_2}^{(t)}$，生成逻辑值 $\mathbf{z}_{k_2}^{(t)}$。然后，转换 $\mathbf{z}_{k_2}^{(t)}$ as $\mathbf{x}_{k_2}^{(t)}$.
  对于 $m \in \{1, ..., L\}$，执行
    [UE]
    通过上行链路发送 $\mathbf{x}^{(t)}[m]$ 。
    [BS]
    对每个信号执行零迫解码，如式（2）所示。
  结束循环
  [BS]
  将信号重构为梯度和逻辑值，如下所示 $\hat{\mathbf{g}}_{k_1}^{(t)}$ and 逻辑值 $\hat{\mathbf{z}}_{k_2}^{(t)}$.
  使用以下方法将参数和逻辑值聚合为 $\hat{}\mathbf{g}^{(t)}$ 和 $\hat{}\mathbf{z}^{(t)}$ 式（3）和式（4）。
  优化 FL/FD 权重，采用 阻尼牛顿法，并将损失函数设为 (18)，如式 (19) 所示。
结束循环

一个在第 w 个 epoch 更新的无约束标量。目标是最小化公共数据上的交叉熵损失

$$\mathcal{L}(s^{(w)}) = F(D_{\text{pub}}^{(t)}; \boldsymbol{\theta}^{(t)} + \sigma(s^{(w)})\mathbf{d}_{\text{fl}} + (1 - \sigma(s^{(w)}))\mathbf{d}_{\text{fd}}). \tag{18}$$

我们采用阻尼牛顿法求解此最小化问题，因为该方法只需几次迭代即可快速收敛 [14]，具体如下

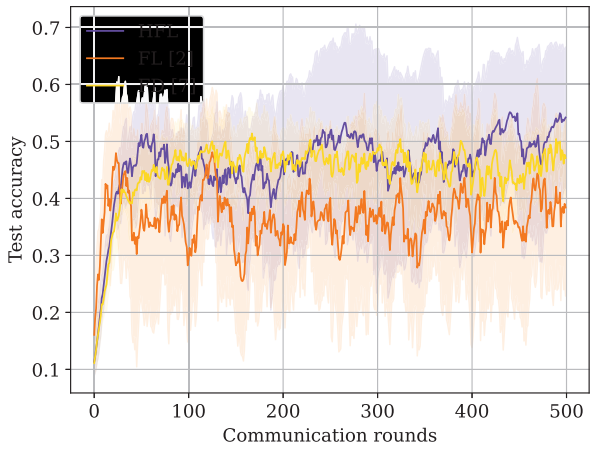
$$s^{(w+1)} = s^{(w)} - \eta_3 \frac{\partial \mathcal{L}(s^{(w)})/\partial s^{(w)}}{\partial^2 \mathcal{L}(s^{(w)})/\partial (s^{(w)})^2}, \tag{19}$$

其中，一阶和二阶导数通过中心有限差分近似，η_3 为阻尼因子。最终值 $s^\star$ 决定权重 $\alpha = \sigma(s^*)$，并在每一轮次重新计算。

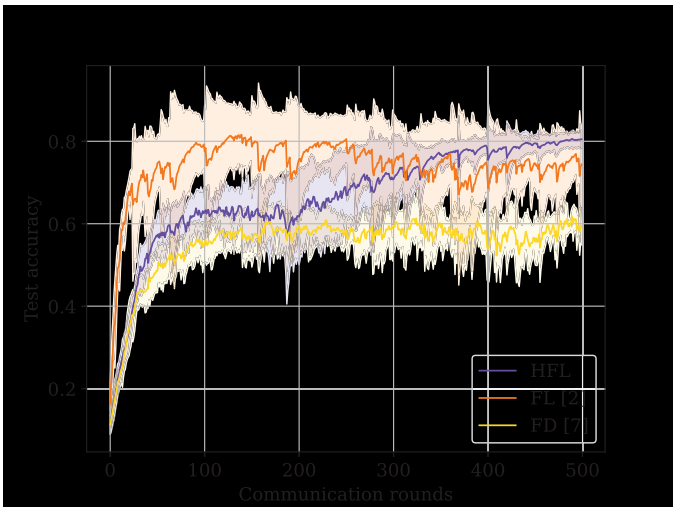
所提出的轻量级方法使混合联邦学习能够在每一轮次自适应地反映信道的瞬时状况。混合联邦学习整体流程的概要见算法1。

IV. 仿真结果

在本节中，我们通过在低信噪比下与联邦学习和联邦蒸馏进行比较，展示所提方法的性能。随后，我们验证所提供的自由度。在我们的系统中，考虑基站配备$N = 30$根天线，且 $K = 30$ 个用户设备均参与分布式学习，不进行调度，以解决包含 $C = 10$ 个类别的 MNIST 分类问题。在传输阶段不进行零填充，即设置$L = P/2 = CP_{\text{pub}}/2 = 39755$。为突出分层联邦学习训练的抗噪性，我们将低信噪比 ρ 设置为 -20dB 和 -15dB，



(a) 在 $\rho = -20$ dB下对比混合联邦学习、联邦学习和联邦蒸馏。



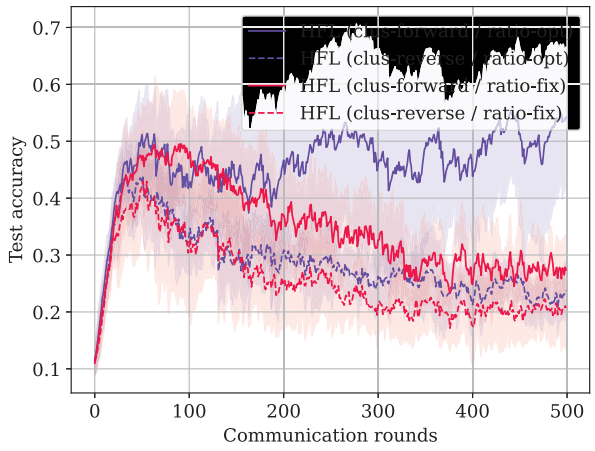
(b) 在 $\rho = -15$ dB下对比混合联邦学习、联邦学习和联邦蒸馏。

图2：低信噪比下混合联邦学习、联邦学习和联邦蒸馏的对比。

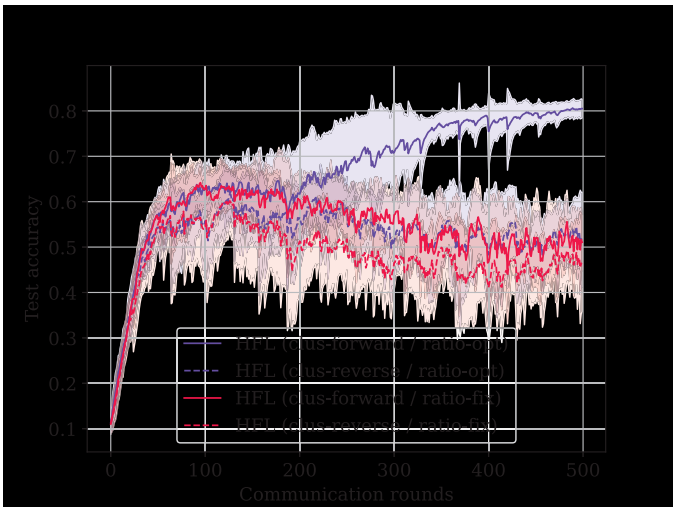
然后每个通信轮次评估一次测试准确率。模型训练的学习率和阻尼因子分别设为 $\eta_1 = \eta_2 = 0.01$和 $\eta_3 = 0.1$。本地训练和知识蒸馏的轮数为 1，阻尼牛顿法的迭代次数为 30，温度为 $\tau = 2$。

A. 比较联邦学习、联邦蒸馏和混合联邦学习

在图 2 中，我们比较了 FL、FD 和 HFL 在低信噪比下的测试准确率，信噪比分别为 $\rho = -20$dB 和 $\rho = -15$dB。FL 使用 FedAvg [2], , 而 FD 遵循 [10]。在更严苛的 $\rho = -20$dB 信噪比下，FL 的准确率低于 FD，这与其对信道噪声的敏感性一致：信道噪声会在聚合前破坏高维梯度。FD 受影响较小，因为它只传输低维逻辑值。HFL 的准确率最高，且比 FD 收敛更快。当信噪比提高到 $\rho = -15$dB 时，FL 的性能恢复并超过 FD，但收敛后仍低于 HFL；与此同时，FD 的收敛速度仍然较慢，且最终准确率较低。HFL 的准确率最高，但收敛速度比 FL 慢。在这两种信噪比下，HFL 均取得了最高的测试准确率



（a） $\rho = -20$ dB 时自由度的影响。



（b） $\rho = -15$ dB 时自由度的影响。

图 3：低信噪比下自由度的影响：聚类和比率选择。

通过结合基于梯度和基于逻辑值的更新，既继承了 FD 的抗噪能力，又利用了 FL 的学习性能。

B. 自由度的影响

在图3中，我们在四种配置下评估 HFL，以验证所提出的自由度选择方案：用户设备聚类（‘clus-forward’、‘clus-reverse’）和权重选择（‘weight-opt’、‘weight-fix’）。在聚类方面，以噪声增强因子为标准自适应地划分用户设备。‘clus-forward’遵循我们提出的规则，而‘clus-reverse’则采用相反的映射方式，即基站在 $\mathcal{I}_k^{(t)} = 0$ 时设置 $q_k > q^\star$，在 $\mathcal{I}_k^{(t)} = 1$ 时设置 $q_k \leq q^\star$。在权重选择方面，‘weight-opt’每轮调整权重，而‘weight-fix’则将其固定为 $\sigma(s^*) = 0.5$。正如预期，同时采用‘clus-forward’和‘weight-opt’可获得最高的测试准确率，证明了该方案对噪声的鲁棒性。此外，‘clus-forward’的准确率优于‘clus-reverse’，进一步支持了我们的启发式分配策略：将高噪声用户设备分配给联邦学习，将低噪声用户设备分配给联邦蒸馏。

V. 结论

本文提出了一种新颖的分层联邦学习框架，可根据无线信道条件调整训练策略，从而缓解联邦学习与联邦蒸馏各自的互补性弱点。我们首先介绍了整体框架并推导了其收敛性。随后，我们利用分层联邦学习的自由度，通过自适应用户设备聚类和权重选择，分别确定每个用户设备传输梯度还是逻辑值，以及联邦学习和联邦蒸馏贡献的权重；前者由 Jenks优化处理，后者则通过阻尼牛顿法处理。尽管结构简单，该框架不同于以往的工作，因为它能够针对每个用户设备，根据信道条件选择要传输的信息。数值结果表明，在充分利用其自由度时，分层联邦学习在低信噪比下始终能取得提升。未来的工作包括为过空中分层联邦学习设计零迫之外的波束成形器，并引入同时依据信道条件和本地数据分布确定的额外自由度。

参考文献

[1] 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 37 卷，第 3 期，第 50–60 页，2020 年 5 月。[2] B. McMahan、E. Moore、D. Ramage、S. Hampson 和 B. A. y Arcas，《基于去中心化数据的深度网络通信高效学习》，载于《人工智能与统计》，2017 年，第 1273–1282 页。[3] W. Y. B. Lim 等，《移动边缘网络中的联邦学习：全面综述》，IEEE通信调查与教程，第 22 卷，第 3 期，第 2031–2063 页，2020 年 4 月。[4] E. Jeong、S. Oh、H. Kim、J. Park、M. Bennis 和 S.-L. Kim，《通信高效的端侧机器学习：非独立同分布私有数据下的联邦蒸馏与增强》，arXiv 预印本 arXiv:1811.11479，2018 年 11 月。[5] L. Li、J. Gou、B. Yu、L. Du 和 Z. Y. D. Tao，《联邦蒸馏：综述》，arXiv 预印本 arXiv:2404.08564，2024 年 4 月。[6] Y. Mu、N. Garg 和 T. Ratnarajah，《大规模MIMO网络中的联邦蒸馏：动态训练、收敛性分析与信道感知学习》，IEEE认知通信与网络学报，2024 年 3 月。[7] G. Gad、E. Gad、Z. M. Fadlullah、M. M. Fouda 和 N. Kato，《带宽受限网络中通过联合知识蒸馏与差分隐私实现通信高效且保护隐私的联邦学习》，第 17 586–17 601 页，2024 年 11 月。[8] T. Lin、L. Kong、S. U. Stich 和 M. Jaggi，《联邦学习中稳健模型融合的集成蒸馏》，神经信息处理系统进展，第 33 卷，第 2351–2363 页，2020 年。[9] J.-H. Ahn、O. Simeone 和 J. Kang，《面向异构数据分布式边缘学习的无线联邦蒸馏》，载于 2019 年 IEEE 第 30 届个人、室内和移动无线电通信年度国际研讨会（PIMRC），2019 年 11 月，第 1–6 页。[10] L. Liu、J. Zhang、S. Song 和 K. B. Letaief，《采用主动数据采样的通信高效联邦蒸馏》，载于 ICC 2022—IEEE国际通信会议，2022 年 8 月，第 201–206 页。[11] H. Xing、O. Simeone 和 S. Bi，《无线设备到设备网络上的联邦学习：算法与收敛性分析》，IEEE通信领域精选期刊，第 39 卷，第 12 期，第 3723–3741 页，2021 年 10 月。[12] G. F. Jenks，《统计制图中的数据模型概念》，《国际制图年鉴》，第 7 卷，第 186–190 页，1967 年。[13] A. Heim、E. Gr¨oller 和 C. Heinzl，《AccuStripes：用于直观比较单变量数据分布的自适应分箱》，arXiv 预印本 arXiv:2207.13663，2022 年 7 月。[14] S. Boyd 和 L. Vandenberghe，《凸优化》。剑桥大学出版社，2004 年。