FEDCOVA：针对噪声标签的鲁棒联邦协方差学习

Xiangyu Zhong
香港中文大学信息工程系
xyzhong@ie.cuhk.edu.hk

Xiaojun Yuan
电子科技大学无线通信国家重点实验室
xiyuan@uestc.edu.cn

Ying-Jun Angela Zhang
香港中文大学信息工程系
yjzhang@ie.cuhk.edu.hk

arXiv:2603.04062v1 [cs.LG] 4 Mar 2026

摘要

分布式数据集中的噪声标签会导致严重的局部过拟合，进而损害联邦学习（FL）中的全局模型。现有解决方案大多依赖于筛选干净设备或与公共干净数据集对齐，而非赋予模型本身鲁棒性。本文提出*FedCova*，一种无需依赖外部资源的联邦协方差学习框架，通过从特征协方差这一全新视角增强模型的内在鲁棒性，消除对外部资源的依赖。具体而言，FedCova 将数据编码到具有判别性且鲁棒的特征空间，以容忍标签噪声。基于互信息最大化，我们设计了一种新颖的联邦有损特征编码目标，该目标仅依赖类别特征协方差，并包含误差容忍项。利用由协方差刻画的特征子空间，我们构建了子空间增强型联邦分类器。FedCova 通过协方差统一了三个关键过程：（1）训练网络进行特征编码；（2）直接根据学习到的特征构建分类器；（3）基于特征子空间纠正噪声标签。我们在异构数据分布下的对称和非对称噪声设置中实现了 FedCova。在 CIFAR-10/100 和真实世界噪声数据集 Clothing1M 上的实验结果表明，与最先进方法相比，FedCova 具有更优的鲁棒性。

1 引言

探索对噪声标签的鲁棒性对机器学习至关重要，因为它能够增强大规模模型训练对海量粗粒度标注数据集的适应能力 Chen 等（2019）。传统方法主要通过若干离线 抗噪策略改进训练过程，通常会检测、筛选或纠正噪声样本 Cheng 等人（2022）；Ren 等人（2018）；Englesson 和 Azizpour（2021）；Xia 等人（2024）。这些启发式方法的鲁棒性往往较为脆弱，尤其是在标签噪声严重时。为解决这一问题，近年来许多算法通过利用辅助干净数据集或构建多模型结构来拓展学习框架 Chen 等人（2019）；Zheng 等人（2024）。尽管这些扩展能够提升性能，但它们仍依赖额外资源 ，而非通过更深入地理解数据来增强模型固有的鲁棒性。

在这一更广泛的背景下，噪声标签下的联邦学习（FL）成为一个更具挑战性的场景。随着范式向边缘人工智能转变，人们越来越期待其实现无处不在但又具备本地性、相互连接但又能保护隐私 Tan 等人（2023）；Chen 等人（2024）。作为一种典型的分布式范式，FL 促进了边缘设备上的协同学习，并通过数据本地性保护隐私 McMahan等人（2017）；Kairouz 等人（2021）。然而，分布式训练的性质使 FL 过度依赖有限且局部化的数据 Kairouz 等人（2021）；Li 等人（2019）。一方面，从数据源的角度来看，边缘设备上收集的标签容易受到

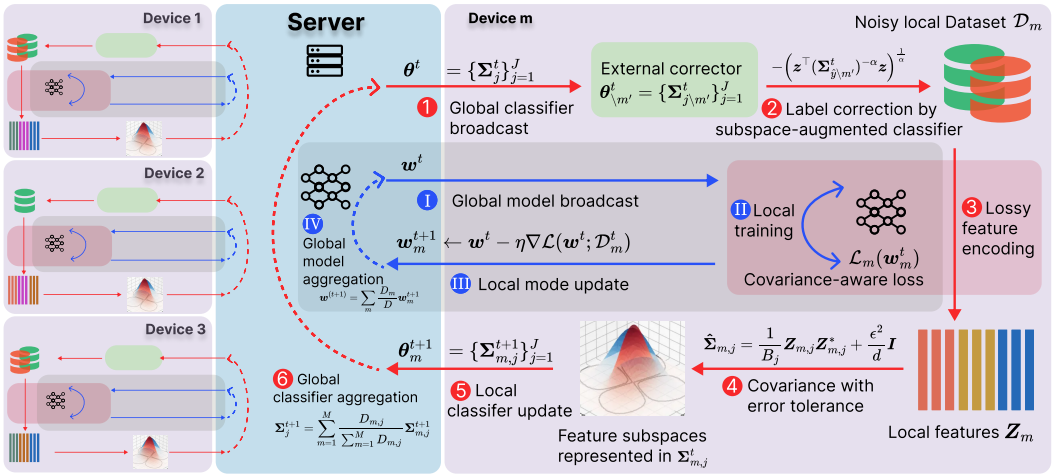


图 1：FedCova 框架概览。绿色数据表示干净数据，红色数据表示标签错误的数据。粉色矩形表示协方差感知特征学习。带有蓝色箭头圆圈的灰色矩形 - 表示传统的 FL 流程，而环绕其周围的红色箭头圆圈 - 则表示 FedCova 在协方差流动的过程中构筑起抵御标签噪声的堡垒 $\boldsymbol{\Sigma}$。具体而言，服务器首先将 全局模型和 全局分类器广播到边缘设备。 随后进行标签校正，之后设备 执行本地特征学习，并将 本地模型和 本地分类器更新至服务器。然后，服务器 对其进行聚合，以开展下一轮迭代。

标注错误、传感器故障和对抗性标签攻击的影响 Song & 等人（2022）；Han 等人（2020）。另一方面，本地模型对有限的噪声数据更加脆弱，导致误导性的过拟合，并在聚合后进一步污染全局模型。随后，模型广播进一步传播噪声更新，在多轮迭代中不断累积。与此同时，现有大多数 FL 研究都集中于异构性缓解，进一步加剧了对准确数据的依赖 Karimireddy 等人（2020）；Zhang 等人（2022）。这些方法在噪声标签下容易出现灾难性的性能下降，因为对噪声数据的过拟合会削弱其分布对齐策略 Lu 等人（2024）；Xu 等人（2022）。

近期针对噪声标签提升 FL 鲁棒性的研究，主要将集中式学习中的离线技术迁移到 FL 框架中，例如样本选择和鲁棒损失正则化 Han 等人（2018）；Song & 等（2022）；Yi & Wu（2019）；Di 等人（2024）；Xu 等人（2022）；Wu 等人（2023a）。然而，这些方法大多忽视了 FL 的独特属性，包括其分布式拓扑所带来的内在交叉验证能力 Ovi 等（2023），以及通过潜在表示共享捕获分布式结构的潜力 Tan 等人（2022）。相比之下，一些研究利用 FL 的分布式特性，例如采用双模型协同训练 Ji 等人（2024）或跨设备蒸馏 Tsouvalas 等人（2024）；Lu 等人（2024），但其效果取决于额外资源，尤其是冗余模型或干净公开数据集。这揭示了一个根本瓶颈：在噪声标签下，模型仍受制于严格的预测-标签对齐，且往往受交叉熵等目标函数限制；这类函数会无意中记忆噪声，从而内在地促使预测趋向可能错误的标签。在校正策略方面，大多数方法通过预测或损失的一致性识别损坏样本，然后依赖干净客户端在训练期间重新标记或重新加权 Tam 等人（2023）；Yang 等（2022）；Xu 等人（2022）。然而，依赖干净数据集并不现实：当干净客户端数量较少，或在非独立同分布（non-i.i.d.）数据条件下仅持有有限样本或类别时，这些方法可能会失效，因为它们并未从根本上利用数据的内在特征。

表示学习强调提取能够捕捉内在数据结构的判别性且富含信息的特征的重要性 Bengio 等人 (2013)；Han 等人 (2020)。受此启发，在存在噪声标签的联邦学习中利用特征空间，可以为鲁棒学习提供更可靠的依据，因为特征统计量较少受到标签噪声的扭曲 Arpit 等人 (2017年)；Zhang 等人 (2023)；Chan 等人 (2022)。对比学习已通过捕捉样本特征之间的相对关系，被应用于噪声标签条件下的集中式学习 Li 等人 (2021)；Yi 等人 (2022)。然而，它需要对比所有样本对，这在联邦学习中不可行，因为数据仍然

分布在各设备上，且跨设备比较受到限制。因此，此类方法在联邦学习中表现欠佳 Zhang 等人 (2023)；Duan 等人 (2022)。在联邦学习中利用特征的尝试通常借助对齐正则化来聚合局部特征均值/质心 Tan 等人 (2022)；Xu 等人。但噪声标签会直接扭曲所估计的质心，将错误信息传播至全局模型，并加重本已由交叉熵损失造成的过拟合。

为弥合现有研究中发现的差距，我们提出了**FedCova**，一种无需依赖外部数据且具有鲁棒性的协方差-**感知联邦**特征学习框架，旨在减轻联邦学习（FL）中噪声标签的影响（概览见图 1）。不同于依赖辅助干净数据或噪声先验的方法，FedCova 无需此类外部依赖。具体而言，FedCova 将联邦学习视为一种特征编码器，通过捕捉特征中固有的统计结构，学习具有判别性且抗噪的表示。我们利用类条件特征的协方差构建特征空间，使其在保持类别可分性的同时对噪声具有鲁棒性。FedCova 通过特征协方差无缝整合了三个关键过程：(1) 用于特征编码的网络训练；(2) 直接基于已学习特征构建内在分类器；(3) 基于特征子空间的标签校正。我们提出了一种基于互信息最大化的信息论训练目标，该目标仅取决于特征协方差，并将其扩展为带有误差容忍项的有损变体。在此目标的驱动下，学习到的编码将预测输出与观测标签之间的依赖关系解耦，同时在正交特征子空间中捕捉类别的判别性统计信息。在服务器端，聚合后的协方差用于对齐全局分类器，并确保各设备间特征表示的一致性。基于这些全局协方差，FedCova 构建最大后验（MAP）分类器，并进一步通过子空间增强对其进行泛化，以增强判别能力。每台设备收到全局分类器后，都会利用该分类器独立设计局部校正器来重新标记，从而有效校正标签，同时避免自校正带来的缺陷。我们的主要贡献总结如下：

- 我们提出 FedCova，这是一种统一的、无依赖的协方差感知型联邦特征学习框架，旨在提升联邦学习的鲁棒性。FedCova 通过特征协方差，整合了特征编码、内在分类器构建和基于子空间的标签校正。
- 我们为联邦有损特征学习提出了一种基于协方差的信息论损失函数。该目标以最大化互信息为基础，约束类条件特征的协方差 结构，并通过误差容忍项增强判别能力和噪声韧性。
- 我们通过协方差 聚合，提出了一种联邦分类器对齐策略。服务器通过子空间增强构建全局最大后验分类器，客户端利用该分类器构建外部校正器，从而校正噪声标签并避免自偏差。
- 我们在 CIFAR-10/100 上针对多种噪声设备比例和样本噪声比，在对称和非对称噪声模式下开展了全面实验；同时，也在具有非独立同分布数据分布的真实噪声数据集 Clothing1M 上进行了实验。结果表明，与最先进方法相比，FedCova 对噪声标签具有更强的鲁棒性。

2 预备知识

联邦特征学习。 使用噪声标签进行训练会破坏输入—标签映射，从而降低性能，并导致过拟合和预测错误。我们不打算仅仅依赖这种直接映射，而是希望探究，判别性特征空间中的表征能否通过调节图像、特征和标签之间的关系，提供更强的鲁棒性。有关相关研究和以往解决方案的详细讨论，请参阅附录 A。

考虑一个由 M 台边缘设备组成的联邦学习（FL）系统，以协作训练一个深度学习模型。该联邦学习系统旨在从用于分类任务的高维多类别图像中学习具有判别性和代表性的特征结构。分别用 $\mathcal{X}$、 $\mathcal{Z}$ 和 $\mathcal{Y}$ 表示数据空间、特征空间和标签空间。令 $\boldsymbol{x} \in \mathcal{X}$、 $\boldsymbol{z} \in \mathcal{Z}$ 和 $y \in \mathcal{Y}$ 分别表示对应随机变量 X、 Z 和 Y 的实例。考虑 $\mathcal{Y} = \{1, 2, \ldots, J\}$，其中 J 是类别数。用 $\mathcal{D}_m = p(\boldsymbol{x},\ y|m)$ 表示设备 $m \in \mathcal{M}$ 的数据集分布，其中 $\mathcal{M}$ 是设备集。设备中样本的联合分布表示为 $p(\boldsymbol{x},\ y,\ m) = p(m)p(\boldsymbol{x},\ y|m) = p(m)p(y|m)p(\boldsymbol{x}|m,\ y)$。所有

设备的全局混合分布为 $\mathcal{D} = \mathbb{E}_{m\sim p(m)}\mathcal{D}_m$。联邦学习系统通过最小化如下损失函数进行训练：

$$\min_{\boldsymbol{w}^t} \quad \mathcal{L}(\boldsymbol{w}^t) := \mathbb{E}_{m\sim p(m)}\mathcal{L}_m(\boldsymbol{w}^t), \tag{1}$$

$$\text{with } \mathcal{L}_m(\boldsymbol{w}^t) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_m}\ell(f_{\boldsymbol{w}^t}(\boldsymbol{x}), y) \tag{2}$$

其中 $f_{\boldsymbol{w}^t}\colon \mathcal{X}\to\mathcal{Z}$ 是一个网络，其全局模型参数 $\boldsymbol{w}^t$ 在第 t个通信轮次中，将输入数据映射为特征表示。具体而言，我们旨在学习数据样本 $\boldsymbol{x}\in\mathbb{R}^D$ 的表示 $\boldsymbol{z}\in\mathbb{R}^d$，其分布为由神经网络 $\boldsymbol{w}^t\in\mathbb{R}^s$参数化的 $p_{\boldsymbol{w}^t}(\boldsymbol{z}|\boldsymbol{x})$。联邦学习网络将 $\boldsymbol{z} = f_{\boldsymbol{w}^t}(\boldsymbol{x})$作为确定性特征编码器输出。通过将其等价表示识别为 $p_{\boldsymbol{w}^t}(\boldsymbol{z}|\boldsymbol{x}) = \delta(\boldsymbol{z} - f_{\boldsymbol{w}^t}(\boldsymbol{x}))$，可以将其视为概率特征编码器的一个特例，其中 $\delta(\cdot)$是狄拉克δ函数。

在通信轮次 t中，每个设备 m 基于本地数据集 $\mathcal{D}_m$ 和 $\boldsymbol{w}^t$，通过随机梯度下降（SGD）进行本地 $\boldsymbol{w}_m^{t+1}\leftarrow \boldsymbol{w}^t - \eta\nabla\mathcal{L}(\boldsymbol{w}^t;\mathcal{D}_m)$，以获得新的模型参数（；），并将其上传至服务器。服务器将本地更新聚合为 $\boldsymbol{w}^{t+1} = \mathbb{E}_{m\sim p(m)}\boldsymbol{w}_m^t$，然后将其广播给边缘设备，以进行下一轮训练。

先验。 通常，学习特征 $\boldsymbol{z}$ 的几何与统计特征在神经网络的潜在层中仍然不明确。为了驱动网络生成信息丰富的特征，以形成能够抵抗噪声标签的鲁棒表示，我们将特征建模为服从某种概率分布。具体而言，我们在所学特征上引入高斯混合（GM）先验，其中每个高斯分量对应一个不同的类别，即

$$\begin{aligned} p(\boldsymbol{z}) &= \sum_{j=1}^{J} p(\boldsymbol{z}, y=j) = \sum_{j=1}^{J} p(j)p(\boldsymbol{z}\mid y=j) \\ &= \sum_{j=1}^{J} \pi_j^t \mathcal{N}(\boldsymbol{\mu}_j^t, \boldsymbol{\Sigma}_j^t), \end{aligned} \tag{3}$$

参数为 $\boldsymbol{\theta}^t = \{\pi_j^t, \boldsymbol{\mu}_j^t, \boldsymbol{\Sigma}_j^t\}_{j=1}^J$，其中 $\pi_j^t = p(j)$表示第 j个类别的分量权重，$\boldsymbol{\mu}_j^t$ 和 $\boldsymbol{\Sigma}_j^t$ 表示第 t 个通信轮次中第 j个类别内特征 $\boldsymbol{z}$ 的均值与协方差矩阵。

与为每个类别分配不同均值并对类别均值施加强惩罚的传统先验Tan等人（2022）不同，我们假设所有分量均值均为零，即 $\boldsymbol{\mu}_j^t = \mathbf{0}, \forall j\in[1,J]$。这种零均值设计旨在消除对显式类别中心的依赖，而类别中心直观上更容易因标签噪声而产生偏差。相反，我们希望系统对特征的协方差结构进行建模。通过消除均值的影响，我们不仅减少了表征中需要估计的参数数量，还促使模型仅通过协方差矩阵捕捉每个类别的独特模式与特征间依赖关系，从而在所学特征中保留数据的内在结构。

3 有损学习目标

信息论损失 我们旨在发现一个表示 $p_{\boldsymbol{w}^t}(\boldsymbol{z}\mid\boldsymbol{x})$，其中包含关于其标签 y 的最大信息。根据 Bell 与 Sejnowski（1995年）提出的信息最大化原理，典型的优化目标是最大化互信息，即衡量两个变量之间共享的信息量，换言之，也就是它们之间的统计依赖关系。基于此，每个设备 m 的目标函数是最大化 Z_m 与 Y_m之间的互信息，换言之，

$$\min_{\boldsymbol{w}_m^t} \mathcal{L}_m(\boldsymbol{w}_m^t) := -I(Z_m; Y_m) = h(Z_m\mid Y_m) - h(Z_m), \tag{4}$$

其中，$h(\cdot)$ 是微分熵函数。我们将 $Z_m\sim\mathcal{N}(\mathbf{0},\Sigma_m)$ 的分布近似为高斯分布，其协方差矩阵为 Σ_m。基于 GM 先验假设，式（4）中的负互信息损失可表示为

$$\mathcal{L}_m(\boldsymbol{w}_m^t) = \frac{1}{2}\sum_{j=1}^{J}\pi_{m,j}^t \log\det\left(\boldsymbol{\Sigma}_{m,j}^t\right) - \frac{1}{2}\log\det\boldsymbol{\Sigma}_m^t. \tag{5}$$

每个客户端均使用式（5）中的目标进行训练，从而使本地模型 $\boldsymbol{w}_m^t$ 学习类别判别特征。请注意，该目标函数与特征均值无关，因为高斯变量的熵仅由其协方差矩阵决定。这一观察与我们关于零均值高斯混合先验的假设一致。此外，从互信息最大化的角度来看，专注于协方差可能使我们能够学习到效果良好的特征。

有损表示。式（5）中的目标旨在最大化特征 Z_m 与其标签 Y_m之间的互信息。然而，当标签存在噪声时，在计算每个类别的协方差 $\boldsymbol{\Sigma}_{m,\ j}^t$ 时，特征与噪声标签之间的耦合仍会对训练产生隐式影响，使模型容易过拟合。为此，我们通过协方差结构在特征空间中引入受控的可变性，提出该目标的一个有损变体。在噪声标签条件下的 FL 中，令 $\mathcal{Y}'$，Y'，y'和 $\mathcal{Z}'$，Z'，$\boldsymbol{z}'$ 分别表示噪声标签和有损特征的空间、随机变量及实例。为容忍噪声标签，我们在特征空间 $\mathcal{Z}'$中引入弹性。在噪声标签 y' 的监督下，将特征从 $\boldsymbol{x}$编码得到的输出 $\boldsymbol{z}'$ 视为准确表示 $\boldsymbol{z}$的噪声版本。因此，对于噪声标签条件下的 FL，我们处理的是 $f'_{\boldsymbol{w}^t}\colon \mathcal{X} \to \mathcal{Z}'$。不失一般性，我们将编码特征的偏差建模为加性高斯误差：

$$\boldsymbol{z}'_m = \boldsymbol{z}_m + \boldsymbol{n}, \ \text{ with } \ \boldsymbol{n} \sim \mathcal{N}\left(\boldsymbol{0}, \frac{\epsilon^2}{d}\boldsymbol{I}\right), \tag{6}$$

其中，ϵ^2 是失真的平方误差容限，$\boldsymbol{n}$ 与特征 $\boldsymbol{z}$ 无关，因此$\mathbb{E}[\|\boldsymbol{z}'_m - \boldsymbol{z}_m\|_2^2] = \epsilon^2$。在训练过程中，对于每个客户端 m，只能访问特征样本以估计协方差矩阵。对于一批 B 训练数据 $\{\boldsymbol{x}_m^i, y_m'^i\}_{i=1}^B$，FL 框架对数据样本 $\boldsymbol{X}_m \triangleq [\boldsymbol{x}_m^1, ..., \boldsymbol{x}_m^B]$ 进行编码，得到有损特征样本 $\boldsymbol{Z}_m \triangleq [\boldsymbol{z}_m^1, ..., \boldsymbol{z}_m^B]$。因此，有损特征的估计协方差矩阵为

$$\hat{\boldsymbol{\Sigma}}_m = \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{z}_m'^i \boldsymbol{z}_m'^{i*} = \frac{1}{B}\boldsymbol{Z}_m\boldsymbol{Z}_m^* + \frac{\epsilon^2}{d}\boldsymbol{I}, \tag{7}$$

其中，$(\cdot)^*$ 是转置操作。类似地，每个类别的协方差矩阵$\hat{\boldsymbol{\Sigma}}_{m,j} = \frac{1}{B_j}\boldsymbol{Z}_{m,j}\boldsymbol{Z}_{m,j}^* + \frac{\epsilon^2}{d}\boldsymbol{I}$，其中 B_j 是 B 中的子批次，包含 j类的样本，而$\boldsymbol{Z}_{m,j} = [\boldsymbol{z}_{m,j}^1, ..., \boldsymbol{z}_{m,j}^{B_j}]$ 是相应的特征样本。因此，式（5）中的损失函数可重新表述为

$$\begin{aligned}\mathcal{L}_m(\boldsymbol{w}_m^t) = &\sum_{j=1}^{J} \frac{B_j}{2B} \log\det\left(\frac{1}{B_j}\boldsymbol{Z}_{m,j}\boldsymbol{Z}_{m,j}^* + \frac{\epsilon^2}{d}\boldsymbol{I}\right) \\ &- \frac{1}{2}\log\det\left(\frac{1}{B}\boldsymbol{Z}_m\boldsymbol{Z}_m^* + \frac{\epsilon^2}{d}\boldsymbol{I}\right).\end{aligned} \tag{8}$$

这实质上促使协方差所表示的每个类别的特征子空间彼此区分，甚至相互正交，从而促进判别。这与分类中的编码率约简理论 Yu 等人（2020）相一致。

备注 3.1。我们的有损学习目标能够抵御噪声标签，这可归因于两个方面的放宽：(1) 模型被引导去学习特征空间的统计结构。我们不再过分强调特征的精确取值（这些取值对应均值统计量），而是允许特征呈现结构化的偏差，例如类高斯分布。对于判别而言，只需从这种偏差中捕捉有用的规律，即探索协方差统计即可。(2) 通过扰动特征协方差实现的有损表示，实际上使每个类别的椭球特征子空间趋于球形，从而可能放宽类别决策边界。也就是说，尽管我们仍力求保留给定标签信息中的最大信息量，但我们获得了一种具有韧性的特征输出；基于从原始数据压缩而来的特征子空间交织，它可能略微偏向另一个类别。有关信息论有损学习目标的详细解读及玩具示例，请参见附录 B。相应的数值分析见附录 E.1。

4 通过协方差聚合的联邦分类器

内在MAP分类器。 对特征进行编码后，我们接下来构建分类器，用于预测和后续校正。在依据式 (8) 中的目标函数完成训练后，编码后的特征自然形成高斯混合聚类。因此，无需另行构建神经网络分类器。相反，我们利用特征的概率结构（如式 (3) 所定义，并通过特征学习进一步强化），设计一个白盒分类器。给定参数$\boldsymbol{\theta}^t = \{\pi_j^t, \boldsymbol{\Sigma}_j^t\}_{j=1}^J$，可通过直接计算每个高斯分量（即每个类别）的似然并选择后验概率最高的类别来完成分类。因此， $\boldsymbol{\theta}^t$ 正是高斯判别分析（GDA）下的内在分类器（Hastie 与 Tibshirani（1996））。然而，出于隐私和通信效率考虑，分布式数据集仍保留在本地设备上，因此所有编码后的特征 $\boldsymbol{z}_m$ 也应留在本地。因此，我们的系统采用联邦分类器：每个客户端根据本地可用的特征独立估计自己的本地分类器参数 $\boldsymbol{\theta}_m^t$ ，随后将这些参数聚合以构建全局分类器 $\boldsymbol{\theta}^t$。在第 t 个通信轮次中，每台设备 m 通过最大似然估计（MLE）估计本地分类器 $\boldsymbol{\theta}_m^t = \{\pi_{m,j}^t, \boldsymbol{\Sigma}_{m,j}^t\}_{j=1}^J$ ，如下所示

$$\pi_{m,j}^t = \frac{D_{m,j}}{D_m}, \ \boldsymbol{\Sigma}_{m,j}^t = \frac{1}{D_{m,j}} \sum_{i=1}^{D_{m,j}} \boldsymbol{z}_m^i \boldsymbol{z}_m^{i*} + \frac{\epsilon^2}{d} \boldsymbol{I}, \tag{9}$$

对于类别 j，其中 $\boldsymbol{z}_m^i$ 是第 i 个通信轮次中设备 m 上第 t 个样本的特征向量，$D_{m,\ j}$ 是设备 m 中第 j 类别的数据量， D_m 是设备 m 的本地数据集大小。完成本地估计后，每台设备将本地分类器 $\boldsymbol{\theta}_m^t$ 上传至服务器。随后，服务器聚合这些本地分类器，得到全局分类器 $\boldsymbol{\theta}^t = \{\pi_j^t, \boldsymbol{\Sigma}_j^t\}_{j=1}^J$ ，如下所示

$$\pi_j^t = \sum_{m=1}^M \frac{D_m}{\sum_{m=1}^M D_m} \pi_{m,j}^t, \boldsymbol{\Sigma}_j^t = \sum_{m=1}^M \frac{D_{m,j}}{\sum_{m=1}^M D_{m,j}} \boldsymbol{\Sigma}_{m,j}^t, \tag{10}$$

对于类别 j，作为高斯消息合并。请注意，通信$\boldsymbol{\theta}_m^t$ 时的隐私问题可通过以下两个原因得到保障：1）它仅表征经过大幅降维的特征，而非原始数据分布。2）在式（9）的协方差矩阵中加入误差容忍项，进一步模糊了相关信息。收到全局分类器后，可依据最大后验概率（MAP）准则进行分类，如下所示

$$\hat{y} = \arg\max_j \left\{ p(y = j \mid \boldsymbol{z}) \propto \pi_j^t \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_j^t) \right\}. \tag{11}$$

子空间增强分类器。 最大后验分类器的设计与训练目标无缝契合，二者均致力于增强类高斯特征子空间中的判别能力。通过展开高斯分布并取对数，式（11）表明 $\boldsymbol{z}$ 属于第 j个类别的概率与$\log\ \pi_j^t - \frac{1}{2}\log\ \|\boldsymbol{\Sigma}_j^t\| - \frac{1}{2}\boldsymbol{z}^\top(\boldsymbol{\Sigma}_j^t)^{-1}\boldsymbol{z}$成正比。由于式（8）_中的目标函数借助不同类别的子空间正交性实现判别，为公平比较，可假设不同特征子空间的体积相等，从而$\|\boldsymbol{\Sigma}_u\| = \|\boldsymbol{\Sigma}_v\| \forall u, v \in [J]$。同时，噪声标签会在 π_j^t中引入直接的估计偏差，而该偏差在训练过程中并未被正则化。因此，我们转而采用马氏距离项 $\boldsymbol{z}^\top(\boldsymbol{\Sigma}_j^t)^{-1}\boldsymbol{z}$ ，并将其推广为子空间增强形式，即

$$p(y = j \mid \boldsymbol{z}) \propto -\left(\boldsymbol{z}^\top (\boldsymbol{\Sigma}_j^t)^{-\alpha} \boldsymbol{z}\right)^{\frac{1}{\alpha}}, \tag{12}$$

其中， α 是增强系数。一般而言，较大的 α 通常会带来更强的判别能力。面对标签噪声时，设置合适的 α 需要在更强的判别能力与更高的噪声标签容忍度之间权衡。更多相关讨论见附录 E.2。

外部校正器。 聚合后，可将全局分类器 $\boldsymbol{\theta}^t = \{\boldsymbol{\Sigma}_j^t\}_{j=1}^J$ 广播到边缘设备，以便对本地数据集中的标签进行校正。我们计划利用 $\boldsymbol{\theta}^t$ 对训练样本进行回溯分类，以检测噪声标签，并在校正轮次中校正噪声可能性较高的标签。基于交叉验证原理（Ovi 等（2023）；Chen 等（2019）；Berrar 等（2019）），对于每个本地设备，我们从全局分类器中提取外部信息

以用作外部校正器。具体而言，设备 m'的外部校正器 $\boldsymbol{\theta}^t_{\setminus m'} = \{\boldsymbol{\Sigma}^t_{j\setminus m'}\}_{j=1}^J$ 计算如下：

$$\boldsymbol{\Sigma}^t_{j\setminus m'} = \frac{\sum_{m=1}^M D_{m,j}\boldsymbol{\Sigma}^t_j - D_{m',j}\boldsymbol{\Sigma}^t_{m',j}}{\sum_{m=1}^M D_{m,j} - D_{m',j}}. \tag{13}$$

We 可以根据相应的外部校正器 $\boldsymbol{\theta}^t_{\setminus m}$ 验证本地数据集 $\mathcal{D}_m$，并进行预测 t 按照式（12）接收 $\hat{}y$ 及其对应的概率 $p(y = \hat{}y \mid \boldsymbol{z}) \propto -\left(\boldsymbol{z}^\top (\boldsymbol{\Sigma}^t_{\hat{y}\setminus m'})^{-\alpha}\boldsymbol{z}\right)^\alpha$。样本重新标注按如下方式进行：

$$(\boldsymbol{x}, y) \leftarrow \{(\boldsymbol{x}, \hat{y}) \mid p(y = \hat{y} \mid \boldsymbol{z}) \geq \eta_c \;\&\; \hat{y} \neq y\}, \tag{14}$$

其中，η_c 是置信度阈值。值得注意的是，从特征编码、分类到校正，我们始终遵循构建并利用鲁棒判别性特征空间这一关键原则。协方差矩阵是表示特征空间的核心，它在边缘设备与服务器之间交换，以实现联邦对齐。整个算法总结如附录 C 中的算法 1 所示。

5 仿真结果

5.1 实验设置

联邦学习系统。 我们在三个数据集上评估算法在图像分类任务中的性能，包括 CIFAR-10、CIFAR-100（Krizhevsky 等（2009））以及真实世界噪声数据集 Clothing1M（Xiao 等（2015））。在网络架构方面，我们对 CIFAR-10 采用 ResNet-18（He 等（2016）），对 CIFAR-100 采用 ResNet-34，对 Clothing1M 采用 ResNet-50。完整的实验设置细节见附录 D.1。

噪声设置与数据异质性。对于 Clothing1M 以外的数据集，我们引入双层噪声方案，以在对称噪声模式和非对称噪声模式下模拟真实的噪声场景。该设置由噪声设备比例（记为 ρ）和样本噪声比例（记为 τ）刻画。在主要实验中，我们将噪声对 (ρ, τ) 在对称噪声模式下设为 {(0.4, 0.5), (0.4, 0.7), (0.6, 0.5), (0.6, 0.7), (0.8, 0.5), (0.8, 0.7)}，在非对称噪声模式下设为 {(0.4, 0.3), (0.4, 0.5), (0.4, 0.7), (0.6, 0.3), (0.6, 0.5), (0.6, 0.7)}。对于数据异质性设置，我们采用同时考虑类别异质性和数据集大小异质性的一般分布集合，二者分别由概率为 p 的伯努利分布和以 $\alpha_{\text{Dir}} > 0$为参数的狄利克雷分布模拟。我们将 $p = 0.5$和 $\alpha_{\text{dir}} = 5$ 设为相对较高的非独立同分布场景。

基线方法。 我们将所提出的算法与多种针对噪声标签的联邦学习（FL）最先进（SOTA）框架进行了比较，包括 FedNed（Lu 等人，2024）、FedNed-（Lu 等人，2024）、FedNoRo（Wu 等人（2023b））、RoFL（Yang 等（2022））、FedCorr（Xu 等人（2022））和 FedAvg（McMahan等人，2017）。此外，我们还在联邦设置下比较了集中式学习中用于噪声标签学习的典型算法：协同教学（Han 等人，2018）和 DivideMix（Li 等人（2020））。基线方法的完整设置详情见附录 D.1。

5.2 性能比较

表1详细分析了不同对称噪声水平设置下，非独立同分布 CIFAR-10 数据集上的模型准确率。总体而言，FedCova 在所有噪声水平下都取得了最高的测试准确率。当噪声相对较低时，例如 $(\rho, \tau) = (0。4, 0。5)$，大多数基线方法仍具有鲁棒性，FedCorr 等方法仍能利用足够多的干净设备来指导训练。然而，一旦干净设备不再占主导地位（即 $\rho \geq 0.5$），对它们的检测就会变得不可靠，性能也会大幅下降；例如，FedCorr 的准确率从 $(\rho, \tau) = (0。4, 0。7)$ 时的 79.24% 降至 $(0。6, 0。5)$ 时的 69.47%，尽管噪声数据量相近。随着噪声进一步增加，FedCova 在更广泛的设置下仍保持强劲表现，凸显了其系统级鲁棒性。这一优势源于其基于协方差的特征对齐，能够产生稳定表示；而大多数基线方法依赖的交叉熵仍会拟合噪声标签。FedNed 通过负向蒸馏降低噪声影响，

表1：在不同对称噪声水平（ρ：噪声设备比例，τ：样本噪声比例）下，CIFAR-10（非独立同分布）上最后五轮的平均准确率和标准差。附加条件：DM：双模型；WU：预热通信轮次；CD：额外干净公共数据集；ED：极高噪声设备。

方法	测试准确率（%）± 标准差（%）						附加条件
	$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.8$		
	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.5$	$\tau = 0.7$	
FedAvg	72.43±4.63	65.16±8.36	64.75±1.46	52.87±3.68	47.98±5.42	22.27±5.04	-
CoteachingFL	78.63±0.58	73.58±0.78	73.34±0.67	65.00±0.72	57.41±0.81	38.86±0.58	DM
DivideMixFL	68.50±0.06	64.58±0.03	66.32±0.16	65.42±0.18	59.88±0.14	53.16±0.05	DM
RoFL	77.78±0.23	72.79±0.21	73.52±0.35	62.51±0.79	58.18±0.14	44.11±0.25	-
FedCorr	84.87±0.66	79.24±0.44	69.47±0.32	67.21±1.69	61.74±1.86	48.15±1.38	WU
FedNoRo	81.51±0.33	80.48±0.30	74.05±0.15	71.93±0.22	63.79±1.20	30.23±1.02	WU
FedNed-	72.90±1.40	64.30±1.01	70.15±1.29	65.71±1.45	52.33±1.82	39.63±1.88	CD
FedNed	82.38±0.13	82.02±0.56	78.16±1.23	77.11±0.45	64.58±1.85	48.98±1.39	CD, ED
FedCova	**86.52±0.23**	**85.50±0.38**	**83.78±0.68**	**80.71±0.56**	**67.21±0.92**	**64.99±0.75**	-

表2：在不同非对称噪声水平下，CIFAR-10（非独立同分布）最后五轮的平均准确率与标准差。

方法	测试准确率（%）± 标准差（%）					
	$\rho = 0.4$			$\rho = 0.6$		
	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$
FedAvg	72.06±4.35	71.23±5.86	67.40±11.2	68.91±5.02	63.53±2.89	43.40±18.2
RoFL	78.87±0.10	77.85±0.53	71.56±0.46	78.45±0.32	66.10±2.68	41.38±1.00
FedCorr	75.21±0.57	62.66±1.18	38.67±1.23	39.37±0.49	34.00±0.29	30.66±0.25
FedNoRo	86.40±0.72	85.72±0.41	84.76±0.62	75.27±0.21	63.88±19.9	33.42±37.8
FedNed-	75.69±0.98	70.38±1.40	65.51±3.48	78.63±0.37	71.22±1.15	66.44±1.44
FedNed	82.30±0.35	79.73±0.43	78.54±0.65	79.77±0.57	78.03±0.34	74.67±0.66
FedCova	**88.00±0.25**	**87.84±0.24**	**87.77±0.19**	**87.83±0.40**	**87.31±0.39**	**87.29±0.26**

在噪声设备较少时表现良好，在 $\rho = 0.4$ 时保持在 85% 以上，在 $\rho = 0.6$ 时约为 77–78%，但噪声设备比率较高时性能会下降。RoFL 得益于特征对齐，但其质心聚合只能捕获均值统计量，仍易受噪声影响。相比之下，FedCova 利用协方差结构，充分发挥了特征空间的判别能力。非对称噪声模式下的数值比较结果汇总于表2。在这种严重的非对称噪声下，FedCorr 和 FedNoRo 等基线方法对高噪声水平更为敏感。相比之下，FedCova 在噪声标签下展现出更强的鲁棒性。当噪声水平从 (0。4，0。3) 变化到 (0。6，0。7) 时，测试准确率仍能保持在 87–88% 左右。这主要归因于 FedCova 所形成的类别子空间正交结构。在非对称噪声下，混淆集中在较少的类别之间，尤其是两个类别之间，因此子空间主要只在这两个类别之间发生混合，这更有利于 FedCova 进行判别。

在类别更多且相互交错更复杂的 CIFAR-100 数据集上进行的实验，进一步加剧了噪声标签条件下联邦学习算法面临的挑战。结果如表3所示。随着类别数量增加，类别之间的判别边界变得更加复杂，难以学习。与前述分析类似，在不同噪声水平设置下，FedCova 仍优于基线方法。此外，我们还在带有噪声标签的真实数据集 Clothing1M 上评估了这些算法，如表4所示。与基线方法相比，FedCova 同样取得了最高准确率。

表1还列出了基线方法的附加条件，这些条件描述了会产生不可忽略资源消耗的额外依赖。FedCova 在不引入额外资源依赖的情况下取得了具有竞争力的性能，保持了无依赖设计。具体而言，FedCorr 和 FedNoRo 依赖本地预热训练，这需要数百次设备–服务器

表3：CIFAR-100（非独立同分布）在不同对称噪声水平下的平均准确率和标准差。

方法	测试准确率（%）± 标准差（%）					
	$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.8$	
	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.5$	$\tau = 0.7$
FedAvg	57.70±2.95	51.21±1.47	41.65±4.32	31.58±6.38	38.49±2.37	18.27±5.24
RoFL	50.71±0.57	48.80±0.16	45.58±0.17	37.89±0.28	29.40±0.62	5.20±0.03
FedCorr	59.66±0.93	55.60±0.45	53.81±0.60	50.01±0.98	49.68±0.46	29.04±0.61
FedNoRo	62.20±0.47	58.57±0.50	42.86±0.16	26.70±1.08	24.43±0.31	14.13±0.26
FedNed-	57.56±0.74	55.35±0.38	51.83±0.86	46.37±0.63	43.44±1.29	32.65±0.60
FedNed	64.67±0.40	61.53±2.09	56.63±0.45	54.84±0.60	49.51±0.95	36.04±0.33
FedCova	**67.63±1.04**	**66.84±0.73**	**60.67±1.11**	**58.46±1.58**	**55.71±0.82**	**40.80±0.78**

表4：Clothing1M（非独立同分布）上的平均准确率和标准差。

	FedAvg	RoFL	FedCorr	FedNoRo	FedNed	**FedCova**
准确率（%）	41.07±1.71	59.75±0.56	56.66±0.23	36.42±0.27	55.80±0.11	**61.42±0.65**

交互，因此通信开销远高于标准联邦训练。运行时间比较的相应分析请参见 D.5节。尽管如此，它们的消融研究表明，预热是最有效的组成部分。FedNed- 和 FedNed 依赖干净的公共数据集作为训练基准。此外，FedNed 还需要存在极高噪声设备，以识别不应学习的内容，这限制了它在某些噪声环境中的适用性。相比之下，我们的框架通过内在特征统计量在内部探索鲁棒性，无需理想的基准数据集或先验，因此适用于多种噪声场景。有关子空间正交性和校正性能的更多实验结果，请参见附录D。

5.3 消融实验

我们评估了 FedCova 在采用和不采用若干关键设计时的性能，具体配置如下：“FedCova w/o ext. corrector”，即不使用外部校正器进行标签校正的 FedCova；“FedCova w/o subs. augment.”，即将子空间增强系数设置为 $\alpha = 1$ 的 FedCova，此时该方法退化为基于高斯混合模型的最大后验分类器所使用的马氏距离；“FedCova w/o classifier align.”，即不采用协方差矩阵的联邦高斯消息合并，而是使用等权重进行聚合的 FedCova；“FedCova w/o error tolerance”，即将误差容忍系数设置为较小值的 FedCova，使特征子空间松弛失效；“FedCova w/o zero mean”，即移除零均值假设，并在聚合中引入均值统计量，作为均值感知分类器的 FedCova。

表5总结了非独立同分布CIFAR-10数据集在噪声水平（ρ, τ）$=(0.6, 0.7)$下的消融实验结果。所有组件均被验证有助于提升准确率。对

表5：FedCova 的关键组件在噪声水平（ρ，τ）=（0.6，0.7）下对 CIFAR-10（非独立同分布）的影响。

配置	准确率（%）
FedCova	80.71±0.56
不含外部校正器的 FedCova	77.88±1.47
FedCova 无子集增强。	79.58±0.60
FedCova 无分类器对齐。	76.86±0.30
FedCova 无误差容忍	69.47±0.51
FedCova（无零均值）	76.93±0.34

特征协方差的关注是FedCova的关键优势，这一点由“FedCova w/o 零均值”和“FedCova w/o 子集增强”的结果所证明。“FedCova w/o 零均值”尤其表明，噪声标签引入的均值偏差会导致性能下降。“FedCova w/o 误差容忍”表现出最大的性能差距，这可能是因为我们的互信息最大化目标所强制的严格正交性在不同本地训练过程中可能有所变化。引入误差容忍单位矩阵既能将特征子空间球面化以抵抗噪声标签，又能在各设备间对齐表示的正交性。同样，跨设备对齐表征的重要性在“FedCova w/o 分类器对齐”中得到进一步证明。“FedCova w/o 外部校正器”凸显了我们校正器设计的有效性以及基于协方差的分类器的判别能力。关于若干超参数影响的进一步补充实验与分析见附录 E。

6 结论

我们提出了 FedCova，一种协方差感知的联邦学习框架，专门用于应对分布式数据集中噪声标签所带来的挑战。通过利用协方差所捕获的特征表示的鲁棒性，FedCova 将特征编码、内在分类器构建和噪声标签校正无缝整合到统一框架中。互信息最大化目标约束了容错的协方差结构，使模型能够学习具有判别性的正交特征子空间，从而增强对标签噪声的韧性。通过协方差聚合，FedCova 有效地在各边缘设备之间对齐联邦分类器。通过引入子空间增强，基于内在最大后验判别推导出广义分类器。此外，本地外部校正器在缓解自偏差的同时促进有效的重新标注。在多个数据集上的大量实验表明，FedCova 在不同标签噪声水平及各类噪声模式下均持续优于最先进的框架，凸显了其鲁棒性。值得注意的是，FedCova 无需现有工作中常见的额外条件，建立了一种由特征协方差下固有鲁棒性驱动的无依赖解决方案。

参考文献

Devansh Arpit、Stanisław Jastrz.ebski、Nicolas Ballas、David Krueger、Emmanuel Bengio、Maxinder S Kanwal、Tegan Maharaj、Asja Fischer、Aaron Courville、Yoshua Bengio 等。深入审视深度网络中的记忆化。载于国际机器学习会议，第 233–242 页。PMLR，2017年。

Anthony J Bell 和 Terrence J Sejnowski。用于盲分离和盲反卷积的信息最大化方法。神经计算，7(6):1129–1159，1995年。

Yoshua Bengio、Aaron Courville 和 Pascal Vincent。表示学习：综述与新视角。*IEEE*模式分析与机器智能汇刊，35(8):1798–1828，2013。

Daniel Berrar 等。交叉验证。，2019。

Duc Bui、Kshitiz Malik、Jack Goetz、Honglei Liu、Seungwhan Moon、Anuj Kumar 和 Kang G Shin。联邦用户表示学习。*arXiv* 预印本 *arXiv:1909.12535*，2019。

Kwan Ho Ryan Chan、Yaodong Yu、Chong You、Haozhi Qi、John Wright 和 Yi Ma。Redunet：基于最大化速率缩减原理的白盒深度网络。机器学习研究期刊，23(114):1–103，2022。

Jiangjie Chen、Xintao Wang、Rui Xu、Siyu Yuan、Yikai Zhang、Wei Shi、Jian Xie、Shuang Li、Ruihan Yang、Tinghui Zhu 等。 从角色人格到个性化：角色扮演语言智能体综述。*arXiv*预印本 *arXiv:2404.18231*，2024。

Pengfei Chen、Ben Ben Liao、Guangyong Chen 和 Shengyu Zhang。理解并利用使用噪声标签训练的深度神经网络。发表于国际机器学习会议，第1062–1070页。PMLR，2019。

De Cheng、Tongliang Liu、Yixiong Ning、Nannan Wang、Bo Han、Gang Niu、Xinbo Gao 和 Masashi Sugiyama。基于流形正则化转移矩阵估计的实例依赖型标签噪声学习。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第16630–16639页，2022。

Zonglin Di、Zhaowei Zhu、Xiaoxiao Li 和 Yang Liu。在欧洲计算机视觉会议上发表《带有本地开放集噪声标签的联邦学习》，第 38–56 页。Springer，2024。

Shaoming Duan、Chuanyi Liu、Zhengsheng Cao、Xiaopeng Jin 和 Peiyi Han。Fed-dr-filter：利用全局数据表示降低噪声标签对联邦学习性能的影响。《未来一代计算机系统》，137：336–348，2022。ISSN 0167-739X。doi：https://doi.org/10.1016/j.future.2022.07.013。

Erik Englesson 和 Hossein Azizpour。用于噪声标签学习的广义詹森-香农散度损失。《神经信息处理系统进展》，34：30284–30297，2021。

Bo Han、Quanming Yao、Xingrui Yu、Gang Niu、Miao Xu、Weihua Hu、Ivor Tsang 和 Masashi Sugiyama。协同教学：使用极高噪声标签进行深度神经网络的鲁棒训练。神经信息处理系统进展，31，2018。

Bo Han、Quanming Yao、Tongliang Liu、Gang Niu、Ivor W Tsang、James T Kwok 和 Masashi Sugiyama。标签噪声表征学习综述：过去、现在与未来。*arXiv*预印本 *arXiv:2011.04406*，2020年。

Trevor Hastie 和 Robert Tibshirani。基于高斯混合的判别分析。皇家统计学会学报*B*辑：统计方法学，58(1):155–176，1996年。Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian 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。

Xinyuan Ji、Zhaowei Zhu、Wei Xi、Olga Gadyatskaya、Zilong Song、Yong Cai 和 Yang Liu。《Fedfixer：缓解联邦学习中的异质标签噪声》。载于 *AAAI*人工智能会议论文集，第 38 卷，第 12830–12838 页，2024 年。

Peter Kairouz、H Brendan McMahan、Brendan Avent、Aurélien Bellet、Mehdi Bennis、Arjun Nitin Bhagoji、Kallista Bonawitz、Zachary Charles、Graham Cormode、Rachel Cummings 等。《联邦学习的进展与开放问题》。机器学习基础与趋势®，14(1–2)：1–210，2021 年。

Sai Praneeth Karimireddy、Satyen Kale、Mehryar Mohri、Sashank Reddi、Sebastian Stich 和 Ananda Theertha Suresh。SCAFFOLD：用于联邦学习的随机控制平均。载于第*37*届国际机器学习会议论文集，第5132–5143页。PMLR，2020年11月。

Alex Krizhevsky、Geoffrey Hinton 等。基于微型图像学习多层特征。2009年。

Jichang Li、Guanbin Li、Feng Liu 和 Yizhou Yu。用于噪声标签识别与纠正的邻域集体估计。载于欧洲计算机视觉会议，第128–145页。Springer，2022年。

Jichang Li、Guanbin Li、Hui Cheng、Zicheng Liao 和 Yizhou Yu。FedDiv：面向含噪声标签联邦学习的协同噪声过滤。载于*AAAI*人工智能会议论文集，第38卷，第3118–3126页，2024a。

Junnan Li、Richard Socher和Steven C.H. Hoi。Dividemix：将噪声标签学习作为半监督学习。载于国际学习表征会议，2020。

Junnan Li、Caiming Xiong和Steven CH Hoi。基于鲁棒表征学习从噪声数据中学习。载于*IEEE/CVF*国际计算机视觉会议论文集，第9485–9494页，2021。

Xiang Li、Kaixuan Huang、Wenhao Yang、Shusen Wang和Zhihua Zhang。FedAvg在非独立同分布数据上的收敛性。*arXiv*预印本 *arXiv:1907.02189*，2019。

Zijian Li、Zehong Lin、Jiawei Shao、Yuyi Mao和Jun Zhang。FedCIR：面向联邦非独立同分布特征的客户端不变表征学习。*IEEE*移动计算汇刊，第1–14页，2024b。ISSN 1558-0660。doi: 10.1109/TMC.2024.3376697。

Yang Lu、Lin Chen、Yonggang Zhang、Yiliang Zhang、Bo Han、Yiu-ming Cheung和Hanzi Wang。通过负向蒸馏实现面向极高噪声客户端的联邦学习。载于*AAAI*人工智能会议论文集，第38卷，第14184–14192页，2024年。

Connor Mclaughlin和Lili Su。通过特征分布适配的个性化联邦学习。神经信息处理系统进展，37:77038–77059，2024年。

Brendan McMahan、Eider Moore、Daniel Ramage、Seth Hampson和Blaise Aguera y Arcas。基于去中心化数据的深度网络通信高效学习。载于 人工智能与统计学，第1273–1282页。PMLR，2017年。

Curtis Northcutt、Lu Jiang和Isaac Chuang。置信学习：估计数据集标签中的不确定性。人工智能研究杂志，70:1373–1411，2021年4月。ISSN 1076-9757。doi: 10.1613/jair.1.12125。440次引用（Semantic Scholar/DOI）[2024-02-06]。

Pretom Roy Ovi、Aryya Gangopadhyay、Robert F. Erbacher 和 Carl Busart。置信联邦学习应对标签翻转数据投毒攻击。载于 Latasha Solomon 和 Peter J. Schwartz（编），多域作战应用中的人工智能与机器学习 *V*，第 32 页，美国奥兰多，2023 年 6 月。国际光学工程学会。ISBN 978-1-5106-6192-9。doi: 10.1117/12. 2663911。0 次引用（Semantic Scholar/DOI）[2024-02-06]。

Mengye Ren、Wenyuan Zeng、Bin Yang 和 Raquel Urtasun。学习对示例进行重加权，以实现鲁棒深度学习。载于国际机器学习会议，第 4334–4343 页。PMLR，2018 年。

Hwanjun Song 等。利用深度神经网络从噪声标签中学习：综述。*IEEE* 神经网络与学习系统汇刊，2022 年。

Kahou Tam、Li Li、Yan Zhao 和 Chengzhong Xu。Fedcoop：面向噪声标签的协作式联邦学习。载于 欧洲人工智能会议 *2023*，第 2298–2306 页。IOS 出版社，2023。

Alysa Ziying Tan、Han Yu、Lizhen Cui 和 Qiang Yang。迈向个性化联邦学习。*IEEE 神经网络与学习系统汇刊*，34(12)：9587–9603，2023。doi：10.1109/TNNLS.2022.3160699。Yue Tan、Guodong Long、Lu Liu、Tianyi Zhou、Qinghua Lu、Jing Jiang 和 Chengqi Zhang。Fedproto：面向异构客户端的联邦原型学习。载于 *AAAI人工智能会议论文集*，第 36 卷，第 8432–8440 页，2022。Daiki Tanaka、Daiki Ikami、Toshihiko Yamasaki 和 Kiyoharu Aizawa。用于噪声标签学习的联合优化框架。载于 *IEEE计算机视觉与模式识别会议论文集*，第 5552–5560 页，2018。Naftali Tishby 和 Noga Zaslavsky。深度学习与信息瓶颈原理。载于 *2015 年 IEEE 信息理论研讨会 (ITW)*，第 1–5 页。IEEE，2015。Naftali Tishby、Fernando C Pereira 和 William Bialek。信息瓶颈方法。*arXiv 预印本 physics/0004057*，2000。Vasileios Tsouvalas、Aaqib Saeed、Tanir Ozcelebi 和 Nirvana Meratnia。将标签混乱化为学习和谐：带噪声标签的联邦学习。*ACM智能系统与技术汇刊*，15(2)：1–26，2024。

Chenrui Wu、Zexi Li、Fangxin Wang 和 Chao Wu。在噪声标签和异构客户端环境下谨慎地进行联邦学习。载于*2023 IEEE多媒体与博览会国际会议（ICME）*，第 660–665 页。IEEE，2023a。

Nannan Wu、Li Yu、Xuefeng Jiang、Kwang-Ting Cheng 和 Zengqiang Yan。Fednoro：通过解决类别不平衡和标签噪声异质性，实现抗噪声联邦学习。*arXiv 预印本 arXiv:2305.05230*，2023b。Yichen Wu、Jun Shu、Qi Xie、Qian Zhao 和 Deyu Meng。通过元软标签校正器学习净化噪声标签。*AAAI人工智能会议论文集*，35(12)：10388–10396，2021年5月。ISSN 2374-3468，2159-5399。doi：10.1609/aaai.v35i12.17244。

Xiaobo Xia、Pengqian Lu、Chen Gong、Bo Han、Jun Yu、Jun Yu 和 Tongliang Liu。《正则截断m估计量：使用噪声标签进行学习》。《*IEEE模式分析与机器智能汇刊*》，第 1–15 页，2024 年。ISSN 0162-8828、2160-9292、1939-3539。doi：10.1109/ TPAMI.2023.3347850。

Tong Xiao、Tian Xia、Yi Yang、Chang Huang 和 Xiaogang Wang。利用大规模噪声标注数据进行图像分类学习。发表于 *CVPR*，2015。Jian Xu、Xinyi Tong 和 Shao-Lun Huang。通过特征对齐和分类器协作实现个性化联邦学习。发表于 第十一届学习表征国际会议。Jingyi Xu、Zihan Chen、Tony QS Quek 和 Kai Fong Ernest Chong。Fedcorr：用于标签噪声校正的多阶段联邦学习。发表于 *IEEE/CVF计算机视觉与模式识别会议论文集*，第 10184–10193 页，2022。Seunghan Yang、Hyoungseob Park、Junyoung Byun 和 Changick Kim。带噪声标签的鲁棒联邦学习。*IEEE智能系统*，37(2)：35–43，2022。Kun Yi 和 Jianxin Wu。用于噪声标签学习的概率端到端噪声校正。发表于 *IEEE/CVF计算机视觉与模式识别会议论文集*，第 7017–7025 页，2019。Li Yi、Sheng Liu、Qi She、A Ian McLeod 和 Boyu Wang。关于噪声标签学习中的对比表示。发表于 *IEEE/CVF计算机视觉与模式识别会议论文集*，第 16682–16691 页，2022。Fuxun Yu、Weishan Zhang、Zhuwei Qin、Zirui Xu、Di Wang、Chenchen Liu、Zhi Tian 和 Xiang Chen。Fed2：特征对齐联邦学习。发表于 第 *27* 届 *ACM SIGKDD* 知识发现与数据挖掘会议论文集，第 2066–2074 页，2021。

Yaodong Yu、Kwan Ho Ryan Chan、Chong You、Chaobing Song 和 Yi Ma。通过最大编码率降低原理学习多样且具有判别性的表示。神经信息处理系统进展，33：9422–9434，2020。

Fengda Zhang、Kun Kuang、Long Chen、Zhaoyang You、Tao Shen、Jun Xiao、Yin Zhang、Chao Wu、Fei Wu、Yueting Zhuang 等。联邦无监督表征学习。《信息技术与电子工程前沿》，24(8):1181–1193，2023。

Lin Zhang、Li Shen、Liang Ding、Dacheng Tao 和 Ling-Yu Duan。通过无数据知识蒸馏微调全局模型，以用于非独立同分布联邦学习。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 10174–10183 页，2022 年。

PengFei Zheng、Yonggang Zhang、Zhen Fang、Tongliang Liu、Defu Lian 和 Bo Han。《Noisediffusion：利用扩散模型校正图像插值中的噪声，超越球面线性插值》。载于第十二届国际学习表征会议，2024年。

A 相关工作

A.1 噪声标签下的联邦学习

随着训练数据规模迅速扩大，噪声标签问题日益突出，其来源包括标注错误、传感器故障或对抗攻击。研究者首先关注带噪声标签的集中式学习 Northcutt 等人（2021）；Chen 等（2019）。主流解决方案大致可分为两类。一类通过引入噪声检测、样本选择和标签校正等离线技术来改进训练过程，旨在为给定的训练网络设计抗噪算法 Cheng 等人（2022）；Ren 等人（2018）；Englesson 和 Azizpour（2021）；Xia 等人（2024）。JointOpt Tanaka 等人（2018）是一个经典范例，它引入联合优化框架，在训练过程中校正噪声样本的标签。具体而言，该方法交替更新模型参数并优化标签。另一类则通过引入额外的资源条件来扩展学习结构，例如假设可以访问干净公开数据集、复制网络以进行交叉验证，或采用无监督学习和元学习等先进框架。Chen 等（2019）；Li 等人（2022）；Wu 等人（2021）；Zheng 等人（2024）。协同教学 Han 等人（2018）是一种代表性框架，它同时训练两个网络，每个网络都会选取少量干净样本来更新另一个网络。协同教学依靠对等网络筛除噪声标签，有效降低了标签噪声的影响并提升鲁棒性。DivideMix Li 等人（2020）进一步推进了这一方法：它将带有标签噪声的训练数据集动态划分为干净子集和噪声子集，同时结合辅助半监督学习算法。

在联邦学习（FL）中面对分布式数据和隐私约束时，服务器无法直接访问本地数据样本，这限制了为集中式学习开发的鲁棒技术在联邦学习中的适用性和有效性。因此，研究人员专注于开发适用于联邦范式的方法。利用联邦学习中设备的原生划分特性，可以实现样本选择和噪声过滤方法（Wu 等人（2023a））。例如，FedCorr（Xu 等人（2022））提出了一种基于维度的噪声过滤策略，利用局部内在维度度量将客户端划分为干净组和噪声组，并进一步训练本地过滤器，通过逐样本训练损失分析，从噪声客户端中区分干净样本。类似地，FedDiv（Li 等人（2024a））通过汇总来自客户端的共享过滤知识来增强噪声过滤，从而改进对各客户端数据集中标签噪声的检测。然而，这些方法仍受限于对输出与标签关系的依赖，因为其学习目标（如交叉熵损失）仍会使模型与可能错误的标签绑定。近期研究尝试通过引入新的噪声标签检测或校正方案来克服这些局限。FedFixer（Ji 等人（2024））为每个设备配置两个复制的本地模型，通过交叉学习实现协同缓解标签噪声并减少过拟合。FedNed（Lu 等人（2024））采用来自极高噪声设备的负向蒸馏，借助基于 Kullback-Leibler（KL）散度设计的新损失函数，引导模型远离错误方向。FedNoRo（Wu 等人（2023b））和标签噪声感知蒸馏（Tsouvalas 等人（2024））也提出了类似的损失函数重新设计，以缓解上述目标函数中的过拟合问题。尽管这些方法有效，但它们需要额外条件，例如额外的重复模型或公共干净数据集，并会增加计算开销，这可能限制其在复杂标签噪声场景中的可扩展性。相比之下，我们的框架首先通过重新设计损失函数来解决传统的输出与标签绑定问题，并进一步消除对补充条件的依赖，从而从根本上引导学习模型实现抵御噪声标签的系统级鲁棒性。

A.2 联邦特征学习

学习和理解特征对于揭示底层数据结构并提升模型泛化至关重要。传统的表示学习强调获取判别性且信息丰富、能够捕捉数据内在结构的特征表示的重要性 Bengio 等人（2013）；Han 等人（2020）。通过将原始数据变换到潜在模式与关系更为明显的特征空间，学习此类特征使模型能够聚焦于数据的统计特性，从而取得有竞争力的性能 Yu 等人（2020）；Li 等人（2024b）；Yu 等人（2021）。这开辟了增强鲁棒性以应对

噪声标签，因为促进能够真正理解输入数据的特征学习，会使模型较少直接受到错误标签的影响 Arpit 等人（2017）；Zhang 等人（2023）。基于这一见解，对比学习等方法 Li 等人（2021）；Yi 等人（2022）在减轻集中式学习框架内噪声标签的不利影响方面已取得相当显著的成功。通过在样本之间的成对损失中惩罚特征相似性，对比学习促使模型捕捉类判别表示，从而增强对标签噪声的鲁棒性。

然而，将这些方法直接应用于联邦学习（FL）颇具挑战，因为数据仍保留在各个设备上，且不允许直接访问不同设备上的样本对。因此，在 FL 场景中，需要对所有样本对进行对比的对比学习等方法往往表现不佳 Zhang 等人（2023）；Duan 等人（2022）。在 FL 中，信息丰富的特征学习主要从联邦表示学习和基于原型的方法的角度进行探索。例如，pFedFDA Mclaughlin & Su（2024）将表示学习建模为生成建模任务，提出一种个性化 FL 新方法，从而能够将全局分类器高效地适配到局部特征分布。FURL Bui 等人（2019）通过在本地训练私有用户嵌入，在 FL 中引入了一种保护隐私且资源高效的方法。为实现局部表示之间的全局一致性，一些工作（如 FedProto Tan 等人（2022）和 FedPAC Xu 等人）聚合学习所得特征的局部均值（质心），并在损失函数中引入正则项以进行对齐。尽管这些方法彰显了 FL 中特征学习的潜力，但它们主要针对数据异质性，通常并未明确解决噪声标签这一挑战。此外，由于交叉熵损失固有的输出与标签依赖关系，这些方法仍易于过拟合。另外，在噪声标签下估计的特征均值可能存在显著偏差，导致噪声检测不准确，并进一步加剧过拟合。

为通过利用特征来应对噪声标签，FedCoop Tam等人（2023）提出了一种协作式联邦学习框架，利用客户端间协作和基于特征的质量评估进行半监督训练，从而有效减轻噪声标签的影响。然而，该方法主要依靠干净子集过滤噪声样本，而非利用特征空间的内在结构。Fed-DR-Filter Duan等人（2022）在联邦学习中提出了一种新颖的过滤方法，根据全局数据表示的相关性识别干净数据。然而，将客户端的全部特征直接传输至服务器会导致过高的通信开销并引发隐私问题，同时服务器还需承担噪声过滤所需的大量计算成本。RoFL Yang 等（2022）通过对齐服务器与客户端之间的类别特征质心（均值），缓解联邦学习中的标签噪声，从而促进各本地模型的决策边界保持一致。该方法还结合样本选择与标签校正，以增强对噪声标注的鲁棒性。然而，如前所述，依赖均值统计量进行对齐会使该方法易受标签噪声影响，尤其是在使用交叉熵损失时，因为被错误标注的数据可能会严重偏置均值特征。相比之下，我们的框架侧重于协方差统计，并设计了一个涵盖从特征编码到判别分类和标签校正的统一系统，从而消除上述限制并增强鲁棒性。

B 理解有损学习目标

本节将进一步阐述第3节中有损学习目标的设计。

B.1 互信息最大化

我们在式（8）中提出的目标函数，其核心思想是最大化学习到的表示 Z 与其对应标签 Y 之间的互信息。下面我们将分步说明。

互信息用于量化已知一个随机变量后，另一个随机变量的不确定性降低了多少。形式上，对于随机变量 Z 和 Y，互信息定义如下：

$$I(Z;Y) = h(Z) - h(Z|Y), \tag{15}$$

其中，$h(Z)$是 Z的微分熵，而 $h(Z|Y)$是给定 Y时 Z 的条件熵。换言之，$h(Z)$衡量 Z中的总不确定性（随机性），$h(Z|Y)$衡量其中的不确定性

Z 在已知 Y之后的情况，而它们的差值 $I(Z;Y)$则体现了由于掌握 Y的信息而减少的关于 Z 的不确定性。

利用这种互信息驱动模型训练的方法被形式化为信息瓶颈 (IB) 原则 Tishby et al. (2000); Tishby & Zaslavsky (2015)。IB 框架提出学习输入 X 的表示 Z，尽可能压缩 X，同时仍保留关于目标 Y的信息。这种权衡通常表示为最大化 $I(Z;Y) - \beta I(Z;X)$，其中 $\beta \geq 0$ 用于平衡充分性（对 Y的信息量）与最小性（对 X的压缩）。由于我们的目标是在联邦特征学习框架中学习类别判别特征，因此式 (4) 中的目标是最大化学习到的表示 Z 与标签 Y之间的互信息，确保特征尽可能多地捕获与目标相关的信息。换言之，我们希望 Z 能够高度准确地预测 Y。通过最大化 $I(Z;Y)$，我们确保表示 Z 保留输入 X 中对预测 Y最有用的信息，同时可能舍弃无关变化。在我们的情形中，我们只关注最大化 $I(Z;Y)$；通过令 $\beta = 0$，这可视为 IB 原则的一个特例。也就是说，我们优先学习能够最大程度保留关于 Y 的信息的表示。具体而言，我们力求在编码后的特征 Z中保留来自 X 的、与标签最相关的信息，确保 Z 包含足够的信息，以便进行准确分类。

假设 Y是一个离散类别标签，在 $\{1, \ldots, J\}$ 中取值，概率为 $\pi_j = P(Y = j)$。基于式（3）中定义的高斯混合先验，对于每个类别 j，表示的条件分布由下式给出

$$Z \mid Y = j \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_j).$$

条件熵 $h(Z|Y)$量化了在已知 Y的情况下 Z 的期望不确定性。由于对于每个 j， $Z|Y = j$ 服从多元高斯分布，因此d维高斯 $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}_j)$的微分熵为

$$h(Z|Y = j) = \frac{1}{2} \log \left[(2\pi e)^d \det(\mathbf{\Sigma}_j) \right]. \tag{16}$$

值得注意的是，熵仅依赖于协方差矩阵 $\mathbf{\Sigma}_j$，而不依赖于均值 $\boldsymbol{\mu}$。因此，给定 Y时 Z 的条件熵是对所有类别取期望，即

$$\begin{aligned} h(Z|Y) &= \sum_{j=1}^{J} \pi_j \, h(Z|Y = j) = \frac{1}{2} \sum_{j=1}^{J} \pi_j \log \left[(2\pi e)^d \det(\mathbf{\Sigma}_j) \right] \\ &= \frac{1}{2} \sum_{j=1}^{J} \pi_j \log \det(\mathbf{\Sigma}_j) + \frac{d}{2} \log(2\pi e). \end{aligned} \tag{17}$$

因此， Z 的边缘分布成为一个高斯混合分布。为便于解析处理，我们通常将 $p(z)$近似为均值为零、协方差为 $\mathbf{\Sigma}$的单一高斯分布。因此，边缘熵 $h(Z)$可计算为

$$h(Z) = \frac{1}{2} \log \left[(2\pi e)^d \det(\mathbf{\Sigma}) \right] = \frac{1}{2} \log \det(\mathbf{\Sigma}) + \frac{d}{2} \log(2\pi e). \tag{18}$$

将式（17）和式（18）中的熵公式代入互信息式（15），我们得到

$$I(Z;Y) = h(Z) - h(Z|Y) = \frac{1}{2} \log \det(\mathbf{\Sigma}) - \frac{1}{2} \sum_{j=1}^{J} \pi_j \log \det(\mathbf{\Sigma}_j). \tag{19}$$

在实践中，我们会像式（5）中的目标函数那样，将其取负后进行最小化。由于高斯变量的熵仅取决于其协方差，而与其均值无关，因此我们的目标函数与特征均值无关，我们可以只关注协方差矩阵。

B.2 为什么 $I(Z;\ Y)$ 有效：与交叉熵损失的关系

或者，我们也可以将特征 Z 与标签 Y之间的互信息重写为

$$I(Z;Y) = H(Y) - H(Y|Z) = H(Y) + \mathbb{E}_{p(X,Y)} \mathbb{E}_{p(Z|X)} \log p(Y|Z). \tag{20}$$

由于我们最大化互信息 $I(Z;Y)$，在给定常数 $H(Y)$ 的情况下，实际上是在以下损失函数下进行训练

$$\mathcal{L}_{MI} = -\mathbb{E}_{p(X,Y)}\mathbb{E}_{p(Z|X)} \log p(Y|Z). \tag{21}$$

在交叉熵损失等常规目标下进行训练时，真实条件分布$p(Y|Z)$通常未知，因此被模型参数化的预测分布$\hat{p}(Y|Z)$所替代（例如神经网络分类器）。从交叉熵的角度来看，一般目标是模型下的期望负对数似然，由下式给出

$$\mathcal{L}_{\mathrm{CE}} = H(p(Y|X), \hat{p}(Y|X)) = -\mathbb{E}_{p(X,Y)} \log \hat{p}(Y|X), \tag{22}$$

其中 $H(p(Y|X), \hat{p}(Y|X))$表示真实条件分布$p(Y|X)$与模型预测$\hat{p}(Y|X)$之间的交叉熵。根据交叉熵关于熵与KL散度的标准分解，我们有

$$H(p(Y|X), \hat{p}(Y|X)) = H(p(Y|X)) + D_{\mathrm{KL}}(p(Y|X)\|\hat{p}(Y|X)) \tag{23}$$

$$= -\mathbb{E}_{p(X,Y)} \log p(Y|X) + \mathbb{E}_{p(X,Y)} \log \frac{p(Y|X)}{\hat{p}(Y|X)}, \tag{24}$$

其中 $D_{\mathrm{KL}}(p(Y|X)\|\hat{p}(Y|X))$是真实条件分布与预测条件分布之间的KL散度，其值始终非负。因此，该分解意味着

$$-\mathbb{E}_{p(X,Y)} \log \hat{p}(Y|X) \geq -\mathbb{E}_{p(X,Y)} \log p(Y|X). \tag{25}$$

考虑预测分布以潜变量 Z 为条件的情形，该潜变量是从 X中提取的中间特征表示，因此上述结果自然推广为

$$-\mathbb{E}_{p(X,Y)} \log \mathbb{E}_{p(Z|X)}\hat{p}(Y|Z) \geq -\mathbb{E}_{p(X,Y)} \log \mathbb{E}_{p(Z|X)}p(Y|Z). \tag{26}$$

在模型的预测分布 $\mathcal{L}_{\mathrm{CE}}$ 下，交叉熵损失始终不小于真实条件分布 p的熵 $Y|Z$。

此外，由于对数函数具有凹性，根据詹森不等式，对于任意固定的（X， Y），有：

$$\mathbb{E}_{p(Z|X)} \log p(Y|Z) \leq \log \mathbb{E}_{p(Z|X)}p(Y|Z). \tag{27}$$

取负并对数据分布求平均后，我们得到：

$$-\mathbb{E}_{p(X,Y)}\mathbb{E}_{p(Z|X)} \log p(Y|Z) \geq -\mathbb{E}_{p(X,Y)} \log \mathbb{E}_{p(Z|X)}p(Y|Z). \tag{28}$$

因此，基于互信息的目标 $\mathcal{L}_{\mathrm{MI}}$ 也是真实条件分布 p的熵 $Y|Z$的上界，与交叉熵损失一样。最小化 $\mathcal{L}_{\mathrm{CE}}$ 或$\mathcal{L}_{\mathrm{MI}}$ 都能从不同角度推动模型逼近真实熵。交叉熵损失侧重于直接改善给定 Z时 Y的预测分布，而互信息目标则强调所学习的特征表示 Z 对于 Y所包含的信息量。因此，将 $\mathcal{L}_{\mathrm{MI}}$ 用作目标是合理的，因为它能鼓励模型提取关于目标信息量最大的特征，这有助于下游预测任务。

B.3 有损表示的解释

式 (5) 中给出的目标函数经过精心设计，用于正则化特征空间的结构，从而通过各类别的主方向增强类别可分性。我们将特征归一化为 $\|\boldsymbol{z}\|_2 = 1$，以关注不同表示之间的方向，使所有特征点都位于特征空间中单位球面的表面上。项 $\log\det(\boldsymbol{\Sigma})$ 反映了特征分布所张成椭球的体积，该体积与其主方差的乘积成正比。通过最小化类别特定协方差矩阵的加权对数行列式之和，即 $\sum_{j=1}^{J} \frac{B_j}{2B} \log\det\left(\frac{1}{B_j}\boldsymbol{Z}_{m,j}\boldsymbol{Z}_{m,j}^{*} + \frac{\epsilon^2}{d}\boldsymbol{I}\right)$，目标函数促使每个类别内的特征分布尽可能紧凑，从而降低类内方差。相反，最大化总体协方差矩阵的对数行列式，即 $\frac{1}{2}\log\det\left(\frac{1}{B}\boldsymbol{Z}_m\boldsymbol{Z}_m^{*} + \frac{\epsilon^2}{d}\boldsymbol{I}\right)$，则促使全局特征分布沿尽可能多的正交方向扩展，从而增大类间方差。因此，该目标函数会隐式地驱使不同类别的主方向彼此正交，因为

这种配置在保持每个类别自身紧凑的同时，使所有类别分布并集所覆盖的总体积最大化。

该正则化具有清晰的信息论解释：它在降低每个类别内部的不确定性与在综合考虑所有类别时增加总体不确定性之间取得平衡。因此，所学特征空间的结构使得来自不同类别的样本沿彼此不同且近乎正交的方向分布，从而显著增强类别之间的可分性。这种结构能够基于特征实现高效的内在分类，因为每个类别在特征空间中占据一个独特且最大程度分离的子空间。因此，式 (5) 中的损失具有理论基础，能够同时促进类内紧致性与类间正交性，从而提升分类性能和特征可解释性。

尽管式 (5) 中的目标对每个类别内的特征结构进行正则化，使其沿特征空间的主方向彼此区分，但它仍然受到输出特征 $\boldsymbol{z}$ 与标签 y（其可能是潜在的噪声标签 y'）之间强依赖关系的影响。当编码特征 $\boldsymbol{z}$ 被赋予错误标签 y'时，它将不可避免地被纳入 $\boldsymbol{\Sigma}_{y'}$的计算之中。为缓解这种依赖，我们通过ˆ在估计协方差中引入误差容忍项（记为 $\boldsymbol{\Sigma}$）来放宽特征子空间的判别正交性，从而得到如式 (8) 所示的变体损失函数。从数学上讲，对于批次协方差 $\frac{1}{B}\boldsymbol{Z}_m\boldsymbol{Z}_m^*$，我们可以将其谱分解表示为

$$\frac{1}{B}\boldsymbol{Z}_m\boldsymbol{Z}_m^* = \boldsymbol{U}_m\boldsymbol{\Lambda}_m\boldsymbol{U}_m^*, \tag{29}$$

其中，$\boldsymbol{U}_m$ 是特征向量矩阵，$\boldsymbol{\Lambda}_m = \text{diag}(\lambda_{m,1}, \ldots, \lambda_{m,d})$包含协方差的特征值。引入误差容忍项 $\frac{\epsilon^2}{d}\boldsymbol{I}$后，估计协方差可分解为

$$\hat{\boldsymbol{\Sigma}}_m = \frac{1}{B}\boldsymbol{Z}_m\boldsymbol{Z}_m^* + \frac{\epsilon^2}{d}\boldsymbol{I} = \boldsymbol{U}_m\left(\boldsymbol{\Lambda}_m + \frac{\epsilon^2}{d}\boldsymbol{I}\right)\boldsymbol{U}_m^*, \tag{30}$$

这意味着每个特征值都扩展为

$$\hat{\lambda}_{m,p} = \lambda_{m,p} + \frac{\epsilon^2}{d}. \tag{31}$$

对特征值进行这种统一平移，实际上平滑了协方差矩阵的谱，降低了任何单个主方向的主导程度。通过引入这种各向同性误差容忍项，类别特定特征子空间之间严格的正交约束得以放宽，进而减轻离群点或错误标记样本的影响，并防止损失函数对标签噪声过度敏感。

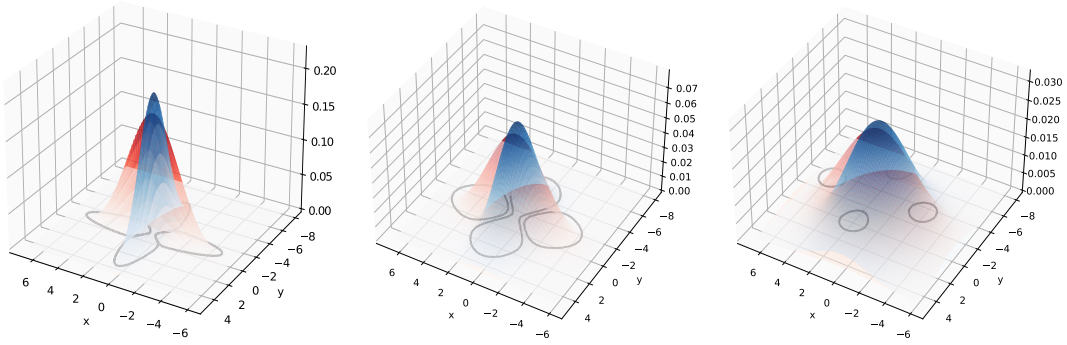
误差容忍度影响的玩具示例。 这里我们通过一个简单的玩具示例，进一步阐明误差容忍项对特征分布及由此产生的决策边界的影响。具体而言，我们考虑两个类别，其特征分布服从二维零均值高斯分布，即 $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_A)$和 $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_B)$，其协方差矩阵为

$$\boldsymbol{\Sigma}_A = \begin{bmatrix} 4.0 & 0 \\ 0 & 0.2 \end{bmatrix}, \quad \boldsymbol{\Sigma}_B = \begin{bmatrix} 0.2 & 0 \\ 0 & 3.0 \end{bmatrix}.$$

这些设置表示特征空间中的两个类别，分别沿相互正交的方向拉长。图 2 展示了误差容限参数 ϵ 取三个不同值时两个类别的概率密度曲面，其中正则化协方差分别为 $\boldsymbol{\Sigma}_A^{(\epsilon)} = \boldsymbol{\Sigma}_A + \frac{\epsilon^2}{d}\boldsymbol{I}$ 和 $\boldsymbol{\Sigma}_B^{(\epsilon)} = \boldsymbol{\Sigma}_B + \frac{\epsilon^2}{d}\boldsymbol{I}$，且 $d = 2$。

根据图 2 中的示意图，我们可以观察到：

- **当 ϵ^2 （无误差容忍项）：** 两个分布具有很强的各向异性，分别沿 x 轴和 y 轴拉长。决策边界（黑色轮廓）在两个分布的交汇处界定清晰，并且对原始协方差的主方向高度敏感。
- **随着 ϵ^2 增大：** 各向同性项开始平滑两个分布之间的差异。形状变得更加圆润，峰值降低，决策边界也明显变得更加平滑，不再紧密贴合原始分布的交集，从而降低了对带有标签噪声的离群点的敏感性。



(a) $\epsilon^2 = 0$ (b) $\epsilon^2 = 2$ (c) $\epsilon^2 = 8$

Fig图 2：向协方差矩阵中添加误差容忍项对概率分布的影响 s.

- **当 ϵ^2较大时：** 两个分布都趋近于各向同性（球形），其密度曲面也有了更大程度的重叠。决策边界进一步平滑，几乎变成圆形，这表明其与原始主方向的严格对齐已大幅放宽，但对噪声的容忍度显著增强。

为进一步说明这些影响，图 4 展示了在不同 ϵ^2 条件下，原始特征空间中的样本分布，以及它们在各类别主轴上的投影。随着 ϵ^2 增大，两个类别的样本都变得更加各向同性，其主方向之间的差异也随之减小，这与 3D 概率曲面中观察到的趋势一致。投影直方图显示，当 $\epsilon^2 = 0$ 时，每个类别沿其自身主轴的分布都呈现出尖锐的峰值，且彼此分离良好，体现出较强的判别能力。随着 ϵ^2 增大，两个类别沿任一主轴的投影分布都变得更加重叠、更加难以区分，这表明误差容忍项的正则化效应能够平滑特征分布，并减弱任何主导方向的影响。这进一步证明，增大 ϵ^2 可通过降低对标签噪声的敏感性来提升鲁棒性。

此示例表明，向协方差矩阵添加误差容忍项 $\frac{\epsilon^2}{d}\boldsymbol{I}$，可通过平滑特征值谱并减轻单个主方向的主导作用来正则化特征分布。因此，决策边界变得更具鲁棒性，也不易受到微小扰动或标签噪声的影响。E.1 节提供了 FedCova 中针对 ϵ^2 的实验性敏感性分析。

C FedCova 算法

D 主要实验详情

D.1 实验设置

总体而言，为了评估抵御噪声标签的基本学习鲁棒性，我们让联邦学习系统处于噪声更集中的场景中。

联邦系统。 为了对鲁棒性进行总体评估，我们在噪声敏感设置下评估联邦学习系统。考虑一个由 $M = 20$ 台设备组成、协同训练模型的联邦学习系统。将一个本地轮次和全部设备参与设为一般情形。需要注意的是，一些在噪声标签下的联邦学习工作 Xu 等人（2022）；Yang 等人（2022）；Wu 等人（2023b）考虑 $M = 100$ 台设备，并将参与比例设为 0.1，根据其提出的方法，这通过仅选择被检测为干净的 10 台设备，间接规避了那些噪声设备。

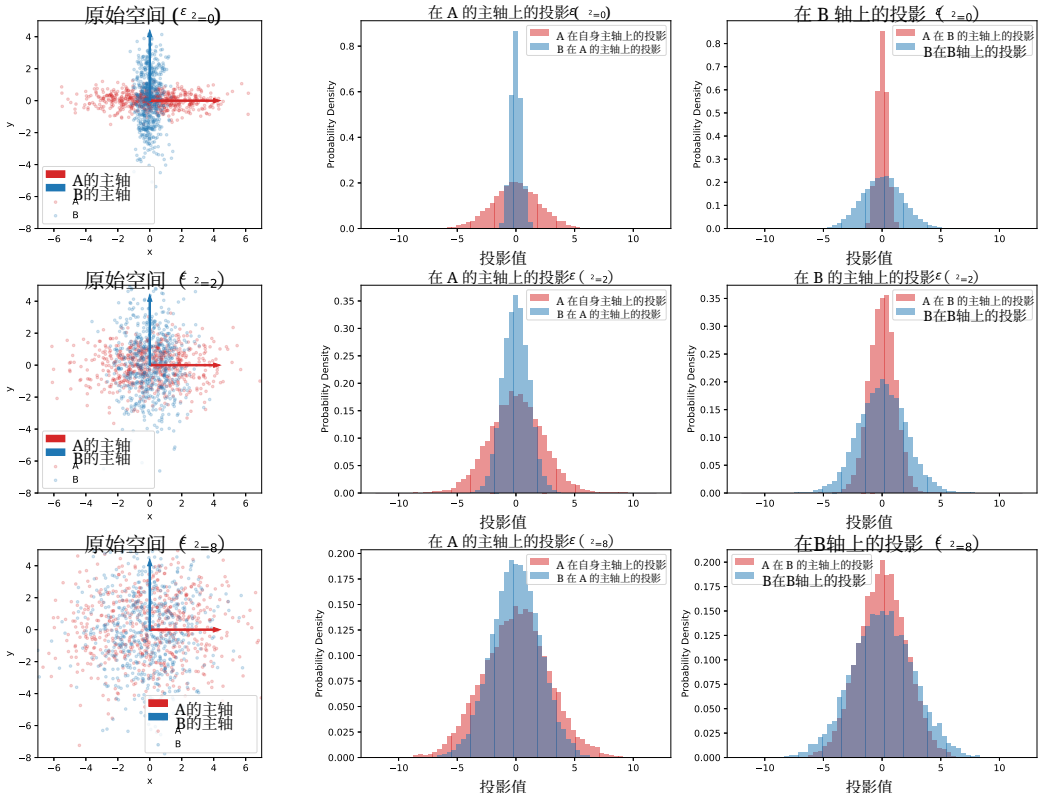


图3： ϵ^2 的效果以样本分布和轴投影的形式呈现。

算法1 FedCova

1：输入：设备数量 M**，类别数量** J**，训练轮次** T**，校正轮次** $\{T_c\}$**，** $\{\mathcal{D}_m^t\}_{m=1}^M$ $D=\sum_m D_m$ η **，数据集大小，以及学习率。 2：初始化：** $t=0$**以及全局模型** $\boldsymbol{w}^{(0)}$**。 3：for** $t\in[0,T]$ **do 4：服务器将全局模型** $\boldsymbol{w}^t$ **和全局分类器** $\boldsymbol{\theta}^t=\{\boldsymbol{\Sigma}_j^t\}_{j=1}^J$**广播给设备； 5：for** $m\in[M]$ **do 6：从服务器接收** $\boldsymbol{w}^t$ **和** $\boldsymbol{\theta}^t=\{\boldsymbol{\Sigma}_j^t\}_{j=1}^J$**。 7：if** $t\in\{T_c\}$**then 8：设备** m **根据式（13）计算其外部校正器** $\boldsymbol{\theta}_{\setminus m}^t=\{\boldsymbol{\Sigma}_{j\setminus m}^t\}_{j=1}^J$**； 9：if** $p\left(y=\hat{y}\mid\boldsymbol{z}\right)\geq\eta_c$ **&** $\hat{y}\neq y$ **then 10：将样本** $(\boldsymbol{x},\ y)\leftarrow(\boldsymbol{x},\ \hat{y})$**重新标记； 11：end if 12：end if 13：设备** m **基于** $\boldsymbol{w}_m^{t+1}\leftarrow\boldsymbol{w}^t-\eta\nabla\mathcal{L}\left(\boldsymbol{w}^t;\ \mathcal{D}_m^t\right)$ **更新本地模型； 14：设备** m **在** $\mathcal{D}_m$**上输出本地特征为** $\boldsymbol{z}_m=f_{\boldsymbol{w}^{t+1}}(\boldsymbol{x}_m)$**； 15：设备** m**基于本地特征估计其本地分类器** $\boldsymbol{\theta}_m^{t+1}=\{\boldsymbol{\Sigma}_{m,j}^{t+1}\}_{j=1}^J$**； 16：设备** m **将其本地模型** $\boldsymbol{w}_m^{t+1}$**及其本地分类器** $\boldsymbol{\theta}_m^{t+1}$**上传至服务器； 17：end for 18：服务器聚合全局模型** $\boldsymbol{w}^{(t+1)}=\sum_m\frac{D_m}{D}\boldsymbol{w}_m^{t+1}$**； 19：服务器根据式（10）聚合全局分类器** $\boldsymbol{\theta}^{t+1}=\{\boldsymbol{\Sigma}_j^{t+1}\}_{j=1}^J$**； 20：end for**

训练任务。 我们在 CIFAR-10、CIFAR-100（Krizhevsky 等（2009））以及真实世界噪声数据集 Clothing1M（Xiao 等（2015））上评估了算法的性能。在网络架构方面，CIFAR-10 使用 ResNet-18（He 等（2016）），CIFAR-

100 使用 ResNet-34，Clothing1M 使用预训练 Resnet-50。为获得特征表示，我们将原始的最终全连接层替换为两个全连接层，从而将特征维度转换为所需维度。

数据异质性。 我们采用一种具有挑战性的数据异质性设置，其中各设备的数据分布为非独立同分布。具体而言，按照 Xu 等人（2022）采用的惯例，我们使用一种通用划分方案，同时考虑类别异质性和数据集大小异质性。为创建非独立同分布的划分，我们首先构造一个大小为 $M \times J$ 的指示矩阵 Ψ，其中每个条目 $\Psi_{m,\ j}$ 表示类别 j 是否包含在设备 m的本地数据集中。每个 $\Psi_{m,\ j}$ 均从固定概率为 p的伯努利分布中采样。对于每个类别 j，我们将 κ_j 定义为包含类别 j样本的设备总数，该值通过对 Ψ的第 j列求和计算得出：$\kappa_j = \sum_{m=1}^{M} \Psi_{m,\ j}$. 接下来，对于每个类别 j，我们从参数为 $\alpha_{\text{Dir}} > 0$的对称狄利克雷分布中采样一个长度为 κ_j的向量 $\boldsymbol{q}_j$。随后，根据概率分布 $\boldsymbol{q}_j$，将类别 j 的样本随机分配给 κ_j 台设备。这种划分方法提供了一种灵活的机制，可用于控制设备间的类别分布以及本地数据集规模的差异。我们在主要实验中将 $p = 0.5$ 和 $\alpha_{\text{dir}} = 5$ 设为相对高度非独立同分布的场景。数据分布如图 4 所示。我们在 E.4 节中调整不同设置以进行其他实验。

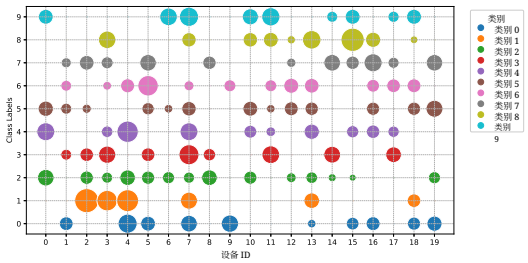
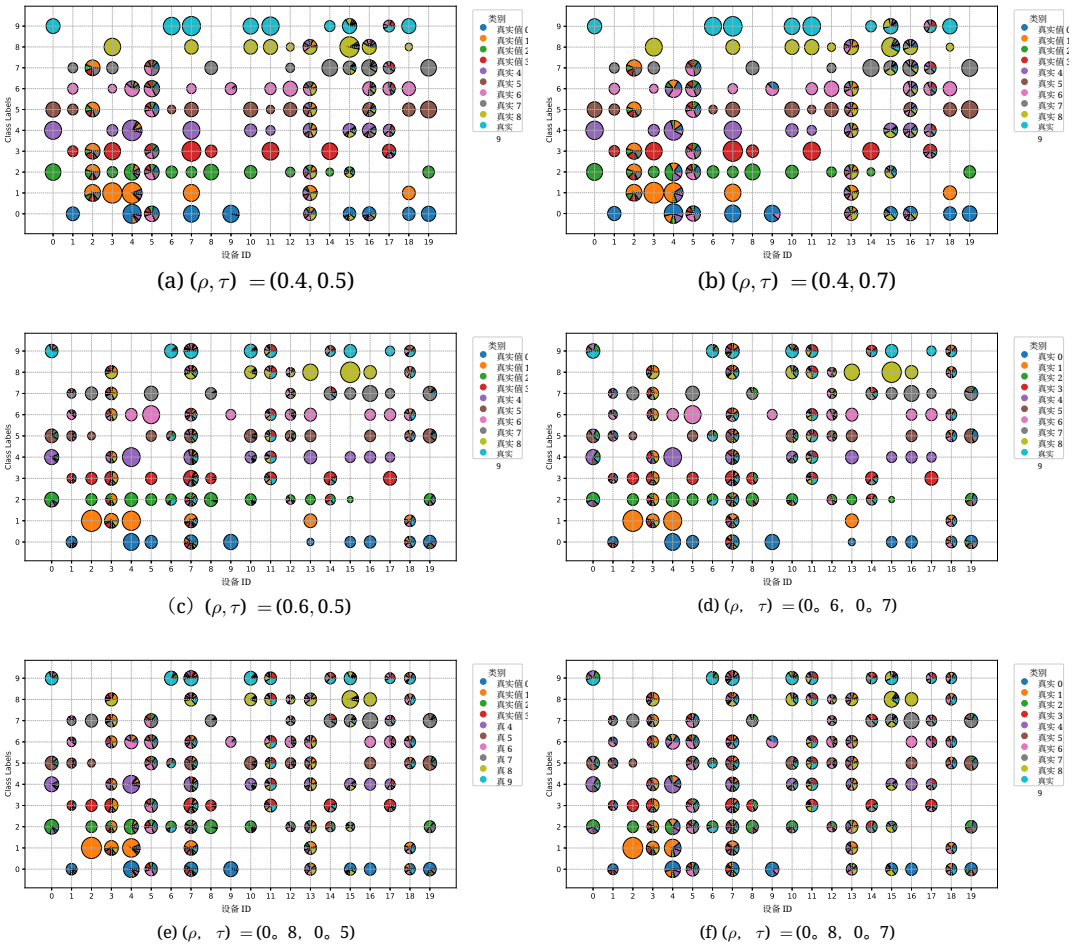


图 4：设备之间的非独立同分布数据分布。

噪声设置。为模拟联邦系统中的真实噪声场景，我们采用由噪声设备比例 ρ 和样本噪声比例 τ 参数化的双层噪声模型。噪声设备比例 ρ 决定包含噪声数据的设备所占比例，而样本噪声比例 τ 则控制每个噪声设备中损坏样本所占比例。对于每个噪声设备，当 $\tau \leq 0.5$ 时，其实际样本噪声比例从 $[0, 2\tau]$ 中均匀采样；当 $\tau \geq 0.5$ 时，则从 $[1-2\tau, 1]$ 中均匀采样，以确保平均样本噪声水平大致以 τ 为中心。此设计使我们能够模拟多样且具有挑战性的噪声条件，以进行鲁棒性评估。我们参照 Song 等人（2022），引入了从对称噪声到非对称噪声的全面噪声模式。在对称噪声模式中，类别被错误标记为随机类别；而在非对称噪声模式中，一个类别主要被翻转为另一个特定类别。在主要实验中，我们在对称噪声模式下采用噪声参数对 (ρ, τ)，取值为 {(0.4, 0.5)、(0.4, 0.7)、(0.6, 0.5)、(0.6, 0.7)、(0.8, 0.5)、(0.8, 0.7)}；在非对称噪声模式下采用 {(0.4, 0.3)、(0.4, 0.5)、(0.4, 0.7)、(0.6, 0.3)、(0.6, 0.5)、(0.6, 0.7)}。图 5 展示了对称噪声模式下噪声数据在设备和类别标签之间的分布。

参数。对于 Resnet-18 和 Resnet-34，我们总共进行 $T = 500$ 个训练轮次；对于预训练 Resnet-50，则进行 $T = 50$ 个训练轮次。学习率设为 $\eta = 0.1$。动量参数设为 $\beta = 0.9$。权重衰减设为 $\gamma = 5e^{-4}$。我们不使用某些基线方法设计中包含的数据混合增强技术。为增强不同设备间的子空间对齐，我们在将全局协方差广播到各设备之前，在服务器端为其加入动量。我们将特征维度设为 $d = 128$。对于校正，我们将校正阈值设为 $\eta_c = 0.5$。校正轮次 $\{T_c\}$ 计划从 $T = 200$ 开始，之后每隔 $\delta = 30$ 轮进行一次校正。在每个校正轮次中，对排名前 k_1 的噪声设备进行校正；对排名前 k_2 的噪声设备进行带外部移除的校正，其中 $k_1 = 10$ 和 $k_2 = 5$。



(a) $(\rho, \tau) = (0.4, 0.5)$ (b) $(\rho, \tau) = (0.4, 0.7)$

(c) $(\rho, \tau) = (0.6, 0.5)$ (d) $(\rho, \tau) = (0.6, 0.7)$

(e) $(\rho, \tau) = (0.8, 0.5)$ (f) $(\rho, \tau) = (0.8, 0.7)$

图 5：不同双层噪声设置的可视化。

计算资源。我们的实验在一台配备一块 NVIDIA A100、一块 NVIDIA RTX A6000 和一块 NVIDIA RTX 3090 GPU 的 GPU服务器上进行。

基线方法。 我们将所提出的算法与几种基线方法进行比较。以下介绍这些方法的主要思路及其具体参数调整。除非另有说明，我们均使用各方法提供的默认最优参数。此处不再赘述通用的 FL 环境设置，包括数据异质性、标签噪声配置等；这些设置遵循上文所述内容，并在相同的随机种子下进行一致模拟，以确保比较公平。

- **FedAvg** McMahan等人（2017年）：FedAvg 是用于聚合分布式本地模型的经典 FL 框架。在每个聚合周期中，客户端将模型更新上传至服务器，由服务器进行加权平均以生成全局模型。FedAvg 不进行标签校正，这体现了 FL 在噪声标签下原有的过拟合问题。
- **RoFL** Yang 等（2022）：RoFL 通过交换按类别划分的特征质心（均值），促进服务器与本地模型之间的协作，从而解决联邦学习中的噪声标签问题。服务器对齐这些质心——它们代表每台设备本地数据的核心特征——并在每个通信轮次中将对齐后的质心广播给选定的客户端。即使客户端之间的噪声分布不同，这一过程也有助于在本地模型之间形成一致的类别决策边界。RoFL 还引入了一种

样本选择方法，用于筛除带有噪声标签的数据，以及一种标签校正方法，用于调整带噪声的实例。

• **FedCorr** Xu 等人（2022）：这是应对噪声标签的最流行框架之一。该框架通过评估设备的局部内在维度，将设备划分为干净类别或噪声类别。此外，它还利用本地训练的过滤器，通过检查单个样本的训练损失，将干净样本与噪声设备区分开来。为提高鲁棒性，该框架在损失函数中引入近端项，并采用数据增强技术。鉴于我们的实验总轮数为 $T = 500$，我们将三个阶段分别设置为 $T_1 = 5$、$T_2 = 250$ 和 $T_3 = 250$。因此，FedCorr 的确切通信开销为$20T_1 + T_2 + T_3 = 600$ 轮。尽管第一阶段会产生较大的通信开销，我们仍将其视为额外的预热阶段。

• **FedNoRo** Wu 等人（2023b）：FedNoRo 是一个两阶段框架，旨在解决类别不平衡和异质标签噪声环境下的鲁棒联邦学习问题。在第一阶段，结合逐类别损失指标与高斯混合模型来识别噪声客户端。在第二阶段，联合采用知识蒸馏和距离感知聚合函数，实现鲁棒的联邦模型更新，从而应对类别不平衡和标签噪声异质性问题。预热阶段设置为 $T_1 = 10$ 轮。

• **FedNed** Lu 等人（2024）：FedNed 通过一种新颖的负向蒸馏策略应对噪声极大的客户端：它利用已识别的噪声客户端的更新，而不是将其丢弃，以引导全局模型避开错误方向。该过程结合伪标签化和基于 KL 散度的损失，增强模型对严重标签噪声的鲁棒性。与所有基线方法不同，在实验中，我们将 FedNed 的参与比例设为 0.5，以确保有极高噪声设备参与。否则，如果所有设备都参与，我们会发现该方法无法区分极高噪声设备。

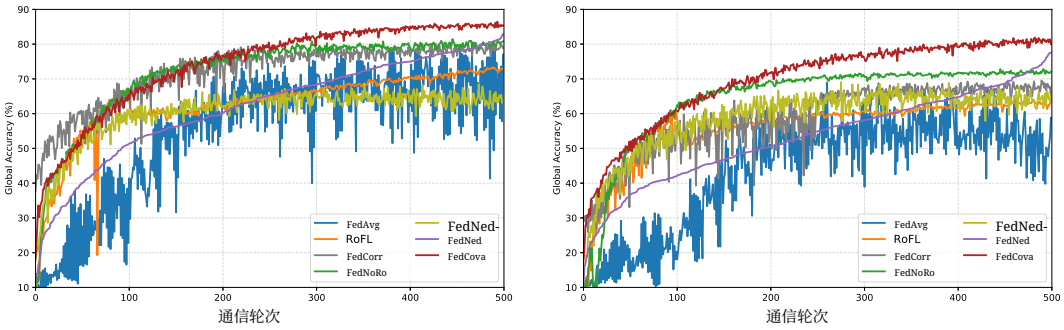
• **FedNed-**Lu 等人（2024）：FedNed- 是 FedNed 的一种退化变体，不再依赖极高噪声设备，因此完全适用于我们所设定的联邦学习严重标签噪声场景。在此配置下，该方法无法区分极高噪声标签，只能利用公共干净数据集来增强噪声韧性。

• **CoteachingFL**：CoteachingFL 将 Co-teaching Han 等人（2018）的范式应用于联邦学习，以应对噪声标签。在协同教学中，两个神经网络同时训练，各自从小批次中选出一部分小损失样本作为干净样本，并用这些样本更新对等网络，通过相互教学减少错误累积。该过程在每个客户端本地执行，以应对噪声标签；随后，两个本地模型通过 FedAvg 框架进行全局聚合。

• **DivideMixFL**：DivideMixFL 将 Li 等人（2020）的 DivideMix 框架应用于联邦学习，以便在噪声标签下进行稳健训练。DivideMix 使用拟合每个样本损失分布的高斯混合模型，动态地将训练数据划分为有标签（干净）集和无标签（含噪声）集，随后采用半监督学习技术，利用两个分歧网络协同划分数据并协同优化标签。类似地，DivideMix 也在每个客户端本地运行，生成的模型再通过 FedAvg 在全局范围内聚合。

D.2 学习曲线

我们在图 6 中绘制了两种噪声设置下各基线方法的准确率曲线随通信轮次的变化。我们观察到，所评估的方法在对噪声标签的鲁棒性方面存在明显差异。在两种设置中，FedCova 始终取得最高的全局准确率，并在整个训练过程中表现出显著的稳定性，大幅优于所有基线方法。特别是，FedCova 不仅收敛更快，而且保持了更平滑、更鲁棒的准确率轨迹，尤其是在噪声增加的情况下，如图 6b 所示。与 FedAvg 等传统聚合方法相比——后者会出现大幅波动且最终准确率低得多——FedCova 对噪声和通信变异性的影响表现出更强的韧性。其他基线方法，包括 FedCorr、FedNed 和 FedNoRo，在早期通信轮次中表现具有竞争力，但随着训练推进最终被 FedCova 超越。总体而言，这些结果凸显了 FedCova 在准确率和鲁棒性方面的优越性，证明了其在异构数据分布下、带有噪声标签的具有挑战性的联邦学习环境中的有效性。



(a) $(\rho,\ \tau) = (0。4,\ 0。7)$　　(b) $(\rho,\ \tau) = (0.6,\ 0.7)$

图6：准确率曲线随通信轮次变化（对比基线方法）。

D.3 子空间正交性

为了定量评估训练过程中不同类别协方差主方向之间正交性的提升，我们计算类别协方差矩阵的主特征向量之间的平均绝对余弦相似度。具体而言，对于每个类别 j 及其协方差矩阵 $\mathbf{\Sigma}_j$，我们提取其主特征向量 $\mathbf{v}j$，对应于最大特征值。然后我们构造矩阵 $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}2, \ldots, \mathbf{v}_J]^\top \in \mathbb{R}^{J\times d}$，其中 J 是类别数量，d 是特征维度。正交矩阵 A 定义为 $A_{uv} = |\mathbf{v}_u^\top \mathbf{v}_v|$，表示类别 u 与 v的主方向之间余弦相似度的绝对值。我们将一种度量——正交性指数 ϕ——定义为

$$\phi = \frac{1}{J(J-1)} \sum_{u\neq v} |\mathbf{v}_u^\top \mathbf{v}_v|, \tag{32}$$

其中，ϕ 值越低，表明类别协方差的主方向越趋于正交，反映出类别间具有更大的多样性和可分性。相反，ϕ 值越高，则表明主方向的对齐程度越高，意味着类别之间的可分性降低。

图 11 展示了在两种噪声设置下，正交性指数 ϕ 和准确率随通信轮次的变化情况。随着训练的推进，ϕ 迅速下降，并在约 100 轮后趋于较低且稳定的数值。这一趋势表明，在我们的目标函数训练下，类别协方差矩阵的主方向变得越来越正交，说明类间判别能力得到增强。值得注意的是，ϕ 的降低与全局准确率显著上升同步发生，如右侧坐标轴所示。这种相关性表明，模型学习更正交（即更具判别性）的类别方向的能力直接带来了性能提升。此外，在较高噪声设置（$\rho,\ \tau$）$= (0.6,\ 0.7)$下，与较低噪声情况相比，ϕ下降得更慢，这反映了噪声对特征对齐的影响，以及误差容忍项所提供的增强容忍度。尽管如此，两种设置下的整体趋势仍然一致，凸显了所提方法在促进类别可分性方面的鲁棒性，并且即使在严重标签噪声条件下也能实现更高的准确率。

D.4 校正性能

我们在图8和图9中全面评估了FedCova在不同标签噪声设置下的校正性能。如图8所示，引入校正后，随着通信轮次的推进，噪声样本总数持续且大幅减少。对于两种噪声设置，$(\rho, \tau) = (0.4, 0.7)$ 和 $(\rho, \tau) = (0.6, 0.7)$，全局噪声率呈现出明显的下降趋势，同时全局测试准确率稳步提升。具体而言，在$(\rho, \tau) = (0.4, 0.7)$下，全局噪声率从27.51%降至6.93%；对于更具挑战性的情形$(\rho, \tau) = (0.6, 0.7)$，其从34.42%降至11.20%。值得注意的是，即便

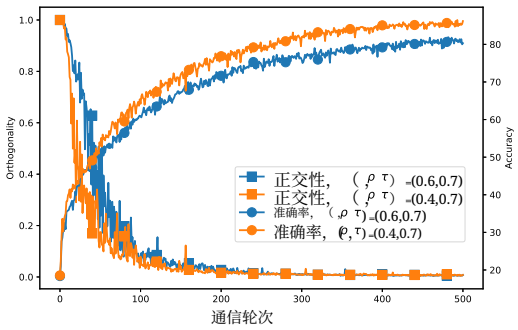


图 7：正交性和准确率随通信轮次变化的曲线。

在初始噪声较高的场景中，校正机制在降低噪声样本并提升准确率方面仍然高度有效，展现出对严重标签污染的鲁棒性。

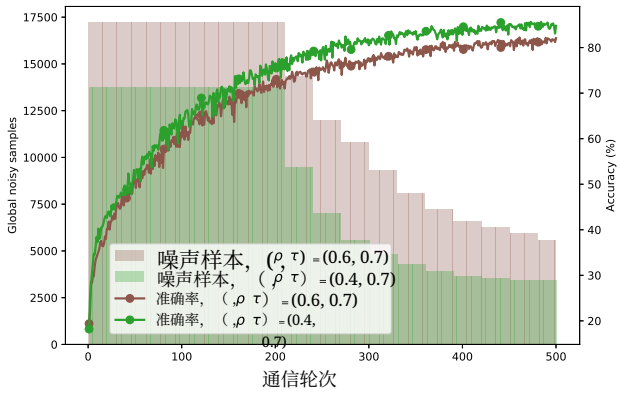
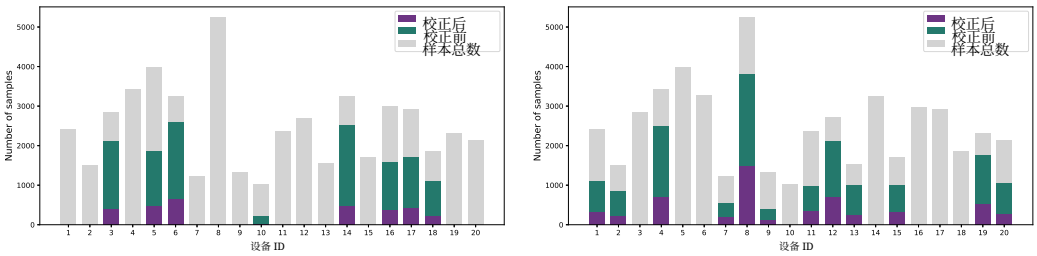


图 8：校正性能和准确率随通信轮次变化的曲线。

图 9 进一步按设备级展示了两种噪声设置下的校正性能。图中显示了每台设备在校正前后的噪声样本数。显然，校正过程显著降低了各设备上的噪声；校正后的噪声样本数（紫色条形）明显少于校正前（绿色条形）。无论设备的本地数据集规模较大还是较小，设备上的样本噪声比例较高还是较低，这一效果都始终如一，表明该校正策略对异质的本地数据分布均有效。



(a) $(\rho, \tau) = (0.4, 0.7)$　　(b) $(\rho, \tau) = (0。6, 0。7)$

图 9：设备上的校正性能。

总体而言，这些结果凸显了我们校正机制的有效性和鲁棒性。噪声整体减少、准确率提高，以及设备层面的去噪效果始终如一，

这些现象在不同程度的异构性和噪声条件下均有所体现，表明在精确分类器和外部校正器的辅助下，我们的周期性校正机制能够通过重新标注数据集，有效减轻标签噪声的影响。

D.5 运行时间比较与实用性分析

我们进一步分析不同鲁棒联邦学习方法的宏观级运行效率，以评估其实际可部署性。尽管计算效率并非本文的主要关注点，但运行时开销是系统层面的重要考量，尤其是在大规模或资源受限联邦场景中。

所有方法均在相同的硬件与软件配置下进行评估。总训练运行时间在单个 NVIDIA RTX A6000 GPU 上进行测量，并在表6中报告。我们还报告了相对于标准 FedAvg 归一化后的运行时间，标准 FedAvg 代表联邦优化中的最小开销基线。

表6：不同方法的总运行时间比较（分钟）。

方法	FedAvg	CoteachingFL	DivideMixFL	RoFL	FedCorr	FedNoRo
运行时间	290	1217	1508	383	643	664
比率	1.0×	4.2×	5.2×	1.3×	2.2×	2.3×

方法	FedNed-	FedNed	**FedCova**
运行时间	332	390	**467**
比率	1.1×	1.3×	**1.6**×

如表6所示，CoteachingFL 和 DivideMixFL 的运行开销最高，主要是因为它们在整个学习过程中需要同时训练两个模型，这显著增加了计算和通信开销。相比之下，RoFL 和 FedNed 的开销相较于 FedAvg 较为适中。

在鲁棒方法中，FedCova 实现了均衡的效率表现。与 FedCorr 和 FedNoRo 相比，FedCova 大幅缩短了总运行时间。这一改进主要归功于移除了高开销的预热阶段；在其他方法中，这些阶段需要额外进行数百轮设备—服务器通信，之后才能启用有效的鲁棒性机制。

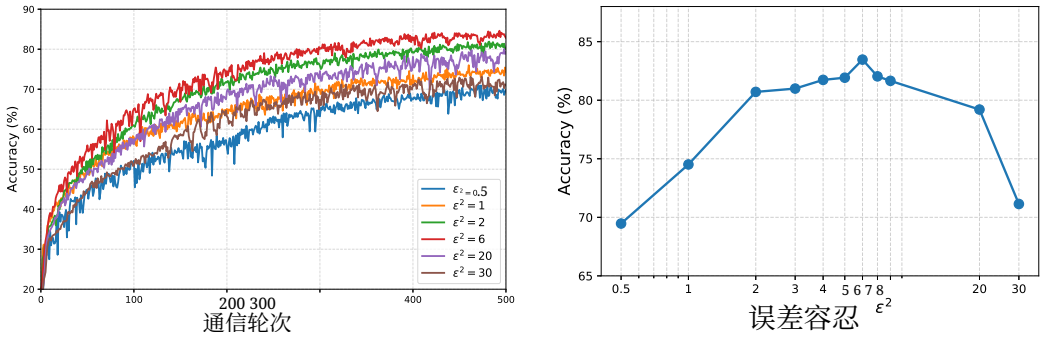
解读运行时间结果时，也应结合各方法所作的假设。虽然 FedNed 的运行时间略短于 FedCova，但它依赖于可用的干净公共数据集。在现实世界的联邦学习场景中，这种假设往往不切实际，而且实际上提供了外部监督。相比之下，FedCova 无需依赖干净的辅助数据或漫长的预热流程，即可在内部处理标签噪声。

总体而言，FedCova 在鲁棒性与计算效率之间实现了良好的权衡，因此更适合在现实的联邦学习约束下实际部署。

E 补充实验结果

E.1 误差容限的影响 ϵ^2

图 10 系统地研究了误差容限参数 ϵ^2 的影响，该参数作为各类别协方差矩阵的各向同性正则化项 $\frac{\epsilon^2}{d}\boldsymbol{I}$ 引入。实验在噪声设置 $(\rho,\ \tau) = (0。6,\ 0。7)$ 下进行，其余参数均采用默认值。该正则化用于平滑协方差的特征值谱，减轻任一单独主方向占据主导地位的情况，并有效放宽类别特定特征子空间之间严格的正交约束。此机制旨在防止损失函数对数据中的微小扰动过于敏感，从而增强模型对噪声样本或错误标记样本的鲁棒性。



(a) 准确率与通信轮次　　(b) 可实现的准确率

图 10：不同误差容忍度 ϵ^2下的准确率。

如图 10a 所示，将 ϵ^2 从 0。5 增至 2，可显著提升收敛速度和最终准确率，因为增加的正则化使模型能够更好地容忍标签噪声和离群点。图 10b 进一步证实了这一点：在此范围内，可达到的准确率随 ϵ^2 大幅提升，并在约 $\epsilon^2 = 6$ 处达到峰值。这表明，适度的正则化使模型能够关注特征空间中关键的统计结构，而非过拟合特定特征值或噪声标签。然而，当 ϵ^2 超出最佳范围后，准确率便开始下降。具体而言，当取值较大时，例如 $\epsilon^2 = 20$ 和 $\epsilon^2 = 30$，可达到的准确率明显降低。这可归因于过度平滑：它过度放宽了类别边界，削弱了所学习特征表示的判别能力。换言之，引入 ϵ^2 对提升鲁棒性和稳定训练至关重要，但误差容忍度过大会使特征子空间之间的差异减弱，从而损害类别可分性。

总体而言，这些结果凸显了误差容限参数的双重作用：它通过正则化协方差谱，增强对噪声和离群点的鲁棒性；但要在模型灵活性与可分性之间取得平衡，就需要仔细调节该参数。观察到的准确率趋势验证了纳入 $\frac{\epsilon^2}{d}\boldsymbol{I}$ 项的理论动机。

从隐私角度进行讨论 ϵ。由于在公式（6）中将 $\boldsymbol{n} \sim \mathcal{N}\left(\mathbf{0}, \frac{\epsilon^2}{d}\boldsymbol{I}\right)$ 添加为高斯噪声，因此这种噪声注入方式与差分隐私（DP）的隐私保护机制相一致。从特征的角度来看，我们因此不可避免地引入了相应的噪声。考虑到隐私问题，这种权衡远不止噪声韧性与模型可分性之间的取舍，还涉及性能与隐私保护水平之间的平衡。随着 ϵ 增大，隐私保护程度提高，但由于特征受到更多噪声干扰，准确率也面临更大挑战。

E.2 增强系数 α的影响

在本节中，我们研究子空间增强分类器中增强系数 α 的不同设置如何影响模型在标签噪声下的性能。该实验在噪声设置 $(\rho, \tau) = (0.6, 0.7)$ 下进行，其余所有超参数均设为默认值。回顾可知，子空间增强分类器中的增强系数 α 控制广义分类器中的指数，其形式如公式12所示。具体而言，$\alpha = 1$ 恢复了最大后验分类器所使用的标准马氏距离，而 $\alpha = 2$ 对应于Chan等人（2022）提出的投影分类器。一般而言，较大的 α 会通过放大各类别特定特征子空间之间的差异来增强分类器的判别能力，但也可能降低对标签噪声和离群点的容忍度。

如图 11 所示，测试准确率对 α 的依赖呈非单调变化。总体而言，随着 α 从 1 增至 8，准确率不断提高，这凸显了子空间增强在噪声标签条件下提升判别能力的效果。最佳性能出现在 $\alpha = 7$。$\alpha = 2$ 下的特殊表现（主实验所采用的设置）可能归因于默认超参数（包括 ϵ^2）是在该设置下调优的，这表明

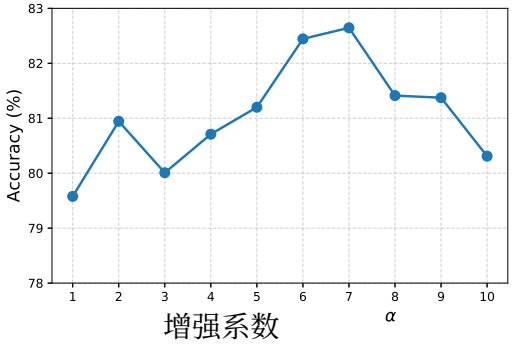


图 11：不同增强系数 α下的准确率。

为获得最佳结果，需要协同调整误差容忍度和判别强度。超过 $\alpha = 7$ 后，进一步增大 α 会导致准确率开始下降。这一趋势表明，虽然增大 α 最初会通过惩罚协方差次要轴方向上的偏差来增强类别分离，但过大的 α 可能使分类器对协方差中的微小扰动和估计噪声过于敏感，从而削弱其在噪声标签下的鲁棒性。

这些结果证明了我们的子空间增强策略是有效的：引入增强系数 α 可增强模型的判别能力，并在标签噪声下显著提升性能。此外，适当协调增强带来的判别能力与误差容忍所产生的放宽效果，有助于在标签噪声下实现均衡的性能。

E.3 特征维度的影响

我们分析特征维度 d 的选择如何影响我们基于协方差的特征学习框架中判别能力与对噪声标签的鲁棒性之间的权衡。实验采用噪声配置（ρ, τ）$= (0.6, 0.7)$，而所有其他超参数均保持默认设置。

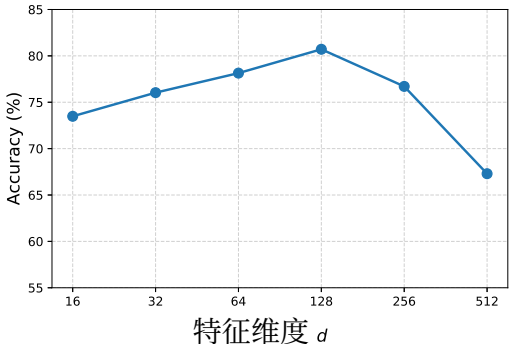


图 12：不同特征维度 d下的准确率。

如图 12 所示，分类准确率首先随着 d 增加，在 $d = 128$ 处达到峰值，随后随着 d 持续增大而下降。这一趋势凸显了两个关键现象。一方面，增大特征维度能够提供更丰富的表示空间，使协方差矩阵可以捕捉更加细致的类内和类间变化，从而增强模型的判别能力。另一方面，过大的特征维度可能引入过参数化，使协方差估计变得不太可靠——尤其是在存在噪声标签且每类样本量有限的情况下。在不同设备之间实现主方向对齐也会变得越来越困难。在这种情况下，模型可能更容易受到噪声影响，导致准确率下降。

计算复杂度和存储开销。 为评估 FedCova 的实际可部署性，我们分析其计算和存储复杂度。由于 B 表示批大小， d 表示特征维度， J 表示类别数量。骨干网络（例如 ResNet-18）占据了主要计算成本，每次迭代需要 $\mathcal{O}(10^9)$ FLOPs。FedCova 额外引入了两个步骤：（1）协方差累积：计算协方差矩阵需要 $\mathcal{O}(Bd^2)$。（2）损失计算：计算 log det(Σ) 及其梯度需要通过 Cholesky分解执行 $\mathcal{O}(d^3)$。在 $d = 128$ 的情况下， $\mathcal{O}(d^3)$ 项对应于 ≈ 2 MFLOPs，比骨干网络计算小三个数量级。此外，这些矩阵运算可以使用标准库（例如 cuBLAS）在 GPU 上自然地并行执行，确保其不会成为瓶颈。对于存储开销，客户端必须存储 J 个大小为 $d \times d$ 的协方差矩阵。存储复杂度为 $\mathcal{O}(Jd^2)$。对于 $J = 10$ 和 $d = 128$，协方差矩阵的存储需求约为 0.6 MB（使用 32位浮点数）。与模型本身的内存占用相比（例如，ResNet-18 占用 ≈ 45 MB），这对于现代移动设备而言微不足道。对于包含 $J = 100$ 个类别的数据集，总内存开销约为 6 MB（相比之下，ResNet-34 占用 ≈ 83 MB）。这完全在资源受限型移动设备的容量范围内。

通信开销分析。在我们的框架中传输的特征协方差矩阵，对于 J 个类别中的每个类别，矩阵大小为 $d \times d$。以主要实验所采用的设置为例，其中 $d = 128$ 和 $J = 10$，需要传输的协方差参数总数为 $J \times d^2$ =$10 \times 128^2 = 163,840$。相比之下，标准 ResNet-18 模型的参数量约为 11。2 百万。因此，传输所有类别协方差矩阵所带来的额外通信开销仅约占模型参数的 1。4%，小了两个数量级。这表明，在实际的联邦学习场景中，相对于模型整体规模，额外通信开销可以忽略不计。

综合来看，这些结果凸显了谨慎选择特征维度 d 的必要性，以平衡分类可区分性、对噪声的鲁棒性和通信
效率。

E.4 不同数据异质性设置

我们进一步通过改变非独立同分布划分参数（p， α_{dir}），研究 FedCova 在不同数据异质性下的性能。如技术部分所述， p 控制每个设备包含特定类别的概率，而 α_{dir} 决定每个类别在不同设备间的样本量变化程度。这种灵活的划分方案使我们能够模拟类别分布和本地数据集规模在不同程度上的异构性。

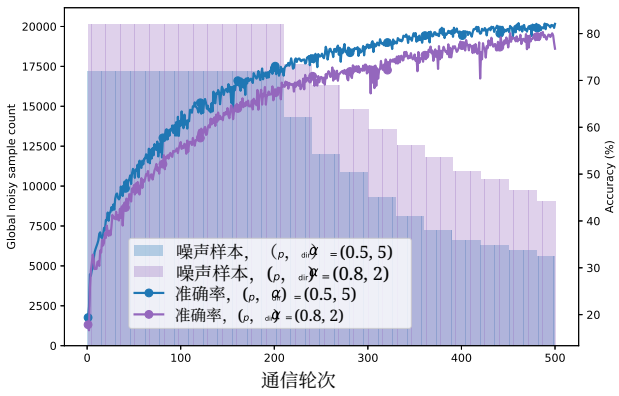


图13：不同数据异质性设置下的校正性能和准确率 .

图 13 展示了两种具有代表性的异质性设置下的校正性能和准确率：(p， α_{dir}) =(0.5，5)（每台设备上的类别多样性较高，数据规模更均衡）和 (p， α_{dir}) =(0.8，2)（每台设备上的类别多样性较低，数据规模更偏斜）。实验在噪声设置 (ρ， τ) =(0.6，0.7) 下进行，其他所有超参数均设为默认值。如图 13 所示，我们的校正框架在训练过程中持续降低全局噪声率。具体而言，在 (p, α_{dir}) = (0.5，5) 下，全局噪声率从 34.42% 降至 11.20%，而

对于 $(p,\ \alpha_{\text{dir}}) = (0.8,\ 2)$，则从 40.33% 降至 18.16%。这表明，即使在更具挑战性的异质性设置下，我们的周期性校正机制仍然有效。即使在更具挑战性的异质性设置 $(p,\ \alpha_{\text{dir}}) = (0.8,\ 2)$ 下，我们的方法仍能显著降低噪声率，并保持较高的准确率，展现出对严重数据异质性的鲁棒性。