arXiv:2601.03584v1 [cs.LG] 7 Jan 2026

局部梯度调节在客户端异质性下稳定联邦学习

Ping Luo[1] Jiahuan Wang[1] Ziqing Wen[1] Tao Sun[1*] Dongsheng Li[1*]

[1]并行与分布式计算国家重点实验室，计算机科学与技术学院，
国防科技大学，长沙，410073，中国

luoping@nudt.edu.cn, wangjiahuan@nudt.edu.cn, zqwen@nudt.edu.cn,
suntao.saltfish@outlook.com, dsli@nudt.edu.cn

摘要

联邦学习（FL）支持分布式客户端协同训练模型，无需共享原始数据，但在现实部署中，统计异质性从根本上对其稳定性构成挑战。本文发现，客户端异质性主要通过扭曲客户端优化期间的局部梯度动态，使联邦学习不稳定，导致系统性漂移在多个通信轮次中不断累积，并阻碍全局收敛。这一观察表明，局部梯度是稳定异质联邦学习系统的关键调节杠杆。基于这一洞见，我们提出一种通用的客户端侧视角，在不增加额外通信开销的情况下调节局部梯度的贡献。受群体智能启发，我们通过探索-收敛梯度再聚合（ECGR）实现这一思路，在对齐良好和未对齐的梯度分量之间取得平衡，从而保留有效更新，同时抑制不稳定因素。理论分析和大量实验（包括在 LC25000 医学影像数据集上的评估）表明，在异质数据分布下，对局部梯度动态进行调节能够持续稳定多种最先进方法中的联邦学习。

1 引言

联邦学习（FL） [McMahan 等人， 2017] 已成为一种分布式机器学习范式，无需客户端共享其原始数据即可实现协同模型训练。随着数据孤岛和日益严格的隐私法规持续限制集中式学习，联邦学习通过将敏感数据保留在本地、仅交换模型更新，提供了一种有效的解决方案。近年来，联邦学习在众多领域取得了显著成功，包括计算机视觉、自然语言处理和推荐系统 [Kairouz 等人， 2021]。特别是，其保护隐私的特性使联邦学习在医疗和医疗保健应用中极具吸引力，例如跨机构医学图像分析 [Lee 等人， 2024], 电子健康记录建模 [Sadilek 等人， 2021], 以及疾病风险预测 [Dayan 等人， 2021], ，这些领域的数据共享往往受到严格限制。这些进展凸显了联邦学习作为大规模、注重隐私的智能系统基础所具有的实践潜力。

尽管前景广阔，联邦学习在现实场景中仍面临根本性挑战，其中最突出的是客户端之间普遍存在的统计异质性 [Ma 等人， 2022]。在实践中，客户端数据很少满足独立同分布（IID），这违背了 FedAvg 等经典联邦优化算法所依赖的一项关键假设 [McMahan 等人， 2017]。非独立同分布数据分布会显著减缓收敛、导致训练不稳定，并使模型性能大幅下降 [Wang 等人， 2020]。在本地训练期间，异构数据会产生特定于客户端的更新方向，这些方向可能与全局最优点显著偏离 [Zhang 等人， 2021]。此类梯度差异的累积是优化难度的主要来源，最终限制了联邦学习在现实部署中的有效性和可扩展性。

*通讯作者。贡献相同。

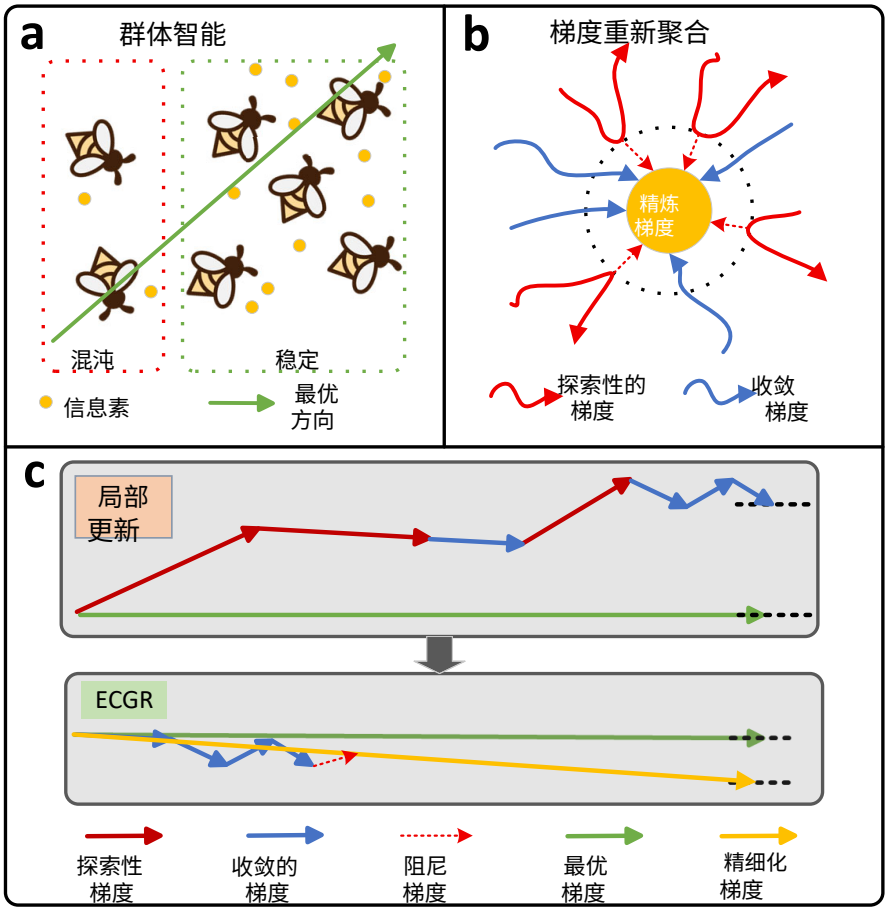


图 1：**ECGR** 策略的框架。**(a)** 蜜蜂群体智能示意图：觅食路径通常包含混沌方向和稳定方向，其中稳定方向主导群体行为。**(b)** 受群体智能启发，联邦学习中的局部梯度被分为探索梯度 和收敛梯度，并重新聚合，使收敛梯度主导最终更新。**(c)**聚合梯度的二维可视化，展示了 ECGR 如何减少数据异质性引起的梯度偏差。

在我们此前的工作中 [Luo 等人， 2025]，，我们系统研究了联邦学习在非独立同分布数据下的优化行为，并识别出导致性能下降的一个关键机制。具体而言，统计异质性最直接地体现在其对局部梯度的影响上：非独立同分布数据会改变本地训练期间客户端更新的方向和幅度，从而引发显著的客户端漂移。这种梯度层面的扭曲并非仅由全局聚合造成，而是在本地出现并随着训练轮次不断累积，因此有必要重新审视通信之前局部更新的形成方式。

受这一洞见的启发，本研究的核心思想是在客户端直接处理局部梯度，以减轻客户端漂移，同时不引入任何额外的通信开销，也不修改现有的联邦学习协议。受蜜蜂觅食行为中观察到的群体智能启发 [Tereshko 和 Loengarov，2005年，Kalavakonda 等人， 2025], 我们提出了一种新颖的梯度重新聚合策略，称为 ECGR（探索-收敛梯度再聚合）。如图 1 (a) 所示，尽管一部分蜜蜂充当探索者，沿偏离最优路径的路线飞行，但群体行为最终由携带稳定信息素信号和一致方向信息的蜜蜂主导。将这一机制映射到联邦学习中的单个客户端（图 1 (b)），我们将每个局部梯度视为一只蜜蜂，而梯度中编码的信息则对应于信息素信号。与最优更新方向偏离较大的局部梯度被识别为探索梯度；尽管含有噪声，它们仍包含对模型收敛至关重要的信息。相比之下，梯

度与最优方向高度一致的梯度被定义为收敛梯度，并作为更新的主要贡献者。如图 1 (c) 所示，ECGR 保留收敛梯度的全部贡献，同时通过阻尼细化机制从探索梯度中提取有用信息。随后，将所得梯度重新缩放，使其范数与原始局部更新的范数相匹配，从而获得更加稳定、稳健的优化轨迹。

综上，这些结果确立了一种面向联邦学习的通用客户端优化视角：通过提炼局部梯度来缓解客户端漂移，并从理论和实证两方面证明了 ECGR 在异质数据分布下的有效性。

2 相关工作

非独立同分布数据下的联邦优化。近年来，大量研究致力于缓解统计异质性对联邦学习的不利影响。早期工作主要通过修改聚合规则或引入控制变量来纠正有偏的本地更新，以应对非独立同分布数据。代表性方法包括 FedProx [Li 等人， 2020], 通过近端项约束本地更新；以及 SCAFFOLD [Karimireddy 等人， 2020], 通过使用控制变量减少客户端漂移。随后，研究人员探索了自适应聚合和归一化策略，例如 Fed-Nova [Wang 等人， 2020] 和 FedAvgM [Hsu 等人， 2019], 以稳定异质数据分布下的收敛。近年来，研究人员提出了个性化联邦学习和聚类联邦学习方法，通过学习多个面向特定客户端或群组的模型，显式考虑客户端异质性 [Fallah 等人，2020年、Ghosh 等人， 2022]。尽管这些方法已被证明有效，但它们通常着眼于全局聚合或客户端参与，而本地优化动态的结构在很大程度上仍未得到考察。

利用联邦学习中的局部梯度。除了以聚合为核心的策略外，近年来也有研究开始探讨如何利用局部梯度信息来改进联邦训练。多项研究利用梯度统计信息来指导客户端选择或加权，优先选择更新信息量更大或更可靠的客户端 [Nishio and Yonetani, 2019, Tang et al., 2022, Li et al.，2022]。其他研究则利用局部梯度来识别高质量或具有代表性的数据子集，从而降低噪声或存在偏差的本地样本所带来的影响 [Schutte et al., 2024, Li et al., 2021a]。此外，研究者还探索了基于梯度的筛选机制，以便在异构环境中检测掉队客户端或异常更新 [Pillutla et al.，2022]。这些方法表明，局部梯度蕴含着丰富的数据质量和优化行为信息。然而，现有大多数方法都是间接利用梯度——用于客户端或数据选择——而非直接处理局部梯度集本身。相比之下，近期少数研究开始考虑显式的梯度级操作，例如梯度裁剪 [Zhou et al.，2025], 过滤 [Han et al.，2024], 或重新加权 [Li et al.,2023], ，以提高鲁棒性。我们的工作与这一新兴方向一致，但有着根本区别：我们在客户端侧对局部梯度进行结构化蒸馏，从噪声更新中提取有用信息，而不丢弃这些更新，也不增加通信开销。

计算病理学中的联邦学习。由于医学影像数据具有敏感性、规模庞大且分散于不同机构，计算病理学已成为联邦学习的重要应用领域 [Adnan et al.，2022]。近期研究已证明，联邦学习可用于全切片图像分析 [Li et al., 2021b], 肿瘤分类 [Al-Asfoor et al., 2024], 以及预后预测 [Feng et al., 2024, Tahir et al.，2025] ，并应用于分布式病理中心。由于染色方案、扫描仪、患者人口统计特征和临床实践各不相同，该领域的非独立同分布数据问题尤为突出 [Xiang et al., 2023, Lu et al.，2021]。为应对这些挑战，既有研究已探索适用于病理数据的领域自适应、归一化和个性化联邦学习策略 [An- tunes et al., 2022, Lu et al., 2022a]。然而，异质局部梯度引发的优化不稳定性仍是一个关键瓶颈。通过在聚合前直接提炼局部梯度，所提出的 ECGR 策略提供了一种互补的优化视角，非常适合计算病理学以及其他隐私敏感型医疗应用所固有的异质性。

3 方法

我们首先介绍联邦平均（FedAvg）[McMa- han et al., 2017] 训练过程中的相关定义和符号，包括局部更新和全局更新阶段。在此基础上，我们提出一种称为*ECGR*的新机制，用于改进局部更新策略。随后，我们将其整体工作流程形式化，并详细介绍相应的算法设计。

3.1 预备知识：联邦平均（FedAvg）

客户端与数据集。

考虑 N 个客户端，每个客户端均关联一个本地数据集 $\mathcal{D}_i \subset \mathcal{D}$（$i = 1$、2、.....、$N$），其中 $\boldsymbol{x}_i \in \mathcal{D}_i$ 表示一个训练样本。客户端采样权重遵循 FedAvg 中的常规设定，即

$$p_i = \frac{|\mathcal{D}_i|}{|\mathcal{D}|}, \quad \text{with } |\mathcal{D}| = \sum_{i=1}^{N} |\mathcal{D}_i|. \tag{1}$$

通信轮次与局部更新。

联邦学习过程共进行 $T \geq 1$ 个通信轮次，服务器在第 $t = 0$、1、.....、T 轮维护全局参数 $\boldsymbol{w}_t$。在每一轮 t 中，客户端 i 进行 E 个本地轮次的训练，对应于批次大小为 B 的 $\tau_i = E\frac{|\mathcal{D}_i|}{B}$（$\tau_i \geq 1$）次本地随机梯度下降迭代。令 $\boldsymbol{w}_{(t,i)}^{\lambda}$ 表示迭代 $\lambda = 0$、1、.....、τ_i 时的本地模型参数。局部更新规则为

$$\boldsymbol{w}_{(t,i)}^{\lambda+1} = \boldsymbol{w}_{(t,i)}^{\lambda} - \eta_l \nabla F_i(\boldsymbol{w}_{(t,i)}^{\lambda}; \boldsymbol{x}_{s_i}), \tag{2}$$

其中，η_l 是本地学习率，$F_i(\cdot)$是本地损失函数，$s_i = \{1, 2, \ldots, \tau_i\}$ 表示小批次的排列。

本地与全局聚合梯度。

对于每个客户端 i，首先收集在各个小批次上计算出的梯度，然后将其聚合以获得

$$\boldsymbol{g}_{(t,s_i)} := \sum \underbrace{\left\{ \eta_l \nabla F_i(\boldsymbol{w}_{(t,i)}^{\lambda}; \boldsymbol{x}_{s_i}) \right\}_{\lambda=1}^{\tau_i}}_{\text{local gradient set}}. \tag{3}$$

在每个局部梯度集内对梯度进行本地聚合（通常采用平均）后，即可得到本地更新后的训练梯度 $\boldsymbol{g}_{(t,s_i)}$。随后将其传输至参数服务器进行全局聚合（通常采用加权平均），如下所示：

$$\boldsymbol{G}_t = \sum_{i=1}^{N} p_i \boldsymbol{g}_{(t,s_i)} \tag{4}$$

需要注意的是，前述所有操作均在各个客户端上执行，而此步骤及后续的全局更新则在参数服务器上进行。

全局更新。

在第 t 轮次获得全局梯度 $\boldsymbol{G}_t$ 后，通过一次直接的随机梯度下降步骤更新全局模型：

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_g \boldsymbol{G}_t. \tag{5}$$

其中，η_g 是全局学习率。在本文中，我们将所有客户端的局部学习率 η_l 设为相同，并固定全局学习率 $\eta_g = 1$。

3.2 探索梯度-收敛梯度重新聚合（ECGR）

在我们提出的ECGR方法中，每个客户端的局部梯度集在局部聚合之前会进行有放回的选择性采样。选择策略包括三个步骤：

1. **幅值排序**：根据局部梯度的幅值选取排名前一半的梯度，确保所得聚合向量的 ℓ_2 差异最小。这些选出的梯度称为收敛梯度。2. **衰减提取**：幅值排序操作后剩余的梯度称为探索梯度。将每个梯度乘以阻尼因子 $\beta \in [0,1]$（实验中通常为 $\beta = 0.1 \sim 0.5$），以减小其影响。3. **重新聚合**：对前述步骤得到的收敛梯度和探索梯度进行局部聚合，并对所得向量重新缩放，使其范数与原始聚合梯度的 ℓ_2 范数一致。

下面给出各个步骤的正式定义。

幅值排序

我们采用 Lu等人 [2022b] 提出的基于集群的贪心策略，从局部梯度集 $\left\{\eta_l \nabla F_i(\boldsymbol{w}_{(t,i)}^{\lambda}; \boldsymbol{x}_{s_i})\right\}$中依次采样一半的梯度。选出的梯度构成一个子集 $\pi_i = \{e_1, e_2, \ldots, e_k\}$，其中 e_λ 表示排序后由排列 s_i 所确定的索引，且 $k = \lfloor \tau_i/2 \rfloor$。首先，令 $\boldsymbol{S}_0 = \mathbf{0}$ 和 $R_0 = s_i$。在第 λ 步（$\lambda \in [1,k]$），我们选择

$$e_\lambda = \arg\min_{e_\lambda \in R_{\lambda-1}} \left\| \boldsymbol{S}_{\lambda-1} + \eta_l \nabla F_i(\boldsymbol{w}_{(t,i)}^{e_\lambda}; \boldsymbol{x}_{s_i}) \right\|, \tag{6}$$

并更新

$$\boldsymbol{S}_\lambda := \boldsymbol{S}_{\lambda-1} + \eta_l \nabla F_i(\boldsymbol{w}_{(t,i)}^{e_\lambda}; \boldsymbol{x}_{s_i}), \qquad R_\lambda := R_{\lambda-1} \setminus \{e_\lambda\}. \tag{7}$$

最后，我们得到

$$\pi_i = \arg\min_{\pi_i \subset s_i} \left\| \boldsymbol{g}_{(t,\pi_i)} \right\|, \qquad \boldsymbol{g}_{(t,\pi_i)} = \sum_{\lambda=1}^{\lfloor \tau_i/2 \rfloor} \eta_l \nabla F_i(\boldsymbol{w}_{(t,i)}^{\lambda}; \boldsymbol{x}_{\pi_i}). \tag{8}$$

在此步骤中，所选集合 π_i 代表客户端梯度集中的“收敛”部分，因为其中包含与全局下降趋势方向一致的梯度，同时过滤掉受局部噪声或离群值主导的梯度。这种选择有助于稳定优化过程，从而加快全局收敛并提升泛化能力。然而，Luo 等人 [2025]的实验发现表明，仅应用此步骤可能会导致有益梯度信息的丢失。因此，还需要进一步通过衰减提取来提取其他有用梯度。

衰减提取

通过幅值排序步骤获得梯度索引集 π_i 后，剩余的梯度集可直接推导如下：

$$\pi_i' = s_i \setminus \pi_i, \qquad \boldsymbol{g}_{(t,\pi_i')} = \sum_{\lambda=1}^{\lfloor \tau_i/2 \rceil} \eta_l \nabla F_i(\boldsymbol{w}_{(t,i)}^{\lambda}; \boldsymbol{x}_{\pi_i'}). \tag{9}$$

相比之下，集合 π_i' 对应于“探索性”梯度，其中包含的分量可能仍有助于全局收敛，但也含有更多随机或有偏信息。为平衡探索性与稳定性，这些梯度会乘以阻尼因子 $\beta \in [0,1]$，以减轻潜在有害更新的影响，同时保留有益的探索方向，从而增强模型鲁棒性并防止过早收敛。

重新聚合

从幅值排序 π_i 和对齐排序 π_i'获得梯度子集后，下一步是将它们组合起来，形成重新聚合梯度。

$$\boldsymbol{g}'_{(t,s_i)} = \gamma_i(\boldsymbol{g}_{(t,\pi_i)} + \beta \boldsymbol{g}_{(t,\pi_i')}), \qquad \gamma_i = \|\boldsymbol{g}_{(t,s_i)}\| / \|\boldsymbol{g}_{(t,\pi_i)} + \beta \boldsymbol{g}_{(t,\pi_i')}\| \tag{10}$$

如公式（10）所示，重新聚合过程通过阻尼因子 $\beta \in [0,1]$平衡“收敛”与“探索”成分，生成精炼局部更新 $\boldsymbol{g}_{(t,\pi_i)} + \beta \boldsymbol{g}_{(t,\pi_i')}$。此处引入缩放系数 γ_i，以确保重新聚合梯度保持与原始梯度 $\boldsymbol{g}_{(t,s_i)}$ 相同的下降幅度，同时允许进行方向调整。这种设计意味着，我们的 ECGR方法只改变局部梯度的聚合方向，而不改变其整体更新强度，从而维持优化稳定性以及各客户端之间的一致性。

最后，我们将各客户端重新聚合后的局部梯度输入标准 FedAvg 流程，以获得全局聚合梯度 $\boldsymbol{G}'_t$，并执行全局模型更新 $\boldsymbol{G}'_t$，从而进行全局更新：

$$\boldsymbol{G}'_t = \sum_{i=1}^{N} p_i \boldsymbol{g}'_{(t,s_i)}, \qquad \boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \boldsymbol{G}'_t. \tag{11}$$

此步骤确保所提出的 ECGR 机制与传统联邦优化流程完全兼容，且不会增加额外的同步或通信开销。通过纳入经过方向精炼的局部更新，全局模型将沿着更加稳定、一致的下降轨迹前进，从而有效减轻异质或含噪局部数据带来的不利影响，并加快各轮次的收敛。

3.3 算法描述

算法1： FedAvg-ECGR

要求： 全局轮次总数 T，本地数据集 $\mathcal{D}_i$ ($\boldsymbol{x}_i \in \mathcal{D}_i$)，本地迭代次数 τ_i，初始化权重 $\boldsymbol{w}_0$，初始化顺序 s_i（客户端 i），学习率 $\eta > 0$

对于每轮 $t = 0, \ldots, T-1$ **执行**
  参数服务器将全局模型 $\boldsymbol{w}_t$ 发送给所有参与客户端;
  对于每个客户端 $i = 1, \ldots, N$ **do**
    对于每次本地迭代 $\lambda = 0, 1, \ldots, \tau_i$ **执行**
      初始化本地模型 $\boldsymbol{w}^{\lambda}_{(t,i)} \leftarrow \boldsymbol{w}_t$;
      局部更新 $\boldsymbol{w}^{\lambda+1}_{(t,\ i)} = \boldsymbol{w}^{\lambda}_{(t,\ i)} - \eta \nabla F_i(\boldsymbol{w}^{\lambda}_{(t,\ i)}; \boldsymbol{x}_{s_i})$;
    end
    存储局部梯度集 $\left\{ \eta \nabla F_i(\boldsymbol{w}^{\lambda}_{(t,i)}; \boldsymbol{x}_{s_i}) \right\}_{\lambda=1}^{\tau_i}$;
    ECGR:
      $\pi_i \leftarrow$取最小值 $_{\pi_i \subset s_i} \left\| \boldsymbol{g}_{(t,\pi_i)} \right\|$ # 幅值排序
      $\pi_i' = s_i \setminus \pi_i$, $\beta \boldsymbol{g}_{(t,\pi_i')}$ # 衰减提取
      $\boldsymbol{g}'_{(t,\ s_i)} = \gamma_i(\boldsymbol{g}_{(t,\pi_i)} + \beta \boldsymbol{g}_{(t,\pi_i')})$ # 重新聚合
  end
  参数服务器从所有客户端接收 $\boldsymbol{g}'_{(t,s_i)}$;
  全局聚合 $\boldsymbol{G}'_t = \sum_{i=1}^{N} p_i \boldsymbol{g}'_{(t,\ s_i)}$;
  全局更新 $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \boldsymbol{G}'_t$;
结束
返回 $\boldsymbol{w}_T$;

为验证所提出的**ECGR** 策略的有效性，我们将其融入经典的**FedAvg**算法中，形成算法1所示的 **FedAvg-ECGR** 算法。通过这一整合，我们能够

评估ECGR如何增强标准联邦学习框架中的优化行为。在该框架中，中央服务器协调多个分布式客户端，在不交换原始数据的情况下协同训练共享模型。

在每一全局通信轮次中，参数服务器将当前全局模型 $\boldsymbol{w}_t$ 传输给所有参与客户端。每个客户端 i 在其私有数据集 $\mathcal{D}_i$上执行 τ_i 次本地更新，生成局部梯度序列 $\{\eta\nabla F_i(\boldsymbol{w}_{(t,i)}^{\lambda};\boldsymbol{x}_{s_i})\}_{\lambda=1}^{\tau_i}$ （算法1第1–8行）。

完成本地训练后，客户端执行**ECGR** 过程，通过三步流程细化局部梯度：（1）幅值排序 选择范数较小的梯度以构成子集 π_i，（2）衰减提取以衰减因子 β缩放互补子集 π_i'，（3）重新聚合 将两个子集合并，得到调整后的本地梯度$\boldsymbol{g}'_{(t,s_i)}$ （算法1第9–10行）。随后，每个客户端将 $\boldsymbol{g}'_{(t,s_i)}$ 传输至服务器，服务器执行加权聚合以得到$\boldsymbol{G}'_t$，并据此更新全局模型 $\boldsymbol{w}_{t+1}$ （算法1第11–14行）。该过程重复进行直至收敛，得到最终全局模型 $\boldsymbol{w}_T$。

如图所示，基线 FedAvg 框架对应算法1中的第1–7行和第11–15行。所提出的**ECGR** 机制通过在每个客户端引入额外的本地操作（第8–10行）扩展了该框架，该操作作为一种有效且轻量级的梯度细化步骤。

这种修改带来了两项主要优势 ：

- **通信效率**。 与 FedAvg 相比，*ECGR* 不会产生额外的通信成本，因为每个客户端仍然只向服务器上传单个聚合梯度 $\boldsymbol{g}'_{(t,\ s_i)}$。这一特性对于带宽受限的联邦环境尤为理想。

- **结构兼容性**。 *ECGR* 保留了 FedAvg 的原有结构，包括局部和全局更新流程，从而确保与现有基于 FedAvg 框架的联邦学习系统实现无缝兼容。

然而，*ECGR* 引入了两项额外成本。首先是存储开销：如算法1第8行所示，*ECGR* 的存储需求约为 FedAvg 的 τ_i 倍，因为必须保留所有局部梯度以供选择，而不是在每次更新后将其丢弃。其次是计算开销：第9行中的梯度选择复杂度为 $O(\tau_i!)$，略高于 FedAvg。

尽管如此，*ECGR* 仍遵循联邦学习的核心设计原则——以低成本的本地计算和内存换取客户端与中央服务器之间更低的通信成本。ECGR 在其他最先进的联邦学习算法上的扩展见附录B节。

4 实证评估

本节将在多个广泛使用的图像分类基准和一项真实世界的医学图像诊断任务上，全面评估所提出的 *MAGS* 算法的有效性。我们进一步将其性能与经典及最先进的联邦学习基线方法进行比较，以展示其在准确率和稳定性方面的优势。完整的实现和实验设置已公开发布于 https://github.com/NUDTPingLuo/ECGR，以便于复现和后续研究。

4.1 基准图像分类

数据集与模型设置。我们在 MNIST [LeCun 等人， 1998], Fashion- MNIST（Fashion-MNIST） [Xiao 等人， 2017], CIFAR-10 和 CIFAR-100 [Krizhevsky， 2009]上进行了实验。MNIST 和 Fashion-MNIST 包含 6 万张尺寸为 28×28的灰度训练图像，而 CIFAR-10 和 CIFAR-100 包含 5 万张尺寸为 32×32的 RGB训练图像。每个数据集均包含 1 万张测试图像，用于评估模型性能。**模型**。对于 MNIST 和 Fashion-MNIST，我们采用经典的 LeNet架构 [LeCun 等人， 1998], 该架构由两个卷积层和三个全连接层组成。前向传播过程中，每个卷积层和全连接层之后均应用 ReLU激活，并在卷积层之后进行最大池化。对于 CIFAR-10 和 CIFAR-100，我们使用一种针对 32×32 RGB 图像设计的更深层卷积神经网络（CNN）。该网络由三个卷积块组成，其数量逐渐增加

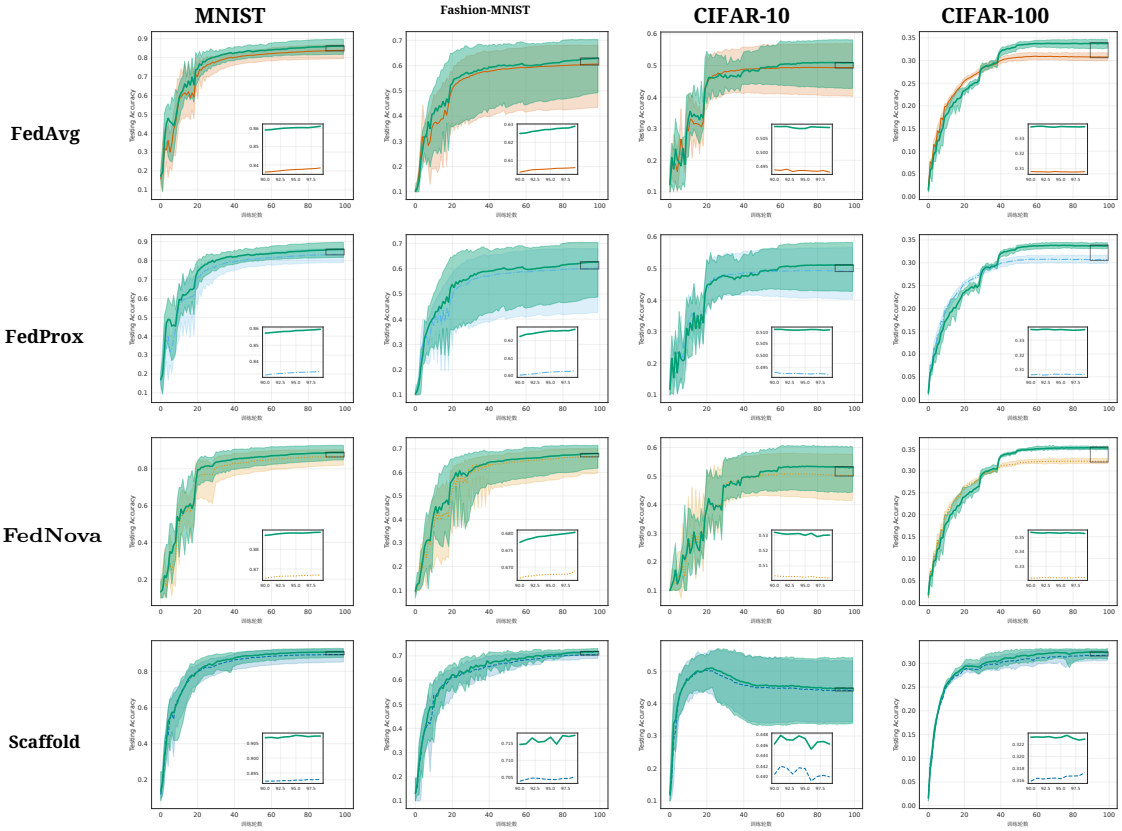


图 2：不同数据集上各联邦学习算法的全局模型测试准确率曲线。每一行对应一种算法，每一列展示特定数据集上的结果。绿色实线表示 ECGR 扩展变体的准确率变化轨迹，其余曲线则表示相应的标准基线。每幅图均显示测试准确率的均值及上下界，这些数据根据使用 5 个不同随机种子的运行结果计算得出。

通道维度（32→64→128→256），每个模块均包含卷积、批归一化、ReLU 激活和最大池化层。最后的特征图会被展平并输入全连接层，以输出类别预测；同时在输出端应用对数 Softmax 函数，以确保数值稳定性。

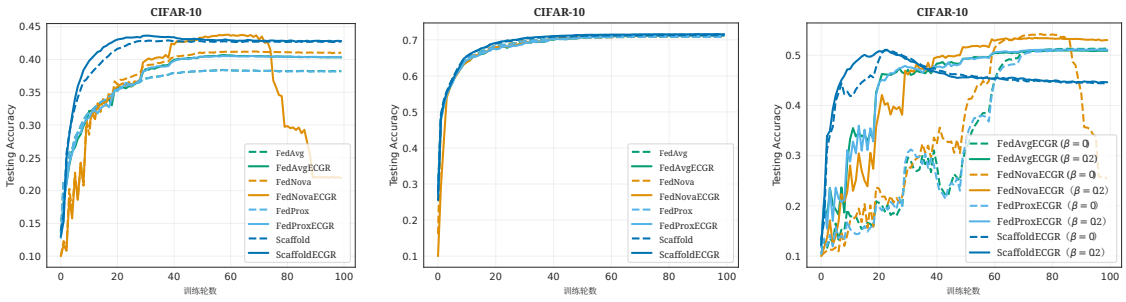
设置。所有实验均在配备 NVIDIA RTX 4070 Ti 图形处理器的工作站上进行，模拟 10 个联邦客户端。每个客户端在每个全局轮次中使用其本地数据集训练一个轮次。每个客户端上的本地训练数据均根据狄利克雷分布从总训练集中采样，以模拟极端的非独立同分布场景，其浓度参数 α 设置为 0。01。对于每个数据集，使用五个不同的随机种子（0、1、42、999 和 2025）生成不同的狄利克雷划分，以确保所报告结果的统计可靠性。每个客户端至少需要包含两个批次的数据，以确保能够计算至少两个局部梯度用于选择。所有模型均使用带动量 0。9 的随机梯度下降（SGD）进行训练，全局训练轮次总数设置为 $T = 100$。所有数据集均采用一致的训练超参数配置：学习率初始值为 0。001，每 10 个轮次衰减为原来的二分之一；所有实验的批次大小均固定为 128。

基线方法。我们将所提出的方法与几种具有代表性的联邦学习基线方法进行比较，包括 Fe- dAvg [McMahan 等人、 2017], FedProx [Li 等人、 2020], FedNova [Wang 等人、 2020] 以及 Scaffold [Karim- ireddy 等人 2020]。FedAvg 在客户端之间执行简单的模型平均。FedProx 通过在本地目标函数中添加近端项扩展了 FedAvg，从而在统计异质性条件下稳定训练。Fed- Nova 进一步对局部更新进行归一化，以消除由不同的局部 epoch 数量所导致的目标不一致性。Scaffold 通过引入控制变量来校正局部更新，从而缓解非独立同分布场景下的客户端漂移。为进行公平比较，所有方法均在相同的实验设置下进行训练，相应的扩展变体见第 B 节。

结果与讨论。我们在 ECGR 策略中采用 $\beta = 0.2$ 阻尼因子，图 2 展示了其在测试数据集上的性能。以下是几个值得注意的观察结果：(i) 对于最终的全局模型准确率，ECGR 在所有选定的数据集和基线方法上均持续提升性能，

相较于对应的基线方法，准确率绝对提升约 1%–2%，从而证明了其整体有效性。(ii) ECGR 的收敛轨迹与对应基线方法的轨迹高度一致，这表明 ECGR 沿着各方法原有的轨迹优化优化过程，与算法1中所述的流程逻辑一致。

4.2 消融实验



(a) $\eta = 0.0001$ (b) $\alpha = 1$ (c) $\beta = 0$对比 $\beta = 0.2$

图3：在 CIFAR-10 上针对学习率 η、数据异质性水平 α和 ECGR 阻尼系数 β开展的消融研究。所有曲线均报告了五次独立运行的平均测试准确率，随机种子分别为 0、1、42、999 和 2025年。

我们对 ECGR 进行了消融实验，以考察不同因素如何影响其在 CIFAR-10数据集上的性能。除非另有说明，所有超参数均与第 4.1 节中的设置保持一致，仅消融实验相关设置除外。更多消融实验结果见附录 C。

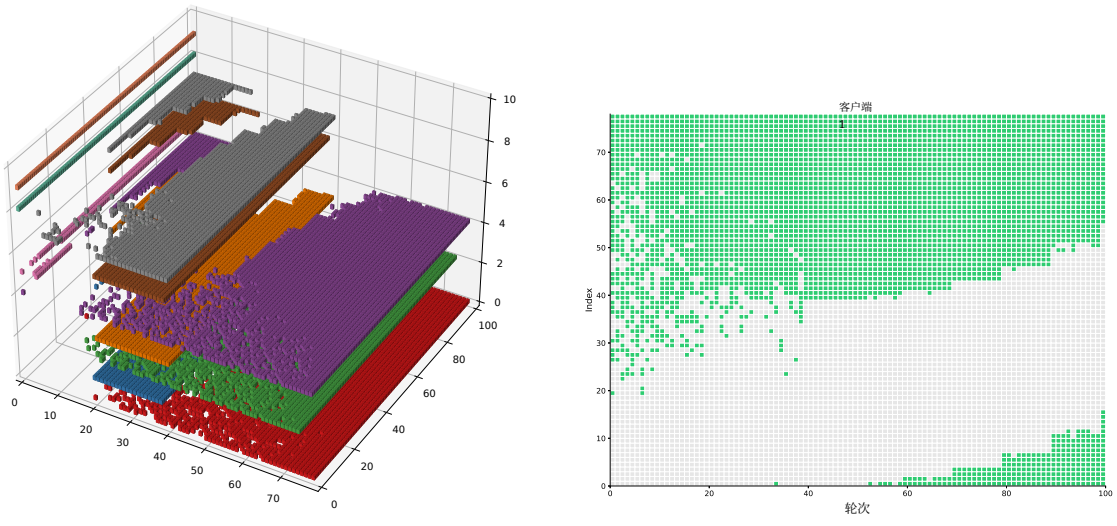
学习率敏感性。 如图 3(a)所示，所有基线方法达到的测试准确率上限始终低于图 2 中的结果，这表明在此设置下优化机制更具挑战性。在这种情况下，将 ECGR 与 FedAvg 和 FedProx 集成后，最终测试准确率仍有显著提升；而 FedNova 则因越过全局最优点而导致性能下降

；Scaffold-ECGR 仅有小幅提升。这可以归因于 FedAvg 和 FedProx 不会显式调整本地梯度或全局梯度，因此与 ECGR 机制天然兼容。相比之下，FedNova 会进行梯度重缩放，这与 ECGR 中的衰减提取步骤部分重叠，导致优化越过全局最优点时，CIFAR-10 上的准确率在最后几个轮次迅速下降。此外，Scaffold 引入控制变量来校正本地更新，这会干扰 ECGR 的幅值排序机制，因此性能提升相对有限。

独立同分布与非独立同分布。如图 3(b)所示，图中展示了狄利克雷分布下$\alpha = 1$ 时的收敛曲线，此时各客户端的局部数据分布接近独立同分布设置。在这种情况下，基线方法及其对应的 ECGR 增强变体的收敛行为高度一致。与图 2中的结果相比，这些观察表明，ECGR 在独立同分布条件下保留了正常的训练动态，同时有效提升了非独立同分布设置下的收敛性能。

丢弃与提取。我们进一步研究了提取操作在所提出的 ECGR 策略中的作用。在 ECGR 中，提取步骤是处理探索性梯度的关键组成部分，其强度由阻尼因子 β控制。根据公式（10），当 β 趋近于 1 时，ECGR 的效果会逐渐减弱。因此，我们比较了两种具有代表性的设置：直接丢弃探索性梯度（$\beta = 0$）和使用适中的阻尼因子应用衰减提取（$\beta = 0.2$）。如图 3(c)所示，当 $\beta = 0$时，ECGR 策略下的测试准确率曲线表现出较慢的早期收敛速度和较大的振荡。此外，FedNova-ECGR 往往会越过全局最优点，这与图 3(a)中观察到的情况相似。这些结果表明，衰减提取机制对于提高收敛稳定性至关重要。

4.3 客户端每轮梯度选择的可视化



(a) 全局三维概览。　　(b) 代表性客户端 1。

图 4：CIFAR-10数据集上 ECGR 每轮梯度选择的可视化。(a) 展示所有客户端梯度选择模式的全局三维概览。每个立方体表示在特定索引（x 轴）和训练轮次（y 轴）下选取的梯度，客户端维度沿 z 轴编码。(b) 展示代表性节点客户端 1 的选择行为细节。

设置。我们展示了所提出的 ECGR 策略在 CIFAR-10数据集上的幅值排序操作。具体流程如下：(i) 根据训练顺序，为客户端训练过程中获得的局部梯度编制索引；(ii) 根据幅值排序机制，将编制索引的梯度分为探索性梯度和收敛梯度；(iii) 突出显示与收敛梯度对应的索引并将其可视化，以展示 ECGR 的梯度选择行为。

结果与讨论。图 4 展示了三维可视化结果，以及代表单个客户端的相应二维可视化。我们观察到，收敛梯度通常出现在本地训练的后期阶段，这揭示了一个值得关注的现象：在基于随机梯度下降的优化过程中，训练过程本质上伴随着一个初始探索阶段，随后进入收敛阶段。这种行为与图 1 所示的群体智能模式十分相似。此外，这些观察结果进一步验证了 ECGR 的核心原则，即强调收敛梯度的主导作用，同时有效利用探索梯度中包含的信息，以增强联邦学习优化过程。

4.4 组织病理图像分析

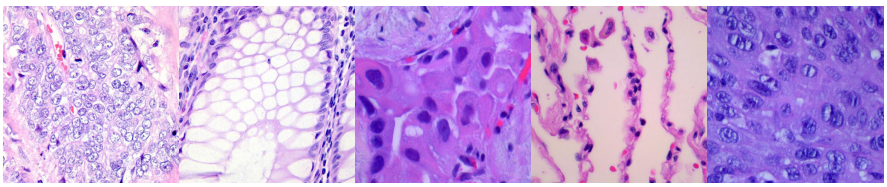


图 5：LC25000数据集中的代表性图像块。从左到右依次为：结肠腺癌、结肠良性组织、肺腺癌、肺良性组织和肺鳞状细胞癌。

在本实验中，我们在非独立同分布数据分布设置下，使用 LC25000 数据集 [Borkowski et al.，2019] 评估了 ECGR，以更好地反映现实的临床部署场景。

数据集。 我们采用了 LC25000 数据集，这是一个公开的结肠和肺组织病理图像块档案。该数据集共包含 25,000 个彩色图像块，均匀分布于五个类别：结肠腺癌、结肠良性组织、肺腺癌、肺良性组织和肺鳞状细胞癌 [Borkowski et al.，2019]。所有图像的尺寸均为 224×224 像素。在本研究中，我们按类别整理数据，以创建模拟联邦学习环境的客户端数据集。每个类别的图像块按约 80/20 的比例划分为训练集和测试集，并确保来自同一源图像的样本被划分到同一个子集中，以避免数据泄漏。

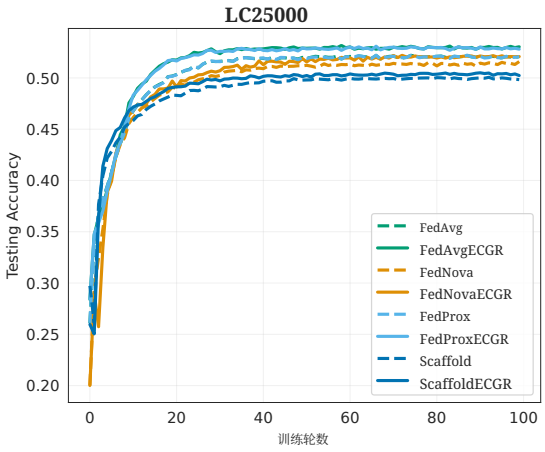
模型。对于所有方法，我们均采用标准的 ResNet-18 神经网络架构 [He 等人提出的， 2016], 由 TorchVision 软件包 [实现， 2016],并使用随机初始化的权重。

实验设置。本节的实验设置与第 4.1 节一致，但衰减提取的阻尼因子 β 除外。根据第 4.2 节中的观察结果，为*FedNova-ECGR* 赋予较大的值 β （即 $\beta = 0.5$），而所有其他基于 ECGR 的基线方法均采用统一设置 $\beta = 0.2$。

结果。如图 6 所示，所提出的 ECGR 策略应用于医学数据集和大规模模型时仍然有效。尽管将 ECGR 集成到各种基线方法中所带来的性能提升相对有限（范围为 0.4% 至 1%），但 ECGR 仍提供了一种通用且切实可行的优化机制，能够在不同设置下持续提升联邦训练效果。

5 结论与未来工作

在本研究中，我们从梯度层面视角考察了统计异质性条件下联邦学习面临的优化挑战。我们发现，局部梯度是非独立同分布数据导致客户端漂移的主要机制，并据此提出了一种通用的客户端优化框架。该框架完全基于本地梯度集合运行，无需修改联邦通信协议，也不会增加通信开销。在此框架下，我们提出了 ECGR——一种受群体智能启发的梯度重新聚合策略。该策略将局部梯度分解为探索性组件和收敛性组件，并调整二者的贡献，以生成更稳定、更鲁棒的更新。理论分析和大量实证结果均表明，ECGR 能够有效缓解客户端漂移，并可无缝集成到现有联邦学习算法中。



算法	最终准确率	最佳准确率
FedAvg	0.5209	0.6150
FedAvg-ECGR	**0.5304**	**0.6186**
FedNova	0.5160	0.6351
FedNova-ECGR	**0.5208**	**0.6399**
FedProx	0.5207	0.6154
FedProx-ECGR	**0.5281**	**0.6210**
Scaffold	0.4984	0.6387
Scaffold-ECGR	**0.5024**	**0.6415**

图6：使用 ResNet-18模型在 LC25000数据集上进行联邦学习的结果。左图：各基线及其对应的ECGR增强变体的测试准确率曲线，取前文定义的五个随机种子下的平均值。右表：全局模型在最后一轮的平均测试准确率，以及训练过程中各基线及其对应 ECGR变体达到的最高准确率。

我们的梯度分解视角与越来越多有选择地利用局部更新中有效组件的既有研究密切相关，并得到这些研究的支持。例如，AdaComp [Chen 等人， 2018] 自适应地仅传输最显著的梯度元素，从而在保持优化保真度的同时减少通信开销。同样， [Sattler 等人， 2019] 通过仅上传幅值较大的梯度来筛选本地更新，从而在异质性条件下突出关键更新。除了梯度压缩之外，FedSkip [Fan 等人， 2022] 将客户端更新分解为全局共享组件和本地特定组件；FedPer [Arivazhagan 等人， 2019] 则将神经网络拆分为共享基础层和个性化层，并仅在客户端之间聚合前者。这些基于不同动机提出的代表性方法共同表明，有选择地提炼、划分或重新加权本地更新，是应对联邦学习中异质性的一条可行且有效的途径，也为本研究探讨的核心思想提供了独立验证。

展望未来，我们认识到，ECGR采用的梯度划分策略只是更广阔设计空间中的一种粗略实现。理论上可能存在一种最优方法，能够划分并重新组合本地梯度集合，从而更精确地平衡稳定性、偏差与收敛效率。未来研究的一个重要方向是对这种最优性进行形式化，并开发有原则的梯度分解与重新聚合机制，最终目标是在异构数据条件下缩小联邦学习与集中式学习之间的性能差距。更广泛地说，我们希望本研究提出的视角——将局部梯度视为具有结构的对象，而非不可分割的更新——能够启发联邦优化、分布式学习及相关领域的进一步探索。

数据可用性

所有使用的数据集均可公开获取。MNIST [LeCun 等人， 1998], Fashion-MNIST [Xiao 等人， 2017],CIFAR-10 和 CIFAR-100 [Krizhevsky， 2009] 是机器学习图像分类中常用的基准数据集。LC25000数据集共包含 25,000 个彩色图像块，平均分布于五个类别：结肠腺癌、结肠良性组织、肺腺癌、肺良性组织和肺鳞状细胞癌 [Borkowski 等人， 2019]。

代码可用性

Ping Luo 已提供所提出框架的 Python 代码（URL: https://github.com/NUDTPingLuo/ECGR）。

参考文献

Mohammed Adnan、Shivam Kalra、Jesse C Cresswell、Graham W Taylor 和 Hamid R Tizhoosh。用于医学图像分析的联邦学习和差分隐私。*Scientific Reports*，12(1):1953，2022年。Muntasir Al-Asfoor、Mohammed Hamzah Alsalihi 和 Kevin Maher。基于联邦学习的脑肿瘤分类。载于*2024*年第十届优化与应用国际会议（*ICOA*），第 1–4 页。IEEE，2024年。Rodolfo Stoffel Antunes、Cristiano Andr´e da Costa、Arne K¨uderle、Imrana Abdullahi Yari 和 Bj¨orn Eskofier。医疗保健领域的联邦学习：系统综述与架构提案。*ACM Transactions on Intelligent Systems and Technology*（*TIST*），13(4):1–23，2022年。Manoj Ghuhan Arivazhagan、Vinay Aggarwal、Aaditya Kumar Singh 和 Sunav Choudhary。带有个性化层的联邦学习。*arXiv* 预印本 *arXiv:1912.00818*，2019年。Andrew A. Borkowski、Marilyn M. Bui、L. Brannon Thomas、Catherine P. Wilson、Lauren A. DeLand 和 Stephen M. Mastorides。肺癌和结肠癌组织病理图像数据集（lc25000），2019年。

Chia-Yu Chen、Jungwook Choi、Daniel Brand、Ankur Agrawal、Wei Zhang 和 Kailash Gopalakrishnan。AdaComp：面向数据并行分布式训练的自适应残差梯度压缩。载于*AAAI*人工智能大会论文集，第 32 卷，2018年。

Ittai Dayan、Holger R Roth、Aoxiao Zhong、Ahmed Harouni、Amilcare Gentili、Anas Z Abidin、Andrew Liu、Anthony Beardsworth Costa、Bradford J Wood、Chien-Sung Tsai 等。用于预测新冠肺炎患者临床结局的联邦学习。*Nature Medicine*，27(10)：1735–1743，2021年。

Alireza Fallah、Aryan Mokhtari 和 Asuman Ozdaglar。个性化联邦学习：一种元学习方法。*arXiv* 预印本 *arXiv:2002.07948*，2020年。

Ziqing Fan、Yanfeng Wang、Jiangchao Yao、Lingjuan Lyu、Ya Zhang 和 Qi Tian。Fedskip：利用联邦跳跃聚合应对统计异质性。在*2022*年 *IEEE*国际数据挖掘会议（*ICDM*），第131–140页。IEEE，2022年。

Bao Feng、Jiangfeng Shi、Liebin Huang、Zhiqi Yang、Shi-Ting Feng、Jianpeng Li、Qinxian Chen、Huimin Xue、Xiangguang Chen、Cuixia Wan 等。用于识别胃癌术后复发高危患者的鲁棒联邦学习模型。《style id='1'>自然·通讯，15(1):742，2024年。

Avishek Ghosh、Jichan Chung、Dong Yin 和 KannanRamchandran。聚类联邦学习的高效框架。《style id='1'>IEEE信息论汇刊，68(12):8076–8091，2022年。

Jialiang Han、Yudong Han、Xiang Jing、Gang Huang 和 Yun Ma。Degafl：面向跨机构联邦学习的去中心化梯度聚合。《style id='1'>IEEE并行与分布式系统汇刊，2024年。

何恺明、Xiangyu Zhang、任少卿和孙剑。用于图像识别的深度残差学习。发表于*IEEE*计算机视觉与模式识别会议论文集（*CVPR*），2016年6月。

Tzu-Ming Harry Hsu、Hang Qi 和 Matthew Brown。衡量非相同数据分布对联邦视觉分类的影响。*arXiv* 预印本 *arXiv:1909.06335*，2019年。Peter Kairouz、H Brendan McMahan、Brendan Avent、Aur´elien 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。Rohan Reddy Kalavakonda、Peyman Dehghanzadeh、Junjun Huan、Soumyajit Mandal 和 Swarup Bhunia。融合智能：融合自然智能与人工智能的范式。*IEEE*物联网期刊，12(15):30548–30563，2025年。doi：10.1109/JIOT.2025.3572367。

Sai Praneeth Karimireddy、Satyen Kale、Mehryar Mohri、Sashank Reddi、Sebastian Stich 和 Ananda Theertha Suresh。Scaffold：联邦学习的随机控制平均。载于国际机器学习会议，第5132–5143页。PMLR，2020年。Alex Krizhevsky。从 Tiny Images 学习多层特征。技术报告，多伦多大学，多伦多，2009年。Yann LeCun、L´eon Bottou、Yoshua Bengio 和 Patrick Haffner。应用于文档识别的基于梯度的学习。《*IEEE*论文集》，86(11):2278–2324，1998年。Edward H Lee、Michelle Han、Jason Wright、Michael Kuwabara、Jacob Mevorach、Gang Fu、Olivia Choud-hury、Ujjwal Ratan、Michael Zhang、Matthias W Wagner 等。介绍一种用于儿童脑肿瘤的联邦学习人工智能平台的国际研究。《自然·通讯》，15(1):7615，2024年。Anran Li、Lan Zhang、Juntao Tan、Yaxuan Qin、Junhao Wang 和 Xiang-Yang Li。用于联邦学习的样本级数据选择。载于*IEEE INFOCOM 2021——IEEE*计算机通信会议，第1–10页，2021a。doi：10.1109/INFOCOM42981.2021.9488723。Bin Li、Yin Li 和 Kevin W Eliceiri。用于全视野切片图像分类的双流多实例学习网络与自监督对比学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第14318–14328页，2021b。Chenning Li、Xiao Zeng、Mi Zhang 和 Zhichao Cao。PyramidFL：用于高效联邦学习的细粒度客户端选择框架。载于第*28*届年度国际移动计算与网络会议论文集，第158–171页，2022年。Tian Li、Anit Kumar Sahu、Manzil Zaheer、Maziar Sanjabi、Ameet Talwalkar 和 Virginia Smith。异构网络中的联邦优化。《机器学习与系统论文集》，2:429–450，2020年。Zexi Li、Tao Lin、Xinyi Shang 和 Chao Wu。重新审视神经网络联邦学习中的加权聚合。载于国际机器学习会议，第19767–19788页。PMLR，2023年。Ming Y Lu、Drew FK Williamson、Tiffany Y Chen、Richard J Chen、Matteo Barbieri 和 Faisal Mahmood。全视野切片图像上的数据高效弱监督计算病理学。《自然·生物医学工程》，5(6):555–570，2021年。Ming Y Lu、Richard J Chen、Dehan Kong、Jana Lipkova、Rajendra Singh、Drew FK Williamson、Tiffany Y Chen 和 Faisal Mahmood。面向十亿像素全视野切片图像的计算病理学联邦学习。医学图像分析，76:102298，2022a。Yucheng Lu、Wentao Guo 和 Christopher M De Sa。Grab：寻找优于随机重排且具有可证明性的更佳数据排列。《神经信息处理系统进展》，35:8969–8981，2022b。Ping Luo、Xiaoge Deng、Ziqing Wen、Tao Sun 和 Dongsheng Li。Bherd：通过选择有益的本地梯度集合加速联邦学习。*IEEE*计算机汇刊，2025年。Xiaodong Ma、Jia Zhu、Zhihao Lin、Shanxuan Chen 和 Yangjie Qin。解决联邦学习中非独立同分布数据问题的最新综述。《未来一代计算机系统》，135:244–258，2022年。H. Brendan McMahan、Eider Moore、Daniel Ramage、Seth Hampson 和 Blaise Aguera y Arcas。基于去中心化数据的通信高效深度网络学习。载于第*20*届国际人工智能与统计会议（*AISTATS*）论文集，2017年。Takayuki Nishio 和 Ryo Yonetani。移动边缘中具有异构资源的联邦学习客户端选择。载于*ICC 2019——2019*年*IEEE*国际通信会议（*ICC*），第1–7页。IEEE，2019年。Krishna Pillutla、Sham M Kakade 和 Zaid Harchaoui。联邦学习的鲁棒聚合。《*IEEE*信号处理汇刊》，70:1142–1154，2022年。

Adam Sadilek、Luyang Liu、Dung Nguyen、Methun Kamruzzaman、Stylianos Serghiou、Benjamin Rader、Alex Ingerman、Stefan Mellem、Peter Kairouz、Elaine O Nsoesie 等。《基于联邦学习的隐私优先健康研究》。*NPJ*数字医学，4(1):132，2021年。Felix Sattler、Simon Wiedemann、Klaus-Robert M¨uller 和 Wojciech Samek。《基于非独立同分布数据的鲁棒且通信高效的联邦学习》。*IEEE*神经网络与学习系统汇刊，31(9):3400–3413，2019年。Philip Schutte、Valentina Corbetta、Regina Beets-Tan 和 Wilson Silva。《FedGS：用于异质医学图像分割的联邦梯度缩放》。载于国际医学图像计算与计算机辅助干预会议，第246–255页。施普林格，2024年。Nurfaidah Tahir、Chau-Ren Jung、Shin-Da Lee、Nur Azizah、Wen-Chao Ho 和 Tsai-Chung Li。《基于联邦学习的死亡率预测模型：系统综述与荟萃分析》。《医学互联网研究杂志》，27:e65708，2025年。Minxue Tang、Xuefei Ning、Yitu Wang、Jingwei Sun、Yu Wang、Hai Li 和 Yiran Chen。《FedCor：面向异质联邦学习的基于相关性的主动客户端选择策略》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10102–10111页，2022年。Valery Tereshko 和 Andreas Loengarov。《蜜蜂觅食动态中的集体决策》。《计算与信息系统》，9(3):1，2005年。TorchVision。《TorchVision：PyTorch 的计算机视觉库》。https://github.com/pytorch/vision，2016年。Jianyu Wang、Qinghua Liu、Hao Liang、Gauri Joshi 和 H Vincent Poor。《解决异质联邦优化中的目标不一致问题》。《神经信息处理系统进展》，33:7611–7623，2020年。Jinxi Xiang、Xiyue Wang、Xinran Wang、Jun Zhang、Sen Yang、Wei Yang、Xiao Han 和 Yueping Liu。《利用全切片图像上的弱监督学习自动诊断和分级前列腺癌》。《生物学与医学中的计算机》，152:106340，2023年。Han Xiao、Kashif Rasul 和 Roland Vollgraf。《Fashion-MNIST：用于机器学习算法基准测试的新型图像数据集》。*arXiv*预印本 *arXiv:1708.07747*，2017年。Jie Zhang、Song Guo、Zhihao Qu、Deze Zeng、Yufeng Zhan、Qifeng Liu 和 Rajendra Akerkar。《资源受限条件下针对非独立同分布数据的自适应联邦学习》。*IEEE*计算机汇刊，71(7):1655–1667，2021年。Hao Zhou、Hua Dai、Siqi Cai、Geng Yang 和 Yang Xiang。《海报：面向异质联邦学习的个性化差分隐私自适应梯度裁剪》。载于*2025*年*ACM SIGSAC*计算机与通信安全会议论文集，第4740–4742页，2025年。

附录

A ECGR 梯度误差降低的假设、定义与定理

我们对这一论点进行了形式化论证：ECGR方法能够缩小局部梯度与理论最优梯度之间的差异。其核心洞见在于，ECGR的重新聚合能够抑制局部梯度方差，同时控制诱导偏差，从而得到更接近真实最优更新方向的局部梯度估计。

A.1 联邦学习优化目标

假设 1（*L*-光滑性）若可微函数 $F\colon \mathbb{R}^d \to \mathbb{R}$ 的梯度是 *L-Lipschitz* 连续的，则称其为 L-光滑，即：

$$\|\nabla F(\boldsymbol{w}) - \nabla F(\boldsymbol{w}')\| \le L\|\boldsymbol{w} - \boldsymbol{w}'\|, \quad \forall \boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d.$$

定理 1 在假设 *1* 和定义 *1*（其中 $L > 1$）成立的条件下，*FedAvg* 的优化目标可定量表述为：对所有客户端 i，最小化局部梯度 $\boldsymbol{g}_{(t,s_i)}$ 与期望全局梯度 $\nabla F(\boldsymbol{w}_t)$ 之间的偏差。

证明：在轮次 $t+1$，根据标准的 L-光滑性不等式，FedAvg 方法下的全局损失函数 $F(\boldsymbol{w}_{t+1})$ 可通过泰勒定理展开为

$$\begin{aligned}
F(\boldsymbol{w}_{t+1}) &\le F(\boldsymbol{w}_t) - \langle \nabla F(\boldsymbol{w}_t), \boldsymbol{w}_{t+1} - \boldsymbol{w}_t \rangle + \frac{L}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2 \\
&= F(\boldsymbol{w}_t) - \langle \nabla F(\boldsymbol{w}_t), \boldsymbol{G}_t \rangle + \frac{L}{2}\|\boldsymbol{G}_t\|^2 \\
&= F(\boldsymbol{w}_t) - \frac{1}{2}\|\nabla F(\boldsymbol{w}_t)\|^2 - \frac{1}{2}\|\boldsymbol{G}_t\|^2 + \frac{1}{2}\|\boldsymbol{G}_t - \nabla F(\boldsymbol{w}_t)\|^2 + \frac{L}{2}\|\boldsymbol{G}_t\|^2 \\
&= F(\boldsymbol{w}_t) - \frac{1}{2}\|\nabla F(\boldsymbol{w}_t)\|^2 + \frac{1}{2}\|\boldsymbol{G}_t - \nabla F(\boldsymbol{w}_t)\|^2 + \frac{L-1}{2}\|\boldsymbol{G}_t\|^2 \\
&\le F(\boldsymbol{w}_t) - \frac{1}{2}\|\nabla F(\boldsymbol{w}_t)\|^2 + \frac{1}{2}\sum_{i=1}^{N} p_i \|\boldsymbol{g}_{(t,s_i)} - \nabla F(\boldsymbol{w}_t)\|^2 + \frac{L-1}{2}\sum_{i=1}^{N} p_i \|\boldsymbol{g}_{(t,s_i)}\|^2
\end{aligned}$$

最后一个不等式由詹森不等式得出。上述表达式建立了轮次（$t+1$）中损失函数的上界。根据假设 1，∇F（$\boldsymbol{w}_t$）仅取决于整个数据集 D 和当前全局模型 $\boldsymbol{w}_t$，因此在轮次 t期间保持为一个固定但未知的值。为了收紧这一上界，有必要最小化以下各项：$\|\boldsymbol{g}_{(t,s_i)} - \nabla F_i(\boldsymbol{w}_t)\|^2$ 以及 $\|\boldsymbol{g}_{(t,s_i)}\|^2$。根据 ℓ_2 范数的定义，第一项衡量局部梯度与期望梯度之间的偏差，而第二项反映局部梯度的大小。这两个量本质上是相互耦合的。

因此，一种简单而有效的方法是在减小每个局部梯度与期望梯度之间偏差的同时，保持其大小，从而提高整体优化过程的稳定性和收敛性。

■

A.2 保持局部梯度大小

定理 2（梯度幅值保持）给定 *ECGR* 的重新聚合公式为 $\boldsymbol{g}'_{(t,s_i)} = \gamma_i(\boldsymbol{g}_{(t,\pi_i)} + \beta \boldsymbol{g}_{(t,\pi'_i)})$，其中 $\gamma_i = \|\boldsymbol{g}_{(t,s_i)}\| / \|\boldsymbol{g}_{(t,\pi_i)} + \beta \boldsymbol{g}_{(t,\pi'_i)}\|$，重新聚合后的局部

梯度与原始局部梯度的幅值保持不变：

$$\begin{aligned}\|\boldsymbol{g}'_{(t,s_i)}\|^2 &= \|\gamma_i(\boldsymbol{g}_{(t,\pi_i)} + \beta\boldsymbol{g}_{(t,\pi'_i)})\|^2 \\ &= \|(\boldsymbol{g}_{(t,\pi_i)} + \beta\boldsymbol{g}_{(t,\pi'_i)})\|^2 \frac{\|\boldsymbol{g}_{(t,s_i)}\|^2}{\|\boldsymbol{g}_{(t,\pi_i)} + \beta\boldsymbol{g}_{(t,\pi'_i)}\|^2} \\ &= \|\boldsymbol{g}_{(t,s_i)}\|^2.\end{aligned}$$

该定理表明，ECGR 保留了每个局部聚合梯度的大小（即“长度”），从而确保 FedAvg 的优化动态不会受到扭曲。

A.3 ECGR 的误差界缩减

在本小节中，我们证明 ECGR 会减小局部梯度与全局真实梯度之间的偏差，即 $\|\boldsymbol{g}'_{(t,s_i)} - \nabla F(\boldsymbol{w}_t)\|^2 < \|\boldsymbol{g}_{(t,s_i)} - \nabla F(\boldsymbol{w}_t)\|^2$。这一结果需要若干额外假设、精确定义和中间引理，下面将逐一介绍并证明。

定义 1（梯度记号）为简化分析，我们引入以下定义 .

- $\boldsymbol{a} = \boldsymbol{g}_{(t,\pi_i)}$：收敛梯度
- $\boldsymbol{b} = \boldsymbol{g}_{(t,\pi'_i)}$：探索梯度
- $\boldsymbol{\mu} = \nabla F(\boldsymbol{w}_t)$：真实梯度
- $\boldsymbol{c} = \boldsymbol{a} + \boldsymbol{b}$：原始聚合梯度
- $\boldsymbol{v} = \boldsymbol{a} + \beta\boldsymbol{b}$：*ECGR*组合梯度
- $\gamma = \frac{\|\boldsymbol{c}\|}{\|\boldsymbol{v}\|}$：缩放因子

定义 2（方向一致性）对于任意向量 $\boldsymbol{x}$、$\boldsymbol{z}$，定义方向一致性函数：

$$Align(\boldsymbol{x}, \boldsymbol{z}) = \frac{\langle \boldsymbol{x}, \boldsymbol{z} \rangle}{\|\boldsymbol{x}\|\|\boldsymbol{z}\|}$$

关键假设

假设2（收敛梯度优越性）

$$\theta_a = \angle(\boldsymbol{a}, \boldsymbol{\mu}) < \theta_b = \angle(\boldsymbol{b}, \boldsymbol{\mu})$$

等价地，$Align$（$\boldsymbol{a}$、$\boldsymbol{\mu}$）$> Align$（$\boldsymbol{b}$、$\boldsymbol{\mu}$）

核心引理和详细证明

引理1（方向一致性单调性引理）对于任意向量 $\boldsymbol{x}$、$\boldsymbol{y}$、$\boldsymbol{z}$，如果：

$$\frac{\langle \boldsymbol{x}, \boldsymbol{z} \rangle}{\|\boldsymbol{x}\|} > \frac{\langle \boldsymbol{y}, \boldsymbol{z} \rangle}{\|\boldsymbol{y}\|}$$

则函数：

$$f(\beta) = \frac{\langle \boldsymbol{x} + \beta\boldsymbol{y}, \boldsymbol{z} \rangle}{\|\boldsymbol{x} + \beta\boldsymbol{y}\|}$$

在 $[0, 1]$ 上严格递减，适用于 $\beta < 1$。

证明： 步骤 *1*：函数定义和导数计算

设 $\boldsymbol{u}(\beta) = \boldsymbol{x} + \beta \boldsymbol{y}$，则：

$$f(\beta) = \frac{\langle \boldsymbol{u}(\beta), \boldsymbol{z} \rangle}{\|\boldsymbol{u}(\beta)\|}$$

计算导数：

$$f'(\beta) = \frac{d}{d\beta} \left(\frac{\langle \boldsymbol{u}, \boldsymbol{z} \rangle}{\|\boldsymbol{u}\|} \right)$$

使用商法则：

$$f'(\beta) = \frac{\langle \boldsymbol{y}, \boldsymbol{z} \rangle \|\boldsymbol{u}\| - \langle \boldsymbol{u}, \boldsymbol{z} \rangle \cdot \frac{d}{d\beta}\|\boldsymbol{u}\|}{\|\boldsymbol{u}\|^2}$$

其中：

$$\frac{d}{d\beta}\|\boldsymbol{u}\| = \frac{d}{d\beta}(\langle \boldsymbol{u}, \boldsymbol{u} \rangle^{1/2}) = \frac{1}{2}\langle \boldsymbol{u}, \boldsymbol{u} \rangle^{-1/2} \cdot 2\langle \boldsymbol{u}, \boldsymbol{y} \rangle = \frac{\langle \boldsymbol{u}, \boldsymbol{y} \rangle}{\|\boldsymbol{u}\|}$$

代入：

$$f'(\beta) = \frac{\langle \boldsymbol{y}, \boldsymbol{z} \rangle \|\boldsymbol{u}\| - \langle \boldsymbol{u}, \boldsymbol{z} \rangle \cdot \frac{\langle \boldsymbol{u}, \boldsymbol{y} \rangle}{\|\boldsymbol{u}\|}}{\|\boldsymbol{u}\|^2} = \frac{\langle \boldsymbol{y}, \boldsymbol{z} \rangle \|\boldsymbol{u}\|^2 - \langle \boldsymbol{u}, \boldsymbol{z} \rangle \langle \boldsymbol{u}, \boldsymbol{y} \rangle}{\|\boldsymbol{u}\|^3}$$

设分子为：

$$M(\beta) = \langle \boldsymbol{y}, \boldsymbol{z} \rangle \|\boldsymbol{u}\|^2 - \langle \boldsymbol{u}, \boldsymbol{z} \rangle \langle \boldsymbol{u}, \boldsymbol{y} \rangle$$

由于分母 $\|\boldsymbol{u}\|^3 > 0$，$f'(\beta)$ 的符号由 $M(\beta)$ 决定。

步骤 *2*：分析 $M(1)$ 的符号

在 $\beta = 1$，$\boldsymbol{u} = \boldsymbol{x} + \boldsymbol{y}$ 时，我们有：

$$M(1) = \langle \boldsymbol{y}, \boldsymbol{z} \rangle \|\boldsymbol{x} + \boldsymbol{y}\|^2 - \langle \boldsymbol{x} + \boldsymbol{y}, \boldsymbol{z} \rangle \langle \boldsymbol{x} + \boldsymbol{y}, \boldsymbol{y} \rangle$$

展开所有项：

$$\begin{aligned} M(1) = {} & \langle \boldsymbol{y}, \boldsymbol{z} \rangle (\|\boldsymbol{x}\|^2 + 2\langle \boldsymbol{x}, \boldsymbol{y} \rangle + \|\boldsymbol{y}\|^2) \\ & - (\langle \boldsymbol{x}, \boldsymbol{z} \rangle + \langle \boldsymbol{y}, \boldsymbol{z} \rangle)(\langle \boldsymbol{x}, \boldsymbol{y} \rangle + \|\boldsymbol{y}\|^2) \end{aligned}$$

完全展开：

$$\begin{aligned} M(1) = {} & \langle \boldsymbol{y}, \boldsymbol{z} \rangle \|\boldsymbol{x}\|^2 + 2\langle \boldsymbol{y}, \boldsymbol{z} \rangle \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \langle \boldsymbol{y}, \boldsymbol{z} \rangle \|\boldsymbol{y}\|^2 \\ & - \langle \boldsymbol{x}, \boldsymbol{z} \rangle \langle \boldsymbol{x}, \boldsymbol{y} \rangle - \langle \boldsymbol{x}, \boldsymbol{z} \rangle \|\boldsymbol{y}\|^2 \\ & - \langle \boldsymbol{y}, \boldsymbol{z} \rangle \langle \boldsymbol{x}, \boldsymbol{y} \rangle - \langle \boldsymbol{y}, \boldsymbol{z} \rangle \|\boldsymbol{y}\|^2 \end{aligned}$$

合并同类项：

$$\begin{aligned} M(1) = {} & \langle \boldsymbol{y}, \boldsymbol{z} \rangle \|\boldsymbol{x}\|^2 + \langle \boldsymbol{y}, \boldsymbol{z} \rangle \langle \boldsymbol{x}, \boldsymbol{y} \rangle \\ & - \langle \boldsymbol{x}, \boldsymbol{z} \rangle \langle \boldsymbol{x}, \boldsymbol{y} \rangle - \langle \boldsymbol{x}, \boldsymbol{z} \rangle \|\boldsymbol{y}\|^2 \end{aligned}$$

重新整理：

$$M(1) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle (\langle \boldsymbol{y}, \boldsymbol{z} \rangle - \langle \boldsymbol{x}, \boldsymbol{z} \rangle) + \|\boldsymbol{x}\|^2 \langle \boldsymbol{y}, \boldsymbol{z} \rangle - \|\boldsymbol{y}\|^2 \langle \boldsymbol{x}, \boldsymbol{z} \rangle$$

步骤 *3*：利用给定条件证明 $M(1) < 0$

给定条件：

$$\frac{\langle \boldsymbol{x}, \boldsymbol{z} \rangle}{\|\boldsymbol{x}\|} > \frac{\langle \boldsymbol{y}, \boldsymbol{z} \rangle}{\|\boldsymbol{y}\|}$$

等价地：

$$\langle \boldsymbol{x}, \boldsymbol{z} \rangle \|\boldsymbol{y}\| > \langle \boldsymbol{y}, \boldsymbol{z} \rangle \|\boldsymbol{x}\| \quad \text{(Condition)}$$

考虑两种情况：

情况 1： $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \geq 0$

由柯西-施瓦茨不等式：$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\|$。因此：

$$\begin{aligned} M(1) &\leq \|\boldsymbol{x}\|\|\boldsymbol{y}\|(\langle \boldsymbol{y}, \boldsymbol{z} \rangle - \langle \boldsymbol{x}, \boldsymbol{z} \rangle) + \|\boldsymbol{x}\|^2 \langle \boldsymbol{y}, \boldsymbol{z} \rangle - \|\boldsymbol{y}\|^2 \langle \boldsymbol{x}, \boldsymbol{z} \rangle \\ &= \|\boldsymbol{x}\|\|\boldsymbol{y}\| \langle \boldsymbol{y}, \boldsymbol{z} \rangle - \|\boldsymbol{x}\|\|\boldsymbol{y}\| \langle \boldsymbol{x}, \boldsymbol{z} \rangle + \|\boldsymbol{x}\|^2 \langle \boldsymbol{y}, \boldsymbol{z} \rangle - \|\boldsymbol{y}\|^2 \langle \boldsymbol{x}, \boldsymbol{z} \rangle \\ &= \langle \boldsymbol{y}, \boldsymbol{z} \rangle (\|\boldsymbol{x}\|\|\boldsymbol{y}\| + \|\boldsymbol{x}\|^2) - \langle \boldsymbol{x}, \boldsymbol{z} \rangle (\|\boldsymbol{x}\|\|\boldsymbol{y}\| + \|\boldsymbol{y}\|^2) \end{aligned}$$

由已知条件：

$$\langle \boldsymbol{x}, \boldsymbol{z} \rangle \|\boldsymbol{y}\| > \langle \boldsymbol{y}, \boldsymbol{z} \rangle \|\boldsymbol{x}\|$$

将两边同时乘以正数（$\|\boldsymbol{x}\| + \|\boldsymbol{y}\|$）：

$$\langle \boldsymbol{x}, \boldsymbol{z} \rangle \|\boldsymbol{y}\|(\|\boldsymbol{x}\| + \|\boldsymbol{y}\|) > \langle \boldsymbol{y}, \boldsymbol{z} \rangle \|\boldsymbol{x}\|(\|\boldsymbol{x}\| + \|\boldsymbol{y}\|)$$

即：

$$\langle \boldsymbol{x}, \boldsymbol{z} \rangle (\|\boldsymbol{x}\|\|\boldsymbol{y}\| + \|\boldsymbol{y}\|^2) > \langle \boldsymbol{y}, \boldsymbol{z} \rangle (\|\boldsymbol{x}\|^2 + \|\boldsymbol{x}\|\|\boldsymbol{y}\|)$$

因此 $M(1) < 0$。

情形 2： $\langle \boldsymbol{x}, \boldsymbol{y} \rangle < 0$

在这种情形下：

- 第一项：$\langle \boldsymbol{x}, \boldsymbol{y} \rangle (\langle \boldsymbol{y}, \boldsymbol{z} \rangle - \langle \boldsymbol{x}, \boldsymbol{z} \rangle) < 0$（因为 $\langle \boldsymbol{x}, \boldsymbol{y} \rangle < 0$ 且 $\langle \boldsymbol{y}, \boldsymbol{z} \rangle - \langle \boldsymbol{x}, \boldsymbol{z} \rangle < 0$）
- 第二项：$\|\boldsymbol{x}\|^2 \langle \boldsymbol{y}, \boldsymbol{z} \rangle - \|\boldsymbol{y}\|^2 \langle \boldsymbol{x}, \boldsymbol{z} \rangle < 0$（由给定条件可知）

因此 $M(1) < 0$。在这两种情况下，我们都有 $M(1) < 0$。

步骤 *4*：证明对所有 $\beta \in [0, 1]$，$f'(\beta) < 0$

由于 $M(\beta)$ 是 β 和 $M(1) < 0$ 的连续函数，通过分析 $M(\beta)$ 的二次函数性质，我们可以证明 $M(\beta) \leq 0$ 在 $[0, 1]$, 上，并且当 $\beta < 1$ 时，$M(\beta) < 0$。因此：

$$f'(\beta) = \frac{M(\beta)}{\|\boldsymbol{u}\|^3} < 0 \quad \text{for } \beta \in [0, 1)$$

也就是说，$f(\beta)$ 在 $[0, 1]$ 上严格单调递减。引理得证。 ■

主定理

定理3（ECGR误差缩减定理）在假设*2*下，对于 $0 \leq \beta < 1$，我们有：

$$\|\boldsymbol{g}'_{(t,s_i)} - \nabla F(\boldsymbol{w}_t)\|^2 < \|\boldsymbol{g}_{(t,s_i)} - \nabla F(\boldsymbol{w}_t)\|^2$$

证明： 步骤*1*：应用引理*1*

根据假设2：

$$\frac{\langle \boldsymbol{a}, \boldsymbol{\mu} \rangle}{\|\boldsymbol{a}\|} > \frac{\langle \boldsymbol{b}, \boldsymbol{\mu} \rangle}{\|\boldsymbol{b}\|}$$

根据引理1，对于 $0 \leq \beta < 1$:

$$\frac{\langle \boldsymbol{a} + \beta \boldsymbol{b}, \boldsymbol{\mu} \rangle}{\|\boldsymbol{a} + \beta \boldsymbol{b}\|} > \frac{\langle \boldsymbol{a} + \boldsymbol{b}, \boldsymbol{\mu} \rangle}{\|\boldsymbol{a} + \boldsymbol{b}\|}$$

即：

$$\frac{\langle \boldsymbol{v}, \boldsymbol{\mu} \rangle}{\|\boldsymbol{v}\|} > \frac{\langle \boldsymbol{c}, \boldsymbol{\mu} \rangle}{\|\boldsymbol{c}\|}$$

步骤*2*：误差比较

由于 $\|\gamma \boldsymbol{v}\| = \|\boldsymbol{c}\| = R$，我们有：

$$\|\gamma \boldsymbol{v} - \boldsymbol{\mu}\|^2 = R^2 - 2\gamma \langle \boldsymbol{v}, \boldsymbol{\mu} \rangle + \|\boldsymbol{\mu}\|^2$$
$$\|\boldsymbol{c} - \boldsymbol{\mu}\|^2 = R^2 - 2\langle \boldsymbol{c}, \boldsymbol{\mu} \rangle + \|\boldsymbol{\mu}\|^2$$

差值：

$$\|\gamma \boldsymbol{v} - \boldsymbol{\mu}\|^2 - \|\boldsymbol{c} - \boldsymbol{\mu}\|^2 = 2[\langle \boldsymbol{c}, \boldsymbol{\mu} \rangle - \gamma \langle \boldsymbol{v}, \boldsymbol{\mu} \rangle]$$

步骤 *3*：证明误差缩减

由步骤 1 得：

$$\frac{\langle \boldsymbol{v}, \boldsymbol{\mu} \rangle}{\|\boldsymbol{v}\|} > \frac{\langle \boldsymbol{c}, \boldsymbol{\mu} \rangle}{R}$$

代入 $\gamma = \frac{R}{\|\boldsymbol{v}\|}$：

$$\gamma\langle \boldsymbol{v}, \boldsymbol{\mu}\rangle > \langle \boldsymbol{c}, \boldsymbol{\mu}\rangle$$

因此：

$$\langle \boldsymbol{c}, \boldsymbol{\mu}\rangle - \gamma\langle \boldsymbol{v}, \boldsymbol{\mu}\rangle < 0$$

代入差值公式：

$$\|\gamma \boldsymbol{v} - \boldsymbol{\mu}\|^2 < \|\boldsymbol{c} - \boldsymbol{\mu}\|^2$$

即：

$$\|\boldsymbol{g}'_{(t,s_i)} - \nabla F(\boldsymbol{w}_t)\|^2 < \|\boldsymbol{g}_{(t,s_i)} - \nabla F(\boldsymbol{w}_t)\|^2$$

定理得证。 ■

B 补充算法

算法 2： FedProx-ECGR

要求： 全局轮次总数 T，本地数据集 $\mathcal{D}_i$（$\boldsymbol{x}_i \in \mathcal{D}_i$），本地迭代次数 τ_i，初始化权重 $\boldsymbol{w}_0$，初始化顺序 s_i，客户端 i，学习率 $\eta > 0$，近端系数 $\mu > 0$

对于每一轮 $t = 0, \ldots, T-1$ **执行**

  参数服务器将全局模型 $\boldsymbol{w}_t$ 发送给所有参与客户端;

  对于每个客户端 $i = 1, \ldots, N$ **do**

    对于每次本地迭代 $\lambda = 0, 1, \ldots, \tau_i$ **执行**

      初始化本地模型 $\boldsymbol{w}^{\lambda}_{(t,i)} \leftarrow \boldsymbol{w}_t$;

      带近端项的局部更新:

      $\boldsymbol{w}^{\lambda+1}_{(t,\ i)} = \boldsymbol{w}^{\lambda}_{(t,\ i)} - \eta\left(\nabla F_i(\boldsymbol{w}^{\lambda}_{(t,i)}; \boldsymbol{x}_{s_i}) + \mu(\boldsymbol{w}^{\lambda}_{(t,i)} - \boldsymbol{w}_t)\right)$;

    end

    存储局部梯度集 $\left\{\eta\left(\nabla F_i(\boldsymbol{w}^{\lambda}_{(t,i)}; \boldsymbol{x}_{s_i}) + \mu(\boldsymbol{w}^{\lambda}_{(t,i)} - \boldsymbol{w}_t)\right)\right\}_{\lambda=1}^{\tau_i}$;

    ECGR:

      $\pi_i \leftarrow \arg\min_{\pi_i \subset s_i} \left\|\boldsymbol{g}_{(t,\pi_i)}\right\|$ # 幅值排序

      $\pi'_i = s_i \setminus \pi_i$, $\beta \boldsymbol{g}_{(t,\pi'_i)}$ # 衰减提取

      $\boldsymbol{g}'_{(t,\ s_i)} = \gamma_i\ (\boldsymbol{g}_{(t,\pi_i)} + \beta \boldsymbol{g}_{(t,\pi'_i)}$ # 重新聚合

  end

  参数服务器接收来自所有客户端的 $\boldsymbol{g}'_{(t,s_i)}$;

  全局聚合 $\boldsymbol{G}'_t = \sum_{i=1}^{N} p_i \boldsymbol{g}'_{(t,\ s_i)}$;

  全局更新 $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \boldsymbol{G}'_t$;

结束

返回 $\boldsymbol{w}_T$;

FedProx、FedNova 和 Scaffold 的 ECGR扩展变体分别见算法 2、算法 3 和算法 4。与算法1一致，ECGR 不会改变原有基线方法的基本训练流程，而是增加一个梯度选择阶段。这一设计同样保留了算法1中强调的通信效率和结构兼容性 特性。

对于本文未涉及的其他先进联邦学习算法，以及联邦优化的未来发展，都可以遵循本节阐述的设计原则，轻松构建 ECGR 扩展。

算法 3： FedNova-ECGR

要求： 总全局轮次 T，本地数据集 $\mathcal{D}_i$（$\boldsymbol{x}_i \in \mathcal{D}_i$），本地迭代次数 τ_i，初始化权重 $\boldsymbol{w}_0$，客户端 i 处的初始化顺序 s_i，学习率 $\eta > 0$

对每一轮 $t = 0, \ldots, T-1$ **执行**

  参数服务器将全局模型 $\boldsymbol{w}_t$ 发送给所有参与客户端;

  对于每个客户端 $i = 1, ..., N$ **do**

    初始化本地模型 $\boldsymbol{w}^0_{(t,i)} \leftarrow \boldsymbol{w}_t$；

    对于每次本地迭代 $\lambda = 0、1、.....、\tau_i - 1$ **执行**

      局部更新：$\boldsymbol{w}^{\lambda+1}_{(t,\ i)} = \boldsymbol{w}^{\lambda}_{(t,\ i)} - \eta \nabla F_i(\boldsymbol{w}^{\lambda}_{(t,\ i)};\ \boldsymbol{x}_{s_i})$；

    end

    存储局部梯度集 $\left\{\eta \nabla F_i(\boldsymbol{w}^{\lambda}_{(t,i)}; \boldsymbol{x}_{s_i})\right\}_{\lambda=1}^{\tau_i}$;

    ECGR:

$$\pi_i \leftarrow \arg\min_{\pi_i \subset s_i} \left\|\boldsymbol{g}_{(t,\pi_i)}\right\| \quad \text{\# 幅值排序}$$

$$\pi_i' = s_i \setminus \pi_i, \quad \beta \boldsymbol{g}_{(t,\pi_i')} \quad \text{\# 衰减提取}$$

$$\boldsymbol{g}'_{(t,\ s_i)} = \gamma_i(\boldsymbol{g}_{(t,\pi_i)} + \beta \boldsymbol{g}_{(t,\pi_i')}) \quad \text{\# 重新聚合}$$

    按局部步数归一化：$\boldsymbol{g}'_{(t,s_i)} \leftarrow \frac{\boldsymbol{g}'_{(t,\ s_i)}}{\tau_i}$;

  end

  参数服务器接收$\boldsymbol{g}'_{(t,s_i)}$ 以及 τ_i，来自所有客户端;

  计算有效步长：$\tau_{\text{eff}} = \sum_{i=1}^{N} p_i \tau_i$;

  聚合归一化梯度：$\boldsymbol{G}'_t = \tau_{\text{eff}} \sum_{i=1}^{N} p_i \boldsymbol{g}'_{(t,\ s_i)}$;

  全局更新 $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \boldsymbol{G}'_t$;

结束

返回 $\boldsymbol{w}_T$;

算法 4： Scaffold-ECGR

需要： 全局轮次总数 T、本地数据集 $\mathcal{D}_i$ （$\boldsymbol{x}_i \in \mathcal{D}_i$）、本地迭代次数 τ_i、初始化权重 $\boldsymbol{w}_0$、全局控制变量 c、本地控制变量 c_i、学习率 $\eta > 0$

对每一轮 $t = 0, \ldots, T-1$ **执行**

  参数服务器向所有参与客户端发送（$\boldsymbol{w}_t, c$）；

  对于每个客户端 $i = 1, ..., N$ **do**

    初始化本地模型 $\boldsymbol{w}_{(t,i)}^0 \leftarrow \boldsymbol{w}_t$;

    对于每次本地迭代 $\lambda = 0, 1, \ldots, \tau_i - 1$ **执行**

      带控制校正的局部更新:

$$\boldsymbol{w}_{(t,i)}^{\lambda+1} = \boldsymbol{w}_{(t,i)}^{\lambda} - \eta \left(\nabla F_i \left(\boldsymbol{w}_{(t,i)}^{\lambda}; \boldsymbol{x}_{s_i} \right) - c_i + c \right)$$

    end

    存储校正后的本地梯度集 $\left\{ \eta \left(\nabla F_i(\boldsymbol{w}_{(t,i)}^{\lambda}; \boldsymbol{x}_{s_i}) - c_i + c \right) \right\}_{\lambda=1}^{\tau_i}$;

    ECGR:

$$\pi_i \leftarrow \arg\min_{\pi_i \subset s_i} \left\| \boldsymbol{g}_{(t,\pi_i)} \right\| \quad \text{\# 幅值排序}$$

$$\pi_i' = s_i \setminus \pi_i, \quad \beta \boldsymbol{g}_{(t,\pi_i')} \quad \text{\# 衰减提取}$$

$$\boldsymbol{g}'_{(t,s_i)} = \gamma_i \left(\boldsymbol{g}_{(t,\pi_i)} + \beta \boldsymbol{g}_{(t,\pi_i')} \right) \quad \text{\# 重新聚合}$$

    更新局部控制变量:

$$c_i' = c_i - c + \frac{1}{\tau_i \eta} \left(\boldsymbol{w}_t - \boldsymbol{w}^{\tau_i}(t,i) \right)$$

  end

  参数服务器接收 $\boldsymbol{g}'_{(t,s_i)}$ 以及来自所有客户端的 c_i'；

  全局聚合： $\boldsymbol{G}_t' = \sum_{i=1}^{N} p_i \boldsymbol{g}'_{(t,s_i)}$;

  全局模型更新： $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \boldsymbol{G}_t'$;

  更新全局控制变量： $c \leftarrow \sum_{i=1}^{N} p_i c_i'$;

结束

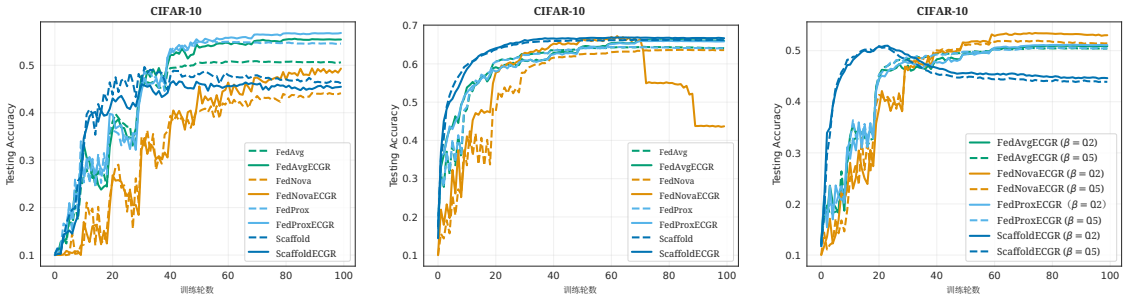
返回 $\boldsymbol{w}_T$;

C 附加结果

C.1 基准图像分类

本节提供了在标准图像分类数据集上进行基准测试的补充实验结果。具体而言，我们展示了在 CIFAR-10 上更全面的消融研究，以补充正文中报告的分析。此外，我们还进一步提供了在 MNIST、Fashion-MNIST 和 CIFAR-100 上的基线消融实验，这些内容未在正文中讨论。

CIFAR-10



(a) $\eta = 0.01$ (b) $\alpha = 0.1$ (c) $\beta = 0.5$与 $\beta = 0.2$对比

图 7：在 CIFAR-10 上针对学习率 η、数据异质性水平 α和 ECGR 阻尼系数 β进行的消融实验。所有曲线均报告了使用随机种子 0、1、42、999 和 2025年进行的五次独立运行的平均测试准确率。

如图 7(a)所示，与图 3(a)相比，平均准确率的上限有所提高；然而，Scaffold-ECGR 的性能进一步下降。相比之下，FedAvg 和 FedProx 的 ECGR 变体仍然有效，并且 FedNova 中观察到的灾难性准确率下降消失了。这些结果表明，对于不同的基线方法——尤其是那些直接操纵局部梯度的方法——仔细调节学习率至关重要。

与图 3(b)相比，图 7(b)显示 FedNova 再次出现灾难性的准确率下降。这表明，在不同的狄利克雷划分下，FedNova-ECGR 需要更加敏感且自适应的超参数调优。尽管如此，我们的 ECGR 策略仍然带来了小幅性能提升，这进一步表明，随着数据异质性程度的增加，其优势会变得更加明显。

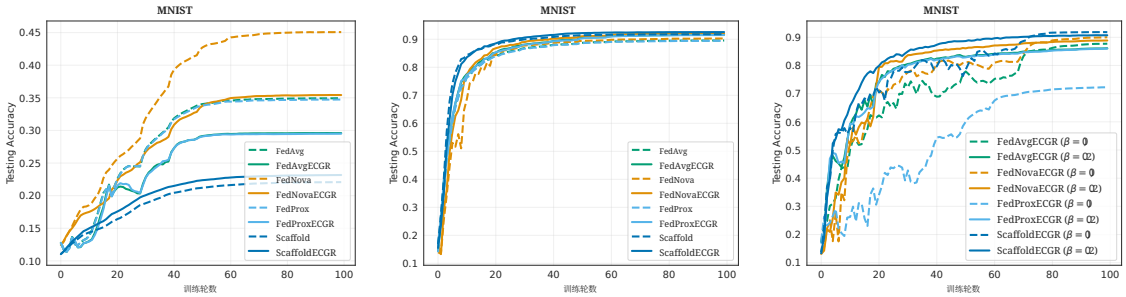
图 7 (c) 显示，随着阻尼系数 β 增大，ECGR 带来的性能提升逐渐减弱，最终退化至基线。然而，图 3 表明，β 过小会导致准确率振荡，并减缓训练早期阶段的提升。因此，选择合适的 β 对于平衡 ECGR 带来的收益与因丢弃过多梯度信息而导致的不稳定性至关重要。

MNIST

由于 MNIST数据集过于简单，且 LeNet模型相对较小，训练动态对学习率的选择变得极为敏感。因此，当学习率设置过低时，基线方法无法收敛，如图 8 (a) 所示。在这种情况下，ECGR 策略无法带来有效的改进。

如图 8 (b) 所示，结果与我们在 CIFAR-10数据集上的发现一致：在高度非独立同分布数据设置下，ECGR 表现更佳。

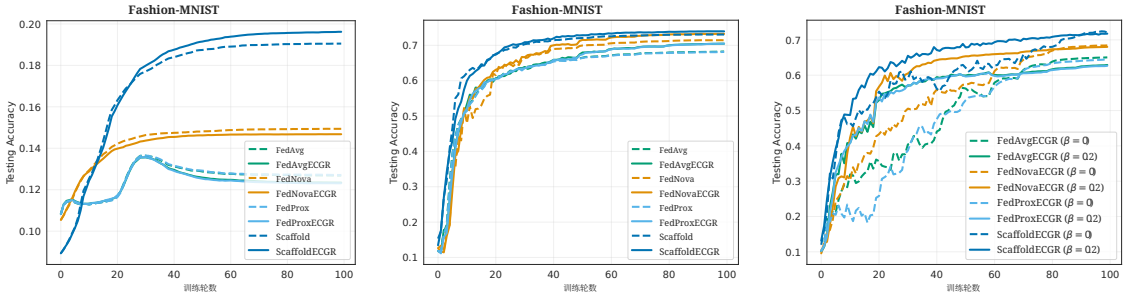
如图 8 (c) 所示，MNIST 上的结果与 CIFAR-10 上观察到的趋势一致：丢弃探索性梯度（即 $\beta = 0$）会导致训练早期出现明显的准确率振荡和性能下降。然而，由于 MNIST数据集较为简单，且 LeNet模型的参数规模较小——这两点共同降低了优化难度，减轻了梯度信息丢失带来的负面影响——对于大多数基线方法（FedProx-ECGR 除外），$\beta = 0$ 这一设置反而意外地带来了更好的最终性能。



(a) $\eta = 0.0001$ (b) $\alpha = 0.1$ (c) $\beta = 0$对比 $\beta = 0.2$

图 8：针对学习率 η、数据异质性水平 α 和 ECGR 阻尼系数 β，在 MNIST 上进行的消融研究。所有曲线均报告了五次独立运行的平均测试准确率，随机种子分别为 0、1、42、999 和 2025年。

Fashion-MNIST



(a) $\eta = 0.0001$ (b) $\alpha = 0.1$ (c) $\beta = 0$与 $\beta = 0.2$相比

图 9：在不同学习率 η、数据异质性水平 α 和 ECGR 阻尼系数 β 下，Fashion-MNIST 上的消融研究。所有曲线均报告了 5 次独立运行的平均测试准确率，随机种子为 0、1、42、999 和 2025年。

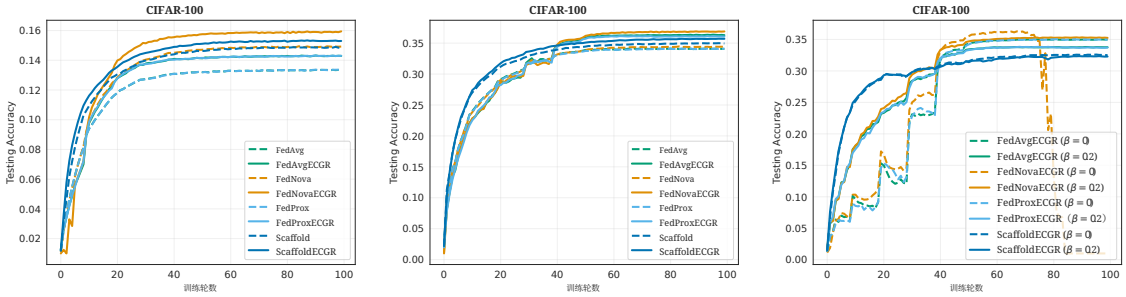
Fashion-MNIST 上的分析结果与 MNIST 上观察到的总体趋势一致。

CIFAR-100

从 CIFAR-10 和 CIFAR-100 得出的结论大体一致，学习率 η 相关的结论除外。与 MNIST 和 Fashion-MNIST 类似，CIFAR-100 上平均测试准确率的上限显著下降。然而，与这些较简单的数据集不同，ECGR 仍能带来明显的性能提升。这可归因于 CIFAR-100 更高的复杂度和更丰富的语义多样性，以及所采用 CNN 模型更大的容量；这些因素使探索梯度阶段与收敛梯度阶段之间的区别更加明显，并使 ECGR 能够更好地利用这一结构。

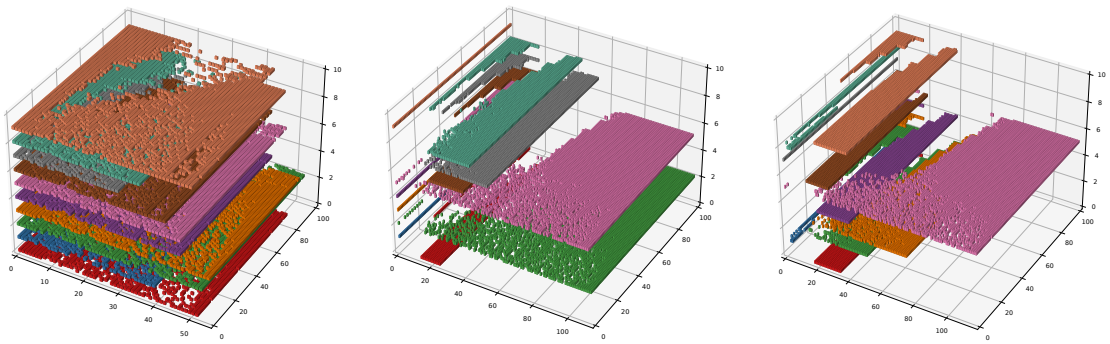
C.2 ECGR 中梯度选择的补充可视化

本节对第 4.3 节中展示的 CIFAR-10 梯度选择可视化进行扩展。我们首先给出独立同分布设置下所选梯度的三维视图，随后展示使用不同随机种子得到的三维可视化结果。接着，我们进一步展示种子为 42 时其他数据集的三维视图。



(a) $\eta = 0.0001$ (b) $\alpha = 0.1$ (c) $\beta = 0$对比 $\beta = 0.2$

图10：关于学习率 η、数据异质性水平 α 和 ECGR 阻尼系数 β 的 CIFAR-100 消融实验。所有曲线均报告了使用随机种子 0、1、42、999 和 2025 进行五次独立运行所得的平均测试准确率。



(a) $\alpha = 1$ (b) 随机种子 0 (c) 随机种子 2025年

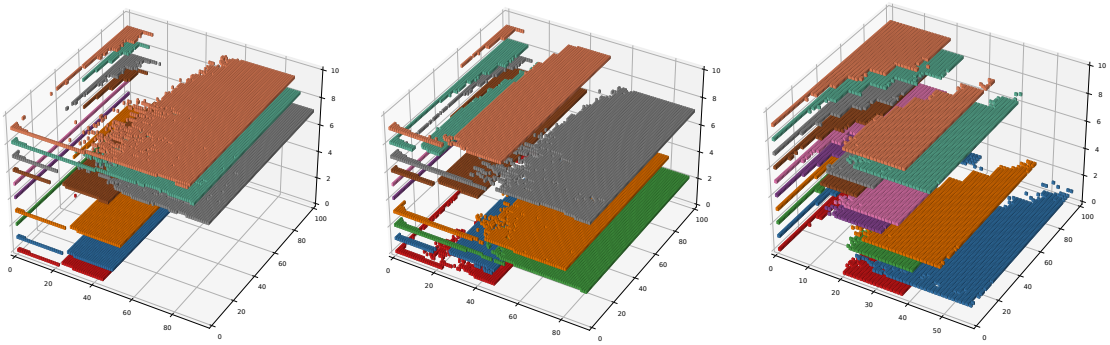
图11：ECGR 在 CIFAR-10 上独立同分布（$\alpha = 1$，随机种子 = 42）和非独立同分布（$\alpha = 0.01$，随机种子 = 0 和 2025年）设置下的梯度选择可视化。

CIFAR-10

如图 11 (a) 所示，在近似 IID 的设置下（即 $\alpha = 1$），ECGR 所做的梯度选择往往类似于随机选择。这是因为，在独立同分布数据分布下，每个客户端计算出的梯度已经接近最优梯度。因此，探索梯度与收敛梯度之间的差异变小，导致可区分性减弱，选择也更加均匀混杂。

图 11 (b) 和 (c) 中的可视化结果表明，在非独立同分布设置下，不同随机种子引起的客户端数据分布变化对 ECGR 的梯度选择行为影响甚微。在所有种子下，ECGR 始终倾向于选择较晚本地迭代中的梯度作为面向收敛的梯度，这与随机梯度下降的经典收敛行为一致。

MNIST、Fashion-MNIST 和 CIFAR-100



(a) MNIST　(b) Fashion-MNIST　(c) CIFAR-100

图 12：ECGR 在三个数据集——MNIST、Fashion-MNIST 和 CIFAR-100——上的每轮梯度选择可视化。所有可视化均在相同的实验设置下生成，Dirichlet 异质性参数为 $\alpha = 0.01$，随机种子为 42。

如图 12 所示，MNIST 和 Fashion-MNIST 数据集呈现出的梯度选择行为与 CIFAR-10 上观察到的情况一致。然而，对于 CIFAR-100，在训练早期，ECGR 选择的收敛梯度往往对应较晚的本地迭代次数；而在训练后期——全局模型接近收敛时——所选收敛梯度则转向较早的本地迭代次数。其他三个数据集（MNIST、Fashion-MNIST 和 CIFAR-10）也呈现出这一现象的较弱版本。

相反，这种行为可以解释为：在训练后期，全局模型逐渐接近收敛状态。因此，探索性梯度与收敛梯度之间的判别差距会缩小，导致所选的收敛梯度转向较早的本地迭代次数。由于 CIFAR-100 的任务复杂度较高，这种近收敛行为出现得更早，因此这一转变在该数据集上更加明显。