中国科学: 信息科学 2024 年 第 54 卷 第 3 期: 566–581
SCIENTIA SINICA Informationis
论文

联邦学习在高度数据异构场景下的泛化鲁棒性增强

万伟[1,4,5,6,7], 胡胜山[1,4,5,6,7*], 陆建荣[1,4,5,6,7], 李明慧[2], 周子淇[1,4,5,6,7], 金海[3,4,5,8]

1. 华中科技大学网络空间安全学院, 武汉 430074
2. 华中科技大学软件学院, 武汉 430074
3. 华中科技大学计算机科学与技术学院, 武汉 430074
4. 大数据技术与系统国家地方联合工程研究中心, 武汉 430074
5. 服务计算技术与系统教育部重点实验室, 武汉 430074
6. 分布式系统安全湖北省重点实验室, 武汉 430074
7. 湖北省大数据安全工程技术研究中心, 武汉 430074
8. 集群与网格计算湖北省重点实验室, 武汉 430074
* 通信作者. E-mail: hushengshan@hust.edu.cn

收稿日期: 2023–04–18; 修回日期: 2023–09–06; 接受日期: 2023–12–01; 网络出版日期: 2024–03–08

国家自然科学基金 (批准号: U20A20177) 和湖北省技术创新计划重点研发专项 (批准号: 2021BAA032) 资助项目

摘要 联邦学习 (federated learning, FL) 是一种以保护客户隐私数据为中心的分布式处理网络, 为解决隐私泄露问题提供了前景良好的解决方案. 然而, FL 的一个主要困境是高度非独立同分布 (non-independent and identically distributed, non-IID) 的数据会导致全局模型性能很差. 尽管相关研究已经探讨了这个问题, 但本文发现当面对 non-IID 数据、不稳定的客户端参与以及深度模型时, 现有方案和标准基线 FedAvg 相比, 只有微弱的优势或甚至更差, 因此严重阻碍了 FL 的隐私保护应用价值. 为解决这个问题, 本文提出了一种对 non-IID 数据鲁棒的优化方案: FedUp. 该方案在保留 FL 隐私保护特点的前提下, 进一步提升了全局模型的泛化鲁棒性. FedUp 的核心思路是最小化全局经验损失函数的上限来保证模型具有低的泛化误差. 大量仿真实验表明, FedUp 显著优于现有方案, 并对高度 non-IID 数据以及不稳定和大规模客户端的参与具有鲁棒性.

关键词 分布式网络, 联邦学习, 异构优化, 泛化性, 鲁棒性, 隐私保护

1 引言

联邦学习 [1] (federated learning, FL) 是一种具有广阔应用前景的分布式学习范式, 其特点是所有本地客户端可以通过共享本地的模型到一个云服务器上来协同学习高性能的全局模型. FL 不会将客户端的数据私自移动到服务器, 在注重客户数据隐私安全的同时致力于提供一个高精度的全局共享模

引用格式: 万伟, 胡胜山, 陆建荣, 等. 联邦学习在高度数据异构场景下的泛化鲁棒性增强. 中国科学: 信息科学, 2024, 54: 566–581, doi: 10.1360/SSI-2023-0107
Wan W, Hu S S, Lu J R, et al. Enhancing generalization robustness of federated learning in highly heterogeneous environments (in Chinese). Sci Sin Inform, 2024, 54: 566–581, doi: 10.1360/SSI-2023-0107

型, 因此可以在不泄露用户隐私的前提下为跨地域的客户提供高质量的深度学习模型服务. 目前, FL 已成为隐私保护等安全领域 [2,3] 的重要研究主题, 并在金融 [4]、推荐系统 [5] 和医学图像分析 [6] 等多个强隐私相关的领域上取得了巨大成功. FedAvg 是 FL 的一个标准算法, 其基本思想是: 服务器在每个通信轮次中向一组活跃的客户端广播其全局模型, 这些客户端随后利用全局模型和本地数据集来训练最新的本地模型, 以用来更新服务器维持的全局模型. 然而, FedAvg 的主要难点是在高度非独立同分布 (non-independent and identically distributed, non-IID) 数据集上进行模型训练. 由于来自不同地区的客户端可能具有完全不同的生活习惯, 因此它们的数据集可能在数据类型和数据规模上高度异构. 而这些问题已被证明会导致 FedAvg 的模型准确率非常差 [7~9], 这严重阻碍了 FL 隐私保护的实际使用价值.

在解决 non-IID 数据导致准确率降低的问题上, 现有工作已经取得了一些富有成效的研究成果. 根据对 FedAvg 的修改方式, 这些方法可以大致分为以下 4 个类别: (1) 操纵本地训练 [10~13]. 该方法的目标是通过修改训练数据 [14~16] 或随机梯度下降 (stochastic gradient descent, SGD) 过程 [17~22], 使得本地模型和全局模型更加一致. (2) 操纵聚合过程 [23~26]. 该方法旨在修改聚合策略, 以减少由模型不一致性引起的低准确率. (3) 操纵客户选择模式 [27~30]. 该方法的目标是为 FL 选择高质量的客户群体. (4) 操纵模型输出模式 [31~34]. 该方法旨在将 FedAvg 的单模型输出形式修改为多模型输出模式, 为不同客户提供个性化的模型. 本文研究专注于面向 non-IID 问题为 FL 提供单一高质量的全局模型, 因此着重对比研究前三类方法. 这些方法使用共享数据、个性化训练、动量和控制变量等方式, 在一定程度上改善了 FL 性能. 然而, 当面对高度 non-IID 数据且客户参与状态十分不稳定的环境时, 这些措施往往难以实现良好的性能. 例如, Mime [10] 和 SCAFFOLD [11] 使用动量和控制变量来强制本地梯度朝着全局梯度的方向更新, 这使得它们在 non-IID 数据下的准确率提升十分有限甚至更差. 事实上, 如表 1 所示, 现有方法主要适用于微弱的 non-IID 数据分布、小规模且稳定的客户端参与、简单的模型和额外的梯度通信. 否则, 这些方法对 FedAvg 的性能提升有限, 甚至可能产生负面影响. 因此, 如何进一步部署一种对 FL 有显著改进的异构优化算法仍然面临很大的挑战.

为了解决以上问题, 本文进一步弥补了异构优化算法与其实际使用之间的差距. 为此, 本文提出了 FedUp, 一种易于实现的异构优化算法. FedUp 的思路是为每一个客户最小化全局损失函数的上界来学习具有较低泛化误差的全局模型. 具体而言, FedUp 首先通过以 Lipschitz 平滑条件为基础的二次函数来动态地界定全局损失的上界. 然后, FedUp 允许每个客户端最小化经验损失和该动态上界之间的一组线性组合, 而不仅仅是只最小化本地经验损失函数. 为了验证其对异构数据的泛化鲁棒性, 本文在 3 个 non-IID 数据集 (FEMNIST, CIFAR-10 和 CIFAR-100) 上进行了大量的实验, 结果表明, 与现有最先进的方法 (如 SCAFFOLD [11], MOON [35], FedDyn [9], FedSMOO [19] 和 FedPVR [20]) 相比, FedUp 具有显著的优势, 并且对高度 non-IID 数据以及不稳定 (例如 0.2% 的采样率) 和大规模客户端参与 (例如 1020 个客户端) 具有鲁棒性. 本文的主要贡献如下:

(1) 本文对当前最先进的 FL 方法进行了全面比较. 实验分析发现, 它们主要适用于温和的 non-IID 数据分布、小规模客户、稳定的客户端参与和额外的梯度上传. 通过对 FL 优化方法进行深入的研究, 本文发现它们需要对 FL 的收敛行为进行繁重且复杂的操控, 这使得它们不得不依赖于这些不切实际的假设.

(2) 本文提出了 FedUp, 一种只需要对 FedAvg 做轻微的修改就可以动态地最小化估计的全局损失函数的上界.

(3) 本文在高度 non-IID 数据和不稳定以及大规模客户端参与情况下对 FedUp 进行了大规模的实验验证, 结果表明, FedUp 在所有情况下都显著优于最先进的方法, 并且对这种设置高度鲁棒.

表 1　FedUp 与现有方案的假设条件比较, 其中 ✓ 和 × 分别表示对应的方法是否满足相应的条件

Table 1 Comparison of our FedUp with existing schemes. Note that ✓ and × mean that the methods can and cannot satisfy the corresponding conditions, respectively

Method	Highly non-IID data	Unstable client participation	Deep model	Large-scale clients	Complex dataset	No additional gradient communication	No additional mapping header
FedProx	×	✓	×	×	✓	✓	✓
FedSGD	✓	✓	×	✓	✓	✓	✓
FedSMB	✓	✓	×	✓	✓	✓	✓
SCAFFOLD	×	×	✓	×	✓	×	✓
FedDC	×	✓	✓	✓	✓	×	✓
Mime	×	×	×	×	×	×	✓
MimeLite	×	✓	✓	×	✓	×	✓
FedDyn	×	×	✓	×	✓	✓	✓
MOON	✓	✓	✓	×	✓	✓	×
FedSplit	×	×	×	×	✓	✓	✓
STEM	×	×	✓	×	×	✓	✓
FedSMOO	×	✓	✓	✓	✓	×	✓
FedPVR	×	×	✓	×	✓	×	✓
FedUp (ours)	✓	✓	✓	✓	✓	✓	✓

2　背景

本文考虑一个集中式 FL 系统, 其中中央服务器协调所有分布式客户, 并使用经典的 FedAvg [1] 优化算法协同训练全局模型. 假设 n 个客户参与 FL 的训练过程, 其中客户的数据集分别表示为 $D_1, D_2, D_3, \ldots, D_n$. 此外, 存在一个服务器协调所有客户以协作最小化以下全局损失函数:

$$\min_{\boldsymbol{X}\in\mathbb{R}^d}\left[F(\boldsymbol{X}) := \sum_{i=1}^{n}\frac{1}{n}F_i(\boldsymbol{X})\right], \tag{1}$$

其中 $F_i(\boldsymbol{X}) = \sum_{s=1}^{|D_i|}\frac{1}{|D_i|}L(\boldsymbol{X}, \xi_{i,s})$ 是第 i 个客户端的损失函数, L 是一个损失函数, 例如交叉熵损失, $\xi_{i,s}$ 是客户数据集 D_i 里的一个数据样本. 在第 t 轮, 所有客户端用 SGD 等优化器求解 $F_i(\boldsymbol{X})$:

$$\boldsymbol{X}_i^{(t,k+1)} = \boldsymbol{X}_i^{(t,k)} - \eta\nabla F_i(\boldsymbol{X}_i^{(t,k)}), \tag{2}$$

其中 η 是学习率, $\boldsymbol{X}_i^{(t,k)}$ 是客户端 i 在第 k 个本地 SGD 步骤时的本地模型, $\nabla F_i(\boldsymbol{X}_i^{(t,k)})$ 表示在 $\boldsymbol{X}_i^{(t,k)}$ 处的 Mini-batch 梯度. 随后服务器用 FedAvg 算法聚合所有本地模型 $\{\boldsymbol{X}_i^{(t,K)}\ |i\in[n]\}$:

$$\boldsymbol{X}^{t+1} = \sum_{i=1}^{n}\frac{1}{n}\boldsymbol{X}_i^{(t,K)}, \tag{3}$$

其中 K 代表 SGD 更新的总步数. 服务器也可以通过聚合所有本地梯度来更新得到全局模型:

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^{t} - \eta\sum_{i=1}^{n}\frac{1}{n}\boldsymbol{g}_i^t, \tag{4}$$

其中 $\boldsymbol{g}_i^t = \sum_{k=1}^{K}\nabla F_i^{(t,k)}$; K 为本地的 SGD 迭代步数.

3　相关研究

当前的工作围绕 non-IID 数据导致性能下降的问题展开了一系列有效的探索. 根据它们对 FL 的操纵方式, 可以大致将这些方法分为 4 个类别: 操纵本地训练、操纵聚合过程、操纵客户选择模式以及操纵模型输出模式.

操纵本地训练. 导致 FedAvg 性能低下最直接的因素是本地客户数据集高度的异构性, 这种异构性使得客户本地优化目标不一致从而导致聚合的全局模型难以兼顾所有本地客户的信息和数据特点. 因此在该类别中, 现有工作致力于在本地训练阶段中对 FL 做修改, 主要涉及本地数据和本地 SGD 训练的修改, 从而消除客户优化目标的不一致. 在本地数据的修改上, 现有工作致力于使用共享的数据集混合到本地训练数据集中 [14~16], 从而直接减少客户数据的不一致. 但是共享的数据通常需要从客户端提供, 有的甚至要求共享数据集的数据标签类型涵盖所有客户数据标签 [14] 或者客户本地数据集的表征 [15], 这给客户带来了极大的隐私泄露担忧. 在本地 SGD 训练上, 一种简单的想法是只进行一步 SGD 训练, 就像 FedSMB 和 FedSGD [36] 中所做的那样, 从而保证 FL 与传统的数据集中式训练表现相同. 然而, 它们需要大量的通信开销, 不适用于具有 dropout 或批归一化的深度模型. 更有技术难度的方式是通过操纵本地梯度 [10~12,18,19] 或者本地优化目标 [7,9,13,17,20] 来迫使全局模型收敛到全局损失函数的最优点. 然而, 精确控制 FL 的收敛行为并不容易, 现有的方案不能很好地实现这个目标, 主要是由于它们依赖于各种不切实际的假设. 具体而言, 为了操纵本地梯度, Mime 及其轻量级版本 MimeLite [10] 需要使用准确的全局信息 (即所有客户端的梯度) 替换 SCAFFOLD [11] 中使用的不准确的动量和控制变量. 结果, 上述方案必须在稳定的客户端参与或类似 FEMNIST [37] 的简单数据集中才能有效工作. 此外, 它们需要更多的通信开销, 并带来更大的隐私泄露风险. 例如, 在每个通信轮次中, Mime 中的客户端需要上传两个额外的梯度, 这使得它们的训练数据更容易被重构而泄露用户的隐私信息; FedDyn [9] 需要稳定的客户端参与, 并在 non-IID 数据集中表现得很糟糕.

操纵聚合过程. 经典的 FL 基准算法 FedAvg 在聚合时, 直接根据客户数据的规模占比对客户模型或者梯度进行加权平均. 现有研究表明, 这种简单的聚合方式会导致 FL 全局模型收敛点和目标函数的最优收敛点不一致 [8]. 为此, 许多工作 [8,23~26] 致力于调整聚合策略来提高 FedAvg. FedBN [26] 简单地移除标准化层 (batch normalization layer) 的参数后再平均本地模型. FedNova [8] 聚合归一化的梯度. FedMA [23] 对本地模型进行层次匹配聚合. 总的来说, 这些方法都是在本地优化目标已经不一致后才进行修正的, 只能缓解异构带来的性能下降. 本文的方法在本地训练开始之前, 尽可能地让所有客户的优化目标函数一致, 因此能从根源上大幅消除异构导致的目标不一致. 此外, 本文的方法发生在本地训练阶段, 和这些方法兼容互补.

操纵客户选择模式. FedAvg 在选择参与聚合的客户时使用随机采样的方式, 无法保证选择到最优的客户, 甚至可能选择到模型质量很差的客户, 从而导致 FL 性能下降. 因此, 现有研究致力于探索更好的客户选择方法, 以提高 FL 的收敛速度和模型准确率 [27~30]. 这些工作通常采取具有优化搜索特点的算法来自适应地决定最佳客户端. 例如, 文献 [27] 利用强化学习来动态地选择每轮训练过程中表现积极的客户; 文献 [28] 基于贪心算法, 要求服务器在预设的时间内筛选尽可能多的高质量模型. 文献 [29] 采用探索与开发策略 (exploration-exploitation strategy) 来探索和利用新的高质量客户. 文献 [30] 利用重要性采样来保证选择的客户能聚合出无偏的全局梯度. 总的来说, 这些方法主要集中在训练之前的客户选择阶段, 而本文的方法则是在训练过程中通过优化客户的本地损失函数来改善 FL 的性能. 因此, 通过联合使用这些方法, 可以在客户选择和本地训练阶段同时对 FL 进行优化, 从而更全面地改善模型的性能.

操纵模型输出模式. 传统的 FL 将所有客户的数据信息汇集到一个全局模型中, 并最终为所有客户输出一个统一的全局模型. 然而, 由于客户数据的高度异构性, 一个共享的全局模型很难涵盖每个客户的数据信息和特点. 因此, 现有的方法旨在将这种单一输出模式改为多模型输出形式, 以为具有不同数据分布的客户提供更加个性化的模型. 在这个方向上, 一种常见的方法是基于聚类的 FL [31~34,38], 通过将具有相似数据分布的客户端放入同一组中, 然后专注于为每个组提供更符合其数据特点的个性

化全局模型. 然而这种方法通常需要等 FL 完成训练后再进行聚类分组, 引入了巨大的通信开销. 个性化 FL [39~41] 则更为精细地为每个客户维护一个个性化的本地模型, 以避免全局模型在客户本地数据集上准确率差的问题. 本文的研究主要集中在经典的 FL 场景, 并且特别关注获取一个单一且高精度的全局模型.

表 1 对现有相关工作进行了全面比较. 通过表中调查结果可以观察到每种方法都有其独特的优缺点. 它们主要依赖于适度的 non-IID 数据、小规模客户数量、稳定的客户端参与、简单的模型和额外的梯度交换. 否则, 正如本文在第 6.2 小节的实验结果所示, 当这些假设不满足时, 它们相对于基线 FedAvg 几乎没有任何优势甚至性能更差. 因此总的来说, 在不牵涉到额外的隐私泄露风险的情况下, 部署能显著提升 FL 性能的异构优化算法仍然远未达到实际应用的程度.

4 现有算法分析

SCAFFOLD, Mime 和 FedDyn 在文献中展现出了优越的理论保证和性能, 然而这些方法依赖于许多不切实际的假设. 为此, 本小节将深入探讨这 3 种先进的方案, 并给出它们在实际场景中失效的模式以及原因分析.

SCAFFOLD [12]. SCAFFOLD 旨在利用控制变量对每一个批次上的梯度 $\nabla F_i(\boldsymbol{X}_i^{(t,k)})$ 进行双重修正, 使其趋近于全局梯度 $\nabla F(\boldsymbol{X}^t)$. 具体而言, SCAFFOLD 首先需要使用与 $\nabla F_i(\boldsymbol{X}_i^{(t,k)})$ 相似的客户端控制变量 c_i^t 大幅消除 $\nabla F_i(\boldsymbol{X}_i^{(t,k)})$, 即通过计算 $\nabla F_i(\boldsymbol{X}_i^{(t,k)}) - c_i^t$ 使 $\nabla F_i(\boldsymbol{X}_i^{(t,k)})$ 被 c_i^t 消除. 然后, 它通过全局控制变量 c^t 来进一步对 $\nabla F_i(\boldsymbol{X}_i^{(t,k)})$ 矫正, 即 $\nabla F_i(\boldsymbol{X}_i^{(t,k)}) - c_i^t + c^t$, 以控制本地梯度朝向具有全局信息的全局控制变量 c^t, 从而使每个客户端具有与传统数据集中式训练相似的收敛行为. 然而, 参与状态不稳定的客户端会导致 c_i^t 过于陈旧 (如, 为以前轮次的梯度 $\boldsymbol{g}_i^1, \boldsymbol{g}_i^2, \ldots, \boldsymbol{g}_i^{t-1}$ 中的任意一个), 这使得在聚合时, 全局控制变量和客户端控制变量不能相互消除, 即对于全局梯度 $\boldsymbol{g}^t - \frac{1}{n}\sum_i^n c_i^t + c^t$, 有 $\boldsymbol{g}^t - \frac{1}{n}\sum_i^n c_i^t + c^t \neq \boldsymbol{g}^t$. 因此, 异常陈旧的全局梯度 $\boldsymbol{g}^t - \frac{1}{n}\sum_i^n c_i^t + c^t$ 会导致全局模型在异构数据集上的性能非常差, 这在本文的实验和之前的工作 [10~12, 16] 中都可以证实.

Mime [10]. Mime 直接利用 $\nabla F_i(\boldsymbol{X}^t, \xi^k)$ 和 $\nabla F(\boldsymbol{X}^t, \xi^k)$ 分别替代 SCAFFOLD 中陈旧的客户端控制变量和全局控制变量, 从而强迫 Mini-batch 梯度 $\nabla F_i(\boldsymbol{X}_i^{(t,k)}, \xi^k)$ 更接近于 $\nabla F(\boldsymbol{X}^t, \xi^k)$, 即利用如下公式更新本地模型: $\nabla F_i(\boldsymbol{X}_i^{(t,k)}, \xi^k) - \nabla F_i(\boldsymbol{X}^t, \xi^k) + \nabla F(\boldsymbol{X}^t, \xi^k)$. 然而, 在异构数据集中, 本地梯度更新反而会向有害的方向推进. 这是因为成功的校正需要假设 $\nabla F_i(\boldsymbol{X}^t, \xi^k)$ 与 $\nabla F_i(\boldsymbol{X}_i^{(t,k)}, \xi^k)$ 相似, 但在数据分布异构时该假设不成立.

为了验证上述观点, 本文通过消融实验来展示 $\nabla F_i(\boldsymbol{X}^t, \xi^k)$ 和 $\nabla F(\boldsymbol{X}^t, \xi^k)$ 如何影响校正效果. 如表 2 所示, 在简单的数据集 FEMNIST 上, $A - B + C$ 可以显著地校正 A 和 C 相似 (将余弦相似度从 0.3405 增加到 0.5451). 这是因为 A 和 B 的相似度很高, 为 0.7514, 使得 A 被 $-B$ 大幅抵消. 此时, C 可以进一步有效地将 $A - B$ 校正接近 C. 然而, 在高度异构的数据集 CIFAR-10 上, 使用 $-B$ 来校正 A 是适得其反的 (即将余弦相似度从 0.0057 降低到 −0.0111), 因为高度异构可以致使 B 与 A 大相径庭, 导致 $A - B$ 与 C 之间的差异更大.

FedDyn [9]. FedDyn 可以确保全局模型收敛于全局目标函数的驻点, 但是它严重依赖于具有丰富计算资源和能稳定参与的客户端, 并且在异构数据集上表现不佳. 具体来说, 它的本地梯度是 $\nabla F_i(\boldsymbol{X}_i^{(t,k)}) - \nabla F_i(\boldsymbol{X}_i^{(s,*)}) + \alpha(\boldsymbol{X}_i^{(t,k)} - \boldsymbol{X}^t)$, 其中 α 是一个常数, $\boldsymbol{X}_i^{(s,*)}$ 可以为客户端 i 在前 $t-1$ 轮中目标函数的任意一个驻点 (即 $\boldsymbol{X}_i^{(s,*)} \in \{\boldsymbol{X}_i^{(1,*)}\boldsymbol{X}_i^{(2,*)}, \ldots, \boldsymbol{X}_i^{(t-1,*)}\}$). 然而, 为了得到 $\boldsymbol{X}_i^{(s,*)}$, 所有客户端都需要训练它们的本地模型直到收敛才能上传, 这显著增加了客户端的计算负担. 其次, 为了

表 2 在 FEMNIST 和 CIFAR-10 上用不同梯度组合校正 $\nabla F_i(\boldsymbol{X}_i^{(t,k)}, \xi^k)$ 的消融研究. 本文计算每个客户端在 40 个 Mini-batch 上的余弦相似度的平均值

Table 2 Ablation study of Mime with different gradient combinations to rectify $\nabla F_i(\boldsymbol{X}_i^{(t,k)}, \xi^k)$ over FEMNIST and CIFAR-10. We compute the mean of the cosine similarities over 40 batches in each client and then report the average of the mean similarities [a)]

Corrective setting	Cosine similarity	
	FEMNIST	CIFAR-10
A and C (not correcting A)	0.3405	0.0057
$A-B+C$ and C (using $-B+C$ to correct A as Mime)	0.5451	−0.0111
$A-B$ and C (only using $-B$ to correct A)	0.0182	−0.2062
A and B (not correcting A)	0.7514	0.0316

a) Let A, B, and C denote $\nabla F_i(\boldsymbol{X}_i^{(t,k)}, \xi^k)$, $\nabla F_i(\boldsymbol{X}^t, \xi^k)$, and $\nabla F(\boldsymbol{X}^t, \xi^k)$, respectively.

表 3 全文所使用符号总结表

Table 3 Summary of notations used in the paper

Parameter	Description
n, m, i	Total number of clients, number of sampled clients, client index
T, t; K, k	Number of communication round, index of round; SGD steps, index of SGD step
$\boldsymbol{X}^t$, $\boldsymbol{X}_i^{(t,k)}$	Global model for round t, client i model for SGD step k in round t
D, ξ; F_i, F	Dataset, sample; client i optimization objective, global optimization objective
∇, μ, α, η	Derivation, smoothing coefficient, FedUp hyperparameter, learning rate

确保所有的本地模型都收敛于最优全局模型, 必须使用最新的 $\boldsymbol{X}_i^{(s,*)}$, 即 $\boldsymbol{X}_i^{(s,*)} = \boldsymbol{X}_i^{(t-1,*)}$, 这只有当 FL 每轮选择相同的客户端时才能满足.

总之, 现有的方法意识到本地优化目标和全局优化目标不一致 (即对于所有客户端 $F_i(\boldsymbol{X}) \neq F(\boldsymbol{X})$) 会导致 FL 出现高的收敛误差, 从而通过对本地梯度的强制控制来纠正有偏差的收敛过程. 然而, 当这种目标不一致在一开始就存在时, 后期再去调整 FL 的模型来靠近理想的最优模型十分困难. 这使它们不得不依赖于各种假设, 从而在异构和不稳定的实际场景中失效.

5 算法设计

5.1 设计动机

为了能够避免以往方案需要对 FL 做的各种复杂的控制, 本文旨在设计一种实用且性能优越的异构优化算法. FedUp 背后的想法非常直观易懂: 如果每个客户在进行经验风险最小化的同时能最小化其本地模型在全局数据集 $D = D_1 \bigcup D_2 \bigcup \cdots \bigcup D_n$ 上的经验风险损失, 那 FL 的性能将得到显著提升. 为了便于阅读, 表 3 总结了本文使用的符号含义.

观察到如果假设 FL 的优化目标满足经常使用的 Lipschitz 平滑假设, 那么全局损失函数 $F(X)$ 的上界可以为

$$F(\boldsymbol{X}) \leqslant F(\boldsymbol{X}^t) + [\nabla F(\boldsymbol{X}^t)]^\top (\boldsymbol{X} - \boldsymbol{X}^t) + \frac{\mu}{2}\|\boldsymbol{X} - \boldsymbol{X}^t\|_2^2, \tag{5}$$

其中 μ 为 Lipschitz 系数. 注意到, 该上界是一个关于模型 $\boldsymbol{X}$ 的二次函数, 同样连续且可导, 这意味着每一个客户可以利用该上界作为其本地优化目标, 从而在最小化模型 $\boldsymbol{X}$ 在其本地数据集上的经验风

算法 1 Complete algorithm description of FedUp

Input: FL communication round T; randomly initialized global model X^0; learning rate η; hyperparameter α;
Output: Final global model $\boldsymbol{X}^{T+1}$;

for $t = 1, 2, \ldots, T$ **do**
  Randomly select m clients into empty set $\mathcal{M}$, such that $\mathcal{M} \subseteq \{1, 2, \ldots, n\}$;
  The server issues $\boldsymbol{X}^t$ to the clients;
  for each client i in $\mathcal{M}$ **do**
    Set the optimization objective function as $h_i(\boldsymbol{X}) = F(\boldsymbol{X}^t) + \frac{\alpha}{\eta}[(\boldsymbol{X}^{t-1} - \boldsymbol{X}^t)]^\top(\boldsymbol{X} - \boldsymbol{X}^t) + \frac{\alpha}{2}\|\boldsymbol{X} - \boldsymbol{X}^t\|_2^2$;
    Perform K SGD updates $\boldsymbol{X}_i^{(t,k+1)} = \boldsymbol{X}_i^{(t,k)} - \eta\nabla h_i(\boldsymbol{X}_i^{(t,k)})$;
    Upload the latest local model $\boldsymbol{X}_i^{(t,K)}$;
  end for
  The server aggregates the local model to get the new global model $\boldsymbol{X}^{t+1} = \sum_{i=1}^m \frac{1}{m}\boldsymbol{X}_i^{(t,K)}$;
end for
return $\boldsymbol{X}^{T+1}$;

险误差时, 能同时最小化本地模型 $\boldsymbol{X}$ 在全局数据集 D 上的经验风险误差. 虽然式 (5) 并不能直接应用于实际场景下的 FL, 但它不需要强制控制梯度或模型来纠正 FL 的收敛而只需要最小化全局模型的经验误差, 这使得它有极大的潜力应用到不稳定和异构的 FL 场景中.

5.2 方案设计

式 (5) 提供了从本地优化目标的角度来开发一种适用于高度异构场景下的 FL 优化算法. 然而, 由于客户没有拥有全局数据集 D, 所以它们无法计算出全局梯度 $\nabla F(\boldsymbol{X}^t)$, 从而无法使用式 (5) 作为优化目标. 为了解决这个问题, 本文的核心想法是利用每一轮 FL 更新过程中服务器下发的全局梯度来动态地估计 $\nabla F(\boldsymbol{X}^t)$, 从而不需要交互额外的梯度信息以及牵涉到任何有关客户端的敏感数据信息. 具体地, 在每一轮 FL 过程中, 由于本地客户接收到的最新全局梯度 $\frac{\alpha}{\eta}(\boldsymbol{X}^{t-1} - \boldsymbol{X}^t)$ 记录了所有客户的数据信息, 因此可以把它看成是全局梯度 $\nabla F(\boldsymbol{X}^t)$ 的一个估计. 最小化该梯度估计的全局损失函数上界意味着保证该全局梯度尽可能在 D 上达到零, 从而全局模型在 D 上具有更好的泛化误差. 因此, 结合客户端的经验风险损失, FedUp 的每个客户端可以动态地获得如下优化目标:

$$F_i(\boldsymbol{X}) + \frac{\alpha}{\eta}[(\boldsymbol{X}^{t-1} - \boldsymbol{X}^t)]^\top(\boldsymbol{X} - \boldsymbol{X}^t) + \frac{\alpha}{2}\|\boldsymbol{X} - \boldsymbol{X}^t\|_2^2, \tag{6}$$

其中 α 是一个取值大于 0 的超参数. 注意, 由于 $F(\boldsymbol{X}^t)$ 为常数项, 不影响优化求解, 因此本文将其移除. 算法 1 提供了 FedUp 的完整流程.

值得一提的是, FedProx 有类似于 FedUp 的优化目标, 然而不同的是, FedProx 仅仅是简单地在本地经验风险损失上加了一个近端项 $\frac{\alpha}{2}\|\boldsymbol{X} - \boldsymbol{X}^t\|_2^2$, 而 FedUp 则是第一次尝试使用估计的基于 Lipschitz 条件的二次函数来动态地减少本地模型在 D 上的全局经验风险损失的误差上界, 因此带来了显著的性能提升. FedProx 仅仅保持本地模型和全局模型的相似性, 无法有效减少本地模型在 D 上的全局经验风险损失的误差, 反而会降低收敛速度 [8].

前文说到, 本文利用了每一轮 FL 更新过程中本地客户接收到的最新全局梯度 $\frac{\alpha}{\eta}(\boldsymbol{X}^{t-1} - \boldsymbol{X}^t)$ 来估计 $\nabla F(\boldsymbol{X}^t)$, 因此估计准确性至关重要. 接下来本文将从理论上分析估计误差的上界. 首先本文遵循文献 [10, 11] 的 3 个最基本的假设, 如下所示.

假设1 存在常数 $G > 0$ 和 $B \geqslant 1$, 使得 $\frac{1}{n}\sum_{i=1}^n \|\nabla F_i(\boldsymbol{X})\|_2^2 \leqslant G^2 + B^2\|\nabla F(\boldsymbol{X})\|_2^2$, $\forall \boldsymbol{X}$.

假设2 $\nabla F_i(\boldsymbol{X}; \xi_i)$ 是 F_i 的无偏梯度且有方差 $\mathbb{E}_{\xi_i}[\|\nabla F_i(\boldsymbol{X}; \xi_i) - \nabla F_i(\boldsymbol{X})\|_2^2] \leqslant \sigma^2$, $\forall i, \boldsymbol{X}$.

假设3 $\{F_i\}$ 是 μ- 平滑, 满足 $\|\nabla F_i(\boldsymbol{X}) - \nabla F_i(\boldsymbol{Y})\|_2 \leqslant \mu \|\boldsymbol{X} - \boldsymbol{Y}\|_2, \forall i, \boldsymbol{X}, \boldsymbol{Y}$.

定理1 (估计误差分析) 客户在第 t 轮使用接收到的最新全局梯度 $\frac{\alpha}{\eta}(\boldsymbol{X}^{t-1} - \boldsymbol{X}^t)$ 来估计 $\nabla F(\boldsymbol{X}^t)$ 时, 估计误差将随着迭代轮次的增加而逐渐趋于一个由超参数 α 决定的常数, 当 α 随着迭代而不断减小趋近于零时, 误差将趋近于零. 具体而言, 在非凸情况下, 如果采样率为 $\frac{m}{n}$, 通信总轮次为 T, SGD 步数为 K, 学习率 η 足够小有 $\eta = \mathcal{O}\left(\frac{1}{T}\right)$, 则估计误差有如下上界:

$$\mathbb{E}_t\left[\left\|\frac{\alpha}{\eta}(\boldsymbol{X}^{t-1} - \boldsymbol{X}^t) - \nabla F(\boldsymbol{X}^t)\right\|_2^2\right] \leqslant \min\left\{\mathcal{O}\left(\frac{1}{T^2}\right), \frac{1}{2\mu^2 K^2}\right\} C_1 + \mathcal{O}\left(\frac{C_2}{mK\sqrt{T}} + \frac{G}{T^{2/3}} + \frac{B^2}{T}\right) C_3 + \frac{K\alpha^2\sigma^2}{n}, \tag{7}$$

其中 $\mathcal{O}(*)$ 表示关于 $*$ 的高阶无穷小量, $C_1 = 18\mu^2\alpha^2K^2G^2 + 9K\mu^2\alpha^2\sigma^2$, $C_2 = (\sigma^2 + K\left(1 - \frac{m}{n}\right)G^2)^2$, $C_3 = 18\alpha^2\mu^2K^2\eta^2B^2 + (\alpha K - 1)^2$. 证明参见补充材料. 可以看到, 定理 1 刻画出了使用最新全局梯度 $\frac{\alpha}{\eta}(\boldsymbol{X}^{t-1} - \boldsymbol{X}^t)$ 来估计 $\nabla F(\boldsymbol{X}^t)$ 的有效性, 当通信轮次 T 无穷大时, 误差上界的前两项将会逐渐消失, 只剩下最后一项 $\alpha^2\frac{K\sigma^2}{n}$. 此时, 由于 α 是可以调节的超参数, 可以将其取为一个与迭代轮次 t 呈反比例的数, 则随着迭代的进行, $\alpha^2\frac{K\sigma^2}{n}$ 也将趋近于 0. 这表明, 只要训练轮次足够多, $\frac{\alpha}{\eta}(\boldsymbol{X}^{t-1} - \boldsymbol{X}^t)$ 对 $\nabla F(\boldsymbol{X}^t)$ 的估计将会越来越准确.

6 实验

6.1 实验设置

本文使用主流的 FL 平台 FedML [42] 来评估 FedUp 算法以及 9 种其他最先进的方案: MOON [35], FedProx [7], STEM [12], FedDyn [9], MimeLite [10], SCAFFOLD [11], FedSMOO [18], FedPVR [19] 和 FedAvg [1].

6.1.1 non-IID 数据划分

本文按照 FedML [42] 的方法, 使用当前主流的狄利克雷 (Dirichlet) 数据分配法为客户创建 non-IID 数据. 具体而言, 通过选择 $p_l \sim \text{Dirichlet}_n(q)$ 为 n 个客户端制造 non-IID 数据划分, 并为每个客户端 i 分配 $p_{(l,i)}$ 比例的标签为 l 的样本. 这里, q 是狄利克雷分布的浓度参数, 它控制了 non-IID 程度, q 越小意味着 non-IID 程度越大. 需要注意的是, 现有的方法, 诸如 FedDyn [9], FedProx [7] 和 FedDC [17], 只评估了一个轻度 non-IID 场景, 即所有客户端具有相同的数据集规模. 在本文的划分中, 客户端之间的数据集在规模和数据类型上都有很大的差异. 这种划分使得客户之间数据集高度异构, 由于空间有限, 本文将具体的划分结果放到了补充材料中.

6.1.2 数据集和模型

本文考虑 3 个数据集: FEMNIST [37], CIFAR-10 [43] 和 CIFAR-100 [44]. 对于 FEMNIST 数据集, 本文使用一个包含 2 个卷积层的卷积神经网络 (convolutional neural network, CNN), 其中第 1 层包含 10 个大小为 5×5 的卷积核, 第 2 层包含 20 个大小为 5×5 的卷积核, 然后是一个包含 320 个神经元的全连接层. 对于 CIFAR-10 和 CIFAR-100 数据集, 本文遵循 FedDyn [9] 的方法, 使用基于组归一化技术 [44] 的 ResNet18 [45] 网络结构. 更详细的介绍参见补充材料.

表 4　当采样率为 10% 时, 在 3 个具有不同 non-IID 程度数据集上的测试准确率 (%) 对比情况表; 注意 q 越小代表着越高的 non-IID 程度

Table 4 Comparison results of the testing accuracy (%) in a 10% sampling rate over three non-IID datasets with different non-IID degrees. Note that a smaller q indicates a higher non-IID degree[a)]

Method	FEMNIST (3400 clients), CNN	CIFAR-10 (100 clients), ResNet18				CIFAR-100 (100 clients), ResNet18					
	non-IID	$q=0.3$	$q=0.4$	$q=0.6$	IID	$q=0.02$	$q=0.1$	$q=0.3$	$q=0.4$	$q=0.6$	IID
FedAvg	78.68	72.52	73.13	75.99	80.24	27.53	33.85	39.41	39.48	41.14	41.68
FedUp	**82.74**	**78.44**	**79.15**	**81.42**	82.38	**28.89**	**40.69**	**44.94**	**45.85**	**46.61**	46.65
FedProx	78.56	73.84	74.96	76.94	80.12	18.48	33.96	39.85	37.62	38.37	41.78
SCAFFOLD	79.15	70.67	72.50	75.62	78.72	1.00	13.69	34.43	34.99	37.44	38.44
MOON	79.23	71.97	73.67	75.77	80.03	26.36	34.14	38.81	40.14	41.31	43.27
MimeLite	78.96	74.52	75.53	78.11	81.21	28.06	34.33	39.99	41.71	41.68	41.77
FedDyn	80.42	58.70	69.11	77.35	**83.31**	4.04	3.75	33.33	37.67	43.34	**46.67**
STEM	77.88	19.26	50.03	60.42	65.33	7.97	12.67	18.70	19.230	20.91	21.56
FedSMOO	78.32	78.14	76.02	79.05	82.16	27.95	35.09	41.37	43.04	44.46	44.84
FedPVR	77.62	74.06	75.82	78.90	81.64	1.00	20.95	39.03	39.67	43.16	44.38

a) Bold indicates the highest accuracy for the same comparison.

6.1.3　参数设置

为了保证公平性, 本文对所有方法都采用 SGD 来运行. 对于 FEMNIST, 本文设置总客户端数为 $n=3400$, 采样率为 10% (即每轮通信中将随机选择 10% 的客户端), 学习率 $\eta=0.02$, 批次大小为 20, 全局轮数 $T=1000$, 本地轮数为 1. 对于 CIFAR-10, 本文设置 $n=100$, 采样率为 10%, $\eta=0.2$, 批次大小为 32, $T=1500$, 本地轮数为 1. 对于 CIFAR-100, 本文设置 $n=100$, 采样率为 10%, $\eta=0.2$, 批次大小为 16, $T=1000$, 本地轮数为 1. 为了进行合理比较, 本文精心调优了 MOON (μ), FedProx (μ), STEM (c), FedDyn (α) 和 MimeLite (β) 的超参数; 这里所使用的符号遵循文献的原始符号. 更详细的介绍参见补充材料.

6.1.4　评估指标

本文使用全局模型测试准确率 (testing accuracy) 来评估所有方法. 测试准确率是分类器准确分类样本数量与测试数据集总样本数量之比.

6.2　实验结果

6.2.1　大规模准确率比较

表 4 比较了所有方法在 3 个不同 non-IID 数据集上的测试准确率. 结果表明, FedUp 在 non-IID 情况下始终表现优于所有竞争对手. 具体来说, FedAvg, FedProx, SCAFFOLD, MimeLite, FedSMOO, FedPVR 和 MOON 在 3 个数据集上保持相近的准确率, 但在不同数据集上, 它们的准确率平均低于 FedUp 约 2%∼5%. STEM 除了在 CIFAR-100 上 q 为 0.02 和 0.1 时比 FedDyn 准确率高之外, 在其他数据集上表现一直最差. FedDyn 在 IID 设置中表现略优于 FedUp, 但在 non-IID 情况下准确率较低. 这是因为 non-IID 设置会加剧不稳定参与客户对 FedDyn 的负面影响.

6.2.2　对不稳定客户参与具有鲁棒性

不稳定的客户端参与, 会严重影响模型的收敛和性能. 为验证 FedUp 在这种情况下的稳定性, 本文在 FEMNIST (3400 个客户端) 和 CIFAR-10 (1000 个客户端) 上考虑了 3 种更小的采样率 (即 0.2%,

表 5 在 non-IID 的 FEMNIST 和 CIFAR-10 数据集上, 采样率对测试准确率 (%) 的影响. 注意, 本文通过设置 $q = 0.8$ 为 1000 个客户端划分 CIFAR-10, 而 FedDyn 需要同时运行 1000 个 ResNet18 模型, 这远远超出了计算资源的上限, 因此将它省去

Table 5 Impact of sampling rates on the testing accuracy (%) over non-IID FEMNIST and CIFAR-10 datasets. Note that we partition CIFAR-10 for 1000 clients by setting $q = 0.8$. And FedDyn needs to run 1000 ResNet18 models simultaneously, which exceeds the upper limit of our computation resources. We thus omit it

Method	FEMNIST, CNN, 3400 clients			CIFAR-10, ResNet18, 1000 clients		
	Sampling rate			Sampling rate		
	0.2%	0.5%	1%	0.2%	0.5%	1%
FedAvg	78.88	79.38	79.65	70.70	74.09	75.81
FedUp	**81.27**	**82.43**	**82.71**	**72.72**	**77.41**	**82.24**
FedProx	78.53	79.38	79.88	50.02	63.53	67.97
SCAFFOLD	78.71	79.74	79.96	70.71	76.13	80.13
MOON	78.69	78.99	79.05	70.42	72.27	73.90
MimeLite	78.26	78.87	79.98	53.91	70.68	75.26
FedDyn	79.77	80.01	80.84	–	–	–
STEM	77.58	77.68	78.88	10.00	10.00	10.00
FedSMOO	79.56	80.64	80.95	71.37	76.64	80.76
FedPVR	78.70	79.93	80.56	70.20	76.59	80.21

0.5% 和 1%). 表 5 的结果表明, 当采样率降低到 0.2% 时, 所有方法都无法改善基线 FedAvg, 而本文提出的 FedUp 表现更好, 比 FedAvg 高至少 2% 的准确性. FedUp 能在更不稳定的客户端参与情况下具有鲁棒性, 这是因为最小化全局损失函数的上界可以大大减小模型泛化的误差方差.

6.2.3 对高度 non-IID 数据具有鲁棒性

图 1 展示了所有方法在不同 non-IID 程度下完整的收敛曲线. 可以观察到, 在所有 non-IID 的设置中, FedUp 在所有数据集上都实现了最好的准确率, 而所有其他方法只实现了和 FedAvg 相似的性能, 有的方法甚至有更差的性能. 注意到, MimeLite 的收敛速度比本文提出的方法快, 但它未能显著提高模型性能. 这是因为使用动量只能加速收敛速度, 但无法极大地缓解本地模型之间的差异. 本文提出的 FedUp 独辟蹊径地最小化全局损失的上界, 使得 FL 更可能有效地收敛到全局损失函数更低的驻点, 从而降低了测试误差.

6.2.4 对大规模客户参与具有鲁棒性

为了验证 FedUp 对大规模客户同时参与聚合的可扩展性, 本文考虑在具有 3400 个客户端的 FEMNIST 上同时采样大规模的客户端参与聚合: 选择 500 个客户参与聚合和选择 1020 个客户参与聚合. 图 2 详细绘制了在测试集上的准确率曲线. 可以观察到, 当大规模客户端参与时, 现有的方法只能达到与 FedAvg 相似的性能或者无法收敛 (例如 STEM 和 SCAFFOLD). 本文提出的 FedUp 比所有竞争对手更快地收敛到最优点, 并在这两种情况下实现了更高的准确性. 需要注意的是, 本小节没有把 FedDyn 纳入对比中, 这是因为它需要大量的计算开销, 远远超出了计算资源可支持的上限.

6.2.5 场景适用性分析

当前, FedUp 算法在实验中使用的数据集包括 FEMNIST, CIFAR-10 和 CIFAR-100, 这些数据集

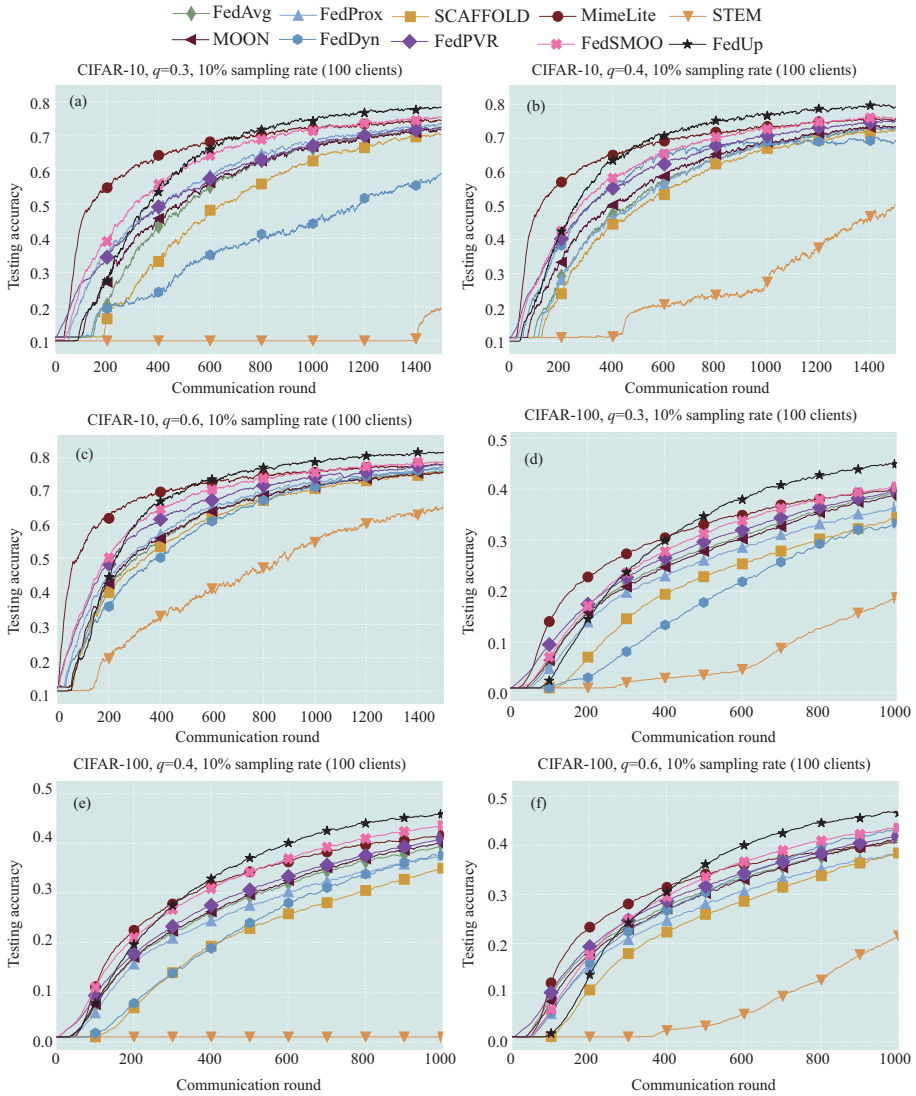

图 1　(网络版彩图) non-IID 程度对 CIFAR-10 和 CIFAR-100 数据集上全局模型测试准确率的影响

Figure 1　(Color online) Impact of non-IID degrees on the testing accuracy of global model over CIFAR-10 and CIFAR-100. (a) CIFAR-10, $q = 0.3$; (b) CIFAR-10, $q = 0.4$; (c) CIFAR-10, $q = 0.6$; (d) CIFAR-100, $q = 0.3$; (e) CIFAR-100, $q = 0.4$; (f) CIFAR-100, $q = 0.6$

都是图像领域的, 并且被用于分类任务. 因此, 尚不清楚 FedUp 算法是否适用于其他非图像和非分类任务的场景. 为了进一步验证 FedUp 算法的应用范围, 本节在经典的 Shakespeare 文本数据集上进行了下一个词预测任务. Shakespeare 数据集 [7] 源自威廉 · 莎士比亚戏剧作品的文集《威廉 · 莎士比亚全集》. 在这个数据集中, 每个会说话的角色都被视为一个客户, 因此由于每个客户的说话风格不同, 该数据集自然而然地具有高度异构性. 在本节实验中, 遵循文献 [7] 的设置, 使用一个包含 100 个隐藏单元, 一个 8 维嵌入层和总共 80 个类的两层 LSTM 分类器来学习这个数据集. 图 3 展示了 FedUp 和基线 FedAvg 在准确率方面的比较情况. 可以观察到, 在文本预测任务中, FedUp 相对于基线 FedAvg 表现出更快的收敛速度和更高的准确率. 这表明 FedUp 在不同的数据类型和任务上具有广泛的适用性. 这是因为 FedUp 只是修改了 FedAvg 的优化目标且不需要关于底层任务和数据的先验知识, 因此

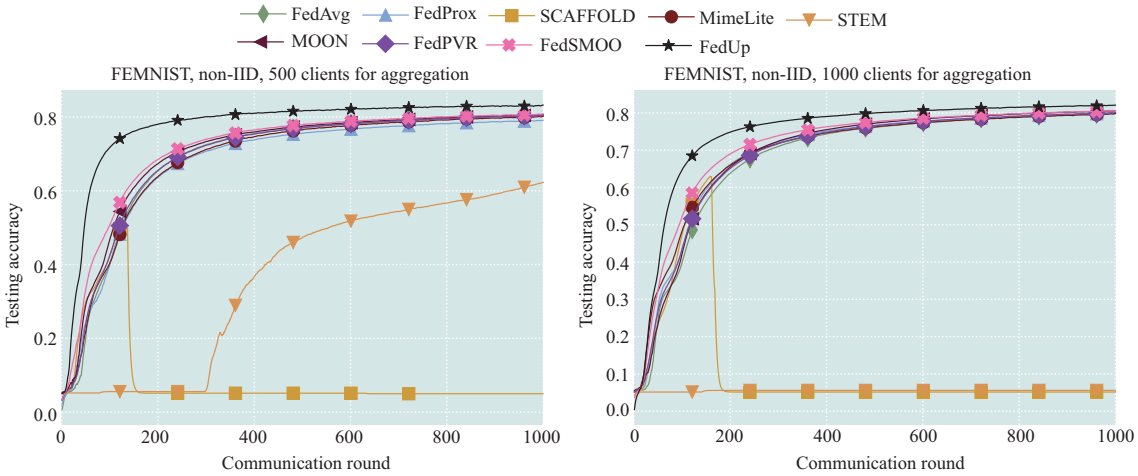

图 2　(网络版彩图) 在 non-IID 的 FEMNIST 数据集上大规模客户端对测试准确率的影响

Figure 2　(Color online) Impact of the large-scale clients on the testing accuracy over non-IID FEMNIST

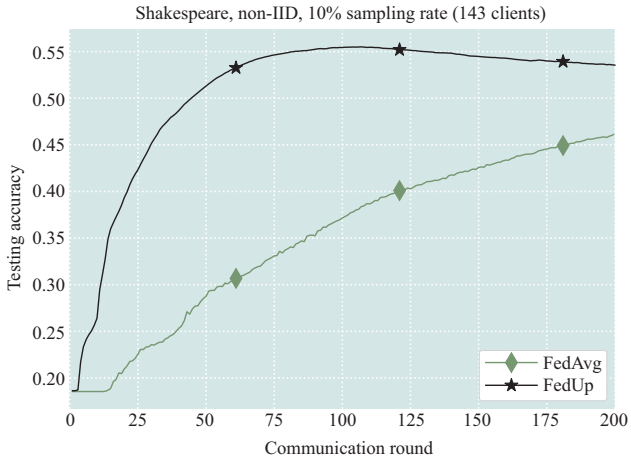

图 3　(网络版彩图) Shakespeare 数据集上测试准确率

Figure 3　(Color online) Testing accuracy on Shakespeare dataset

通常来说, FedUp 应该会和 FedAvg 一样具有广泛的适用场景. 总的来说, 无论是面对文本预测还是其他图像分类任务, FedUp 都是一个值得考虑的选项, 可以在各种不同的数据类型和任务中得到应用.

7　讨论和未来的工作

虽然 FedUp 是一种实用且有效的方法, 可以克服现有 FL 方法不切实际的局限性, 但据本文所知, FedUp 使用的梯度, 即 $\boldsymbol{X} - \boldsymbol{X}^t$, 在文献中还没有有效的假设来约束其有界, 因此, 在解释其收敛速度等性能方面提出了理论上的挑战. 以前的方法通常通过直接使用 $\nabla F(\boldsymbol{X}^t, \xi^k)$ 这个准确的梯度, 或利用控制变量来规避 $\boldsymbol{X} - \boldsymbol{X}^t$ 的使用, 但是这些解决方案需要不切实际的假设, 例如稳定的客户端参与和简单的数据集, 严重限制了它们的可行性, 正如第 3 节所讨论的那样.

相比之下, FedUp 试图开发一种实用的方法, 可以应用到只有全局梯度 $\boldsymbol{X} - \boldsymbol{X}^t$ 可以利用的现实

场景中, 因此不得不直接利用该全局梯度来作为 $\nabla F(\boldsymbol{X}^t)$ 的估计. 然而, 保证这种梯度有界需要进一步的理论创新, 这超出了本文的研究范围. 本文把这一重要任务留给接下来的工作.

总的来说, 最近在 FL 领域的研究进展取得了重大的理论突破 [9~11], 但这些突破都以牺牲实用性为代价, 比如, 假设准确的 $\nabla F(\boldsymbol{X}^t, \xi^k)$ 可用后, Mime 的收敛速度能超过任何数据集中式训练方法收敛速度的下界, 但却只能应用于简单的数据集中. 为了真正推动 FL 的发展, 必须同时考虑算法在现实场景中的实用性和可解释性. 然而, 到目前为止, 这一重要领域在很大程度上仍未得到探索, 因此需要开发兼顾实用性和可解释性的 FL 方法. 本文相信, FedUp 在借鉴理论客观依据的同时优先考虑实用性, 可以为未来旨在推进这一重要问题的工作提供灵感来源.

8 结束语

本文对现有的 FL 优化方法进行了全面比较, 并揭示了它们在解决非独立同分布数据时存在的许多不切实际的限制. 为了弥补 FL 和其实际使用中存在的巨大差距, 本文提出了 FedUp 来致力于最小化全局损失函数的上界, 使得 FL 能收敛到测试误差更小的驻点. 为了验证其有效性和鲁棒性, 本文使用了真实生活中存在的大规模数据集进行了大量实验, 结果表明, FedUp 在面对非独立同分布数据, 大规模和不稳定的客户参与时远优于最先进的算法.

补充材料 本文的补充材料见网络版 infocn.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.

参考文献

1 Mcmahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 2017. 1273–1282

2 Gao S, Yuan L P, Zhu J M, et al. A blockchain-based privacy-preserving asynchronous federated learning. Sci Sin Inform, 2021, 51: 1755–1774 [高胜, 袁丽萍, 朱建明, 等. 一种基于区块链的隐私保护异步联邦学习. 中国科学: 信息科学, 2021, 51: 1755–1774]

3 Feng J, Cai Q Z, Jiang Y. Towards training time attacks for federated machine learning systems. Sci Sin Inform, 2021, 51: 900–911 [冯霁, 蔡其志, 姜远. 联邦学习下对抗训练样本表示的研究. 中国科学: 信息科学, 2021, 51: 900–911]

4 Byrd D, Polychroniadou A. Differentially private secure multi-party computation for federated learning in financial applications. In: Proceedings of the 1st ACM International Conference on AI in Finance, New York, 2020. 1–9

5 Liang F, Yang E Y, Pan W K, et al. Survey of recommender systems based on federated learning. Sci Sin Inform, 2022, 52: 713–741 [梁锋, 羊恩跃, 潘微科, 等. 基于联邦学习的推荐系统综述. 中国科学: 信息科学, 2022, 52: 713–741]

6 Xu J, Glicksberg B S, Su C, et al. Federated learning for healthcare informatics. J Healthc Inform Res, 2021, 5: 1–19

7 Li T, Sahu A K, Zaheer M, et al. Federated optimization in heterogeneous networks. In: Proceedings of Machine Learning and Systems, Austin, 2020. 429–450

8 Wang J, Liu Q, Liang H, et al. Tackling the objective inconsistency problem in heterogeneous federated optimization. In: Proceedings of the 33rd Annual Conference in Neural Information Processing Systems, Virtual, 2020. 7611–7623

9 Acar D A E, Zhao Y, Matas R, et al. Federated learning based on dynamic regularization. In: Proceedings of the 9th International Conference on Learning Representations, Virtual Event, 2021. 1–20

10 Karimireddy S P, Jaggi M, Kale S, et al. Breaking the centralized barrier for cross-device federated learning.

In: Proceedings of the 34th Annual Conference on Neural Information Processing Systems, Virtual, 2021. 28663–28676

11 Karimireddy S P, Kale S, Mohri M, et al. SCAFFOLD: stochastic controlled averaging for federated learning. In: Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 2020. 5132–5143

12 Khanduri P, Sharma P, Yang H, et al. STEM: a stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning. In: Proceedings of the 34th Annual Conference on Neural Information Processing Systems, Virtual, 2021. 6050–6061

13 Pathak R, Wainwright M J. FedSplit: an algorithmic framework for fast federated optimization. In: Proceedings of the 33rd Annual Conference Advances in Neural Information Processing Systems, Virtual, 2020. 7057–7066

14 Zhao Y, Li M, Lai L, et al. Federated learning with non-IID data. 2018. ArXiv:1806.00582

15 Itahara S, Nishio T, Koda Y, et al. Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-IID private data. IEEE Trans Mobile Comput, 2023, 22: 191–205

16 Eunjeong J, Seungeun O, Hyesung K, et al. Communication-efficient on-device machine learning: federated distillation and augmentation Under non-iid private data. 2018. ArXiv:1811.11479

17 Gao L, Fu H, Li L, et al. FedDC: federated learning with non-IID data via local drift decoupling and correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, 2022. 10112–10121

18 Sun Y, Shen L, Chen S X, et al. Dynamic regularized sharpness aware minimization in federated learning: approaching global consistency and smooth landscape. In: Proceedings of the 44th International Conference on Machine Learning, Honolulu, 2023. 32991–33013

19 Li B, Mikkel N S, Tommy S A, et al. On the effectiveness of partial variance reduction in federated learning with heterogeneous data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, 2023. 3964–3973

20 Dai Y T, Chen Z Y, Li J N, et al. Tackling data heterogeneity in federated learning with class prototypes. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence, Washington, 2023. 7314–7322

21 Guo Y X, Tang X Y, Lin T. FedBR: improving federated learning on heterogeneous data via local learning bias reduction. In: Proceedings of the 44th International Conference on Machine Learning, Honolulu, 2023. 12034–12054

22 Wang S, Xu Y Q, Wang Z G, et al. Beyond ADMM: a unified client-variance-reduced adaptive federated learning framework. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence, Washington, 2023. 10175–10183

23 Wang H Y, Mikhail Y, Sun Y K, et al. Federated learning with matched averaging. In: Proceedings of the 8th International Conference on Learning Representations, Addis Ababa, 2020

24 Chen H Y, Chao W L. FedBE: making Bayesian model ensemble applicable to federated learning. 2020. ArXiv:2009.01974

25 Gong X, Sharma A, Karanam S, et al. Ensemble attention distillation for privacy-preserving federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2021. 15076–15086

26 Li X, Jiang M, Zhang X, et al. FedBN: federated learning on non-IID features via local batch normalization. In: Proceedings of the 8th International Conference on Learning Representations, Addis Ababa, 2020. 1–16

27 Wang H, Kaplan Z, Niu D, et al. Optimizing federated learning on non-iid data with reinforcement learning. In: Proceedings of the 39th IEEE Conference on Computer Communications, Toronto, 2020. 1698–1707

28 Nishio T, Yonetani R. Client selection for federated learning with heterogeneous resources in mobile edge.

In: Proceedings of the IEEE International Conference on Communications, Shanghai, 2019. 1–7

29 Lai F, Zhu X, Madhyastha H V, et al. Oort: efficient federated learning via guided participant selection. In: Proceedings of the 15th USENIX Symposium on Operating Systems Design and Implementation, 2021. 19–35

30 Rizk E, Vlaski S, Sayed S H. Optimal importance sampling for federated learning. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, 2021. 3095–3099

31 Sattler F, Muller K R, Samek W. Clustered federated learning: model-agnostic distributed multitask optimization under privacy constraints. IEEE Trans Neural Netw Learn Syst, 2021, 32: 3710–3722

32 Briggs C, Fan Z, Andras P. Federated learning with hierarchical clustering of local updates to improve training on non-IID data. In: Proceedings of the International Joint Conference on Neural Networks, Glasgow, 2020. 1–9

33 Collins L, Hassani H, Mokhtari A, et al. Exploiting shared representations for personalized federated learning. In: Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 2021. 2089–2099

34 Zhang J, Hua Y, Wang H, et al. FedALA: adaptive local aggregation for personalized federated learning. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence, Washington, 2023. 11237–11244

35 Li Q, He B, Song D. Model-contrastive federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2021. 10713–10722

36 Nasirigerdeh R, Bakhtiari M, Torkzadehmahani R, et al. Federated multi-mini-batch: an efficient training approach to federated learning in non-IID environments. 2020. ArXiv:2011.07006

37 Caldas S, Duddu S M K, Wu P , et al. LEAF: a benchmark for federated settings. 2018. ArXiv:1812.01097

38 Ma X, Zhu J, Lin Z, et al. A state-of-the-art survey on solving non-IID data in federated learning. Future Generation Comput Syst, 2022, 135: 244–258

39 Zhang X, Li Y, Li W, et al. Personalized federated learning via variational Bayesian inference. In: Proceedings of the International Conference on Machine Learning, Baltimore, 2022. 26293–26310

40 Marfoq O, Neglia G, Vidal R, et al. Personalized federated learning through local memorization. In: Proceedings of the International Conference on Machine Learning, Baltimore, 2022. 15070–15092

41 Ni X M, Shen X Y, Zhang H. Adaptive personalized federated learning for heterogeneous data: a method based on parameter decomposition and continual learning. Sci Sin Inform, 2022, 52: 2306–2320 [倪宣明, 沈鑫圆, 张海. 面向异构数据的自适应个性化联邦学习 —— 一种基于参数分解和持续学习的方法. 中国科学: 信息科学, 2022, 52: 2306–2320]

42 He C, Li S, So J, et al. FedML: a research library and benchmark for federated machine learning. 2020. ArXiv:2007.13518

43 Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. 2009. ArXiv:2009.12532

44 Wu Y, He K. Group normalization. In: Proceedings of the 15th European Conference, Munich, 2018. 3–19

45 He K M, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 770–778

Enhancing generalization robustness of federated learning in highly heterogeneous environments

Wei WAN[1,4,5,6,7], Shengshan HU[1,4,5,6,7*], Jianrong LU[1,4,5,6,7], Minghui LI[2],
Ziqi ZHOU[1,4,5,6,7] & Hai JIN[3,4,5,8]

1. *School of Cyber Science and Engineering, Huazhong University of Science and Technology, Wuhan* 430074*, China*;
2. *School of Software Engineering, Huazhong University of Science and Technology, Wuhan* 430074*, China*;
3. *School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan* 430074*, China*;
4. *National Engineering Research Center for Big Data Technology and System, Wuhan* 430074*, China*;
5. *Services Computing Technology and System Lab, Wuhan* 430074*, China*;
6. *Hubei Key Laboratory of Distributed System Security, Wuhan* 430074*, China*;
7. *Hubei Engineering Research Center on Big Data Security, Wuhan* 430074*, China*;
8. *Cluster and Grid Computing Lab, Wuhan* 430074*, China*
* Corresponding author. E-mail: hushengshan@hust.edu.cn

Abstract Federated learning (FL) is a distributed processing network that focuses on protecting client privacy data, providing a promising solution for addressing privacy leakage issues. However, a major quagmire in FL is to train clients' models over significantly non-independent and identically distributed (non-IID) data, which would lead to a low-performance global model. Although this issue has been investigated by many previous works, this paper finds that they have little or no performance improvement over the standard baseline FedAvg when facing highly non-IID data, unstable client participation, and deep models, seriously hindering the privacy protection application value of FL. To address this issue, a new solution called FedUp has been proposed. FedUp is a robust optimization solution for non-IID FL that improves the generalization robustness of the global model while retaining the privacy protection characteristics of FL. FedUp minimizes the upper bound of the global empirical loss function to ensure that the models exhibit smaller generalization errors. Simulation experiments show that FedUp achieves significant advantages over state-of-the-art methods, and is robust to highly non-IID data as well as unstable and large-cohort client participation. This solution has the potential to improve the performance of FL and make it more practical for privacy protection applications.

Keywords distributed network, federated learning, heterogeneous optimization, generalization, robustness, privacy protection