异质且相关的客户端可用性下的联邦学习

Angelo Rodio*、Francescomaria Faticanti*、Othmane Marfoq*†、Giovanni Neglia*、Emilio Leonardi‡
*Inria，蔚蓝海岸大学´，法国。电子邮箱：{firstname.lastname}@inria.fr,
†埃森哲实验室，索菲亚-安提波利斯，法国。电子邮箱：{firstname.lastname}@accenture.com,
‡都灵理工大学，都灵，意大利。电子邮箱：{firstname.lastname}@polito.it

arXiv:2301.04632v1 [cs.LG] 11 Jan 2023

摘要——**移动设备和物联网设备产生的海量数据推动了联邦学习（FL）的发展。这一框架允许此类设备（或客户端）在不共享本地数据的情况下协同训练机器学习模型。联邦学习算法（如FedAvg）会迭代地聚合客户端基于各自数据集计算的模型更新。客户端可能表现出不同程度的参与度，并且这种参与度通常会随时间变化，且与其他客户端相关。本文首次针对异质且相关的客户端可用性下类似 FedAvg 的联邦学习算法开展收敛分析。我们的分析揭示了相关性如何不利地影响算法的收敛速率，以及聚合策略如何缓解这种影响，但代价是使训练趋向于有偏模型。在理论分析的指导下，我们提出了 CA-Fed，这是一种新的联邦学习算法，旨在平衡最大化收敛速度与最小化模型偏差这两个相互冲突的目标。为此，CA-Fed 会动态调整分配给每个客户端的权重，并可能忽略可用性低且相关性高的客户端。我们的实验结果表明，在合成数据集和真实数据集上，与当前最先进的 AdaFed 和 F3AST 相比，CA-Fed 均能获得更高的时间平均准确率和更低的标准差。**

关键词—**联邦学习，分布式优化。**

I. 引言

移动设备和物联网设备生成的海量数据，推动了分布式机器学习训练范式的兴起 [1], [2]。联邦学习（FL） [3]–[6] 是一种新兴框架，地理位置分散的设备（或客户端）可以参与训练共享的机器学习（ML）模型，而无需共享其本地数据。提出联邦学习旨在降低收集大量数据的总体成本，同时保护用户可能敏感的私人信息。在最初的联邦平均算法（FedAvg） [4], 中，中央服务器从可用客户端集合中随机选择一部分客户端，并向其广播共享模型。被选中的客户端在各自的本地数据集上执行若干独立的随机梯度下降（SGD）步，然后将本地模型更新发送回服务器。随后，服务器聚合收到的客户端更新，生成新的全局模型，并开始新一轮训练。在 FedAvg 的每次迭代中，服务器通常会随机抽取数百台设备参与训练 [7], [8]。

本研究得到了法国政府的资助，资助来源包括法国国家研究署（ANR）通过“未来投资”项目 3IA Côte d’Azur（编号：ANR-19-P3IA-0002）提供的资金，以及法国邮政集团作为 Inria基金会赞助方，在 FedMalin Inria挑战赛框架下提供的资助。本研究的初步版本已被 IEEE INFOCOM 2023 接收。

在真实场景中，客户端的可用性/活动状态由协调服务器无法控制且难以预测的外部因素决定。例如，通常只有处于空闲状态、正在充电且连接到宽带网络的智能手机才允许参与训练过程 [4], [9]。这些资格要求可能导致设备的可用性在时间和空间上相关[7], [10]–[12]。例如，时间相关性可能源于智能手机连续充电数小时，随后在当天剩余时间内不符合参与资格。类似地，由可再生能源供电的传感器，其活动状态可能取决于随时间呈现固有相关性的自然现象（例如太阳光）。空间相关性则指不同客户端之间的相关性，这通常是用户地理分布不同所导致的。例如，处于同一时区的客户端通常表现出相似的可用性模式，例如受到一天中不同时段的影响。

众所周知，数据采样过程中的时间相关性会对机器学习训练的性能产生负面影响，即使是在集中式设置中也是如此 [13], [14]，并且可能导致灾难性遗忘：在训练的最后阶段使用的数据可能会对最终模型产生不成比例的影响，“抹去”先前学到的信息 [15], [16]。在联邦学习中也观察到了灾难性遗忘：同一地理区域内的客户端，其本地数据分布更加相似，而且客户端参与遵循循环的每日模式（从而导致空间相关性） [7], [10], [11], [17]。尽管有这些证据，目前仍缺乏对时间和空间相关的客户端参与情况下联邦学习算法收敛性的理论研究。

本文首次对异质且相关的客户端可用性条件下的 FedAvg [4] 进行了收敛分析。我们假设客户端的时间和空间可用性遵循任意有限状态马尔可夫链：这一假设刻画了客户端活动彼此相关的真实场景，同时仍使系统具有解析可处理性。我们的理论分析：(i) 通过一个额外项量化相关性对算法收敛速度的负面影响，该项取决于马尔可夫链的谱性质；(ii) 指出两个相互冲突的目标之间存在权衡：缓慢收敛至最优模型，或快速收敛至有偏模型，即收敛至一个最小化不同于初始目标的目标函数的模型。受理论分析结果的启发，

我们提出 CA-Fed，这是一种动态为客户端分配权重的算法，能够在最大化收敛速度和最小化模型偏差之间实现良好权衡。有趣的是，CA-Fed 可以选择忽略可用性低且时间相关性高的客户端。我们的实验结果表明，排除这类客户端是应对联邦学习中异质且相关的客户端可用性的一种简单而有效的方法。事实上，尽管 CA-Fed 达到的最高准确率与最先进的方法 F3AST [18] 和 AdaFed [19]，相当，但其测试准确率的时间平均值更高，且随时间变化的变异性更小。

本文其余部分安排如下。第二节描述联邦学习中客户端可用性相关性问题，并讨论主要相关工作。第 III 节对异质且相关的客户端参与情况下的 FedAvg 进行收敛分析。第四节介绍我们提出的相关性感知联邦学习算法 CA-Fed。我们在第五节评估 CA-Fed，并在合成数据和真实世界数据上将其与最先进的方法进行比较。第七节总结全文。

II. 背景与相关工作

我们考虑一个由 $\mathcal{K}$ 个 N 客户端组成的有限集合。每个客户端 $k \in \mathcal{K}$ 持有一个本地数据集 D_k。客户端旨在共同学习全局机器学习模型的参数 $\boldsymbol{w} \in W \subseteq \mathbb{R}^d$（例如，神经网络架构的权重）。在训练过程中，参数为 $\boldsymbol{w}$ 的模型在数据样本 $\xi \in D_k$ 上的质量由损失函数 $f(\boldsymbol{w}; \xi)$ 衡量。在中央服务器的协调下，客户端求解以下优化问题：

$$\min_{\boldsymbol{w} \in W \subseteq \mathbb{R}^d} \left[F(\boldsymbol{w}) := \sum_{k \in \mathcal{K}} \alpha_k F_k(\boldsymbol{w}) \right], \tag{1}$$

其中，$F_k(\boldsymbol{w}) := \frac{1}{|D_k|} \sum_{\xi \in D_k} f(\boldsymbol{w}; \xi)$ 是在客户端 k 的本地数据集上计算的平均损失，$\boldsymbol{\alpha} = (\alpha_k)_{k \in \mathcal{K}}$ 是满足 $\sum_k \alpha_k = 1$ 的正系数。它们表示中央服务器分配给每个客户端 k 的目标重要性。通常，$(\alpha_k)_{k \in \mathcal{K}}$ 与客户端数据集的大小 $|D_k|$ 成正比，因此，式 (1) 中的目标函数 F 等于在所有客户端本地数据集的并集 $D = \cup_{k \in \mathcal{K}} D_k$ 上计算的平均损失。

在第三节中详述的适当假设下，问题 (1) 存在唯一解。我们用 $\boldsymbol{w}^*$（分别用 F^*）表示 F 的最小化器（分别表示其最小值）。此外，对于 $k \in \mathcal{K}$，F_k 在 W 上存在唯一最小化器。我们用 $\boldsymbol{w}_k^*$（分别用 F_k^*）表示 F_k 的最小化器（分别表示其最小值）。

问题 (1) 通常通过迭代算法求解 [4], [8]，这类算法 [4], [8] 需要服务器与客户端之间进行多轮通信。在第 $t > 0$ 轮，服务器将全局模型的最新估计值 $\boldsymbol{w}_{t,0}$ 广播给可用客户端集合 (A_t)。客户端 $k \in A_t$ 通过执行 $E \geq 1$ 步本地随机梯度下降（SGD），利用其本地数据更新全局模型：

$$\boldsymbol{w}_{t,j+1}^k = \boldsymbol{w}_{t,j}^k - \eta_t \nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k) \quad j = 0, \ldots, E-1, \tag{2}$$

其中 $\eta_t > 0$ 是经过适当选择的学习率，称为本地学习率；$\mathcal{B}_{t,j}^k$ 是在第 t 轮、第 j 步从客户端 k 的本地数据集中采样得到的随机批次；$\nabla F_k(\cdot, \mathcal{B}) := \frac{1}{|\mathcal{B}|} \sum_{\xi \in \mathcal{B}} \nabla f(\cdot, \xi)$ 是本地梯度 ∇F_k 的无偏估计量。随后，每个客户端将其本地模型更新 $\Delta_t^k := \boldsymbol{w}_{t,E}^k - \boldsymbol{w}_{t,0}^k$ 发送给服务器。服务器计算 $\Delta_t := \sum_{k \in A_t} q_k \cdot \Delta_t^k$，即客户端本地更新的加权平均，其中使用了非负的聚合权重 $\boldsymbol{q} = (q_k)_{k \in \mathcal{K}}$。聚合权重的选择定义了一种聚合策略（我们稍后会讨论不同的聚合策略）。聚合后的更新 Δ_t 可视为 $-\nabla F(\boldsymbol{w}_{t,0})$ 的代理量；服务器将其应用于全局模型：

$$\boldsymbol{w}_{t+1,0} = \mathbf{Proj}_W(\boldsymbol{w}_{t,0} + \eta_s \cdot \Delta_t) \tag{3}$$

其中，$\mathbf{Proj}_W(\cdot)$ 表示在集合 W 上的投影，而 $\eta_s > 0$ 是经过适当选择的学习率，称为服务器学习率。[1]

聚合更新 Δ_t 通常是 $-\nabla F(\boldsymbol{w}_{t,0})$ 的有偏估计量，其中每个客户端 k 均按照其在集合 A_t 中的出现频率以及其聚合权重 q_k 按比例计入。事实上，在第三节中规定的适当假设下，可以证明（见定理 2），由（2）和（3）所描述的更新规则收敛于有偏全局目标函数 F_B 的唯一极小值点；该目标函数同时取决于客户端的可用性（即序列 $(A_t)_{t>0}$）和聚合策略（即 $\boldsymbol{q} = (q_k)_{k \in \mathcal{K}}$）：

$$F_B(\boldsymbol{w}) := \sum_{k=1}^N p_k F_k(\boldsymbol{w}), \text{ with } p_k := \frac{\pi_k q_k}{\sum_{h=1}^N \pi_h q_h}, \tag{4}$$

其中，$\pi_k := \lim_{t \to \infty} \mathbb{P}(k \in A_t)$ 表示客户端 k 的渐近可用性。系数 $\boldsymbol{p} = (p_k)_{k \in \mathcal{K}}$ 可以解释为服务器在训练过程中赋予每个客户端 k 的有偏重要性，通常不同于目标重要性 $\boldsymbol{\alpha}$。下文中，$\boldsymbol{w}_B^*$（分别为 F_B^*）表示 F_B 的最小化器（分别为最小值）。

在一些大规模联邦学习应用中，例如训练 Google 键盘的下一个词预测模型，每个客户端最多只参与一轮训练。协调器通常在每轮中选取几百个客户端，进行几千轮（例如，参见 [5,表 2]），但可用客户端集合可能包含数亿台 Android 设备。在这种情况下，除非掌握有关每个客户端可用性的某些先验信息，否则很难应对潜在的偏差。无论如何，服务提供商也可以将联邦学习用于规模小得多的客户端集合（例如，安装了特定应用的智能手机用户）。在这种情况下，客户端会多次参与训练：协调服务器可以跟踪每个客户端的可用性，并尝试补偿其参与情况中可能存在的严重异质性。

此前关于联邦学习 [4], [17]–[19], [22]–[25] 的大量研究都探讨了这一问题，并在不同的

[1] 这一聚合规则（3）也已在其他研究中得到考虑，例如，[20], [21]。在其他联邦学习算法中，服务器会对客户端的本地模型求平均。对（3）稍作修改即可得到这一聚合规则。

客户端可用性假设（即关于 $(A_t)_{t>0}$ 的假设）下，设计了聚合策略，通过适当选择 $\boldsymbol{q}$，使 Δ_t 的估计无偏。文献 [22] 首次分析了客户端部分参与时 FedAvg 在非独立同分布数据上的表现。他们的分析涵盖两种情况：一种是从 $\mathcal{K}$ 中以均匀随机、不放回的方式抽取活跃客户端，并赋予其等于目标重要性的聚合权重（如 [4]中所假设）；另一种是从 $\mathcal{K}$ 中以概率 $\boldsymbol{\alpha}$独立同分布、有放回地抽取活跃客户端，并赋予相同权重（如 [23]中所假设）。然而，文献 [4], [22], [23] 忽略了客户端随机可用性所引入的方差。[24]的作者仅考虑更新重要的客户端（以更新范数衡量），从而降低了这类方差。文献 [17] 和 [25] 分别通过聚类采样和软聚类采样降低了聚合方差。

一些近期研究 [18], [19], [26]并未积极追求无偏目标函数的优化。相反，它们推导收敛误差的上界，并提出启发式方法来最小化这些上界，但这可能会引入一定偏差。我们的工作采用了类似的思路：将我们的算法与 [18] 中的 F3AST 和 [19]中的 AdaFed 进行比较。

我们研究的新颖之处在于考虑了客户端可用性动态中的空间相关性和时间相关性。正如引言中所述，客户端的资格条件也会引入这类相关性，例如，智能手机处于充电状态并连接到宽带网络。迄今为止，人们一直忽视相关性的影响，可能是因为研究联邦学习算法收敛时会增加额外的复杂性。据我们所知，唯一的例外是 [18]，该研究通过针对客户端可用性不相关和正相关的情形分别提出两种不同的算法，初步探讨了空间相关性问题（随着相关程度变化，这两种算法之间并不存在平滑过渡）。

时间相关性对集中式随机梯度方法的影响已在 [12]–[14], [27] 中得到研究：这些工作探讨了一种随机梯度下降变体，其中样本按照马尔可夫链抽取。参考文献 [12] 将其分析扩展到联邦学习场景，其中每个客户端都按照马尔可夫链抽取样本。相比之下，我们的工作并不假设数据采样存在相关性，而是假设客户端的可用性存在相关性。不过，我们的一些证明技巧与这一研究方向所采用的技巧相似，尤其是我们借鉴了 [14]中的一些结果。

III. 分析

A. 主要假设

我们考虑一个时隙系统，其中一个时隙对应一个联邦学习通信轮次。我们假设客户端在时隙 $t \in \mathbb{N}$ 上的可用性遵循离散时间马尔可夫链$(A_t)_{t\geq 0}$。[2]

[2]在第 III 节-D 中，我们将重点讨论这样的情形：该链是 N 相互独立的马尔可夫链的叠加，每个客户端对应一条链。

假设 1. 有限状态空间 $[M]$ 上的马尔可夫链 $(A_t)_{t\geq 0}$ 是时齐、不可约且非周期的。其转移矩阵为 $\boldsymbol{P}$，平稳分布为 $\boldsymbol{\pi}$。

文献中已采用马尔可夫链来建模随机网络的动态，其中图中的某些节点或边可以在激活与非激活状态之间切换 [28], [29]。先前的马尔可夫假设在提供高度灵活性的同时，仍保证了系统的解析可处理性。马尔可夫过程的当前分布与平稳分布之间的距离动态，可以通过其转移矩阵的谱性质来刻画 $\boldsymbol{P}$ [30]。令 $\lambda_2(\boldsymbol{P})$ 表示 $\boldsymbol{P}$ 按绝对值排序后的第二大特征值。先前的工作 [14] 已证明：

$$\max_{i,j\in[M]} |[\boldsymbol{P}^t]_{i,j} - \pi_j| \leq C_P \cdot \lambda(\boldsymbol{P})^t, \quad \text{for } t \geq T_P, \tag{5}$$

其中参数 $\lambda(\boldsymbol{P}) := (\lambda_2(\boldsymbol{P}) + 1)/2$，以及 C_P，T_P均为正数，其取值见 [14,引理 1]。[3] 注意，$\lambda(\boldsymbol{P})$量化了马尔可夫过程（A_t）$_{t\geq 0}$的相关性：$\lambda(\boldsymbol{P})$越接近 1，马尔可夫链收敛到其平稳分布的速度越慢。

在我们的分析中，我们作出以下额外假设。令 $\boldsymbol{w}^*$，$\boldsymbol{w}_B^*$ 分别表示 F和 F_B在 W上的极小化点。

假设 2。假设类 W 是凸的、紧的，并且其内部包含极小化点 $\boldsymbol{w}^*$、$\boldsymbol{w}_B^*$、$\boldsymbol{w}_k^*$。

以下假设涉及客户端的局部目标函数 $\{F_k\}_{k\in\mathcal{K}}$。假设 3 和 4 是凸优化 [31, 第 4.1、4.2]节文献中的标准假设。假设 5 是联邦优化算法 [8,第 6.1]节分析中的标准假设。

假设 3（L-光滑性）。本地函数 $\{F_k\}_{k=1}^N$具有 *L-Lipschitz*连续梯度：

$$F_k(\boldsymbol{v}) \leq F_k(\boldsymbol{w}) + \langle \nabla F_k(\boldsymbol{w}), \boldsymbol{v} - \boldsymbol{w}\rangle + \frac{L}{2}\|\boldsymbol{v} - \boldsymbol{w}\|_2^2 \quad \forall \boldsymbol{v}, \boldsymbol{w} \in W$$

假设 4（强凸性）。本地函数 $\{F_k\}_{k=1}^N$是 μ-强凸的：

$$F_k(\boldsymbol{v}) \geq F_k(\boldsymbol{w}) + \langle \nabla F_k(\boldsymbol{w}), \boldsymbol{v} - \boldsymbol{w}\rangle + \frac{\mu}{2}\|\boldsymbol{v} - \boldsymbol{w}\|_2^2 \quad \forall \boldsymbol{v}, \boldsymbol{w} \in W$$

假设 5（有界方差）。每个设备中的随机梯度方差有上界：$\mathbb{E}\|\nabla F_k(\boldsymbol{w}_t^k, \xi_t^k) - \nabla F_k(\boldsymbol{w}_t^k)\|^2 \leq \sigma_k^2 \quad k = 1, \ldots, N$

假设 2–5 推出本地函数满足以下性质，这些性质由引理 1 描述（证明见附录 B）。

引理 1. 在假设 *2–5* 下，存在常数 D、G 和 $H > 0$，使得对于 $\boldsymbol{w} \in W$ 和 $k \in \mathcal{K}$，有：

$$\|\nabla F_k(\boldsymbol{w})\| \leq D, \tag{6}$$

$$\mathbb{E}\|\nabla F_k(\boldsymbol{w}, \xi)\|^2 \leq G^2, \tag{7}$$

$$|F_k(\boldsymbol{w}) - F_k(\boldsymbol{w}_B^*)| \leq H. \tag{8}$$

[3]注意，只要 $\lambda(\boldsymbol{P}) \in (\lambda_2(\boldsymbol{P}), 1)$，(5) 就适用于 $\lambda(\boldsymbol{P})$ 的不同定义。对 $\lambda(\boldsymbol{P})$ 的具体选择会改变常数 C_P和 T_P。

与其他研究类似， [8], [22], [23], [32], 我们引入了一个度量，用于量化客户端本地数据集的异质性：

$$\Gamma \coloneqq \max_{k\in\mathcal{K}}\{F_k(\boldsymbol{w}^*) - F_k^*\}. \tag{9}$$

如果本地数据集相同，则本地函数 $\{F_k\}_{k\in\mathcal{K}}$彼此一致，并且与 F一致； $\boldsymbol{w}^*$ 是每个本地函数的最小化器，且 $\Gamma = 0$。一般而言，本地数据集所抽取自的分布越接近， Γ 就越小。

B. 主要定理

定理 1（证明见附录 A）将目标全局目标函数的误差分解为有偏全局目标函数的优化误差和偏差误差之和。

定理 1 （总误差的分解）。 在假设 *2–4* 下，目标全局目标函数的优化误差$\epsilon = F(\boldsymbol{w}) - F^*$ 可被界定如下：

$$\epsilon \le \underbrace{2\kappa^2(F_B(\boldsymbol{w}) - F_B^*)}_{\coloneqq\epsilon_{opt}} + \underbrace{2\kappa^4\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}}\Gamma}_{\coloneqq\epsilon_{bias}}, \tag{10}$$

where $\kappa \coloneqq L/\mu$, *and* $\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}} \coloneqq \sum_{k=1}^N (\alpha_k - p_k)^2/p_k$.

下面的定理 2 证明了，与有偏目标函数 F_B相关的优化误差 ϵ_{opt} ，在由方案（3）确定的轨迹上渐近地趋于零。不消失的偏差误差 ϵ_{bias} 刻画了$F(\boldsymbol{w})$与 $F_B(\boldsymbol{w})$之间的差异。该项取决于目标概率分布 $\boldsymbol{\alpha} = (\alpha_k)_{k\in\mathcal{K}}$ 与 $\boldsymbol{p} = (p_k)_{k\in\mathcal{K}}$ 之间的卡方散度 $\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}}$ ，以及Γ，后者量化了本地函数的异质性程度。当所有本地函数都相同（$\Gamma = 0$）时，偏差项 ϵ_{bias} 也会消失。对于 $\Gamma > 0$，偏差误差仍可通过分配给设备的聚合权重加以控制。特别地，当$q_k \propto \alpha_k/\pi_k$, $\forall k \in \mathcal{K}$时，偏差项会消失。由于该选择会渐近地抵消偏差误差，我们将其称为无偏聚合策略。

然而，在实践中，联邦学习训练的迭代次数有限 T （通常只有几百次 [5], [7]），因此前述渐近分析可能并不适用。在这种情况下，无偏聚合策略可能并非最优，因为最小化 ϵ_{bias} 并不一定能使总误差 $\epsilon \le \epsilon_{\text{opt}} + \epsilon_{\text{bias}}$ 最小化。这促使我们分析优化误差ϵ_{opt}。

定理 2 （优化误差的收敛性 ϵ_{opt}）。设假设 *1–5* 成立，且常数 $M, L, D, G, H, \Gamma, \sigma_k, C_P, T_P, \lambda(\boldsymbol{P})$ 如上所定义。设 $Q = \sum_{k\in\mathcal{K}} q_k$。设步长满足：

$$\sum_t \eta_t = +\infty, \quad \sum_t \ln(t)\cdot\eta_t^2 < +\infty. \tag{11}$$

令 T 表示通信轮次总数。对于 $T \ge T_P$，期望优化误差的上界如下：

$$\mathbb{E}[F_B(\bar{\boldsymbol{w}}_{T,0}) - F_B^*] \le \frac{\frac{1}{2}\boldsymbol{q}^\intercal\boldsymbol{\Sigma}\boldsymbol{q} + \upsilon}{\boldsymbol{\pi}^\intercal\boldsymbol{q}} + \psi + \frac{\phi}{\ln(1/\lambda(\boldsymbol{P}))}}{(\sum_{t=1}^T \eta_t)}, \tag{12}$$

where $\bar{\boldsymbol{w}}_{T,0} \coloneqq \frac{\sum_{t=1}^T \eta_t \boldsymbol{w}_{t,0}}{\sum_{t=1}^T \eta_t}$, *and*

$$\begin{aligned}
&\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_k^2\pi_k \textstyle\sum_t \eta_t^2),\\
&\upsilon = \tfrac{2}{E}\left\|\boldsymbol{w}_{0,0} - \boldsymbol{w}^*\right\|^2 + \tfrac{1}{4}MQ\textstyle\sum_t(\eta_t^2 + \frac{1}{t^2}),\\
&\psi = 4L(EQ+2)\Gamma\textstyle\sum_t \eta_t^2 + \frac{2}{3}(E-1)(2E-1)G^2\sum_t \eta_t^2,\\
&\mathcal{J}_t = \min\left\{\max\left\{\left\lceil \ln\left(2C_P H t\right)/\ln\left(1/\lambda(\boldsymbol{P})\right)\right\rceil, T_P\right\}, t\right\},\\
&\phi = 2EDGQ\textstyle\sum_t \ln(2C_P H t)\eta_{t-\mathcal{J}_t}^2.
\end{aligned}$$

定理 2（证明见附录 B）证明了在相关客户端参与的情况下，期望有偏目标函数 F_B 收敛至其最小值 F_B^* 。我们的上界 (12) 通过因子 $\ln(1/\lambda(\boldsymbol{P}))$ 刻画相关性的影响：相关性越高，收敛速率越慢。具体而言，我们发现 (12) 的分子与 $\boldsymbol{q}$ 存在二次与线性的分式依赖关系。最小化 ϵ_{opt} 通常会得到与最小化 ϵ_{bias} 不同的 $\boldsymbol{q}$ 选择。

C. 最小化总误差 $\epsilon \le \epsilon_{opt} + \epsilon_{bias}$

我们的分析指出，在最小化 ϵ_{opt} 或 ϵ_{bias} 之间存在权衡。我们的目标是找到最优的聚合权重 $\boldsymbol{q}^*$，使 (10) 中总误差 $\epsilon(\boldsymbol{q})$ 的上界最小：

$$\begin{aligned}
\underset{\boldsymbol{q}}{\text{minimize}} \quad & \epsilon_{\text{opt}}(\boldsymbol{q}) + \epsilon_{\text{bias}}(\boldsymbol{q});\\
\text{subject to} \quad & \boldsymbol{q} \ge 0,\\
& \|\boldsymbol{q}\|_1 = Q.
\end{aligned} \tag{13}$$

在附录 E 中，我们证明 (13) 是一个凸优化问题，可以用拉格朗日乘数法求解。然而，由于 (10) 和 (12) 中的常数（例如 L、μ、 Γ、 C_P）通常取决于具体问题，且在训练过程中难以估计，因此该解在实际中并无实用价值。特别是， Γ 带来了特殊的困难，因为它是根据目标目标函数 F 的最小化器定义的，而联邦学习算法通常最小化的是有偏函数 F_B。此外，(10) 中的上界与 [32], 中的上界类似，在将某些 q_k 设为 0 时会发散，但这仅仅是证明技巧导致的表象。下面给出一个更具实际意义的结果（证明见附录 C）：

定理 3 (总误差的另一种分解 ϵ)。 在与定理 *1* 相同的假设下，令 $\Gamma' \coloneqq \max_k\{F_k(\boldsymbol{w}_B^*) - F_k^*\}$。以下结论成立：

$$\epsilon \le \underbrace{2\kappa^2(F_B(\boldsymbol{w}) - F_B^*)}_{\coloneqq\epsilon_{opt}} + \underbrace{8\kappa^4 d_{TV}^2(\boldsymbol{\alpha},\boldsymbol{p})\Gamma'}_{\coloneqq\epsilon'_{bias}}, \tag{14}$$

其中 $d_{TV}(\boldsymbol{\alpha},\boldsymbol{p}) \colon = \frac{1}{2}\sum_{k=1}^N|\alpha_k - p_k|$ 是概率分布 $\boldsymbol{\alpha}$ 与 $\boldsymbol{p}$ 之间的全变差距离。

新常数 Γ' 根据 $\boldsymbol{w}_B^*$定义，因此在训练过程中更容易计算。然而， Γ' 取决于 $\boldsymbol{q}$，因为它是在 F_B的最小点处计算的。这种依赖关系使得对 (14) 右侧进行最小化更具挑战性（例如，相应的问题不是凸问题）。我们分别研究两项 ϵ_{opt} 和 ϵ'_{bias} 的最小化，并从中获得一些启示，据此设计新的联邦学习算法 CA-Fed。

D. 最小化ϵ_{opt}

ϵ_{opt} 的最小化仍然是一个凸优化问题（附录 D）。具体而言，在最优点，非负权重根据 $q_k^* = a(\lambda^*\pi_k - \theta^*)$设定，其中 a、 λ^*和 θ^* 为正数（见 (29)）。由此可知，可用性较低的客户端在聚合中获得的权重也较小。具体来说，这表明可以将可用性最低的客户端排除在聚合之外，从而得出以下指导原则：

准则 A：为了加快收敛，我们可以排除，即设定 $q_k^* = 0$，可用性最低的客户端 π_k。

这一准则可以从直觉上得到解释：参与度较低的客户端所提供的更新可能过于零散，以至于联邦学习算法无法跟踪其局部目标函数。它们就像噪声，会拖慢算法的收敛。排除这些客户端，使其不参与，可能更有利。

我们观察到，聚合权重 $\boldsymbol{q}$ 的选择不会影响客户端的可用性过程，尤其不会影响$\lambda(\boldsymbol{P})$。但是，如果算法排除了一些客户端，我们可以考虑马尔可夫链的状态空间，该状态空间仅指定其余客户端的可用状态；这条马尔可夫链可能具有不同的谱性质。为明确起见，本文在此处（以及后文中）考虑以下特例：每个客户端 k 的可用性都按照一个两状态马尔可夫链 $(A_t^k)_{t\geq 0}$ 演变，其转移概率矩阵为 $\boldsymbol{P_k}$ ，且这些马尔可夫链彼此独立。在这种情况下，聚合过程由乘积马尔可夫链 $(A_t)_{t\geq 0}$ 描述，其转移矩阵为 $\boldsymbol{P} = \bigotimes_{k\in\mathcal{K}} \boldsymbol{P_k}$，并且 $\lambda(\boldsymbol{P}) = \max_{k\in\mathcal{K}} \lambda(\boldsymbol{P_k})$，其中 $\boldsymbol{P_i} \bigotimes \boldsymbol{P_j}$ 表示矩阵 $\boldsymbol{P_i}$ 与 $\boldsymbol{P_j}$ [30, 的克罗内克积。练习 12.6]。在这种设定下，可以通过考虑由聚合权重非零的客户端所定义的缩减状态空间，重新定义马尔可夫链 $(A_t)_{t\geq 0}$ ，即$\boldsymbol{P'} = \bigotimes_{k'\in\mathcal{K}|q_{k'}>0} \boldsymbol{P_{k'}}$，并且 $\lambda(\boldsymbol{P'}) = \max_{k'\in\mathcal{K}|q_{k'}>0} \lambda(\boldsymbol{P_{k'}})$；与所有客户端都参与聚合的情况相比，这种状态空间可能更小。这些考量引出了以下准则：

准则 B：为了加快收敛，我们可以排除（即设置 $q_k^* = 0$） $\lambda(\boldsymbol{P_k})$ 较大的客户端。

直觉也支持这一准则。 $\lambda(\boldsymbol{P_k})$ 较大的客户端往往会长时间处于可用或不可用状态。由于梯度方法众所周知的灾难性遗忘问题 [33], [34],，这些客户端可能会在训练期间的最后阶段出现时，不公平地将算法引向其本地目标。此外，它们在早期阶段的参与可能毫无用处，因为它们的贡献会在长期缺席期间被遗忘。联邦学习算法可能会受益于直接忽略这类客户端。

我们注意到，准则 B 严格适用于这一特定情形：客户端动态彼此独立（且不存在空间相关性）。我们没有提供相应的

算法 1： CA-Fed（相关性感知联邦学习）

输入： $\boldsymbol{w}_0$, η_0, $\boldsymbol{\alpha}$, $\boldsymbol{q}^{(0)}$, $\{\eta_t\}_{t=1}^T$, η_s, E, β, τ

初始化 $\hat{\boldsymbol{F}}^{(0)}$, $\hat{\boldsymbol{F}}^*$, $\hat{\Gamma}'^{(0)}$, $\hat{\boldsymbol{\pi}}^{(0)}$, 以及 $\hat{\boldsymbol{\lambda}}^{(0)}$;

对于 $t = 1, \ldots, T$ **执行**

  接收活跃客户端集合 A_t、损失向量 $\boldsymbol{F}^{(t)}$;

  更新 $\hat{\boldsymbol{F}}^{(t)}$, $\hat{\Gamma}'^{(t)}$, $\hat{\boldsymbol{\pi}}^{(t)}$, 以及 $\hat{\boldsymbol{\lambda}}^{(t)}$;

  初始化 $\boldsymbol{q}^{(t)} = \frac{\boldsymbol{\alpha}}{\hat{\boldsymbol{\pi}}^{(t)}}$;

  $\boldsymbol{q}^{(t)} \leftarrow$ 获取$(\boldsymbol{q}^{(t)}, \boldsymbol{\alpha}, \hat{\boldsymbol{F}}^{(t)}, \hat{\boldsymbol{F}}^*, \hat{\Gamma}'^{(t)}, \hat{\boldsymbol{\pi}}^{(t)}, \hat{\boldsymbol{\lambda}}^{(t)})$;

  $\boldsymbol{q}^{(t)} \leftarrow$ get $(\boldsymbol{q}^{(t)}, \boldsymbol{\alpha}, \hat{\boldsymbol{F}}^{(t)}, \hat{\boldsymbol{F}}^*, \hat{\Gamma}'^{(t)}, \hat{\boldsymbol{\pi}}^{(t)}, \hat{\boldsymbol{\pi}}^{(t)})$;

  对于 客户端 $\{k \in A_t;\ q_k^{(t)} > 0\}$，并行 **执行**

    对于 $j = 0, \ldots, E-1$ **执行**

      $\boldsymbol{w}_{t,j+1}^k = \boldsymbol{w}_{t,j}^k - \eta_t \nabla F_k(\boldsymbol{w}_{t,j}^k, \mathcal{B}_{t,j}^k)$;

    $\Delta_t^k \leftarrow \boldsymbol{w}_{t,E} - \boldsymbol{w}_{t,0}$;

  $\boldsymbol{w}_{t+1,0} \leftarrow \mathbf{Proj}_W(\boldsymbol{w}_{t,0} + \eta_s \sum_{k\in A_t} q_k^{(t)} \cdot \Delta_t^k)$;

函数 get $(\boldsymbol{q}, \boldsymbol{\alpha}, \boldsymbol{F}, \boldsymbol{F}^*, \Gamma, \boldsymbol{\pi}, \rho)$：

  $\mathcal{K} \leftarrow$ 按降序排列在 ρ 中;

  $\hat{\epsilon} \leftarrow \langle \boldsymbol{F} - \boldsymbol{F}^*, \boldsymbol{\pi} \tilde{\odot} \boldsymbol{q} \rangle + d^2{}_{TV}(\boldsymbol{\alpha},\ \boldsymbol{\pi} \tilde{\odot} \boldsymbol{q}) \cdot \Gamma$;

  对于 $k \in \mathcal{K}$ **执行**

    $q^+ \leftarrow 0$;

    $\hat{\epsilon}^+ \leftarrow \langle \boldsymbol{F} - \boldsymbol{F}^*, \boldsymbol{\pi} \tilde{\odot} \boldsymbol{q}^+ \rangle + d_{TV}^2(\boldsymbol{\alpha},\ \boldsymbol{\pi} \tilde{\odot} \boldsymbol{q}^+) \cdot \Gamma$;

    如果 $\hat{\epsilon} - \hat{\epsilon}^+ \geq \tau$ **则**

      $\hat{\epsilon} \leftarrow \hat{\epsilon}^+$;

      $\boldsymbol{q} \leftarrow \boldsymbol{q}^+$;

  返回 $\boldsymbol{q}$

准则来处理客户端存在空间相关性的情况（这一任务留待未来研究）。然而，在这一更一般的情形下，可以忽略准则 B，但仍参考准则 A 和 C；也可以在客户端存在空间相关性的情况下继续考虑准则 B（见第 VI-B 节的讨论）。

E. 最小化 ϵ'_{bias}

当目标重要性 $\boldsymbol{\alpha}$ 与有偏重要性 $\boldsymbol{p}$ 之间的全变差距离为零时，(14) 中的偏差误差 ϵ'_{bias} 消失，即当 $q_k \propto \alpha_k/\pi_k$, $\forall k \in \mathcal{K}$时。随后，排除对优化误差贡献最大且尤其会减慢收敛速度的客户端（准则 A 和 B）后，我们可以为剩余客户端分配与其可用性成反比的聚合权重，从而最小化偏差误差 ϵ'_{bias}。

准则 C：为降低偏差误差，我们为未被先前准则排除的客户端设置 $q_k^* \propto \alpha_k/\pi_k$。

IV. 所提出的算法

第二节中的准则 A 和 B 表明，最小化 ϵ_{opt} 可能导致一些可用客户端被排除在聚合步骤 (3) 之外，尤其是可用性低和/或相关性高的客户端。对于其余客户端，准则 C 提议将其聚合权重设为与可用性成反比，以降低偏差误差 ϵ'_{bias}。受这些见解启发，我们提出了 CA-Fed，这是一种客户端采样与聚合策略，考虑了联邦学习中客户端可用性相关性问题，详见

算法 1。在训练过程中，CA-Fed 会学习应排除哪些客户端，以及如何设置其他客户端的聚合权重，从而在 ϵ_{opt} 和 ϵ'_{bias}之间实现良好的权衡。准则 A 和 B 指明了要移除哪些客户端，而第 t 轮要移除的客户端确切数量则通过最小化 $\epsilon^{(t)}$ 来确定，并将其作为式 (14) 中上界的代理：[4]

$$\epsilon^{(t)} := F_B(\boldsymbol{w}_{t,0}) - F_B^* + d_{TV}^2(\boldsymbol{\alpha}, \boldsymbol{p})\Gamma'. \tag{15}$$

A. CA-Fed 的核心步骤

在每个通信轮次 t，服务器都会向所有活跃客户端发送当前模型 $\boldsymbol{w}_{t,0}$，而每个客户端 k 都会返回基于一个样本批次 $\mathcal{B}_t^k$ 计算出的当前损失的噪声估计 $F(\boldsymbol{w}_{t,0}, \xi)$（第 3 行）。服务器利用这些值以及当前可用客户端集合 A_t 的信息，改进对每个客户端损失（$\hat{\boldsymbol{F}}^{(t)} = (\hat{F}_k^{(t)})_{k\in\mathcal{K}}$）和每个客户端损失最小值（$\hat{\boldsymbol{F}}^* = (\hat{F}_k^*)_{k\in\mathcal{K}}$）的估计，以及对 Γ'，π_k, λ_k 和 $\epsilon^{(t)}$ 的估计，分别记为 $\hat{\Gamma}'^{(t)}$、$\hat{\pi}_k^{(t)}$、$\hat{\lambda}_k^{(t)}$ 和 $\hat{\epsilon}^{(t)}$（可能的估计器将在下文说明）（第 4 行）。

服务器决定是否排除可用性模式表现出高度相关性（高 $\hat{\lambda}_k^{(t)}$）的客户端（第 6 行）。首先，服务器按照 $\hat{\boldsymbol{\lambda}}^{(t)}$ 降序考察所有客户端（第 14 行），并评估排除这些客户端（第 17 行）后，$\hat{\epsilon}^{(t)}$ 是否会下降超过阈值 $\tau \geq 0$（第 19 行）。然后，服务器按照 $\hat{\boldsymbol{\pi}}^{(t)}$ 升序考察客户端，并重复相同的过程，以便可能排除一些可用性较低（低 $\hat{\pi}_k^{(t)}$）的客户端（第 7 行）。

选定参与客户端（满足 $q_k > 0$ 的客户端）后，服务器会通知它们按照（2）继续更新当前模型（第 9–10 行），而其他可用客户端则保持空闲。最后，按照（3）聚合模型更新（第 12 行）。

B. 估计器

我们现在简要讨论估计量 $\hat{F}_k^{(t)}$、$\hat{F}_k^*$、$\hat{\Gamma}'^{(t)}$、$\hat{\pi}_k^{(t)}$ 和 $\hat{\lambda}_k^{(t)}$ 的可能实现方式。服务器对客户端本地损失的估计（$\hat{\boldsymbol{F}}^{(t)} = (\hat{F}_k^{(t)})_{k\in\mathcal{K}}$）可通过参数为 $\beta \in (0, 1]$: 的自回归滤波器，根据接收到的活跃客户端损失（$\boldsymbol{F}_{(t)} = (F_k^{(t)})_{k\in A_t}$）获得。

$$\hat{\boldsymbol{F}}^{(t)} = (\mathbf{1} - \beta \mathbb{1}_{A_t}) \odot \hat{\boldsymbol{F}}^{(t-1)} + \beta \mathbb{1}_{A_t} \odot \boldsymbol{F}^{(t)}, \tag{16}$$

其中，$\odot$ 表示向量之间的逐分量乘法，$\mathbb{1}_{A_t}$ 是一个 N 维二进制向量，当且仅当客户端 k 在轮次 t 处于活跃状态时，其第 k 个分量等于 1，即 $k \in A_t$。服务器可以跟踪客户端的损失最小值，并将 F_k^* 估计为 $\hat{F}_k^* = \min_{s\in[0,t]} \hat{F}_k^{(s)}$。$F_B(\boldsymbol{w}_{t,0})$、$F_B^*$、$\Gamma'$ 和 $\epsilon^{(t)}$ 的值可按如下方式估计：

$$\hat{F}_B^{(t)} - \hat{F}_B^* = \langle \hat{\boldsymbol{F}}^{(t)} - \hat{\boldsymbol{F}}^*, \hat{\boldsymbol{\pi}}^{(t)} \tilde{\odot} \boldsymbol{q}^{(t)} \rangle, \tag{17}$$

$$\hat{\Gamma}'^{(t)} = \max_{k\in\mathcal{K}} (\hat{F}_k^{(t)} - \hat{F}_k^*), \tag{18}$$

$$\hat{\epsilon}^{(t)} = \hat{F}_B^{(t)} - \hat{F}_B^* + d_{TV}^2(\boldsymbol{\alpha}, \hat{\boldsymbol{\pi}}^{(t)} \tilde{\odot} \boldsymbol{q}^{(t)}) \cdot \hat{\Gamma}'^{(t)}. \tag{19}$$

[4]根据式 (14)，可以合理地引入一个超参数，对求和中的优化项和偏差项的相对重要性进行加权。我们将在第 VI-A 节讨论对 CA-Fed 的这项额外优化。

其中 $\boldsymbol{\pi} \tilde{\odot} \boldsymbol{q} \in \mathbb{R}^N$，使得 $(\boldsymbol{\pi} \tilde{\odot} \boldsymbol{q})_k = \frac{\pi_k q_k}{\sum_{h=1}^N \pi_h q_h}$，$k \in \mathcal{K}$。

对于 $\hat{\pi}_k^{(t)}$，服务器只需记录客户端 k 截至时间 t 处于可用状态的总次数，并使用带有贝塔先验的贝叶斯估计器计算 $\hat{\pi}_k^{(t)}$，即 $\hat{\pi}_k^{(t)} = (\sum_{s\leq t} \mathbb{1}_{k\in A_s} + n_k)/(t + n_k + m_k)$，其中 n_k和 m_k是贝塔先验的初始参数。

对于 $\hat{\lambda}_k^{(t)}$，服务器可以假设客户端的可用性按照具有两个状态（可用和不可用）的马尔可夫链演变，记录相应的状态转移次数，并通过贝叶斯估计器估计转移矩阵 $\hat{\boldsymbol{P}}_k^{(t)}$，方法与估计 $\hat{\pi}_k^{(t)}$ 类似。最后，通过计算 $\hat{\boldsymbol{P}}_k^{(t)}$ 的特征值获得 $\hat{\lambda}_k^{(t)}$。

C. CA-Fed 的计算/通信开销

CA-Fed 旨在提高训练收敛性，而非降低其计算和通信开销。不过，排除部分可用客户端会降低总体训练成本，正如本节将以神经网络训练为例进行讨论。

未被选中参加训练的可用客户端，只需在当前模型上对单个批次评估一次本地损失，而无需执行 E 梯度更新；后者大约需要 $2 \times E - 1$ 更多计算量（因为需要进行前向和反向传播）。对于已选中的客户端，计算损失不会带来额外计算开销，因为无论如何，它们都应在首次本地梯度更新期间进行前向传播。

在通信方面，被排除的客户端只需传输损失这一单个标量，其大小远小于模型更新。相反，参与客户端会传输本地损失和模型更新。不过，这项额外开销可以忽略不计，而且很可能完全被排除客户端带来的通信节省所抵消。

V. 实验评估

A. 实验设置

a) 联邦系统模拟器：在我们的实验中，我们模拟具有不同时间相关性水平的客户端可用性动态。我们将每个客户端的活跃状态建模为一个具有状态空间 $\mathcal{S} = \{$“活跃”，“非活跃”$\}$的二状态齐次马尔可夫过程。我们使用$p_{k,\,s}$ 表示客户端 $k \in \mathcal{K}$ 保持在状态 $s \in \mathcal{S}$的概率。

为了模拟联邦学习系统中存在的统计异质性，我们考虑一种实验设置，其中包含两个互不相交的客户端组 G_i，$i = 1, 2$，并为它们分别指定两个不同的数据分布 $\mathcal{P}_i$，$i = 1, 2$，具体定义将在后文给出。令 $r_i = |G_i|/N$，$i = 1, 2$ 表示组 $i = 1, 2$ 中的客户端比例。为了模拟现实联邦系统中客户端可用性模式的异质性，我们将每组客户端均匀随机地分为两类：“可用性较高”的客户端，其处于活跃状态的稳态概率为 $\pi_{k,\text{ active}} = 1/2 + g$；以及“可用性较低”的客户端，其稳态概率为 $\pi_{k,\text{ active}} = 1/2 - g$，其中 $g \in$

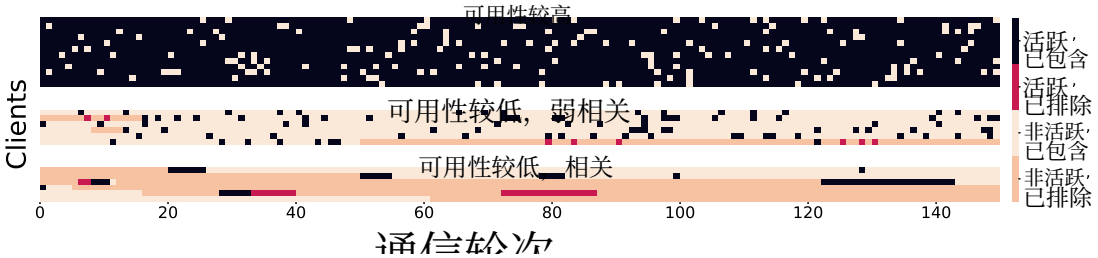


图 1：合成数据集上的客户端活动及 CA-Fed 的客户端选择。

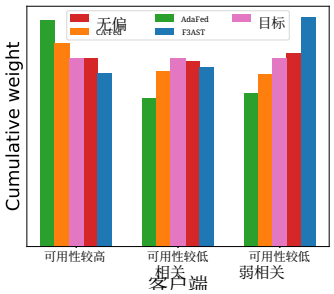
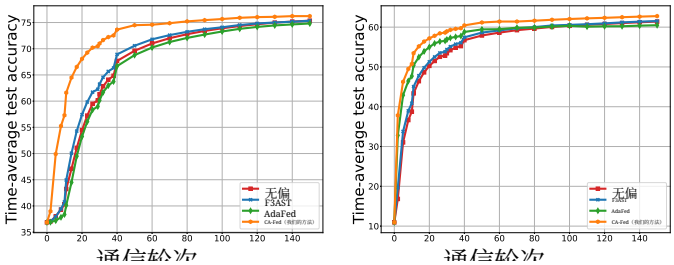


图 2：在合成数据集的整个训练过程中，不同算法赋予各客户端的重要性。



（a）合成数据　　（b）MNIST

图 3：测试准确率与通信轮次的关系。

$(0,\ 1/2)$ 是一个用于控制客户端可用性异质性的参数。此外，我们将每类客户端均匀随机地分为两个子类："相关"客户端倾向于持续保持相同状态（$\lambda_k = \nu$ 其 ν 值接近 1），以及"弱相关"客户端，它们保持当前状态和改变状态的可能性几乎相同（$\lambda_k \sim \mathcal{N}(0,\ \varepsilon^2)$，其中 ε 接近 0）。在我们的实验中，我们假设 $r_1 = r_2 = 1/2$、$g = 0.4$、$\nu = 0.9$ 和 $\varepsilon = 10^{-2}$。

b) 数据集和模型：所有实验均在一个二分类合成数据集（见附录 F）和真实世界的 MNIST 数据集上进行，使用 [35], $N = 24$ 客户端。对于 MNIST 数据集，我们在两个客户端组之间引入统计异质性（即使两个分布 $\mathcal{P}_1$ 和 $\mathcal{P}_2$ 不同），采用与 [36]相同的方法：1）为每个客户端分配训练数据总量中的一个随机子集；2）通过随机交换两对标签，修改第二组客户端的数据。我们保留 MNIST 原始的训练/测试数据划分，并将训练数据集的 20% 用作验证数据集。对于合成数据集和真实世界的 MNIST 数据集，我们均使用带有参数为 10^{-2}的岭惩罚的线性分类器，其目标函数为强凸函数。

c) 基准方法：我们将算法 1 中定义的 CA-Fed 与无偏聚合策略进行比较。在该策略中，所有活跃客户端都会参与，并根据其可用性获得与之成反比的权重；此外，我们还将其与第二节讨论的最先进联邦学习算法进行比较：F3AST [18]和 AdaFed [19]。我们通过网格搜索调整学习率 η，η_s 搜索网格为 η：$\{10^{-3}, 10^{-2.5}, 10^{-2}, 10^{-1.5}, 10^{-1}\}$，$\eta_s$ ：$\{10^{-2}, 10^{-1.5}, 10^{-1}, 10^{-0.5}, 10^{0}\}$。对于 CA-Fed，我们使用了$\tau = 0$，$\beta = 0.2$。我们假设所有算法都可以访问预言机，以获取每个客户端的真实可用性参数。在实践中，无偏聚合策略、AdaFed 和 F3AST 依赖于对 $\pi_{k,\ \text{active}}$的精确了解，而 CA-Fed 则依赖于对 $\pi_{k,\ \text{active}}$ 和 λ_k的精确了解。 [5]

B. 实验结果

图 1 展示了在合成数据集上进行一次训练过程期间每个客户端的可用性。CA-Fed 选中的客户端（相应地，被排除的客户端）以黑色（相应地，红色）突出显示。我们观察到，被排除的客户端往往是平均可用性较低或相关性较高的客户端。

图 2 展示了不同算法在一次完整训练过程中为每个客户端赋予的重要性 p_k（随时间取平均）k。我们观察到，除无偏算法外，所有算法都偏离了目标重要性 $\boldsymbol{\alpha}$。如指导线 A 和 B 所示，CA-Fed 倾向于偏向"可用性较高"的客户端群体，而牺牲"可用性较低"的客户端。

图 3 显示了训练得到的模型截至第 t 轮的时间平均准确率，该结果取自三次不同运行的平均值。在两个数据集上，CA-Fed 的准确率均为最高，比排名第二的算法（F3AST）高出约一个百分点。表 I 列出了各算法在三次运行中的平均值：训练期间达到的最高测试准确率、训练期间达到的时间平均测试准确率，以及训练期间后半段的标准差。结果表明，尽管 CA-Fed 达到的最高准确率与无偏基线以及最先进的 AdaFed 和 F3AST 相当，但其时间平均准确率比排名第二的算法（F3AST）高出 1.24 个百分点，且标准差也更小（1.5×）。

[5]作者已在以下地址公开其代码和数据：https://github.com/arodio/CA-Fed。

表 I：Synthetic / MNIST 数据集上的最高测试准确率和时间平均测试准确率，以及它们的标准差。

	测试准确率		
	最高值	时间平均值	标准差
无偏	78.94 / 64.87	75.32 / 61.39	0.48 / 1.09
F3AST	78.97 / 64.91	75.33 / 61.52	0.40 / 0.94
ADAFED	78.69 / 63.77	74.81 / 60.48	0.59 / 1.37
CA-FED	**79.03 / 64.94**	**76.22 / 62.76**	**0.28 / 0.61**

VI. 讨论

本节讨论有关我们算法的一些一般性问题和说明。

A. 控制被排除客户端的数量

定理 1 和定理 3 表明，条件数 κ^2 在最小化总误差 ϵ方面可以发挥重要作用。我们的算法使用总误差的一个代理量（$\epsilon^{(t)}$）。为了考虑 κ^2的影响，我们可以引入一个超参数，用于权衡式 (15) 中优化误差和偏差误差的相对重要性：

$$\epsilon'^{(t)} \coloneqq F_B(\boldsymbol{w}_{t,0}) - F_B^* + \bar{\kappa}^2 \cdot d_{TV}^2(\boldsymbol{\alpha}, \boldsymbol{p})\Gamma'.$$

$\bar{\kappa}^2$ 较小时，会加大对偏差项的惩罚，从而优先降低优化误差，导致 CA-Fed 排除更多客户端。另一方面， $\bar{\kappa}^2$较大时，CA-Fed 倾向于纳入更多客户端。渐近地，当 $\bar{\kappa}^2 \to +\infty$时，CA-Fed 将退化为无偏基线。为了进一步提升 CA-Fed 的性能，可以对 $\bar{\kappa}^2$ 的取值进行更精细的调整。

B. 存在空间相关性时的 *CA-Fed*

尽管 CA-Fed 主要是为处理时间相关性而设计的，但在存在空间相关性时，它也不一定表现不佳。

考虑以下存在空间相关性的场景：客户端被分组成簇，每个簇 $c \in \mathcal{C}$ 都由一个底层马尔可夫链描述，该链决定簇中所有客户端何时可用或不可用；不同簇的马尔可夫链彼此独立。令 λ_c 表示簇 c 的马尔可夫链中模最大的第二大特征值。在这种情况下，需要排除簇 $\bar{c} = \text{argmax}_{c\in\mathcal{C}}\, \lambda_c$ 中的所有客户端，以降低聚合马尔可夫链的特征值。

在这种设置下，CA-Fed 会为同一簇中的所有客户端估算出相似的特征值，随后会正确地开始考虑排除簇 $\bar{c}$ 中的客户端，并可能依次移除该簇中的所有客户端。这些分析表明，即使存在空间相关性，CA-Fed 仍可能正常运行。

C. 关于 *CA-Fed* 的公平性

在训练阶段排除某些客户端的策略（例如 CA-Fed），可能自然会引发公平性方面的担忧。文献中对联邦学习公平性的概念尚无统一定义 [37, 第 8] 章：公平性目标可以通过在式（1）中适当选择目标权重来实现。例如，对每个客户端设置相等的权重 α_k ，即可实现客户端级公平性；将权重 α_k 设置为与本地数据集大小 $|D_k|$成正比，则可实现逐样本公平性。假设式（1）中的全局目标函数确实也体现了公平性方面的考量，那么 CA-Fed 本质上是公平的，因为它能确保训练得到的模型的性能目标函数尽可能接近其最小值。

VII. 结论

本文首次针对客户端可用性异质且相关的情形，对类似 FedAvg 的联邦学习算法进行了收敛分析。分析量化了相关性对算法收敛速率的不利影响，并指出了偏差与收敛速度之间普遍存在的权衡。在理论分析的指导下，我们提出了 CA-Fed，这是一种新的联邦学习算法，旨在平衡最大化收敛速度与最小化模型偏差这两个相互冲突的目标。实验结果表明，自适应地排除时间相关性高且可用性低的客户端，是应对联邦学习中客户端可用性异质且相关问题的有效方法。

附录

A. 定理 *1* 的证明

我们将目标函数的优化误差上界表示为有偏目标函数的优化误差加上一个偏差项：

$$\begin{aligned} F(\boldsymbol{w}) - F^* &\overset{(a)}{\leq} \tfrac{1}{2\mu}\left\|\nabla F(\boldsymbol{w})\right\|^2 \overset{(b)}{\leq} \tfrac{L^2}{2\mu}\left\|\boldsymbol{w} - \boldsymbol{w}^*\right\|^2 \\ &\overset{(c)}{\leq} \tfrac{L^2}{\mu}(\left\|\boldsymbol{w} - \boldsymbol{w}_B^*\right\|^2 + \left\|\boldsymbol{w}_B^* - \boldsymbol{w}^*\right\|^2) \\ &\overset{(d)}{\leq} \underbrace{\tfrac{2L^2}{\mu^2}(F_B(\boldsymbol{w}) - F_B^*)}_{:=\epsilon_{\text{opt}}} + \underbrace{\tfrac{2L^2}{\mu^2}(F(\boldsymbol{w}_B^*) - F^*)}_{:=\epsilon_{\text{bias}}}, \end{aligned}$$

其中，(a)、(b) 和 (d) 由假设3、4得出，而不等式 (c) 由 $(a+b)^2 \leq 2a^2 + 2b^2$ 得出。具体而言，(b) 要求 $\nabla F_k(\boldsymbol{w}_k^*) = 0$。定理 2 进一步阐述了优化误差$\epsilon_{\text{opt}}$。现在我们展开 ϵ_{bias}:

$$\begin{aligned} \left\|\nabla F(\boldsymbol{w}_B^*)\right\| &\overset{(e)}{=} \left\|\textstyle\sum_{k=1}^N (\alpha_k - p_k)\nabla F_k(\boldsymbol{w}_B^*)\right\| \\ &\overset{(f)}{\leq} L\textstyle\sum_{k=1}^N |\alpha_k - p_k|\left\|\boldsymbol{w}_B^* - \boldsymbol{w}_k^*\right\| \\ &\overset{(g)}{\leq} L\sqrt{\tfrac{2}{\mu}}\textstyle\sum_{k=1}^N \frac{|\alpha_k - p_k|}{\sqrt{p_k}}\sqrt{p_k(F_k(\boldsymbol{w}_B^*) - F_k^*)}, \end{aligned} \tag{20}$$

其中，(e) 用到了 $\nabla F_B(\boldsymbol{w}_B^*) = 0$；$(f)$ 先应用三角不等式，再应用 L-光滑性；(g) 则由μ-强凸性得出。此外，(f) 要求 $\nabla F_k(\boldsymbol{w}_k^*) = 0$。与 (g) 中的 [32], 类似，我们将分子和分母同时乘以 $\sqrt{p_k}$。经过直接计算可得：

$$\begin{aligned} \left\|\nabla F(\boldsymbol{w}_B^*)\right\|^2 &\overset{(h)}{\leq} \tfrac{2L^2}{\mu}\Big(\textstyle\sum_{k=1}^N \frac{|\alpha_k - p_k|}{\sqrt{p_k}}\sqrt{p_k(F_k(\boldsymbol{w}_B^*) - F_k^*)}\Big)^2 \\ &\overset{(i)}{\leq} \tfrac{2L^2}{\mu}\Big(\sum_{k=1}^N \tfrac{(\alpha_k - p_k)^2}{p_k}\Big)\Big(\sum_{k=1}^N p_k(F_k(\boldsymbol{w}_B^*) - F_k^*)\Big) \\ &\overset{(j)}{\leq} \tfrac{2L^2}{\mu}\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}}\Gamma, \end{aligned}$$

其中，(*i*) 用到了柯西–施瓦茨不等式，而 (*j*) 用到了：

$\sum_{k=1}^{N} p_k(F_k(\boldsymbol{w}_B^*) - F_k^*) \le \sum_{k=1}^{N} p_k(F_k(\boldsymbol{w}^*) - F_k^*) \le \Gamma.$

最后，由于 F 的强凸性，我们得出：

$$F(\boldsymbol{w}_B^*) - F^* \le \frac{1}{2\mu} \left\|\nabla F(\boldsymbol{w}_B^*)\right\|^2 \le \frac{L^2}{\mu^2} \chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}} \Gamma. \quad \square$$

B. 定理 *2* 的证明

1) 补充符号：设 $\boldsymbol{w}_{t,\,j}^k$ 为设备 k 在全局轮次 t、本地迭代 j 时计算出的模型参数向量。我们定义：

$$g_t(A_t) = \textstyle\sum_{k\in A_t} q_k \sum_{j=0}^{E-1} \nabla F_k(\boldsymbol{w}_{t,j}^k, \xi_{t,j}^k),$$

和 $\bar{g}_t(A_t) = \mathbb{E}_{\xi|A_t}[g_t(A_t)]$.

根据式 (2) 和 (3)，CA-Fed 的更新规则为：

$$\boldsymbol{w}_{t+1,0} = \mathbf{Proj}_W(\boldsymbol{w}_{t,0} - \eta_t g_t(A_t)). \tag{21}$$

2) 关键引理和结果：我们给出有助于证明主要定理的引理和结果。

引理 *1* 的证明。 W 的有界性给出了一个关于 $(\boldsymbol{w}_{t,\,0})_{t\ge 0}$ 的上界，该上界基于 (2) 和 (3) 中的更新规则。由 $\{F_k\}_{k\in\mathcal{K}}$ 的凸性可得：

$$D \coloneqq \sup_{\boldsymbol{w}\in W, k\in\mathcal{K}} \left\|\nabla F_k(\boldsymbol{w})\right\| < +\infty.$$

条目 (6)、(8) 可直接由前面的观察得出。结合 (6) 和假设 5 可得条目 (7)：

$$\mathbb{E}\left\|\nabla F_k(\boldsymbol{w}, \xi)\right\|^2 \le D^2 + \max_{k\in\mathcal{K}}\{\sigma_k^2\} \coloneqq G^2. \quad \square$$

引理 2（异构客户端可用性下的收敛）。设本地函数 $\{F_k\}_{k\in\mathcal{K}}$ 为凸函数，且假设 *3*、*5* 成立。若 $\eta_t \le \frac{1}{2L(EQ+1)}$，则有：

$$\begin{aligned}\textstyle\sum_t \eta_t \mathbb{E}[\sum_{k\in A_t} q_k \left(F_k(\boldsymbol{w}_{t,0}) - F_k(\boldsymbol{w}_B^*)\right)] \le \\ + \tfrac{2}{E}\left\|\boldsymbol{w}_{0,0} - \boldsymbol{w}_B^*\right\|^2 + 2\textstyle\sum_{k=1}^N \pi_k q_k^2 \sigma_k^2 \sum_t \eta_t^2 \\ + \tfrac{2}{3}\textstyle\sum_{k=1}^N \pi_k q_k (E-1)(2E-1) G^2 \sum_t \eta_t^2 \\ + 2L(EQ+2)\textstyle\sum_{k=1}^N \pi_k q_k \Gamma \sum_t \eta_t^2 := C_1 < +\infty.\end{aligned}$$

引理 *2* 的证明。

$$\begin{aligned}\left\|\boldsymbol{w}_{t+1,\,0} - \boldsymbol{w}_B^*\right\|^2 &= \left\|\mathbf{Proj}_W\left(\boldsymbol{w}_{t,\,0} - \eta_t g_t\right) - \mathbf{Proj}_W\left(\boldsymbol{w}_B^*\right)\right\|^2 \\ &\le \left\|\boldsymbol{w}_{t,0} - \eta_t g_t - \boldsymbol{w}_B^* + \eta_t \bar{g}_t - \eta_t \bar{g}_t\right\|^2 = A_1 + A_2 + A_3,\end{aligned}$$

其中：

$$\begin{aligned}A_1 &= \left\|\boldsymbol{w}_{t,0} - \boldsymbol{w}_B^* - \eta_t \bar{g}_t\right\|^2, \\ A_2 &= 2\eta_t \langle \boldsymbol{w}_{t,0} - \boldsymbol{w}_B^* - \eta_t \bar{g}_t, \bar{g}_t - g_t\rangle, \\ A_3 &= \eta_t^2 \left\|g_t - \bar{g}_t\right\|^2.\end{aligned}$$

注意 $\mathbb{E}[A_2] = 0$。我们利用 [22] 中的关键步骤对 A_1、A_3 进行上界估计：

(1) 如果每台设备上的随机梯度方差有上界，则 $g_t(A_t)$ 的方差也有上界：

$$\begin{aligned}A_3 &= \mathbb{E}_{\mathcal{B}|A_t}\left\|g_t - \bar{g}_t\right\|^2 = \\ &= \textstyle\sum_{k\in A_t} q_k^2 \sum_{j=0}^{E-1} \mathbb{E}_{\mathcal{B}|A_t}\left\|\nabla F_k(\boldsymbol{w}_{t,j}^k, \xi_{t,j}^k) - \nabla F_k(\boldsymbol{w}_{t,j}^k)\right\|^2 \\ &\le E\textstyle\sum_{k\in A_t} q_k^2 \sigma_k^2;\end{aligned}$$

(2) 本地模型 $\boldsymbol{w}_{t,\,E}^k$ 与全局模型 $\boldsymbol{w}_{t,\,0}$ 之间的距离有上界，因为随机梯度的平方范数期望有上界：

$$\begin{aligned}&\mathbb{E}_{\mathcal{B}|A_t}\textstyle\sum_{k\in A_t} q_k \sum_{j=0}^{E-1}\left\|\boldsymbol{w}_{t,j}^k - \boldsymbol{w}_{t,0}\right\|^2 = \\ &= \mathbb{E}_{\mathcal{B}|A_t}\textstyle\sum_{k\in A_t} q_k \sum_{j=1}^{E-1} \eta_t^2 \left\|\sum_{j'=0}^{j-1} \nabla F_k(\boldsymbol{w}_{t,j'}^k, \xi_{t,j'}^k)\right\|^2 \\ &\le \eta_t^2 \textstyle\sum_{k\in A_t} q_k \sum_{j=1}^{E-1} j \sum_{j'=0}^{j-1} \mathbb{E}_{\mathcal{B}|A_t}\left\|\nabla F_k(\boldsymbol{w}_{t,j'}^k, \xi_{t,j'}^k)\right\|^2 \\ &\le \eta_t^2 \textstyle\sum_{k\in A_t} q_k G^2 \sum_{j=1}^{E-1} j^2 \\ &= \tfrac{1}{6}\eta_t^2 \textstyle\sum_{k\in A_t} q_k E(E-1)(2E-1)G^2. \quad \square\end{aligned}$$

引理 3 (经过 $\mathcal{J}_t$ 步后的优化误差)。假设条件 *1*、*2* 成立，本地函数 $\{F_k\}_{k\in\mathcal{K}}$ 为凸函数，D、H 的定义如 (6)、(8) 所示，且 $\mathcal{J}_t$ 的定义如定理 *2* 所示。则：

$$\begin{aligned}&\textstyle\sum_t \eta_t \mathbb{E}[\sum_{k\in A_t} q_k (F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_{t,0}))] \\ &\le EDGQ \textstyle\sum_t \mathcal{J}_t \eta_{t-\mathcal{J}_t}^2 \sum_{k=1}^N \pi_k q_k := \frac{C_3}{\ln(1/\lambda(\boldsymbol{P}))} < +\infty.\end{aligned}$$

为证明引理 3，我们引入以下结果：

$$\left|F_k(\boldsymbol{v}) - F_k(\boldsymbol{w})\right| \le D \cdot \left\|\boldsymbol{v} - \boldsymbol{w}\right\|, \ \forall \boldsymbol{v}, \boldsymbol{w} \in W, \tag{22}$$

$$\mathbb{E}_{\mathcal{B}_{t,0}^k, \ldots, \mathcal{B}_{t,E-1}^k}\left\|\boldsymbol{w}_{t+1,0} - \boldsymbol{w}_{t,0}\right\| \le \eta_t G E(\textstyle\sum_{k\in A_t} q_k). \tag{23}$$

由 $\{F_k\}_{k\in\mathcal{K}}$ 的凸性可得公式 (22)：

$$\langle \nabla F_k(\boldsymbol{v}), \boldsymbol{v} - \boldsymbol{w}\rangle \le \left\|F_k(\boldsymbol{v}) - F_k(\boldsymbol{w})\right\| \le \langle \nabla F_k(\boldsymbol{w}), \boldsymbol{v} - \boldsymbol{w}\rangle;$$

根据柯西–施瓦茨不等式可得：

$$\begin{aligned}\left|F_k(\boldsymbol{v}) - F_k(\boldsymbol{w})\right| &\le \max\{\left\|\nabla F_k(\boldsymbol{v})\right\|, \left\|\nabla F_k(\boldsymbol{w})\right\|\}\left\|\boldsymbol{v} - \boldsymbol{w}\right\| \\ &\le D \cdot \left\|\boldsymbol{v} - \boldsymbol{w}\right\|.\end{aligned}$$

结合公式（7）和（21），可得公式（23）：

$$\begin{aligned}&\mathbb{E}_{\mathcal{B}|A_t}\left\|\boldsymbol{w}_{t+1,0} - \boldsymbol{w}_{t,0}\right\| \le \\ &\le \eta_t \mathbb{E}_{\mathcal{B}|A_t}\left\|\textstyle\sum_{k\in A_t} q_k \sum_{j=0}^{E-1} \nabla F_k(\boldsymbol{w}_{t,j}^k, \xi_{t,j}^k)\right\| \\ &\le \eta_t \textstyle\sum_{k\in A_t} q_k \sum_{j=0}^{E-1} \mathbb{E}_{\mathcal{B}|A_t}\left\|\nabla F_k(\boldsymbol{w}_{t,j}^k, \xi_{t,j}^k)\right\| \\ &\le \eta_t G E(\textstyle\sum_{k\in A_t} q_k).\end{aligned}$$

引理 *3* 的证明。$\mathcal{J}_t$ 通信轮次后局部目标函数的演变上界为：

$$\begin{aligned}&\textstyle\sum_t \eta_t \mathbb{E}[\sum_{k\in A_t} q_k (F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_{t,0}))] \\ &\overset{(a)}{\le} D\textstyle\sum_t \eta_t \mathbb{E}[\sum_{k\in A_t} q_k \mathbb{E}_B\left\|\boldsymbol{w}_{t-\mathcal{J}_t,0} - \boldsymbol{w}_{t,0}\right\|] \\ &\overset{(b)}{\le} D\textstyle\sum_t \eta_t \sum_{d=t-\mathcal{J}_t}^{t-1} \mathbb{E}[\sum_{k\in A_t} q_k \mathbb{E}_B\left\|\boldsymbol{w}_{d,0} - \boldsymbol{w}_{d+1,0}\right\|] \\ &\overset{(c)}{\le} EDG\textstyle\sum_t \sum_{d=t-\mathcal{J}_t}^{t-1} \eta_t \eta_d \mathbb{E}[\sum_{k\in A_t} q_k \sum_{k'\in A_d} q_{k'}] \\ &\overset{(d)}{\le} \tfrac{EDG}{2}\textstyle\sum_t \sum_{d=t-\mathcal{J}_t}^{t-1} (\eta_t^2 + \eta_d^2) \mathbb{E}[\sum_{k\in A_t} q_k \sum_{k'\in A_d} q_{k'}] \\ &\overset{(e)}{\le} EDGQ\textstyle\sum_t \mathcal{J}_t \eta_{t-\mathcal{J}_t}^2 \sum_{k=1}^N \pi_k q_k := \frac{C_3}{\ln(1/\lambda(\boldsymbol{P}))},\end{aligned}$$

其中，(*a*) 根据式 (22) 得出；(*b*) 应用了三角不等式；(*c*) 使用了式 (23)；(*d*) 应用了柯西–施瓦茨不等式；(*e*) 使用了 $\eta_t < \eta_d \le \eta_{t-\mathcal{J}_t}$ 和 $\sum_{k=1}^N q_k = Q$。 □

3) 证明核心：证明分为两个主要步骤：

1. $\sum_t \eta_t \sum_{k=1}^N \pi_k q_k \mathbb{E}[F_B(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_B^*)] \le C_2 + \frac{C_3}{\ln(1/\lambda(\boldsymbol{P}))}$;
2. $\sum_t \eta_t \sum_{k=1}^N \pi_k q_k \mathbb{E}[F_B(\boldsymbol{w}_{t,0}) - F_B(\boldsymbol{w}_{t-\mathcal{J}_t,0})] \le \frac{C_3}{\ln(1/\lambda(\boldsymbol{P}))}$.

步骤 *1*。结合引理 2 和引理 3，可得：

$$\sum_t \eta_t \mathbb{E}[\sum_{k\in A_t} q_k (F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*))] \le C_1 + \frac{C_3}{\ln(1/\lambda(\boldsymbol{P}))}.$$

常数 $\mathcal{J}_t$ 在 [14], 中引入，是分析中的一个重要参数，并且经常使用。结合定理 2 中的定义和方程 (5)，可得：

$$\left|[P^{\mathcal{J}_t}]_{i,j} - \pi_j\right| \le C_P \lambda(\boldsymbol{P})^{\mathcal{J}_t} \le \frac{1}{2Ht}, \quad \forall i,j \in [M]. \tag{24}$$

假设 $t \ge T_P$。我们推导出一个重要的下界：

$$\begin{aligned}
&\mathbb{E}_{A_t|A_{t-\mathcal{J}_t}}[\textstyle\sum_{k\in A_t} q_k (F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*))] \\
&\overset{(a)}{=} \textstyle\sum_{\mathcal{I}=1}^M \mathbb{P}(A_t{=}\mathcal{I}|A_{t-\mathcal{J}_t}) \sum_{k\in\mathcal{I}} q_k (F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*)) \\
&\overset{(b)}{=} \textstyle\sum_{\mathcal{I}=1}^M \ [P^{\mathcal{J}_t}]_{A_{t-\mathcal{J}_t},\mathcal{I}} \ \sum_{k\in\mathcal{I}} q_k \left(F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*)\right) \\
&\overset{(c)}{\ge} \textstyle\sum_{\mathcal{I}=1}^M \ \left(\pi(\mathcal{I}) - \frac{1}{2Ht}\right) \sum_{k\in\mathcal{I}} q_k (F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*)) \\
&\overset{(d)}{\ge} (\textstyle\sum_{k=1}^N \pi_k q_k) \cdot (F_B(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_B^*) - \frac{1}{2t} MQ,
\end{aligned} \tag{25}$$

其中，(*a*) 是条件期望的定义，(*b*) 利用了马尔可夫性质，(*c*) 由 (24) 得出，而 (*d*) 则由 (8) 得出。取全期望：

$$\begin{aligned}
&(\textstyle\sum_{k=1}^N \pi_k q_k) \sum_t \eta_t \mathbb{E}[F_B(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_B^*] \\
&\quad \le \textstyle\sum_t \eta_t \mathbb{E}[\sum_{k\in A_t} q_k (F_k(\boldsymbol{w}_{t-\mathcal{J}_t,0}) - F_k(\boldsymbol{w}_B^*))] \\
&\qquad + \frac{1}{4} MQ \textstyle\sum_t (\eta_t^2 + \frac{1}{t^2}) = C_2 + \frac{C_3}{\ln(1/\lambda(\boldsymbol{P}))},
\end{aligned} \tag{26}$$

where $C_2 = C_1 + \frac{1}{4} MQ \sum_t (\eta_t^2 + \frac{1}{t^2})$.

步骤 *2*。直接计算可得（与引理 3 类似）：

$$(\textstyle\sum_{k=1}^N \pi_k q_k) \sum_t \eta_t \mathbb{E}[F_B(\boldsymbol{w}_{t,0}) - F_B(\boldsymbol{w}_{t-\mathcal{J}_t,0})] \le \frac{C_3}{\ln(1/\lambda(\boldsymbol{P}))}.$$

将步骤 1 和 2 相加，并应用詹森不等式：

$$\begin{aligned}
&(\textstyle\sum_{t=1}^T \eta_t)(\sum_{k=1}^N \pi_k q_k)\, \mathbb{E}[F_B(\bar{\boldsymbol{w}}_{T,0}) - F_B^*] \le \\
&(\textstyle\sum_{k=1}^N \pi_k q_k) \sum_{t=1}^T \eta_t\, \mathbb{E}[F_B(\boldsymbol{w}_{t,0}) - F_B^*] \le C_2 + \frac{2C_3}{\ln(1/\lambda(\boldsymbol{P}))},
\end{aligned}$$

其中 $\bar{\boldsymbol{w}}_{T,0} := \frac{\sum_{t=1}^T \eta_t \boldsymbol{w}_{t,0}}{\sum_{t=1}^T \eta_t}$，常数见 (12)。 □

C. Proof of Theorem 3

按照定理 1 的相同步骤，将 (20) 展开为：

$$\begin{aligned}
\|\nabla F(\boldsymbol{w}_B^*)\| &\le L\sqrt{\tfrac{2}{\mu}} \textstyle\sum_{k=1}^N |\alpha_k - p_k| \sqrt{(F_k(\boldsymbol{w}_B^*) - F_k^*)} \\
&\le 2L\sqrt{\tfrac{2}{\mu}} d_{TV}(\boldsymbol{\alpha}, \boldsymbol{p}) \sqrt{\Gamma'},
\end{aligned}$$

其中，$d_{TV}(\boldsymbol{\alpha}, \boldsymbol{p}) := \frac{1}{2} \sum_{k=1}^N |\alpha_k - p_k|$ 是概率测度 $\boldsymbol{\alpha}$ 与 $\boldsymbol{p}$之间的全变差距离。 □

D. 最小化ϵ_{opt}

公式 12 定义了以下优化问题：

$$\begin{aligned}
\underset{\boldsymbol{q}}{\text{minimize}} \quad & f(\boldsymbol{q}) = \tfrac{\frac{1}{2}\boldsymbol{q}^\intercal \boldsymbol{A}\boldsymbol{q} + B}{\boldsymbol{\pi}^\intercal \boldsymbol{q}} + C; \\
\text{subject to} \quad & \boldsymbol{q} \ge 0, \\
& \boldsymbol{\pi}^\intercal \boldsymbol{q} > 0, \\
& \|\boldsymbol{q}\|_1 = Q.
\end{aligned}$$

我们通过添加变量 $s := 1/\boldsymbol{\pi}\ \boldsymbol{q}$，然后替换 $\boldsymbol{y} := s\boldsymbol{q}$，来重写该问题。注意，目标函数是凸函数的透视函数，因此是凸的：

$$\min_{\boldsymbol{y},s} \quad f(\boldsymbol{y}, s) = \tfrac{1}{2s} \boldsymbol{y}^\intercal \boldsymbol{A}\boldsymbol{y} + Bs + C \tag{27a}$$

$$\text{s.t.} \quad \boldsymbol{y} \ge 0,\ s > 0,\ \boldsymbol{\pi}^\intercal \boldsymbol{y} = 1,\ \|\boldsymbol{y}\|_1 = Qs. \tag{27b}$$

拉格朗日函数 $\mathcal{L}$如下：

$$\begin{aligned}
\mathcal{L}(\boldsymbol{y}, s, \lambda, \theta, \boldsymbol{\mu}) &= \tfrac{1}{2s} \boldsymbol{y}^\intercal \boldsymbol{A}\boldsymbol{y} + Bs + C + \\
&+ \lambda(1 - \boldsymbol{\pi}^\intercal \boldsymbol{y}) + \theta(\|\boldsymbol{y}\|_1 - Qs) - \boldsymbol{\mu}^\intercal \boldsymbol{y}.
\end{aligned} \tag{28}$$

由于约束 $s > 0$ 定义了一个开集，(27b) 中各约束定义的集合不是闭集。然而，解不会位于边界 $s = 0$ 上，因为 $\mathcal{L}^* \to +\infty$ 当 $s \to 0^+$时成立，因此我们可以考虑 $s \ge 0$。 y_k^* 的 KKT 条件为：

$$\text{if } y_k^* > 0\text{: } y_k^* = \tfrac{s^*}{A[kk]}(\lambda^* \pi_k - \theta^*);\ y_k^* = 0 \text{ otherwise.} \tag{29}$$

由于 $\lambda^* \ge 0$， π_k 较小的客户端可能具有 $q_k^* = 0$。

E. $\epsilon_{opt} + \epsilon_{bias}$的凸性

在附录 D 中，我们证明了 $\epsilon_{\text{opt}}(\boldsymbol{q})$ 是凸的。为证明 $\epsilon_{\text{bias}}(\boldsymbol{q})$ 也是凸的，我们需要研究 $\chi^2_{\boldsymbol{\alpha}\|\boldsymbol{p}} = \sum_{k=1}^N (f_k \circ g_k)(\boldsymbol{q})$ 的凸性，其中 $f_k(p_k) = (p_k - \alpha_k)^2/p_k$，且 $g_k(\boldsymbol{q}) = (\pi_k q_k)/\sum_{h=1}^N \pi_h q_h$。我们观察到 $f_k(p_k)$ 是凸的，而 $g_k(\boldsymbol{q})$ 是线性分式函数 [38] 的一个特例。通过直接检查，可以证明 $(f_k \circ g_k)(\boldsymbol{q})$ 在 dom($f_k \circ g_k$) $= \{\boldsymbol{q}$ 上是凸的： $\|\boldsymbol{q}\|_1 = Q > 0\}$。

F. 合成数据集

我们的合成数据集按以下方式生成：

1) 对于客户端 $k \in \mathcal{K}$，从参数为 1/2的伯努利分布中采样群组标识 i_k ； 2) 从 d维正态分布中采样模型参数 $\boldsymbol{w}^* \sim \mathcal{N}(\boldsymbol{0}, I_d)$； 3) 对于客户端 $k \in \mathcal{K}$和样本索引 $j \in \{1, \ldots, 150\}$，从 d维正态分布中采样客户端输入数据 $\boldsymbol{x}_k^{(j)} \sim \mathcal{N}(\boldsymbol{0}, I_d)$； 4) 对于满足条件的客户端 $k \in \mathcal{K}$ $i_k = 0$ $j \in$和样本 $\{1, \ldots, 150\}$ $y_k^{(j)}$索引，从参数等于 sigmoid($\langle \boldsymbol{w}^*, \boldsymbol{x}_k^{(j)} \rangle$) 的伯努利分布中采样真实标签；5) 对于满足条件 $i_k = 1$的客户端 $k \in \mathcal{K}$ 和样本索引 $j \in \{1, \ldots, 150\}$，从参数等于0.8·sigmoid($\langle \boldsymbol{w}^*, \boldsymbol{x}_k^{(j)} \rangle$)+0.2·(1−sigmoid($\langle \boldsymbol{w}^*, \boldsymbol{x}_k^{(j)} \rangle$))的伯努利分布中采样真实标签 $y_k^{(j)}$。

参考文献

[1] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen 和 J. S. Rellermeyer, “关于分布式机器学习的综述”，*ACM ComputingSurveys(CSUR)*，第 53 卷，第 2 期，第 1–33 页，2020年。

[2] S. Wang、T. Tuor、T. Salonidis、K. K. Leung、C. Makaya、T. He 和 K. Chan，“边缘与学习相遇：面向资源受限分布式机器学习的自适应控制”，载于*2018*年 *IEEE INFOCOM——IEEE*计算机通信会议。IEEE，2018年，第63–71页。

[3] J. Konečný、H. B. McMahan、F. X. Yu、P. Richtarik、A. T. Suresh 和 D. Bacon，“联邦学习：提高通信效率的策略”，载于*NIPS* 隐私多方机器学习研讨会，2016年，https://arxiv.org/abs/1610.05492。

[4] B. McMahan、E. Moore、D. Ramage、S. Hampson 和 B. A. y Arcas，“从去中心化数据中高效地学习深度网络”，载于人工智能与统计学。PMLR，2017年，第 1273–1282 页。

[5] P. Kairouz、H. B. McMahan、B. Avent、A. Bellet、M. Bennis、A. N. Bhagoji、K. Bonawitz、Z. Charles、G. Cormode、R. Cummings等，“联邦学习的进展与未解问题”，机器学习基础与趋势®，第 14 卷，第 1–2 期，第 1–210 页，2021年。

[6] T. Li、A. K. Sahu、A. Talwalkar 和 V. Smith，“联邦学习：挑战、方法与未来方向”，*IEEE*信号处理杂志，第 37 卷，第 3 期，第 50–60 页，2020年。

[7] H. Eichner、T. Koren、B. McMahan、N. Srebro 和 K. Talwar，“半循环随机梯度下降”，发表于国际机器学习大会。PMLR，2019年，第 1764–1773 页。

[8] J. Wang、Z. Charles、Z. Xu、G. Joshi、H. B. McMahan、M. Al-Shedivat、G. Andrew、S. Avestimehr、K. Daly、D. Data 等，“联邦优化指南”，*arXiv* 预印本 *arXiv:2107.06917*，2021年。

[9] K. Bonawitz、V. Ivanov、B. Kreuter、A. Marcedone、H. B. McMahan、S. Patel、D. Ramage、A. Segal 和 K. Seth，“实用的安全聚合用于保护隐私的机器学习”，载于《2017 ACM SIGSAC 计算机与通信安全会议论文集》，2017年，第1175–1191页。

[10] Y. Ding、C. Niu、Y. Yan、Z. Zheng、F. Wu、G. Chen、S. Tang 和 R. Jia，《分块循环数据上的分布式优化》，*arXiv*预印本 *arXiv:2002.07454*，2020年。

[11] C. Zhang、Y. Xie、H. Bai、B. Yu、W. Li 和 Y. Gao，“联邦学习综述：从集中到分布式”，*人工智能评论*，2021年。

[12] T. T. Doan，《局部随机逼近：联邦学习与分布式多任务强化学习算法的统一视角》，*arXiv* 预印本 *arXiv:2006.13460*，2020年。

[13] T. T. Doan、L. M. Nguyen、N. H. Pham 和 J. Romberg，《加速马尔可夫梯度下降的收敛速度及其在强化学习中的应用》，*arXiv* 预印本 *arXiv:2002.02873*，2020年。

[14] T. Sun、Y. Sun 和 W. Yin，《关于马尔可夫链梯度下降》，神经信息处理系统进展，第31卷，2018年。

[15] M. McCloskey 和 N. J. Cohen，《连接主义网络中的灾难性干扰：序列学习问题》，载于学习与动机心理学，G. H. Bower 编。Academic Press，1989年，第24卷，第109–165页。

[16] J. Kirkpatrick、R. Pascanu、N. Rabinowitz、J. Veness、G. Desjardins、A. A. Rusu、K. Milan、J. Quan、T. Ramalho、A. Grabska-Barwinska 等，《克服神经网络中的灾难性遗忘》，美国国家科学院院刊，第114卷，第13期，第3521–3526页，2017年。

[17] M. Tang、X. Chen、W. Guo 和 X. Wang，“联邦学习中的基于相关性的客户端选择策略”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年。

[18] M. Ribero, H. Vikalo, 和 G. De Veciana，“联邦学习：在间歇性客户端可用性和时变通信约束下”，*arXiv* 预印本 *arXiv:2205.06730*，2022年。[19] L. Tan、X. Zhang、Y. Zhou、X. Che、M. Hu、X. Chen 和 D. Wu，“AdaFed：使用自适应聚合权重优化参与感知联邦学习”，*IEEE*网络科学与工程汇刊，2022年。[20] A. Nichol、J. Achiam 和 J. Schulman，“关于一阶元学习算法”，*arXiv* 预印本 *arXiv:1803.02999*，2018年。[21] S. J. Reddi、Z. Charles、M. Zaheer、Z. Garrett、K. Rush、J. Konečný、S. Kumar 和 H. B. McMahan，“自适应联邦优化”，载于国际学习表征会议，2021年。[22] X. Li、K. Huang、W. Yang、S. Wang 和 Z. Zhang，“FedAvg 在非独立同分布数据上的收敛性”，载于国际学习表征会议，2019年。

[23] T. Li、A. K. Sahu、M. Zaheer、M. Sanjabi、A. Talwalkar、和 V. Smith，“异构网络中的联邦优化”，机器学习与系统会议论文集，第2卷，第429–450页，2020年。

[24] W. Chen、S. Horvath 和 P. Richtarik，“面向联邦学习的最优客户端采样”，*arXiv* 预印本 *arXiv:2010.13723*，2020年。

[25] Y. Fraboni、R. Vidal、L. Kameni 和 M. Lorenzi，“聚类采样：降低方差并提高联邦学习中客户端选择的代表性”，发表于国际机器学习大会。PMLR，2021年，第3407–3416页。

[26] Y. Jee Cho、J. Wang 和 G. Joshi，《联邦学习中有偏客户端选择的理解》，载于第*25*届人工智能与统计学国际会议论文集。PMLR，2022年，第10 351–10 375页。

[27] T. T. Doan、L. M. Nguyen、N. H. Pham、和J. Romberg，《有限随机梯度下降在马尔可夫随机性下的时间分析，”*arXiv*预印本 *arXiv:2003.10973*，2020年。

[28] A. Mokhtari 和 A. Ribeiro，《随机拟牛顿的全局收敛性》，*运筹学与工程学报*，2021年。

[29] H. Ohno，《动力学流》，载于《亨利·庞加莱研究所年刊（*B*辑）：概率与统计》，第 33 卷，第 4 期。爱思唯尔，1997 年，第 497–528 页。

[30] D. A. Levin and Y. Peres, 马尔可夫链与混合时间，美国数学会，2017年，第107卷。

[31] L. Bottou, F. E. Curtis 和 J. Nocedal，《大规模机器学习的优化方法，》*SIAM Review*，第60卷，第2期，第223–311页，2018年。

[32] J. Wang, Q. Liu, H. Liang, G. Joshi 和 H. V. Poor，“解决异构联邦优化中的目标不一致问题”，神经信息处理系统进展，第 33 卷，第 7611–7623 页，2020年。

[33] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville 和 Y. Bengio，“基于梯度的神经网络中灾难性遗忘的实证研究”，载于 国际学习表征会议，2013，arXiv 预印本 arXiv:1312.6211。

[34] R. Kemker, M. McClure, A. Abitino, T. Hayes, 和 C. Kanan，“衡量神经网络中的灾难性遗忘”，发表于*AAAI*人工智能大会论文集，第 32 卷，第 1 期，2018 年。

[35] Y. LeCun 和 C. Cortes，“MNIST手写数字数据库”，2010年。

[36] F. Sattler、K.-R. Müller 和 W. Samek，“聚类联邦学习：隐私约束下与模型无关的分布式多任务优化”，*IEEE* 神经网络与学习系统汇刊，第 32 卷，第 8 期，第 3710–3722 页，2020年。

[37] H. Bayerkhorn，联邦学习：方法与应用的全面概述。SpringerCham，2022年。

[38] S. Boyd 和 L. Vandenberghe，凸优化。剑桥大学出版社，2004年。