arXiv:2305.04630v1 [cs.LG] 8 May 2023

通过空中计算实现无线网络中的联邦学习*

Halil Yigit Oksuz[1,2,3], Fabio Molinari[1],Henning Sprekeler[2,3],Jörg Raisch[1,2]

摘要—**在多智能体系统中，智能体可以通过交换各自估计的模型参数，协作学习一个模型，而无需交换智能体本地持有的数据。这种策略通常称为**联邦学习，**主要出于两个原因：（i）避免共享可能规模庞大的数据集，从而提高资源效率；（ii）保障本地智能体数据的隐私。采用一种名为“空中计算”的超越5G通信策略，还可以进一步提高效率。该策略利用无线信道的干扰特性。标准通信方案通过让不同智能体在不同的时隙或频隙发送信号来避免干扰，而空中计算无需如此，从而节省资源。在这种情况下，接收信号是发送信号的加权和，权重未知（即衰落信道系数）。该领域的现有研究旨在重建这些未知系数。与此不同，本文提出的方法无需通过复杂的编解码方案重建信道系数，从而同时提升效率和隐私。**

I. 引言

过去十年间，机器学习相关主题因其在许多应用领域的成功而备受关注（例如， [1]–[3]）。随着计算能力不断提升，设备变得更小巧，却也更加智能、性能更强，能够处理海量数据和更复杂的计算。

数据通常分布在这些智能设备或智能体上，例如智能手机、个人计算机和分布式数据中心。在集中式学习方法中，所有数据都会收集到一台计算能力强大的机器上，例如云端单元或服务器，并在其上训练模型 [4]。然而，对于需要保护隐私、提高资源效率和实现低时延的特定应用而言，这种方法可能并不合适。例如，出于隐私方面的顾虑，智能手机用户可能不愿分享自己的数据；此外，随着智能体数量增加，必须分配更多通信资源来分享这些数据】【。智能体的 data 与中央单元。然而，也可以在不进行集中式数据收集的情况下训练模型。在这种情况下，智能体分别使用各自的数据训练模型，并将各自训练好的模型（以参数向量的形式）共享给中央单元。通常，与共享所有本地数据集相比，共享参数向量所需的通信资源更少。这种方法通常称为联邦学习，能够实现更具隐私性和通信效率的学习（参见 [5]–[8]）。

尽管联邦学习通过避免共享各个数据集来提升隐私，但与中央单元共享本地模型参数仍可能泄露敏感信息 [9]–[11]。可以采用加密或差分隐私等常见技术来解决这一隐私问题 [12], ，但代价是效率和性能降低 [9], [13]。

随着智能体数量和模型参数数量的增加，整个系统的通信负载也会随之增加 [10], [11], [14]。为解决这一问题，可以让智能体进行多步本地训练，或对模型更新进行量化 和/或稀疏化 ，以实现高效压缩；另一种方法是在特定时间步仅允许部分智能体传输 [15]–[19]。然而，从资源效率的角度看，这些技术大多并不高效，因为它们会增加带宽需求或通信轮次，通常会导致总吞吐量和学习速度下降，正如[20], [21]中也观察到的那样。

相反，我们考虑采用所谓的空中计算（ Over-The-Air Computation）方法，例如， [22]，以提高通信效率。当多个智能体同时在同一频段内传输时，信号会受到干扰这一物理现象的影响。标准通信协议通过正交传输信号来避免干扰：在时分多址（TDMA）中，为智能体分配不同的时隙以供其传输；而在频分多址（FDMA）中，则为不同用户分配不同的频段。

空中计算的理念是利用干扰，而非与之对抗，从而提高通信效率。例如，在一个由 N 个智能体组成的系统中，每个智能体都需要向中央单元传输一个m维参数向量 $\theta \in \Re^m$ ，采用时分多址或频分多址需要进行 mN 次正交传输（通过时分或频分复用），而空中计算则需要

* 本研究由德国研究联合会（DFG，德国研究基金会）通过德国卓越战略资助，项目为 EXC 2002/1“智能科学”，项目编号 390523135。[1] H.Y. Oksuz、F. Molinari 和 J. Raisch 任职于德国柏林工业大学控制系统组。

{oksuz@tu-berlin.de},{molinari,raisch}@control.tu-berlin.de。[2] H.Y. Oksuz、H. Sprekeler 和 J. Raisch 任职于德国柏林工业大学智能科学卓越集群，地址：Marchstr. 23, 10587, 柏林，德国。[3] H.Y. Oksuz 和 H. Sprekeler 任职于德国柏林工业大学认知过程建模组。

仅采用m 正交传输。由于其优势，联邦学习已由 [23]–[25] 通过空中计算实现。

有鉴于此，本文的主要贡献可概括如下：

- 不同于 [23]和 [25],的研究，我们不假设已知信道系数（也不假设能够重构这些系数），而是提出了一种能够应对其未知特性的算法。我们无需额外进行预处理来重构信道，因此所提出的方案在时间和资源方面更加高效。

- 由于信道特性未知，隐私得到了固有保障。中央单元无法重构各个智能体传输的信息。

本文其余部分安排如下：第二节介绍问题设置。第三节提出所设计的联邦学习算法并证明其收敛性。第四节给出一个数值示例。第五节给出总结性评论。

符号说明

实数集记为 $\Re$，$\Re^m$ 表示 m 维欧几里得空间。$\mathbb{N}$ 和 $\mathbb{N}_0$ 分别表示自然数集和非负整数集。对于向量 $\mathrm{x} \in \Re^m$，x^T 表示其转置。向量 x 的二范数记为 $||\mathrm{x}||$。随机变量 p 的期望值记为 $\mathbb{E}[\mathrm{p}]$。给定有限集 S，其基数记为 $|\mathrm{S}|$。对于可微函数 $\mathrm{f}: \Re^m \to \Re$，$\nabla \mathrm{f}(\mathrm{x})$ 表示函数 f 在 $\mathrm{x} \in \Re^m$处的梯度。将 $\mathrm{x} \in \Re^m$ 投影到非空闭凸集 $\mathrm{S} \subset \Re^m$ 所得结果记为 $\mathrm{P}_S(\mathrm{x})$，其中 $\mathrm{P}_S(\mathrm{x}) = \mathrm{argmin}_{s\in S}||\mathrm{s}-\mathrm{x}||$。

II. 问题描述

A. 带约束的联邦学习

考虑一个由N 个智能体组成的系统，这些智能体连接到服务器或云端单元，目标是执行一项机器学习任务。每个智能体都可以访问数据集的不同部分，并具有各自的局部代价函数。我们用i号智能体可访问的数据集合$D_i = \{d_i^n\}_{n=1}^{|D_i|}$表示，并用 $\mathcal{L}_i(d_i^n, \theta)$表示参数为 $\theta \in \Re^m$的模型的代价函数值。在监督学习情境下，数据集D_i 由输入和目标组成的样本对构成，即$d_i^n = (u_i^n, z_i^n)$，其中u_i^n和z_i^n分别表示输入数据和目标数据。此外，智能体i 的私有局部代价函数可表示为

$$f_i(\theta) = \frac{1}{|D_i|}\sum_{n=1}^{|D_i|}\mathcal{L}_i(d_i^n, \theta), \tag{1}$$

若全局代价函数定义为所有局部代价函数的平均值，即

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} f_i(\theta), \tag{2}$$

那么，整个系统的目标是协同求解以下约束优化问题

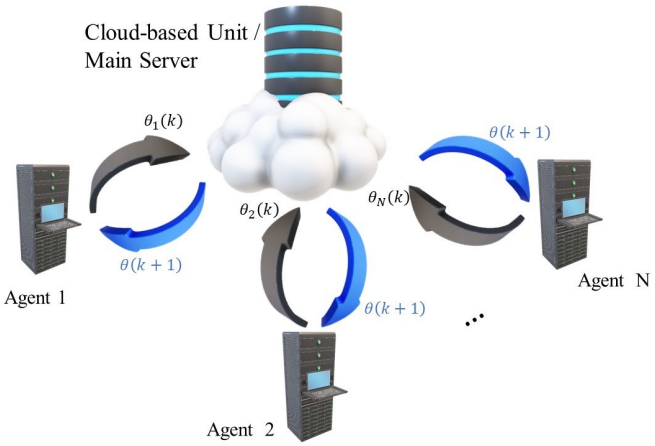


图 1：联邦学习场景示意图。

$$\theta^* = \arg\min_{\theta\in\Theta}\mathcal{L}(\theta) = \arg\min_{\theta\in\Theta}\frac{1}{N}\sum_{i=1}^{N} f_i(\theta), \tag{3}$$

其中，$\Theta \subset \Re^m$ 是一个非空约束集。下文中，我们将 $\mathcal{L}(\theta^*)$记为 $\mathcal{L}^*$。此外，最优解集合定义为

$$\Theta^* = \{\theta \in \Theta \mid \frac{1}{N}\sum_{i=1}^{N} f_i(\theta) = \mathcal{L}^*\}.$$

若整个数据集，即

$$D = D_1 \cup D_2 \cup \cdots \cup D_N$$

如果服务器已知这些数据，就可以采用基于集中式梯度下降的优化方法来解决上述全局学习任务 [3]。然而，在所考虑的联邦学习方法中，每个智能体只能访问自己的数据集，并基于该数据集训练自己的模型。经过若干本地优化步骤后，局部参数估计的当前版本会被传输至服务器进行聚合。聚合后的版本随后会传输给各智能体，用于下一步优化（见图 1）。

B. 空中通信

在无线通信系统中，无线多址信道（WMAC）模型已被广泛用于描述多个发射器通过衰落信道与单个接收器之间的通信，例如 [26]，[27]。在本文中，我们采用 [20],[28] 所描述的基于 WMAC 的通信模型。在该模型中，多个（假设为 N 个）智能体在同一频段内同时发送信息 $\{\mathbf{s}_i(k) \in \Re^m\}_{i=1}^N$。这些信息受到信道的干扰，并在接收器处叠加，即接收到的信息为

$$\mathbf{r}^{rec}(k) = \sum_{i=1}^{N}\alpha_i(k)\mathbf{s}_i(k), \tag{4}$$

其中，$\alpha_i(k)$ 是未知的时变正信道系数。

算法1： FedCOTA

初始化： $\theta(0) \in \Theta$.

学习循环：

1：**对**每个时间步$k \in \mathbb{N}_0$ **执行**

2：服务器广播 $\theta(k)$ 3：每个智能体i 更新其本地变量：

$$\theta_i(k) = \theta(k) - \eta(k)\nabla f_i\big(\theta(k)\big) \tag{5}$$

$$\rho_i(k) = 1 \tag{6}$$

4: 每个智能体i 传输 $\theta_i(k)$ 和 $\rho_i(k)$

5：服务器接收：

$$\theta^{\text{rec}}(k) = \sum_{i=1}^{N} \alpha_i(k)\theta_i(k) \tag{7}$$

$$\rho^{\text{rec}}(k) = \sum_{i=1}^{N} \alpha_i(k)\rho_i(k) \tag{8}$$

6：服务器更新：

$$\theta(k+1) = \mathbf{P}_\Theta\left(\frac{\theta^{\text{rec}}(k)}{\rho^{\text{rec}}(k)}\right) \tag{9}$$

7: **结束循环**

如上所述，将空中计算用于联邦学习可带来以下优势：

- **隐私**：信道系数$\{\alpha_i(k)\}_{Ni=1}$ 在 (4) 中未知，因此中央单元不可能单独从$\mathbf{r}_{rec}(k)$重构 $\{s_i(k)\}_{Ni=1}$ ；
- **效率**：如 [21], 所示，随着发射器数量的增加，空中计算能够实现优于线性增长的通信成本扩展。

与 [25], 提出的方法不同，我们的方法不需要了解信道系数，也无需在每个时间步重建这些系数，从而消除了一项复杂性来源。

III. 联邦约束空中学习（FEDCOTA）算法

本节首先介绍所提出的算法，该算法使智能体能够通过空中通信进行联邦学习。随后，我们将研究所提出算法的收敛性质。

A. 算法

FedCOTA算法总结于算法1中。开始时，服务器向智能体广播 $\theta(0) \in \Theta$。随后，在迭代过程中，每个智能体使用局部更新规则 (5) 计算自己的参数向量 $\theta_i(k)$。之后，所有智能体同时（并在同一频段内）将其局部参数向量 $\theta_i(k)$ 发送至中央单元。然后，所有智能体再次同时在同一频段内发送一个常数 $\rho_i(k) = 1$ 。由于无线信道的叠加特性（参见第二节-B），服务器在时间步 k 接收到 (7) 和 (8)。最后，服务器计算 (9)，从而得到$\theta(k+1)$，可将其写为

$$\begin{aligned}\theta(k+1) &= \mathbf{P}_\Theta\left(\frac{\theta^{\text{rec}}(k)}{\rho^{\text{rec}}(k)}\right) = \mathbf{P}_\Theta\left(\sum_{i=1}^{N}\frac{\alpha_i(k)}{\sum_{i=1}^{N}\alpha_i(k)}\theta_i(k)\right)\\ &= \mathbf{P}_\Theta\left(\sum_{i=1}^{N} h_i(k)\theta_i(k)\right)\end{aligned} \tag{10}$$

其中，$\mathbf{P}_\Theta$ 是到集合 Θ 上的投影算子，$\alpha_i(k)$ 是未知的时变正实信道系数，而 $h_i(k) = \alpha_{i(k)}\Sigma_{Ni=1}\alpha_{i(k)}$ 是相应的归一化信道系数。根据构造，$h_i(k)$ 为正，且

$$\sum_{i=1}^{N} h_i(k) = 1, \tag{11}$$

对所有 $k \geq 0$。

B. 预备知识

以下假设也常见于该领域的相关论文，并将在全文中成立。

假设 *1*：约束集 $\Theta \subset \Re^m$ 是凸且紧的。因此（参见 [29, 定理 2.41，第 40] 页），Θ 也因此是闭且有界的。

假设*2*： 代价函数 $f_i(\theta)$ 连续可微且强凸，其梯度是 L-利普希茨连续的，即，对于任意 θ_1，$\theta_2 \in \Theta \subset \Re^m$ 且对所有$i = 1, 2, .., N$，以下不等式成立：

(i) $\exists \mu_i > 0$ 使得

$$f_i(\theta_2) - f_i(\theta_1) \geq \nabla f_i(\theta_1)^T(\theta_2 - \theta_1) + \frac{\mu_i}{2}||\theta_2 - \theta_1||^2, \tag{12}$$

(ii) $\exists L_i > 0$ 使得

$$||\nabla f_i(\theta_1) - \nabla f_i(\theta_2)|| \leq L_i||\theta_1 - \theta_2||. \tag{13}$$

注释*1*： 注意，全局代价函数 $\mathcal{L}(\theta)$ 也因此可微且为强凸函数，其梯度为 L-Lipschitz连续梯度，即

$$\mathcal{L}(\theta_2) - \mathcal{L}(\theta_1) \geq \nabla\mathcal{L}(\theta_1)^T(\theta_2 - \theta_1) + \frac{\mu}{2}||\theta_2 - \theta_1||^2$$

$$||\nabla\mathcal{L}(\theta_1) - \nabla\mathcal{L}(\theta_2)|| \leq L||\theta_1 - \theta_2||,$$

其中 $\mu = \frac{1}{N}\sum_{i=1}^{N}\mu_i$ 且 $L = \frac{1}{N}\sum_{i=1}^{N}L_i$（见 [30]）。

注释*2*： 注意，假设1和假设2意味着局部和全局代价函数在Θ上具有有界梯度。

注意，假设1和假设2已广泛应用于联邦学习和分布式优化文献中，用以说明问题（3）的解存在，以及所提出算法的收敛性质（见[3], [25], [31]–[35], 及其中的参考文献）。

假设*3*： 对于所有 $k \geq 0$，算法1中的步长选为 $\eta(k) = \frac{\eta_c}{\sqrt{k+1}}$，其中 $0 < \eta_c \leq \frac{1}{L}$。

备注 *3*： 在假设3下，步长 $\eta(k)$ 递减，并满足 $\sum_{k=0}^{\infty}\eta(k) = \infty$ 且 $\lim_{k\to\infty}\eta(k) = 0$。

许多以往研究（例如， [3],）也采用了与关于步长的假设3类似的条件，

[34]–[40], 及其中的参考文献）来证明精确收敛至最优解。

引理 *1*： 若$f(\cdot)$连续可微、凸，且具有L-Lipschitz连续梯度，则对任意 $\theta_1, \theta_2 \in \Re^m$，下列不等式成立：

$$0 \le f(\theta_2) - f(\theta_1) - \nabla f(\theta_1)^T(\theta_2 - \theta_1) \le \frac{L}{2}||\theta_1 - \theta_2||^2.$$

证明： 该结论可直接由 [35,定理 2.1.5]得出。 ■

注*4*： 由引理 1 和强凸性的定义（12），可得 $\mu \le L$ 若函数连续可微、强凸且具有L-Lipschitz连续梯度。因此，根据假设3， $\eta_c \le \frac{1}{L} \le \frac{1}{\mu}$ 成立。

假设 *4*： 假设未知的时变正实信道系数是同一概率分布的独立实现（在时间和智能体之间），即 $\forall k \in \mathbb{N}_0$, $\alpha_i(k) \sim \mathcal{D}\big(\bar{\alpha}, \text{方差}(\alpha)\big)$, 其中 $\bar{\alpha}$和方差(α)分别表示分布 $\mathcal{D}$的均值和方差。

如 [20], [28], 所述，我们在此参考 [41, 第 2.3 章、第 2.4]章以及 [42, 第 5.4], 章，因此将信道系数视为相同概率分布的独立实现（见[43]）。

引理 *2*: 在假设 4 下， $\mathbb{E}[h_i(k)] = \frac{1}{N}$ 对所有 $i = 1,2,\ldots,N$和 $k \ge 0$均成立。

证明：由 (11) 可得 $\mathrm{h_i(k)} = \alpha\mathrm{i(k)}\,\Sigma\mathrm{Ni=1}\,\alpha\mathrm{i(k)}$，以及 $\sum_{i=1}^{N} \mathrm{h}_i\mathrm{(k)} = 1$。根据假设 4，每个 α_i(k) 均从相同的分布中采样，因此每个 h_i(k) 的底层分布相同，即 $\mathbb{E}[\mathrm{h}_i\,\mathrm{(k)}] = \mathbb{E}[\mathrm{h}_j\mathrm{(k)}]$ 对所有 $\mathrm{i, j} \in \{1, 2, \cdots, \mathrm{N}\}$ 和 $\forall \mathrm{k} \ge 0$ 均成立。此外，由于期望的线性性，我们有

$$1 = \mathbb{E}\Big[\sum_{i=1}^{N} h_i(k)\Big] = \sum_{i=1}^{N} \mathbb{E}[h_i(k)] = N\mathbb{E}[h_i(k)],$$

这意味着 $\mathbb{E}[h_i(k)] = \frac{1}{N}$对所有$i = 1, 2, \ldots, N$均成立。 ■

*C. FedCOTA*算法的收敛性分析

首先，我们给出证明FedCOTA算法收敛所需的预备结果。考虑

$$C(k) = C_k = 1 - \eta(k)\mu. \tag{14}$$

注意，在假设3下，$C_k \ge 0$ 对$k \in \mathbb{N}_0$成立。由于$\eta(k)$递减，C_k 随k递增，因此只需证明$C_0 \ge 0$。这可由$C_0 = 1 - \eta_c\mu$和 $\eta_c \le \frac{1}{\mu}$ 立即得到（见备注4），故$C_0 \ge 0$。还需注意，由于C_k 递增，$C_k > 0$ 对所有$k \ge 1$均成立。

引理*3*： 在假设3下，$\lim_{k\to\infty}(C_k)^k = 0$。

证明： 令$C_k = 1 - \frac{Q}{\sqrt{k+1}}$，其~~中~~$Q = \eta_c\mu > 0$，则可写为

$$\lim_{k\to\infty}(C_k)^k = \lim_{k\to\infty}\Big(1 - \frac{Q}{\sqrt{k+1}}\Big)^k. \tag{15}$$

注意到 $\forall x \in \Re$，我们有 $1 + x \le e^x$。因此， $\forall k \in \mathbb{N}_0$

$$C_k = 1 - \frac{Q}{\sqrt{k+1}} \le e^{-\frac{Q}{\sqrt{k+1}}}. \tag{16}$$

此外，如上文所述，$C_k \ge 0$，因此

$$0 \le (C_k)^k \le e^{-\frac{Qk}{\sqrt{k+1}}}. \tag{17}$$

由于当$k \to \infty$时，(17)右侧的项趋于零，因此我们已证明 $\lim_{k\to\infty}(C_k)^k = 0$。 ■

Lemma 4: Under Assumption 3, $\prod_{k=0}^{\infty} C_k = 0$.

证明： 利用关系式 $1 + x \le e^x$, $\forall x \in \Re$，可得

$$\begin{aligned}\prod_{t=0}^{k} C_t &= \prod_{t=0}^{k}\big(1 - \eta(t)\mu\big) \\ &\le \prod_{t=0}^{k} \mathrm{e}^{-\eta(t)\mu} = \mathrm{e}^{-\sum_{t=0}^{k}\eta(t)\mu}\end{aligned} \tag{18}$$

对两边取$k \to \infty$ 时的极限，得到

$$\lim_{k\to\infty}\prod_{t=0}^{k} C_t \le \mathrm{e}^{-\mu\sum_{t=0}^{\infty}\eta(t)} = 0 \tag{19}$$

since by Assumption 3, $\mu\sum_{t=0}^{\infty}\eta(t) = \infty$. ■

引理*5*： 在假设3下，对于$k \to \infty$,

$$\sum_{t=0}^{k-2}\Big(\prod_{l=t+1}^{k-1} C_l\Big)\eta^2(t) + \eta^2(k-1)$$

收敛到任意小的正值 ε。

证明： 见附录。 ■

现在我们可以介绍主要结果。

定理 *1*： 假设 1、2、3 和 4 成立。则，$\exists \theta^* \in \Theta^*$使得对于$k \to \infty$， $\mathbb{E}\big[||\theta(k) - \theta^*||^2\big]$ 可以任意小。

证明： 对于任意 $\theta^* \in \Theta^*$，利用（5）、（10）、（11）以及投影$\mathbf{P}_\Theta$的非扩张[1] 性质，可得

$$\begin{aligned}||\theta(k+1) - \theta^*||^2 &= \Big|\Big|\mathbf{P}_\Theta\Big(\sum_{i=1}^{N} h_i(k)\theta_i(k)\Big) - \mathbf{P}_\Theta(\theta^*)\Big|\Big|^2 \\ &\le \Big|\Big|\sum_{i=1}^{N} h_i(k)\Big(\theta(k) - \eta(k)\nabla f_i(\theta(k))\Big) - \theta^*\Big|\Big|^2 \\ &= ||\theta(k) - \theta^*||^2 \\ &\quad - 2\eta(k)\sum_{i=1}^{N} h_i(k)\nabla f_i^T(\theta(k))(\theta(k) - \theta^*) \\ &\quad + \Big|\Big|\eta(k)\sum_{i=1}^{N} h_i(k)\nabla f_i(\theta(k))\Big|\Big|^2.\end{aligned} \tag{20}$$

[1] $||\mathbf{P}_\Theta(x) - \mathbf{P}_\Theta(y)|| \le ||x - y||$ 对所有$x, y \in \Re^m$ 均成立，前提是 Θ 为非空闭凸集（见 [44]）。

注意，约束集 Θ 的有界性和 $\nabla f_i(\theta(k))$ 的利普希茨连续性意味着存在 $\exists D_\Theta, M > 0$，使得

$$\left|\left|\theta(k)-\theta^*\right|\right| \le D_\Theta, \tag{21}$$
$$\left|\left|\nabla f_i(\theta(k))\right|\right| \le M. \tag{22}$$

因此，(20) 右侧最后一项的上界可写为

$$\begin{aligned}\left|\left|\eta(k)\sum_{i=1}^{N} h_i(k)\nabla f_i(\theta(k))\right|\right|^2 &= \eta^2(k)\left|\left|\sum_{i=1}^{N} h_i(k)\nabla f_i(\theta(k))\right|\right|^2 \\ &\le \eta^2(k)\sum_{i=1}^{N} h_i(k)\left|\left|\nabla f_i(\theta(k))\right|\right|^2 \\ &\le \eta^2(k)M^2,\end{aligned} \tag{23}$$

这由 (11)、(22) 以及函数 $||\cdot||^2$ 的凸性可得。随后，对 (20) 两边取期望，并利用期望的线性性，可得

$$\begin{aligned}\mathbb{E}\left[||\theta(k+1)-\theta^*||^2\right] \le{}& \mathbb{E}\left[\left|\left|\theta(k)-\theta^*\right|\right|^2\right] \\ &- 2\eta(k)\mathbb{E}\Big[\sum_{i=1}^{N} h_i(k)\nabla f_i^T(\theta(k))\left(\theta(k)-\theta^*\right)\Big] \\ &+ \eta^2(k)M^2,\end{aligned} \tag{24}$$

其中，右侧第二项可写为

$$\begin{aligned}&-2\eta(k)\mathbb{E}\Big[\sum_{i=1}^{N} h_i(k)\nabla f_i^T(\theta(k))\left(\theta(k)-\theta^*\right)\Big] \\ &= -2\eta(k)\sum_{i=1}^{N}\mathbb{E}\left[h_i(k)\nabla f_i^T(\theta(k))\left(\theta(k)-\theta^*\right)\right].\end{aligned} \tag{25}$$

注意，$h_i(k)$ $(i=1,2,\cdots,N)$ 的统计特性与 $h_i(t)$ 在 $t<k$ 时无关，这意味着 $\theta(k)$ 和 $h_i(k)$ 在时刻 k 是统计独立的，因为 $\theta(k)$ 的统计特性仅取决于 $h_i(t)$ 在 $t<k$ 时的情况，以及 $i=1,2,\cdots,N$。因此，利用式 (2)、引理 2 以及期望的线性性质，我们可以进一步将式 (25) 写为

$$\begin{aligned}&-2\eta(k)\sum_{i=1}^{N}\mathbb{E}\left[h_i(k)\nabla f_i^T(\theta(k))\left(\theta(k)-\theta^*\right)\right] \\ &= -2\eta(k)\sum_{i=1}^{N}\mathbb{E}\left[h_i(k)\right]\mathbb{E}\left[\nabla f_i^T(\theta(k))\left(\theta(k)-\theta^*\right)\right] \\ &= -2\eta(k)\sum_{i=1}^{N}\frac{1}{N}\mathbb{E}\left[\nabla f_i^T(\theta(k))\left(\theta(k)-\theta^*\right)\right] \\ &= -2\eta(k)\mathbb{E}\Big[\sum_{i=1}^{N}\frac{1}{N}\nabla f_i^T(\theta(k))\left(\theta(k)-\theta^*\right)\Big] \\ &= -2\eta(k)\mathbb{E}\left[\nabla\mathcal{L}^T(\theta(k))\left(\theta(k)-\theta^*\right)\right].\end{aligned} \tag{26}$$

此外，根据假设2（$f_i(\theta)$ 和 $\mathcal{L}(\theta(k))$ 的强凸性），我们有

$$\begin{aligned}-2\eta(k)\nabla\mathcal{L}^T(\theta(k))\left(\theta(k)-\theta^*\right) \le{}& -2\eta(k)\left(\mathcal{L}(\theta(k))-\mathcal{L}(\theta^*)\right) \\ &-\eta(k)\mu||\theta(k)-\theta^*)||^2.\end{aligned} \tag{27}$$

注意，对于所有 $k\ge 0$，$\theta^*\in\Theta^*$（约束集中的最优点），均有 $\mathcal{L}(\theta(k))-\mathcal{L}(\theta^*)\ge 0$。结合对 (27) 两边取期望，可得

$$-2\eta(k)\mathbb{E}\left[\nabla\mathcal{L}^T(\theta(k))\left(\theta(k)-\theta^*\right)\right] \le -\eta(k)\mu\mathbb{E}\left[||\theta(k)-\theta^*)||^2\right]. \tag{28}$$

利用 (25)–(28)，可将 (24) 改写为

$$\begin{aligned}\mathbb{E}\left[||\theta(k+1)-\theta^*||^2\right] &\le \left(1-\eta(k)\mu\right)\mathbb{E}\left[\left|\left|\theta(k)-\theta^*\right|\right|^2\right] \\ &\quad+\eta^2(k)M^2 \\ &= C_k\mathbb{E}\left[||\theta(k)-\theta^*||^2\right]+\eta^2(k)M^2.\end{aligned} \tag{29}$$

递归关系 (29) 可写为

$$\begin{aligned}\mathbb{E}\left[\left|\left|\theta(k)-\theta^*\right|\right|^2\right] \le{}& \left(\prod_{t=0}^{k-1}C_t\right)\mathbb{E}\left[\left|\left|\theta(0)-\theta^*\right|\right|^2\right] \\ &+M^2\left(\sum_{t=0}^{k-2}\Big(\prod_{l=t+1}^{k-1}C_l\Big)\eta^2(t)+\eta^2(k-1)\right).\end{aligned} \tag{30}$$

注意，根据引理4，$\prod_{t=0}^{\infty}C_t=0$ 成立。此外，根据引理5和假设3，对式（30）两侧取极限，即可完成证明。 ■

IV. 数值示例

现在，我们将 FedCOTA算法应用于联邦逻辑回归问题。在该问题中，智能体系统中的每个智能体只能访问自己的本地数据集，却要共同完成一项全局二分类任务。设第i个智能体可用的数据集为$D_i=\{d_i^n\}_{n=1}^{|D_i|}$，其中$d_i^n=(u_i^n,z_i^n)\in\Re^m\times\{0,1\}$，而$u_i^n$和$z_i^n$分别表示第$i$个智能体可用的输入数据和标签。注意，所有智能体的数据都分为两个不同的数据类别，分别标记为 0 或 1；它们的目标是利用现有数据在 $\Re^m$ 中找到一个分离超平面，以便智能体能够区分来自不同类别的一些未见数据。为了以联邦方式实现这一目标，每个智能体都将交叉熵用作其局部代价函数，可表示为

$$\begin{aligned}f_i(\theta,d_i)=\lambda||\theta||_2^2-\frac{1}{|D_i|}\Bigg(&\sum_{n=1}^{|D_i|}z_i^n\log\left(S(\theta^Tu_i^n)\right) \\ &+(1-z_i^n)\log\left(1-S(\theta^Tu_i^n)\right)\Bigg)\end{aligned} \tag{31}$$

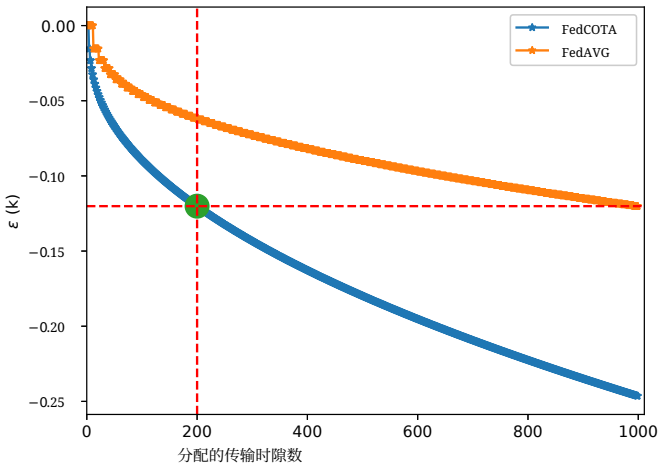


图 2：全局逻辑回归模型与联邦逻辑回归模型的比较。前者在整个数据集上进行训练，而后者通过 FedCOTA 和 FedAVG 进行训练。

其中，$\hat{z}_i^n = S(\theta^T u_i^n)$是第$i$个智能体对标签$z_i^n$的局部估计，$S(x) = \frac{1}{1+e^{-x}}$是 Sigmoid函数，$-||\theta||_2$是参数向量 θ的L_2-范数，λ是正则化超参数，其中 $\lambda = 0$ 表示不进行正则化，而 $\lambda = 1$ 表示正则化程度达到最大。注意，仿真中选取了 $\lambda = 0.0001$ ，且式（31）给出的总体代价函数是强凸的。此外，由于参数向量 θ始终处于闭且有界的（约束）集 Θ中，式（31）的二阶导数有界，因此其梯度是Lipschitz连续的。另外，对于每个智能体，步长统一设为 $\eta(k) = \frac{1}{\sqrt{k+1}}$，此时总体代价函数为

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} f_i(\theta, d_i) \tag{32}$$

其中，N 表示系统中的智能体数量。我们考虑一个由 10 个智能体组成的系统，每个智能体共有 $|D_i| = 100$ 个训练样本。参数向量的维度为 $m = 3$，即 $\theta \in \Re^3$，其中也包括一个偏置项。目标是求解 $\theta^* = \text{argmin}_{\theta\in\Theta}\mathcal{L}(\theta)$，其中 $\Theta = \{\theta \in \Re^3 | \ ||\theta(k)|| \leq 15, \forall k \geq 0\}$ 是约束集。

为了获得所需的参数向量 θ_d，我们训练了一个能够访问整个数据集的全局逻辑回归模型。我们预计参数向量 $\theta(k)$ 会趋于收敛 θ_d。可以通过计算 $\varepsilon(k) = \log_{10}\left(\frac{\theta}{\theta}_{(k)-\theta_d(0)-\theta_d}\right)$ 来评估这一点，其中每个通信轮次对应一个 $k \geq 0$。如图 2 所示，使用 FedCOTA 算法的智能体的参数向量趋于收敛到 θ_d。

请注意，图 2 还比较了 Fed-COTA 算法与标准 FedAVG算法 [5]，后者采用时分多址方案进行通信智能体与中央单元之间的通信。在这种情况下，每个通信轮次都会为每个智能体分配单独的时隙，以传输其参数向量。服务器收到参数向量后，会计算它们的平均值并将结果发送回智能体。由于我们考虑的是一个包含 $N = 10$ 个智能体的系统，每个智能体每轮通信需要 10 个时隙来传输更新后的参数向量，而 FedaOTA算法只需 2 个时隙（见式 (7) 和 (8)），因此其速度是 FedAVG算法的 5 倍（见图 2）。

V. 结论

本文研究了通过空中计算实现联邦学习的问题，该方法在通信效率和隐私方面带来了显著提升。我们通过考虑时变步长，研究了所提基于梯度的算法的收敛性质。随后，我们给出了一些数值示例，以说明理论结果。

我们目前的研究方向是使用随机梯度下降，这种方法在训练样本数量较大时更有效。此外，未来的工作还将研究智能体之间完全分布式通信的情形，以及用户间数据分布非独立同分布（non-iid）的情形。

VI. 附录：引理5的证明

根据假设3，C_k 递增且非负。因此，我们有

$$\sum_{t=0}^{k-2}\Big(\prod_{l=t+1}^{k-1} C_l\Big)\eta^2(t) + \eta^2(k-1) \leq \sum_{t=0}^{k-1}(C_{k-1})^{k-t-1}\eta^2(t). \tag{33}$$

此外，同样根据假设3，$\eta(k)$递减，且$\lim_{k\to\infty}\eta(k) = 0$。因此，对于任意给定的 $\varepsilon > 0$，存在一个时间步$k_0 > 0$，使得 $\eta(k) \leq \varepsilon, \forall k \geq k_0$。由于$\eta(k)$递减，我们有

$$\eta(k) < \eta(k-1) = \frac{1-C_{k-1}}{\mu}. \tag{34}$$

将式 (34) 两边同乘 $\eta(k)$，可得对所有$k \geq k_0$

$$\eta^2(k) < \frac{(1-C_{k-1})\varepsilon}{\mu}. \tag{35}$$

对于 $k > k_0 + 1$，(33) 右侧可有界为

$$\sum_{t=0}^{k-1}(C_{k-1})^{k-t-1}\eta^2(t) \leq \sum_{t=0}^{k_0}(C_{k-1})^{k-t-1}\eta^2(t) + \sum_{t=k_0+1}^{k-1}(C_{k-1})^{k-t-1}\eta^2(t). \tag{36}$$

$\eta(k)$ 的递减性意味着，对于 $k \geq 0$，$\eta^2(0) \geq \eta^2(k)$。由 (14) 可知，对于所有 $k \geq 1$，$C_{k-1} \leq 1$。因此，(36) 右侧的第一项可写为

$$\begin{aligned}\sum_{t=0}^{k_0}\left(C_{k-1}\right)^{k-t-1}\eta^2(t) &\leq \eta^2(0)\sum_{t=0}^{k_0}\left(C_{k-1}\right)^{k-t-1}\\ &= \eta^2(0)\sum_{t=0}^{k_0}\left(C_{k-1}\right)^{k-k_0+t-1}\\ &= \eta^2(0)\left(C_{k-1}\right)^{-k_0}\left(C_{k-1}\right)^{k-1}\sum_{t=0}^{k_0}\left(C_{k-1}\right)^{t}\\ &\leq \eta^2(0)\left(C_{k-1}\right)^{-k_0}\left(C_{k-1}\right)^{k-1}(k_0+1),\end{aligned} \tag{37}$$

其中，$\lim_{k\to\infty}(C_{k-1})^{k-1} = 0$ 由引理3成立。因此，我们有

$$\lim_{k\to\infty}\sum_{t=0}^{k_0}\left(C_{k-1}\right)^{k-t-1}\eta^2(t) = 0. \tag{38}$$

此外，对于 $k > k_0+1$，式 (35) 也成立，这使我们能够求得式 (36) 右端第二项的一个上界，即

$$\begin{aligned}\sum_{t=k_0+1}^{k-1}\left(C_{k-1}\right)^{k-t-1}\eta^2(k) &\leq \frac{(1-C_{k-1})\varepsilon}{\mu}\sum_{t=k_0+1}^{k-1}\left(C_{k-1}\right)^{k-t-1}\\ &= \frac{(1-C_{k-1})\varepsilon}{\mu}\left(\frac{1-\left(C_{k-1}\right)^{k-k_0-1}}{1-C_{k-1}}\right)\\ &\leq \frac{\varepsilon}{\mu}.\end{aligned} \tag{39}$$

利用（37）–（39），并对（36）两边取极限，得到

$$\lim_{k\to\infty}\sum_{t=0}^{k-1}\left(C_{k-1}\right)^{k-t-1}\eta(t) \leq \frac{\varepsilon}{\mu}. \tag{40}$$

由于 ε 是任意的，我们完成了证明。

参考文献

[1] Y. LeCun、Y. Bengio 和 G. Hinton，《深度学习》，《自然》，第521卷，第7553期，第436–444页，2015年。[2] I. Goodfellow、Y. Bengio 和 A. Courville，《深度学习》。MIT出版社，2016年。[3] A. Nedic，《网络中凸机器学习问题的分布式梯度方法：分布式优化》，《*IEEE*信号处理杂志》，第37卷，第3期，第92–101页，2020年。[4] J. Chen 和 X. Ran，《边缘计算中的深度学习：综述》，*IEEE* 会刊，第107卷，第8期，第1655–1674页，2019年。[5] B. McMahan、E. Moore、D. Ramage、S. Hampson 和 B. A. y Arcas，《基于去中心化数据的深度网络通信高效学习》，载于《人工智能与统计学》。PMLR，2017年，第1273–1282页。[6] V. Smith、C.-K. Chiang、M. Sanjabi 和 A. S. Talwalkar，《联邦多任务学习》，《神经信息处理系统进展》，第30卷，2017年。[7] Q. Yang、Y. Liu、T. Chen 和 Y. Tong，《联邦机器学习：概念与应用》，《*ACM*智能系统与技术汇刊（*TIST*）》，第10卷，第2期，第1–19页，2019年。

[8] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, B. McMahan 等，《迈向大规模联邦学习：系统设计》，《机器学习与系统会议录》，第 1 卷，第 374–388 页，2019 年。[9] H. B. McMahan, D. Ramage, K. Talwar 和 L. Zhang，《学习差分隐私循环语言模型》，arXiv 预印本 arXiv:1710.06963，2017 年。[10] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings 等，《联邦学习的进展与开放问题》，《机器学习基础与趋势》，第 14 卷，第 1–2 期，第 1–210 页，2021 年。[11] 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。[12] C. Dwork, F. McSherry, K. Nissim 和 A. Smith，《在隐私数据分析中校准噪声与敏感度》，载于《密码学理论会议》。施普林格，2006 年，第 265–284 页。[13] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal 和 K. Seth，《面向隐私保护机器学习的实用安全聚合》，载于《2017年ACM SIGSAC计算机与通信安全会议论文集》，2017 年，第 1175–1191 页。[14] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh 和 D. Bacon，《联邦学习：提升通信效率的策略》，arXiv 预印本 arXiv:1610.05492，2016 年。[15] D. Alistarh, D. Grubic, J. Li, R. Tomioka 和 M. Vojnovic，《QSGD：通过梯度量化与编码实现高通信效率的随机梯度下降》，《神经信息处理系统进展》，第 30 卷，2017 年。[16] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat 和 C. Renggli，《稀疏化梯度方法的收敛性》，《神经信息处理系统进展》，第 31 卷，2018 年。[17] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie 和 R. Pedarsani，《FedPAQ：一种采用周期平均和量化的高通信效率联邦学习方法》，载于《国际人工智能与统计学会议》。PMLR，2020 年，第 2021–2031 页。[18] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar 和 S. Cui，《高通信效率联邦学习》，《美国国家科学院院刊》，第 118 卷，第 17 期，2021 年。[19] A. Mitra, R. Jaafar, G. Pappas 和 H. Hassani，《联邦学习中的线性收敛：应对客户端异质性与稀疏梯度》，《神经信息处理系统进展》，第 34 卷，2021 年。[20] F. Molinari, N. Agrawal, S. Sta´nczak 和 J. Raisch，《衰落无线信道上的最大一致性》，《IEEE网络系统控制汇刊》，第 8 卷，第 2 期，第 791–802 页，2021 年。[21] M. Frey, I. Bjelakovi´c 和 S. Sta´nczak，《相关信道中的空中计算》，《IEEE信号处理汇刊》，第 69 卷，第 5739–5755 页，2021 年。[22] G. Zhu, J. Xu, K. Huang 和 S. Cui，《面向大规模物联网无线数据聚合的空中计算》，《IEEE无线通信》，第 28 卷，第 4 期，第 57–65 页，2021 年。[23] K. Yang, T. Jiang, Y. Shi 和 Z. Ding，《通过空中计算实现联邦学习》，《IEEE无线通信汇刊》，第 19 卷，第 3 期，第 2022–2035 页，2020 年。[24] M. M. Amiri 和 D. Gündüz，《无线边缘的机器学习：基于空中计算的分布式随机梯度下降》，《IEEE信号处理汇刊》，第 68 卷，第 2155–2169 页，2020 年。[25] T. Sery, N. Shlezinger, K. Cohen 和 Y. C. Eldar，《基于异构数据的空中联邦学习》，《IEEE信号处理汇刊》，第 69 卷，第 3796–3811 页，2021 年。[26] R. Ahlswede，《多向通信信道》，载于《第二届信息论国际研讨会：察赫卡佐尔，亚美尼亚，苏联，1971年9月2日至8日》，1973 年。[27] A. Giridhar 和 P. Kumar，《迈向无线传感器网络中网内计算理论》，《IEEE通信杂志》，第 44 卷，第 4 期，第 98–107 页，2006 年 4 月。[28] F. Molinari, N. Agrawal, S. Sta´nczak 和 J. Raisch，《采用半双工通信技术的集群网络中的空中最大一致性》，《IEEE网络系统控制汇刊》，2022 年。[29] W. Rudin 等，《数学分析原理》。麦格劳-希尔，纽约，1976 年，第 3 卷。[30] K. Yuan, Q. Ling 和 W. Yin，《论去中心化梯度下降的收敛性》，《SIAM优化杂志》，第 26 卷，第 3 期，第 1835–1854 页，2016 年。

[31] C. Xi, V. S. Mai, R. Xin, E. H. Abed 和 U. A. Khan，《行随机矩阵下有向图优化的线性收敛》，*IEEE* 自动控制汇刊，第 63 卷，第 10 期，第 3558–3565 页，2018 年。[32] R. Xin, S. Pu, A. Nedi´c 和 U. A. Khan，《基于一阶方法的去中心化优化通用框架》，*IEEE* 会刊，第 108 卷，第 11 期，第 1869–1889 页，2020 年。[33] V. S. Mai 和 E. H. Abed，《使用行随机矩阵在加权有向图上进行分布式优化》，载于 *2016*年美国控制会议 *(ACC)*。IEEE，2016 年，第 7165–7170 页。[34] H. Li, Q. Lü 和 T. Huang，《时变一般非平衡有向图上的分布式投影次梯度算法》，*IEEE* 自动控制汇刊，第 64 卷，第 3 期，第 1309–1316 页，2018 年。[35] Y. Nesterov，《凸优化导论：基础课程》。施普林格科学与商业媒体，2003 年，第 87 卷。[36] A. Nedi´c 和 A. Olshevsky，《时变有向图上的分布式优化》，*IEEE* 自动控制汇刊，第 60 卷，第 3 期，第 601–615 页，2014 年。[37] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin 和 K. H. Johansson，《分布式优化综述》，《控制年度评论》，第 47 卷，第 278–305 页，2019 年。[38] P. Wang、P. Lin、W. Ren 和 Y. Song，“采用更一般步长的分布式次梯度多智能体优化”，*IEEE* 自动控制汇刊，第 63 卷，第 7 期，第 2295–2302 页，2017 年。[39] P. Xie、K. You、R. Tempo、S. Song 和 C. Wu，“时变非平衡有向图上的不等式约束分布式凸优化”，*IEEE* 自动控制汇刊，第 63 卷，第 12 期，第 4331–4337 页，2018 年。[40] V. S. Mai 和 E. H. Abed，“具有行随机性和约束正则性的有向图上的分布式优化”，《自动化》，第 102 卷，第 94–104 页，2019 年。[41] D. Tse 和 P. Viswanath，《无线通信基础》。剑桥大学出版社，2012 年。[42] A. F. Molisch，《无线通信》。约翰·威立父子出版公司，2012 年。[43] B. Sklar，“移动数字通信系统中的瑞利衰落信道。I. 特性”，《*IEEE*通信杂志》，第 35 卷，第 7 期，第 90–100 页，1997 年。[44] D. Bertsekas、A. Nedic 和 A. Ozdaglar，《凸分析与优化》。Athena Scientific，2003 年，第 1 卷。