跨孤岛联邦学习的自适应激励：一种多智能体强化学习方法

Shijing Yuan[1] *Hongze Liu*[1] *Hongtao Lv*[2] *Zhanbo Feng*[1] *Jie Li*[1⋆] *Hongyang Chen*[3] *Chentao Wu*[1]

[1]中国上海交通大学人工智能研究院、教育部人工智能重点实验室、计算机科学与工程系[2]中国济南山东大学软件学院[3]中国之江实验室图计算研究中心

arXiv:2302.07493v1 [cs.LG] 15 Feb 2023

摘要

跨孤岛联邦学习（FL）是一种典型的联邦学习，使组织（例如金融或医疗实体）能够基于彼此隔离的数据训练全局模型。合理的激励是鼓励组织贡献数据的关键。然而，现有关于跨机构联邦学习激励的研究未考虑环境动态性（例如训练过程中所训练全局模型的精度，以及数据归属不确定客户端的数据）。此外，其中大多数研究假设组织会共享私人信息，这并不现实。为克服这些局限，我们提出了一种用于跨机构联邦学习的新型自适应机制，旨在动态训练环境中激励组织贡献数据，从而最大化其长期收益。该机制基于多智能体强化学习，无需组织提供私人信息，即可从势博弈历史中学习近乎最优的数据贡献策略。实验表明，我们的机制能够实现自适应激励，并有效提高组织的长期收益。

关键词— 跨机构联邦学习，自适应激励，多智能体强化学习，势博弈

1. 引言

联邦学习（FL）是一种分布式学习范式，使客户端能够基于相互隔离的数据训练全局模型。跨机构联邦学习是一类典型的联邦学习，其中客户端由不同组织（如金融公司和制药公司）构成 [1]。在跨机构联邦学习中，组织贡献数据以训练全局模型（例如药物发现和疫苗开发等长期跨孤岛联邦学习过程），其中组织拥有的本地数据可能会动态变化 [2]。然而，组织可能不愿参与训练。这是因为组织的商业份额可能会受到其潜在竞争对手的损害，而这些竞争对手同样能够从全局模型中获利 [3]。因此，合理的激励机制对于鼓励客户端为训练贡献数据资源至关重要 [4]。

关于跨机构联邦学习激励机制的相关研究，重点关注客户端贡献评估 [5, 6], 公平性 [7, 8], 个性化激励 [9], 隐私开销 [10, 11], 以及社会福利最大化 [2, 12, 13]。然而，上述研究要么忽略了组织贡献的数据资源可能动态变化这一事实，要么依赖于训练所得全局模型的精度与数据贡献量之间的函数形式（精度函数）。然而，在实践中，精度函数未知且会随时间变化 [14]。此外，上述研究假设参与训练的组织会与其他组织共享其私人信息，例如盈利能力和训练开销，这在真实的跨机构联邦学习场景中并不现实。

为克服上述局限性，我们为跨机构联邦学习设计了一种新颖的基于多智能体强化学习（MARL）的自适应机制，以在真实的动态训练环境中实现自适应激励。具体而言，我们提出了一种基于收益再分配的自适应机制，以鼓励组织自适应地贡献数据。重要的是，组织训练收益的计算仅需要精度函数的值，而不需要其具体形式。在所提出的机制下，我们证明了组织为最大化其个人收益而进行的交互构成一个加权势博弈。为了确定使长期收益最大化的数据贡献策略，我们将组织之间的交互建模为一个多智能体部分可观测马尔可夫决策过程（POMDP）：各组织在不了解其他组织私人信息的情况下贡献数据，以最大化个人收益。随后，我们提出了一种基于MARL的分布式方法，该方法结合了策略梯度和微分神经计算机，称为MPGD。该方法无需任何其他组织的私人信息，即可从历史势博弈中学习在动态训练环境下近乎最优的数据贡献策略。

我们的主要贡献总结如下。

- 据我们所知，我们是首个针对真实的动态训练环境设计跨机构联邦学习自适应激励机制的研究者。具体而言，我们的机制有利于实际应用，因为：a）它不需要任何组织的私人信息；b）它不依赖于训练所得全局模型精度与数据贡献之间的任何特定函数形式。

- 我们提出了一种基于 MARL 的去中心化算法 MPGD，用于激励跨机构联邦学习，该算法鼓励各组织在真实且动态的训练环境中分布式地决定其数据贡献。

- 基于近真实平台和真实数据集开展的实验表明，我们的机制能够在真实的动态训练环境中实现自适应激励，并有效提高组织的长期收益。

*通讯作者。

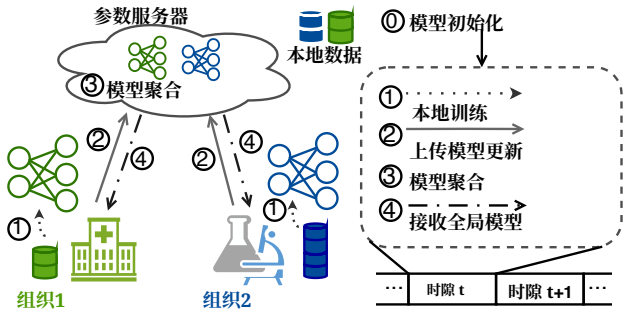

图 1。长期跨孤岛联邦学习的流程，其中组织动态贡献数据，以最大化各自的收益。

2. 激励机制

本节提出一种基于收益再分配的自适应激励机制，并证明组织之间为实现收益最大化而进行的交互构成一个加权势博弈。

如图 1 所示，我们考虑一个由 $\mathcal{O} = \{o_n\}_{n\in\mathcal{N}}$ 个组织组成的长期跨孤岛联邦学习场景，其中 $\mathcal{N} = \{1,\dots,N\}$。在参数服务器的协助下，每个组织 $o_n \in \mathcal{O}$ 都贡献其本地数据，以参与长期训练过程，并期望获得训练后的全局模型。训练过程被离散为时隙 $\mathcal{I} = \{0,1,\dots,T-1\}$。在每个时隙 t，$o_n \in \mathcal{O}$ 贡献 $x_n^{[t]}$ 数据以进行本地训练，其中 $x_n^{[t]} \in [0, |\mathcal{D}_n^{[t]}|]$，$\mathcal{D}_n^{[t]}$ 和 $|\mathcal{D}_n^{[t]}|$ 分别是 o_n 的本地数据集及其大小（数据样本的数量）。本地训练完成后，o_n 将更新后的模型上传至参数服务器进行模型聚合。

为了刻画参与训练的组织的收益，我们量化其在每个时隙中从全局模型获得的收入以及训练开销。

1) 全局模型带来的收入。令 $d_n = \frac{x_n}{|\mathcal{D}_n|}$ 为 o_n 的数据贡献策略，并将 $\boldsymbol{d}_{-n}^{[t]} = \{d_j^{[t]}\}_{j\in\mathcal{N},j\neq n}$ 表示为该时隙内除 o_n 之外所有组织的数据贡献策略。第 t^{th} 个时隙结束后训练得到的全局模型精度可表示为 $P^{[t]}(d_n^{[t]}, \boldsymbol{d}_{-n}^{[t]})$。注意，$P^{[t]}(d_n^{[t]}, \boldsymbol{d}_{-n}^{[t]})$[1] 随时隙 t 变化，并且所有组织都可以在第 t^{th} 个时隙结束时观测到。令 p_n 为 o_n 从全局模型单位性能中获得的利润，因此 o_n 从全局模型获得的收入可表示为 $p_n P^{[t]}(d_n^{[t]}, \boldsymbol{d}_{-n}^{[t]})$ [2,13]。

2) 训练开销。在 t^{th} 时隙中，o_n 的训练开销由能耗 $E_n^{[t]}$ 和通信开销 $C_n^{[t]}$ 组成，其中 $E_n^{[t]} = v_n d_n^{[t]} |\mathcal{D}_n|$，$v_n$ 是 o_n [15,16] 每个训练样本的单位能耗。

接下来，我们提出一种基于收益再分配的自适应激励机制，以鼓励组织贡献其本地数据。其动机在于，组织之间潜在的竞争可能会阻碍它们贡献数据，因为组织的竞争者也能从全局模型中获利，从而可能损害其商业份额 [3,17]。所提出的收益再分配使数据贡献较少的组织能够补偿数据贡献较多的组织。因此，o_n 获得的收益再分配为 $r_n^{[t]}(d_n^{[t]}, \boldsymbol{d}_{-n}^{[t]}) = \sum_{j\in\mathcal{N}} \alpha^{[t]}(d_n^{[t]} - d_j^{[t]})$，其中 $\alpha^{[t]}$ 表示收益再分配的强度。显然，$r_n^{[t]}$ 会随数据贡献 $\{d_n^{[t]}, \boldsymbol{d}_{-n}^{[t]}\}$ 自适应变化。

3) 参与训练的收益。最后，o_n 在 t^{th} 时隙的收益可表述为

$$u_n^{[t]}(d_n^{[t]}, \boldsymbol{d}_{-n}^{[t]}) = p_n P^{[t]}(d_n^{[t]}, \boldsymbol{d}_{-n}^{[t]}) - v_n d_n^{[t]} |\mathcal{D}_n| - C_n^{[t]} + r_n^{[t]}. \quad (1)$$

因此，方程 (1) 只需要 P 的值 $^{[t]}$ 而无需知道其具体形式。

基于上述机制，我们应用博弈论 [18] 来分析每个组织的最优数据贡献策略，因为博弈论是分析多个客户端之间交互的有力方法。具体而言，追求收益最大化的各组织之间的交互可以表述为一个非合作博弈 $\mathcal{G} = \{\mathcal{N}, \{d_n^{[t]}\}_{n\in\mathcal{N}}, \{u_n^{[t]}\}_{n\in\mathcal{N}}\}$，其中每个组织 o_n 在给定 $\boldsymbol{d}_{-n}^{[t]}$ 的情况下做出决策 $d_n^{[t]}$ 以最大化其收益

$$\max_{d_n^{[t]}} \quad u_n^{[t]}(d_n^{[t]}, \boldsymbol{d}_{-n}^{[t]}) \quad \text{s.t. } d_n^{[t]} \in [0,1]. \quad (2)$$

因此，为实现 $\mathcal{G}$ 的均衡 [18]，[19]，其概念定义如下。**定义1（纳什均衡）**。$\mathcal{G}$ 的纳什均衡（*NE*）是一个策略组合 $\boldsymbol{d}^{[t]*} = \{d_n^{[t]*}\}_{n\in\mathcal{N}}$，其中没有任何 o_n 能够通过单方面改变其策略来进一步增加自身收益

$$u_n(d_n^{[t]*}, \boldsymbol{d}_{-n}^{[t]*}) \geq u_n(d_n^{[t]*}, \boldsymbol{d}_{-n}^{[t]*}), \quad \forall n \in \mathcal{N}. \quad (3)$$

为证明纳什均衡的存在性，我们引入加权势博弈的定义如下。

定义 2（加权势博弈）。博弈 $\mathcal{G}$ 是加权势博弈 *[18]*，若存在势函数 $\boldsymbol{U}(d_n^{[t]}, \boldsymbol{d}_{-n}^{[t]})$，使得每个 o_n 的策略偏离所导致的收益变化等于势函数的变化乘以一个非负权重因子 w_n，即：

$$w_n \left[\boldsymbol{U}(\boldsymbol{d}^{[t]\prime}) - \boldsymbol{U}(\boldsymbol{d}^{[t]})\right] = u_n(\boldsymbol{d}^{[t]\prime}) - u_n(\boldsymbol{d}^{[t]}), \quad (4)$$

其中，$\boldsymbol{d}^{[t]} = \{d_n^{[t]}, \boldsymbol{d}_{-n}^{[t]}\}$ 和 $\boldsymbol{d}^{[t]\prime} = \{d_n^{[t]\prime}, \boldsymbol{d}_{-n}^{[t]}\}$。注意，每个势博弈都必然存在纳什均衡。

定理 1（$\mathcal{G}$ 是加权势博弈）。$\mathcal{G}$ 是一个加权势博弈，其势函数为式 *5* 中给出的势函数。 g p f g q $( )$

$$\boldsymbol{U}(\boldsymbol{d}^{[t]}) = \boldsymbol{P}\left(\boldsymbol{d}^{[t]}\right) - \sum_{n\in\mathcal{N}} \left[\frac{v_n d_n^{[t]} |\mathcal{D}_n| + C_n^{[t]}}{p_n} - \sum_{j\in -\boldsymbol{n}} \frac{r_{n,j}^{[t]}}{p_n}\right]. \quad (5)$$

证明。为简洁起见，这里省略上标 $[t]$。令任意 o_n 的决策从 d_n 变为 d_n'，则有

$$\boldsymbol{U}(d_n', \boldsymbol{d}_{-n}) = \boldsymbol{P}\left(d_n', \boldsymbol{d}_{-n}\right) - \sum_{j\in\mathcal{N},j\neq n} \frac{v_j d_j |\mathcal{D}_n| + C_j}{p_j}$$

$$- \frac{v_n d_n' |\mathcal{D}_n| + C_n}{p_n} + \sum_{j\in\mathcal{N}} \frac{r_{n,j}(d_n', d_j)}{p_n} + \sum_{\substack{j\in\mathcal{N},\\ j\neq i}} \sum_{k\in\mathcal{N}} \frac{r_{j,k}(d_j, d_k)}{p_n}.$$

Subtracting Eq. (6) from Eq. (5), we obtain (6)

$$p_n \left[\boldsymbol{U}(d_n, \boldsymbol{d}_{-n}) - \boldsymbol{U}(d_n', \boldsymbol{d}_{-n})\right] = u_n(d_n, \boldsymbol{d}_{-n}) - u_n(d_n', \boldsymbol{d}_{-n}), \quad (7)$$

hence $\mathcal{G}$ is a weighted potential game that the NE exists. □

Although $\mathcal{G}$ is proved to be a weighted potential game, the NE of

[1]本文将训练后的全局模型 $P^{[t]}(d_n^{[t]}, \boldsymbol{d}_{-n}^{[t]})$ 的精度称为精度函数。根据 [2, 12, 13]，$P^{[t]}(d_n^{[t]}, \boldsymbol{d}_{-n}^{[t]})$ 是具有强凸损失函数的凸函数。也就是说，给定 $\boldsymbol{d}_{-n}$，随着 o_n 投入的数据量增加，$P(d_n, \boldsymbol{d}_{-n})$ 会提高，但增长率会降低。因此，我们假设 $P(d_n, \boldsymbol{d}_{-n})$ 满足 $\nabla_{d_n} P(d_n, \boldsymbol{d}_{-n}) \geq 0$。

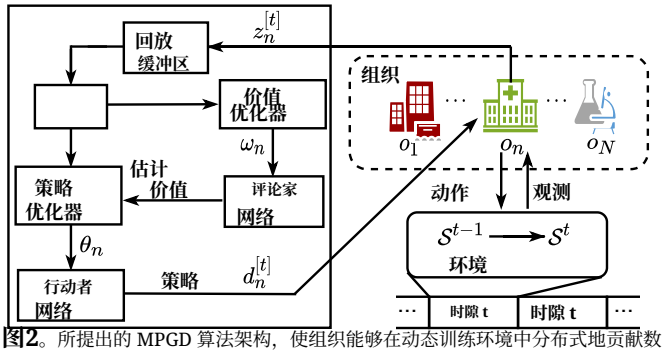

图2。所提出的 MPGD 算法架构，使组织能够在动态训练环境中分布式地贡献数据，而无需获取其他组织的私人信息。

$\mathcal{G}$ 无法通过传统方法（例如，基于动态响应的算法 [19]）实现，原因如下。*1*）在实际场景中，组织可能不知道其他组织的盈利能力等私人信息 p_n、通信开销 $C_n^{[t]}$和训练开销 $E_n^{[t]}$。*2*）训练得到的全局模型的精度 $P^{[t]}$（$d_n^{[t]}$，$\boldsymbol{d}_{-n}^{[t]}$）与数据贡献 $\{d_n^{[t]}, \boldsymbol{d}_{-n}^{[t]}\}$之间的函数关系是一个难以进行定量描述的"黑箱" *[13]*。基于上述观察，我们在第3节中将组织之间的长期交互建模为多智能体部分可观测马尔可夫决策过程（POMDP），并进一步提出一种基于多智能体强化学习的分布式算法，即 **MPGD**。

3. 基于多智能体强化学习的算法

3.1. 部分可观测马尔可夫决策过程

我们将组织贡献数据以最大化长期收益的过程表述为一个多智能体部分可观测马尔可夫决策过程，即 $\mathcal{M} =< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{Z}, \mathcal{R}, \mathcal{T} >$，其中$\mathcal{S} = \{\mathcal{S}_n = \{P^{[t]}, C_n^{[t]}, E_n^{[t]}, \alpha^{[t]}\}_{t\in\mathcal{I}}\}_{n\in\mathcal{N}}$表示状态空间，$\mathcal{A} = \{\mathcal{A}_n = \{d_n^{[t]}\}_{t\in\mathcal{I}}\}_{n\in\mathcal{N}}$表示动作空间，$\mathcal{P} = \{\mathcal{S}_n \times \mathcal{A} \times \mathcal{S}_n \to [0,1]\}_{n\in\mathcal{N}}$是状态转移概率函数集，$\mathcal{Z} = \{\mathcal{Z}_n = \{z_n^{[t]}\}_{t\in\mathcal{I}}\}_{n\in\mathcal{N}}$是观测空间，$z_n^{[t]} = [h_n^{[t-1]}, \ldots, h_n^{[t-H]}]$是 o_n在第 t^{th} 个时隙的观测，$h_n^{[k]} = \{\boldsymbol{d}_{-n}^{[k]}, C_n^{[k]}, \alpha^{[k]}, P^{[k]}\}$，$k\in[t-1, t-H]$，$\mathcal{R} = \{\mathcal{R}_n = \{u_n^{[t]}(d_n^{[t]}, \boldsymbol{d}_{-n}^{[t]})\}_{t\in\mathcal{I}}\}_{n\in\mathcal{N}}$是奖励空间，$\mathcal{T} = \{\{\mathcal{S}_n \times \mathcal{O}_n \to [0,1]\}_{t\in\mathcal{I}}\}_{n\in\mathcal{N}}$表示观测函数集。值得注意的是，$z_n^{[t]}$包含其他组织 $\boldsymbol{d}_{-n}^{[k]}$的数据贡献策略、$o_n$的通信开销 $C_n^{[k]}$、收益再分配强度 $\alpha^{[k]}$，以及此前 H 个时隙全局模型的精度 $P^{[k]}$。

3.2. 算法设计

接下来，我们提出一种基于多智能体强化学习的去中心化算法，称为**MPGD**，该算法结合策略梯度方法与微分神经计算机（DNC）[20]，以确定数据贡献策略 $\{\{d_n^{[t]}\}_{t\in\mathcal{I}}\}_{n\in\mathcal{N}}$，从而在动态环境中最大化组织的长期收益。具体而言，o_n由 θ_n 参数化的数据贡献策略为 $\pi_{\theta_n}: \mathcal{Z}_n \times \mathcal{A}_n \to [0,1]$，因此 o_n的优化目标为

$$\begin{aligned}\theta_n^* &= \arg\max_{\theta_n} J_n(\pi_{\theta_n}) \\ &= \arg\max_{\theta_n} \mathbb{E}\left[V^{\pi_{\theta_n}}(\boldsymbol{z}_n^0) \mid \rho_n^0\right] \\ &= \arg\max_{\theta_n} \mathbb{E}\left[Q^{\pi_{\theta_n}}(\boldsymbol{z}_n^0, d_n^0) \mid \rho_n^0, \pi_{\theta_n}\right],\end{aligned} \tag{8}$$

其中，$J_n(\pi_{\theta_n})$是目标函数，$V^{\pi_{\theta_n}}(\boldsymbol{z}_n) = \mathbb{E}\left[R_n^t \mid \boldsymbol{z}_n^t = \boldsymbol{z}_n, \Pi, \mathcal{P}, \mathcal{T}\right]$是观测的价值函数，$Q^{\pi_{\theta_n}}(\boldsymbol{z}_n, d_n) = \mathbb{E}\left[R_n^t \mid \boldsymbol{z}_n^t = \boldsymbol{z}_n, d_n^t = d_n, \Pi, \mathcal{P}, \mathcal{T}\right]$是观测和动作的价值函数，$\rho_n^0$是 z_n的初始观测概率分布，$\Pi = \{\pi_{\theta_n}\}_{n\in\mathcal{N}}$表示所有组织策略的集合，$R_n^t = \sum_{l=t}^{T} \gamma^{l-t} u_n^t$是折扣期望未来奖励，$\gamma \in [0,1]$是折扣因子。

如图2所示，**MPGD** 由以下部分组成：参数为 ω的评论家网络，其中每个观测都映射为一个特征向量；参数为 θ 的行动者网络，该网络根据观测输出动作，以近似策略；用于存储历史博弈记录的经验回放缓冲区；以及DNC模块。DNC模块是一种循环神经网络 [20]，有助于组织的策略更快地收敛至纳什均衡。为加速MPGD的收敛，我们对行动者网络和评论家网络进行如下优化。

3.2.1. 行动者网络的优化

我们采用策略梯度定理 [20] 来优化演员网络，该定理对策略梯度进行裁剪。

命题1。 *MPGD*的策略梯度为

$$\begin{aligned}\nabla_{\theta_n} J_n &= \mathbb{E}_{\pi_{\theta_n}, \rho_n^1(\boldsymbol{z}_n)}\left[\nabla_{\theta_n} \log \pi_{\theta_n}(\boldsymbol{z}_n, d_n) Q^{\pi_{\theta_n}}(\boldsymbol{z}_n, d_n)\right] \\ &= \mathbb{E}_{\pi_{\theta_n}, \rho_n^1(\boldsymbol{z}_n)}\left[\nabla_{\theta_n} \log \pi_{\theta_n}(\boldsymbol{z}_n, d_n) A^{\pi_{\theta_n}}(\boldsymbol{z}_n, d_n)\right] \\ &\approx \mathbb{E}_{\pi_{\widehat{\theta_n}}, \rho_n^1(\boldsymbol{z}_n)}\left[f_n \nabla_{\theta_n} \log \pi_{\theta_n}(\boldsymbol{z}_n, d_n) A^{\pi_{\theta_n}}(\boldsymbol{z}_n, d_n)\right],\end{aligned} \tag{9}$$

where $f_n = \frac{\pi_{\theta_n}(z_n|d_n)}{\pi_{\widehat{\theta_n}}(z_n|d_n)}$, $A^{\pi_{\theta_n}}(\boldsymbol{z}_n, d_n) = Q^{\pi_{\theta_n}}(\boldsymbol{z}_n, d_n) - V^{\pi_{\theta_n}}(\boldsymbol{z}_n)$, *$\hat{\theta}$ is the policy parameter for sampling, and* $\rho_n^1(z_n)$ *is the observation distribution.*

证明 根据定义，$V^\pi(s) = \sum_d \pi(s,d) Q^\pi(s,d)$，我们有

$$\nabla_\theta V^\pi(s) = \sum_{s\in\mathcal{S}} d^\pi(s) \sum_d \pi(s,d) \nabla_\theta \log \pi(s,d) Q^\pi(s,d), \tag{10}$$

其中 $d^\pi(s) = \sum_{t=0}^{\infty} \gamma^t P(s^t = s)$是折扣状态分布。因此，当我们对状态 $s \sim (1-\gamma) d^\pi(s)$和动作 $d \sim \pi(s, d)$取期望，可得

$$\begin{aligned}\nabla_\theta V^\pi(s) &= \frac{1}{1-\gamma} \sum_{s\in\mathcal{S}} d_\gamma^\pi \sum_d \pi(s,d) \nabla_\theta \log \pi(s,d) Q^\pi(s,d) \\ &= \frac{1}{1-\gamma} \mathbb{E}_{\pi(\theta)}\left[\nabla_\theta \log \pi(s,d) Q^\pi(s,d)\right],\end{aligned} \tag{11}$$

其中 $d_\gamma^\pi = (1-\gamma) d^\pi(s)$。由于反映隐状态 s 的观测 z 是从观测分布 $\rho^1(z)$中采样得到的，因此可得

$$\nabla_\theta J = \mathbb{E}\left[V^{\pi_\theta}(z) \mid \rho\right] = \mathbb{E}_{\pi_\theta, \rho^1(z)}\left[\nabla_\theta \log \pi_\theta(z,d) Q^{\pi_\theta}(z,d)\right], \tag{12}$$

这证实了式（9）。 □

此外，策略梯度以及相对于 θ_n 的估计梯度可计算如下

$$\begin{aligned}&\nabla_{\theta_n} J_n \approx \mathbb{E}_{\pi_{\widehat{\theta_n}}, \rho_n^1(\boldsymbol{z}_n)}\left[\nabla_{\theta_n} \log \pi_{\theta_n}(\boldsymbol{z}_n, d_n) C(\boldsymbol{z}_n, d_n)\right], \\ &\nabla_{\theta_n} \hat{J}_n = \frac{1}{D} \sum_{t=0}^{D-1} \nabla_{\theta_n} \log \pi_{\theta_n}(\boldsymbol{z}_n^t, d_n^t) C(\boldsymbol{z}_n^t, d_n^t), \text{ where} \\ &C(\boldsymbol{z}_n, d_n) = \min\left\{f_n A^{\pi_{\theta_n}}(\boldsymbol{z}_n, d_n), \eta(f_n) A^{\pi_{\theta_n}}(\boldsymbol{z}_n, d_n)\right\},\end{aligned} \tag{13, 14}$$

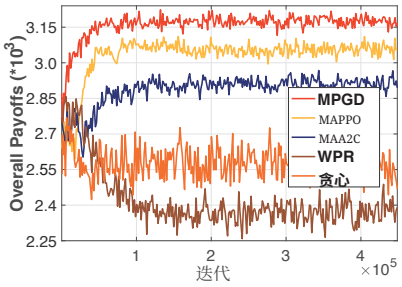

图 3。四种不同方案下的总体收益。

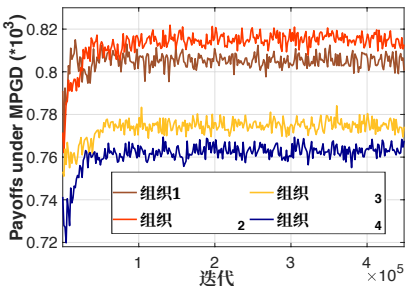

图 4。MPGD 下各组织的收益。

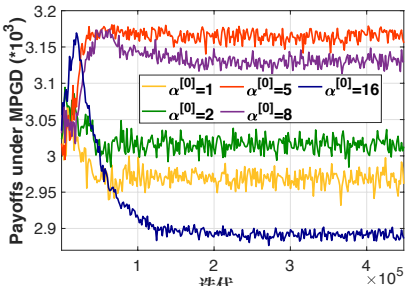

图 5。收益再分配强度对组织收益的影响。

$$\eta(d)=\begin{cases}1+\epsilon, & d>1+\epsilon\\ d, & 1-\epsilon\le d\le 1+\epsilon,\\ 1-\epsilon, & d<1-\epsilon\end{cases}\tag{15}$$

且 ϵ 是一个可调参数。演员网络通过随机梯度下降进行更新，即$\theta_n \leftarrow \theta_n + l_{a,n}\nabla_{\theta_n}\hat{J}_n$，其中$l_{a,n}$是演员网络更新的学习率，$D$ 是小批量大小。

3.2.2. 评论家网络的优化

评论家网络的目标函数为

$$F_n(\omega_n)=\mathbb{E}_{\boldsymbol{z}_n\sim\rho_n^1(\boldsymbol{z}_n)}\left[-V_{\omega_n}(\boldsymbol{z}_n)+\mathbb{E}_{\boldsymbol{z}'_n,d_n}\left[u+V_{\omega_n}(\boldsymbol{z}'_n)\right]\right]^2,\tag{16}$$

其中，$\boldsymbol{z}'_n$ 是 $\boldsymbol{z}_n$的下一时刻观测。同样，相对于 ω_n 的估计梯度计算如下：

$$\nabla_{\omega_n}\hat{F}_n=\frac{1}{D}\sum_{t=0}^{D-1}\left[V_{\omega_n}(\boldsymbol{z}_n^t)-Y_n^t\right]\frac{\mathrm{d}V_{\omega_n}(\boldsymbol{z}_n)}{\mathrm{d}\omega_n},\tag{17}$$

其中，$Y_n^t=R_n^t-\gamma^{D-t}R_n(D)+\gamma^{D-t}V_{\omega_n}(z_n(D))$，且 D 表示小批量的大小。评论家网络以学习率$l_{c,\ n}$进行更新，即$\omega_n \leftarrow \omega_n - l_{c,\ n}\nabla_{\omega_n}\hat{F}_n$。

4. 实验与讨论

实验设置。为评估组织的长期收益，我们在一个近乎真实的联邦学习平台上，使用 FedAvg 算法和 FMNIST 与 CIFAR- 数据集，对 $|\mathcal{N}|=4$ 个客户端开展实验。根据 [16, 24], , o_n 的盈利能力和通信开销分别设为 $p_n \sim N$(1000，10) 和 $C_n^{[t]} \sim N$(0。5，0。02)。数据样本数量设为 $|\mathcal{D}_n| \sim N$(2000，50)，每个单位样本的能耗设为 $v_n \sim N$(4，0。2)。为避免过高的训练开销，收益再分配强度 α 设为随 $P^{[t]}$ 的增益而衰减，因为其会随时隙 t 的增加而降低，即 $\alpha^{[t]}=\frac{P^{[t]}-P^{[t-1]}}{P^{[t-1]}-P^{[t-2]}}\alpha^{[0]}$，其中 $\alpha^{[0]}=5$。演员网络和评论家网络均包含两个隐藏全连接层，节点数分别为 210 和 50。

基线方法。我们将 MPGD 与最先进的方法进行比较，包括 **MAA2C** [20], , 这是一种多智能体版本的优势演员-评论家（A2C）算法，融合了 DNC 和 A2C；**MAPPO** [25],，一种用于多智能体训练的近端策略优化（PPO）算法；**贪心**，其中每个组织在每个时隙都选择能够获得最大奖励的数据贡献决策；以及**WPR** [16], , 组织在不采用所提出的激励机制的情况下参与训练，这意味着收益再分配项 $r_n^{[t]}$不包含在 o_n 的收益 $u_n^{[t]}$中。

图 3 展示了不同方案下的收敛情况和长期收益。可以看到，与其他基线方法相比，MPGD 实现了最高的长期收益和最快的收敛速度。与 WPR 不同，MPGD 下的收益逐步增加并迅速收敛。这一结果表明，所提出的基于收益再分配的激励机制能够有效鼓励组织提供数据参与训练，从而提升组织的长期效用。

图 4 展示了 MPGD 下各组织的收益。我们发现，由于数据贡献策略不断更新，所有组织的收益持续提高，并在 10^5 次迭代后收敛。这表明，所提出的机制无需获取任何组织的私人信息，也不需要对 $P^{[t]}$ 的具体形式作出假设，即可实现动态自适应激励$d_n^{[t]}$，$\boldsymbol{d}_{-n}^{[t]}$。

图 5 展示了 α 的初始值对组织总体收益的影响。我们可以得出一个有趣的结论：α 的初始值适当，有利于最大化组织的总体收益。如图 5 所示，将 $\alpha^{[0]}$ 从 1 增加到 16 并不总能提高组织的总体收益。原因在于，虽然较大的 $\alpha^{[0]}$ 能够鼓励组织贡献更多数据，但更多数据也意味着更高的训练开销。

5. 结论

在本研究中，我们为跨孤岛联邦学习设计了一种自适应激励机制 MPGD，能够在真实的动态训练环境中鼓励组织自适应地贡献数据。MPGD 易于在实践中应用，因为它不依赖任何组织的私人信息，也不假设精度函数具有任何特定的函数形式。实验结果表明，与典型基线方法相比，MPGD 能够有效提高长期收益。

6. 参考文献

[1] Qinbin Li、Zeyi Wen、Zhaomin Wu、Sixu Hu、Naibo Wang、Yuan Li、Xu Liu 和 Bingsheng He，《联邦学习系统综述：数据隐私与保护的愿景、炒作与现实》，*IEEE*计算机汇刊，第 1–1 页，2021，doi:10.1109/TKDE.2021.3124599。[2] Ning Zhang、Qian Ma 和 Xu Chen，《促进跨孤岛联邦学习中的长期合作：重复博弈视角》，*IEEE Trans. Mob Comput.*，第 1–1 页，2022，doi:10.1109/TMC.2022.3148263。[3] Yufeng Zhan、Peng Li、Song Guo 和 Zhihao Qu，《联邦学习激励机制设计：挑战与机遇》，*IEEE Netw.*，第 35 卷，第 4 期，第 310–317 页，2021。[4] Yufeng Zhan、Jie Zhang、Zicong Hong、Leijie Wu、Peng Li 和 Song Guo，《联邦学习激励机制设计综述》，*IEEE Trans.Emerg. Topics Comput.*，2021。

[5]Yihao Xue、Chaoyue Niu、Zhenzhe Zheng、Shaojie Tang、Chengfei Lv、Fan Wu 和 Guihai Chen，《探讨个体客户端对联邦学习的影响》，载于*AAAI*会议论文集，2021，第 10560–10567 页。[6] Hongtao Lv、Zhenzhe Zheng、Tie Luo、Fan Wu、Shaojie Tang、Lifeng Hua、Rongfei Jia 和 Chengfei Lv，《联邦学习中用户贡献的无数据评估》，载于*WiOPT*会议论文集，2021年10月，第 81–88 页。[7] Han Yu、Zelei Liu、Yang Liu、Tianjian Chen、Mingshu Cong、Xi Weng、Dusit Niyato 和 Qiang Yang，《一种公平性感知的联邦学习激励方案》，载于*AAAI/ACM AIES*会议论文集，2020，第 393–399 页。[8] Tiansheng Huang、Weiwei Lin、Wentai Wu、Ligang He、Keqin Li 和 Albert Y Zomaya，《一种具有公平性保障的联邦学习高效客户端选择方案》，*IEEE*并行与分布式系统汇刊，第 32 卷，1564，2020 7，第 1552 期。[9] Shuyu Kong、You Li 和 Hai Zhou，《联邦学习激励机制》，*arXiv*预印本 *arXiv:2205.10951*，2022。[10] Peng Sun、Haoxuan Che、Zhibo Wang、Yuwei Wang、Tao Wang、Liantao Wu 和 Huajie Shao，《Pain-fl：面向联邦学习的个性化隐私保护激励》，*IEEE*通信选定领域期刊，第 39 卷，第 12 期，第 3805–3820 页，2021年10月。[11] Ningning Ding、Zhixuan Fang 和 Jianwei Huang，《针对多维私人信息的高效联邦学习最优合同设计》，*IEEE*通信选定领域期刊，第 39 卷，第 1 期，第 186–200 页，2020年11月。[12] Jianan Chen、Qin Hu 和 Honglu Jiang，《跨孤岛联邦学习中的社会福利最大化》，载于*IEEE*国际声学、语音与信号处理会议论文集，2022，第 4258–4262 页。[13] Ming Tang 和 Vincent WS Wong，《跨孤岛联邦学习的激励机制：公共物品视角》，载于*IEEE*计算机通信会议论文集，2021，第 1–10 页。[14]Sai Praneeth Karimireddy、Satyen Kale、Mehryar Mohri、Sashank Reddi 和 Stich，《Scaffold：联邦学习的随机控制平均算法》，载于*ACM ICML*会议论文集，2020年7月，第 5132–5143 页。[15] Paolo Di Lorenzo、Claudio Battiloro、Mattia Merluzzi 和 Sergio Barbarossa，《无线网络边缘自适应联邦学习的动态资源优化》，载于*IEEE*国际声学、语音与信号处理会议论文集，2021，第 4910–4914 页。[16] Yufeng Zhan、Peng Li、Leijie Wu 和 Song Guo，《L4l：联邦学习中由经验驱动的计算资源控制》，*IEEE*计算机汇刊，第 71 卷，第 4 期，第 971–983 页，2021年6月。[17] Xiaohu Wu 和 Han Yu，《Mars-fl：促进竞争者在联邦学习中协作》，*IEEE*大数据汇刊，第 1–11 页，2022。[18] Yong Huat Chew、Boon-Hee Soong 等，《势博弈理论》，Springer，2016。[19] Brian Swenson、Ryan Murray 和 Soummya Kar，《势博弈中的最佳响应动态》，*SIAM*控制与优化期刊，第 56 卷，第 4 期，第 2734–2767 页，2018年5月。[20] Yufeng Zhan、Song Guo、Peng Li 和 Jiang Zhang，《基于深度强化学习的边缘计算卸载博弈》，*IEEE*计算机汇刊，第 69 卷，第 6 期，第 883–893 页，2020。

[21] Fan Lai、Yinwei Dai、Sanjay S. Singapuram、Jiachen Liu、Xi- angfeng Zhu、Harsha V. Madhyastha 和 Mosharaf Chowd- hury，《FedScale：大规模联邦学习的模型与系统性能基准测试》，载于 *Proc. ACM ICML*，2022 年。[22] Roland Vollgraf、Han Xiao、Kashif Rasul，《Fashion-MNIST 数据集》，https://github.com/zalandoresearch/fashion-mnist，访问日期：2022年1月。[23] Vinod Nair、Alex Krizhevsky 和 Geoffrey Hinton，《CIFAR-10 数据集》，https://www.cs.toronto.edu/kriz/cifar.html，访问日期：2022年1月。[24] Canh T Dinh、Nguyen H Tran、Minh NH Nguyen、Choong Seon Hong、Wei Bao、Albert Y Zomaya 和 Vincent Gramoli，《无线网络上的联邦学习：收敛分析与资源分配》，*IEEE/ACM Trans Netw.*，第 29 卷，第 1 期，第 398–409 页，2020 年。[25] Chao Yu、Akash Velu、Eugene Vinitsky、Yu Wang、Alexan- dre Bayen 和 Yi Wu，《PPO 在合作式多智能体游戏中的惊人有效性》，*arXiv preprint arXiv:2103.01955*，2021年。