基于强化学习的可持续联邦学习激励机制设计

艾秋媛　詹志坚　王　聪　宋　洁*

摘　要：随着数据在互联网、物联网和人工智能等技术中的广泛应用，数据共享成为促进经济和科技发展的关键引擎之一。然而，由于数据隐私和法律等多方面的顾虑，数据共享面临挑战。联邦学习作为一种新兴的机器学习范式，以保护数据隐私的同时促进多方协作而备受关注。本文关注跨孤岛的长期联邦学习合作，旨在解决数据所有者参与合作的成本和风险问题。本文首先建立了动态博弈模型，考虑了联邦客户端之间的互动策略；然后，提出了一种基于强化学习的激励机制，通过中央计划者为不同训练期设定激励，有效地促进客户端的参与。实验证明，该激励方案在提高系统总收益和控制激励成本方面具有显著效果。本文为可持续联邦学习提供了一种有效的激励设计，有望推动数据共享和合作模型在不同领域的应用。

关键词：数据共享；联邦学习；激励机制；稳定合作
中图分类号：TP18
JEL 分类号：C45；C72

一、引　言

随着互联网、物联网和人工智能等技术的快速发展，数据已经成为推动经济和科技发展的关键引擎之一。通过数据分析，企业可以深入了解客户行为、市场趋势和运营绩效，从而做出更明智的决策，提高整体绩效。然而，数据不足是数据分析中的一个重大挑战，以市场营销为例，训练一个精准触达用户的模型可能需要来自运营商、社交网站、购物网站等多方的数据，而一个机构内的数据可能无法完全满足需求。数据流通是解决这一问题的关键，通过数据流通，各领域、各行业的信息得以互相连接和交换，跨界合作和产业融合不断涌现，从而为经济增长带来新的活力。然而，数据共享面临数据隐私相关的伦理、法律等多方面顾虑，全面而多样化的数据集始终相对稀缺，亟待有效的数据共享机制。

在此背景下，联邦学习（Federated Learning，简称 FL）作为一种新的机器学习

* 艾秋媛，北京大学工学院，E-mail: aiqiuyuan@stu. pku. edu. cn；詹志坚，北京大学前沿交叉学科研究院，E-mail: zhanzhijian@stu. pku. edu. cn；王聪（通信作者），北京大学光华管理学院，E-mail: wangcong@gsm. pku. edu. cn；宋洁，北京大学工学院，E-mail: songjie@coe. pku. edu. cn。作者感谢国家自然科学基金重点项目（72131001）、国家自然科学基金青年项目（72101007）、国家自然科学基金专项项目（72241420）对本文研究的资助。作者感谢匿名审稿人和编辑部的宝贵意见，当然文责自负。

范式开始崭露头角（Yang et al., 2020）。联邦学习的参与者能够在保护数据隐私的前提下训练本地模型，并与服务器或其他对等点交换模型参数，从而充分发挥数据融通价值。一方面，这项技术可以克服传统方法的局限性，将机器学习过程与数据所有者的数据获取、存储、训练过程分开，为数据的安全与隐私保护提供有力支持。另一方面，它为众多参与者提供了一个协作训练全局模型的机会，帮助参与者通过模型参数交换方式完成数据共享，这也能成为间接构成数据市场的方式。这种“数据可用但不可见”的方法已成为数据密集型领域的一种流行方法，如智能制造（Han et al., 2019）、医疗保健（Brisimi et al., 2018；Huang et al., 2019）和自动驾驶汽车（Saputra et al., 2019；Wang et al., 2021）。通过联邦学习，可以创建统一的多源数据协作生态系统，促进数据共享、合作建模，进而可以有效地促进管理决策的效率及效果。

很多联邦学习任务并非一蹴而就，而是需要多方的持续合作才能实现可持续发展。例如，在多个医疗机构共同建立联邦学习模型来检测和管理慢性病时，需要持续积累临床数据，学习病例变化，才能提高模型的鲁棒性和预测能力，反映最新的医学知识和实践。然而，联邦学习领域初期研究主要关注技术方法设计（Bao et al., 2023；Yu et al., 2023），而将各方对数据资源的贡献作为理所当然的前提条件。这种假设很容易受到现实场景的挑战。例如，参与联邦学习通常需要消耗数据所有者的计算和通信资源等，这些方面的成本阻碍了客户端对联邦做出贡献。如果采用完全开放共享的模式建立联邦，理性的客户端更倾向于“搭便车”，即不贡献数据而仍享受他人数据带来的好处，这可能导致合作的崩溃（Bolton and Dewatripont, 2004）。此外，即使客户端出于利他主义愿意贡献数据，但对其他人可能“搭便车”的担忧也将减少它们贡献的意愿（Choi and Robertson, 2019）。

面对以上挑战，许多研究者开始关注联邦学习激励机制设计工作。Tu et al.（2022）从经济学和博弈论的角度综述了目前的联邦学习激励机制设计工作，他们将现有研究总结为四类：基于博弈的方案、基于拍卖的方案、基于合同的方案和基于匹配理论的方案。Zeng et al.（2022）概述了联邦学习激励机制的问题框架，对 Shapley 值、Stackelberg 博弈、拍卖、合同和强化学习等现有技术框架进行了分类讨论。另外，他们提出了多维博弈理论模型，用于研究参与者的经济行为，并展示了模型在跨孤岛联邦学习场景中的适用性。Gupta and Gupta（2023）对 2019—2022 年的文献进行分析，概述了基于博弈论的联邦学习激励模型在利润最大化、身份验证、隐私管理、信任管理和威胁检测方面的最新研究。以上综述对于联邦学习激励机制设计方面的工作进行了清晰、全面的分类、总结，并提供了对未来研究的见解。现有工作更多地聚焦于跨设备（Cross-Device）的联邦学习中（Zeng et al., 2022），且考虑单次合作关系的居多。随着数据市场的逐渐活跃，越来越多的组织和企业希望通过跨孤岛（Cross-Silo）联邦学习来共享数据，提升模型质量。在跨孤岛联邦学习

中，由于参与者本身也有一定的数据积累，其策略性选择相较于跨设备联邦学习中参与方的选择更为复杂多样，即参与方既可以选择参与公共训练，也可以仅通过本地训练来提升自己的模型效用。此外，当合作从单期扩展为长期时，时间一致性问题可能会导致可信的不对称均衡不成立，参与者可能有动机延迟其贡献或选择“搭便车”（Kessing，2007）。本文关注跨孤岛的长期（Long-Term）联邦学习过程，建立了动态博弈模型来刻画联邦客户端之间的互动策略，并提出一种基于强化学习的激励机制以促进理性参与者的贡献，以期提升联邦学习系统的总体收益。

在本文中，我们建立了动态博弈模型来刻画联邦客户端之间的互动策略。首先，我们将跨孤岛的长期联邦学习合作过程分为若干个模型训练期，联邦客户端在每个训练期都能收集到新数据，并利用它们不断提升机器学习的模型精度。其次，客户端在每一个合作期都拥有两种可能的策略选择，既可以参与公共联邦训练，也可以自留数据本地训练。在每期合作完成时，客户端都能基于自己的模型精度获得相应的收益。由于不参与公共训练的客户端无法得到当期更新的公共模型参数，客户端在每个训练期开始时都面临着参与成本与潜在收益之间的权衡。由于模型中所蕴含的信息量随着客户端的投入不断积累，客户端在整个长期联邦学习合作过程中也面临着在各期之间如何分配资源的跨期决策问题。

基于以上背景假设，我们建立了博弈树来考虑客户端之间的博弈问题，客户端在每个训练期的决策都建立在对已发生的合作历史的完全知悉以及对未来行动的理性预期的基础上。通过逆向递推方法，我们可解得客户端的均衡策略，它呈现间隔投入的特征。显然，合作的延迟并不利于模型精度的快速提升，进而会降低客户端的期望收益。基于此，我们设计了基于强化学习的动态激励方案，根据客户端的合作进程为不同训练期设定激励。首先，我们将联邦学习的组织方视为中央计划者，负责在合作的每个训练期开始前发布当期激励，以促进联邦客户端的投入。深度强化学习（DRL）智能体辅助中央计划者决策激励方案，而联邦客户端则成为与智能体交互的环境。一方面，我们仔细设计了 DRL 方法的状态、动作和奖励，使其充分包含联邦学习合作过程的信息。另一方面，我们引入双深度 Q 网络（Double DQN，简称 DDQN）、优先级回放（Prioritized Replay）、噪声网络（Noisy Net）来改进传统的深度 Q 网络（Deep Q-Network，简称 DQN）方法的效用。通过充分的实验，我们验证了该方案在提升系统总收益和控制激励成本方面的有效性。合理的激励成本惩罚能够使 DRL 智能体收敛到最具性价比的激励方案，这种激励方案能够在客户端贡献意愿不高的合作期精准激励，相同预算下的系统收益远高于固定激励的方案。

余文结构安排如下：第二部分概述国内外研究现状，第三部分建立长期联邦学习的动态博弈模型，第四部分提出基于强化学习的动态激励算法，第五部分汇报实验结果与发现，第六部分总结全文并提出未来研究的展望。

二、国内外研究现状

（一）联邦学习激励机制

联邦学习框架的提出引起了机器学习和人工智能领域的广泛兴趣，近年来已有许多研究致力于研究联邦学习合作中的激励机制设计。最初被提出时联邦学习用于跨设备联邦学习（Cross-Device FL）的场景下，一般由公司或组织发起机器学习模型训练项目，各设备将自身数据用于参与训练并获取奖励，例如金钱、应用程序积分等（McMahan et al.，2017）。这种模式在谷歌的 Gboard（Hard et al.，2018）和苹果的 iOS13（Bhowmick et al.，2018）等产品中得到成功应用。该模式保障用户和数据使用方之间不存在原始数据共享，在涉及用户个人信息的场景（例如，用户行为分析、用户输入法建模、人脸建模等）下具有显著优势。在跨设备联邦学习中，任务发布者无法保证移动设备诚实地完成所有轮次的训练，因此需要合理的激励机制来驱使设备进行稳定的数据投入。该场景下的联邦客户端并没有自己训练学习模型的需要，因此它只有参与和不参与联邦公共训练这两种选择，策略相对简单。目前，已经有许多研究关注了该场景下激励机制的设计，主要采用契约理论（Ding et al.，2020；Li et al.，2023；Liu et al.，2023）、拍卖理论（Cong et al.，2020；Zeng et al.，2020；Deng et al.，2022）等，保证客户端诚实地上报自己的相关条件，并关注了联邦项目的公平性（Xu et al.，2021）和隐私性（Sun et al.，2021）。

随着联邦学习的不断发展，越来越多的企业和组织开始认识到其在构建数据市场中的价值和作用，现阶段的讨论与研究也在逐步向广义联邦学习转向。广义联邦学习，即跨孤岛联邦学习（Cross-Silo FL），其核心在于通过合作建立模型解决数据孤岛问题，该模式下的联邦客户端为独立公司或组织。该模式主要应用于自动驾驶技术、医疗数据分析、金融机构合作等领域，通过模型优化等合作过程实现数据共享（Kaissis et al.，2020）。在跨孤岛联邦学习中，参与者既可以选择参与公共训练，也可以选择通过本地训练来提升自己的模型效用，因此在长期合作中存在跨期决策的问题，相比跨设备联邦学习中的客户端策略要更加复杂多样，本文的研究也聚焦于此类场景。目前，该领域已经有了一些有价值的探索。Yu et al.（2020）提出了一种适用于跨孤岛联邦学习的利润分享方案，该方案考虑了客户端等待收益到账的时间成本，并强调了各客户端收益的公平性以确保最大化集体效用。Lim et al.（2020）使用一个分层激励机制同时考虑了跨设备和跨孤岛的激励，通过倒推法，先利用契约理论制定合同进行跨设备的激励，然后再使用合并和拆分算法求解联盟博弈来建立跨孤岛的合作，在整个过程中实现公平性和激励相容性。Shi et al.（2022）通过一种高效的基于 Shapley 值的贡献评估方法、一种基于声誉以及局部和全局梯度分布的

新型激励机制，实现了奖励公平。然而，以上研究尚未重点关注联邦客户端在整个合作过程中的复杂动态策略，而本文更加关注多期合作中客户端的跨期选择问题，并进一步引入强化学习方法来设计动态的激励策略。

（二）长期联邦学习合作的建模

初期的联邦学习合作研究大多仅关注一次性的静态合作，即将对合作协议和激励机制的分析限制在单次博弈的框架之内，而没有充分考虑主体间跨期资源配置上的策略平衡。近年来，越来越多的研究开始关注长期联邦学习合作中更复杂的主体互动与激励方案。Wu et al.（2022）基于 VCG 拍卖模型考虑了长期联邦学习中的可持续合作，但其场景侧重于数据手机平台从分散的移动用户设备中收集数据，属于跨设备的联邦学习。Zhang et al.（2022）关注了客户端在长期跨孤岛联邦学习中的自私行为，并提出了一种带有惩罚方案的合作策略，使得这种策略成为重复博弈的纳什均衡，以减少“搭便车”现象。然而，他们的研究忽视了数据积累和跨时间决策问题。Bi et al.（2023）通过建立长期囚徒困境模型揭示了联邦客户端什么时候更愿意建立联邦学习伙伴关系，并使用一种基于惩罚的合同机制来促进联邦客户端对模型的贡献。尽管该研究关注到了长期合作中的数据积累，但其仅从多期重复博弈角度进行建模，其联邦客户端的策略局限于多期重复博弈中简单的以牙还牙、残酷触发策略。相比以上文献，本文通过建立动态博弈模型更细致地刻画了联邦客户端在长期合作过程中的动态互动策略。在建模过程中，我们尤其受到研发合作博弈和自愿公共品提供领域相关研究的启发。

联邦学习合作协议与研发合作博弈有许多共通之处：联邦学习中共同训练的公共模型可以视为一种知识外溢的形式，而知识外溢性引起的“搭便车”问题是研发合作博弈的核心。减轻“搭便车”问题的机制涉及奖励互惠行为和惩罚“搭便车”行为的回报结构，相关分析通常关注企业在创新激励和知识外溢阻碍之间的策略平衡（Cellini and Lambertini，2009；Yap et al.，2014）。基于此，本文在模型中设定了中央服务器，它可以根据联邦客户端是否在当轮贡献数据来差异化地反馈参数，这使得不贡献的客户端无法获得最新的模型参数，为预防“搭便车”行为带来了机制优势。

同时，联邦学习场景下的长期可持续合作也与经济学中自愿提供公共产品有共通之处，其中共同训练的联邦学习模型可以视为公共产品。自愿提供公共产品的领域在经典的动态博弈文献中有许多研究，这些研究致力于理解多方合作的演变、公共品贡献决策随时间的动态变化以及合作协议和激励对参与个体行为策略的影响。Fershtman and Nitzan（1991）最早开始系统性地探讨长期合作的稳定性、“搭便车”问题以及过去贡献对未来决策的影响。在此基础上，许多后续研究（Yildirim，2006；Kessing，2007）将自愿提供公共产品建模为能够被一维状态变量描述的马尔科夫问

题，其中参与方的策略是时间的连续函数，此时，动态优化问题中的经典方法，如马尔可夫完美均衡、反向归纳法和哈密尔顿函数法，可以用于求解均衡问题。在其启发下，本文将长期联邦学习中客户端动态合作的博弈问题刻画为由多维状态变量刻画的马尔科夫问题。由于维度上升，我们选择利用博弈树的方法来求解均衡。我们在建模时刻画了长期动态互动中策略选择的复杂性，充分考虑了联邦客户端在本地训练与合作训练间的选择以及在跨期资源分配方面的权衡。

（三）强化学习方法及其在联邦学习中的应用

强化学习（RL）是一种机器学习范式，它涉及一个智能体在与环境的交互中学习如何采取行动以达到某个目标。深度 Q 网络（DQN）算法是求解此类复杂连续决策问题广为采用的一种方法（Mnih et al.，2013；Mnih et al.，2015；Ling et al.，2023）。后续各种扩展性方法被提出，以提升其稳定性和效用。例如，双深度 Q 网络（van Hasselt et al.，2016；Cui et al.，2023）算法通过解耦行动选择和行动评估解决了 Q 值高估问题；优先级回放（Schaul et al.，2015；Saglam et al.，2023）通过带有优先级的数据重放，提升了数据效率；噪声网络（Fortunato et al.，2017；Zemzem and Tagina，2023）使用随机网络层进行探索，相比之前的贪婪探索更具系统性。本文也吸取了以上技巧，并在我们的场景中获得了良好的收敛性和出色的模型表现。

近年来，强化学习被创新地用于联邦学习的激励机制设计。在联邦学习训练中，中央计划者可以被建模为强化学习智能体，执行客户端选择和激励分配等动作，以助力可持续高质量的联邦学习合作的进行。Zhan et al.（2020）为中央计划者和联邦客户端分别训练了强化学习智能体，以助其决策最优定价策略和最优训练策略，基于强化学习的策略能够解决信息不对称问题并且其效果好于其他基线策略。Wang et al.（2020）提出了 SFAC 架构，一种用于无人机辅助移动群智感知的联邦学习框架，利用基于两层强化学习的激励机制，促进无人机高质量模型共享。Wu et al.（2022）提出了一种具有长期在线 VCG 拍卖机制的联邦学习，该机制采用基于经验的深度强化学习算法来获得最优策略。以上研究显示了强化学习在联邦学习激励机制设计中的应用潜力，然而以上研究都是聚焦于跨设备联邦学习，其中的联邦客户端不具有本地训练或参与公共训练的复杂选择。本文同样应用深度强化学习算法来进行中央计划者的激励方案设计，并为其设计了信息更加全面的系统状态量，得到了高性价比的激励方案。

综上所述，本文在现有文献的基础上，细致地刻画了长期跨孤岛联邦学习合作过程中的策略互动和资源配置决策过程，设计相应的深度强化学习算法对联邦学习的各参与方进行动态的、时变的激励，为可持续联邦学习的激励机制设计提供理论及实践保证。

三、联邦学习场景的动态博弈建模

联邦学习规避了跨实体原始数据的直接共享，利用参数聚合技术，如广泛使用的 FedAvg，在不同实体间传递机器学习模型信息。这种方法使各数据持有方可以在不向外传输敏感数据的情况下合作进行模型训练。

联邦学习的架构由一个中央服务器和多个联邦客户端构成。通常，联邦学习过程始于中央服务器指定学习任务、建立初始化的模型架构并招募共同训练模型的多个联邦客户端。服务器将初始模型分发给每个客户端，客户端将各自使用其本地数据独立训练本地模型。客户端的本地模型训练至收敛后，再将更新的模型参数（如 CNN 模型的梯度参数）发送回服务器。服务器采用 FedAvg 等参数聚合方法来合并这些梯度并以此更新全局模型。在后续迭代中，服务器将更新的模型参数重新分发给每个客户端，客户端继续自适应地训练其本地模型。这个迭代过程一直持续到全局模型收敛。在此过程中，通过传递模型参数，客户端之间实现了间接的数据交流。

本文聚焦于跨孤岛的长期联邦学习合作场景。为刻画客户端的跨期策略互动，我们进行如下场景建模：联邦客户端通过中央服务器进行联邦学习合作，在离散的 T 个时期内合作训练一个联邦模型。在此过程中，数据被视为一种同质的可投入资源，客户端随时间不断自然获得新数据，并用以更新模型。在 T 期内，模型不重置，而是基于新的训练数据进行更新。在每个时期，客户端可以选择是否参与当期的联邦学习合作。如果客户端选择不参与合作，它能够利用本期收集的数据在上一期模型的基础上进行本地的模型训练，并在该期结束时获得当前本地模型的收益。如果客户端选择参与合作，则可以不断训练并和中央服务器交换参数直到当期模型收敛。我们假设客户端参与联邦学习，即与中央服务器交换训练参数（以下简称“投入”）会引致额外成本 c（包括传输成本、数据风险等），但作为回报可以获取中央服务器回传的当期最新公共模型参数。这样，“投入”的成本和“投入”的潜在收益就构成了客户端在长期联邦学习合作中的权衡因素，支配着客户端的策略互动。

这里值得指出的是，由于本文聚焦于客户端之间的策略互动，即在每个周期内是否选择将本地更新参数传输至中央服务器，尽管在联邦学习中实际上传输的是模型参数而非直接“投入”原始数据，为表意方便，我们仍将客户端的行动简称为“投入”数据。

（一）数据收集过程与动态投入决策

该模型可用于刻画多个联邦客户端的策略互动，为了简化叙述，我们在下文中以 2 个联邦客户端为例，详细介绍 T 期联邦学习合作的动态过程。如前所述，在 T

期联邦学习过程中，数据可以视为同质的可投入资源，客户端每期开始时获取一定量的数据，并可以选择是否将当期获得的新数据“投入”中央服务器以进行联邦模型训练：若选择“不投入”，则当期获得的新数据仅用于本地服务器上训练并更新本地模型；若选择“投入”，则在本地训练的基础上，额外耗费成本 c 与中央服务平台进行参数交互，并收获其回传的最新公共模型参数。既往研究更多着眼于单次的合作决策，我们希望深入刻画长期的动态过程，具体地，我们对该过程引入如下详细假设：

（1）数据收集：在 T 个离散时期内，联邦客户端双方在每期开始时均可自然获得 1 单位新数据，该单位新数据仅可用于当期行动决策，即当期数据仅能投入本地训练或共享至联邦模型。

（2）模型演化：在 T 期的合作过程中，联邦客户端双方及中央服务器各自保有自身模型，即两个本地模型和一个联邦学习模型。这些模型都随着时间推移和新数据加入而进行不同程度的更新，即：随着联邦学习过程的进行，这些模型相当于是在逐渐增大的训练数据上训练得到的。因此在长期联邦学习中，模型 m_t 的演化可以用其包含的训练数据量 x_t 的演化来刻画。记联邦学习产生的三个模型系列（客户端 A 的本地模型、客户端 B 的本地模型、联邦模型）分别为 $\{m_t^{\mathrm{A}}\}_{t=1}^{T}$，$\{m_t^{\mathrm{B}}\}_{t=1}^{T}$ 和 $\{m_t^{\mathrm{P}}\}_{t=1}^{T}$，相应地记这些模型所包含的数据量序列为 $\{x_t^{\mathrm{A}}\}_{t=1}^{T}$，$\{x_t^{\mathrm{B}}\}_{t=1}^{T}$ 和 $\{x_t^{\mathrm{P}}\}_{t=1}^{T}$，下文中将着重论述各 $\{x_t\}$ 的演化过程。

（3）数据聚合与模型效用：在博弈建模中各方的学习模型均随数据累积而演化，在统计意义上，模型精度与训练集规模正相关，我们参考 Zhang et al.（2020），采用广义恒弹性（CES）替代生产函数来刻画多来源数据共同训练模型的精度：

$$f(x_1, \cdots, x_n) = X^{\rho} = \left[\left(\sum_{i=1}^{n} (x_i)^r \right)^{\frac{1}{r}} \right]^{\rho}$$

其中，$x_i(i = 1, 2, \cdots, n)$ 为 n 组非同源数据的数据量，X 为利用这 n 组非同源数据进行机器学习的等效数据量，$f(X)$ 为模型精度，r, ρ 分别为替代率参数和边际效应参数。我们进一步假设：

①模型精度边际递减，即 CES 生产函数中 $0 < \rho < 1$。此时精度效用 f 是等效数据量 X 的递增的凹函数。

②数据完全替代，即 CES 生产函数中 $r = 1$。该假设意味着每个时期、每个客户端获得的 1 单位数据均为同质且完全替代的，即先后用 a 单位数据和 b 单位数据训练模型的效果与用 $(a + b)$ 单位数据训练模型的效果相同。

（4）投入决策与参数回传：

①在每个时期开始时，每个客户端都获得 1 单位的新数据，并决定是否将新数据“投入”给中央服务器或仅用于本地训练。以 $u_t^i \in \{0, 1\}$ 表示联邦客户端 i 在第 t 期选择是否“投入”的决策变量，$u_t^i = 1$ 表示客户端 i 在第 t 期选择“投入”。

②所有客户端决策过后，中央服务器基于所有客户端的“投入”进行模型参数聚合，从而将公共模型 m_{t-1}^{P} 更新至模型 m_t^{P}。具体地，由于该过程中联邦模型 m^{P} 吸收了当期“投入”的新信息，故对应的等效数据量变化关系为 $x_t^{\mathrm{P}}=x_{t-1}^{\mathrm{P}}+u_t^{\mathrm{A}}+u_t^{\mathrm{B}}$。

③联邦模型当期聚合完成后，m_t^{P} 的参数被发送给本期的“投入”者。即只有当期选择“投入”行动的博弈参与者才能收到 m_t^{P} 的更新反馈参数。

④在收到服务器的反馈后，客户端 i 能够聚合公共模型和上期本地模型中所包含的数据信息，并将本地模型 m_{t-1}^{i} 更新至模型 m_t^{i}。具体地，若第 t 期客户端 i 选择“不投入”（$u_t^i=0$），则由于不涉及从中央服务器获得数据信息，其当期模型中的等效数据量 x_t^i 仅在上期基础上增加当期自身获得的 1 单位；反之，若第 t 期客户端 i 选择“投入”（$u_t^i=1$），则其模型中的等效信息量包括累积至当期的 t 单位自身收集的数据以及中央服务器回传参数中蕴含的对手累积“投入”数据量 $\sum_{\tau=1}^{t}u_\tau^{-i}$。综上，在第 t 期，博弈双方持有的等效数据量可以用如下递推方式给出见式（1）：

$$\begin{cases} x_0^i = 0 \\ x_t^i = u_t^i \cdot \left(t+\sum_{\tau=1}^{t}u_\tau^{-i}\right)+(1-u_t^i)\cdot(x_{t-1}^i+1) \end{cases} \tag{1}$$

基于上述分析，在联邦学习中客户端与中央服务器的数据传输关系见图 1。客户端可以选择当期是否对联邦模型进行（单位）参数投入（$u_t^i\in\{0,1\}$）；双方选择完毕后，服务器根据当期是否投入将聚合后的联邦模型参数回传给当期投入者；最后，所有客户端根据当期获得的所有新信息更新模型。

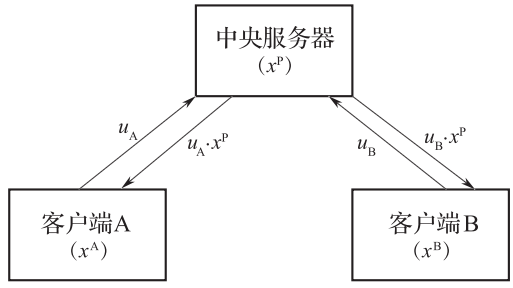


图 1　双客户端联邦学习参数传输框架

注：本图展示了 1 个中央服务器 +2 个联邦客户端的情形：三方持有的模型分别记为 x^{A}, x^{B}, x^{P}；在每一时期，联邦客户端各自选择是否贡献 1 单位数据，即选择 u^{A}, $u^{\mathrm{B}}\in\{0,1\}$；选择完毕后，中央服务器进行参数聚合及回传，回传方向上的 $u_i\cdot x^{\mathrm{P}}$ 表示中央服务器仅回传给当期参与贡献的客户端，即 $u_i=0$ 时无回传。

（二）联邦客户端的效用度量

在该动态博弈中，博弈参与方，即联邦客户端在每期都根据其自身模型的精度

获得效用，在每个时期 t，客户端 i 的即时收益取决于时刻 t 的模型准确性，从而取决于其当前等效数据量 x_t^i，同时“投入”中央服务器会引致额外的成本 c，即传输成本或风险成本。因此，我们定义联邦客户端在第 t 期的即期收益为：

$$F(x_t^i, u_t^i, c) = f(x_t^i) - u_t^i \cdot c \tag{2}$$

理性的博弈参与方 i 会最大化自身各期内的效用折现之和，双方折现率统一为 γ，则客户端的优化目标为最大化 T 期内的折现总收益 $\sum_{t=1}^{T} \gamma^{t-1} F(x_t^i, u_t^i, c)$。

（三）博弈树与均衡

基于前文提出的背景假设，该博弈为动态差分博弈，我们通过以下方式考虑其闭环策略均衡，即任意一方客户端在任何一期的“投入 – 不投入”策略是基于其在第 t 期时对已发生的合作历史的完全知悉以及对未来行动的理性预期，即非事前决定的，而是随着双方合作历史动态决策的。该博弈具有较为复杂的状态转移和收益规则，因此我们通过图 2 的博弈树来直观地呈现求解闭环均衡策略的逆向递推法。

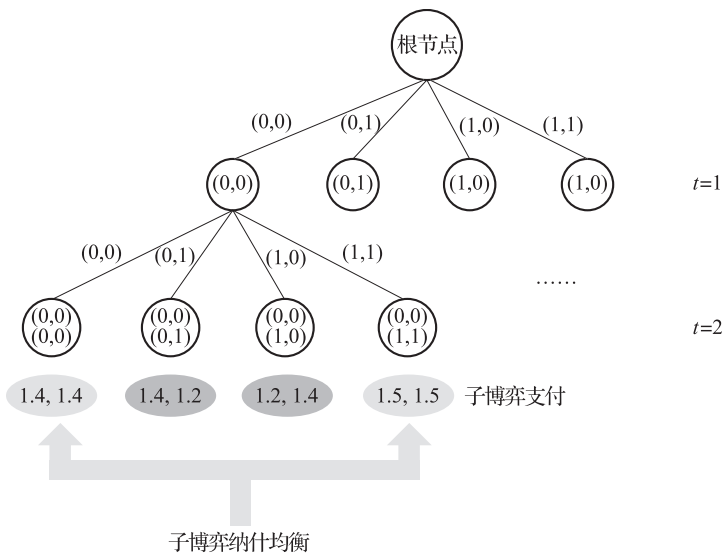


图 2　双客户端动态博弈树框架

注：本图展示的博弈树为效用函数 $f(x) = \sqrt{x}$，$c = 0.2$ 时双客户端博弈前两轮的情形，其中以 0 表示行动“不投入”，1 表示行动“投入”。

T 期双人动态博弈中所有可能出现的互动情况可由深度为 T 的四叉树刻画，博弈树的第 t 层有 2^t 个节点，表示博弈截至第 t 期结束可能出现的所有行动组合，即截至第 t 期所有可能的历史路径。每个节点衍生出 4 个子节点，分别对应之后一期双方的四种决策组合：“$u_t^A = 0, u_t^B = 0$”“$u_t^A = 0, u_t^B = 1$”“$u_t^A = 1, u_t^B = 0$”“$u_t^A = 1, u_t^B = 1$”。图 2 展示了一个深度 $T = 2$ 期的博弈树，其中例如 $t = 2$ 层节点“(0, 0)(1, 0)”

表示双方在第 1 期均未投入，第 2 期 A 投入 B 不投入的对应历史。每个节点都会根据当前历史对博弈双方产生支付（图 2 中仅展示了 $t=2$ 期的支付），理性客户端会最大化自身在各期的支付折现总和。

为求解合理的博弈均衡，本文采用序贯理性和序贯均衡的概念。序贯理性要求博弈者在自身的每个信息集上都采取自身利益最大化的行动。针对本文博弈模型而言，由于在长期联邦学习合作中，前期的投入与否会影响后期的收益，因此序贯理性要求博弈者同时关注投入决策的即期效应和长期影响，只有当博弈进行到最后一期 $t=T$ 时，由于此时投入与否不再存在未来长期影响，博弈者的理性决策才为最大化当前历史下第 T 期即期收益。此时第 T 期子博弈的均衡可由纳什均衡简单刻画：如图 2 所示，在 (0,0) 节点所对应的子博弈中，“(0,0)(0,0)”和“(0,0)(1,1)”均为子博弈纳什均衡，即在第 1 期双方“都不投入”的条件下，第 2 期双方“都投入”或“都不投入”均构成子博弈纳什均衡。

根据该博弈的嵌套结构，可以用动态规划的想法求解博弈均衡，每个节点处除了根据其对应的历史对博弈双方产生即期收益，还会在历史基础上展开子博弈。因此，每个节点对客户端而言有两部分价值：一是即期收益，二是其对应子博弈产生的未来收益期望。根据这一点，可以构建求解闭环策略均衡的逆向递推法：①对于 $t=T$ 的叶子节点，节点价值仅为对应其历史的即期收益。②对于非叶子节点，其节点价值等于该节点对应历史的即期收益，加上其可构成纳什均衡的子节点的节点价值折现。一个子博弈可能存在多个子博弈均衡，由于本文旨在研究激励对自发博弈均衡的影响，我们仅考虑其中双方收益最高的均衡，例如前述例子中的“(0,0)(0,0)”和“(0,0)(1,1)”均为“(0,0)”节点子博弈纳什均衡，我们仅考虑优势均衡“(0,0)(1,1)”。进而，我们可使用值函数和动态规划法搜索闭环策略均衡。该算法求解的闭环策略均衡（反馈策略均衡）的核心在于求解以下递归地最大化 $F(u_t^i, u_t^{-i} \mid H_t) + \delta V(H_t, u_t^i, u_t^{-i})$。其中，$H_t$ 代表博弈历史，它与当期策略情况 u_t^i, u_t^{-i} 一起决定当期的等效数据量 x_t^i；F 为节点即期收益函数，V 为节点综合收益函数。通过建立一个四叉博弈树，我们可以计算每条路径的价值，进而解得博弈均衡，具体过程见过程 1（Procedure 1）。算法第 1 ~ 3 行递归地生成一个 T 层的四叉树，并为每个节点标记动作属性，记录该节点所代表的客户端的当期“投入”情况，以便于历史策略追溯；第 4 ~ 6 行从根节点根据历史路径计算每个节点对应的节点状态，即根据式（1）计算等效数据量 $(x_t^{\mathrm{A}}, x_t^{\mathrm{B}})$；第 7 ~ 14 行计算节点价值，首先根据式（2）定义即期收益函数，然后从叶子节点开始向上递归计算节点价值。对于叶子节点，由于博弈在该期结束，因此节点价值就是即期收益值；而其他节点的价值则包括即期收益和未来收益期望两部分。我们根据该节点的子节点价值矩阵解得子博弈纳什均衡，将该均衡子节点的价值作为未来收益期望。记录每个子博弈纳什均衡后，即得到了整个博弈的均衡路径。

Procedure 1: Generate Game Tree (GT)

Input: tree depth T; cooperation cost c; discount rate γ
Output: Game Tree GT

```
Generate Gt nodes;
for t =1 to T do
    recursively generate GT nodes with node.action in ([0,0],[0,1],[1,0],[1,1]) of depth t;
Calculate node.state;
for nodes with t = 1 to T do
    calculate node.state according to equation(1);
Calculate node.value;
define function payoff(node.state,node.action,c) according to equation(2);
for nodes with t =T to 1 do
    if t == T then
        node.value=payoff(node.state,node.action.c);
    else
        select node.BestChild;
        node.value=payoff(node.state,node.action,c)+γ node.BestChild.value;
```

四、基于强化学习的激励方案

强化学习（RL）是智能体根据环境采取行动以获得最大奖励的学习过程，它能够从过去的经验中学习策略，因而被广泛应用于博弈中。在每个时间步 t，强化学习智能体观察状态 s_t，并执行动作 a_t。执行动作后，环境状态转移到下一个状态 s_{t+1}，强化学习智能体收到奖励 r_t。状态转换和奖励遵循马尔可夫过程，是一个离散时间随机控制过程，由序列 $s_1, a_1, r_1, s_2, \cdots, r_{t-1}, s_t, a_t, r_t, \cdots$表示。强化学习的目标是最大化预期累积折现回报 $\sum_{t=1}^{T} \gamma^{t-1} r_t$，其中 γ 是折现因子，用于将未来的奖励折现。

本文将联邦学习合作组织方作为中央计划者，负责在每个合作期发布当期激励，以促进联邦客户端的投入。进而，可为中央计划者训练一个帮助其决策激励值的深度强化学习（DRL）智能体。在每个合作期 t 开始时，中央计划者首先观察联邦学习系统的当前状态 s_t，并发布其愿意为本期贡献者提供的奖励，也即动作 a_t。随后，联邦客户端结合当前合作进程和本期奖励额度，基于前述博弈树决定是否在本期合作，本期训练完成后，即可计算得到 r_t, s_{t+1}。在一次联邦学习合作的 T 期中，不断重复上述过程，即可收集到 T 组训练数据。DRL 的目标是找到最优策略，即将任意状态 s_t 映射到最大化累积折现奖励期望的动作 a_t，在本文的场景中表现为在联邦学习的任意进程下选择下一期的最优激励。

（一）DRL 的状态、动作与奖励设计

DRL 方法的成功依赖于状态、行动和奖励的精心设计。由于联邦学习系统是一个较为复杂的系统，包含各种信息，因此定义其状态是相对困难的。在本文中，我们在设计中选择了一种包含全信息的定义方式。

状态包括三个方面：联邦客户端的历史合作情况、联邦合作的进程情况、中央服务器的历史行动与奖励情况。以一个 6 期的合作为例，其对应 29 维的状态。其中，前 12 维是两个联邦客户端分别的历史合作情况，1 代表合作，0 代表不合作；中间 5 维代表的是双方的累计投入量和累计积累量以及当前的合作轮次；最后 12 维则代表中央服务器在 6 期中的历史行动与奖励情况。状态在初始化时被设置为全 0。举例来说，State = [1,1,0,0,0,0,0,1,1,0,0,0,0,0,4,2,4,2,2,0.06,0.904,0,1.55,0,0,0,0,0,0,0,0,0,0]，这代表两个联邦客户端截至目前在第 1、2 期进行了合作，当前合作进行了 2 轮，双方的历史总投入为 2 次，累计等效数据量为 4，中央服务器在第 1 期给出的激励为 0.06，当期奖励 $r_1 = 0.904$，第 2 期给出的激励为 0，当期奖励 $r_2 = 1.55$。这种状态设置既可以体现当前的合作进程，也可以体现合作历史中的行动和奖励情况，包含了合作进行到目前所涉及的全部信息。

动作是本轮中央计划者发布的激励，动作空间是连续且无界的。为了简化，我们将范围限制为 $[0,c]$，并将其离散化为仅精确到小数点后两位。值得注意的是，如果客户端并没有选择合作，那么本轮实际付出的激励为 0，因为不合作的客户端不会得到激励。

在涉及中央计划者的奖励时，我们考虑了联邦客户端的收益和中央计划者为激励付出的成本。我们将两个联邦客户端和一个中央计划者作为一个系统来看，联邦客户端通过激励获得的额外收益与中央计划者付出的激励成本是相互抵消的，这个部分不产生额外的效用。而由于更积极的合作所带来的额外收益则是总体社会福利的真正来源。我们希望中央计划者能够以相对小的激励成本换取联邦客户端更多的合作，因此在奖励中加入了一个关于成本的惩罚项。t 时刻的奖励 r_t 如下：

$$r_t = [f(x_t^{\mathrm{A}}) - (c - a_t)u_t^{\mathrm{A}} + f(x_t^{\mathrm{B}}) - (c - a_t)u_t^{\mathrm{B}} - \lambda\, a_t(u_t^{\mathrm{A}} + u_t^{\mathrm{B}})]/2 \tag{3}$$

其中，$f(x_t^{\mathrm{A}}) - (c - a_t)u_t^{\mathrm{A}}$ 代表其联邦客户端 A 在第 t 期的收益，$f(x_t^{\mathrm{A}})$ 代表其模型的效用，而 $-(c - a_t)u_t^{\mathrm{A}}$ 代表其受到激励后贡献数据的成本；同样地，$f(x_t^{\mathrm{B}}) - (c - a_t)u_t^{\mathrm{B}}$ 则代表联邦客户端 B 在第 t 期的收益，$f(x_t^{\mathrm{B}})$ 代表其模型的效用，而 $-(c - a_t)u_t^{\mathrm{B}}$ 代表其受到激励后贡献数据的成本；最后一项 $-\lambda\, a_t(u_t^{\mathrm{A}} + u_t^{\mathrm{B}})$ 则代表中央计划者进行激励行为所付出的成本，其中 λ 是预算约束，可以理解为是关于成本的惩罚项，λ 越高意味着中央计划者对激励成本的敏感度越高。

（二）基于优先级回放的双深度噪声 Q 网络的激励算法

本文基于深度 Q 网络（DQN）方法（Mnih et al., 2013, 2015）来训练中央计划者的激励机制，该方法是强化学习领域的经典方法，也能够很好地满足我们的需求。其核心思想是学习一个动作值函数（Q 函数），表示在给定状态下采取某个动作的预期回报。Q 函数的更新如下：

$$Q(s, a) = Q(s, a) + \alpha \cdot [r + \gamma \cdot \max Q, (s', a') - Q(s, a)] \tag{4}$$

其中，$Q(s, a)$是在状态 s 下采取动作 a 的 Q 值；α 是学习率；r 是环境返回的即时奖励；γ 是折现因子，表示对未来奖励的重视程度；s'和 a'是采取动作 a 后获得的下一状态和下一动作。

DQN 引入了深度学习，通过使用深度神经网络来逼近和学习 Q 函数。这个深度神经网络以接收状态作为输入，输出每个可能动作的 Q 值。在每个训练回合中，中央计划者根据 Q 函数选择 Q 值最大的动作，即当期激励预算，并进行发布，联邦参与者收到激励预算后决定本期是否贡献，本期联邦学习合作完成后则可得到当期奖励和下一个状态。这一组经验被存储到回放缓冲区中，然后抽样进行训练，这能够打破样本之间的时序关联，提高数据的利用效率。另外，值得一提的是 DQN 通过引入一个目标 Q 网络，来减缓训练过程中的目标值变化，提高稳定性。尽管 DQN 已经在 Atari 游戏等领域展现出强大的性能，但将它直接应用于我们的场景时仍难以获得良好的效果。DQN 方法存在着 Q 值高估、随机采样忽略区别经验重要性以及贪婪探索不够系统化的问题。因此，我们进一步采用了双深度 Q 网络（DDQN）、优先级回放、噪声网络的改进方案。

1. 双深度 Q 网络

我们在算法中采用 DDQN 方法，它是经典 DQN 方法的改进。DQN 背后的主要思想是使用神经网络来近似 Q 函数，该函数表示给定状态下每个可能动作的未来奖励期望。通过训练网络以最小化预测 Q 值和目标 Q 值之间的差异，DQN 可以学习在给定环境中做出最佳决策。在传统的 DQN 方法中，用于动作选择和动作评估的 Q 值是使用同一组参数来估计的。其更新规则可以表示为：

$$Q(s_t, a_t) \leftarrow r_t + \gamma \cdot \max_{a'} Q'(s_{t+1}, a') \tag{5}$$

这可能会导致 Q 值的高估，从而对学习过程和学习策略的质量产生负面影响。DDQN 通过解耦动作选择和动作评估过程来解决此问题。在 DDQN 中，选择下一时刻 Q 值最大的动作 a'是由主网络 Q 完成的，然后再用目标网络 Q′去计算该动作 a'对应的 Q 值。具体而言，DDQN 的更新规则可以描述为：

$$Q(s_t, a_t) \leftarrow r_t + \gamma \cdot Q'(s_{t+1}, \operatorname{argmax}_{a'} Q(s_{t+1}, a')) \tag{6}$$

这样，任何一个网络对 Q 值的高估都有机会被另一个网络修正，从而降低了 Q 值的过高估计，提高了算法的稳定性。

2. 优先级回放

在传统的 DQN 算法中，经验回放缓冲区中的样本被均匀地随机采样，这忽略了不同样本的重要性区别。优先级回放引入了样本的优先级概念，使得那些对网络参

数更新影响较大的样本被更多采样。衡量样本的重要性需要用到一个指标——Temporal Difference（TD）误差，该误差反映了当前网络对于该样本的预测与实际目标的差距。TD 误差越大，表明该样本对网络的训练具有更高的信息量，因此被赋予更高的优先级。优先级回放方法根据样本的优先级，确定每个样本被抽样的概率，优先级越高的样本被选中的概率越大，但仍保留一定的随机性。在实际算法中，如果每次采样都对优先级进行排序并选择最大值会浪费大量的计算资源。常见的解决方案是使用 SumTree 的数据结构来存储样本的优先级，其中树的叶子节点存储着每个样本的优先级，而父节点的值等于其子节点的和。SumTree 的采样过程为：①根据根节点的优先级和采样样本数，划分采样区间，然后在这些区间中均匀采样，得到所要选取的样本的优先级；②从根节点开始，逐层将样本的优先度和节点的优先度进行对比，最终可以得到所要采样的叶子样本。这种树状结构本质上采用了概率式的选择方式，相比线性结构，这种方法不用逐一检查数据，大大提高了运行效率。

3．噪声网络

在强化学习中，提高智能体的探索能力是一个关键问题。传统的 DQN 通常采用 ε – 贪心策略，即以 ε 的概率采取随机策略，以增加系统的探索性。而噪声网络通过对神经网络的参数引入噪声，从而增强模型的探索能力。神经网络全连接层的前向计算公式为：

$$\hat{y} = w\hat{x} + b \tag{7}$$

其中，w 是权重，b 是偏置，$\hat{x}$和$\hat{y}$分别代表神经网络全连接层的输入和输出。在传统的 DQN 中，这些参数通常是确定的。但是在噪声网络中，通过为这些参数引入高斯噪声可以得到新的网络 $\tilde{Q}$。具体的计算公式为：

$$\hat{y} = (\mu_w w + \sigma_w \odot N_w)\hat{x} + \mu_b + \sigma_b \odot N_b \tag{8}$$

其中，μ_w, μ_b是权重和偏置的均值，σ_w, σ_b是它们的标准差，N_w, N_b 是高斯噪声。

在每个回合（Episode）开始时，在智能体与环境进行交互之前，采样噪声。在整个回合中，噪声网络的参数保持不变。ε – 贪心策略在动作空间上加噪声，这导致了在同一个回合内，在相同或相似的状态下，智能体不一定采取相同的动作。而噪声网络是在参数空间上而非动作空间上加噪声，这能够保证在同一个回合内动作与状态的一致性，使得探索与状态相关，更加系统化，从而提高强化学习任务中的探索效率。

结合以上方法，我们为中央计划者设计了基于优先级回放的双深度噪声 Q 网络激励算法，见算法 1（Algorithm 1）。算法第 1 ~3 行初始化了回放缓冲区、Q 网络和目标 Q 网络。同时，初始化动作值函数 Q 和目标动作值函数$\hat{Q}$，它们在初始化时使用随机的权重。从第 4 行开始，进入训练回合的循环，第 5 ~6 行表示在每个训练回合开始时，首先初始化状态为全 0，并根据算法 1 生成用于联邦客户端决策的博弈

树。从第 7 行开始进入一次联邦合作的时间步循环，对于每个时间步，首先利用 Q 网络根据当前状态 s_t 选择本期动作 a_t，即当期发布的奖励；随后，根据过程 2（Procedure 2）更新博弈树并获得客户端的决策情况 $u_t^{\mathrm{A}}, u_t^{\mathrm{B}}$；客户端完成决策后即可根据状态转移方程和奖励函数得到下期状态 s_{t+1} 和本期奖励 r_t；将当期的经验（s_t, a_t, r_t, s_{t+1}）放入回放缓冲区并设置状态为新状态。当回放缓冲区中有足够的样本后，从内存 D 中基于优先级采样方法采样一个批次的经验，用于更新 Q 网络。每结束一次合作的时间步循环，重置用于引入噪声的 Q 网络的噪声；每隔 U 个训练回合，将目标动作值函数 $\hat{Q}$ 的权重更新为当前 Q 函数的权重。完成所有训练回合后，算法结束。其中用到的博弈树更新与客户端决策算法在强化学习中属于智能体（中央计划者）与环境（联邦客户端）交互的接口，具体见过程 2（Procedure 2）。首先，提取出状态变量中包含的联邦客户端的历史决策信息，根据历史路径匹配到博弈树中所对应的节点位置。客户端当前处在该节点处进行决策，由于中央计划者发布了激励 a_t，那么客户端在计算下一期的即期收益时将成本参数重置为 $c' = c - a_t$。根据 c' 更新当前节点所有子节点的价值，然后重新接触均衡子节点，该节点的动作则对应着客户端的决策 $u_t^{\mathrm{A}}, u_t^{\mathrm{B}}$。这里值得指出的是，每次更新博弈树只重置了当前深度节点的价值，并不影响后续节点的价值，而节点价值的计算只与子节点的价值相关而与父节点价值无关，故而在一次合作期间不需要因为这种节点价值更新而更新重置博弈树。因此，在算法 1（Algorithm 1）中，我们只在每次时间步循环开始前生成博弈树，这提高了算法的效率。

Algorithm 1: DRL-based Incentive Mechanism

Input: cooperation rounds T; cooperation cost c; discount rate γ, memory capacity N; episodes E; target update U; batch size B; penalty coefficient λ

Output: action-value function Q

Initialize replay memory D to capacity N;
Initialize action-value function Q with random weights θ;
Initialize target action-value function $\hat{Q}$ with weights $\theta^{-}=\theta$;
for *episode* = 1 *to* E **do**
  Initialize s_1= [0]∗len(s_1)
  GT= **Generate Game Tree** (T,c,γ)
  for t= 1 *to* T **do**
    a_t= arg max$_a$ $Q(s_t, a;\theta)$
    $u_t^{\mathrm{A}}, u_t^{\mathrm{B}}$= **Update GT and Decide Strategy**(s_t,a_t,GT,c,γ)
    calculate r_t and s_{t+1} with $u_t^{\mathrm{A}}, u_t^{\mathrm{B}}$, γ
    store transition (s_t,a_t,r_t,s_{t+1}) in D
    set $s_t = s_{t+1}$
    if *the size of memory* $D \geq$ *batch size* B **then**
      Prioritized sample batch of transitions(s_j, a_j, r_j, s_{j+1}) from D
      $$y = \begin{cases} r_j & \text{if episode ends at } j+1, \\ r_j + \gamma \cdot \boldsymbol{Q}'(\boldsymbol{s}_{t+1}, \arg\max_{a'} \boldsymbol{Q}(\boldsymbol{s}_{t+1}, \boldsymbol{a}';\boldsymbol{\theta});\boldsymbol{\theta}^{-}) & \text{otherwise.} \end{cases}$$
      perform a gradient descent step on$(y_j - \boldsymbol{Q}(s_j, a_j;\theta))^2$ with respect to the network parameters θ
    Reset **Noise** for Q-net;
    Every U episodes reset $\hat{Q} = Q$;

Procedure 2: Update Game Tree(GT) and Decide Strategy

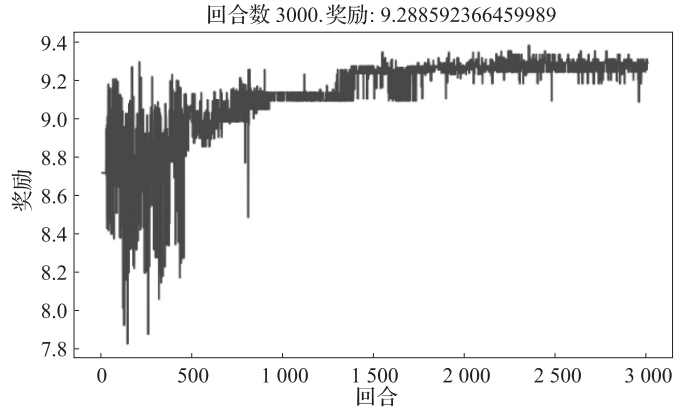
Input: state s_t; action a_t; Game Tree GT; cooperate cost c; discount rate γ
Output: strategies of clients u_t^A, u_t^B

```
Get the historical strategies from s_t;
Match them with node.action in GT to reach the CurrentNode;
Update CurrentNode.children.value with c'=c−a_t and γ;
for nodes in CurrentNode.children do
    if t == T then
        node.value=payoff (node.state,node.action,c');
    else
        select node.BestChild;
        node.value=payoff(node.state,node.action,c')+γ node.BestChild.value;
Select CurrentNode.BestChild;
u_t^A, u_t^B=CurrentNode.BestChild.action;
```

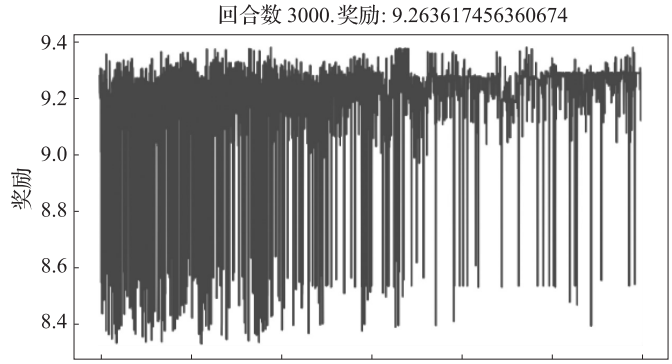
五、仿真结果与分析

（一）实验设置

假设有 2 个组织作为联邦客户端参与到长期联邦合作中，一次合作的总期数为 T，每期参与合作的成本为 c，效用函数为 $f(X)=\sqrt{X}$，其中 X 为等效数据量。通过无激励的博弈预实验，我们选择 $T=6$ 来进行主要实验，更大的 T 需要消耗更多的计算资源，而更小的 T 则无法包含丰富的均衡形式。我们在预实验中测试了成本 c 从 0 到 0.5 时的结果，最终选取 $c=0.25, 0.35, 0.45$ 三个值作为不同成本参数下长期联邦学习合作博弈的代表来进行后续实验。同时，研究中以网格搜索的形式进行了充分的预实验以进行超参数取值调优，结果表明模型表现对于折现因子 γ 不敏感，这可能是由于模型合作期数相对较短、合作环境相对稳定。基于此，在主要结果展示部分，我们仅展示 $\gamma=0.9$ 时的结果。此外，预实验也证明在本文的场景下模型表现对于折现因子 γ 不敏感，基于此，在主要结果展示部分，我们仅展示 $\gamma=0.9$ 时的结果。Q 网络的输入维数即状态维数与总期数相关，当 $T=6$ 时，输入为 29 维状态值，Q 网络的输出与成本 c 相关，我们将输出简化为$[0, c]$之间保留两位小数的值，因此输出维数为 $100c+1$。在输入和输出层之间，Q 网络还包括两个全连接线性层，一个包含 290 个神经元的特征层和一个包含 40 个神经元的价值层。在强化学习的训练中，我们设置训练回合 $E=10\,000$，经验缓冲区内存大小 $N=10\,000$，训练批量 $B=128$，目标 Q 网络更新间隔 $U=100$。另外，预算约束 λ 也是一个值得讨论的参数，我们在后面的实验中讨论了其对结果的影响。我们首先取 $\lambda=1.6$，$c=0.45$ 测试了算法的收敛性。图 3a 和图 3b 分别展示了 DDQN + Noisy Net 与 DQN 下，模型收益随训练回合的变化，显然 DDQN + Noisy Net 具有良好的收敛性，而 DQN 直接应用则难以取得良好的效果。因此，在后文中，我们主要展示 DDQN + Noisy Net 下的结果。



a）DDQN+Noisy Net



b）DQN

图 3　算法收敛性的比较

注：本图为作者数值模拟实验的结果。容易看出，在同样的训练回合数下，DDQN + Noisy Net 方法的奖励相较于 DQN 方法收敛性更佳。

（二）实验结果与分析

由于强化学习结果具有一定随机性，我们对所有实验重复 5 次，并取结果的平均值进行展示。我们在 $c = 0.25, 0.35, 0.45$ 三种不同的合作成本参数下进行了实验，它们对应着不同成本下的无激励均衡状态。

以下为表述方便，我们将客户端在 T 期内投入决策的均衡路径以长度为 T 的 0 – 1序列表示，即以序列第 t 位上的 0 表示客户端在第 t 期选择不投入，1 表示客户端在第 t 期选择投入。例如，对于本实验中 $T = 6$ 的情形，长度为 6 的 0 – 1 序列“010101”表示客户端在且仅在第 2，4，6 期选择投入。

总体而言，联邦客户端在无激励时会呈现间隔投入的特点。$c = 0.25$ 时，联邦客户端的闭环均衡策略为 111010；$c = 0.35$ 时，为 110100；$c = 0.45$ 时，为 010100。尽管三者有不同的策略，但总体都呈现出投入 – 不投入交替出现的模式，在后面的分析中可以进一步看到它们的共性特点。

在展示引入额外激励后的结果之前，我们先对相关概念进行简要说明。我们所研究的联邦学习合作系统，包括联邦客户端和中央计划者两方。在合作过程中，联邦客户端的总收益包括两个部分，一是联邦合作所带来的基于模型精度的收益，二是中央计划者提供的激励；而中央计划者的总收益为负值，代表其为联邦客户端提供激励所带来的开支。中央计划者为联邦客户端提供的激励在合作系统内部流动，激励大小本身并不计入系统的净收益。我们将系统净收益定义为联邦客户端和中央计划者的收益之和，即联邦客户端的总收益减去中央计划者所支出的激励额。那么，由于联邦客户端收到的激励和中央计划者的激励开支在计算系统净收益时相互抵消，因此系统净收益仅与最后的合作路径相关。我们将最高系统净收益所对应的合作路径简称为最优路径。通过枚举计算可知，$c=0.25$ 对应的最优路径为 111110，而 $c=0.35$ 和 0.45 时的最优路径为 111100。容易看出，这些最优路径的特点是前期连续合作，即所有有益的合作都尽早达成的路径才是最优路径，这一点是符合直觉的。

1．参数 λ 与均衡路径

为便于理解系统收益情况，我们首先展示参数 λ 与均衡路径之间的关系。λ 是强化学习奖励设置中对激励额的惩罚系数，代表对激励额的约束，中央计划者对激励的预算随着 λ 的增大而逐渐紧缩。我们选择 $\lambda=[1, 1.2, 1.4, 1.6, 1.8, 2, 3, 5, 10, 20]$ 这些值进行激励实验。图 4 展示的是投入积累量的路径，可以看到，在三种成本参数下，当 λ 超过某个阈值之后，联邦客户端的投入路径会退回无激励时的投入路径，如图中“-●-”线所示；而当 $\lambda=1$ 时，联邦客户端的投入路径则转向最优路径，如图中“-■-”线所示。当 $c=0.25$ 时，只存在以上两种路径情况。当 $c=0.35$ 时，在两种情况之间出现了一个过渡形态，即 5 次试验中有 2 次选择了最优路径而 3 次选择了初始路径，因此均值结果在两者之间。这是因为，$\lambda=2$ 时，以一定激励换取第 3 期的合作或者放弃该期的激励以节省成本所带来的 RL 奖励较为接近，因此不同的结果都可能产生。当 $c=0.45$ 时，也存在这种情况，在 λ 从 1 变化到 2 的过程中，中央计划者逐渐放弃了在第 3 期进行激励，而直到 λ 增大至 20，第 1 期的激励也被放弃了。

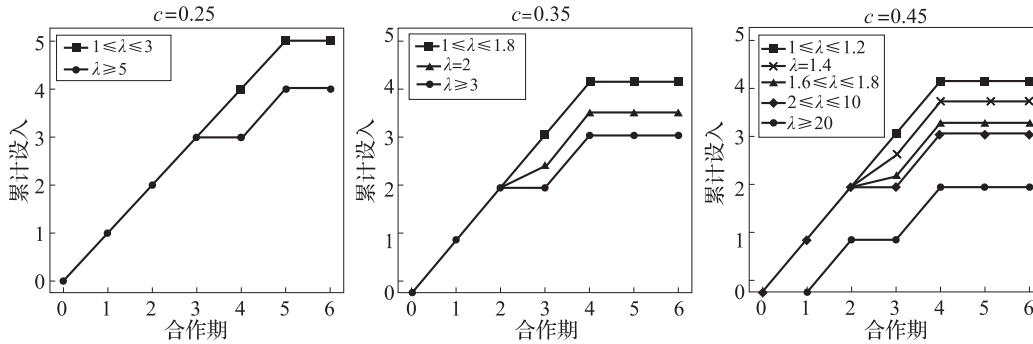


图 4　参数 λ 与均衡路径

注：本图为作者数值实验的结果。由于强化学习结果具有一定的随机性，展示结果为重复实验 5 次取平均值。

2. 参数 λ 与系统收益

在图 5 中，我们采用整柱高度代表激励后联邦客户端总收益的提升量，浅色部分的高度代表中央计划者的激励支出，整柱高度减去浅色部分则代表激励后系统净收益的提升量，图中用深色部分表示。图中折线代表着系统净收益的提升量与激励总额之比，我们称之为激励性价比。

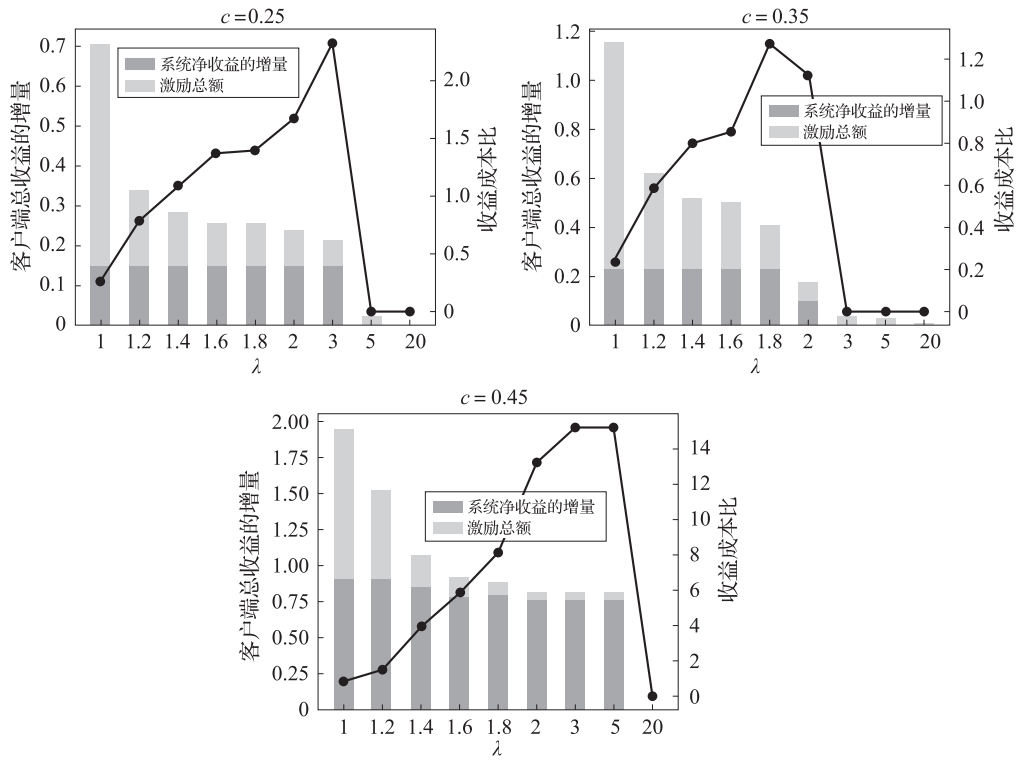


图 5　参数 λ 与系统收益

注：本图为作者数值实验的结果。由于强化学习结果具有一定的随机性，展示结果为重复实验 5 次取平均值。左轴（柱状图）为激励后福利的总增量和净增量，右轴（折线图）为收益成本比，即单位激励产生的净收益提升量。

首先，激励总额和联邦客户端总收益增量随着 λ 参数的增加有减小的趋势。这是因为 λ 越大，代表强化学习的奖励中对激励成本的惩罚项越大，使得中央计划者在支付激励时更加保守。

系统净收益的增量关于 λ 参数单调不增。由于联邦客户端收到的激励和中央计划者的激励开支相互抵消，因此系统净收益仅与最后的合作路径相关，因而随着激励额度的增加，净收益提升呈阶梯性而非连续性的增加。对照图 4 可以看出，当预算约束 λ 在一定范围内增大时，如果合作路径尚未改变，那么系统净收益也不变；而随着 λ 继续增大，激励额进一步收紧，直至中央计划者开始选择放弃对某些时期的激励，合作期数减少，净收益也随之降低。

激励性价比则与 λ 形成先增后减的关系。$\lambda=1$ 时，计划者不受预算约束，进而直接给出大额激励，大大拉低了激励性价比。当 λ 增大到一个阈值，合作路径退化回原均衡路径时，系统净收益为 0，因此激励性价比也降至 0。在 λ 从 1 到这个阈值之间，随着约束收紧，激励总额减少，而系统净收益不变或者减少，激励性价比先增后减，存在一个最具性价比的 λ。对于不同的 c 参数，最具激励性价比的 λ 也是不同的。在 $c=0.25$ 时，一定程度地降低激励预算并未降低净收益，$\lambda=3$ 对应的激励性价比最高。在 $c=0.35$ 时，$\lambda=1.8$ 对应的激励性价比最高，这是因为当 $\lambda=2$ 时，计划者有一定概率放弃激励，导致了净收益的大幅下跌。在 $c=0.45$ 时，$\lambda=3$ 对应的激励性价比最高，尽管此时系统与最高系统净收益略有差距，但激励的开支被大幅减少了。在下文中，我们均以最高性价比的预算约束 λ 对应的激励方案为例进行讨论。

3．最高性价比激励方案的特点

图 6 展示了最高性价比激励方案下的中央计划者策略和联邦客户端策略。其中，实线代表激励额，点线代表联邦客户端的策略路径，点黑线和点灰线分别代表激励前与激励后。可以看到激励额总体与无激励时客户端的参与意愿负相关，而具体激励额大小则由具体情况决定。在客户端本就自发选择投入的合作期，激励额趋近于 0。而对于客户端选择不投入的合作期，激励方案中有两种可能的选择。如果该期合作能够带来较大收益，且促成该期合作不需要很高的激励成本，则选择足以使得联邦客户端在该期投入的激励值，例如 $c=0.25$ 时在第 4 期的激励、$c=0.35$ 时在第 3 期的激励以及 $c=0.45$ 时在第 1 期的激励；反之，如果在此处激励性价比不高，那么结果则是放弃该期的激励，直接归 0。由于数据在模型精度提升中的效用边际递减，越后期的投入所带来的收益越小，直至不足以覆盖参与的成本，此时合作无益。在三种成本下，第 6 期投入都是收益小于成本的，因此没有方案选择在第 6 期激励。又如，$c=0.45$ 时，尽管第 3 期投入的收益是大于成本的，但要促使客户端在该期投入需要花费较大的激励额，这降低了激励性价比，因此该方案同样放弃了第 3 期的激励。总的来说，$c=0.25$ 和 $c=0.35$ 时，最高性价比激励方案使得联邦客户端选择了最优路径；$c=0.45$ 时，最高性价比激励方案使联邦客户端向最优路径偏移，但并未完全达到最优路径。值得一提的是，在 λ 减小到 1.2 时，其对应的激励方案并未放弃该合作期，而是将客户端激励到了最优路径上，尽管其性价比有所降低，但获得了最高的系统净收益。因此，中央计划者可以根据自己的需求和偏好决定合适的 λ 取值，以获得最高的系统净收益或者最高的激励性价比。

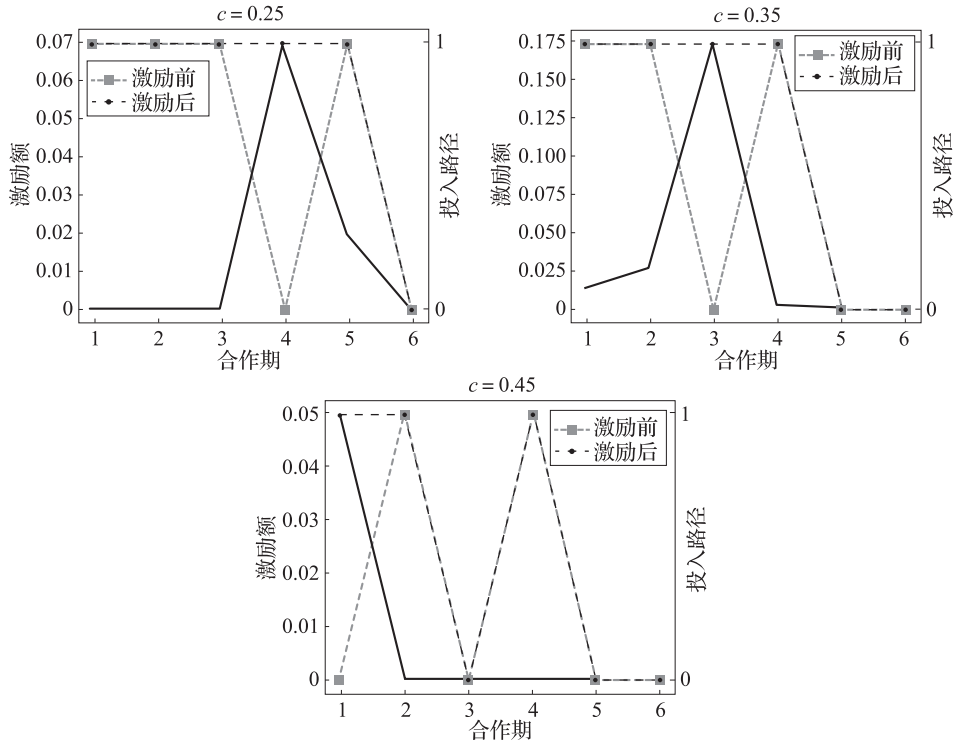


图 6 最高性价比激励方案下的中央计划者策略和联邦客户端策略

注：本图为作者数值实验的结果。由于强化学习结果具有一定的随机性，展示结果为重复实验 5 次取平均值。实线、点黑线和点灰线三种折线分别表示激励额度、激励前合作意愿、激励后合作意愿随合作期数的变化。

4．RL 方案与固定激励方案对比

我们所提出的基于 RL 的激励方案是一种高效的方案，能够以较低的激励开支，换得较大的系统净收益增长。如图 7 所示，对比在每期进行固定激励的方案，基于 RL 的激励方案能有更高的激励性价比。固定激励的方案简单易操作，即每期为投入的联邦客户端提供固定的激励额度。以 $c = 0.35$ 的场景为例，我们选择了四种有代表性的固定激励方案（fixed0 – fixed3），它们分别对应不同的合作路径。如表 1 所示，随着单期激励额度的提升，联邦客户端会逐步增加其投入期数（体现为投入路径序列中“1”的个数增加），但却不一定能换得系统净收益的提升。在该场景下，无论选择何种额度的固定激励方案都无法使客户端的投入路径变为最优路径，这是由这个动态博弈本身的性质所决定的。对于固定激励方案本身来说，如果激励额度过低则无法对博弈结果造成影响；而如果激励额度过高则会导致联邦客户端不计成本地投入，然而随着数据的效用边际递减，系统收益将逐渐无法覆盖投入成本，这也会带来更低的系统净收益。而基于 RL 的激励方案能够根据合作的进程确定动态的激励额度，以便针对性地引导联邦客户端走向最优的投入路径，因而获得了远超固定激励方案的激励性价比。

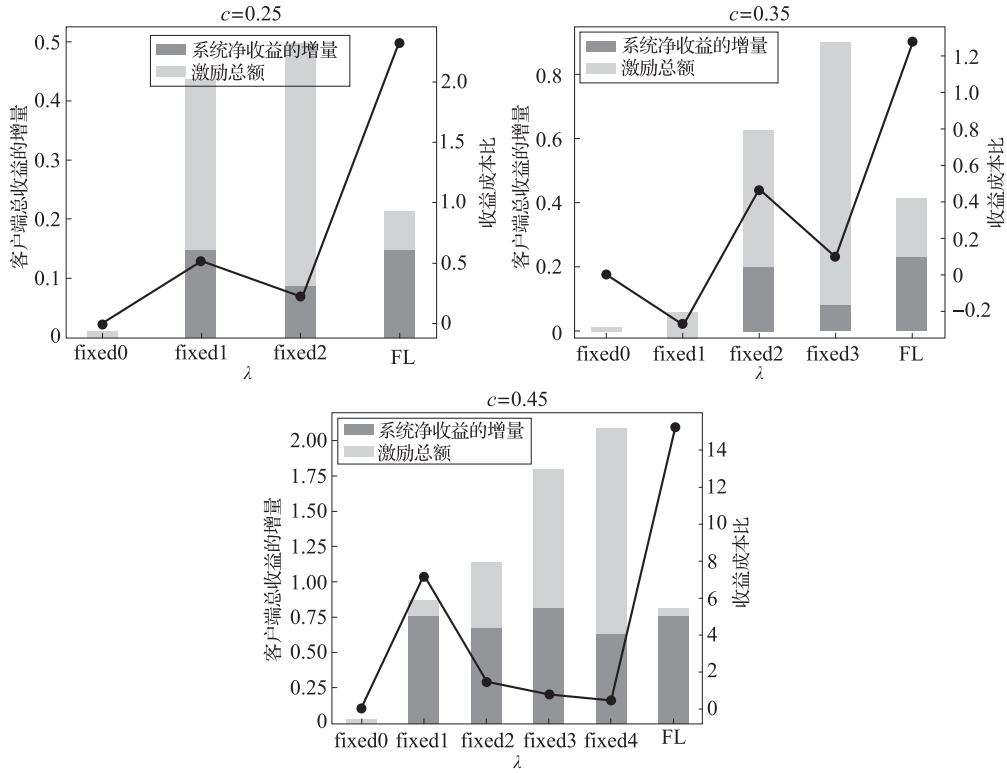


图 7　RL 方案与固定激励方案的激励结果对比

注：本图为作者数值实验的结果。由于强化学习结果具有一定的随机性，展示结果为重复实验 5 次取平均值。左轴（柱状图）为激励后福利的总增量和净增量，右轴（折线图）为收益成本比，即单位激励产生的净收益提升量。

表 1　$c=0.35$ 时各激励方案的结果

方案	单期激励额度	总激励（最低）	投入路径	系统净收益	系统净收益提升	激励性价比
fixed0	[0.00, 0.04)	0	110100	9.660	0	0
fixed1	[0.04, 0.13)	0.079	110110	9.640	−0.020	−0.250
fixed2	[0.13, 0.20)	0.427	111110	9.860	0.200	0.470
fixed3	[0.20, 0.25]	0.819	111111	9.740	0.080	0.097
RL	浮动	0.181	111100	9.890	0.230	1.270

注：fixed0 – fixed3 数据来源于直接穷举计算，RL 数据来源于强化学习结果。投入路径简化表示为 0 – 1序列，例如长度为 6 的 0 – 1 序列“010101”表示客户端在且仅在第 2，4，6 期选择投入。

5．多联邦客户端合作的扩展

上文中，我们以双客户端的合作为主说明了激励方案的相关特点，而本激励方案同样适用于多联邦客户端合作的场景。在多客户端合作的场景下，算法流程无须进行改动，相关的状态与奖励设置细节做对应调整即可。以三个客户端为例，状态

设置需包括三个客户端的历史行为，奖励更改为三个客户端的收益与激励成本的加权之和，另外客户端的策略生成树也由四叉树变为八叉树。在实验中，为减少计算资源的消耗，我们将合作期数设定为 4 期。在不进行激励的情况下，当 $c=0.35$ 时，联邦客户端在前 3 期投入；当 $c=0.45$ 时，联邦客户端仅在第 2、3 期投入。我们分别在这两种成本设置下进行了实验，且 λ 取 1.2 和 2。结果显示，当 $c=0.35$ 时，任何期的激励额均为 0；而当 $c=0.45$ 时，在第 1 期激励至合作发生，其主要结论与双客户端时无异。

六、结论与展望

联邦学习作为一种新兴的机器学习范式，允许参与方在不共享原始数据的情况下进行模型训练，有助于解决信息孤岛问题。然而，长期稳定的联邦合作实践还面临着一些挑战，本文聚焦于联邦学习合作的激励机制设计问题。首先，我们将各方通过联邦学习进行长期数据分析合作过程建模为一个动态博弈过程，该模型使我们能够细致刻画联邦学习客户端的跨期选择。我们采用博弈树来刻画动态博弈策略，并通过逆向递推解出均衡，结果显示，博弈均衡表现出间歇性合作的模式。这种模式延迟了合作进程，造成了系统总收益的损失。为了解决这一问题，我们设计了一种基于深度强化学习算法的激励机制。在这个激励机制中，中央服务器充当中央计划者，通过为联邦学习客户端提供激励，引导参与方转移合作策略至最优合作路径。这一机制的核心思想是通过动态调整激励，使得每个参与方都倾向于采取更加合作的策略，从而提升整个联邦学习系统的净收益。基于模拟实验，本文验证了这一激励方案的有效性。我们发现系统净收益仅与合作路径相关，一定程度地收紧激励预算会引起激励额和联邦客户端总收益的降低，但不一定会引起系统净收益的下降。因此，激励性价比关于激励预算约束先增后减，适当收紧激励预算约束有助于获得更高的激励性价比，但过分收紧预算则无法达到理想的激励效果。最高性价比的激励方案主要是在联邦客户端缺乏合作意愿的合作期进行适当激励，以促进高收益期的合作。同时，在合作收益无法覆盖成本的时期，该方案并不采取激励。与传统的固定激励方案相比，我们设计的基于深度强化学习的激励机制能够基于合作进程动态地确定每一时期的激励额，因而能够在控制激励支出的情况下最大程度地促进合作。

本文的研究为数据共享与隐私保护问题提供了新的视角。在传统的数据合作中，由于隐私和安全的顾虑，数据提供方通常不愿意分享原始数据。而联邦学习作为一种去中心化的学习方式，允许参与方在不共享敏感信息的情况下进行模型训练，从而有效地解决了数据共享的问题。通过提出的激励机制，本文的研究可

以鼓励参与方更积极地参与合作，进一步释放数据的活力。同时，本文设计的基于深度强化学习的激励机制在概念上是创新性的。通过动态调整激励，中央计划者能够更好地适应联邦学习合作的动态变化，使得每个联邦客户端都能在自身利益的基础上做出最优的合作决策。通过模拟实验的验证，我们发现本文设计的激励机制能够在激励性价比上显著优于传统的固定激励方案。这意味着，中央计划者通过支付较少的激励就能使联邦客户端为整个训练提供更多的信息，进而提升模型性能。这不仅提高了合作效率，还使得联邦学习系统更具有可持续性，为长期合作奠定了坚实的基础。本文为联邦学习的组织者提供了一种补贴或奖励转移机制来吸引那些可能因为成本敏感性而犹豫参与的数据所有机构，为促进社会数据潜能的释放提供了强大工具。

本文也存在一些不足之处有待在后续研究中深入探索。首先，可以在强化学习的状态设置中进一步纳入总激励预算。在实际应用中，联邦学习组织方可能会面临有限的激励预算，需要合理分配给各个参与方。通过将这一限制纳入模型，可以更好地指导激励的分配，使得整个联邦学习系统能够更加经济高效地运作。其次，可考虑为联邦学习客户端训练学习和预测未来激励的强化学习智能体。本文主要侧重于中央计划者，也即联邦学习组织方对客户端的激励进行设计，而参与方的反馈和学习过程也是一个重要的方面。通过引入智能体来建模客户端的学习和决策过程，我们可以更好地理解参与方是如何适应激励机制的，从而进一步优化激励设计。

综上，本文不仅在理论上揭示了长期联邦学习合作中的动态博弈特性，也提出了创新的激励机制来提高合作效率，为在当前信息时代背景下更好地实现数据共享和合作提供了新的思路和方法。

参考文献

[1] BAO Y, PENG Y, WU C, 2023. Deep learning-based job placement in distributed machine learning clusters with heterogeneous workloads[J]. IEEE/ACM Transactions on Networking, 31(2): 634-647.

[2] BHOWMICK A, DUCHI J, FREUDIGER J, KAPOOR G, ROGERS R, 2018. Protection against reconstruction and its applications in private federated learning[J]. arXiv Preprint. DOI: 10. 48550/arXiv. 1812. 00984.

[3] BI X, GUPTA A, YANG M, 2023. Understanding partnership formation and repeated contributions in federated learning: an analytical investigation[J]. Management Science. DOI: 10. 1287/mnsc. 2023. 00611.

[4] BOLTON P, DEWATRIPONT M, 2004. Contract theory[M]. Cambridge: MIT Press.

[5] BRISIMI T S, CHEN R, MELA T, OLSHEVSKY A, PASCHALIDIS I, SHI W, 2018. Federated learning of predictive models from federated electronic health records[J]. International Journal of Medical Informatics, 112: 59-67.

[6] CELLINI R, LAMBERTINI L, 2009. Dynamic R&D with spillovers: competition vs cooperation[J]. Journal of Economic Dynamics and Control, 33(3): 568-582.

[7] CHOI T, ROBERTSON P J, 2019. Contributors and free-riders in collaborative governance: a computational ex-

ploration of social motivation and its effects[J]. Journal of Public Administration Research and Theory, 29(3): 394-413.

[8] CONG M, YU H, WENG X, YIU S M, 2020. A game-theoretic framework for incentive mechanism design in federated learning[M]//YANG Q, FAN L, YU H. Federated learning: privacy and incentive. Cham: Springer, 205-222.

[9] CUI K, HAO R, HUANG Y, LI J, SONG Y, 2023. A novel convolutional neural networks for stock trading based on DDQN algorithm[J]. IEEE Access, 11: 32308-32318.

[10] DENG Y, LYU F, REN J, CHEN Y C, YANG P, ZHOU Y, ZHANG Y, 2022. Improving federated learning with quality-aware user incentive and auto-weighted model aggregation[J]. IEEE Transactions on Parallel and Distributed Systems, 33(12): 4515-4529

[11] DING N, FANG Z, HUANG J, 2020. Optimal contract design for efficient federated learning with multi-dimensional private information[J]. IEEE Journal on Selected Areas in Communications, 39(1): 186-200.

[12] FERSHTMAN C, NITZAN S, 1991. Dynamic voluntary provision of public goods[J]. European Economic Review, 35(5): 1057-1067.

[13] FORTUNATO M, AZAR M G, PIOT B, MENICK J, OSBAND I, GRAVES A, MNIH V, MUNOS R, HASSABIS D, PIETQUIN O, BLUNDELL C, LEGG S, 2017. Noisy networks for exploration[J]. arXiv Preprint. DOI: 10. 48550/arXiv. 1706. 10295.

[14] GUPTA R, GUPTA J, 2023. Federated learning using game strategies: state-of-the-art and future trends[J]. Computer Networks, 225. DOI: 10. 1016/j. comnet. 2023. 109650.

[15] HAN X, YU H, GU H, 2019. Visual inspection with federated learning[C]//KARRAY F, CAMPILHO A, YU A. Image analysis and recognition: 16th international conference, ICIAR 2019, Waterloo, oN, Canada, August 27-29, 2019, proceedings, part II. Cham: Springer, 52-64.

[16] HARD A, RAO K, MATHEWS R, RAMASWAMY S, BEAUFAYS F, AUGENSTEIN S, EICHNER H, KIDDON C, RAMAGE D, 2018. Federated learning for mobile keyboard prediction[J]. arXiv Preprint. DOI: 10. 48550/arXiv. 1811. 03604.

[17] HUANG L, SHEA A L, QIAN H, MASURKAR A, DENG H, LIU D, 2019. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records [J]. Journal of Biomedical Informatics, 99. DOI: 10. 1016/j. jbi. 2019. 103291.

[18] KAISSIS G A, MAKOWSKI M R, RÜCKERT D, BRAREN R F, 2020. Secure, privacy-preserving and federated machine learning in medical imaging[J]. Nature Machine Intelligence, 2(6): 305-311.

[19] KESSING S G, 2007. Strategic complementarity in the dynamic private provision of a discrete public good[J]. Journal of Public Economic Theory, 9(4): 699-710.

[20] LI L, YU X, CAI X, XIN H, 2022. Contract-theory-based incentive mechanism for federated learning in health crowdsensing[J]. IEEE Internet of Things Journal, 10(5): 4475-4489.

[21] LIM W Y B, XIONG Z, MIAO C, NIYATO D, YANG Q, LEUNG C, POOR H V, 2020. Hierarchical incentive mechanism design for federated machine learning in mobile networks[J]. IEEE Internet of Things Journal, 7(10): 9575-9588.

[22] LING J, XIA J, ZHU F, GAO C, 2023. DQN-based resource allocation for NOMA-MEC-aided multi-source data stream[J]. EURASIP Journal on Advances in Signal Processing. DOI: 10. 1186/s13634-023-01005-2.

[23] LIU Y, TIAN M, CHEN Y, XIONG Z, LEUNG C, MIAO C, 2023. A contract theory based incentive mechanism for federated learning[M]//RAZAVI-FAR R, WANG B, TAYLOR M E, YANG Q. Federated and transfer learning. Cham: Springer, 117-137.

[24] MCMAHAN H B, MOORE E, RAMAGE D, HAMPSON S, ARCAS B A Y, 2017. Communication-efficient learning of deep networks from decentralized data[R]. International Conference on Artificial Intelligence and Statistics. DOI: 10. 48550/arXiv. 1602. 05629.

[25] MNIH V, KAVUKCUOGLU K, SILVER D, GRAVES A, ANTONOGLOU I, WIERSTRA D, RIEDMILLER M,

2013. Playing atari with deep reinforcement learning[J]. arXiv Preprint. DOI: 10. 48550/arXiv. 1312. 5602.

[26] MNIH V, KAVUKCUOGLU K, SILVER D, RUSU A A, VENESS J, BELLEMARE M G, GRAVES A, RIEDMILLER M, FIDJELAND A K, OSTROVSKI G, PETERSEN S, BEATTIE C, SADIK A, ANTONOGLOU I, KING H, KUMARAN D, WIERSTRA D, LEGG S, HASSABIS D, 2015. Human-level control through deep reinforcement learning[J]. Nature, 518: 529-533.

[27] SAGLAM B, MUTLU F B, CICEK D C, KOZAT S S, 2023. Actor prioritized experience replay[J]. Journal of Artificial Intelligence Research, 78: 639-672.

[28] SAPUTRA Y M, HOANG D T, NGUYEN D N, DUTKIEWICZ E, MUECK M D, SRIKANTESWARA S, 2019. Energy demand prediction with federated learning for electric vehicle networks[R]. 2019 IEEE Global Communications Conference (GLOBECOM). DOI: 10. 1109/GLOBECOM38437. 2019. 9013587.

[29] SCHAUL T, QUAN J, ANTONOGLOU I, SILVER D, 2015. Prioritized experience replay[J]. arXiv Preprint. DOI:10. 48550/arXiv. 1511. 05952.

[30] SHI Z, ZHANG L, YAO Z, LYU L, 2022. FedFAIM: a model performance-based fair incentive mechanism for federated learning[J]. IEEE Transactions on Big Data, (99): 1-13.

[31] SUN P, CHE H, WANG Z, WANG Y, 2021. Pain-FL: personalized privacy-preserving incentive for federated learning[J]. IEEE Journal on Selected Areas in Communications, 39(12): 3805-3820.

[32] TU X, ZHU K, LUONG N C, NIYATO D, 2022. Incentive mechanisms for federated learning: from economic and game theoretic perspective[J]. IEEE Transactions on Cognitive Communications and Networking, 8(3): 1566-1593.

[33] VAN HASSELT H, GUEZ A, HESSEL M, MNIH V, SILVER D, 2016. Learning values across many orders of magnitude[R]. Proceedings of the 30th International Conference on Neural Information Processing Systems. DOI: 10. 48550/arXiv. 1602. 07714.

[34] WANG W, MEMON F H, LIAN Z, YIN Z, GADEKALLU T R, PHAM V Q, DEV K, SU C, 2021. Secure-enhanced federated learning for AI-empowered electric vehicle energy prediction[J]. IEEE Consumer Electronics Magazine, 12(2): 27-34.

[35] WANG Y, SU Z, ZHANG N, BENSLIMANE A, 2020. Learning in the air: secure federated learning for UAV-assisted crowdsensing[J]. IEEE Transactions on Network Science and Engineering, 8(2): 1055-1069.

[36] WU L, GUO S, LIU Y, HONG Z, 2022. Sustainable federated learning with long-term online VCG auction mechanism[R]. 2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS). DOI: 10. 1109/ICDCS54860. 2022. 00091.

[37] XU X, LYU L, MA X, MIAO C, FOO C, LOW B K, 2021. Gradient driven rewards to guarantee fairness in collaborative machine learning[J]. Neural Information Processing Systems, 34: 16104-16117.

[38] YANG Q, LIU Y, CHENG Y, KANG Y, CHEN T, YU H, 2020. Federated learning[M]. Berlin : Springer.

[39] YAP Y J, LUCKRAZ S, TEY S K, 2014. Long-term research and development incentives in a dynamic Cournot duopoly[J]. Economic Modelling, 39: 8-18.

[40] YILDIRIM H, 2006. Getting the ball rolling: voluntary contributions to a large-scale public project[J]. Journal of Public Economic Theory, 8(4): 503-528.

[41] YU H, LIU Z, LIU Y, CHEN T, 2020. A fairness-aware incentive scheme for federated learning[R]. AAAI/ACM Conference on AI, Ethics, and Society. DOI: 10. 1145/3375627. 3375840.

[42] YU H, YANG S, ZHU S, 2019. Parallel restarted SGD with faster convergence and less communication: demystifying why model averaging works for deep learning[R]. AAAI Conference on Artificial Intelligence. DOI: 10. 1609/aaai. v33i01. 33015693.

[43] ZEMZEM W, TAGINA M, 2023. Improving exploration in deep reinforcement learning for stock trading[J]. International Journal of Computer Applications in Technology, 72(4): 288-295.

[44] ZENG R, ZENG C, WANG X, LI B, 2022. Incentive mechanisms in federated learning and a game-theoretical ap-

proach[J]. IEEE Network, 36(6): 229-235.

[45] ZENG R, ZHANG S, WANG J, CHU X, 2020. FMore: an incentive scheme of multi-dimensional auction for federated learning in MEC[R]. Singapore: 2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS). DOI: 10. 1109/ICDCS47774. 2020. 00094.

[46] ZHAN Y, LI P, QU Z, ZENG D, 2020. A learning-based incentive mechanism for federated learning[J]. IEEE Internet of Things Journal, 7(7): 6360-6368.

[47] ZHANG J, CHEN B, CHENG X, BINH H T T, YU S, 2020. PoisonGAN: generative poisoning attacks against federated learning in edge computing systems[J]. IEEE Internet of Things Journal, 8(5): 3310-3322.

[48] ZHANG N, MA Q, CHEN X, 2022. Enabling long-term cooperation in cross-silo federated learning: a repeated game perspective[J]. IEEE Transactions on Mobile Computing, 22(7): 3910-3924.

Incentive Mechanism Design for Sustainable Federated Learning Based on Reinforcement Learning

Qiuyuan Ai
(College of Engineering, Peking University)

Zhijian Zhan
(Academy for Advanced Interdisciplinary Studies, Peking University)

Cong Wang *
(Guanghua School of Management, Peking University)

Jie Song
(College of Engineering, Peking University)

Summary: As Internet, Internet of Things (IoT), and Artificial Intelligence (AI) technologies rapidly evolve, data has become a critical driving force behind economic and technological advancement. Companies can leverage data analysis to gain comprehensive insights into customer behavior, market trends, and operational performance, thereby making informed decisions and enhancing overall performance. However, a single organization's data may not be sufficient for comprehensive data analysis, posing a significant challenge. For instance, developing an accurate marketing model to target users may necessitate data from multiple sources, such as telecom operators, social networking sites, and e-commerce platforms. This data scarcity necessitates data-sharing mechanisms, which are often fraught with concerns surround data privacy, ethics, and legality. In this regard, Federated Learning (FL)—a novel machine learning paradigm—has garnered increasing attention. FL participants can train local models, safeguard data privacy, and exchange only model parameters with servers or other peers, fully capitalizing on the value of data. This "data-available-but-not-visible" approach is gaining popularity in data-intensive fields.

Many FL tasks cannot be accomplished in a single instance and require sustained collaboration among multiple parties. For example, in the joint development of an FL model across multiple medical institutions to detect and manage chronic diseases, continuous accumulation of clinical data, learning from case changes, and model robustness and predictability improvements are necessary to reflect the latest medical knowledge and practices. Current literature on FL cooperative behavior and incentive mechanisms, however, primarily focuses on cross-device federated learning and considers only one-off cooperation. This modeling is inadequate for characterizing practical cross-silo long-term FL patterns. On the one hand, cross-silo FL participants, who also accumulate a certain amount of data, have more complex and diverse strategic options compared to those in cross-device FL. Participants can choose to participate in public training or solely improve their

* Corresponding Author: Cong Wang, Guanghua School of Management, Peking University, E-mail: wangcong@gsm.pku.edu.cn.

model utility through local training. On the other hand, when cooperation transitions from a one-off to a long-term scenario, time inconsistency issues may lead to free-riding behaviors, incentivizing participants to delay data contributions while enjoying the benefits of others' contributions. To address these limitations, this study concentrates on the long-term cross-silo FL process, establishing a dynamic game model to characterize federated clients' interactive strategies and proposing a reinforcement learning-based incentive mechanism to encourage rational participant contribution, aiming to boost the FL system's overall revenue.

This paper first establishes a dynamic game model to characterize federated clients' long-term interactive strategies. We devise a cooperation contract in which the central server only transmits the aggregated parameters to current training period contributors. With the long-term cross-silo FL cooperation process divided into several model training periods, clients have two strategic choices in each period: to participate in public federated training or to retain data for local training only. At the end of each period, clients receive feedback parameters from the central server and gain corresponding benefits based on their local models' accuracy. In this framework, clients face a trade-off between participation costs and potential early contribution benefits. Given the information accumulation in the model with the client's input, clients also confront a cross-period decision-making problem regarding resource allocation throughout the entire long-term FL cooperation process. Based on these background assumptions, this paper establishes a game tree to consider the game solution, where clients' decisions in each training period are based on full knowledge of past cooperation and rational expectations of future actions. Through backward induction, we solve for the client's equilibrium strategy, which exhibits intermittent contribution gaps, clearly deviating from the socially optimal cooperative pattern.

Building on the above game analysis, this paper subsequently designs a dynamic incentive scheme based on reinforcement learning, setting incentives for different training periods based on clients' cooperation progress. Firstly, we regard the FL organization as a central planner responsible for issuing incentives before each training period to encourage federated client input. The Deep Reinforcement Learning (DRL) agent assists the central planner in making incentive decisions, with federated clients serving as the environment with which the agent interacts. On the one hand, we meticulously design the state, action, and reward of the DRL method to fully encompass the information of the federated learning cooperation process. On the other hand, we introduce enhancements to the traditional Deep Q-Network (DQN) method, such as Double Deep Q-Network (DDQN), prioritized replay, and noisy network, to augment the method's performance. Through extensive experiments, we verify the scheme's effectiveness in improving the system's total revenue and controlling incentive costs. Reasonable incentive cost penalties can guide the DRL agent towards the most cost-effective incentive scheme, accurately incentivizing low-willingness cooperation periods of clients, and the system revenue under the same budget significantly surpasses that of fixed incentives.

This paper not only theoretically uncovers the dynamic patterns in long-term cross-silo federated learning cooperation but also proposes innovative incentive mechanisms to enhance cooperation efficiency, offering fresh insights and methodologies for effectively facilitating data sharing and cooperation in the contemporary information era.

Keywords: Data Sharing; Federated Learning; Incentive Mechanism; Stable Cooperation

JEL Classification: C45; C72