arXiv:2305.03097v2 [cs.LG] 4 Oct 2024

联邦集成引导离线强化学习

Desik Rengarajan* **Nitin Ragothaman DileepKalathil Srinivas Shakkottai**电气与计算机工程系，德克萨斯农工大学

摘要

我们研究联邦离线强化学习（RL）问题。在这一场景中，分布式学习智能体必须仅利用按照不同且未知的行为策略生成的小规模预收集数据集，协作学习高质量的控制策略。简单地将标准离线强化学习方法与标准联邦学习方法结合起来解决这一问题，可能会导致策略表现不佳。为此，我们提出联邦集成引导离线强化学习算法（FE-DORA），利用集成学习方法提炼客户端的集体智慧。我们开发了 FEDORA 代码库，以便在联邦学习平台上利用分布式计算资源。我们表明，在各种复杂的连续控制环境和真实世界数据集上，FEDORA 的表现显著优于其他方法，包括基于合并后数据池进行的离线强化学习。最后，我们在移动机器人上展示了 FEDORA 在真实世界中的性能。我们的代码和实验视频见 https://github.com/DesikRengarajan/FEDORA。

1 引言

联邦学习是一种协作式学习方法：客户端将本地训练的模型（而非数据）分享给联邦协调方，协调方定期整合这些模型，并将联邦模型返回给客户端以供进一步优化 [10, 30]。联邦学习近期在监督学习应用中取得了成功，因为它能够利用每个客户端的少量数据训练出性能良好的模型，同时保护隐私并减少通信资源的使用。人们也开始关注将联邦学习用于在线 强化学习（RL）：客户端通过与各自环境进行连续交互来学习，并在客户端之间联邦化学习到的策略 [12, 21, 23]。然而，与真实世界系统进行此类在线交互往往不可行，而且每个客户端可能仅拥有按照客户端特定行为策略收集的预先采集运行数据。联邦离线 RL的核心问题在于：如何仅利用客户端上由异质策略收集的此类离线数据，在不实际共享任何数据的情况下学习最优策略。

离线强化学习算法 [17]，（如 CQL [15] 和 TD3-BC [5] ）采用演员-评论家学习方法，仅利用各客户端现有的数据集。然而，在我们的场景中，将这种方法应用于客户端的许多小型数据集，会在各客户端产生质量各异（且未知）的策略集成，以及准确度不一的相应评论家。我们将看到，使用 FedAvg [20] 等标准联邦方法，直接联邦聚合此类通过离线强化学习训练的策略和评论家，可能会得到比各个组成策略更差的策略。因此，我们指出联邦离线强化学习面临以下基本挑战：（i）集成异质性：异构的客户端数据集会产生性能水平各异的策略。关键在于汲取策略集成的集体智慧，而非对它们取平均值。（ii）悲观价值

*通讯作者。电子邮箱：desik.29@gmail.com

第 38 届神经信息处理系统大会（NeurIPS 2024）。

计算：离线强化学习采用悲观方法来计算数据中缺乏充分表征的动作的价值，以尽量减少分布偏移（从而降低采取这些动作的概率）。然而，联邦聚合必须积极地从评论家集成中提取最高价值（并由此促进高价值动作）。（iii）数据异质性：与其他联邦学习场景一样，在联邦轮次之间，基于各客户端异构数据执行多个本地梯度步，可能会导致模型产生偏差。我们必须对本地策略进行正则化，以减少这种偏移。

在这项工作中，我们提出了联邦集成引导离线强化学习算法（FEDORA），该算法通过协作生成高质量的控制策略和评论家函数。FEDORA 仅使用本地数据（其质量未知）来评估客户端策略的表现，并在每一轮联邦过程中生成各组成策略的加权组合，使整体目标最大化，同时通过权重的熵进行正则化。联邦化客户端评论家时也采用相同的方法。以这种方式遵循最大熵原则，能够生成联邦策略和评论家，从集成中提取集体智慧。为此，FEDORA 以集成学习的方式，根据各客户端策略的相对优值构建联邦策略和评论家。FEDORA 确保联邦评论家和每个客户端的本地评论家在评估时保持乐观，从而设定具有挑战性的目标供训练使用。它通过同时依据联邦策略和本地数据集对客户端策略进行正则化，来应对数据异质性。最后，FEDORA 根据数据集所能生成的策略质量，逐步降低对该数据集的依赖，从而削弱无关数据的影响。据我们所知，尚无其他工作系统地识别离线联邦强化学习的这些根本挑战，也没有工作设计方法来明确应对其中的每一项挑战。

我们提出了一个框架，可在单个系统上或分布式计算资源上实现 FEDORA。我们在多种 MuJoCo环境和真实世界数据集上评估 FEDORA，并表明其性能优于其他几种方法，包括在汇总数据集上进行离线强化学习。我们还通过在 TurtleBot 机器人 [1] 上开展真实世界实验，展示了 FEDORA 的出色性能。

2 相关工作

离线强化学习：离线强化学习旨在从由行为策略生成的固定数据集 [17] 中学习策略。离线强化学习方法面临的一项关键挑战是分布偏移问题：学习策略的状态-动作访问分布可能不同于生成离线数据的行为策略所对应的分布。众所周知，这种分布偏移可能导致学习策略表现不佳 [17]。离线强化学习算法解决这一问题的一种常见方法是通过对演员或评论家进行正则化，学习一个接近生成数据的行为策略的策略 [5, 7, 16, 14, 33]。一些离线强化学习算法会在整个数据集或其子集上执行加权的行为克隆或模仿学习 [31, 22, 3]。 [38, 37] 提出了数据再平衡方法，旨在优先选择高回报转移；这些方法可与离线强化学习算法结合，以缓解异构数据场景中的分布偏移问题。

联邦学习： [20] 提出了 FedAvg，这是一种联邦策略，客户端在不共享数据的情况下协作学习联合模型。在 [25] 中提出了 FedAvg 的推广版本。联邦学习中的一个关键问题是数据异质性，即客户端拥有非独立同分布数据，这会导致收敛不稳定且缓慢 [30, 11, 18]。为解决数据异质性问题， [18]提出了 FedProx，这是 FedAvg 的一种变体，其中引入了近端项，以减少本地模型与服务器模型之间的偏差。

联邦强化学习：近期，联邦学习已扩展至在线强化学习领域。 [12] 分析了联邦表格型 Q 学习的性能。 [23] 将传统在线强化学习算法与 FedAvg 结合，应用于多个场景。一些研究提出，根据性能指标调整 FedAvg 的加权方案，例如乒乓球游戏中的回合长度 [21] 或过去 10 个训练回合的平均回报 [19] ，以提升性能或实现个性化。 [32]提出了一种方法，通过关注客户端的性能指标（如平均奖励、平均损失和命中率）来计算权重，并将其用于边缘缓存应用。 [8] 使用 Transformer 编码器学习在线强化学习场景中智能体之间的上下文关系。 [9] 提出了一种替代的联邦方法，通过奖励塑形在客户端之间共享信息。 [34]提出了基于 KL 散度的本地策略与全局策略正则化方法，以解决在线强化学习场景中的数据异质性问题。

在离线强化学习场景中， [39]通过采用基于线性马尔可夫决策过程构建的多站点马尔可夫决策过程模型来形式化离线联邦学习，并提出了联邦动态治疗方案算法。然而，这种方法需要先完成本地训练，然后只进行一步联邦平均。与这项工作不同，我们的方法不假设马尔可夫决策过程是线性的；这一假设会限制许多现实问题中的应用。此外，我们采用标准的联邦学习理念，定期进行联邦聚合，并在其后执行多轮本地更新。据我们所知，我们是首个为数据异构的客户端提出通用联邦离线强化学习算法的工作。

3 预备知识

联邦学习： 联邦学习的目标是最小化以下目标函数 ，

$$F(\theta) = \mathbb{E}_{i\sim\mathcal{P}}\left[F_i(\theta)\right], \tag{1}$$

其中， θ 表示联邦（服务器）模型的参数， F_i 表示客户端 i 的本地目标函数， $\mathcal{P}$ 表示客户端集合 $\mathcal{N}$ 上的分布。FedAvg 算法 [20]是以联邦方式求解公式 (1) 的一种常用方法。FedAvg 将训练过程划分为多个轮次，在每轮开始时 t，服务器将其当前模型 θ^t 广播给所有客户端，每个客户端将其当前本地模型初始化为当前服务器模型。客户端在各自的数据集 $\mathcal{D}_i$ 上执行多次本地更新，以获得更新后的本地模型 θ_i^t。随后，服务器根据各客户端本地数据集的大小对这些本地模型进行加权平均，从而获得下一轮联邦学习所使用的服务器模型θ^{t+1} ，即

$$\theta^{t+1} = \sum_{i=1}^{|\mathcal{N}|} w_i\theta_i^t, \quad w_i = \frac{|\mathcal{D}_i|}{|\mathcal{D}|}, \quad |\mathcal{D}| = \sum_{i=1}^{|\mathcal{N}|}|\mathcal{D}_i|. \tag{2}$$

强化学习：我们使用马尔可夫决策过程（MDP）框架对 RL 进行建模，该框架表示为一个元组 $(\mathcal{S}, \mathcal{A}, R, P, \gamma, \mu)$，其中 $\mathcal{S}$是状态空间， $\mathcal{A}$是动作空间， R： $\mathcal{S}\times\mathcal{A}\to\mathbb{R}$是奖励函数， P： $\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$ 表示转移概率函数，即在状态 s下执行动作 a转移到状态 s'的概率， γ是折扣因子， μ是初始状态 s_0的分布。策略 π是将状态映射到动作（确定性策略）或动作上的分布（随机策略）的函数。RL 的目标是最大化策略 π的无限时域折扣奖励，其定义为 $J(\pi) = \mathbb{E}_{\pi,P,\mu}[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)]$，即执行策略 π所获得的期望累积折扣奖励。策略 π 在状态 s下执行动作 a 时的状态-动作价值函数（或 Q 函数）是采取动作 a 处于状态 s 并在此后遵循策略 π 所获得的期望累积折扣奖励：$Q^\pi(s,a) = \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)|s_0=s, a_0=a\right]$。

离线强化学习： 离线强化学习的目标是仅使用由行为策略 π_b 收集的静态转移数据集 $\mathcal{D}$（s， a， r， s'）来学习策略 π，而不与环境进行任何额外交互。离线强化学习算法通常会针对行为策略进行某种正则化，以确保学习策略不会偏离行为策略。这样做是为了防止分布偏移；这是离线强化学习中的一个重大问题，因为学习策略与行为策略之间的差异可能导致对数据集 [16, 17]中未出现的状态-动作对进行错误的 Q 值估计。

我们的方法与大多数离线强化学习算法兼容，例如 CQL [15] 或 TD3-BC [5]。我们选择 TD3-BC 进行说明，原因是它简单易用，并且在基准问题上具有出色的实证表现。TD3-BC算法是TD3算法 [6]的行为克隆（BC）正则化版本。TD3-BC中的策略通过TD3目标函数与行为克隆损失的线性组合进行更新，其中TD3目标函数确保策略得到改进，而BC损失则防止分布偏移。更具体地说，TD3-BC目标函数可写为

$$\pi = \arg\max_{\pi} U_{\mathcal{D}}(\pi), \tag{3}$$

$$\text{where} \quad U_{\mathcal{D}}(\pi) = \mathbb{E}_{s,a\sim\mathcal{D}}\left[\lambda Q^\pi(s,\pi(s)) - (\pi(s)-a)^2\right], \tag{4}$$

且 λ 是一个超参数，用于确定行为克隆项的相对权重。

4 联邦离线强化学习

在现实世界的离线强化学习应用中，数据通常来自多个具有不同（未知）专业水平的智能体（客户端）所采用的策略。客户端通常不愿共享数据。我们的目标是仅利用此类离线数据，学习底层强化学习问题的最优策略，同时不让客户端了解其数据质量，也不让他们彼此或与服务器共享数据。此外，客户端和服务器都无法访问底层模型或环境。我们将客户端集合记为 $\mathcal{N}$。每个客户端 $i \in \mathcal{N}$都拥有离线数据集 $\mathcal{D}_i = \{(s_j,\ a_j,\ r_j,\ s'_j)_{j=1}^{m_i}\}$
根据行为策略 π_i^b生成。我们假设所有客户端的底层马尔可夫决策过程模型 P和奖励函数 $R(\cdot,\cdot)$均相同，而离线数据集 $\mathcal{D}_i$之间的统计差异仅由收集数据时所使用的行为策略 π_i^b不同所致。

在 FedAvg 等标准联邦学习算法中，每个客户端会先进行多次参数更新，然后再将参数发送给服务器。已知，在联邦学习中执行多次本地更新能够显著降低通信成本，同时不影响收敛解的最优性 [10, 30]。在联邦离线强化学习中，由于每个客户端都必须利用本地离线数据 $\mathcal{D}_i$执行多步策略评估和策略更新，因此，考虑与标准离线强化学习算法目标一致的客户端目标函数是合理的。我们选用 TD3-BC 算法 [5], 使用的目标函数，即式（3） $U_{\mathcal{D}_i}$ 中给出的目标函数，作为客户端目标函数。作出这一选择，是因为 TD3-BC 目标函数简单，并且已在多种环境中取得了良好的实证效果。类似于式（1）中给出的标准联邦学习目标，我们现在可以将联邦离线强化学习目标定义为

$$U(\pi_{\text{fed}}) = \sum_{i=1}^{|\mathcal{N}|} w_i U_{\mathcal{D}_i}(\pi_{\text{fed}}), \tag{5}$$

其中 w_i 为待确定的权重。

一种在不共享数据的情况下利用不同用户经验的方法，是将现有的联邦学习技术与离线强化学习算法相结合。这种天真的联邦策略是否足以协同学习出优秀的联邦策略？此外，联邦是否真的有必要？本节旨在了解联邦离线强化学习面临的挑战，并设计一种算法框架来应对这些挑战。

我们首先通过示例说明设计联邦离线强化学习算法时面临的问题。我们考虑 MuJoCo 中的 Hopper环境 [28], , 其中有 $|\mathcal{N}| = 10$、 $|\mathcal{D}_i| = 5000$，并使用 D4RL数据集 [4] 中的数据。不过，我们不使用所有客户端由同一策略生成的数据，而是考虑这样一种设置：五个客户端使用 hopper-expert-v2数据集中的数据（该数据集由经过充分训练的（专家）SAC策略生成），另外五个客户端使用 hopper-medium-v2数据集中的数据（该数据集由经过部分训练的（中等）策略生成，其性能仅达到专家策略的三分之一）。客户端和服务器都不知道数据的质量（专家或中等）。图1展示了多种算法的性能对比，其中均值和标准差是根据 4 个随机种子计算得出的。

合并所有数据（集中式）： 在监督学习中，将数据集中起来进行学习是理想的情形。然而，如图 1 所示，在离线强化学习中，使用由不同的行为策略生成的合并数据进行集中式训练可能会适得其反。这与 [36] 的结论一致：汇集来自专业水平不同的行为策略的数据，会加剧学习策略与各数据集之间的分布偏移，导致性能不佳。在其他离线强化学习文献 [5, 16] 中，也观察到了因合并数据而导致的类似性能下降。我们还探索了采用数据再平衡的集中式算法，并观察到 FEDORA 仍然表现更优（见附录 B.9）。

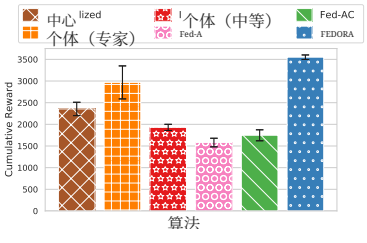


图1：联邦式和集中式离线强化学习算法的性能对比。

我们想进一步强调，在联邦设置中，合并所有客户端的数据只是假设性的基础，因为数据分布在各个客户端之间，无法合并。

个体离线强化学习： 在此，智能体使用各自的数据集进行离线强化学习，而不与其他智能体合作。在图 1 中，我们观察到，拥有专家数据或中等数据的客户端都无法很好地学习，并且

表现出较大的标准差。这一现象可能是因为没有任何客户端拥有足够的数据来学习良好的策略。

朴素联邦离线强化学习：一种简单的联邦方法是将离线强化学习目标用作本地目标，并应用 FedAvg（式（2））。然而，离线强化学习算法通常由两个部分组成——演员和评论家。事先并不清楚应联邦化哪个部分，因此我们开展实验，分别仅联邦化演员（Fed-A）或同时联邦化演员和评论家（Fed-AC）。令人意外的是，如图 1 所示，这些朴素策略得到的联邦策略表现不如个体离线强化学习。

4.1 联邦离线强化学习面临的问题

Our 该示例说明，在设计可行的联邦 e
fede离线强化学习算法时，必须解决若干根本性问题，包括：

1. 集成异质性：在异构数据上进行离线强化学习，会得到一组质量各异的策略。关键在于利用这些不同策略所包含的信息，而不是简单地对它们取平均。然而，在每个客户端进行一步本地梯度更新后，再像 FedAvg 那样使用权重进行联邦聚合，$w_i = |\mathcal{D}_i|/|\sum_{i=1}^{|\mathcal{N}|} |\mathcal{D}_i|$，等价于使用所有客户端的合并数据集求解离线强化学习问题 [30]。如图 1 所示，由此产生的分布偏移会导致性能不佳。我们应如何对客户端学到的策略集成进行最优联邦聚合？

2. 悲观价值计算： 大多数离线强化学习算法都会针对离线数据加入悲观项，以尽量减少分布偏移。仅使用本地数据和这一悲观项来训练客户端评论家，可能会使其对在自身数据集中代表性不足、但在其他客户端数据中代表性良好的动作持悲观态度。我们如何有效利用联邦评论家与本地计算的评论家，为每个客户端的离线强化学习设定有进取性的目标点？

3. 数据异质性： 联邦学习要求各客户端在联邦训练前执行多个本地梯度步，以提高通信效率。然而，过多的训练轮次会使客户端的本地模型偏向其数据集。这种客户端漂移现象在联邦（监督）学习中广为人知，并可能导致策略无法达到全局最优。反过来，这可能导致联邦策略的表现不如仅使用客户端数据进行本地训练，如图 1 所示。我们应如何对本地策略进行正则化以防止这种情况？

5 FEDORA 设计方法

我们希望开发一种联邦集成引导离线强化学习算法（FEDORA），以系统地解决第 4 节中概述的问题。我们的方法由三个基本要求驱动。首先，客户端共同拥有一组质量各异（且未知）的本地策略集成，而服务器必须在联邦过程中利用其中蕴含的集体知识。其次，必须使用依赖本地数据进行策略评估的评论家集成来评估这些策略的质量。最后，每轮联邦结束后，客户端必须通过离线强化学习更新其本地策略，同时利用本地数据和收到的联邦策略。

使用 FedAvg 最大化式 (5) 中的联邦离线强化学习（RL）目标，会按照式 (2) 设置权重，即根据每个客户端数据集的大小对其贡献进行加权。这等同于使用所有客户端的合并数据集来解决离线强化学习问题。然而，这种方法会加剧影响离线强化学习算法的分布偏移问题，导致性能不佳。 [36] 从分析和实证两方面验证了这一问题。我们在图 1 中展示了这一现象：基于汇总数据进行离线强化学习会得到次优策略。 [36] 建议通过识别可能带来策略改进的样本来谨慎共享数据。然而，在联邦离线强化学习设置中，我们无法在客户端之间共享任何数据。

我们的解决方案是遵循最大熵原则，选择最能体现当前对各客户端策略相对优值认知的权重。为避免权重集中于少数当前表现最佳的客户端，我们对权重添加熵

reg正则化，温度参数为 β，由此得到以下目标函数：

$$U(\pi_{\text{fed}}) = \sum_{i=1}^{|\mathcal{N}|} w_i U_{\mathcal{D}_i}(\pi_{\text{fed}}) - \frac{1}{\beta} \sum_{i=1}^{|\mathcal{N}|} w_i \log w_i. \tag{6}$$

随后，我们可以通过拉格朗日对偶方法证明，该目标在以下情况下达到最大值

$$w_i = \frac{e^{\beta U_{\mathcal{D}_i}(\pi_{\text{fed}})}}{\sum_{i=1}^{|\mathcal{N}|} e^{\beta U_{\mathcal{D}_i}(\pi_{\text{fed}})}}. \tag{7}$$

基于熵正则化目标所提出的这类 soft-max 权重，我们现在设计 FEDORA，以满足上述三个要求中的每一项。

在下文中，$\pi_i^{(t,k)}$ 表示联邦的第 t 轮中客户端 i 经过 k 步本地策略更新后的策略。由于所有客户端都会在每轮联邦开始时将其本地策略初始化为联邦策略，因此对于每个客户端 i，都有 $\pi_i^{(t,0)} = \pi_{\text{fed}}^t$。我们还记 $\pi_i^t = \pi_i^{(t,K)}$，其中 K 是本地更新的最大次数。由于所有客户端都会将其本地评论家初始化为联邦评论家，因此对于本地评论家，我们同样可以定义 $Q_i^{(t,k)}$、$Q_i^{(t,0)} = Q_{\text{fed}}^t$和 $Q_i^t = Q_i^{(t,K)}$。

5.1 基于客户端策略的集成导向学习

我们首先需要一种近似 $U_{\mathcal{D}_i}(\pi_{\text{fed}})$ 的方法，以便确定方程（7）中所示的客户端 i的权重 w_i。我们利用最终本地策略 $J_i^t = \mathbb{E}_{s\sim\mathcal{D}_i}\left[Q_i^t(s,\ \pi_i^t(s))\right]$,的性能，该性能同样刻画了客户端 i处的相对性能，并将其作为 $U_{\mathcal{D}_i}(\pi_{\text{fed}})$。的代理指标。这里，$Q_i^t$是第 t 轮经过 K 次本地更新后的本地评论家函数。在离线强化学习中，很难直接获得这样一个无偏的本地评论家 Q_i^t，因为我们无法访问环境来执行策略并评估其性能。我们将在后文介绍计算 Q_i^t 和 π_i^t 的方法。本地估计值 J_i^t的准确性高度依赖于 i处可用的数据样本数量，因此，按照联邦平均的通常做法，我们在计算权重时需要考虑数据集 $|\mathcal{D}_i|$的规模。因此，客户端权重和联邦策略更新为

$$w_i^t = \frac{e^{\beta J_i^t}|\mathcal{D}_i|}{\sum_{i=1}^{|\mathcal{N}|} e^{\beta J_i^t}|\mathcal{D}_i|}, \quad \pi_{\text{fed}}^{t+1} = \sum_{i=1}^{|\mathcal{N}|} w_i^t \pi_i^t. \tag{8}$$

5.2 评论家训练的联邦乐观机制

我们的算法中的评论家发挥着两大作用。首先，每个客户端进行策略更新时的离线强化学习需要利用本地数据进行策略评估。其次，每轮联邦过程中，评论家进行的策略评估决定了客户端 i 的本地策略在集成学习中的权重w_i^t t。我们希望每个客户端的本地评论家既能利用所有客户端的评论家集成所蕴含的知识，又能针对用于策略评估的本地数据进行调优。基于离线数据的评论家会受到外推误差的影响，因为对于本地数据集中未出现过的状态-动作对，评论家会给出错误估计，这会严重影响联邦离线强化学习中的演员-评论家式策略更新。由于联邦策略是由本地策略集合推导而来，因此它可能会采取任何客户端的本地数据集中都未出现过的动作。若每轮通信开始时都将本地策略初始化为联邦策略，这个问题会更加严重。我们引入联邦乐观机制这一概念来训练本地评论家，使评论家能够借鉴群体智慧并受到鼓励而采取乐观估计。我们通过两个步骤实现这种联邦乐观机制。

首先，我们采用集成导向的评论家联邦，根据客户端 i 在轮次 t 的本地评论家的优值对其进行加权，从而计算出联邦评论家：

$$Q_{\text{fed}}^{t+1} = \sum_{i=1}^{|\mathcal{N}|} w_i^t Q_i^t. \tag{9}$$

这种熵正则化平均确保了策略良好的客户端评论家能够显著影响联邦评论家。

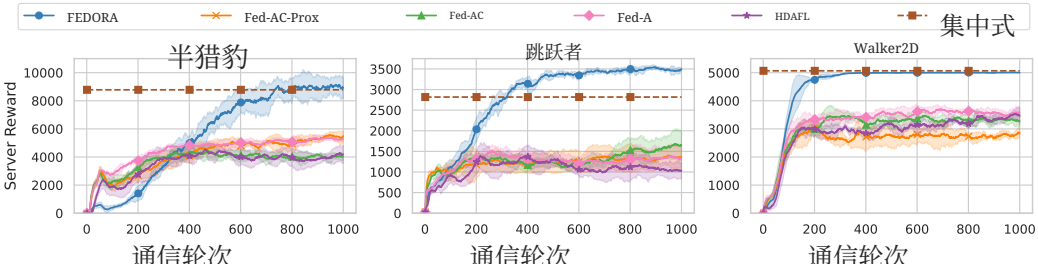


图2：在不同 MuJoCo环境中对算法的评估。

其次，对于本地评论家更新，我们将目标值设为本地评论家和联邦评论家中的较大值，即 $\tilde{Q}_i^{(t,k)}(s,a) = \max\left(Q_i^{(t,k)}(s,a)\ Q_{\text{fed}}^t(s,a)\right)$，其中$\tilde{Q}_i^{(t,k)}(s,a)$ 是联邦的第 t^{th} 轮在进行 k 次本地评论家更新后，状态 s 和动作 a 的目标值。这确保本地评论家具有系统所见的乐观（但很可能可行的）目标。使用这一乐观目标计算贝尔曼误差后，我们按如下方式更新本地评论家

$$Q_i^{(t,k+1)} = \arg\min_Q \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}_i}[(r+\gamma\tilde{Q}_i^{(t,k)}(s',a') - Q(s,a))^2], \tag{10}$$

其中 $a' = \pi_i^{(t,k)}$。实际上，我们在单次梯度更新后获得 $Q_i^{(t,k+1)}$。

5.3 面向异构数据的近端策略更新

虽然乐观评论家对于设定积极的估计至关重要，但它可能会错误地估计 $\tilde{Q}_i^{(t,k)}$ 的价值。因此，同时针对本地数据和联邦策略对本地策略更新进行正则化至关重要。对于相对于本地离线数据的正则化，我们采用与 TD3-BC 算法相同的方法，并定义本地损失函数 $\mathcal{L}_{\text{local}}(\pi) = \mathbb{E}_{(s,a)\sim\mathcal{D}_i}[-Q_i^{(t,k)}(s,\ \pi(s)) + (\pi(s) - a)^2]$。随后，我们在公式（11）中定义演员损失 $\mathcal{L}_{\text{actor}}$，其中第二项是相对于联邦策略的正则化。本地策略使用 $\mathcal{L}_{\text{actor}}$进行更新，

$$\mathcal{L}_{\text{actor}}(\pi) = \mathcal{L}_{\text{local}}(\pi) + \mathbb{E}_{(s,a)\sim\mathcal{D}_i}[(\pi(s) - \pi_{\text{fed}}^t(s))^2], \quad \pi_i^{t,k+1} = \arg\min_\pi \mathcal{L}_{\text{actor}}(\pi). \tag{11}$$

5.4 衰减本地数据的影响

FEDORA 在其策略更新公式（11）中结合使用本地数据损失和近端项。然而，本地数据损失可能会妨碍更新后策略的性能，因为本地数据集可能是根据非专家行为策略生成的。因此，如果客户端通过降低 $\mathcal{L}_{\text{actor}}$ 中 $\mathcal{L}_{\text{local}}$ 的影响而导致更新后策略的性能下降，则必须衰减其本地数据的影响。为此，我们首先在轮次 t 使用联邦评论家和本地数据评估联邦策略的性能。在此评估中，我们使用代理估计 $J^{\text{fed},\ i\ \ t} = \mathbb{E}_{s\sim\mathcal{D}_i}[Q_{\text{fed}}^t(s,\ \pi_{\text{fed}}^t(s))]$。我们将该值与更新后策略的性能 J_i^t 进行比较，后者使用更新后的评论家获得。该差值为我们提供了本地数据所带来改进的估计值。如果 $J^{\text{fed},\ i\ \ t} \geq J_i^t$，我们将 $\mathcal{L}_{\text{local}}$ 的影响衰减一个因子 δ。

我们在算法1和2中总结了 FEDORA。我们想强调的是，在我们的算法中，与任何离线强化学习场景一样，客户端或服务器都无法访问环境或马尔可夫决策过程。此外，我们还想指出，客户端并不了解自己所持有数据的质量。

算法1客户端 i 的算法概要

函数 train_client(π^t_{fed}, Q^t_{fed})
  $\pi_i^{(t,0)} = \pi^t_{\text{fed}}$, $Q_i^{(t,0)} = Q^t_{\text{fed}}$
  for $1 \le k < K$ **执行**
    通过一个梯度步更新评论家，关于公式（10）
    通过一个梯度步更新演员，关于公式（11）
  结束循环
  如果 $J^{\text{fed}}_{\ i\ \ t} \ge J_i^t$，则 $\mathcal{L}_{\text{local}}$ 按 δ 衰减
结束函数

算法 2 服务器算法概要

初始化 π^1_{fed} Q^1_{fed}
对 $t \in 1 \ldots$ **执行**
  将 π^t_{fed} 和 Q^t_{fed} 发送至 $i \in \mathcal{N}$
  采样 $\mathcal{N}_t \subset \mathcal{N}$
  对 $i \in \mathcal{N}_t$ **执行**
    i.train_客户端 (π^t_{fed}, Q^t_{fed})（客户端侧）
  结束循环
  为 $\mathcal{N}_t$ 中的客户端计算 $\pi^{t+1}{}_{\text{fed}}$ 和 $Q^{t+1}{}_{\text{fed}}$ 分别使用公式 (8) 和 (9)。
结束循环

6 实验评估

我们开展实验以回答三个广泛的问题：**(i) 比较性能：** 对于由异质行为策略生成的客户端数据，FEDORA 与其他方法相比表现如何？**(ii) 对客户端更新和数据质量的敏感性：** 性能如何取决于客户端的本地梯度步数、参与联邦的智能体数量的随机性，以及客户端数据的质量？**(iii) 消融实验：** 性能如何取决于 FEDORA 的不同组件？我们在 Flower 联邦学习平台 [2] 上实现了 FEDORA，该平台支持跨设备学习。我们还提供了可在单台机器上运行的仿真设置（见附录A）。

基线方法： 我们考虑以下基线方法。**(i) Fed-A:** 所有客户端的本地目标均遵循 TD3-BC（公式 3）。服务器对演员的参数执行 FedAvg，而每个客户端在本地学习评论家。**(ii) Fed-AC:** 所有客户端的本地目标均遵循 TD3-BC，服务器对演员和评论家的参数执行 FedAvg。**(iii) Fed-AC-Prox:**我们在 Fed-AC 中加入近端项；已有研究表明，当客户端的数据具有异构性时，该项有助于联邦监督学习 [18]。**(iv) 异质数据感知联邦学习 (HDAFL)** 我们将 HDAFL [35]扩展到离线强化学习场景：将演员网络划分为通用部分和客户端特定部分，并在每轮中仅对通用部分进行联邦聚合。**(v) 集中式：**我们合并所有客户端的数据，在汇总数据上执行离线强化学习（TD3-BC）。

6.1 仿真环境实验

实验设置：我们重点研究这样一种场景：多个客户端协同学习以解决同一任务，但各客户端用于收集数据的行为策略可能有所不同。我们在客户端数量为 $|\mathcal{N}| = 50$，且每个客户端的本地数据集大小为 $|\mathcal{D}_i| = 5000$的条件下开展实验。在这 50 个客户端中，25 个客户端获得来自 D4RL [4] 专家数据集的数据，其余 25 个客户端则获得来自 D4RL 中等数据集的数据。客户端（以及服务器）均不了解各自数据集的质量。此外，客户端和服务器均无法访问环境。我们从中随机选择 $|\mathcal{N}_t| = 20$ 个客户端参与每一轮 t 联邦训练。服务器从客户端获取权重，具体为 $|\mathcal{N}_t|$，并计算联邦权重 $\pi^{t+1}{}_{\text{fed}}$ 和 $Q^{t+1}{}_{\text{fed}}$。对于每张图，我们使用四个不同的随机种子评估性能。我们在三个 MuJoCo 任务上评估 FEDORA 和基线方法的性能：跳跃者、半猎豹和 Walker2D。在每一轮联邦训练中，所有算法的每个客户端均执行 20 个本地训练轮次，在我们的实验设置中，这大约相当于 380 个本地梯度步。

FEDORA 的比较性能：在图2中，我们绘制了服务器/联邦策略在每个通信/联邦轮次中的累积回合奖励。我们观察到，FEDORA 的表现优于所有联邦基线方法，其性能达到或优于集中式训练。此外，即使经过许多通信轮次的训练，联邦基线方法仍未能学习到良好的服务器策略，并且与 FEDORA 相比，其性能停留在较低水平，这凸显了异构数据会损害其性能。

为了理解来自多个行为策略的数据对集中式训练的影响，我们考虑这样一种场景：50 个客户端参与联邦学习，每个客户端的数据集大小为 $|D_i| = 5000$，其中 25 个客户端拥有专家数据，另外 25 个客户端拥有随机数据，即由一个.....生成的数据。

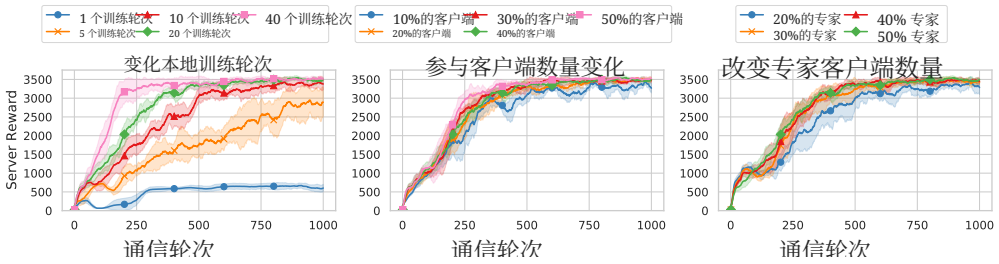
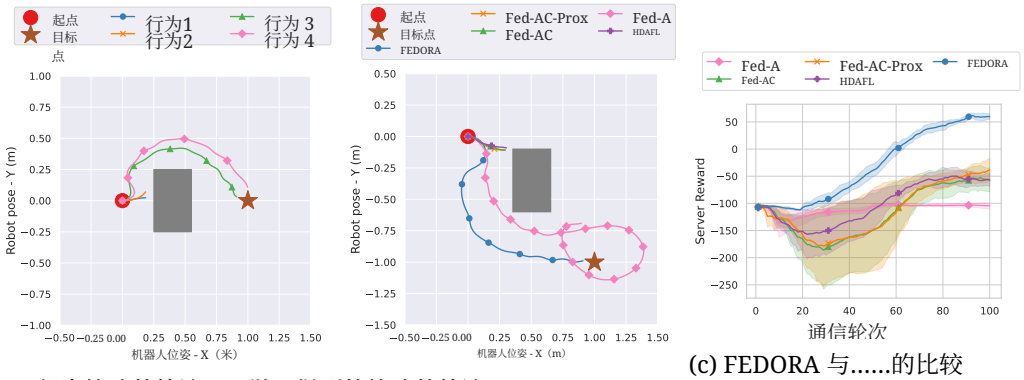


图 4：在 FEDORA 中改变 (a) 本地梯度步数、(b) 每轮次参与的客户端数量，以及 (c) 专家客户端数量的影响。

随机策略。从图 3 中可以看出，与 FEDORA 相比，结合所有客户端的数据会导致性能下降。这一观察结果表明，使用通过多种行为策略收集的数据进行集中式训练可能会产生不利影响。



(a) 行为策略的轨迹 (b) 学习得到的策略的轨迹

(c) FEDORA 与……的比较联邦基线算法

图 5：存在障碍物时，FEDORA 与其他联邦基线方法在移动机器人导航任务中的评估。

对客户端更新和数据质量的敏感性：我们在与图2相同的设置下，研究 FEDORA 对 Hopper 环境中客户端更新频率和数据质量的敏感性。增加本地训练步数可以提高通信效率，但在异构数据条件下会因客户端漂移 [11]而产生不利影响。在图4(a)中，我们研究了改变本地训练轮数的影响。我们观察到，增加训练轮数会加快学习速度，这表明 FEDORA 能够有效地从异构数据中学习。由于通信/计算方面的限在线a，并非所有客户端都能参与联邦的每一轮。

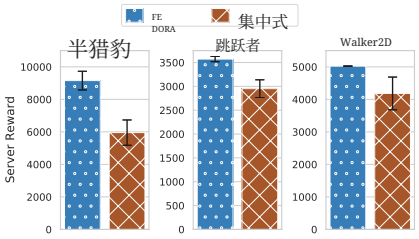


图3：FEDORA 与异构数据集中式训练的比较。

束。在图 4(b) 中，我们研究了参与联邦的客户端比例所产生的影响。我们观察到，FEDORA 对联邦过程中客户端比例的变化具有较强的鲁棒性。最后，在图 4(c) 中，我们通过改变拥有专家数据集的客户端比例，研究了数据异质性的影响。我们观察到，即使只有 20% 的客户端拥有专家级数据，FEDORA 仍能取得良好的性能。我们在附录 B 中展示了若干消融研究和额外实验。

6.2 TurtleBot 上的真实世界实验

我们在 TurtleBot [1], a 双轮差速驱动机器人（图 6）上评估了 FEDORA 的性能，以利用分布在多个机器人（客户端）上的离线数据，通过协作学习控制策略，使机器人能够在避开障碍物的同时导航至航点。该场景与多种现实应用相关，例如分布在不同住宅中的清洁机器人，它们旨在协作

学习控制策略，利用分布在不同机器人上的数据进行导航并避开障碍物。协作学习至关重要，因为单个机器人可能没有足够的数据可供学习，或者未曾遇到足够多样的场景。此外，联邦学习克服了在机器人之间共享数据所带来的隐私问题。

我们使用四种具有不同专业水平的行为策略在真实世界中收集数据（图5(a)）。我们在20个客户端上训练100个通信轮次，每个通信轮次包含20个本地训练轮次（见图5(c)）。图5(b)展示了真实世界中由不同算法学习到的策略所获得的轨迹，只有FEDORA能够通过避开障碍物成功到达目标点。更多细节见附录C。我们在 https://github.com/DesikRengarajan/FEDORA 提供了实验视频。

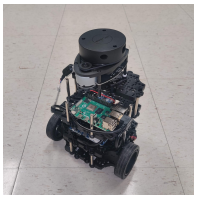


图6：TurtleBot3 Burger。

7 结论

我们提出了一种联邦离线强化学习方法，该方法考虑了在客户端生成数据的策略集成质量的异质性。我们通过系统地开发一种性能良好的集成引导方法FEDORA，解决了多个具有挑战性的问题。该方法提取策略和评论家的集体智慧，并抑制对无关本地数据的过度依赖。我们在多个仿真和真实世界任务中展示了其性能。

8 伦理声明与社会影响

在这项工作中，我们提出了一种用于联邦离线强化学习的新算法。联邦离线强化学习领域有望在不共享数据的情况下，既保护隐私又减少通信需求，从而推动强化学习算法的广泛应用。在整个研究过程中，我们没有涉及人类受试者或人类生成的数据。因此，我们认为本研究方法不存在任何伦理问题。

尽管强化学习在应用于有益于社会的系统方面前景广阔，但将其应用于涉及人类交互的环境时必须谨慎。这是因为此类场景中的保证具有概率性，因此必须确保相关风险处于可接受范围内，以实现安全部署。

9 局限性与未来工作

在这项工作中，我们研究了联邦离线强化学习问题。我们假设所有客户端共享相同的马尔可夫决策过程模型（转移核和奖励模型），而离线数据集之间的任何统计差异都源于收集数据时所使用的行为策略不同。未来，我们计划将研究范围扩展到客户端具有不同转移模型和奖励模型的场景。为此，我们计划将离线元强化学习的思路扩展到联邦学习场景。此外，我们计划进一步探索联邦离线强化学习中的个性化问题。我们还认为，我们的方法也可能适用于联邦监督学习，尤其是在数据来源质量各异的情况下；我们打算在未来将其作为一条独立研究方向进行正式探索。

10 致谢

本研究部分得到了美国国家科学基金会资助 CNS 2312978、ECCS 2038963、美国陆军研究办公室资助 W911NF-19-1-0367 以及 NSF-CAREER-EPCN-2045783 的支持。本材料中表达的任何观点、研究结果、结论或建议均为作者的观点，不一定反映资助机构的意见。

本研究的部分工作使用了德克萨斯农工大学高性能研究计算中心提供的先进计算资源。

参考文献

[1] Robin Amsters 和 Peter Slaets。Turtlebot 3 作为机器人教育平台。载于《机器人教育：当前研究与创新》，第 170–181 页。Springer，2020 年。[2] Daniel J Beutel、Taner Topal、Akhil Mathur、Xinchi Qiu、Titouan Parcollet 和 Nicholas D Lane。Flower：一个友好的联邦学习研究框架。*arXiv* 预印本 *arXiv:2007.14390*，2020 年。[3] Xinyue Chen、Zijian Zhou、Zheng Wang、Che Wang、Yanqiu Wu 和 Keith Ross。Bail：面向批量深度强化学习的最佳动作模仿学习。神经信息处理系统进展，33：18353–18363，2020 年。[4] Justin Fu、Aviral Kumar、Ofir Nachum、George Tucker 和 Sergey Levine。D4RL：用于深度数据驱动强化学习的数据集。*arXiv* 预印本 *arXiv:2004.07219*，2020 年。[5] Scott Fujimoto 和 Shixiang Shane Gu。离线强化学习的极简方法。神经信息处理系统进展，34：20132–20145，2021 年。[6] Scott Fujimoto、Herke Hoof 和 David Meger。解决演员-评论家方法中的函数逼近误差。载于国际机器学习会议，第 1587–1596 页，2018 年。[7] Scott Fujimoto、David Meger 和 Doina Precup。无需探索的离策略深度强化学习。载于国际机器学习会议，第 2052–2062 页，2019 年。[8] Liam Hebert、Lukasz Golab、Pascal Poupart 和 Robin Cohen。Fedformer：强化学习中结合注意力的情境联邦。*arXiv* 预印本 *arXiv:2205.13697*，2022 年。[9] Yiqiu Hu、Yun Hua、Wenyan Liu 和 Jun Zhu。基于奖励塑形的联邦强化学习。*IEEE Access*，9：67259–67267，2021 年。[10] Peter Kairouz、H Brendan McMahan、Brendan Avent、Aurélien Bellet、Mehdi Bennis、Ar- jun Nitin Bhagoji、Kallista Bonawitz、Zachary Charles、Graham Cormode、Rachel Cummings 等。联邦学习的进展与开放问题。《机器学习基础与趋势®》，14(1–2)：1–210，2021 年。[11] Sai Praneeth Karimireddy、Satyen Kale、Mehryar Mohri、Sashank Reddi、Sebastian Stich 和 Ananda Theertha Suresh。Scaffold：用于联邦学习的随机控制平均方法。载于国际机器学习会议，第 5132–5143 页，2020 年。[12] Sajad Khodadadian、Pranay Sharma、Gauri Joshi 和 Siva Theja Maguluri。联邦强化学习：马尔可夫采样下的线性加速。载于国际机器学习会议，第 10997–11057 页，2022 年。[13] Ilya Kostrikov、Ashvin Nair 和 Sergey Levine。基于隐式 Q 学习的离线强化学习。*arXiv* 预印本 *arXiv:2110.06169*，2021 年。[14] Aviral Kumar、Justin Fu、Matthew Soh、George Tucker 和 Sergey Levine。通过自举误差缩减稳定离策略 Q 学习。神经信息处理系统进展，32，2019 年。[15] Aviral Kumar、Aurick Zhou、George Tucker 和 Sergey Levine。面向离线强化学习的保守 Q 学习。神经信息处理系统进展，33：1179–1191，2020 年。[16] Aviral Kumar、Aurick Zhou、George Tucker 和 Sergey Levine。面向离线强化学习的保守 Q 学习。神经信息处理系统进展，33：1179–1191，2020 年。[17] Sergey Levine、Aviral Kumar、George Tucker 和 Justin Fu。离线强化学习：教程、综述及开放问题展望。*arXiv* 预印本 *arXiv:2005.01643*，2020 年。[18] Tian Li、Anit Kumar Sahu、Manzil Zaheer、Maziar Sanjabi、Ameet Talwalkar 和 Virginia Smith。异构网络中的联邦优化。机器学习与系统会议论文集，2：429–450，2020 年。

[19] Hyun-Kyo Lim、Ju-Bong Kim、Ihsan Ullah、Joo-Seong Heo 和 Youn-Hee Han。用于多设备精确控制的联邦强化学习加速方法。*IEEEAccess*，9:76296–76306，2021。[20] Brendan McMahan、Eider Moore、Daniel Ramage、Seth Hampson 和 Blaise Aguera 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，第1273–1282页，2017。[21] Chetan Nadiger、Anil Kumar 和 Sherine Abdelhak。面向快速个性化的联邦强化学习。载于*2019*年*IEEE*第二届人工智能与知识工程国际会议（*AIKE*），第123–127页。IEEE，2019。[22] Xue Bin Peng、Aviral Kumar、Grace Zhang 和 Sergey Levine。优势加权回归：简单且可扩展的离策略强化学习。*arXiv*预印本 *arXiv:1910.00177*，2019。[23] Jiaju Qi 、 Qihao Zhou 、 Lei Lei 和 Kan Zheng。联邦强化学习：技术、应用与开放挑战。*arXiv*预印本 *arXiv:2108.11887*，2021。[24] Rongjun Qin、Songyi Gao、Xingyuan Zhang、Zhen Xu、Shengkai Huang、Ze wen Li、Weinan Zhang 和 Yang Yu。NeoRL：面向离线强化学习的近真实世界基准。*arXiv*预印本 *arXiv:2102.00714*，2021。[25] Sashank J. Reddi、Zachary Charles、Manzil Zaheer、Zachary Garrett、Keith Rush、Jakub Koneˇcnýˇ、Sanjiv Kumar 和 Hugh Brendan McMahan。自适应联邦优化。载于国际学习表征会议，2021。[26] J ohn Schulman、Sergey Levine、Pieter Abbeel、Michael Jordan 和 Philipp Moritz。信赖域策略优化。载于第*32*届国际机器学习会议论文集，第37卷，第1889–1897页，2015。[27] Denis Tarasov、Alexander Nikulin、Dmitry Akimov、Vladislav Kurenkov 和 Sergey Kolesnikov。CORL：面向研究的深度离线强化学习库。载于第三届离线强化学习研讨会：离线强化学习作为“发射台”，2022。[28] Emanuel Todorov、Tom Erez 和 Yuval Tassa。MuJoCo：用于基于模型的控制的物理引擎。载于*IEEE/RSJ*智能机器人与系统国际会议，第5026–5033页，2012。[29] José R Vázquez-Canteli、Sourav Dey、Gregor Henze 和 Zoltán Nagy。CityLearn：为需求响应和城市能源管理领域的多智能体强化学习研究建立标准。*arXiv*预印本 *arXiv:2012.10504*，2020。[30] Jianyu Wang、Zachary Charles、Zheng Xu、Gauri Joshi、H Brendan McMahan、Maruan Al-Shedivat、Galen Andrew、Salman Avestimehr、Katharine Daly、Deepesh Data 等。联邦优化指南。*arXiv*预印本 *arXiv:2107.06917*，2021。[31] Qing Wang、Jiechao Xiong、Lei Han、Han Liu、Tong Zhang 等。针对批量历史数据的指数加权模仿学习。神经信息处理系统进展，2018。[32] Xiaofei Wang、Ruibin Li、Chenyang Wang、Xiuhua Li、Tarik Taleb 和 Victor CM Leung。面向设备到设备辅助异构协作边缘缓存的注意力加权联邦深度强化学习。*IEEE*通信领域精选期刊，39(1):154–169，2020。[33] Yifan Wu、George Tucker 和 Ofir Nachum。行为正则化离线强化学习。*arXiv*预印本 *arXiv:1911.11361*，2019。[34] Zhijie Xie 和 Shenghui Song。Fedkl：通过惩罚 KL 散度应对联邦强化学习中的数据异质性。*IEEE*通信领域精选期刊，41(4):1227–1242，2023。[35] Lixuan Yang、Cedric Beliard 和 Dario Rossi。异质数据感知联邦学习。*arXiv*

*prep**rint arXiv:2011.06393*，2020年。

[36] Tianhe Yu、Aviral Kumar、Yevgen Chebotar、Karol Hausman、Sergey Levine 和 Chelsea Finn。面向多任务离线强化学习的保守数据共享。神经信息处理系统进展，34:11501–11516，2021。[37] Yang Yue、Bingyi Kang、Xiao Ma、Gao Huang、Shiji Song 和 Shuicheng Yan。离线优先经验回放。*arXiv* 预印本 *arXiv:2306.05412*，2023。[38] Yang Yue、Bingyi Kang、Xiao Ma、Zhongwen Xu、Gao Huang 和 Shuicheng Yan。通过数据再平衡增强离线强化学习。*arXiv* 预印本 *arXiv:2210.09241*，2022。[39] Doudou Zhou、Yufeng Zhang、Aaron Sonabend-W、Zhaoran Wang、Junwei Lu 和 Tianxi Cai。联邦离线强化学习。*arXiv* 预印本 *arXiv:2206.05581*，2022。

附录

我们在附录中展示了若干结果和细节，以说明 FEDORA 的性能。这些内容包括实验设置的详细信息（附录A）、用于研究 FEDORA 不同组成部分并展示其在不同设置下性能的补充实验（附录B），以及使用 TurtleBot 开展的真实世界实验的详细信息（附录C）。

A 实验设置

算法实现： 我们使用 PyTorch 框架编写本文中的算法，基于一个公开可用的 TD3-BC 实现。演员网络和评论家网络均包含两个隐藏层，每层大小为 256，并使用 ReLU 非线性函数。我们使用 0.99 的折扣因子，客户端使用 Adam优化器以 3×10^{-4} 的学习率更新其网络。训练 FEDORA 时，我们固定衰减率 $\delta = 0.995$ 和温度 $\beta = 0.1$。在集中式设置中，TD3-BC 训练 5×10^5 个时间步。联邦训练和集中式训练的批量大小均为 256。

客户端的训练数据由从D4RL数据集采样的轨迹组成。在只有部分客户端参与联邦某一轮次的情况下，我们从整个客户端集合中均匀采样所需数量的客户端。

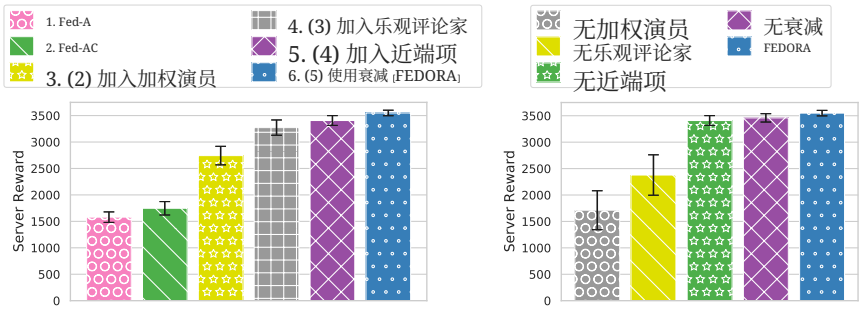
联邦结构： 我们基于 Flower 联邦学习平台 [2],实现了 FEDORA。该平台支持在具有异构软件栈、计算能力和网络带宽的设备之间进行学习。Flower 管理客户端与服务器之间的所有通信，并使我们能够轻松实现 FEDORA 自定义的服务器端和客户端算法。然而，由于 Flower 面向监督学习，因此每轮联邦中只会传输和接收一个模型，而我们希望联邦化策略和评论家模型。我们通过将两个模型简单地附加在一起，并在服务器端和客户端分别进行适当的打包与解包，解决了这一限制。

虽然“FEDORA-over-Flower”是在分布式计算资源上协同工作的有效解决方案，我们也希望有一种能够在单台机器上运行的仿真设置。此方法依次在每个选定的客户端上执行 FEDORA，然后执行联邦步骤，从而能够在理想化的联邦设置中评估 FEDORA 的各个组成部分。

计算资源： 在 MuJoCo环境（如图2所示）上进行的每次运行，若在单台机器（AMD Ryzen Threadripper 3960X 24核处理器，
2 块 NVIDIA 2080Ti GPU）。使用 Flower 框架在分布式计算环境中运行时，这段时间可以大幅缩短。

B 额外实验

B.1 单个算法组件的重要性



(a) 顺序添加的影响
一次处理一个算法组件

(b) 移除单个个体的效果
FEDORA 的算法组件

图 7：消融研究。

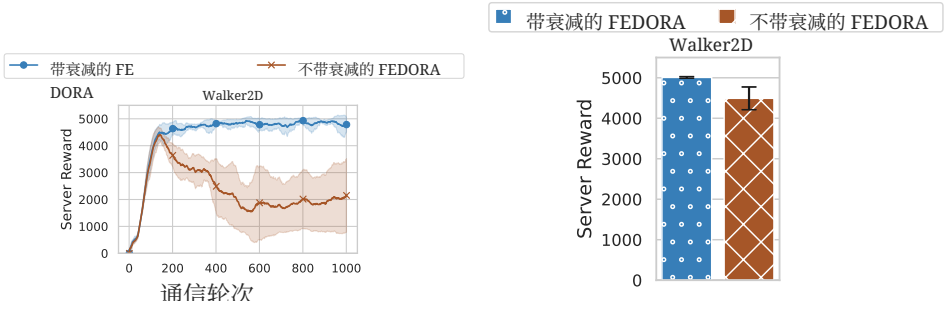
我们进行了消融研究，以考察算法的不同组件，并了解它们对联邦策略性能的相对影响。我们使用第 4 节所述的实验框架，其中包含 10 个客户端和 Hopper 环境，并绘制联邦策略在 4 个随机种子上的平均性能及标准差。消融实验采用两种方式进行：（a）从 Fed-A（仅对演员进行联邦化的朴素方法）开始逐步构建 FEDORA，每次添加一个新的算法组件并评估其性能；（b）每次从 FEDORA 中排除一个组件，并评估由此得到的算法。

我们在图 7a 中观察到，使用客户端策略的优先级加权平均来计算联邦策略（Eq. (8)），并采用乐观评论家（Eq. (9) - (10)），能够显著提升联邦策略的性能。这与我们的直觉一致：最重要的是提取可供联邦使用的策略和评论家中的集体智慧，并确保评论家设定乐观的目标点。近端项通过选择接近本地数据集或联邦策略中所见动作的动作，帮助正则化本地策略更新（公式（11））。此外，逐渐减弱本地更新的影响，使本地策略能够通过选择本地数据集中未出现的动作来利用联邦策略的视角。

从图 7b 可以看出，从 FEDORA 中移除优先级加权演员会导致性能下降幅度最大，其次是乐观评论家。同样，这与我们的直觉一致，即这两者的影响最为重要。去除近端项和本地衰减也会导致服务器性能下降，同时标准差增大。

B.2 Walker环境中衰减机制的消融实验

我们在图8中研究了衰减本地数据（5.4）影响这一做法在Walker2D环境中的效果。尽管图7所示的实验（针对Hopper环境）中，衰减机制似乎只带来了小幅提升，但我们观察到，它在Walker2D环境中带来了显著提升。



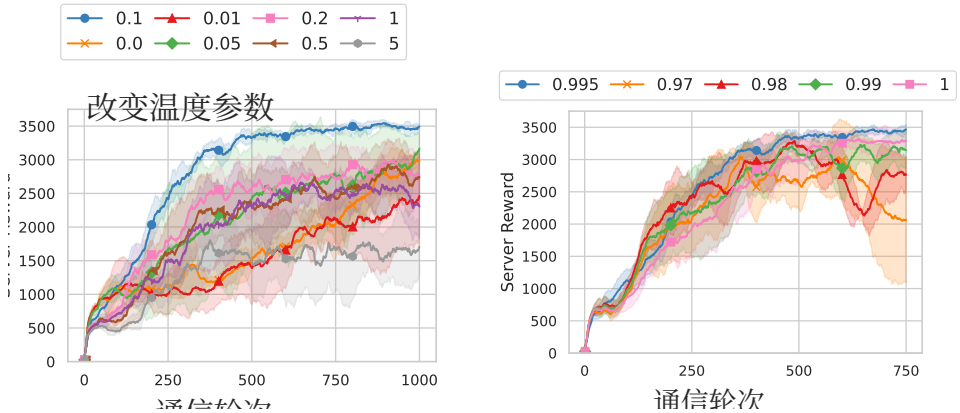
（a）训练曲线　　（b）峰值表现

图8：Walker2d环境中衰减机制的消融研究（设置与图7类似）。

B.3 超参数扫描

在图 9a 中，针对不同的 β 取值运行 FEDORA，其中 β 是联邦的温度参数。我们考虑与图 1 类似的场景：采用 Hopper-v2 环境，共有 10 个客户端，其中 $|D_i| = 5000$ 参与联邦；5 个客户端拥有专家数据，另外 5 个拥有中等数据。当 $\beta = 0$ 时，这相当于采用均匀加权方案，在联邦过程中不考虑各客户端所含数据的质量。随着 $\beta \to \infty$，加权方案趋向于最大加权方案，此时联邦策略与数据质量最高的单个客户端的策略相同。

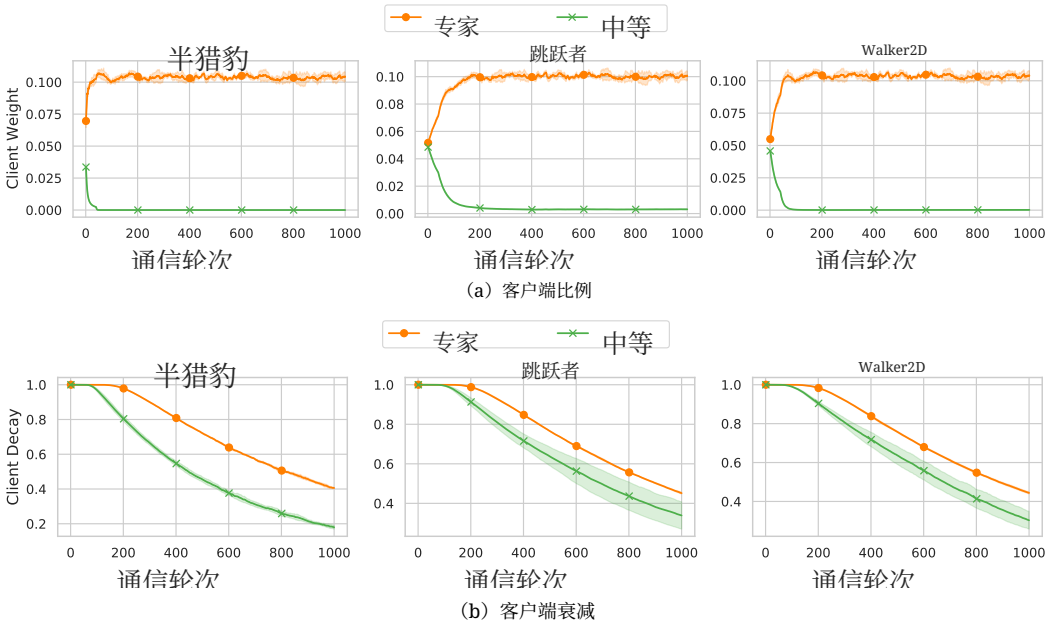
在图 9b 中，我们针对衰减参数 δ 的不同取值运行 FEDORA。衰减参数通过降低 $\mathcal{L}_{\text{local}}$ 的影响，来控制本地数据在 $\mathcal{L}_{\text{actor}}$ 中的影响。在图 9b 中，我们考虑与图 1 类似的场景：采用 Hopper-v2 环境，10 个客户端拥有 $|D_i| = 5000$ 并参与联邦，其中 5 个客户端拥有专家数据，另外 5 个客户端拥有中等数据。我们对所有算法进行 750 轮联邦训练。我们观察到，FEDORA 对 δ 取值的变化具有鲁棒性。



（a）变化 β　　（b）变化 δ

图 9：超参数扫描

B.4 客户端性能分析



（a）客户端比例

（b）客户端衰减

图10：联邦过程中客户端表现的分析。在每轮联邦中，计算参与的专家客户端和中等客户端的性能指标平均值。

我们在 MuJoCo环境 上训练 FEDORA，采用与第 6 节类似的设置：每轮联邦训练中，随机选择 50 个客户端中的 20 个参与。我们的目标是分析拥有专家数据和中等数据的客户端对学习过程的贡献。与之前一样，客户端和算法都不了解数据质量。

我们绘制了每轮联邦中参与的专家和中等数据集客户端的平均权重 w_i^t，如图10a所示。我们观察到，中等客户端的权重降至 0，而专家客户端的权重升至 0。1。这一发现强调了客户端是根据其相对优值进行组合的。

在图 10b 中，我们绘制了参与的专家和中等数据集客户端（第 5.4 节）中与 $\mathcal{L}_{\text{local}}$ 相关的衰减值均值。随着训练的推进，两组客户端的衰减值都在下降。每当联邦策略的性能本地估计值 $J^{\text{fed}}_{i,t}$

高于更新后的本地策略性能估计值 J_i^t时，衰减值就会下降。衰减值降低意味着，随着轮次 t 推进，联邦策略更常带来优于本地策略的性能提升。因此，仅使用本地数据进行训练会造成不利影响，而参与联邦学习有助于学习到更优的策略。

B.5 使用 CityLearn 进行联邦离线强化学习实验

真实世界环境通常具有较大的状态空间，且本质上具有随机性。我们在 CityLearn [29] 上开展联邦实验，以评估 FEDORA 在此类大规模系统中的有效性。CityLearn 是一个 OpenAI Gym 环境，旨在实现城市规模能源管理和需求响应，其建模依据是住宅建筑的数据。目标是通过调节冷冻水箱和生活热水来重塑总能源需求曲线；这两者是每栋建筑中热能储存的两种形式。随着社区发展和天气变化，住宅建筑的能源需求也会发生变化。因此，控制器必须定期更新其策略，以实现高效的能源管理。联邦学习使服务于

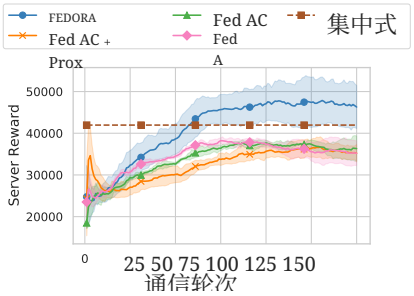


图 11：在 CityLearn 上对算法进行评估。

邻近社区的公用事业公司能够协同训练策略，同时保护用户数据隐私，这促使我们在该环境中使用 FEDORA。

在我们的实验中，共有 10 个客户端，每个客户端拥有 5000 个训练样本，并且都参与联邦中的 150 轮训练。客户端的训练数据来自离线强化学习基准 NeoRL [24]。其中 5 个客户端分别拥有来自 CityLearn High 和 CityLearn Low 数据集的数据，这些数据由软演员-评论家策略采集，该策略分别训练至最佳性能水平的 75% 和 25%。在联邦的每一轮中，每个客户端都会进行 20 个本地训练轮次。每轮联邦结束时，服务器奖励会通过在线评估，并显示在图 11 中。我们观察到，FEDORA 的表现优于其他联邦离线强化学习算法，也优于集中式训练；后者使用 TD3-BC 在汇总自所有客户端的数据上进行学习。这些发现表明，FEDORA 能够在大规模随机环境中取得良好表现。

B.6 多种行为策略及参与联邦的客户端比例的影响

在本节中，我们研究客户端拥有来自多个行为策略的数据时，不同客户端参与联邦的比例所产生的影响。我们考虑这样一种情形：在 Hopper-v2 环境中有 50 个客户端，并且其中，

- 12 个客户端拥有专家数据（来自使用软演员-评论家训练至收敛的策略的样本）。
- 12个客户端拥有中等数据（来自训练后性能约为专家策略三分之一的策略的样本）。
- 14个客户端拥有随机数据（来自随机初始化策略的样本）。
- 12 个客户端拥有某个策略的回放缓冲区中的数据，该策略经过训练，性能达到中等智能体的水平。

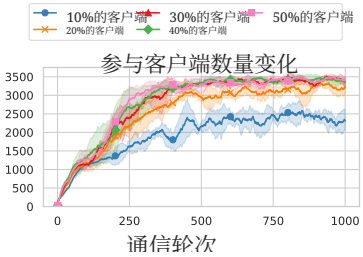


图 12：每轮参与的客户端数量变化对 FEDORA 的影响

我们通过改变每轮联邦中参与的客户端比例来运行 FEDORA。我们观察到，即使参与联邦的客户端比例低至 20%，FEDORA 依然相当稳健。

B.7 可变大小数据集

在图 13 中，我们运行客户端数据集大小可变的 FEDORA，并将其与数据集大小固定的 FEDORA 进行比较。

FEDORA 方差数据集：我们考虑 Hopper-v2 环境中有 10 个客户端的场景。5 个客户端拥有专家数据集，数据集大小分别为 4000，5000，6000，7000，8000；另外 5 个客户端拥有中等数据集，数据集大小分别为 4000，5000，6000，7000，8000。

FEDORA： 我们考虑图1所描述的场景，数据集大小固定为5000。

从图13中我们可以观察到，FEDORA具有鲁棒性，能够处理可变数据集大小。

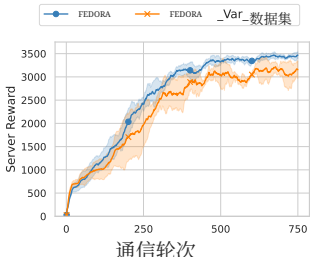


图13：可变数据集大小

B.8 与其他离线强化学习算法进行集中式训练

我们针对跳跃者-v2环境考虑了一个与图3类似的场景，其中有50个客户端，$|D_i| = 5000$参与联邦，其中25个客户端拥有专家数据，另外25个客户端拥有随机数据。我们将不同离线强化学习算法在汇总数据上的表现与FEDORA进行比较。我们选择的算法是离线强化学习保守Q学习（CQL）[15]和采用隐式Q学习（IQL）的离线强化学习 [13]，其实现来自CORL库 [27]。从图14中可以观察到，将来自不同策略的数据汇总会影响这两种离线强化学习算法。

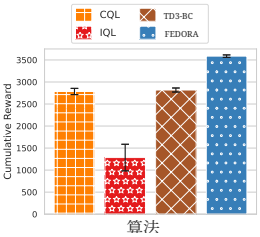


图14：与不同离线强化学习算法的比较

B.9 数据再平衡

我们考虑一个类似于图1的场景，其中采用包含10个客户端的Hopper-v2环境，$|D_i| = 5000$ 参与联邦，其中5个客户端拥有专家数据，5个客户端拥有中等数据。我们比较TD3-BC、采用数据再平衡的TD3-BC（TD3-BC_RB） [38, 37], 和FEDORA的性能。从图15中可以注意到，当数据通过多个行为策略收集时，加入数据再平衡确实有助于提升离线强化学习算法的性能。我们还注意到，采用数据再平衡的TD3-BC性能并未达到FEDORA的性能。我们推测，这可能是由于其采用了更优越的加权机制，

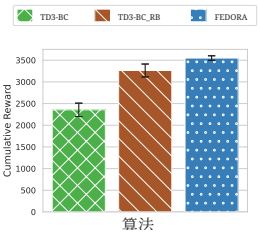


图15：与采用数据再平衡的离线强化学习的比较

而且数据再平衡无法完全解决由来自多个行为策略的数据所引起的分布偏移问题。

B.10 不同的加权机制

我们考虑一个与图 1 类似的场景，其中研究的是 Hopper-v2 环境，有 10 个客户端参与$|D_i| = 5000$联邦，其中 5 个客户端拥有专家数据，另有 5 个客户端拥有中等数据。我们运行所有算法进行 750 轮联邦训练。我们将 FEDORA 的性能与另外两种基线方法进行比较，这两种方法仅将 FEDORA 的加权机制改为基于各客户端数据集的平均奖励。**(1.) FEDORA_RAvg：** 我们根据客户端数据集中的平均奖励组合客户端。我们选择客户端联邦权重i, $w_i = \frac{R_i}{\sum_{k \in \mathcal{N}_t} R_k}$，其中 R_i 对应客户端 i 数据集的平均奖励，$\mathcal{N}_t$ 是第 t 轮参与联邦的客户端集合。**(2.) FE-DORA_RC：** 我们扩展了所提出的加权方案

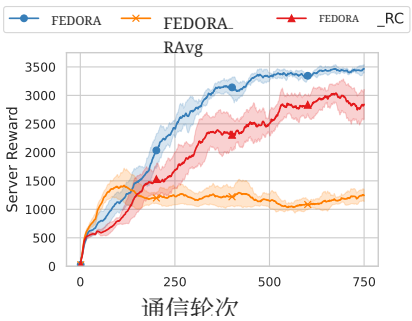


图16：基于数据集中的平均奖励对不同加权机制进行比较

在 [38] 中，将其应用到联邦设置中。在此场景下，我们选择客户端的联邦权重 i，$w_i = \frac{p_i}{\sum_{k\in\mathcal{N}_t} p_k}$。其中 $p_i = \frac{R_i - R_{\min}}{R_{\max} - R_{\min}}$，此处 R_i 对应客户端 i 数据集的平均奖励，$R_{\min} = \min_{i\in\mathcal{N}_t} R_i$，以及 $R_{\max} = \max_{i\in\mathcal{N}_t} R_i$。
从图 16 可以看出，FEDORA 的表现优于两种基线方法。我们发现，根据数据集的平均奖励进行加权无法取得良好效果，这可能是因为各数据集的平均奖励差异不大。例如，拥有专家数据集的客户端，其平均奖励为 3.6，而拥有中等数据集的客户端，其平均奖励为 3.11。因此，仅根据平均奖励进行组合，最终相当于采用均匀加权方案。我们发现，扩展 [38] 中提出的方法确实有所帮助，但这种加权方案仍不如 FEDORA。
FEDROA 的加权方案更优，因为它根据从数据集学习到的策略的表现来组合策略，而不是依据数据集中的平均奖励。换言之，FEDORA 会对能够学习出表现更好的策略的数据集赋予更高权重。

C 真实世界机器人实验详情

C.1 示范数据收集

我们使用信赖域策略优化 [26]，在第 C.2 节所述的移动机器人定制模拟器上训练了四种专业水平各异的行为策略。第一种策略能够进行航点导航，但会与障碍物发生碰撞。第二种策略能够到达航点，同时避开固定位置的障碍物。第三种策略尚未完全学会避开位于不同位置的障碍物。最后，第四种策略能够无碰撞地导航至目标点。我们在真实环境中执行这些行为策略，并改变航点（目标位置）和障碍物的位置，以收集示范数据；随后，我们利用这些数据训练 FE DORA 和其他基线方法。每个客户端的数据集包含 300 个数据点，均由单一行为策略采集。训练完成后，我们在真实环境中使用 TurtleBot 测试学习策略，以确认其可行性。

C.2 模拟器设计

我们基于 OpenAI Gym 框架开发了一款用于移动机器人的一阶模拟器，可用于训练强化学习算法。机器人在二维空间中的位姿由其 X、Y 坐标以及相对于 X 轴的朝向 θ 表示。位姿使用差速驱动运动学进行更新

$$\begin{aligned} x_{t+1} &= x_t + \Delta t\ v\cos\theta_t \\ y_{t+1} &= y_t + \Delta t\ v\sin\theta_t \\ \theta_{t+1} &= \theta_t + \Delta t\ \omega, \end{aligned} \tag{12}$$

其中，(x_t, y_t, θ_t) 是时刻 t 的位姿，v 和 w 分别是机器人的线速度和角速度，Δt 是系统的时间离散步长。
模拟器使用功能性激光雷达检测是否存在障碍物。我们通过对机器人及其周围环境中的障碍物进行离散表示来模拟激光雷达。对于激光雷达周围的每个扫描方向，我们使用 Bresenham直线算法生成由离散点组成的路径。模拟器通过计算每条路径上的点数来确定激光雷达测量值，从机器人开始沿路径前进，直到遇到障碍物或到达最大量程。

奖励函数旨在鼓励有效的航点导航，同时防止碰撞。我们定义了一个边界网格，该网格在起点和目标位置的所有方向上均向外延伸 1 米。在时刻 t，导航至目标位置（x_g，y_g）的奖励函数定义为

$$R_t = \begin{cases} +100, & \text{如果 } |x_t - x_g| \le thresh \text{ 且} \\ -10, & |y_t - y_g| \le thresh \text{ 如果机器人位于边界之} \\ -100, & \text{外 如果机器人发生碰撞 否则} \\ -(\mathtt{c.t.e}_t^2 + \mathtt{a.t.e}_t + \mathtt{h.e}_t) + \sum \mathtt{lidar}_t, & \end{cases} \tag{13}$$

其中 c。t。e_t 为横向偏差，a。t。e_t 为纵向偏差，h。e_t 为航向误差，lidar_t为时间 t的激光雷达测量值数组，而 $thresh$为距离阈值误差，选取为

0。1 米。设时间 t 时到目标的 L-2 距离和朝向目标的航向分别为 d_t^g 和 θ_t^g。那么，我们有

$$\begin{array}{rcl} d_t^g & = & \sqrt{(x_g - x_t)^2 + (y_g - y_t)^2}, \\ \theta_t^g & = & \tan^{-1}\left(\frac{y_g - y_t}{x_g - x_t}\right), \\ \texttt{c.t.e}_t & = & d_t^g \sin(\theta_g - \theta_t), \\ \texttt{a.t.e}_t & = & |x_g - x_t| + |y_g - y_t|, \\ \texttt{h.e}_t & = & \theta_t^g - \theta_t. \end{array} \tag{14}$$

C.3 移动机器人平台

我们在 Robotis TurtleBot3 Burger 移动机器人 [1]，（一种开源差速驱动机器人）上评估训练好的算法。该机器人配备了基于车轮编码器的位姿估计系统，并搭载 RPLIDAR-A1 激光雷达，用于检测障碍物。我们使用 ROS 作为中间件来建立通信。机器人通过无线网络将其状态（位姿和激光雷达信息）传输至计算机，随后计算机将正在执行的策略所建议的相应动作传回机器人。