最优传输能否改进联邦逆强化学习?

David Millard DJM3622@RIT.EDU
Ali Baheri
AKBEME@RIT.EDU美国纽约州罗切斯特理工学院机械工程系

编者： G. Sukhatme、L. Lindemann、S. Tu、A. Wierman、N. Atanasov

arXiv:2601.00309v1 [cs.LG] 1 Jan 2026

摘要

在机器人学和多智能体系统中，自主智能体组成的群体通常会在略有差异的环境中运行，同时追求共同的高层目标。由于动力学差异、隐私限制和通信带宽有限，直接汇集各智能体的数据来学习共享奖励函数通常并不可行。本文提出一种基于最优传输的联邦逆强化学习（IRL）方法。每个客户端首先在本地执行轻量级最大熵逆强化学习，同时遵守其计算能力和隐私方面的限制。随后，通过 Wasserstein 重心融合得到的奖励函数，并在融合过程中考虑其内在几何结构。我们进一步证明，与联邦学习中传统的参数平均方法相比，这种重心融合能够得到更准确的全局奖励估计。总而言之，本研究提供了一个原则明确且通信高效的框架，用于推导可在异构智能体和环境之间泛化的共享奖励。

关键词： 逆向强化学习、联邦学习、最优传输

1. 引言

随着对个性化机器人学需求的不断增长，如何在保护隐私的同时实现经验共享、加速学习，变得愈发重要。这些机器人通常依据专家策略在环境中运行，而环境的动力学会随时间变化——例如，家用吸尘机器人适应新的布局，或自动驾驶汽车在陌生城市中行驶。这类变化会改变底层的转移动力学，但机器人的性能仍应对这些变化保持稳健。此外，在机器人网络中，很可能已有另一个智能体适应了类似的环境。然而，由于每台机器人都在私有域中运行，因此必须防止泄露有关某个客户端环境的任何信息。最后，许多机器人平台主要用于推理或轻量级微调，因此缺乏计算资源，无法全面开展逆强化学习，以恢复适用于其特定运行条件的最优奖励参数。这一场景促使我们研究联邦逆强化学习问题，即以隐私保护的方式，在多个智能体之间共享并聚合学习到的奖励函数。

近期许多研究开始探索逆向强化学习的联邦变体。在许多方法中，模型融合是通过直接进行参数平均来实现的McMahan 等人(2017)，但并未考虑底层奖励景观的几何结构。尽管这些平均方案简单且通信效率高，但当局部参数拟合不佳时，往往会导致次优或不稳定的全局奖励；对于计算资源有限或专家演示不足的客户端而言，这是意料之中的结果。在高度异质的环境中，这种不稳定性可能会通过聚合传播，导致奖励估计不一致甚至不安全，进而降低整体系统性能。此外，最终得到的全局

© 2026 D. Millard 和 A. Baheri。

奖励可能无法在采用不同动力学的客户端之间保持一致的行为语义。这些局限促使人们开发具备几何感知能力的融合机制，使其即使面对含噪、有偏或训练不足的本地估计，也能生成可靠的全局奖励。

为应对这些挑战，我们将离散情形下本地学习到的每个奖励视为其所参数化的对象：定义在共享状态—动作支撑集上的函数。我们在公共格点上显式计算该函数的取值，执行平移与归一化步骤，以获得奖励质量的概率分布；随后，通过计算这些归一化测度的熵正则化瓦瑟斯坦重心来融合客户端。这种考虑几何结构的融合方式保留了空间结构，减弱了弱客户端带来的欠拟合伪影，并在存在异质性的情况下得到稳定且语义一致的全局奖励。最后，我们通过在共享特征基上进行最小二乘法，将融合后的测度投影回参数，使轻量级客户端无需共享轨迹，也不会导致模型不稳定，便能参与其中。我们的贡献如下：

1. 我们提出一个联邦逆强化学习框架，使用熵正则化瓦瑟斯坦重心融合学习到的奖励函数。2. 我们提供理论分析，确立所提重心聚合的稳定性与参数误差界，并证明在局部估计误差有界时，该方法会收敛至真实奖励函数。3. 我们通过离散和连续控制基准上的实证研究表明，与参数平均相比，重心奖励融合能够持续获得更稳定、可迁移性更强的全局奖励。

2. 相关工作

从多个角度研究了如何在动力学存在异质性的多个智能体之间学习奖励函数。在联邦与分布式逆强化学习领域，已有工作提出了去中心化优化方案和收敛保证，旨在无需集中处理专家轨迹即可恢复奖励Gong 等人(2023）；Banerjee 等人(2021）；Hassan 等人(2024）。然而，这些方法通常仍沿用传统联邦学习中的标准参数平均服务器更新，主要着眼于修改本地优化步骤，而非聚合机制本身。除纯联邦设置外，多任务和迁移式逆强化学习方法还会关联相关任务以共享统计信息，例如通过最大因果熵形式化定义、教师—学生架构和群体学习Gleave 和 Habryka(2018）；Noothigattu 等人(2021）；Liu 和 Zhu(2022）。相比之下，我们的形式化定义保留了联邦学习的核心隐私与通信约束，同时明确处理转移动力学和示范分布的异质性，以及可能导致本地奖励参数拟合不足的硬件限制。

经典逆强化学习方法包括对逆问题的早期形式化定义Ng 等人(2000)、通过特征期望匹配实现的学徒式学习Abbeel 和 Ng(2004）、最大熵（MaxEnt）逆强化学习Ziebart 等人(2008，以及生成对抗模仿学习（GAIL）等对抗变体Ho 和 Ermon(2016）。我们采用最大熵本地学习器，因其具有鲁棒性，且其梯度已得到充分研究Snoswell 等人(2020）。此外，最大熵逆强化学习可与线性函数近似相结合；我们在实验过程中采用这种方法，在保持支撑集有意义的同时，使计算复杂度处于可控范围内Song 等人(2025）。尽管我们的框架旨在保持轻量化，但我们明确

假设客户端侧的计算预算有限，这与先前强调嵌入式和移动机器人平台资源受限的研究一致Neuman 等人(2022）；Imteaj 和 Amini(2020）。在更广泛的联邦学习文献中，FedAvg 仍是标准基线McMahan 等人(2017），据我们所知，此前尚无研究在联邦化逆强化学习时提出替代的聚合技术。

最优传输提供了一种具备几何感知能力的框架，用于比较和聚合分布。熵正则化通过 Sinkhorn 算法实现高效计算Cu-turi (2013)，而 Wasserstein 重心则为概率测度的平均提供了一种有原则的定义 Agueh 和 Carlier (2011。近期研究已将最优传输应用于强化学习，用其分析逆强化学习中的奖励歧义Baheri (2023，并基于 Wasserstein 距离开发多智能体强化学习中的共识或策略更新机制 Pereira 和 Amini(2025)； Baheri(2025)； Shahrooei 和 Baheri (2025)；Salgarkar 和 Baheri(2025）。在本研究中，我们对本地学习到的奖励函数计算 Wasserstein 重心，以获得一种兼具可解释性、隐私保护性且能够抵御训练不一致性的融合奖励。

3. 预备知识与问题表述

本节介绍我们方法的组成部分，并定义其所针对的问题领域。

3.1. 逆强化学习

逆强化学习旨在恢复底层奖励函数，以解释马尔可夫决策过程（MDP）中观察到的专家行为。给定一个奖励函数未知的马尔可夫决策过程 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \gamma)$ R^*：$\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 以及专家演示 $\mathcal{D}^E = \{\tau_j\}_{j=1}^N$，其中每条轨迹$\tau_j = (s_0, a_0, \ldots, s_T, a_T)$ 表示一系列状态—动作对，目标是推断出奖励函数 ˆ ˆ R，使得在 R 下的最优策略能够复现专家的行为。形式化地说，逆强化学习 ˆ旨在求得满足 R 的函数

$$\pi^*_{\hat{R}} = \arg\max_{\pi} \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{T} \gamma^t \hat{R}(s_t, a_t)\right], \quad \text{with} \quad \pi^*_{\hat{R}} \approx \pi^*_{R^*}. \tag{1}$$

在我们的工作中，我们旨在寻找一种线性函数近似，使得 ˆ $R_\theta(s_t, a_t) = \theta^\top \phi(s_t, a_t)$，其中 $\phi(s, a)$ 表示特征向量， θ 表示奖励参数。最大熵逆强化学习将轨迹分布建模为

$$p(\tau \mid \theta) = \frac{1}{Z(\theta)} \exp\left(\sum_{t=0}^{T} \gamma^t \theta^\top \phi(s_t, a_t)\right), \tag{2}$$

其中， $Z(\theta)$ 表示用于确保归一化的配分函数。该目标函数旨在最大化此分布下专家轨迹的似然，同时保持最大熵。

3.2. Wasserstein 重心

最优传输提供了一种有理论依据的数学框架，通过考察将一个概率分布转变为另一个所需的最小传输成本，来衡量二者之间的差异。

给定度量空间 $(\mathcal{X}, d)$ 上的测度 μ 和 ν，2-Wasserstein 距离定义为

$$W_2^2(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^2 \, d\gamma(x,y), \tag{3}$$

其中，$\Pi(\mu,\ \nu)$ 是所有耦合的集合，其边缘分布为 μ 和 ν。直观而言，W_2 表示将一个概率分布转换为另一个概率分布所需的最小工作量，不仅体现逐点差异，也捕捉它们之间的几何关系。在此基础上，Wasserstein 重心将算术平均数的概念推广到赋有瓦瑟斯坦度量的概率测度空间。给定一组分布 $\{\mu_i\}_{i=1}^N$ 和总和为 1 的非负权重 $\{\alpha_i\}_{i=1}^N$，其重心 $\bar{\mu}$定义为下式的极小化解

$$\bar{\mu} = \arg\min_{\mu\in\mathcal{P}_2(\mathcal{X})} \sum_{i=1}^{N} \alpha_i W_2^2(\mu,\mu_i), \tag{4}$$

其中，$\mathcal{P}_2(\mathcal{X})$ 表示具有有限二阶矩的概率测度空间。因此，重心代表输入分布在几何意义上最一致的“平均值”，它相对于 Wasserstein 距离处于中心位置。

3.3. 问题表述

我们研究横向联邦强化学习问题 Qi 等人(2021）。我们从 N 个客户端开始，其中每个客户端 i 可以对应单个智能体或一组协作智能体。令 $\mathcal{M}_i = (\mathcal{S},\ \mathcal{A},\ P_i,\ \gamma)$ 表示客户端 i的马尔可夫决策过程，其中 $\mathcal{S}$ 是共享状态空间，$\mathcal{A}$ 是共享动作空间，P_i : $\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$ 表示客户端特有的转移动力学，并且$\gamma\in(0,\ 1)$ 是折扣因子。

假设 1（共享有限支撑与度量）所有客户端共享一个公共的有限状态-动作空间 $X = S\times A$，其中 $|X| = n$ 配备基础度量 d，其直径为 D，最小非零间距为 $\delta > 0$。在连续控制实验中，该有限空间对应于状态—动作联合域上的离散化格点，用于构造传输成本。

接下来，每个客户端 F_i 都可以访问一个包含专家演示数据的本地数据集

$$\mathcal{D}_i^E = \{\tau_j^{(i)}\}_{j=1}^{N_i}, \quad \text{where} \quad \tau_j^{(i)} = (s_0, a_0, s_1, a_1, \ldots, s_T, a_T) \tag{5}$$

表示专家在未知奖励函数 R^* 下生成的一条状态—动作对轨迹：$\mathcal{S}\times\mathcal{A}\to\mathbb{R}$。联邦逆强化学习的根本挑战在于，在不向其他客户端或中央服务器披露任何专家数据或行为信息 ˆ 的情况下，恢复一个能够解释所有客户端专家行为的共享奖励函数，同时兼顾动力学和演示数据两方面的异质性。此外，所恢复的奖励函数 R 不仅应能解释各本地马尔可夫决策过程 $\mathcal{M}_i$ 中观测到的专家演示，还应能泛化到已见和未见环境。形式化地说，ˆR 应满足 ˆ$R\approx R^*$，且在 $\bigcup_{i=1}^N \mathcal{M}_i$，上具备可迁移性，使得对于任意新的马尔可夫决策过程 $\mathcal{M}_{\text{new}} = (\mathcal{S},\ \mathcal{A},\ P_{\text{new}},\ \gamma)$，若其转移动力学 P_{new}取自与 $\{P_i\}_{i=1}^N$ 相同的分布族，则其诱导的最优策略

$$\pi_{\hat{R}}^* = \arg\max_{\pi}\ \mathbb{E}_{\pi,P_{\text{new}}}\left[\sum_{t=0}^{T}\gamma^t \hat{R}(s_t,a_t)\right] \tag{6}$$

在真实奖励 R^*下与专家策略 $\pi_{R^*}^*$ 高度匹配。

假设 2（特征映射与条件数）奖励由一个满列秩的共享特征基 $\Phi \in \mathbb{R}^{n\times d}$ 表示。令 $\sigma_{\min}$（Φ）表示其最小奇异值，并定义κ_Φ：$= \sigma_{\min}$（Φ）$^{-1} < \infty$。

4. 方法

本节介绍我们的方法。与标准联邦学习在多轮迭代中交替进行本地更新和全局聚合不同，我们只执行一次聚合步骤，使用所提出的重心方法将本地学习到的奖励融合一次。

4.1. 客户端学习

每个客户端都使用最大熵逆强化学习框架，独立求解一个本地逆强化学习问题。对于客户端 F_i，我们假设奖励采用线性参数化形式 $R_i(s,a) = \theta_i^T \phi(s,a)$，其中 $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ 是特征映射，$\theta_i \in \mathbb{R}^d$ 是奖励参数。智能体 i 的最大熵逆强化学习目标函数为：

$$\mathcal{L}_i(\theta_i) = \mathbb{E}_{\tau\sim\mathcal{D}_i^E}\left[\sum_{t=0}^{T}\gamma^t\theta_i^T\phi(s_t,a_t)\right] - \log Z_i(\theta_i) - \frac{\lambda}{2}\|\theta_i\|_2^2 \tag{7}$$

其中，$Z_i(\theta_i)$ 是配分函数。目标函数关于 θ_i 的梯度为：$\nabla_{\theta_i}\mathcal{L}_i = \mu_{\mathcal{D}_i^E} - \mu_{\pi_{\theta_i}} - \lambda\theta_i$ 其中，$\mu_{\mathcal{D}_i^E}$表示经验特征期望，$\mu_{\pi_{\theta_i}}$ 表示当前奖励参数所诱导策略下的期望特征计数。

4.2. 聚合

本地学习后，每个客户端 F_i 都拥有参数 θ_i^*，这些参数生成奖励函数 $R_i(s,a) = (\theta_i^*)^T\phi(s,a)$。随后，这些参数会被发送到中央服务器。该服务器假设可以访问一个探测集（一个与客户端环境相似的确定性环境）。接着，我们利用最优传输理论，将这些异构奖励函数聚合为一个共识奖励。我们将每个奖励函数视为一个符号测度，并求其 Wasserstein 重心：

$$R^* = \arg\min_R \sum_{i=1}^{K}\alpha_i W_2^2(R,R_i) \tag{8}$$

其中，α_i 是聚合权重，W_2 是 2-Wasserstein 距离。我们通过正向平移和归一化，将每个奖励函数转换为概率测度，从而使问题易于处理。此外，直接计算 Wasserstein 重心的计算开销过高。我们采用熵正则化，得到一个可通过 Sinkhorn-Knopp 算法高效求解的平滑问题。随后，通过交替优化并结合动量更新来计算 Wasserstein 重心，以加快收敛。

4.3. 参数恢复

计算重心奖励函数 R^* 后，我们通过求解最小二乘问题来恢复相应的参数 θ^*：

$$\theta^* = \arg\min_\theta \sum_{s\in\mathcal{S}}\|R^*(s,a) - \theta^T\phi(s,a)\|_2^2 \tag{9}$$

该问题的闭式解为 $\theta^* = (\Phi^T\Phi)^{-1}\Phi^T r^*$，其中 Φ 是特征矩阵，r^* 是向量化重心奖励。

5. 理论结果

现在，我们分析 Wasserstein 重心融合对学习到的奖励和下游控制性能的影响。

定理1（稳定性与参数误差界）假设 *1* 和 *2* 成立。每个智能体 $i \in \{1, \dots, K\}$ 运行局部最大熵逆强化学习，并输出一个奖励向量 $\hat{r}_i \in \mathbb{R}^n$。固定一个平移$\sigma > 0$，使得 $\hat{r}_i + \sigma\mathbf{1}$ 的所有条目以及共享总体奖励 $r^\star$ 的所有条目均为正，并定义

$$T_\sigma(r) \;=\; \frac{r + \sigma\mathbf{1}}{\mathbf{1}^\top(r + \sigma\mathbf{1})} \in \Delta_n, \quad p_i := T_\sigma(\hat{r}_i), \tag{10}$$

$$p^\star := T_\sigma(r^\star), \tag{11}$$

$$Z(r) := \mathbf{1}^\top(r + \sigma\mathbf{1}), \tag{12}$$

其中 $Z_{\min} := \min\big(\min_i Z(\hat{r}_i)\;\; Z(r^\star)\big)$。令 $\bar{p}$ 为 $\{p_i\}_{i=1}^K$ 在权重 $\alpha \in \Delta_K$下的 2-*Wasserstein* 重心，并通过 $\bar{r}_Z := Z\,\bar{p} - \sigma\mathbf{1}$ 映射回奖励，其中 $Z > 0$为选定的标量。假设局部逆强化学习精度为 $\|\hat{r}_i - r^\star\|_1 \le \varepsilon_i\,(i = 1,\; \dots,\; K)$。则：

$$W_2(\bar{p},\, p^\star) \le \frac{2D}{\sqrt{Z_{\min}}}\Big(\textstyle\sum_{i=1}^K \alpha_i\varepsilon_i\Big)^{1/2}, \tag{13}$$

并且，如果我们选择 $Z = Z(r^\star)$ 使得 $r^\star = Z(r^\star)\,p^\star - \sigma\mathbf{1}$，那么

$$\|\bar{r}_{Z(r^\star)} - r^\star\|_2 \le \frac{2\sqrt{2}\,D\,Z(r^\star)}{\delta\sqrt{Z_{\min}}}\Big(\textstyle\sum_{i=1}^K \alpha_i\varepsilon_i\Big)^{1/2}. \tag{14}$$

此外，令 $\theta^\star = \operatorname{argmin}_\theta \|\Phi\theta - r^\star\|_2$ 且 $\hat{\theta} = (\Phi^\top\Phi)^{-1}\Phi^\top\bar{r}_{Z(r^\star)}$，则参数误差满足

$$\|\hat{\theta} - \theta^\star\|_2 \;\le\; \kappa_\Phi\,\frac{2\sqrt{2}\,D\,Z(r^\star)}{\delta\sqrt{Z_{\min}}}\Big(\textstyle\sum_{i=1}^K \alpha_i\varepsilon_i\Big)^{1/2}, \tag{15}$$

其中 κ_Φ 与假设*2*中的定义相同。

证明 [证明概要] (1) 平移归一化关于 ℓ_1 是 Lipschitz 连续的。对于 $p = T_\sigma(r)$、$q = T_\sigma(s)$，可通过商法则界证明$\|p-q\|_1 \le (2/Z_{\min})\|r-s\|_1$。(2) 从 ℓ_1 到 W_2。在有限度量空间上，$W_2^2(\mu,\nu) \le D^2\,\mathrm{TV}(\mu,\nu) = \frac{D^2}{2}\|\mu-\nu\|_1$。一结合 (1) 可得 $W_2^2(p_i,p^\star) \le (D^2/Z_{\min})\,\varepsilon_i$。(3) 重心收缩。根据三角不等式以及 Wasserstein 重心的最优性（参见 §II-F），$W_2^2(\bar{p},p^\star) \le 4\sum_i \alpha_i W_2^2(p_i,p^\star)$，从而得到 (13)。(4) 回到奖励和参数。由 $Z = Z(r^\star)$，$\bar{r}_{Z(r^\star)} - r^\star = Z(r^\star)(\bar{p} - p^\star)$ 且 $\|\bar{p} - p^\star\|_2 \le \sqrt{2}\,W_2(\bar{p},p^\star)/\delta$，可推出 ˆ (14)。最后$\|\theta - \theta^\star\|_2 \le \kappa_\Phi\,\|\bar{r}_{Z(r^\star)} - r^\star\|_2$，得到 (15)。■

定理1 意味着两个关键性质。首先，聚合不会放大噪声：如果每个本ˆ地奖励都与 ε_i相差不超过 $r^\star$，那么融合奖励 $\bar{r}$ 及其恢复出的参数向量 θ 分别与$O((\sum_i \alpha_i\varepsilon_i)^{1/2})$相差不超过 $r^\star$ 和 $\theta^\star$。其次，这种依赖仅取决于

本地误差的加权平均，因此单个较弱的客户端无法主导融合模型。现在，我们将奖励空间中的这种稳定性与控制性能联系起来。接下来的结果界定了每个客户端在执行由融合奖励 $\bar{r}$ 诱导的策略时，相较于（未知的）针对 $r^\star$ 的最优策略所产生的折扣回报次优差距。

定理 2（重心融合的策略性能界）固定一个客户端 $i \in \{1,\dots,K\}$，其马尔可夫决策过程为 $M_i = (\mathcal{S}, \mathcal{A}, P_i, \gamma)$，并令 $X = \mathcal{S} \times \mathcal{A}$ 为有限的，$|X| = n$。对于任意策略 π，令

$$\rho_i^\pi(x) \;:=\; (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\,\mathbb{P}_{P_i,\pi}\big((s_t,a_t)=x\big), \qquad x \in X, \tag{16}$$

表示归一化的 γ-折扣占用度量（因此 $\|\rho_i^\pi\|_1 = 1$）。对于任意奖励向量 $r \in \mathbb{R}^n$，定义 π 在客户端 i 上的（折扣）回报为 $J_i(\pi; r) := \frac{1}{1-\gamma}\langle \rho_i^\pi, r\rangle$. 令 $\pi_r \in \arg\max_\pi J_i(\pi; r)$ 为奖励 r 在 M_i 上的最优策略。那么对于任意 $r, s \in \mathbb{R}^n$,

$$0 \;\le\; J_i(\pi_r; r) - J_i(\pi_s; r) \;\le\; \frac{2}{1-\gamma}\,\|r-s\|_\infty. \tag{17}$$

特别地，令 $r_\star$ 表示总体奖励，$\bar{r} := \bar{r}_{Z(r_\star)}$ 表示由定理*1*的重心映射和反投影生成的融合奖励，则对于每个客户端 i,

$$0 \;\le\; J_i(\pi_{r_\star}; r_\star) - J_i(\pi_{\bar{r}}; r_\star) \;\le\; \frac{2}{1-\gamma}\,\|r_\star - \bar{r}\|_\infty \;\le\; \frac{2}{1-\gamma}\,\|r_\star - \bar{r}\|_2. \tag{18}$$

结合 (1 8) 与定理*1*的奖励空间界，可得显式的端到端保证

$$J_i(\pi_{r_\star}; r_\star) - J_i(\pi_{\bar{r}}; r_\star) \;\le\; \frac{4\sqrt{2}\,D\,Z(r_\star)}{(1-\gamma)\,\delta\,\sqrt{Z_{\min}}}\left(\sum_{j=1}^{K}\alpha_j\,\varepsilon_j\right)^{1/2}. \tag{19}$$

证明 由于 π_r 使 $J_i(\pi; r)$达到最大值，故不等式 (17) 的左侧立即成立。对于上界,

$$\begin{aligned} J_i(\pi_r; r) - J_i(\pi_s; r) &= \frac{1}{1-\gamma}\left\langle \rho_i^{\pi_r} - \rho_i^{\pi_s},\, r\right\rangle \\ &= \frac{1}{1-\gamma}\left\langle \rho_i^{\pi_r} - \rho_i^{\pi_s},\, r-s\right\rangle \;+\; \frac{1}{1-\gamma}\Big(\langle \rho_i^{\pi_r}, s\rangle - \langle \rho_i^{\pi_s}, s\rangle\Big). \end{aligned} \tag{20}$$

The 第二项为 ≤ 0，因为 π_s 使 $J_i(\pi; s)$ 达到最大值。应用 H¨older 不等式可得 s

$$J_i(\pi_r; r) - J_i(\pi_s; r) \;\le\; \frac{1}{1-\gamma}\,\|\rho_i^{\pi_r} - \rho_i^{\pi_s}\|_1\,\|r-s\|_\infty. \tag{21}$$

由于 ρ_i^π是定义在 X上的概率向量，$\|\rho_i^{\pi_r} - \rho_i^{\pi_s}\|_1 \le \|\rho_i^{\pi_r}\|_1 + \|\rho_i^{\pi_s}\|_1 = 2$，得到 (17)。令$r = r_\star$和 $s = \bar{r}$，得到 (18)，而 $\|\cdot\|_\infty \le \|\cdot\|_2$给出 (18) 中的最后一个不等式。最后，定理1（奖励空间界）给出

$$\|r_\star - \bar{r}\|_2 \;\le\; \frac{2\sqrt{2}\,D\,Z(r_\star)}{\delta\,\sqrt{Z_{\min}}}\Big(\sum_{j=1}^{K}\alpha_j\,\varepsilon_j\Big)^{1/2}, \tag{22}$$

并将其代入（18）得到（19）。 ■

表1：平均成功率（%），均值 ± 标准差。客户端在分布内（客户端环境）和留出（未见环境）上进行评估。 20×20 情形有20个客户端，10×10 情形有10个客户端， 5×5 情形有3个客户端。

	分布内			留出		
环境	**本地**	**Mean**	**重心**	**本地**	**Mean**	**重心**
20×20	73.7 ± 14.0	73.5 ± 14.5	$\mathbf{76.4 \pm 13.4}$	–	75.3 ± 23.0	$\mathbf{81.4 \pm 19.0}$
10×10	80.9 ± 13.9	81.6 ± 14.2	$\mathbf{89.1 \pm 10.7}$	–	76.1 ± 8.0	$\mathbf{80.3 \pm 9.9}$
5×5	84.2 ± 12.6	85.1 ± 11.9	$\mathbf{90.5 \pm 9.8}$	–	78.4 ± 7.3	$\mathbf{82.6 \pm 8.1}$

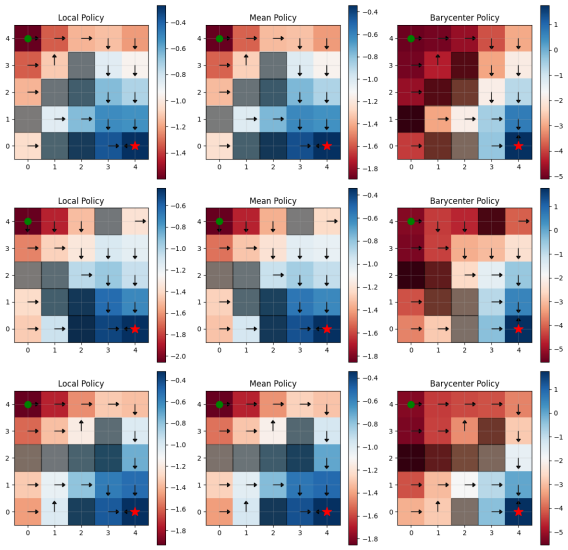


图1：三个异质的 5×5 网格世界客户端的本地奖励、均值奖励与重心奖励对比。本地最大熵逆强化学习会产生对障碍物的虚假吸引；参数平均会模糊语义。Wasserstein 重心融合保留几何结构、抑制伪影，并在各客户端上恢复平滑且与目标一致的奖励场。

不等式（19）从定量角度解释了实证结果中的性能提升，因为重心融合能够收缩奖励误差（定理1），并保证每个客户端上的诱导策略，其回报与专家相比至多相差 $O\Big((\sum_j \alpha_j \varepsilon_j)^{1/2} \Big)$，其中前置因子仅取决于（$\gamma$，$D$，$\delta$，$Z$（$r_\star$） $Z_{\min}$），而与客户端数量及其异质性无关。在本地逆强化学习误差有界的情况下，重心融合会产生：（i）与总体奖励保持接近的全局奖励和参数向量；（ii）一种策略，在总体奖励下，其在每个客户端自身动力学上的折扣回报经证明接近最优。

6. 实证结果

本节介绍两个案例研究。第一个在离散网格世界环境中进行深入分析，第二个则在一组Gymnasium基准测试中评估我们的方法。

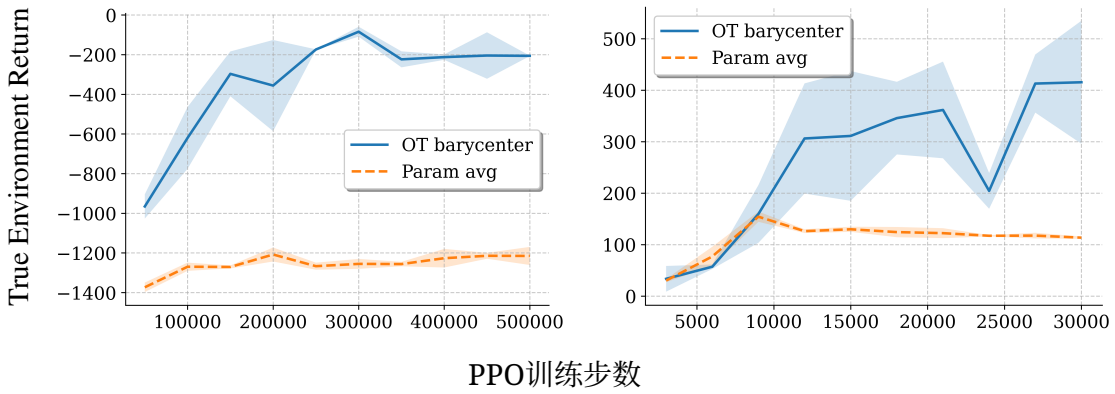


图 2：不同环境下的学习动态。左图展示 Pendulum-v1 的结果，右图展示 CartPole-v1 的结果。曲线显示了参数平均和 Wasserstein重心奖励融合下，均值 ± 标准差的真实环境回报随 PPO训练时间步的变化。

6.1. 网格世界导航

我们首先在随机网格世界导航任务中评估该框架。所有客户端共享相同的状态和动作空间，但障碍物布局和转移随机性各不相同。特征表示采用 ℓ_2 到目标和最近障碍物的距离；发生障碍物碰撞时回合终止。每个智能体的滑移概率设为 $u \sim U(0,\ 0.10)$。表1显示，在所有网格尺寸下，重心融合的表现始终优于未融合的局部奖励和简单的参数平均；在异质性最大的 20×20 环境中，提升幅度最大。融合奖励更加平滑，迁移效果也更好，因此在分布内和留出环境中都取得了更高的成功率。图1 显示，尽管若干本地学习器因覆盖范围有限而给障碍物赋予了虚假的吸引性（负）权重，重心融合仍能恢复全局一致的奖励，并为障碍物距离赋予正系数，从而鼓励更安全的导航。

6.2. Gymnasium基准测试

接下来，我们在 Gymnasium 套件中的连续控制环境上评估重心融合 Felten 等人（2023），其中每个客户端都在略有差异的物理参数（例如扭矩上限或重力）下训练，以模拟异质的机器人和环境动力学。与离散网格世界环境不同，这里的连续状态和动作空间需要能够平滑泛化的特征表示。我们采用径向基函数（RBF）Majdisova 和 Skala（2017）特征，其中心位于采样状态和动作构成的均匀格点上。为便于计算 Wasserstein 重心，我们将这些连续空间离散化，以构建传输成本的有限支撑集。图 2 中的学习曲线显示，与参数平均相比，重心融合可获得更高的回报。

6.3. 消融实验

最后，我们在随机 5×5 网格世界中进行一项消融实验，其中包含10 个异质客户端，以评估定理1提出的论断，具体而言，即没有弱客户端占据主导地位。对于弱客户端所占比例 $p \in \{0.00, 0.25, 0.50\}$，我们均匀抽取 $p \cdot K$ 个客户端，并减少其局部最大熵逆强化学习

表2：对随机 5×5 网格世界中弱客户端所占比例 p 的消融实验。报告三个随机种子下的均值 $\pm$ 标准差成功率（%）。

	分布内		留出	
p	**Mean**	**重心**	**Mean**	**重心**
0.00	81.97 ± 3.10	$\mathbf{81.97 \pm 1.97}$	86.43 ± 1.86	$\mathbf{88.27 \pm 2.21}$
0.25	84.43 ± 3.35	$\mathbf{86.20 \pm 3.15}$	84.83 ± 2.76	$\mathbf{85.50 \pm 3.02}$
0.50	80.17 ± 7.94	$\mathbf{85.27 \pm 2.76}$	82.13 ± 0.91	$\mathbf{85.67 \pm 2.73}$

迭代次数，其减少幅度由随机乘法因子决定，其余客户端则使用完整预算。根据得到的局部参数，我们构建均值融合模型和瓦瑟斯坦重心融合模型，并按照此前相同的方法评估它们的成功率。如表2所总结，当 $p = 0$时，重心融合的表现与均值基线相当；随着 p增加，其性能下降得更慢，尤其在 $p = 0.50$时始终保持优势，这与定理1的稳定性保证一致。

7. 讨论

上述结果表明，在联邦逆强化学习中，重心奖励融合为传统参数平均提供了一种具有几何感知能力的替代方案，并展现出诸多优势。在离散和连续领域中，Wasserstein 重心都提供了一种一致的聚合机制，能够捕捉共享的行为语义，同时减轻训练不足或含噪本地模型的影响。这一特性在异构环境中尤其有价值，因为直接进行参数平均可能会因特征尺度不同、局部偏差或演示数据不足而导致性能下降。尽管该方法展现出良好前景，但也存在一个显著缺点。正如技术附录中所述，可扩展性是一个核心问题。离散熵正则化瓦瑟斯坦重心的内存复杂度随支撑点数量 n 呈 $\mathcal{O}(n^2)$ 增长，因为必须存储并处理一个 $n \times n$ 代价矩阵。因此，基于Sinkhorn算法的重心求解器的每次迭代运行时间同样关于 n 呈二次增长（关于客户端分布数量呈线性增长），随着离散化不断细化，这很快会变得难以承受。在若干环境中，状态-动作空间足够大，以至于采用合理的精细离散化将需要数量多到难以承受的支撑点。在这些情况下，我们被迫采用粗粒度支撑集；而在这种粗粒度离散化下，该方法表现不佳，性能明显低于简单的参数平均。

8. 结论

本文介绍了一种联邦逆强化学习框架，通过 Wasserstein 重心融合局部最大熵逆强化学习奖励，而非采用参数平均。从理论上看，当局部估计误差有界时，重心融合具有稳定性；从实验结果来看，该方法能够在异构客户端之间产生更易解释、可迁移性更强的奖励。尽管可扩展性和覆盖分辨率仍是尚待解决的挑战，但结果表明，具备几何感知能力的奖励聚合可作为资源受限智能体之间实现可靠、隐私保护协作的基础。未来将探索连续重心形式，以及面向大规模真实世界机器人系统的个性化微调。

参考文献

Pieter Abbeel 和 Andrew Y Ng。通过逆强化学习进行学徒学习。载于第二十一届国际机器学习会议论文集，第 1 页，2004 年。Martial Agueh 和 Guillaume Carlier。Wasserstein 空间中的重心。*SIAM*数学分析杂志，43(2):904–924，2011 年。Ali Baheri。通过最优传输理论理解逆强化学习中的奖励歧义。*arXiv* 预印本 *arXiv:2310.12055*，2023 年。Ali Baheri。合作多智能体强化学习的 Wasserstein 重心共识，2025 年。网址https://arxiv.org/abs/2506.12497。Soumya Banerjee、Samia Bouzefrane 和 Amar Abane。混合区块链方法下的身份管理：结合联邦逆强化学习的扩展研究。载于*2021*年*IEEE*第*22*届高性能交换与路由国际会议 *(HPSR)*，第 1–6 页。IEEE，2021 年。Marco Cuturi。Sinkhorn距离：高速计算最优传输。神经信息处理系统进展，第 26 卷，2013 年。Florian Felten、Lucas N. Alegre、Ann Now´e、Ana L. C. Bazzan、El Ghazali Talbi、Gr´egoire Danoy 和 Bruno C. da Silva。用于多目标强化学习的可靠基准测试与研究工具包。载于第*37*届神经信息处理系统大会论文集（*NeurIPS 2023*），2023 年。Adam Gleave 和 Oliver Habryka。多任务最大熵逆强化学习。*arXiv* 预印本 *arXiv:1805.08882*，2018 年。Wei Gong、Linxiao Cao、Yifei Zhu、Fang Zuo、Xin He 和 Haoquan Zhou。应用于智能重症监护病房、具有差分隐私的联邦逆强化学习。*IEEE*物联网期刊，10(21):19117–19124，2023 年。Sheikh Salman Hassan、Yu Min Park、Yan Kyaw Tun、Walid Saad、Zhu Han 和 Choong Seon Hong。提升 6G 卫星网络的频谱效率：基于生成对抗模仿学习的策略学习，通过异步联邦逆强化学习实现。*arXiv* 预印本 *arXiv:2409.18718*，2024 年。Jonathan Ho 和 Stefano Ermon。生成对抗模仿学习。神经信息处理系统进展，第 29 卷，2016 年。Ahmed Imteaj 和 M Hadi Amini。FedAR：面向分布式移动机器人的活动与资源感知联邦学习模型。载于*2020*年*IEEE*第*19*届机器学习与应用国际会议 *(ICMLA)*，第 1153–1160 页。IEEE，2020 年。Shicheng Liu 和 Minghui Zhu。面向多智能体系统的分布式逆约束强化学习。神经信息处理系统进展，第 35 卷：33444–33456，2022 年。Zuzana Majdisova 和 Vaclav Skala。径向基函数近似：比较与应用。应用数学建模，51:728–743，2017 年。

Brendan McMahan、Eider Moore、Daniel Ramage、Seth Hampson 和 Blaise Aguera y Arcas。利用去中心化数据高效通信地学习深度网络。载于人工智能与统计学，第 1273–1282 页。PMLR，2017 年。Sabrina M Neuman、Brian Plancher、Bardienus P Duisterhof、Srivatsan Krishnan、Colby Banbury、Mark Mazumder、Shvetank Prakash、Jason Jabbour、Aleksandra Faust、Guido CHE de Croon 等。微型机器人学习：资源受限机器人机器学习面临的挑战与发展方向。载于*2022 年 IEEE 第四届人工智能电路与系统国际会议（AICAS）*，第 296–299 页。IEEE，2022 年。Andrew Y Ng、Stuart Russell 等。逆向强化学习算法。载于*ICML*，第 1 卷，第 2 页，2000 年。Ritesh Noothigattu、Tom Yan 和 Ariel D Procaccia。向志同道合的教师学习逆向强化学习。载于*AAAI*人工智能大会论文集，第 35 卷，第 9197–9204 页，2021 年。Luiz Pereira 和 M Hadi Amini。使用 Wasserstein 重心的异构联邦强化学习。*arXiv* 预印本 *arXiv:2506.15825*，2025 年。Jiaju Qi、Qihao Zhou、Lei Lei 和 Kan Zheng。联邦强化学习：技术、应用与尚待解决的挑战。*arXiv* 预印本 *arXiv:2108.11887*，2021 年。Chirayu Salgarkar 和 Ali Baheri。距离何时重要：用于多智能体协调的 Wasserstein 信赖域。2025 年。Zahra Shahrooei 和 Ali Baheri。Wasserstein重心软演员-评论家算法。*arXiv* 预印本 *arXiv:2506.10167*，2025 年。Aaron J Snoswell、Surya PN Singh 和 Nan Ye。重新审视最大熵逆向强化学习：新视角与新算法。载于*2020 年 IEEE*计算智能研讨会系列（*SSCI*），第 241–249 页。IEEE，2020 年。Li Song、Qinghui Guo、Irfan Ali Channa 和 Zeyu Wang。基于最大熵的逆向强化学习综述：方法与应用。*Symmetry*，17(10):1632，2025 年。Brian D Ziebart、Andrew L Maas、J Andrew Bagnell、Anind K Dey 等。最大熵逆向强化学习。载于*AAAI*，第 8 卷，第 1433–1438 页。美国伊利诺伊州芝加哥，2008 年。

附录 A. 实验设置

在本节 在实证结果部分中，我们详细介绍所评估的每种 Gymnasium 设置 .

A.1. Pendulum-v1

对于 Pendulum-v1，每个客户端都被分配一个物理参数略有不同的连续单连杆摆。客户端 i 使用按比例缩放的扭矩上限、阻尼系数和引力常数，其数值随 i线性增加，同时在观测中加入少量高斯过程噪声。具体而言，我们将扭矩缩放系数设为 $1.0 + 0.1 \cdot i$，阻尼缩放系数设为 $1.0 + 0.05 \cdot i$，重力设为 $9.81 + 0.2 \cdot i$，观测噪声的标准差设为 $0.005 \cdot i$。这使控制难度和对动作的敏感性存在轻微异质性，同时保持相同的稳定化目标函数。对于逆强化学习和重心融合，我们使用 --n-theta 32 --n-thdot 21 --n-action 5 对 (θ， θ， a) 空间进行离散化。

A.2. CartPole-v1

对于 CartPole-v1，每个客户端都操作一个具有不同物理缩放比例的倒立摆系统。客户端 i会根据客户端的不同，通过缩放因子调整小车作用力大小、重力、小车质量、摆杆质量和摆杆长度。具体而言，我们将作用力大小缩放为 $1.0 + 0.1 \cdot i$，将重力缩放为 $1.0 + 0.005 \cdot i$，将两种质量都缩放为 $1.0 + 0.05 \cdot i$，将摆杆长度缩放为 $1.0 + 0.05 \cdot i$，并加入标准差为 $0.005 \cdot i$ 的高斯观测噪声。这些变化会改变不同客户端平衡任务的难度和敏感度，同时保持奖励结构不变（直至失败，每步奖励为 $+1$ ）。对于重心，我们使用 --n-x 5 --n-xdot 5 --n-theta 7 --n-thetadot 5 --n-action 2 对状态-动作空间进行离散化。

A.3. LunarLanderContinuous-v3

对于 LunarLanderContinuous-v3，我们考虑一艘配备连续推力、处于异质风场中的月球着陆器。每个客户端 i 均设置了不同的风力和湍流级别，同时保持重力不变。我们将风力参数设为 $10.0 + 1.0 \cdot i$ ，将湍流强度设为 $1.0 + 0.2 \cdot i$，这会使索引较大的客户端受到更强的侧向扰动。奖励函数采用 Gymnasium 实现中的标准塑形奖励。对于重心融合，我们使用 --n-ll-x 5 --n-ll-y 5 --n-ll-vx 5 --n-ll-vy 5，在关键状态分量（位置和速度）上构建乘积网格，并对角度、角速度和支腿接触指示量进行粗粒度离散化。

A.4. Acrobot-v1

对于 Acrobot-v1，每个客户端控制一个欠驱动双连杆摆，其连杆质量、长度和力矩尺度因客户端而异。客户端 i 将两根连杆的质量和长度均缩放为原来的 1.0 $+0.05 \cdot i$，进而影响质心位置和整体惯性。我们还根据质量与长度的综合缩放系数对可用力矩进行缩放，因此不同客户端的有效驱动强度各不相同。与其他任务一样，我们在状态中加入标准差为 $0.005 \cdot i$ 的小幅高斯观测噪声。奖励仍采用标准的稀疏终止奖励（成功摆起）。对于 Wasserstein 重心，我们使用 --n-theta1 9 --n-theta2 9 --n-thdot1 5 --n-thdot2 5 --n-action 3，根据关节角、角速度和离散动作对状态—动作空间进行离散化。

附录 B. 超参数

本节记录了专家训练、逆强化学习优化、聚合和评估所用的主要超参数。表 3 列出了各环境的超参数，而表 4列出了我们所用形式化定义的特定超参数。平均融合这一参数没有任何超参数。

表 3：各阶段和环境的通用训练超参数。

超参数	倒立摆	小车	LLC	Acro
专家训练				
专家训练器 / 算法	PPO	PPO	PPO	PPO
专家总步数	50000	200000	300000	5000
专家批量大小	64	64	64	64
专家学习率	$3 \cdot 10^{-3}$	$3 \cdot 10^{-3}$	$3 \cdot 10^{-3}$	$3 \cdot 10^{-3}$
折扣因子 γ	0.99	0.99	0.99	0.99
广义优势估计 λ	0.95	0.95	0.95	0.95
裁剪范围	0.2	0.2	0.2	0.2
客户端数量	6	6	6	6
轨迹生成				
每个客户端的回合数	50	50	50	50
最大回合长度	200	500	1000	500
IRL				
逆强化学习算法	最大熵	最大熵	最大熵	最大熵
逆强化学习迭代次数／步数	50	25	100	25
逆强化学习学习率	$5 \cdot 10^{-3}$	$5 \cdot 10^{-3}$	$5 \cdot 10^{-3}$	$5 \cdot 10^{-3}$
熵权重／温度	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$
特征映射／维度	64	64	64	64
奖励正则化（例如 L2）	l_2	l_2	l_2	l_2
评估				
评估训练器／算法	PPO	PPO	PPO	PPO
评估步数 / 回合/客户端	500000	30000	2000000	25000
随机种子数量	10	10	10	10

附录 C：局限性

图 3 展示了两个连续控制环境 `LunarLanderContinuous-v3` 和`Acrobot-v1`——在这些环境中，重心融合的优势开始减弱。这种粗粒度离散化是维持计算可处理性所必需的，但也不可避免地限制了恢复奖励的精度以及所得策略梯度的平滑性。这揭示了当前框架的两个主要局限。首先是可扩展性：计算 Wasserstein 重心的成本会随着离散化支撑集的维度和分辨率迅速增加，因为即使采用熵正则化，内存和

表 4：所提方法的 Wasserstein 重心融合超参数。

超参数	倒立摆	小车	LLC	Acro
支撑离散化				
状态网格分辨率	32×21	$5^3 \times 7$	$5^4 \times 1^3$	$9^2 \times 5^2$
动作网格分辨率	5	2	3	5
支撑点总数 n	3360	1750	1875	10125
最优传输重心求解器				
熵正则化 ε	0.5	0.5	0.5	0.5
每个重心步骤的 Sinkhorn迭代次数	10	10	10	10
重心迭代次数	50	50	50	50

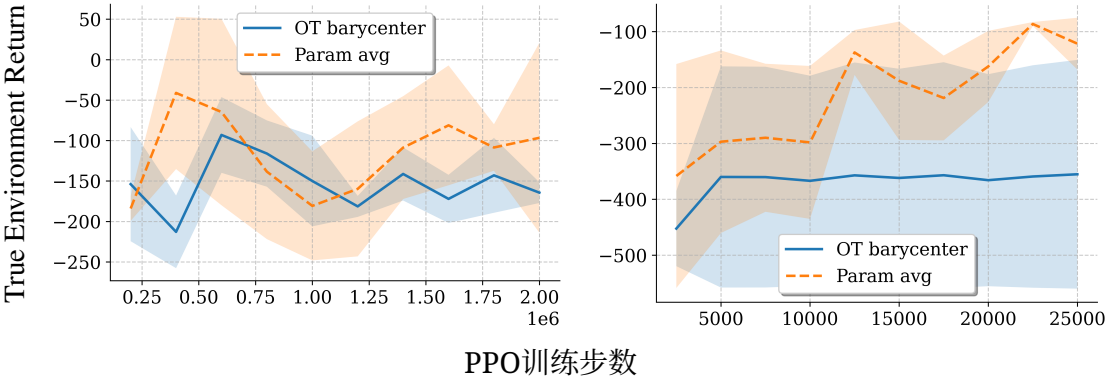


图 3：高维环境中的学习动态。在参数平均和 Wasserstein 重心奖励融合下，真实环境回报的均值 $\pm$ 标准差随 PPO 训练时间步的变化情况。由于支撑集维度较高，这些案例需要采用粗粒度离散化。

运行时间仍会随箱数量呈二次增长。其次是对探索不足状态的分辨率：当专家演示的覆盖范围有限时，融合奖励往往会过度平滑，或回归到熵先验，从而降低其在状态空间低密度区域的保真度。

附录D. 代码可用性

我们的代码仓库采用 Apache License 2.0 许可。所有检查点、输出和数据均可向通讯作者提出申请获取。