LS-IQ：用于逆强化学习的隐式奖励正则化

Firas Al-Hafez[1], **Davide Tateo**[1], **Oleg Arenz**[1],**Guoping Zhao**[2],**Jan Peters**[1,3]
[1] 智能自主系统，运动实验室[32]德国人工智能研究中心 (DFKI)、认知科学中心、Hessian.AI 达姆施塔特工业大学，德国{name.surname}@tu-darmstadt.de

arXiv:2303.00599v1 [cs.LG] 1 Mar 2023

摘要

近期的模仿学习方法直接使用隐式奖励形式学习 Q 函数，而非显式奖励函数。然而，这些方法通常需要对隐式奖励进行正则化以提高稳定性，并且往往会错误处理吸收状态。先前的研究表明，对隐式奖励函数进行平方范数正则化是有效的，但没有对由此产生的算法性质进行理论分析。在这项工作中，我们表明，在策略与专家的混合分布下使用该正则化项，可以提供一个特别有启发性的视角：原始目标函数可以被理解为平方贝尔曼误差最小化，而相应的优化问题则最小化专家分布与混合分布之间有界的 χ^2-散度。这一视角使我们能够解决不稳定问题并恰当处理吸收状态。我们表明，我们的方法——最小二乘逆Q学习（LS-IQ）——优于最先进算法，尤其是在包含吸收状态的环境中。最后，我们提出使用逆动力学模型，仅根据观测进行学习。采用这种方法后，即使在没有专家动作可用的情况下，我们也能保持性能。[1]

1 引言

逆强化学习（IRL）技术旨在从专家示范中稳健地提取行为，并解决经典模仿学习（IL）方法所面临的问题（Ng 等人，1999年；Ziebart 等人，2008年）。在近期的逆强化学习方法中，对抗模仿学习（AIL）方法（Ho & Ermon，2016年；Fu 等人，2018年；Peng 等人，2021年）将奖励和策略的优化置于对抗框架中，已被证明取得了尤为出色的效果。这些方法受到生成对抗网络（GANs）的启发（Goodfellow 等人，2014年），在学习判别器与根据判别器输出计算奖励函数并改进智能体策略之间交替进行。这些显式奖励方法在学习奖励函数和价值函数时需要与环境进行大量交互。近期，研究者提出了隐式奖励方法（Kostrikov 等人，2020年；Arenz 和 Neumann，2020年；Garg 等人，2021年）。这些方法直接学习 Q-函数，从而显著加速策略优化。在隐式奖励方法中，逆软Q学习（IQ-Learn）是当前最先进的方法。该方法通过在专家分布上加入奖励正则化来修改分布匹配目标，从而实现最小化策略分布与专家分布之间的 χ^2-散度。然而，尽管其推导仅考虑了专家分布上的正则化，但在连续控制任务中的实际实现表明，同时对专家分布和策略分布上的奖励进行正则化能够取得显著更好的性能。

本文的贡献有两方面：首先，我们证明，使用这一正则化项时，所得目标函数会最小化专家与专家和策略之间的混合分布之间的 χ^2 散度。随后，我们研究了相对于混合分布进行正则化对 IQ-Learn 理论性质的影响。我们证明该散度有界，这意味着奖励和

[1]代码可在 https://github.com/robfiras/ls-iq 获取

Q函数的取值也有界，从而显著提高了学习稳定性。事实上，所得目标函数对应于最小二乘贝尔曼误差最小化，并且与软Q模仿学习（SQIL）（Reddy 等人，2020）密切相关。其次，我们提出了最小二乘逆Q学习（LS-IQ），这是一种新颖的逆强化学习算法。遵循对 χ^2 正则化项的分析所揭示的理论见解，我们解决了 IQ-Learn 方法中的多种不稳定因素：Q函数尺度的不确定性、Q函数目标值的爆炸，以及奖励偏差（Kostrikov 等人，2019），即假设吸收状态提供零奖励。我们利用 Q函数的结构性质以及基于专家最优性的启发式方法，推导出 LS-IQ 算法。这使得该算法在许多任务上的性能有所提升，并且总体上学习更加稳定，Q函数估计的方差也更小。最后，我们将隐式奖励方法扩展到基于观测的模仿学习场景，通过训练逆动力学模型（IDM）来预测专家动作，因为这些动作不再假定为可用。即使在这一具有挑战性的场景中，我们的方法仍能达到与已知专家动作时相近的性能。

相关工作。绝大多数逆强化学习和模仿学习方法都建立在最大熵（MaxEnt）逆强化学习框架之上（Ziebart，2010年）。特别是，Ho & Ermon (2016) 提出了生成对抗模仿学习（GAIL），将生成对抗网络应用于模仿学习问题。原始方法通过最小化与专家分布之间的詹森-香农散度来实现目标；在 Nowozin 等人（2016）的研究基础上，这一方法被扩展到一般的 f散度（Ghasemipour 等人，2019年）。在 f散度中，皮尔逊 χ^2 散度能够提升生成对抗网络（Mao 等，2017年）和对抗模仿学习（Peng 等人，2021年）的训练稳定性。Kostrikov 等人（2019年）引入回放缓冲区，用于策略和判别器的离策略更新。作者还指出了奖励偏差这一问题，它在许多模仿学习方法中都很常见。事实上，对抗模仿学习方法会隐式地为这些状态赋予零奖励，从而根据所选散度产生生存偏差或终止偏差。Kostrikov 等人（2020年）在先前工作的基础上引入了离线策略评估领域的最新进展（Nachum 等人，2019年）。他们的方法 ValueDice 使用逆贝尔曼算子，以 Q函数表示奖励函数，并最小化与专家分布之间的反向库尔贝克-莱布勒散度（KL散度）。Arenz 和 Neumann（2020年）基于策略的信赖域更新，推导出一种非对抗性形式。他们的方法 O-NAIL 使用标准的软演员评论家（SAC）更新来改进策略（Haarnoja等人，2018）。O-NAIL 可以视为更通用的 IQ-Learn 算法（Garg 等人，2021年）的一个实例；该算法可根据隐式奖励正则项优化不同的散度。Garg 等人（2021年）还表明，与反向KL散度相比，使用 χ^2 散度能让其算法取得更好的性能。Reddy 等人（2020年）提出了一种使用 SAC 并为专家和策略赋予固定二元奖励的方法。Swamy 等人（2021年）为上述许多方法提供了统一的视角，明确指出 GAIL、ValueDice、MaxEnt-IRL 和 SQIL 都可以看作矩匹配算法。最后，Sikchi 等人（2023年）提出一种用于对抗模仿学习的排序损失，通过使用带排序目标值的最小二乘目标函数来训练奖励函数。

2 预备知识

记号。马尔可夫决策过程（MDP）是一个元组（$\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu_0$），其中 $\mathcal{S}$是状态空间，$\mathcal{A}$是动作空间，P：$\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^+$ 是转移核，r：$\mathcal{S} \times \mathcal{A} \to \mathbb{R}$是奖励函数，$\gamma$ 是折扣因子，μ_0：$\mathcal{S} \to \mathbb{R}^+$ 是初始状态分布。在每一步中，智能体从环境中观察状态 $s \in \mathcal{S}$，使用策略 π：$\mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$采样动作 $a \in \mathcal{A}$，并以概率 $P(s'|s,a)$ 转移到下一个状态 $s' \in \mathcal{S}$，同时获得奖励 $r(s,a)$。我们定义占用度量 $\rho_\pi(s,a) = \pi(a|s) \sum_{t=0}^{\infty} \gamma^t \mu_t^\pi(s)$，其中 $\mu_t^\pi(s') = \int_s, _a \mu_t^\pi(s)\pi(a|s)P(s'|s,a)da\,ds$ 是给定 $t > 0$时的状态分布，且 $\mu_0^\pi(s) = \mu_0(s)$。利用占用度量，我们可以将策略 π 下的期望奖励表示为 $\mathbb{E}_{\rho_\pi}[r(s,a)] \triangleq \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t)]$,，其中对于 $t > 0$，$s_0 \sim \mu_0$，$a_t \sim \pi(.|s_t)$，以及 $s_{t+1} \sim P(.|s_t,a_t)$。此外，$\mathbb{R}^{\mathcal{S}\times\mathcal{A}} = \{x$：$\mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$表示状态-动作空间上的函数集合，$\mathbb{R}$ 表示扩展实数 $\mathbb{R} \cup \{+\infty\}$。我们将软价值函数记为 $V(s)$ 和 $Q(s,a)$，而使用 $V(s)$ 和 $Q(s,a)$ 表示不含熵奖励的价值函数。

将逆强化学习视为占用匹配问题。给定一组由专家策略 π_E采样的状态和动作演示，逆强化学习旨在从奖励函数族 $\mathcal{R} = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ 中找到一个奖励函数 $r(s,\ a)$，为

专家策略 π_E 的样本赋予高奖励，并为其他策略赋予低奖励。我们考虑 Ho & Ermon (2016) 提出的框架，该框架从占用匹配问题中推导出带有额外凸奖励正则项 ψ_ρ 的最大熵逆强化学习目标函数：$\mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}$

$$\max_{r\in\mathcal{R}} \min_{\pi\in\Pi} L_\rho(r,\pi) = \max_{r\in\mathcal{R}} \left(\min_{\pi\in\Pi} -\beta H_\rho(\pi) - \mathbb{E}_{\rho_\pi}[r(s,a)] \right) + \mathbb{E}_{\rho_{\pi_E}}[r(s,a)] - \psi_\rho(r), \tag{1}$$

其中，策略空间为 $\Pi = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$，折扣累积熵奖励为 $H_\rho(\pi) = \mathbb{E}_{\rho_\pi}[-\log(\pi(a|s))]$，，而常数 β 用于控制熵奖励。请注意，内层优化是一个最大熵强化学习（RL）目标函数（Ziebart，2010年），其最优策略为

$$\pi^*(a|s) = \frac{1}{Z_s} \exp\left(\tilde{Q}(s,a)\right), \tag{2}$$

其中，$Z_s = \int_{\hat{a}} \exp\ \tilde{Q}(s,\hat{a})\, d\hat{a}$ 是配分函数，$\tilde{Q}(s,a)$ 是软动作价值函数，对于给定的奖励函数，该函数由软贝尔曼算子定义：$(\mathcal{B}^\pi Q)(s,a) = r(s,a) + \gamma \mathbb{E}_{s'\sim P(.|s,a)} V^\pi(s')$，其中 $V^\pi(s') = \mathbb{E}_{a\sim\pi(.|s)}[Q(s,a) - \log\ \pi(a|s)]$。

Garg 等人（2021年）利用逆软贝尔曼算子 $(\mathcal{T}^\pi Q)(s,\ a) = Q(s,\ a) - \gamma \mathbb{E}_{s'\sim P(.|s,\ a)} V^\pi(s')$，将公式 1 从奖励-策略空间变换到 Q-策略空间，从而建立奖励与 Q-函数之间的一一对应关系。该算子可以将目标函数L_ρ 从奖励-策略空间变换到 Q-策略空间，以下记作 $\mathcal{J}_\rho$

$$\max_{r\in\mathcal{R}} \min_{\pi\in\Pi} L_\rho(r,\pi) = \max_{\tilde{Q}\in\tilde{\Omega}} \min_{\pi\in\Pi} \mathcal{J}_\rho(\tilde{Q},\pi)\,, \tag{3}$$

其中 $\tilde{\Omega} = \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ 是 $\tilde{Q}$-函数的空间。此外，他们利用公式 2，根据给定的 Q-函数提取最优策略 $\pi_{\tilde{Q}}$，以去除公式 1 中的内层优化循环，因此

$$\begin{aligned}\max_{\tilde{Q}\in\tilde{\Omega}} \mathcal{J}_\rho(\tilde{Q},\pi_{\tilde{Q}}) = \max_{\tilde{Q}\in\tilde{\Omega}}\ & \mathbb{E}_{\rho_{\pi_E}}\left[\tilde{Q}(s,a) - \gamma \mathbb{E}_{s'\sim P(.|s,a)}[\tilde{V}^\pi(s')]\right] - \beta H_\rho(\pi_{\tilde{Q}}) \\ & - \mathbb{E}_{\rho_\pi}\left[\tilde{Q}(s,a) - \gamma \mathbb{E}_{s'\sim P(.|s,a)}[\tilde{V}^\pi(s')]\right] - \psi_\rho(\tilde{\mathcal{T}}^\pi \tilde{Q}).\end{aligned} \tag{4}$$

实用奖励正则化。Garg 等人（2021年）推导出一个正则化项，对来自专家的状态-动作对上的奖励施加 L_2 范数惩罚，使得 $\psi_{\pi_E}(r) = c\,\mathbb{E}_{\rho_{\pi_E}}\left[r(s,\ a)^2\right]$，其中 c 为正则化常数。然而，在连续动作空间中，该正则化项会导致不稳定。在实践中，Garg 等人（2021年）通过将该正则化项应用于混合分布来解决这一不稳定问题

$$\psi_\rho(r) = \alpha\, c\, \mathbb{E}_{\rho_{\pi_E}}\left[r(s,a)^2\right] + (1-\alpha)\, c\, \mathbb{E}_{\rho_\pi}\left[r(s,a)^2\right], \tag{5}$$

其中 α 通常设为。。需要注意的是，这一变化使得我们无法再直接从公式 1 中提取策略。事实上，公式 5 中的正则化项也取决于策略。以往的研究并未解决这一问题。在接下来的章节中，我们将深入分析这一正则化项，从而解决上述问题并推导出正确的策略更新。在介绍我们的方法之前，我们使用附录 A 中的命题 A.1，将目标函数 L_ρ 和 $\mathcal{J}_\rho$ 从占用度量下的期望转换为状态-动作分布 d_{π_E} 和 d_π下的期望，以下分别记为 L 和 $\mathcal{J}$。

3 最小二乘逆 Q 学习

在本节中，我们介绍所提出的模仿学习算法。该算法基于公式 1 中提出的占用匹配问题，并使用公式 5 所定义的正则化项。我们首先将所得目标函数解释为专家分布与专家和策略的混合分布之间的 χ^2 散度。随后我们将展示，该正则化项使我们能够将原始目标函数转化为一个贝尔曼误差最小化问题，其中专家和策略分别具有固定的二元奖励。具有固定奖励的强化学习问题是一种独特的设定，我们可以利用它来界定 Q函数目标值，为专家状态上的 Q函数提供固定目标值而无需进行自举，并恰当地处理吸收状态。然而，这些技术需要应用于硬 Q函数。因此，通过引入额外的熵评论家，我们将软动作价值函数 $\tilde{Q}$ 切换为硬 Q函数。我们还提出了一个正则化评论家，使我们能够恢复与公式 5 中正则化项相对应的正确策略更新。最后，我们提出使用 IDM 来解决基于观测的模仿学习问题。

3.1 将其解释为统计散度

Ho & Ermon (2016) 表明，他们的正则化项会导致专家与策略的状态-动作分布之间的詹森-香农散度最小化。类似地，Garg 等人（2021）表明，他们的正则化项 $\psi_{\pi_E}(r)$ 会导致 χ^2 散度最小化。然而，公式 5 中提出的正则化项尚未得到研究。我们将证明，该正则化项会使专家状态-动作分布与专家和策略之间的混合分布之间的 χ^2 散度最小化。因此，我们从公式 1 中给出的目标函数出发，并注意到，根据极小极大定理（冯·诺伊曼，1928年），强对偶性——$\max_{r\in\mathcal{R}}\min_{\pi\in\Pi} L = \min_{\pi\in\Pi}\max_{r\in\mathcal{R}} L$——可直接推出，因为 $-H(\pi)$、$-\mathbb{E}_{d_\pi}[r(s,a)]$ 关于 π 是凸的，而 $-\mathbb{E}_{d_\pi}[r(s,a)]$、$\mathbb{E}_{d_{\pi_E}}[r(s,a)]$ 和 $\psi(r)$ 关于 r 是凹的（Ho & Ermon，2016）。我们使用其变分形式表示专家分布与混合分布之间的 χ^2 散度，

$$\begin{aligned}2\chi^2\big(d_{\pi_E}\,\|\,\underbrace{\tfrac{d_{\pi_E}+d_\pi}{2}}_{d_{\text{mix}}}\big) &= \sup_r 2\Big(\mathbb{E}_{d_{\pi_E}}[r(s,a)] - \mathbb{E}_{\tilde{d}}\Big[r(s,a)+\tfrac{r(s,a)^2}{4}\Big]\Big)\\ &= \sup_r \mathbb{E}_{d_{\pi_E}}[r(s,a)] - \mathbb{E}_{d_\pi}[r(s,a)] - c\alpha\mathbb{E}_{d_{\pi_E}}\left[r(s,a)^2\right] - c(1-\alpha)\mathbb{E}_{d_\pi}\left[r(s,a)^2\right], \quad (6)\end{aligned}$$

正则化常数为 $c = 1/2$ 和 $\alpha = 1/2$。现在，如果最优奖励属于 $\mathcal{R}$，则公式 1 中的原始目标函数变为熵正则化的 χ^2 散度最小化问题

$$\max_{r\in\mathcal{R}}\min_{\pi\in\Pi} L = \min_{\pi\in\Pi} 2\chi^2\big(d_{\pi_E}\,\|\,\tfrac{d_{\pi_E}+d_\pi}{2}\big) - \beta H(\pi)\,. \quad (7)$$

公式 7 表明，正则化逆强化学习目标通过优化奖励来匹配某种散度，同时通过优化策略来最小化该散度。当所匹配的散度无界时，最优奖励也会无界，从而导致学习过程中的不稳定性。不同于智能体分布与专家分布之间的 χ^2-散度，关于混合分布的 χ^2-散度有界为 $[0, 1/2]$，如命题 A.3 所示，并且其最优奖励

$$r^*(s,a) = \frac{1}{c}\frac{d_{\pi_E}(s,a) - d_\pi(s,a)}{d_{\pi_E}(s,a) + d_\pi(s,a)}\,, \quad (8)$$

也有界于区间 $[-1/c, 1/c]$ 内，如命题 A.2 所示。

3.2 从强化学习角度看分布匹配

接下来，我们将从一个新颖的视角来阐释公式 4，以便更好地理解正则化项的作用。事实上，对于公式 5 中定义的正则化项，我们可以将该目标函数解释为一个熵正则化最小二乘问题，如下列命题所示：

命题 3.1 若 $r_{\tilde{Q}}(s,a) = (\mathcal{T}^\pi Q)(s,a)$ 是一个 Q-函数的隐式奖励函数，那么对于 $\psi(r_{\tilde{Q}}) = c\,\mathbb{E}_{\tilde{d}}[r_{\tilde{Q}}(s,a)^2]$，其中 $\tilde{d}(s,a) = \alpha d_{\pi_E}(s,a) + (1-\alpha)d_\pi(s,a)$，公式 4 在状态-动作分布下的解等于一个熵正则化最小二乘最小化问题的解，使得 $\arg\min_{\tilde{Q}\in\Omega}\mathcal{L}(Q,\pi_{\tilde{Q}}) = \arg\max_{\tilde{Q}\in\Omega}\mathcal{J}(Q,\pi_{\tilde{Q}})$，其中

$$\mathcal{L}(\tilde{Q},\pi_{\tilde{Q}}) = \alpha\mathbb{E}_{d_{\pi_E}}\left[\left(r_{\tilde{Q}}(s,a) - r_{\max}\right)^2\right] + (1-\alpha)\mathbb{E}_{d_{\pi_{\tilde{Q}}}}\left[\left(r_{\tilde{Q}}(s,a) - r_{\min}\right)^2\right] + \frac{\beta}{c}H(\pi_{\tilde{Q}})\,, \quad (9)$$

其中 $r_{\max} = \frac{1}{2\alpha c}$ 和 $r_{\min} = -\frac{1}{2(1-\alpha)c}$。

证明见附录 A.3。公式 9 中得到的目标函数与最小二乘生成对抗网络（LSGANs）（Mao 等，2017）设定中的目标函数非常相似，其中 $r_{\tilde{Q}}(s,a)$ 可解释为判别器，$r_{\max}$ 可解释为专家样本的目标值，而 $r_{\min}$ 可解释为策略 π下样本的目标值。对于 $\alpha = 0.5$ 和$c = 1$，从而得到 $r_{\max} = 1$ 和 $r_{\min} = -1$，公式 9 与 LSGANs 设定中判别器的目标函数仅在熵项上有所不同。

现在，将隐式奖励函数替换为逆软贝尔曼算子，并重新整理各项，得到

$$\begin{aligned}\mathcal{L}(\tilde{Q},\pi_{\tilde{Q}}) =& \alpha\mathbb{E}_{d_{\pi_E}}\left[\left(\tilde{Q}(s,a) - (r_{\max} + \gamma\mathbb{E}_{s'\sim P(.|s,a)}[\tilde{V}^\pi(s')]\right)^2\right] \quad (10)\\ &+ (1-\alpha)\mathbb{E}_{d_{\pi_{\tilde{Q}}}}\left[\left(\tilde{Q}(s,a) - (r_{\min} + \gamma\mathbb{E}_{s'\sim P(.|s,a)}[\tilde{V}^\pi(s')])\right)^2\right] + \frac{\beta}{c}H(\pi_{\tilde{Q}})\\ =& \alpha\,\delta^2(d_{\pi_E}, r_{\max}) + (1-\alpha)\,\delta^2(d_\pi, r_{\min}) + \frac{\beta}{c}H(\pi_{\tilde{Q}})\,, \quad (11)\end{aligned}$$

其中 δ^2 是平方软贝尔曼误差。我们可以从公式 11中推导出以下结论：

χ^2-混合分布下的正则化逆强化学习可以视为一个强化学习问题，其中专家和策略分别具有固定奖励 $r_{\max}$和 $r_{\min}$。这一洞见有助于我们理解正则化常数 c的重要性：它定义了目标奖励，因此也决定了 Q函数的尺度。由此得到的目标函数与 SQIL 算法密切相关，因为 SQIL 同样使用固定奖励。然而，SQIL 使用 $r_{\max} = 1$和 $r_{\min} = 0$，这在我们的 $\alpha < 1$设定中不可行。熵项看起来是另一个差异，但我们注意到，它不会影响评论家更新，因为此时 $\pi_{\tilde{Q}}$ 是固定的。与 SQIL 一样，通过使用方程 2 提取最大熵策略，可以最大化熵。

在固定奖励设置下稳定训练非常简单。我们可以针对奖励偏差问题给出简洁的解决方案——参见第 3.4 节——并为专家提供固定的 Q-目标值，为策略提供裁剪后的 Q-函数目标值——参见第 3.5 节和第 3.7 节，以显著提高学习稳定性。然而，为了应用这些技术，我们必须通过引入熵评论家，将软动作价值函数切换为硬动作价值函数。此外，我们还展示了如何通过引入正则化评论家，恢复与公式 5 中正则化项相对应的正确策略更新。

3.3 熵与正则化评论家

~ ~我们使用 $Q(s,\ a)\ = Q(s,\ a)\ + \mathcal{H}^{\pi}(s,\ a)$对 Q-函数进行隐式表示，将其分解为硬 Q-函数和熵评论家

$$\mathcal{H}^{\pi}(s,a) = \mathbb{E}_{P,\pi}\left[\sum_{t'=t}^{\infty} -\gamma^{t'-t+1}\beta \log \pi(a_{t'+1}|s_{t'+1}) \middle| s_t = s, a_t = a\right]. \tag{12}$$

这一过程使我们能够保持最大熵表述，同时保留对硬 Q函数进行操作的能力。我们将软逆贝尔曼算子替换为硬最优逆贝尔曼算子 $(\mathcal{T}Q)(s,a)\ = Q(s,a) - \gamma\mathbb{E}_{s'\sim P(.|s,a)}V^{*}(s')$，其中最优价值函数为$V^{*}(s)\ = \max_a Q(s,a)$。

如前所述，公式 5 中引入的正则化项还包含另一个依赖于策略的项。实际上，公式 1 中的内层优化问题——即方括号中的项——不再是纯粹的最大熵问题，而是包含了项 $-k\mathbb{E}_{d_\pi}[r(s,a)^2]$，其中 $k = c(1-\alpha)$。为了将这一项纳入最终的隐式动作价值函数 $Q^{\dagger}(s,a)$，我们再学习一个正则化评论家

$$\mathcal{C}(s,a) = \mathbb{E}_{P,\pi}\left[\sum_{t'=t}^{\infty} \gamma^{t'-t} r(s_{t'},a_{t'})^2 \middle| s_t = s, a_t = a\right]. \tag{13}$$

使得 $Q^{\dagger}(s,a)\ = Q(s,a)\ + \mathcal{H}^{\pi}(s,a)\ + k\,\mathcal{C}(s,a)$。利用 $Q^{\dagger}$，我们可以得到公式 2 中内层最小化问题的精确解。在实践中，我们学习一个将 $\mathcal{H}^{\pi}$ 和 $\mathcal{C}$ 结合起来的评论家 $\mathcal{G}^{\pi}$。我们使用以下目标函数独立训练后者

$$\min_{\mathcal{G}^{\pi}} \delta_{\mathcal{G}}^2 = \min_{\mathcal{G}^{\pi}}\ \mathbb{E}_{d_\pi}\left[(\mathcal{G}^{\pi}(s,a) - (k\, r_Q(s,a)^2 + \mathbb{E}_{\substack{s'\sim P \\ a'\sim\pi}}\left[\gamma(-\beta\log\pi(a'|s') + \mathcal{G}^{\pi}(s',a'))\right]))^2\right], \tag{14}$$

这是一个熵正则化贝尔曼误差最小化问题，其中隐式奖励的平方由 r_Q 按比例缩放，缩放系数为 k。

3.4 吸收状态的处理

IQ-Learn 忽略的另一个技术问题是如何正确处理吸收状态。Garg 等人（2021）通过在逆贝尔曼算子的折扣价值函数前添加一个指示项来处理吸收状态：当 s' 是终止状态时，该指示项为 ν；否则为 $\nu = 1$。

$$(\mathcal{T}_{\text{iq}}^{\pi}Q)(s,a) = Q(s,a) - (1-\nu)\gamma\mathbb{E}_{s'\sim P(.|s,a)}V^{\pi}(s'). \tag{15}$$

该逆贝尔曼算子是通过在假设吸收状态的价值为零的情况下，对 $r(s,\ a)$ 求解正向贝尔曼算子得到的。然而，正如 Kostrikov 等人（2019）所指出的，这一假设可能引入终止偏差或生存偏差；吸收状态的价值也需要进行学习。我们的观点清晰地阐明了公式 15 中逆贝尔曼算子的作用：公式 10 中的目标函数会将转移到吸收状态的 Q-函数回归至 $r_{\max}$ 或 $r_{\min}$。然而，根据公式 9，吸收状态的隐式奖励应回归至 $r_{\max}$ 或 $r_{\min}$。

相反，我们从标准贝尔曼算子推导出逆算子，同时利用吸收状态的价值 s_A 与策略 π无关这一点

$$(\mathcal{T}^{\pi}_{\text{lsiq}}Q)(s,a)=Q(s,a)-\gamma\mathbb{E}_{s'\sim P(.|s,a)}\big((1-\nu)V^{\pi}(s')+\nu V(s_A)\big). \quad (16)$$

我们进一步利用吸收状态的价值可以用闭式形式计算这一点，即 $V(s_A)=\frac{r_A}{1-\gamma}$，其中 r_A 在专家状态上等于 $r_{\max}$，在策略状态上等于 $r_{\min}$。请注意，尽管使用吸收状态的解析价值而非自举，对应的正向贝尔曼算子仍会收敛到相同的 Q-函数，正如我们在附录 A.5 中所示。将公式 16 中的逆算子应用于公式 9 时，我们会正确地将转移至吸收状态的 Q-函数回归到其折扣回报。附录 A.4 展示了最终的完整目标函数。

我们展示了修改后的算子在图 1（上）所示玩具任务中的效果：黑色点质量会出现在四个深蓝色方块之一，并需要到达中间的绿色区域。一旦智能体进入红色区域，回合便会终止。专家总是选择到达绿色区域的最短路径，从不经过红色区域。IQ-Learn 提出的算子未能充分惩罚智能体到达吸收状态，导致 IQ-Learn 智能体无法稳定地到达目标，如图 1（下）的橙色曲线所示。相反，使用我们的算子 $\mathcal{T}_{\text{lsiq}}$时，智能体能够成功完成任务。

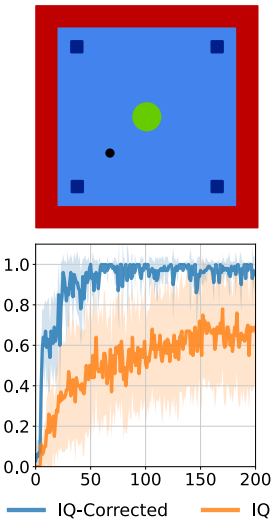

图 1：点质量玩具任务（上）及成功率图（下）。这里，我们比较标准的 IQ-Learn 算子与修改后的算子。

3.5 专家残差最小化的另一种表述

公式 9 中的第一项定义了分布上的平方贝尔曼误差最小化问题，分布 d_{π_E}

$$\alpha\,\delta^2(d_{\pi_E}, r_{\max}) = \alpha\mathbb{E}_{d_{\pi_E}}\left[(r_Q(s,a) - r_{\max})^2\right]. \quad (17)$$

由于采用自举，即使专家策略固定，这种最小化也可能颇具挑战，因为除非轨迹到达吸收状态，否则它无法固定 Q-函数的尺度。对于循环任务的专家数据，这个问题尤为突出，因为我们生成的轨迹具有固定时域。尺度缺乏固定值会增大算法的方差，进而降低其性能。

因此，我们提出一个修改后的目标函数，并分析公式 17。当所有在 d_{π_E} 下可达的 (s,a) 均满足 $r_Q(s,a) = r_{\max}$ 时，该项达到最小值。因此，该项的目标是在专家轨迹上推动奖励趋向 $r_{\max}$。若考虑这一最小值，专家轨迹中的每个转移都有如下 Q-值：

$$Q^{\pi_E}(s,a) = \sum_{t=0}^{\infty}\gamma^t r_{\max} = \frac{r_{\max}}{1-\gamma} = Q_{\max}, \quad \text{with } s,a \sim d_{\pi_E}(s,a). \quad (18)$$

由于我们在专家分布上的最大化目标函数等价于将专家的状态和动作的价值推向 $Q_{\max}$，因此我们提出用固定目标 $Q_{\max}$ 替换自举目标，从而得到以下新的目标函数：

$$\mathcal{L}_{\text{lsiq}}(Q)=\alpha\mathbb{E}_{d_{\pi_E}}\left[(Q(s,a)-Q_{\max})^2\right]+(1-\alpha)\mathbb{E}_{d_{\pi}}\left[\big(Q(s,a)-(r_{\min}+\gamma\mathbb{E}_{s'\sim P(.|s,a)}[V^*(s')])\big)^2\right]. \quad (19)$$

请注意，为清晰起见，我们略去了终止状态处理。完整的目标函数见附录 A.4。此外，我们省略了熵项，因为目前已将其纳入 $\mathcal{H}^{\pi}(s,a)$ 中。这个新的目标函数融入了对专家数据的偏向。因此，它并不严格等价于原始的问题表述。不过，它会将 Q-函数更新至同一个理想目标，同时提供更简单、更稳定的优化景观。从经验上看，我们发现这一修改虽然仅凭直觉得到论证，却对算法的性能产生了非常积极的影响。

3.6 从观测中学习

在许多现实任务中，我们无法获得专家动作，只能获得对专家行为的观测（Torabi 等，2019b）。在这种情况下，通过学习一个仅依赖于当前状态和下一状态的判别器，可以轻松调整对抗模仿学习方法，例如 GAIfO（Torabi 等，2019a）。遗憾的是，将同样的方法应用于学习 Q-函数的隐式奖励算法并不直接。IQ-Learn 方法（Garg 等人，2021年）基于对原始目标函数的简化来执行更新：它不使用专家动作，而是在专家状态下从策略中采样动作。然而，正如我们的实验结果所示，这种重新表述无法在标准基准测试中取得良好性能。

文献中一种常见做法是训练一个 IDM。这种方法此前已用于行为克隆（Torabi 等，2018年；Nair 等，2017年）以及基于示范的强化学习（Guo 等，2019年；Pavse 等，2020年；Radosavovic 等，2021年）。沿用这一思路，我们通过在策略数据上在线训练 IDM，并用它预测专家未被观测到的动作，生成了我们方法的仅观测版本。我们将公式 19 中的目标函数修改为

$$\mathcal{L}_{\text{lsiq-o}}(Q)=\alpha\mathbb{E}_{d_{\pi_E}}\left[\left(Q(s,\Gamma_\omega(s,s'))-Q_{\max}\right)^2\right]+\bar{\alpha}\mathbb{E}_{d_\pi}\left[\left(Q(s,a)-(r_{\min}+\gamma\mathbb{E}_{s'\sim P(.|s,a)}[V^*(s')])\right)^2\right], \quad (20)$$

其中动力学模型为 $\Gamma_\omega(s,\ s')$，其参数为 ω 和 $\bar{\alpha}=(1-\alpha)$。我们省略吸收状态的符号表示，转而参见附录 A.4。请注意，IDM 仅用于评估专家动作，其训练通过求解以下优化问题完成

$$\min_\omega \mathcal{L}_\Gamma(\omega)=\min_\omega \mathbb{E}_{d_\pi,\mathcal{P}}\left[\|\Gamma_\omega(s,s')-a\|_2^2\right], \quad (21)$$

其中，期望是在学习者策略 π生成的状态分布上计算的。尽管训练分布与评估分布之间的不匹配可能会引发问题，但我们的实证评估表明，在基准测试中，我们取得了与动作感知算法相近的性能。附录 B 中将详细介绍这一方法。

3.7 实用算法

本节将具体说明如何实现我们算法的实用版本。我们的方法概览见算法1。在实践中，我们使用参数化函数来近似 Q、 π、 $\mathcal{G}$ 和 Γ，并通过对代理目标函数进行梯度上升来优化后者；这些代理目标函数利用数据集 $\mathcal{D}_\pi$ 和$\mathcal{D}_{\pi_E}$，近似在 d_π 和d_{π_E} 下的期望。此外，我们采用了 Garg 等人（ 2021 ）已提出的目标网络。然而，虽然公式 4 中的目标函数没有说明使用目标网络的直观理由，但公式 11 中的目标函数等价于一个强化学习目标函数，而目标网络是其中一种广为人知的稳定化工具。此外，我们利用了对硬Q函数的访问权限，以及固定奖励目标设定，来计算可能的 Q 值最大值和最小值， $Q_{\min}=\frac{r_{\min}}{1-\gamma}$ 和 $Q_{\max}=\frac{r_{\max}}{1-\gamma}$，并将目标网络的输出裁剪到该范围内。请注意，这对吸收状态同样成立。这样做可以确保目标 Q 始终处于所需范围内，而 IQ-Learn 中往往无法做到这一点。目标裁剪可以防止因使用神经近似器而导致 Q 值爆炸。这项技术使算法能够从较差的价值函数估计中恢复，并防止 Q 函数偏离可容许函数集合。最后，我们发现，即使策略从不访问这些状态和动作，在小型固定专家数据集上训练策略也会使专家轨迹的熵奖励逐渐衰减。为解决这一问题，我们将专家状态上的熵奖励裁剪为策略状态上最大熵的滑动平均值。

算法1 LS-IQ

初始化： Q_θ、 π_ϕ、 $\mathcal{G}_\zeta$ 以及可选的 Γ_ω

对于步骤 t，在 $\{1,\dots,N\}$ **执行**

采样小批次 $\mathcal{D}_\pi$ 和 $\mathcal{D}_{\pi_E}$

（可选）使用 Γ_ω 为 $\mathcal{D}_{\pi_E}$ 预测动作

$\mathcal{D}_{\pi_E}\leftarrow\{\{s,\Gamma_\omega(s,\ s')\ s'\}|\forall\{s,s'\}\in\mathcal{D}_{\pi_E}\}$

使用公式 20 更新 Q-函数

$\theta_{t+1}\leftarrow\theta_t+\kappa_Q\nabla_\theta[\mathcal{J}(\theta,\ \mathcal{D}_\pi,\ \mathcal{D}_{\pi_E})]$

（可选）使用公式 14 更新 $\mathcal{G}$-函数

$\zeta_{t+1}\leftarrow\zeta_t-\kappa_\mathcal{G}\nabla_\zeta[\delta^2_\mathcal{G}(\zeta,\ \mathcal{D}_\pi)]$

使用 KL散度更新策略 π_ϕ

$\phi_{t+1}\leftarrow\phi_t-\kappa_\pi\nabla_\phi[D_{KL}(\pi\phi\|\pi_{\tilde{Q}})]$

（可选）使用公式 21 更新 Γ_ω

$\omega_{t+1}\leftarrow\omega_t-\kappa_\Gamma\nabla_\omega[\mathcal{L}_\Gamma(\omega,\mathcal{D}_\pi)]$

结束循环

在连续动作空间中， Z_s 是难以计算的，因此我们无法直接使用公式 2 提取最优策略。与以往工作（Haarnoja等人，2018；Garg 等人，2021）一样，我们使用参数化策略 π_ϕ，通过最小化 KL散度 $D_{\text{KL}}(\pi_\phi\|\pi_{\tilde{Q}})$ 来近似 $^\pi\tilde{Q}$。在我们的实现中，我们发现没有必要使用双评论家更新。这一选择降低了算法的

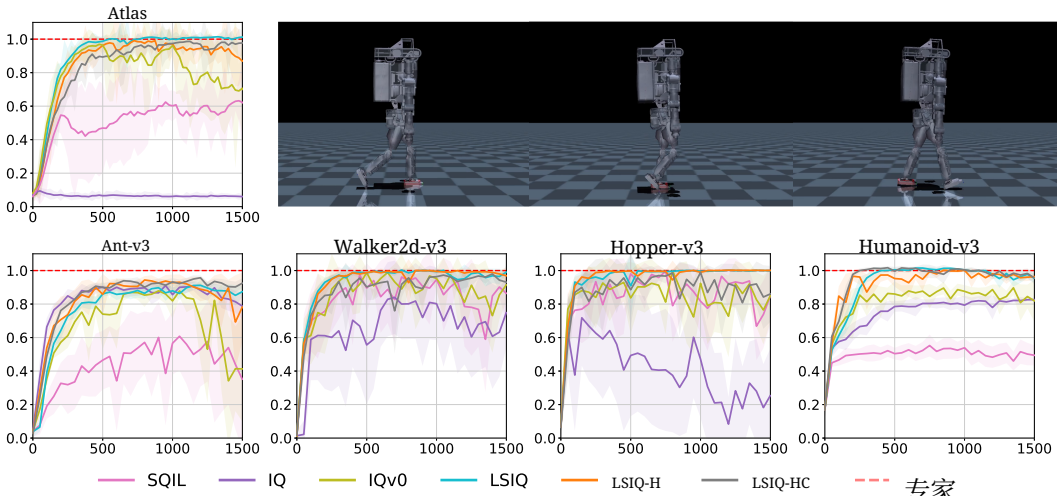

图 2：LS-IQ 不同版本的比较。横轴表示归一化折扣累积奖励，纵轴表示训练步数（$\times 10^3$）。第一行展示了结果以及一个示例轨迹——此处为训练后的 LS-IQ 智能体——该任务是在 Atlas 机器人上进行的运动任务。第二行展示了 4 个 MuJoCo Gym 任务，其中专家的累积奖励为 $\rightarrow$ Hopper:3299.81, Walker2d:5841.73, Ant:6399.04, Humanoid:6233.45

计算和内存需求，使其与 SAC 相当。最后，我们在策略期望中用 $V^{\pi}(s)$ 替换 $V^{*}(s)$，因为在连续动作空间中我们无法获得后者。

4 实验

我们在六个 MuJoCo 环境中评估所提方法：Ant-v3、Walker2d-v3、Hopper-v3、HalfCheetah-v3、Humanoid-v3 和 Atlas。后者是我们提出的一种新型运动环境，详见附录 C.1。我们选取以下基线方法：GAIL（Ho & Ermon，2016年）、VAIL（Peng 等人，2019年）、IQ-Learn（Garg 等人，2021年）和 SQIL（Reddy 等人，2020年）。为确保比较公平，所有方法均在相同的 Mush- roomRL 框架（D'Eramo 等人，2021年）中实现。我们确认，我们的实现取得了与作者原始实现相当的结果。对于各个环境，我们采用原始作者提出的超参数，并在新环境中进行网格搜索。IQ-Learn 的原始实现会根据给定环境评估两种不同的算法变体。我们将这两种变体称为 IQv0（使用望远镜求和（Garg 等人，2021年）评估公式 4 中智能体的期望回报）和 IQ（直接使用公式 4），并在所有环境中评估这两种变体。对于我们的方法，除正则化系数 c 和熵系数 β 外，我们采用与 IQ-Learn 相同的超参数，并针对每个环境调整这两个系数。我们仅考虑等权混合，即 $\alpha = 0.5$。

在我们的第一个实验中，我们对 LSIQ 的不同设计选择进行了消融实验。我们评估了以下变体：LSIQ-HC 使用（组合的）熵评论家和正则化评论家，LSIQ-H 仅使用熵评论家，而 LSIQ 不使用任何额外的评论家，与 IQ-Learn 类似。这些实验使用了十个随机种子和五条专家轨迹。对于 Atlas 环境，我们使用了 100 条轨迹。我们还将 IQ、IQv0 和 SQIL 作为基线，并在图 2 中展示了四种环境的学习曲线。HalfCheetah 环境的学习曲线见附录 C.6。值得注意的是，不采用望远镜求和的 IQ-Learn 在 Atlas、Walker 和 Hopper 上表现不佳；与 Ant 和 HalfCheetah 相比，这些环境更容易出现吸收状态，而 Ant 和 HalfCheetah 几乎总是在固定步数后终止。我们推测，Walker 和 Hopper 上较差的性能是由奖励偏差导致的，因为吸收状态受到的惩罚不足。IQv0 受这一问题的影响较小，因为它将智能体访问过的所有状态都视为初始状态，因此会对这些状态施加更强的奖励惩罚。我们进一步开展了消融研究，以展示所提出技术的影响，包括研究固定目标值的作用、对目标 Q-值进行裁剪、对专家进行熵裁剪，以及

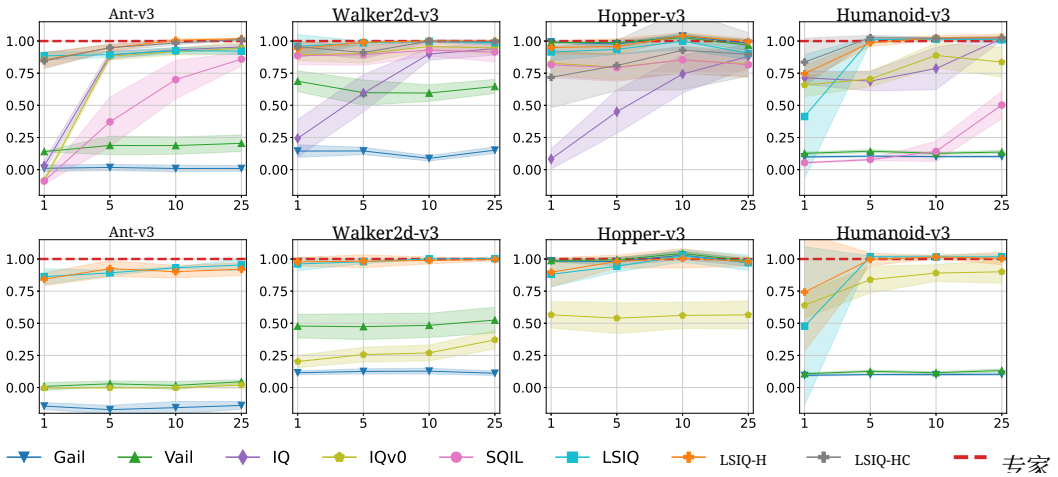

图 3：不同 Mujoco 环境中，专家轨迹数量影响的消融研究。横坐标表示归一化累积奖励，纵坐标表示专家轨迹的数量。第一行展示考虑状态和动作时的性能，第二行展示仅使用状态时的性能。专家累积奖励与图 2 相同。

附录 C 中对吸收状态的处理。我们的结果表明，额外的评论家影响甚微，而固定目标值能显著提升性能。

在主要实验中，我们仅评估 LSIQ 和 LSIQ-H，因为它们在大多数环境中表现最佳。我们针对四种不同数量的专家示范（1、5、10 和 25）将我们的方法与所有基线进行比较，并始终使用五个随机种子。每项实验均在有专家动作和无专家动作的情况下进行。当无法获得动作时，我们在 GAIL 和 VAIL 中使用状态转移判别器（Torabi 等，2019a），在 LSIQ 中使用逆动力学模型（参见第 3.6 节）。相比之下，IQ-Learn 在无法获得专家动作时，会使用当前策略在专家状态上预测的动作。在从观察中学习设定下，我们不评估 SQIL，并省略 IQ 的图表，因为它在任何环境中都无法收敛，而是仅关注 IQv0。图 3 展示了四种环境中，不同示范数量对应的最终期望回报。所有学习曲线（包括 HalfCheetah 环境的曲线）均见附录 C.6（状态-动作设定）和附录 C.5（从观察中学习设定）。实验结果表明，LSIQ 的性能与所有基线相当或更优。尤其是在从观察中学习设定下，LSIQ 表现出色，其回报与观测到状态和动作时的设定相近。

5 结论

受 IQ-Learn 实际实现的启发，我们利用隐式奖励函数以及对专家与策略混合分布施加的平方 L_2 惩罚，推导出一种分布匹配算法。我们证明，该正则化项能够最小化与混合分布之间有界的 χ^2-散度，并由此对 Q函数和策略的更新进行修正。我们的分析揭示了该方法与 SQIL 之间一个有趣的联系——SQIL 并非从对抗式分布匹配目标推导而来；分析还表明，IQ-Learn 存在奖励偏差。基于这些发现，我们提出了一种新方法 LS-IQ，它采用修正后的逆贝尔曼算子来处理奖励偏差、目标裁剪、策略样本的固定奖励目标值，以及专家样本的固定 Q函数目标值。我们还证明，IQ-Learn 的策略优化与对混合分布进行正则化并不一致，并说明了如何通过学习一个额外的正则化评论家来解决这一问题。在实验中，LS-IQ 的表现优于强基线方法，尤其是在从观测中学习时；在这种情况下，我们训练了一个 IDM 来预测专家动作。未来的工作中，我们将量化固定 Q函数目标值所引入的偏差，并探究这一启发式方法为何是稳定学习的关键。我们还将分析 Q函数目标值中的误差传播，并推导关于 Q函数近似误差的理论保证。

致谢

本研究的计算工作在达姆施塔特工业大学的 Lichtenberg 高性能计算机上完成。本研究得到德国科学基金会（DFG）项目资助，项目编号为 SE1042/41-1。本文所述研究部分得到了德国联邦教育与研究部（BMBF）的资助，资助内容属于 KIARA 合作项目的子项目“运行区域建模与探索、AI 辅助系统设计以及技术使用的法律层面”（项目编号：13N16274）。

参考文献

Alexander Amini、Wilko Schwarting、Ava Soleimany 和 Daniela Rus。深度证据回归。载于第三十四届神经信息处理系统大会论文集，线上，2020年12月。

Oleg Arenz 和 Gerhard Neumann。非对抗模仿学习及其与对抗方法的联系。*arXiv*，2020年8月。

Carlo D’Eramo、Davide Tateo、Andrea Bonarini、Marcello Restelli 和 Jan Peters。MushroomRL：简化强化学习研究。《机器学习研究期刊》，22：1–5，2021年。

Justin Fu、Katie Luo 和 Sergey Levine。通过对抗逆强化学习学习鲁棒奖励。载于国际学习表征会议论文集，加拿大温哥华，2018年4月。

Divyansh Garg、Shuvam Chakraborty、Chris Cundy、Jiaming Song 和 Stefano Ermon。IQ-Learn：用于模仿的逆软 Q 学习。载于第三十五届神经信息处理系统大会论文集，澳大利亚悉尼，2021年12月。

Seyed Kamyar Seyed Ghasemipour、Richard Zemel 和 Shixiang Gu。模仿学习方法的散度最小化视角。载于机器人学习会议论文集，日本大阪，2019年11月。

Ian J.Goodfellow、JeanPouget-Abadie、Mehdi Mirza、BingXu、DavidWarde-Farley、SherjilOzair、Aaron Courville 和 Yoshua Bengio。生成对抗网络。载于第二十八届神经信息处理系统大会论文集，加拿大蒙特利尔，2014年12月。

Xiaoxiao Guo、Shiyu Chang、Mo Yu、Gerald Tesauro 和 Murray Campbell。结合专家状态序列的混合强化学习。载于*AAAI* 人工智能大会论文集，第 33 卷，第 3739–3746 页，2019年。

Tuomas Haarnoja、Aurick Zhou、Pieter Abbeel 和 Sergey Levine。软演员-评论家：基于随机演员的离策略最大熵深度强化学习。载于国际机器学习会议论文集，斯德哥尔摩，瑞典，2018年7月。

Jonathan Ho 和 Stefano Ermon。生成对抗模仿学习。载于 NIPS 2016。

ICLR 2019。 国际学习表征会议，新奥尔良，美国，2019年5月。

ICLR 2020。 国际学习表征会议，线上，2020年5月。

Ilya Kostrikov、Kumar Krishna Agrawal、Debidatta Dwibedi、Sergey Levine 和 Jonathan Tomp- son。判别器-行动者-评论家：解决对抗模仿学习中的样本效率低下和奖励偏差问题。载于 ICLR 2019。

Ilya Kostrikov、Ofir Nachum 和 Jonathan Tompson。通过离策略分布匹配进行模仿学习。发表于 ICLR 2020。

Xudong Mao、Qing Li Haoran Xie、Raymond Y.K. Lau、Zhen Wang 和 Stephen Paul Smolley。最小二乘生成对抗网络。发表于*IEEE* 国际计算机视觉会议（*ICCV*）论文集，意大利威尼斯，2017年10月。

Ofir Nachum、Yinlam Chow、Bo Dai 和 Lihong Li。Dualdice：与行为无关的折扣平稳分布校正估计。在第三十三届神经信息处理系统会议论文集中，加拿大温哥华，2019年12月。

Ashvin Nair、Dian Chen、Pulkit Agrawal、Phillip Isola、Pieter Abbeel、Jitendra Malik 和 Sergey Levine。结合自监督学习与模仿，开展基于视觉的绳索操作。在*2017*年*IEEE*机器人与自动化国际会议（*ICRA*）中，第 2146–2153 页。IEEE，2017年。

Andrew Ng、Daishi Harada 和 Stuart Russell。奖励变换下的策略不变性：奖励塑形的理论与应用。在国际机器学习会议论文集中，斯洛文尼亚布莱德，1999年6月。

NIPS 2016。 第三十届神经信息处理系统会议论文集，西班牙巴塞罗那，2016年12月。

Sebastian Nowozin、Botonf Cseke 和 Ryota Tomioka。f-GAN：使用变分散度最小化训练生成式神经采样器。发表于 NIPS 2016。

Brahma S Pavse、Faraz Torabi、Josiah Hanna、Garrett Warnell 和 Peter Stone。RIDM：用于从单个观测示范中学习的强化逆动力学建模。*IEEE*机器人与自动化快报，5(4)：6262–6269，2020年。

Xue Bin Peng、Angjoo Kanazawa、Sam Toyer、Pieter Abbeel 和 Sergey Levine。变分判别器瓶颈：通过约束信息流改进模仿学习、逆强化学习和生成对抗网络。发表于 ICLR 2019。

Xue Bin Peng、Ze Ma、Pieter Abbeel、Sergey Levine 和 Angjoo Kanazawa。AMP：用于风格化基于物理的角色控制的对抗运动先验。*ACM*图形学汇刊，40：1–20，2021年。

Ilija Radosavovic、Xiaolong Wang、Lerrel Pinto 和 Jitendra Malik。用于灵巧操纵的仅状态模仿学习。载于*2021*年 *IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第 7865–7871 页。IEEE，2021年。

Siddharth Reddy、Anca D. Dragan 和 Sergey Levine。SQIL：通过强化学习与稀疏奖励进行模仿学习。载于 ICLR 2020。

Harshit Sikchi、Akanksha Saran、Wonjoon Goo 和 Scott Niekum。用于模仿学习的排序博弈。《机器学习研究汇刊》，2023年。

Gokul Swamy、Sanjiban Choudhury、J. Andrew Bagnell 和 Zhiwei Steven Wu。关于矩与匹配：一种弥合模仿差距的博弈论框架。载于国际机器学习会议论文集，线上，2021年7月。

Faraz Torabi、Garrett Warnell 和 Peter Stone。基于观测的行为克隆。载于第二十七届国际人工智能联合会议（*IJCAI*）论文集，第 4950–4957 页，瑞典斯德哥尔摩，2018年7月。

Faraz Torabi、Garrett Warnell 和 Peter Stone。基于观测的生成对抗模仿。载于国际机器学习会议论文集，加利福尼亚州长滩，2019年7月。

Faraz Torabi、Garrett Warnell 和 Peter Stone。基于观测的模仿学习最新进展。载于第二十八届国际人工智能联合会议（*IJCAI*）论文集，中国澳门，2019年8月。

约翰·冯·诺伊曼。《论博弈理论》。《数学年刊》，100(1):295–320，1928年。由 Sonya Bargman 翻译成英文（Von Neumann，1959年）。

约翰·冯·诺伊曼。论策略博弈理论。《博弈论研究》，4: 13–42，1959年。

Brian D. Ziebart. 运用最大因果熵原理对有目的的自适应行为进行建模。博士论文，华盛顿大学，2010年。Brian D. Ziebart、Andrew Maas、J.Andrew Bagnell 和 Anind K. Dey。最大熵逆强化学习。发表于第二十三届人工智能 *AAAI* 会议论文集，芝加哥，美国，2008年7月。

A 证明与推导

本节给出正文中各项命题的证明。此外，我们还补充两个命题，分别讨论最优奖励函数以及考虑混合分布时 χ^2-散度的界。

A.1 从占用度量到分布

根据命题 A.1，在占用度量下的解 $\arg\max_{Q\in\tilde{\Omega}} \mathcal{J}(Q,\pi_{\tilde{Q}})$ 等于在状态-动作分布下的解 $\arg\max_{\tilde{Q}\in\tilde{\Omega}} \mathcal{J}(Q,\pi_{\tilde{Q}})$。这一结果使我们从现在起可以使用以下分布匹配问题：

$$\max_{\tilde{Q}\in\tilde{\Omega}} \mathcal{J}(\tilde{Q},\pi_{\tilde{Q}}) = \max_{\tilde{Q}\in\tilde{\Omega}} \mathbb{E}_{d_{\pi_E}}\left[r_{\tilde{Q}}(s,a)\right] - \mathbb{E}_{d_{\pi_{\tilde{Q}}}}\left[r_{\tilde{Q}}(s,a)\right] - \psi(r_{\tilde{Q}}) - \beta H(\pi_{\tilde{Q}})\,, \tag{22}$$

其中，我们引入隐式奖励 $r_{\tilde{Q}}(s,a) = Q(s,a) - \gamma\mathbb{E}_{s'\sim P(.|s,a)}[V^{\pi}(s')]$ 为便于理解，β 是控制熵正则化的常数，d_{π_E} 是专家的状态-动作分布，而 d_π 是策略下的状态-动作分布。

命题 A.1 令 $\max_{r\in\mathcal{R}} \min_{\pi\in\Pi} L_\rho(\pi,r)$ 为正则化占用匹配优化问题的对偶问题，令 $L(\pi,r)$ 为正则化分布匹配问题的拉格朗日函数。则有 $L_\rho(\pi,r) \propto L(\pi,r)$ 以及 $\mathcal{J}_\rho(Q,\pi_{\tilde{Q}}) \propto \mathcal{J}(Q,\pi_{\tilde{Q}})$。

命题证明*A.1*。

从任意策略 π 的占用度量定义出发，我们通过积分计算归一化常数：

$$\begin{aligned}
\int_{\mathcal{S}\times\mathcal{A}} \rho_\pi(s,a)dsda &= \int_{\mathcal{S}\times\mathcal{A}} \lim_{T\to\infty} \sum_{t=0}^{T-1} \gamma^t \mu_t^\pi(s)\pi(a|s)\,dsda \\
&= \lim_{T\to\infty} \sum_{t=0}^{T-1} \gamma^t \int_{\mathcal{S}\times\mathcal{A}} \mu_t^\pi(s)\pi(a|s)\,dsda \\
&= \lim_{T\to\infty} \sum_{t=0}^{T-1} \gamma^t \cdot 1 \\
&= \frac{1}{1-\gamma}
\end{aligned} \tag{23}$$

现在我们计算（折扣）状态-动作分布：

$$d_\pi(s,a) = \frac{\rho_\pi(s,a)}{\int_{\mathcal{S}\times\mathcal{A}} \rho(s,a)dsda} = \frac{\rho_\pi(s,a)}{\frac{1}{1-\gamma}} = (1-\gamma)\rho_\pi(s,a) \tag{24}$$

因此，我们有：

$$\rho_\pi(s,a) = \frac{1}{1-\gamma} d_\pi(s,a) \tag{25}$$

将公式 25 代入目标函数的定义，我们得到：

$$\mathcal{J}(\pi) = \frac{1}{1-\gamma}\int_{\mathcal{S}\times\mathcal{A}} d_\pi(s,a)r(s,a)dsda = \frac{1}{1-\gamma}\mathbb{E}_{d_\pi}\left[r(s,a)\right] \tag{26}$$

利用公式 25，可以采用与公式 26 类似的推导方式计算熵和正则化项。将推导出的公式代入公式 1，并合并常数 $\frac{1}{1-\gamma}$，即可证明该命题。■

A.2 混合分布上 χ^2 散度的界

命题 A.2 根据公式*6*所示的变分形式，分布 d_{π_E} 与混合分布 $0.5\,d_{\pi_E} + 0.5\,d_\pi$ 之间的皮尔逊 χ^2-散度对应的最优奖励为

$$r^*(s,a) = \frac{1}{c}\frac{d_{\pi_E}(s,a) - d_\pi(s,a)}{d_{\pi_E}(s,a) + d_\pi(s,a)} \tag{27}$$

并且有界，使得 $r^*(s,\ a)\ \in [-1/c, 1/c]$ 对所有 $s\in\mathcal{S}$、$a\in\mathcal{A}$ 均成立。

命题*A.2*的证明。 此证明沿用了 Goodfellow 等人（2014）中命题1的证明。我们回顾一下， χ^2-散度在公式6中的混合分布上为

$$\begin{aligned}2\chi^2\big(d_{\pi_E}\,\|\,\underbrace{\tfrac{d_{\pi_E}+d_\pi}{2}}_{d_{\text{mix}}}\big)&=\max_{r\in\mathcal{R}}2\Big(\mathbb{E}_{d_{\pi_E}}\left[r(s,a)\right]-\mathbb{E}_{d_{\text{mix}}}\left[r(s,a)+kr(s,a)^2\right]\Big)\\&=\max_{r\in\mathcal{R}}\mathbb{E}_{d_{\pi_E}}\left[r(s,a)\right]-\mathbb{E}_{d_\pi}\left[r(s,a)\right]-k\mathbb{E}_{d_{\pi_E}}\left[r(s,a)^2\right]-k\mathbb{E}_{d_\pi}\left[r(s,a)^2\right]\\&=\max_{r\in\mathcal{R}}\int_{\mathcal{S}}\int_{\mathcal{A}}d_{\pi_E}(s,a)\big(r(s,a)-kr(s,a)^2\big)-d_\pi(s,a)\big(r(s,a)+kr(s,a)^2\big)\,da\,ds,\end{aligned}\tag{28}$$

其中 $k=1/4$适用于传统的 χ^2-散度。我们通过设定 $k=c/2$ 来推广 χ^2-散度。对于任意 a, $b\in\mathbb{R}^+\setminus\{0\}$，函数 $y\to a(y-\frac{c}{2}y^2)-b(y+\frac{c}{2}y^2)$ 在 $\frac{1}{c}\frac{a-b}{a+b}$处取得最大值，而该值属于区间 $[-1/c,1/c]$。

为完成证明，我们注意到，可以在 $Supp(d_\pi)\cup Supp(d_{\pi_E})$ 之外任意定义奖励函数，因为这不会影响散度。■

命题 A.3 分布 d_{π_E} 与混合分布 $0.5\,d_{\pi_E}+0.5\,d_\pi$ 之间的皮尔逊 χ^2-散度有如下界

$$0\le 2\chi^2\big(d_{\pi_E}\,\|\,\underbrace{\tfrac{d_{\pi_E}+d_\pi}{2}}_{d_{mix}}\big)\le\tfrac{1}{c}\,.\tag{29}$$

命题*A.3*的证明。为提高可读性，我们省略符号中对状态和动作的显式依赖，并将 d_{π_E}和 d_π 分别记作 $d_{\pi_E}(s,a)$和 $d_\pi(s,a)$。对于任意散度，下界显然成立，并在 $d_\pi=d_{\text{mix}}=d_{\pi_E}$时取到。为证明上界，我们使用方程 27 中的最优奖励函数，并将其代入方程 28，其中 $k=c/2$

$$\begin{aligned}2\chi^2\big(d_{\pi_E}\|\underbrace{\tfrac{d_{\pi_E}+d_\pi}{2}}_{d_{\text{mix}}}\big)&=\int_{\mathcal{S}}\int_{\mathcal{A}}d_{\pi_E}\left(r^*(s,a)-\tfrac{c}{2}r^*(s,a)^2\right)-d_\pi\left(r^*(s,a)+\tfrac{c}{2}r^*(s,a)^2\right)da\,ds\\&=\int_{\mathcal{S}}\int_{\mathcal{A}}d_{\pi_E}\left(\tfrac{1}{c}\left(\frac{d_{\pi_E}-d_\pi}{d_{\pi_E}+d_\pi}\right)-\tfrac{c}{2}\tfrac{1}{c^2}\left(\frac{d_{\pi_E}-d_\pi}{d_{\pi_E}+d_\pi}\right)^2\right)\\&\quad-d_\pi\left(\tfrac{1}{c}\left(\frac{d_{\pi_E}-d_\pi}{d_{\pi_E}+d_\pi}\right)+\tfrac{c}{2}\tfrac{1}{c^2}\left(\frac{d_{\pi_E}-d_\pi}{d_{\pi_E}+d_\pi}\right)^2\right)da\,ds\\&=\int_{\mathcal{S}}\int_{\mathcal{A}}d_{\pi_E}\left(\frac{2d_{\pi_E}^2-2d_\pi^2-d_{\pi_E}^2+2d_{\pi_E}d_\pi-d_\pi^2}{2c(d_{\pi_E}+d_\pi)^2}\right)\\&\quad-d_\pi\left(\frac{2d_{\pi_E}^2-2d_\pi^2+d_{\pi_E}^2-2d_{\pi_E}d_\pi+d_\pi^2}{2c(d_{\pi_E}+d_\pi)^2}\right)da\,ds\\&=\int_{\mathcal{S}}\int_{\mathcal{A}}\frac{d_{\pi_E}^3+d_\pi^3-d_{\pi_E}d_\pi^2-d_{\pi_E}^2d_\pi}{2c(d_{\pi_E}+d_\pi)^2}da\,ds=\frac{1}{2c}\int_{\mathcal{S}}\int_{\mathcal{A}}\frac{(d_{\pi_E}-d_\pi)^2}{d_{\pi_E}+d_\pi}da\,ds\\&=\frac{1}{2c}\int_{\mathcal{S}}\int_{\mathcal{A}}\frac{d_{\pi_E}^2}{d_{\pi_E}+d_\pi}+\frac{d_\pi^2}{d_{\pi_E}+d_\pi}-2\frac{d_{\pi_E}d_\pi}{d_{\pi_E}+d_\pi}da\,ds\\&=\frac{1}{2c}\Bigg(\underbrace{\mathbb{E}_{d_{\pi_E}}\left[\frac{d_{\pi_E}}{d_{\pi_E}+d_\pi}\right]}_{\le1}+\underbrace{\mathbb{E}_{d_\pi}\left[\frac{d_\pi}{d_{\pi_E}+d_\pi}\right]}_{\le1}+\underbrace{\mathbb{E}_{d_{\pi_E}}\left[\frac{-2d_\pi}{d_{\pi_E}+d_\pi}\right]}_{\le0}\Bigg)\le\frac{1}{c}.\end{aligned}\tag{30}$$

请注意，该界是紧的，因为每个期望的单独界只有同时成立时才能达到。■

命题 A.4 令 $\chi^2\big(d_{\pi_E}\,\|\,\alpha d_{\pi_E}+(1-\alpha)d_\pi\big)$ 为分布 d_{π_E}与混合分布 $\alpha d_{\pi_E}+(1-\alpha)d_\pi$之间的 *Pearson* χ^2-散度。则有：

$$\chi^2\big(d_{\pi_E}\,\|\,\alpha d_{\pi_E}+(1-\alpha)d_\pi\big)\le(1-\alpha)\chi^2\big(d_{\pi_E}\,\|\,d_\pi\big)\,.$$

命题证明 *A.4*。 证明可直接由 f-散度的联合凸性得出：

$$D_f\big(\kappa P_1+(1-\kappa)P_2\,\|\,\kappa Q_1+(1-\kappa)Q_2\big)\le\kappa D_f(P_1\,\|\,Q_1)+(1-\kappa)D_f(P_2\,\|\,Q_2)\,,$$

其中 $\kappa \in [0,1]$ 是混合系数。 χ^2-散度是一种 f-散度，其中 $f(t) = (t-1)^2$。现在使用 χ^2-散度，并令 $P = P_1 = P_2 = Q_2$、 $D = Q_1$ 和 $\alpha = (1-\kappa)$:

$$\chi^2\big(P \,\big\|\, \alpha P + (1-\alpha)D\big) \leq (1-\alpha)\,\chi^2(P \,\|\, D) + \underbrace{\alpha\chi^2(P \,\|\, P)}_{=0}$$
$$\chi^2\big(P \,\big\|\, \alpha P + (1-\alpha)D\big) \leq (1-\alpha)\,\chi^2\big(P \,\big\|\, D\big)\,.$$

令 $P = d_{\pi_E}$ 和 $D = d_\pi$ 即可完成证明。 ■

A.3 从 χ^2正则化的最大熵逆强化学习到最小二乘奖励回归

回顾一下，基于占用度量的熵正则化逆向强化学习可表示为

$$L_\rho(r,\pi) = (-\beta H_\rho(\pi) - \mathbb{E}_{\rho_\pi}[r(s,a)]) + \mathbb{E}_{\rho_{\pi_E}}[r(s,a)] - \psi_\rho(r). \tag{31}$$

还可以回顾一下，利用软逆贝尔曼算子进行变量替换可得:

$$\begin{aligned}\mathcal{J}_\rho(\tilde{Q},\pi_{\tilde{Q}}) = \mathbb{E}_{\rho_{\pi_E}}\Big[\tilde{Q}(s,a) - \gamma\mathbb{E}_{s'\sim P(.|s,a)}[\tilde{V}^\pi(s')]\Big] - \beta H_\rho(\pi_{\tilde{Q}}) \\ - \mathbb{E}_{\rho_\pi}\Big[\tilde{Q}(s,a) - \gamma\mathbb{E}_{s'\sim P(.|s,a)}[\tilde{V}^\pi(s')]\Big] - \psi_\rho(r).\end{aligned} \tag{32}$$

现在可以使用命题 A.1，将目标函数从占用度量转换为分布。

命题的证明 *3.1*。 从公式 22 中的目标函数出发，通过展开期望并重新排列各项，我们得到:

$$\begin{aligned}\mathcal{J}(\tilde{Q},\pi_{\tilde{Q}}) =& \mathbb{E}_{d_{\pi_E}}\left[r_{\tilde{Q}}(s,a)\right] - \mathbb{E}_{d_\pi}\left[r_{\tilde{Q}}(s,a)\right] - c\,\mathbb{E}_{\tilde{d}}\left[r_{\tilde{Q}}(s,a)^2\right] - \beta H(\pi_{\tilde{Q}}) \\ =& \mathbb{E}_{d_{\pi_E}}\left[r_{\tilde{Q}}(s,a)\right] - \mathbb{E}_{d_\pi}\left[r_{\tilde{Q}}(s,a)\right] \\ & - c\alpha\,\mathbb{E}_{d_{\pi_E}}\left[r_{\tilde{Q}}(s,a)^2\right] - c(1-\alpha)\,\mathbb{E}_{d_\pi}\left[r_{\tilde{Q}}(s,a)^2\right] - \beta H(\pi_{\tilde{Q}}) \\ =& -\mathbb{E}_{d_{\pi_E}}\left[c\alpha\, r_{\tilde{Q}}(s,a)^2 - r_{\tilde{Q}}(s,a)\right] - \mathbb{E}_{d_\pi}\left[c(1-\alpha)\, r_{\tilde{Q}}(s,a)^2 + r_{\tilde{Q}}(s,a)\right] - \beta H(\pi_{\tilde{Q}}) \\ =& -c\Big(\alpha\,\mathbb{E}_{d_{\pi_E}}\left[r_{\tilde{Q}}(s,a)^2 - \tfrac{1}{\alpha c} r_{\tilde{Q}}(s,a)\right] \\ & + (1-\alpha)\,\mathbb{E}_{d_\pi}\Big[r_{\tilde{Q}}(s,a)^2 + \tfrac{1}{(1-\alpha)c} r_{\tilde{Q}}(s,a)\Big] + \tfrac{\beta}{c} H(\pi_{\tilde{Q}})\Big)\end{aligned}$$

定义 $r_{\max} = \frac{1}{2\alpha c}$ 和 $r_{\min} = -\frac{1}{2(1-\alpha)c}$ 并配方后，我们得到:

$$\begin{aligned}\mathcal{J}(\tilde{Q},\pi_{\tilde{Q}}) =& -c\Big(\alpha\,\mathbb{E}_{d_{\pi_E}}\left[r_{\tilde{Q}}(s,a)^2 - \tfrac{1}{\alpha c} r_{\tilde{Q}}(s,a)\right] + (1-\alpha)\,\mathbb{E}_{d_\pi}\Big[r_{\tilde{Q}}(s,a)^2 + \tfrac{1}{(1-\alpha)c} r_{\tilde{Q}}(s,a)\Big] \\ & + \tfrac{\beta}{c} H(\pi_{\tilde{Q}})\Big) + \alpha c\,(r_{\max}^2 - r_{\max}^2) + (1-\alpha)c\,(r_{\min}^2 - r_{\min}^2) \\ =& -c\Big(\alpha\,\mathbb{E}_{d_{\pi_E}}\left[r_{\tilde{Q}}(s,a)^2 - \tfrac{1}{\alpha c} r_{\tilde{Q}}(s,a) + r_{\max}^2\right] + (1-\alpha)\,\mathbb{E}_{d_\pi}\left[r_{\tilde{Q}}(s,a)^2\right. \\ & \left. + \tfrac{1}{(1-\alpha)c} r_{\tilde{Q}}(s,a) + r_{\min}^2\right] + \tfrac{\beta}{c} H(\pi_{\tilde{Q}})\Big) + \alpha c\, r_{\max}^2 + (1-\alpha)c\, r_{\min}^2 \\ =& -c\Big(\alpha\,\mathbb{E}_{d_{\pi_E}}\left[\left(r_{\tilde{Q}}(s,a) - \tfrac{1}{2\alpha c}\right)^2\right] + (1-\alpha)\,\mathbb{E}_{d_\pi}\left[\left(r_{\tilde{Q}}(s,a) + \tfrac{1}{2(1-\alpha)c}\right)^2\right] \\ & + \tfrac{\beta}{c} H(\pi_{\tilde{Q}})\Big) + \alpha c\, r_{\max}^2 + (1-\alpha)c\, r_{\min}^2.\end{aligned}$$

最后，我们得到以下结果:

$$\mathcal{J}(\tilde{Q},\pi_{\tilde{Q}}) = -c\Big(\alpha\,\mathbb{E}_{d_{\pi_E}}\left[\left(r_{\tilde{Q}}(s,a) - r_{\max}\right)^2\right] + (1-\alpha)\,\mathbb{E}_{d_\pi}\left[\left(r_{\tilde{Q}}(s,a) - r_{\min}\right)^2\right] + \tfrac{\beta}{c} H(\pi_{\tilde{Q}})\Big) + K, \tag{33}$$

其中 $K = \alpha c\, r_{\max}^2 + (1-\alpha)c\, r_{\min}^2 = \frac{1}{4\alpha c} + \frac{1}{4(1-\alpha)c}$ 是一个固定常数。

比较公式 9 和公式 33 可得

$$\mathcal{J}(\tilde{Q},\pi_{\tilde{Q}}) + K \propto \mathcal{L}(\tilde{Q},\pi_{\tilde{Q}}). \tag{34}$$

鉴于仿射变换（具有正的乘法常数）保持最优解，$\tilde{Q}^* \in \arg\max_{\tilde{Q}\in\Omega} \mathcal{J}(\tilde{Q}, \pi_{\tilde{Q}})$ 是熵正则化最小二乘目标 $\mathcal{L}(\tilde{Q}, \pi_{\tilde{Q}})$ 的解。

■

A.4 处理终止状态的完整 LS-IQ 目标函数

将第 3.4 节中推导的逆贝尔曼算子代入式（9）中定义的最小二乘目标函数，并重新整理各项，可得到以下针对硬 Q-函数的目标函数：

$$\begin{aligned}\mathcal{L}(Q,\pi_Q) = &\alpha\mathbb{E}_{\substack{s,a\sim d_{\pi_E}\\ s'\sim P(.|s,a)}}\left[(1-\nu)\,\left(Q(s,a)-(r_{\max}+\gamma\,V^{\pi}(s')\right)^2\right] \\ &+\alpha\mathbb{E}_{\substack{s,a\sim d_{\pi_E}\\ s'\sim P(.|s,a)}}\left[\nu\,\left(Q(s,a)-(r_{\max}+\gamma\,\tfrac{r_{\max}}{1-\gamma}\right)^2\right] \\ &+(1-\alpha)\mathbb{E}_{\substack{s,a\sim d_{\pi_Q}\\ s'\sim P(.|s,a)}}\left[(1-\nu)\,\left(Q(s,a)-(r_{\min}+\gamma\,V^{\pi}(s')\right)^2\right] \\ &+(1-\alpha)\mathbb{E}_{\substack{s,a\sim d_{\pi_Q}\\ s'\sim P(.|s,a)}}\left[\nu\,\left(Q(s,a)-(r_{\min}+\gamma\,\tfrac{r_{\min}}{1-\gamma})\right)^2\right]+\frac{\beta}{c}H(\pi_Q)\end{aligned} \tag{35}$$

$$\begin{aligned}\mathcal{L}(Q,\pi_Q) = &\alpha\mathbb{E}_{\substack{s,a\sim d_{\pi_E}\\ s'\sim P(.|s,a)}}\left[(1-\nu)\,\left(Q(s,a)-(r_{\max}+\gamma\,V^{\pi}(s')\right)^2\right] \\ &+\alpha\mathbb{E}_{\substack{s,a\sim d_{\pi_E}\\ s'\sim P(.|s,a)}}\left[\nu\,\left(Q(s,a)-\tfrac{r_{\max}}{1-\gamma}\right)^2\right] \\ &+(1-\alpha)\mathbb{E}_{\substack{s,a\sim d_{\pi_Q}\\ s'\sim P(.|s,a)}}\left[(1-\nu)\,\left(Q(s,a)-(r_{\min}+\gamma\,V^{\pi}(s')\right)^2\right] \\ &+(1-\alpha)\mathbb{E}_{\substack{s,a\sim d_{\pi_Q}\\ s'\sim P(.|s,a)}}\left[\nu\,\left(Q(s,a)-\tfrac{r_{\min}}{1-\gamma}\right)^2\right]+\frac{\beta}{c}H(\pi_Q)\,.\end{aligned} \tag{36}$$

现在加入第3.5节中引入的专家分布固定目标，可得：

$$\begin{aligned}\mathcal{L}_{\text{lsiq}}(Q,\pi_Q) = &\alpha\mathbb{E}_{\substack{s,a\sim d_{\pi_E}\\ s'\sim P(.|s,a)}}\left[(1-\nu)\,\left(Q(s,a)-\tfrac{r_{\max}}{1-\gamma}\right)^2\right] \\ &+\alpha\mathbb{E}_{\substack{s,a\sim d_{\pi_E}\\ s'\sim P(.|s,a)}}\left[\nu\,\left(Q(s,a)-\tfrac{r_{\max}}{1-\gamma}\right)^2\right] \\ &+(1-\alpha)\mathbb{E}_{\substack{s,a\sim d_{\pi_Q}\\ s'\sim P(.|s,a)}}\left[(1-\nu)\,\left(Q(s,a)-(r_{\min}+\gamma\,V^{\pi}(s')\right)^2\right] \\ &+(1-\alpha)\mathbb{E}_{\substack{s,a\sim d_{\pi_Q}\\ s'\sim P(.|s,a)}}\left[\nu\,\left(Q(s,a)-\tfrac{r_{\min}}{1-\gamma}\right)^2\right]+\frac{\beta}{c}H(\pi_Q)\end{aligned} \tag{37}$$

$$\begin{aligned}\mathcal{L}_{\text{lsiq}}(Q,\pi_Q) = &\alpha\mathbb{E}_{\substack{s,a\sim d_{\pi_E}\\ s'\sim P(.|s,a)}}\left[\left(Q(s,a)-\tfrac{r_{\max}}{1-\gamma}\right)^2\right] \\ &+(1-\alpha)\mathbb{E}_{\substack{s,a\sim d_{\pi_Q}\\ s'\sim P(.|s,a)}}\left[(1-\nu)\,\left(Q(s,a)-(r_{\min}+\gamma\,V^{\pi}(s')\right)^2\right] \\ &+(1-\alpha)\mathbb{E}_{\substack{s,a\sim d_{\pi_Q}\\ s'\sim P(.|s,a)}}\left[\nu\,\left(Q(s,a)-\tfrac{r_{\min}}{1-\gamma}\right)^2\right]+\frac{\beta}{c}H(\pi_Q)\,,\end{aligned} \tag{38}$$

其中，公式 38 是我们针对硬 Q-函数的完整 LS-IQ 目标函数。对于仅观测设置，我们使用 IDM 预测专家动作。

A.5 前向备份算子的收敛性

如第 3.4 节所述，我们的逆算子，

$$(\mathcal{T}^{\pi}_{\text{lsiq}}Q)(s,a) = Q(s,a) - \gamma\mathbb{E}_{s'\sim P(.|s,a)}\left((1-\nu)V^{\pi}(s') + \nu V(s_A)\right), \tag{39}$$

基于标准的贝尔曼备份算子，但我们不使用自举，而是采用转移至吸收状态时已知的值。现在我们将展示，反复应用相应的前向算子

$$(\mathcal{B}^{\pi}_{\text{lsiq}}Q)(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim P(.|s,a)}\left((1-\nu)V^{\pi}(s') + \nu V(s_A)\right), \tag{40}$$

收敛到 Q 函数。我们的证明基于一种常用技术，该技术通常用于证明标准 Bellman 算子的收敛性，即证明 Q 函数

$Q^\pi(s,a) := r(s,a) + \gamma E_{s'\sim p(s'|s,a)}\mathbb{E}_{a'\sim\pi(a'|s')}Q^\pi(s',a')$ 是我们算子的一个不动点，也就是说，$(\mathcal{B}^\pi_{\text{lsiq}}Q^\pi)(s,a) = Q^\pi(s,a)$，并进一步证明我们的算子是一个压缩映射，

$$||(\mathcal{B}^\pi_{\text{lsiq}}Q_A)(s,a) - (\mathcal{B}^\pi_{\text{lsiq}}Q_B)(s,a)||_\infty \leq \gamma||Q_A(s,a) - Q_B(s,a)||_\infty, \tag{41}$$

其中 $||.||_\infty$ 是最大范数；为简洁起见，此处假设状态和动作的数量有限。

命题 A.5 策略 π 的 Q 函数是 $\mathcal{B}^\pi_{\text{lsiq}}$ 的一个不动点，

$$(\mathcal{B}^\pi_{\text{lsiq}}Q^\pi(s,a)) = Q^\pi(s,a)\,. \tag{42}$$

命题证明*A.5*。 证明直接来自这样一个事实：如果将我们的前向算子应用于策略的实际 Q 函数 $Q^\pi(s,\ a)$，它执行的更新与标准贝尔曼算子相同，因为此时 $V(s_A) := {}^{r_{(s_A)}}{}_{(1-\gamma)} = \mathbb{E}_{a'\sim\pi(.|_A)}Q^\pi(s',\ a')$。因此，

$$\begin{aligned}
(\mathcal{B}^\pi_{\text{lsiq}}Q^\pi(s,a)) &= r(s,a) + \gamma\mathbb{E}_{s'\sim P(.|s,a)}\left((1-\nu)V^\pi(s') + \nu V(s_A)\right) && (43)\\
&= r(s,a) + \gamma\mathbb{E}_{s'\sim P(.|s,a)}\left((1-\nu)\mathbb{E}_{a'\sim\pi(a'|s')}Q^\pi(s',a') + \nu\mathbb{E}_{a'\sim\pi(.|s')}Q^\pi(s',a')\right) && (44)\\
&= r(s,a) + \gamma\mathbb{E}_{s'\sim P(.|s,a)}\mathbb{E}_{a'\sim\pi(.|s')}Q^\pi(s',a') = Q^\pi(s,a)\,. && (45)
\end{aligned}$$

■

命题 A.6 前向算子 $\mathcal{B}^\pi_{\text{lsiq}}$ 是一个压缩映射，

$$||(\mathcal{B}^\pi_{\text{lsiq}}Q_A)(s,a) - (\mathcal{B}^\pi_{\text{lsiq}}Q_B)(s,a)||_\infty \leq \gamma||Q_A(s,a) - Q_B(s,a)||_\infty. \tag{46}$$

命题证明*A.6*。

$$\begin{aligned}
&\left|\left|(\mathcal{B}^\pi_{\text{lsiq}}Q_A)(s,a) - (\mathcal{B}^\pi_{\text{lsiq}}Q_B)(s,a)\right|\right|_\infty && (47)\\
=&\max_{s,a}\left|\gamma\mathbb{E}_{s'\sim P(.|s,a)}\mathbb{E}_{a'\sim\pi(a'|s')}\left[(1-\nu)(Q_A(s',a') - Q_B(s',a'))\right]\right| && (48)\\
\leq&\gamma\max_{s,a}\left|\mathbb{E}_{s'\sim P(.|s,a)}\mathbb{E}_{a'\sim\pi(a'|s')}\left[(Q_A(s',a') - Q_B(s',a'))\right]\right| && (49)\\
\leq&\gamma\max_{s',a'}\left|Q_A(s',a') - Q_B(s',a')\right| && (50)\\
=&\gamma\left|\left|Q_A(s,a) - Q_B(s,a)\right|\right|_\infty && (51)
\end{aligned}$$

■

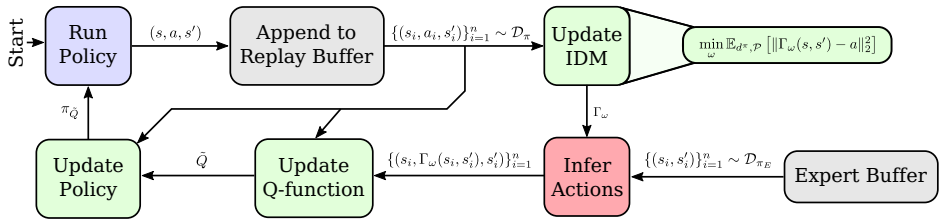

图 4：LS-IQ 中 IDM 的训练流程。

B 从观测中学习

本节将更详细地介绍 IDM。图 4 展示了 LS-IQ 中 IDM 的训练过程。如图所示，IDM 使用智能体生成的回放缓冲区数据，根据状态转移推断动作。因此，我们使用了公式 21 中简单的回归损失。在训练开始时，IDM 使用（随机）智能体生成的转移进行学习。当智能体逐渐接近专家分布时，IDM 则使用更接近专家的转移进行训练。使用 IDM 的直觉源于这样一种认识：虽然两个智能体可能产生不同的轨迹，进而产生不同的状态-动作分布，但它们共享相同的底层动力学。这使得 IDM 能够推断出有关与某个状态转移对应的动作的更多信息，而策略预测的随机动作（Garg 等人（2021）所采用的方法）无法提供这些信息。第 4 节的实验表明，在仅状态场景中，使用 IDM 的性能优于或不逊于基线方法。

虽然我们只介绍了确定性 IDM，但也尝试过随机版本。例如，我们将 IDM 建模为高斯分布，并使用最大似然损失对其进行训练。我们还尝试了完全贝叶斯方法来施加先验：学习正态-逆伽马分布的参数，并使用学生 t 分布预测状态转移的动作，这与 Amini 等人（2020年）的方法相同。然而，随机方法并未带来明显益处，因此我们采用简单的确定性方法。

C 实验

本节介绍环境描述以及因论文篇幅限制而未在正文中呈现的补充结果。

C.1 ATLAS 运动环境

Atlas 运动环境是我们提出的一种新型运动环境。与 Mujoco Gym 任务不同，这种环境旨在让智能体在更真实的任务上接受训练；Mujoco Gym 任务通常具有经过精细调校、专为强化学习智能体设计的动力学特性。Atlas 环境默认固定手臂，因此有 10 个活动关节。每个关节由一个电机通过扭矩控制。状态空间包括所有关节的位置和速度，以及前脚和后脚所受的三维力，因此状态维度为 $D_s = 20 + 2 \cdot 2 \cdot 3 = 32$。动作空间包括每个关节电机的期望扭矩，因此动作维度为$D_a = 10$。此外，还可以启用包含手臂的上半身，将关节和执行器的数量扩展至 27 个。Atlas 环境使用 Mushroom-RL 的 Mujoco 接口实现（D'Eramo 等人，2021年）。

为确保内容完整，我们在图 5 中加入了一个额外的 VAIL 智能体，并展示累积奖励图；这与图 2 中的折扣累积奖励图相对照。用于衡量性能的奖励定义为 $r = \exp(-(v_\pi - v_{\pi_E})^2)$，其中 v_π 是智能体的上身速度，v_{π_E} 是专家的上身速度。专家的行走速度为 $1.25 \frac{m}{s}$。环境、代码和专家数据可在 https://github.com/ robfiras/ls-iq 获取。

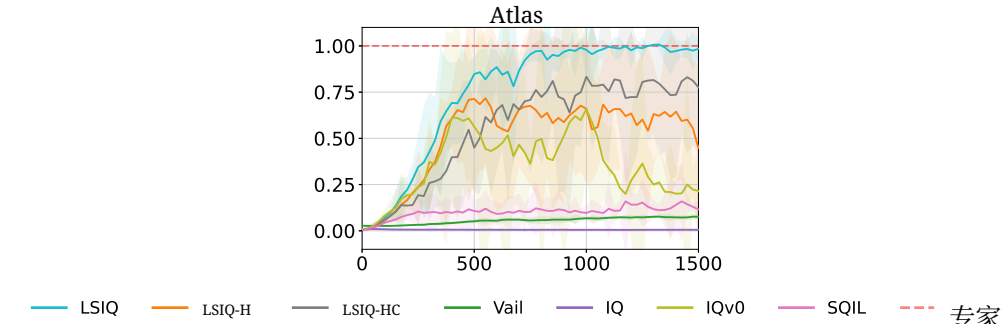

图 5：运动任务的训练结果和示例轨迹——此处为训练得到的 LSIQ 智能体——该任务使用仿真阿特拉斯机器人。横轴表示归一化累积奖励。纵轴表示训练步数（$\times 10^3$）。

C.2 消融研究：吸收状态处理与目标值 Q 裁剪

图6展示了针对采用自举的 LSIQ 智能体，所提出的吸收状态处理方法以及对 Q-函数目标值进行裁剪的影响消融实验。为观察吸收状态处理和裁剪本身的效果，我们没有纳入固定目标值。请注意，固定目标值会隐式处理专家的吸收状态，因为它会为转移至吸收状态的状态和动作提供相同的目标值。我们选择了一个不含熵项和正则化评论家的 LSIQ 智能体。实验在 Humanoid-v3任务上进行，因为 HalfCheetah-v3、Ant-v3 要么没有吸收状态，要么吸收状态极少，而 Walker-v3 和 Hopper-v3 则过于简单，难以观察到这些效果。我们使用的正则化常数为 $c = 0.5$，混合参数为 $\alpha = 0.5$，由此得到的最大奖励为 $r_{\max} = \frac{1}{2\ 1-0.5()0.5} = 2$，最小奖励为 $r_{\min} = -\frac{1}{2\ 1-0.5()0.5} = -2$。由此得到的最大 Q值为 $Q_{\max} = \frac{r_{\max}}{1-\gamma} = 200$，最小 Q值为 $Q_{\min} = \frac{r_{\min}}{1-\gamma} = -200$。各行展示了不同的智能体配置：第一，采用裁剪和吸收状态处理的 LSIQ；第二，采用裁剪但不进行吸收状态处理的 LSIQ；最后，既不裁剪也不处理吸收状态的 LSIQ——这等同于使用对称奖励目标的 SQIL。为更好地理解其影响，图中展示了各配置下每个随机种子的结果。可以看到，第一个 LSIQ 智能体能够成功学会任务，并将转移至吸收状态的 Q值回归至最小Q值 -200。第二种配置不处理吸收状态，因此无法在所有随机种子下学会任务。可以看到，吸收状态的平均 Q值介于 -2 和 -6 之间。仔细观察第二行的前两幅图可以发现，那些未能学会任务的随机种子，其非吸收转移上的平均 Q值接近甚至低于导致吸收状态的状态和动作的平均Q值。这进一步说明了终止状态处理的重要性：它将通向吸收状态的状态和动作的 Q值拉至可能的最低 Q值，从而避免终止偏差。最后，可以看到最后一种 LSIQ 配置存在 Q值爆炸的问题。图中横坐标的刻度范围清楚地体现了这一点。有趣的是，尽管 Q值高得惊人，部分随机种子的表现仍然不错，但这显然与累积奖励图中的高方差相关。

C.3 消融研究：固定目标与熵裁剪的影响

为展示固定目标（参见第3.5节）和熵裁剪（参见3.7）的效果，我们针对所有Mujoco任务中不同版本的LSIQ开展了一系列消融研究。图7展示了仅包含熵评论家的LSIQ版本的结果，图8展示了同时包含熵评论家和正则化评论家的LSIQ版本的结果。从图中可以看出，采用固定目标和熵裁剪的版本表现最佳。尤其值得注意的是，对于需要较高温度参数 β的任务，熵裁剪尤为重要，而Humanoid-v3环境正属于这种情况。

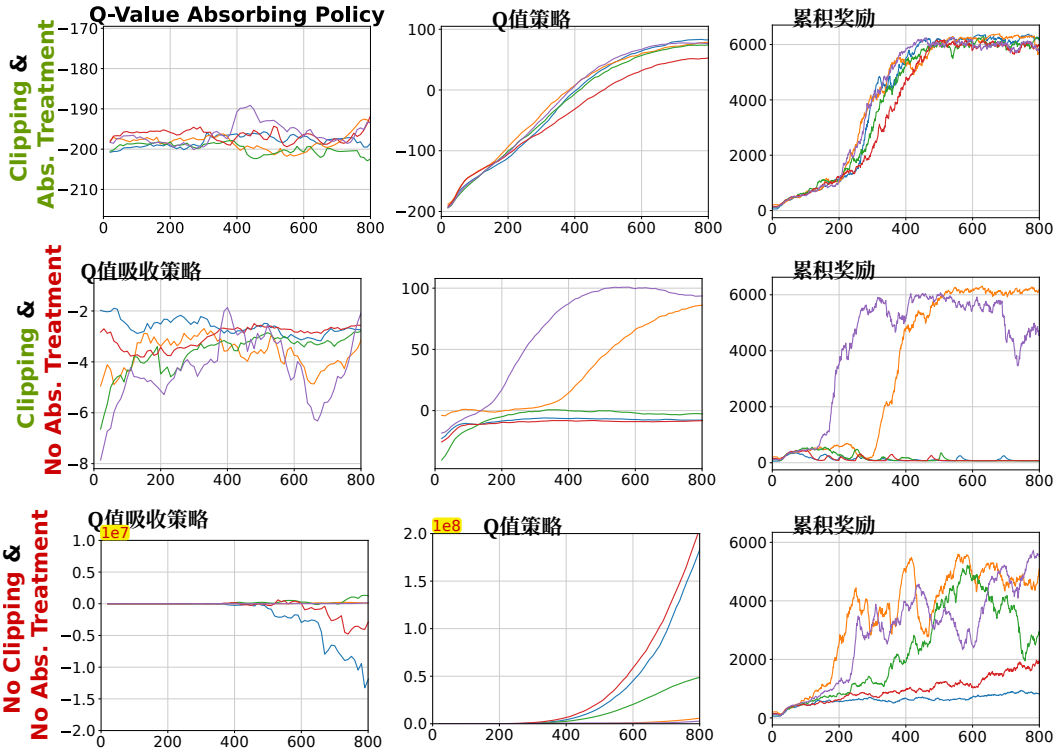

图6：针对吸收状态处理方法和 Q-值目标裁剪效果的消融研究，实验对象为采用自举的 LSIQ 智能体（无固定目标）。实验在 Humanoid-v3任务上进行，专家的累积奖励为 6233。45。每幅图中的多条曲线表示**不同的随机种子**。第一列展示策略访问的、**会进入吸收状态的** 状态和动作的平均 Q-值。第二列展示策略访问的、不会进入吸收状态的所有状态和动作的平均 Q-值。第三列展示累积奖励。各行展示对 LSIQ 智能体进行的消融实验。纵轴表示训练步数（$\times 10^3$）。

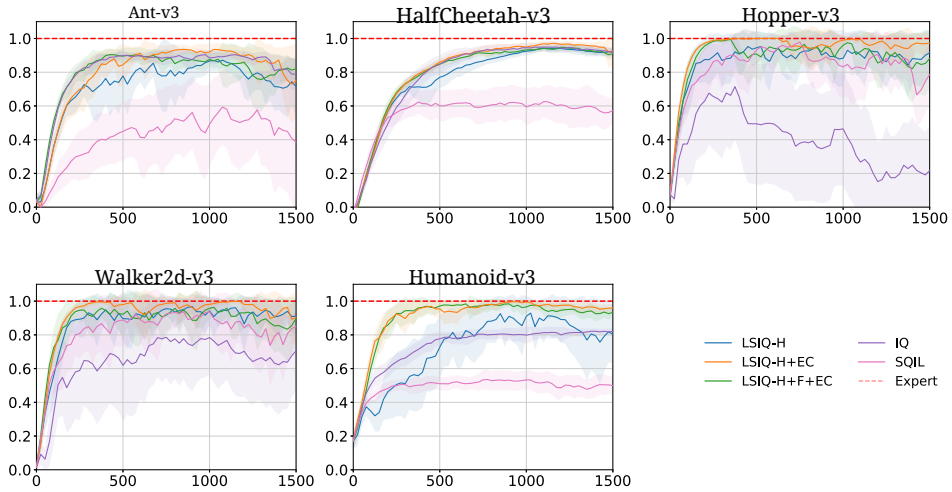

图7：不同版本的**LSIQ-H**（**正则化评论家**）对比：第一种是自举版本 $\rightarrow$LSIQ-H；第二种是带熵裁剪的自举版本 $\rightarrow$LSIQ-H+EC；第三种是带熵裁剪的固定目标版本 $\rightarrow$LSIQ-H+EC+FT。另加入 IQ 和 SQIL 作为参考。横轴表示归一化折扣累积奖励。纵轴表示训练步数（$\times 10^3$）。实验使用**5** 条专家轨迹和五个随机种子进行。专家累积奖励 $\rightarrow$Hopper：3299.81，Walker2d：5841.73，Ant：6399.04，HalfCheetah：12328.78，Humanoid：6233.45

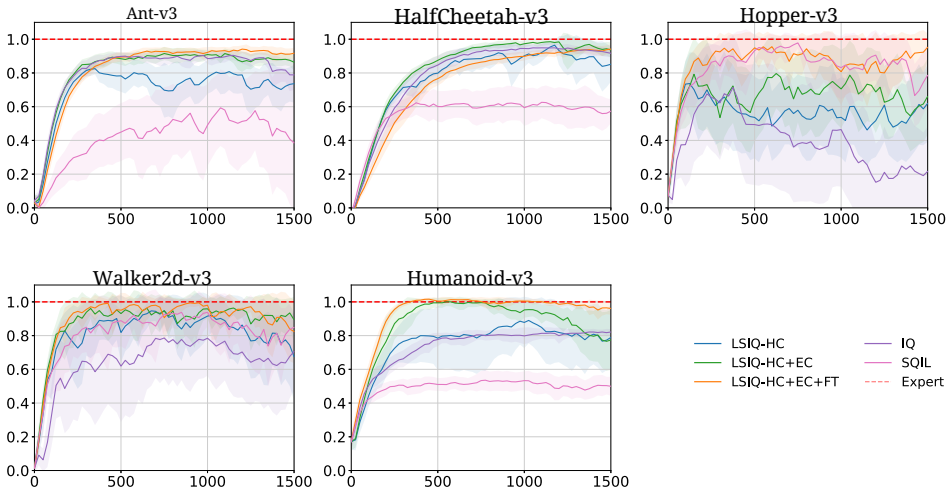

图8：不同版本的**LSIQ-HC（熵+正则化评论家）**对比：首先是自举版本 $\rightarrow$ LSIQ-HC；其次是带熵裁剪的自举版本 $\rightarrow$ LSIQ-HC+EC；第三是带熵裁剪的固定目标版本 $\rightarrow$ LSIQ-HC+EC+FT。加入 IQ 和 SQIL 作为参考。横坐标表示归一化折扣累积奖励。纵坐标表示训练步数（$\times 10^3$）。实验使用**5** 条专家轨迹和五个随机种子进行。专家累积奖励 →Hopper:3299.81, Walker2d:5841.73, Ant:6399.04, HalfCheetah:12328.78, Humanoid:6233.45

C.4 LSIQ 不同版本的所有实验结果

图9展示了图 2 中的所有图表，并补充了 Humanoid-v3 的结果。图10和图11对应图 3，但展示了全部五种环境。相应的学习曲线见附录 C.6 和 C.5。

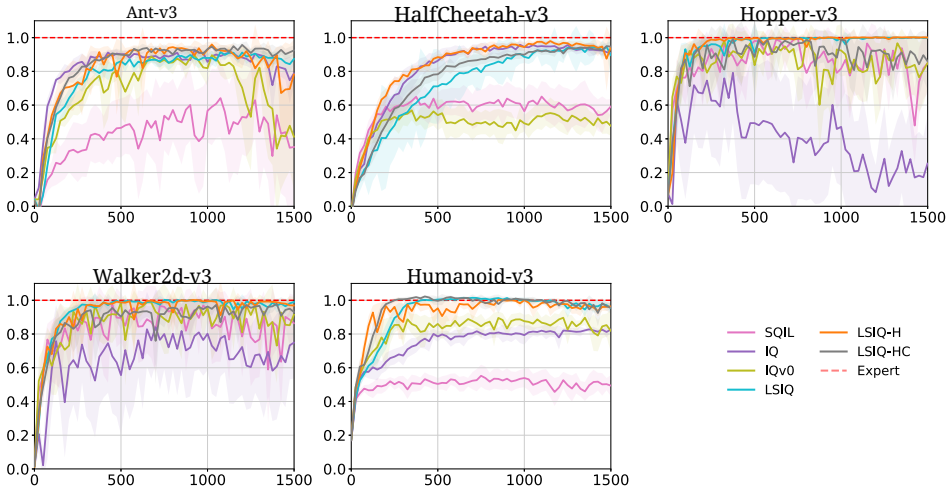

图9：LS-IQ 不同版本的比较。现在还包括 Humanoid-v3环境。横轴表示归一化折扣累积奖励。纵轴表示训练步数（$\times 10^3$）。专家累积奖励 $\rightarrow$ Hopper：3299.81，Walker2d：5841.73，Ant：6399.04，HalfCheetah：12328.78，Humanoid：6233.45

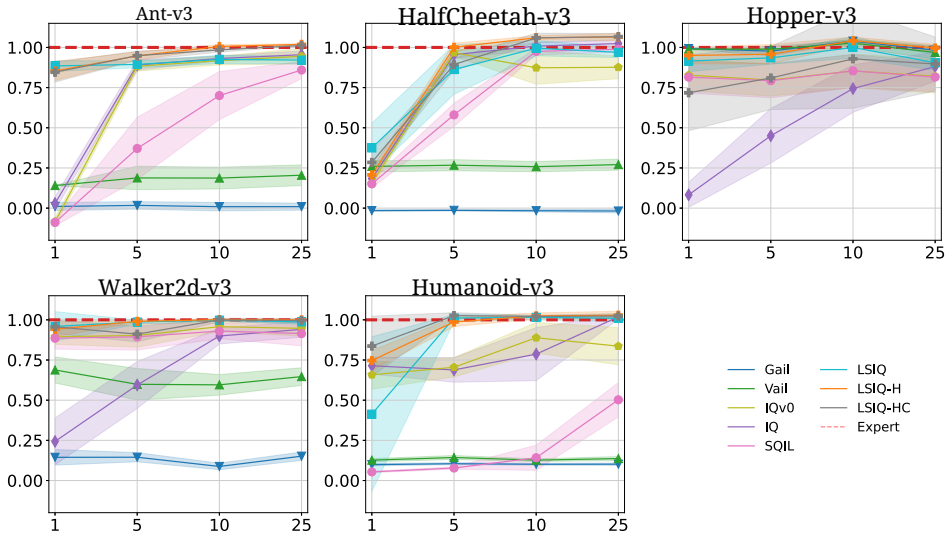

图10：不同 Mujoco 环境中专家轨迹数量影响的比较。智能体获得专家提供的**状态和动作** 。为确保完整性，此处列出了所有图表。横轴表示归一化累积奖励，纵轴表示专家轨迹数量。第一行展示考虑状态和动作时的性能，第二行展示仅使用状态时的性能。每个智能体的训练结果取五个随机种子的平均值。阴影区域表示 95% 置信区间。专家累积奖励 → Hopper:3299.81, Walker2d:5841.73, Ant:6399.04, HalfCheetah:12328.78, Humanoid:6233.45

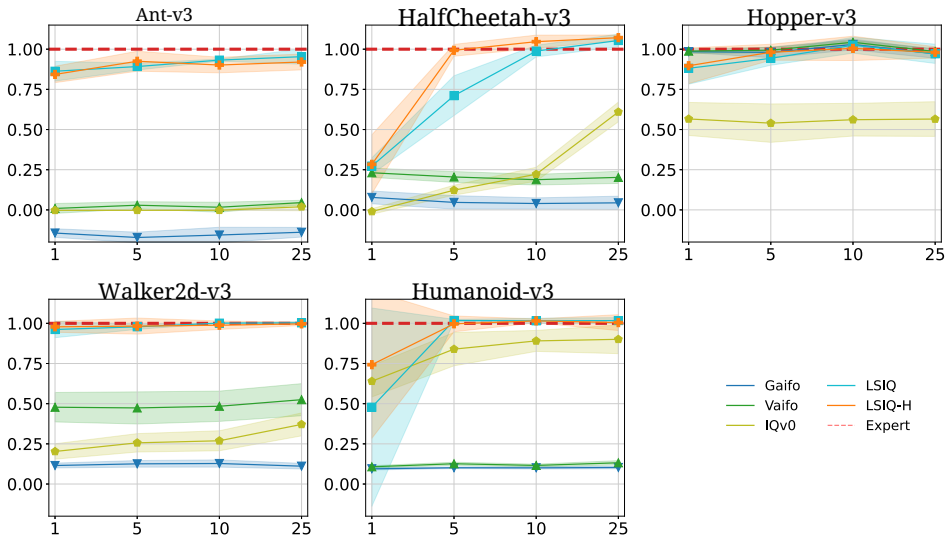

图11：专家轨迹数量对不同 MuJoCo 环境的影响比较。向专家提供**仅专家状态**。为完整起见，此处列出了所有图表。横坐标表示归一化累积奖励，纵坐标表示专家轨迹数量。第一行展示考虑状态和动作时的性能，第二行展示仅使用状态时的性能。每个智能体的训练结果取五个随机种子的平均值。阴影区域表示 95% 置信区间。专家累积奖励 → Hopper：3299.81，Walker2d：5841.73，Ant：6399.04，HalfCheetah：12328.78，Humanoid：6233.45

C.5 仅从状态进行模仿学习——完整训练曲线

从观测中学习实验的学习曲线分别见图 12、13、14 和 15，对应 1、5、10 和 25 个专家示范。

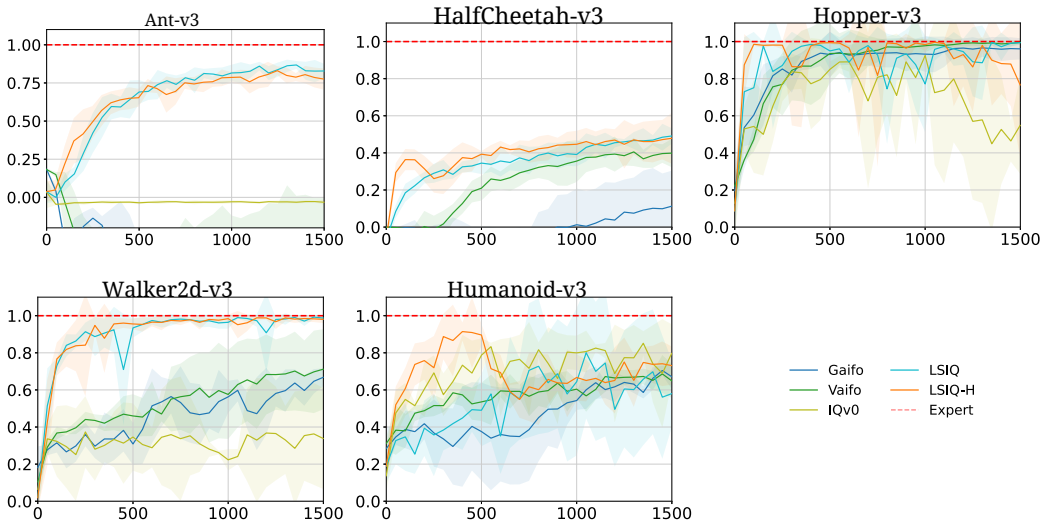

图 12：使用由**1** 条**仅包含状态**的专家轨迹时，不同智能体在 MuJoCo 任务上的训练性能。横坐标表示归一化折扣累积奖励。纵坐标表示训练步数（$\times 10^3$）。每个智能体的训练结果取五个随机种子的平均值。阴影区域表示 95% 置信区间。专家累积奖励 $\rightarrow$ Hopper:3299.81, Walker2d:5841.73, Ant:6399.04, HalfCheetah:12328.78, Humanoid:6233.45

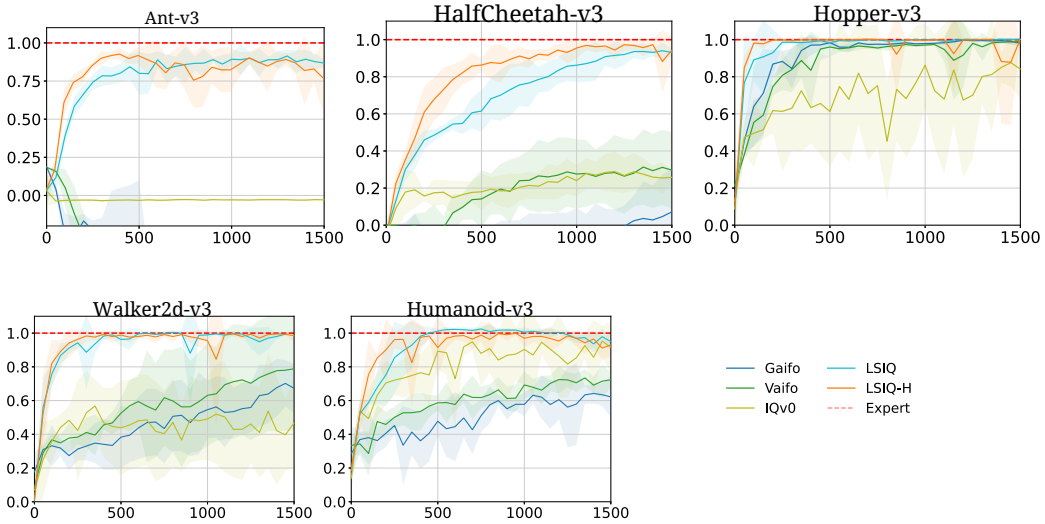

图 13：使用由**5** 条**仅包含状态**的专家轨迹时，不同智能体在 MuJoCo 任务上的训练性能。横坐标表示归一化折扣累积奖励。纵坐标表示训练步数（$\times 10^3$）。每个智能体的训练结果取十个随机种子的平均值。阴影区域表示 95% 置信区间。专家累积奖励 $\rightarrow$ Hopper:3299.81, Walker2d:5841.73, Ant:6399.04, HalfCheetah:12328.78, Humanoid:6233.45

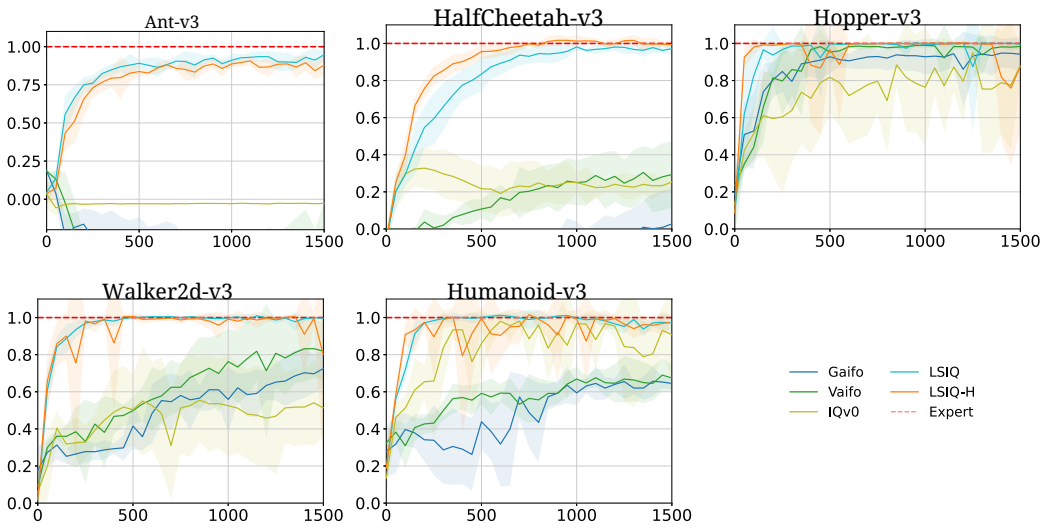

图 14：使用仅由**10** 条**状态**组成的专家轨迹时，不同智能体在 MuJoCo 任务上的训练性能。横坐标表示归一化折扣累积奖励。纵坐标表示训练步数（$\times 10^3$）。每个智能体的训练结果取五个随机种子的平均值。阴影区域表示 95% 置信区间。专家累积奖励 → Hopper:3299.81, Walker2d:5841.73, Ant:6399.04, HalfCheetah:12328.78, Humanoid:6233.45

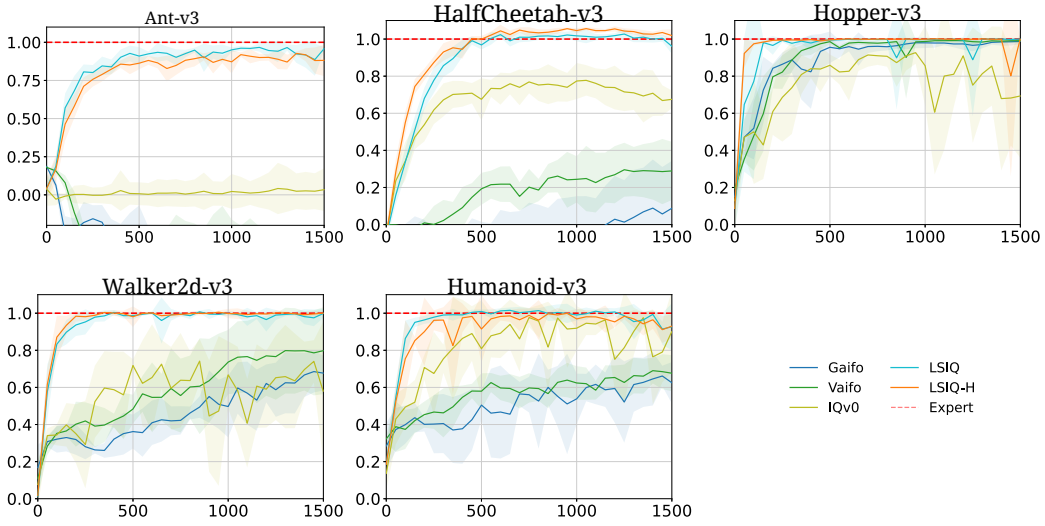

图 15：使用由**25** 条**仅包含状态**的专家轨迹时，不同智能体在 MuJoCo 任务上的训练性能。横坐标表示归一化折扣累积奖励。纵坐标表示训练步数（$\times 10^3$）。每个智能体的训练结果均为五个随机种子的平均值。阴影区域表示 95% 置信区间。专家累积奖励 → Hopper：3299.81，Walker2d：5841.73，Ant：6399.04，HalfCheetah：12328.78，Humanoid：6233.45

C.6 基于状态和动作的模仿学习——完整训练曲线

观测到状态和动作的实验学习曲线分别见图 16、17、18 和 19，对应 1、5、10 和 25 条专家示范。

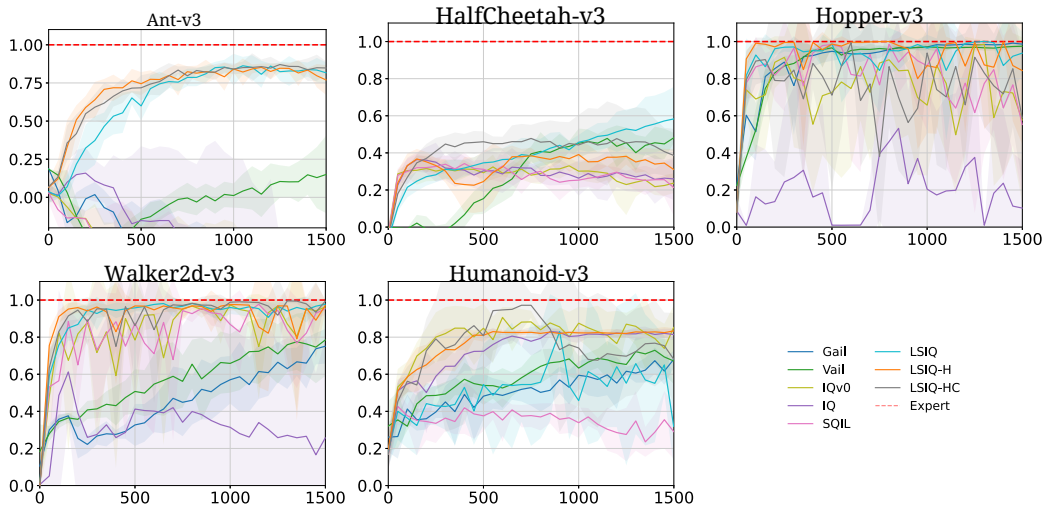

图 16：使用**1** 条专家轨迹时，不同智能体在 MuJoCo 任务上的训练性能。横坐标表示归一化折扣累积奖励。纵坐标表示训练步数（$\times 10^3$）。每个智能体的训练结果取五个随机种子的平均值。阴影区域表示 95% 置信区间。专家累积奖励 $\rightarrow$ Hopper:3299.81, Walker2d:5841.73, Ant:6399.04, HalfCheetah:12328.78, Humanoid:6233.45

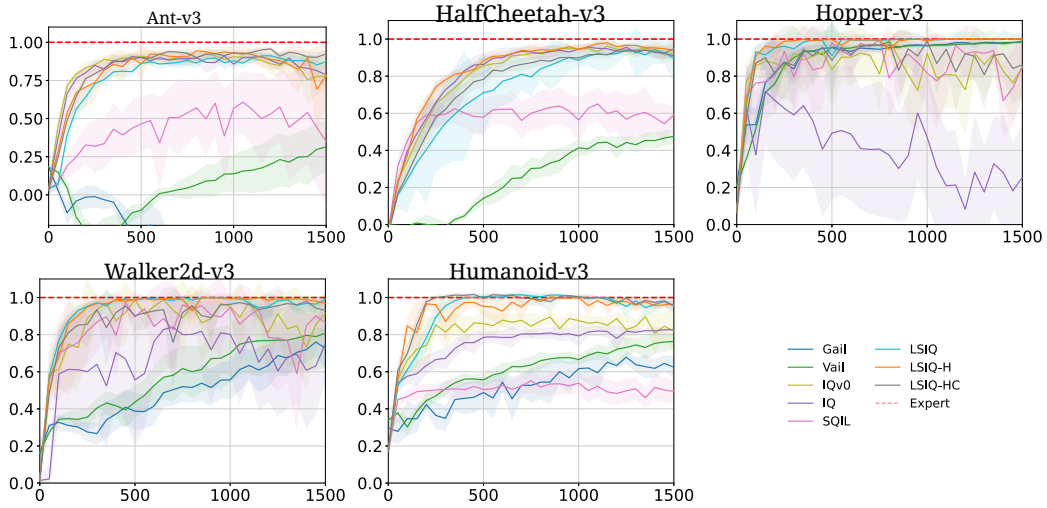

图 17：使用**5** 条专家轨迹时，不同智能体在 MuJoCo 任务上的训练性能。横坐标表示归一化折扣累积奖励。纵坐标表示训练步数（$\times 10^3$）。每个智能体的训练结果取十个随机种子的平均值。阴影区域表示 95% 置信区间。专家累积奖励 $\rightarrow$ Hopper:3299.81, Walker2d:5841.73, Ant:6399.04, HalfCheetah:12328.78, Humanoid:6233.45

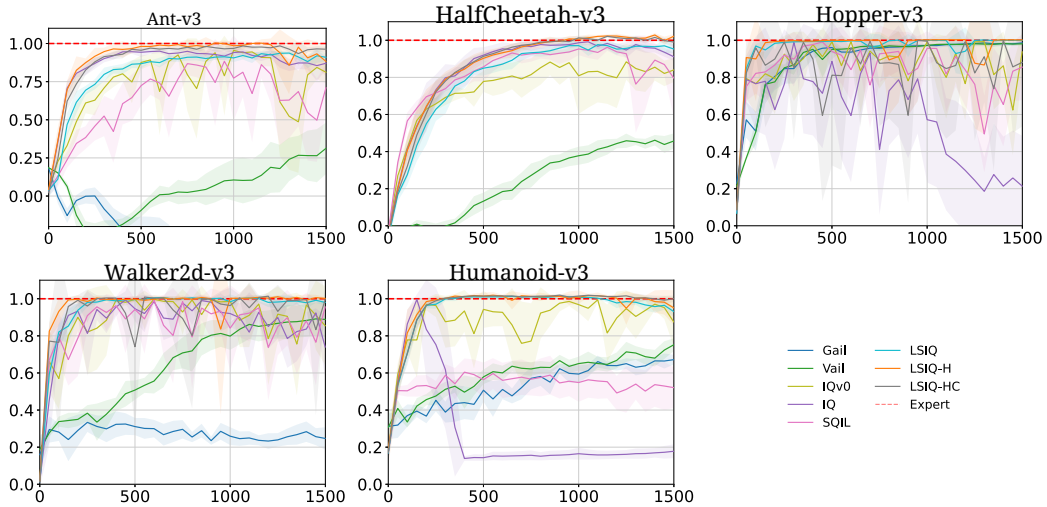

图 18：使用**10** 条专家轨迹时，不同智能体在 MuJoCo 任务上的训练性能。横坐标表示归一化折扣累积奖励。纵坐标表示训练步数（$\times 10^3$）。每个智能体的训练结果均取五个随机种子的平均值。阴影区域表示 95% 置信区间。专家累积奖励 $\rightarrow$ Hopper:3299.81, Walker2d:5841.73, Ant:6399.04, HalfChee- tah:12328.78, Humanoid:6233.45

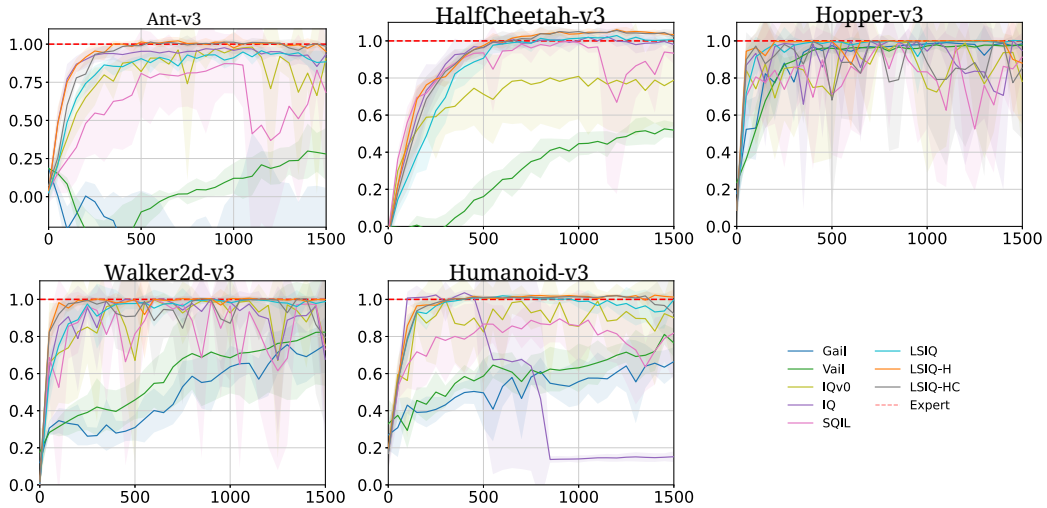

图 19：使用 **25** 条专家轨迹时，不同智能体在 Mujoco 任务上的训练性能。横坐标表示归一化折扣累积奖励，纵坐标表示训练步数（$\times 10^3$）。每个智能体的训练结果为五个随机种子的平均值。阴影区域表示 95% 置信区间。专家累积奖励 $\rightarrow$ Hopper：3299.81，Walker2d：5841.73，Ant：6399.04，HalfCheetah：12328.78，Humanoid：6233.45