文章编号:1671-4598(2025)01-0235-07 DOI:10.16526/j.cnki.11-4762/tp.2025.01.030 中图分类号:TP181 文献标识码:A

带最大熵修正和 GAIL 的 PPO 算法

王泽宁，刘 蕾
（中国电子科技集团公司 第 15 研究所，北京 100083）

摘要：为提高智能体在策略优化过程中的探索性和稳定性，改善强化学习中智能体陷入局部最优和奖励函数设置问题，提出了一种基于最大熵修正和 GAIL 的 PPO 算法；在 PPO 框架内引入最大熵修正项，通过优化策略熵，鼓励智能体在多个可能的次优策略间进行探索，从而更全面地评估环境并发现更优策略；同时，为解决强化学习过程中因奖励函数设置不合理引起的训练效果不佳问题，引入 GAIL 思想，通过专家数据指导智能体进行学习；实验表明，引入最大熵修正项和 GAIL 的 PPO 算法在强化学习任务上取得了良好的性能，有效提升了学习速度和稳定性，且能有效规避因环境奖励函数设置不合理引起的性能损失；该算法为强化学习领域提供了一种新的解决策略，对于处理具有挑战性的连续控制问题具有重要意义。

关键词：强化学习；PPO 算法；生成式对抗模仿学习；深度学习；最大熵学习

PPO Algorithm with Maximum Entropy Correction and GAIL

WANG Zening，LIU Lei
（The 15th Research Institute，China Electronics Technology Group Corporation，Beijing 100083，China）

Abstract：To enhance the exploration and stability of intelligent agents during policy optimization，and to improve the issues of agents falling into optima and setting reward function in reinforcement learning，a proximal policy optimization（PPO）algorithm based on maximum entropy correction and generative adversarial imitation learning（GAIL）is proposed. A maximum entropy correction term is introduced in the PPO framework. By optimizing policy entropy，the agents are encouraged to explore among multiple potential suboptimal policies，thereby comprehensively evaluating the environment and developing more better strategies. Meanwhile，to solve poor training performance caused by unreasonable reward function settings in the reinforcement learning，the GAIL idea is introduced to guide the gents to learn through expert data. Experimental results demonstrate that the PPO algorithm with the maximum entropy correction and GAIL is introduced to achieve remarkable performance in reinforcement learning，effectively improving learning speed and stability while avoiding performance degradation caused by unreasonable reward function settings in the environment. This algorithm provides a novel solution in the field of reinforcement learning，it is of great significance for dealing with challenging continuous control problems.

Keywords：reinforcement learning；PPO algorithm；GAIL；deep learning；maximum entropy learning

0 引言

随着人工智能和机器学习技术的飞速发展，强化学习作为一种重要的学习范式，已经在多个领域展现出其强大的潜力和应用价值。其中，近端策略优化（PPO，proximal policy optimization）算法作为近年来强化学习领域的一大突破[1]，以其高效稳定的性能成为了众多研究者关注的焦点。文献［2］通过自适应地调整信任区域内的剪切范围提高了信任区域内的探索能力，而且具有更好的性能范围。然而，面对复杂的实际问题和多样化的应用场景，PPO 算法也面临着一些挑战，例如探索与利用的平衡和奖励函数设置就是两个重要的问题。

强化学习方法让智能体通过与环境的交互来学习如何做出最优的决策。这一过程中，智能体主要有两个核心行为：探索未知的环境以发现潜在的更高奖励，以及利用已知的信息来最大化当前的收益。平衡好这两者之间的关系是强化学习算法达成良好性能的关键。为了解决这一难题，引入最大熵项作为一种有效的手段为强化学习中平衡探索与利用提供了新的思路。该方法最早由文献［3］提出，最大熵原理的核心思想是在给定约束条件下，选择不确定性最大的策略[4]。在强化学习中，这意味着智能体在做出决策时不仅要考虑当前已知的最

收稿日期:2024-07-22； **修回日期**:2024-09-05。
作者简介:王泽宁(2000-),男,硕士。
引用格式:王泽宁,刘 蕾.带最大熵修正和 GAIL 的 PPO 算法[J].计算机测量与控制,2025,33(1):235-241.

优解，还要考虑到未来可能出现的各种情况，从而保持一定的灵活性和适应性[5]。引入最大熵项对强化学习中平衡探索与利用的好处主要体现在以下几个方面：首先，通过增加策略的不确定性，智能体能够更加积极地探索未知的环境，从而发现更多潜在的高奖励路径。这有助于智能体避免陷入局部最优解，提高整体的性能表现[6]。其次，引入最大熵项还可以在一定程度上提高强化学习算法的收敛速度和稳定性。通过鼓励智能体在探索和利用之间找到一个合理的平衡点。同时，由于智能体在做出决策时会考虑更多的可能性，因此其决策过程也会更加稳定，不易受到环境噪声和随机性的影响[7]。文献［8］通过优化其中的经验回放部分进一步优化了效果。文献［9］通过将最大熵强化学习与进化策略相结合提出了一种简单而有效的方法来稳定学习。

另一方面，在强化学习中，奖励函数扮演着至关重要的角色，它作为智能体行为评价的核心标准，直接决定了智能体在交互过程中如何理解和评估其行为的优劣。然而，在现实世界的应用场景中，由于环境的复杂性和多变性，以及人类认知的局限性和主观性，我们往往难以设计出能够完全反映环境真实需求和期望的奖励函数。奖励函数在设计时很可能会遇到多方面的困难，例如环境可能包含大量的隐藏信息或未知因素，这些信息对于智能体的行为评价至关重要，但往往难以被直接观察和量化；而且人类的期望和需求往往是模糊和多变的，很难用单一的奖励函数来准确描述[10]。即使能够设计出看似合理的奖励函数，也可能因为环境的动态变化或智能体的行为多样性而导致奖励函数失效或产生误导。因此，如何有效改善奖励函数设置问题，成为了提升强化学习算法性能的关键之一。而在强化学习中，模仿学习（IL，imitation learning）作为一种重要的学习范式，为解决奖励函数设置问题提供了全新的视角[11]。模仿学习通过从专家行为数据中学习策略，为我们提供了一种更为直接和高效的学习途径。研究者们提出了生成式对抗模仿学习方法借鉴了生成式对抗网络（GANs，generative adversarial networks）的思想，将模仿学习问题转化为一个生成器和判别器之间的对抗游戏。通过引入生成式对抗网络的思想，生成式对抗模仿学习（GAIL，generative adversarial imitation learning）方法使得智能体能够直接从专家行为数据中学习策略，无需依赖于预定义的奖励函数。这大大简化了强化学习问题的建模过程，同时也提高了学习算法的稳定性和鲁棒性[12]。文献［13］引入第二个判别来训练策略，与指示演示数据的第一个判别器并行区分目标状态以提高策略的学习能力。文献［14］考虑了采样轨迹的不同时刻采取的行动之间可能存在的时间依赖性，提出了两种GAIL 变体并在自动驾驶领域取得了良好的效果。

为了在 PPO 算法框架中结合最大熵思想和模仿学习的优势，本文提出了一种新型的强化学习算法，即在 PPO 算法框架中引入最大熵项，并加入生成式对抗模仿学习的方法改善奖励函数设置问题。本文首先在第二节中详细阐述了如何在 PPO 算法中引入最大熵正则项。随后在第三节中研究了 GAIL 算法与 PPO 的集成方法，展示了如何通过生成对抗网络框架下的对抗学习，从专家数据中提取有用信息，以指导策略优化从而有效避免因奖励函数设置不当导致的性能瓶颈。最后，在本文的第四节中，展示了该算法在基准测试环境中的实验结果。实验结果表明，本文提出的算法在收敛速度、学习性能和泛化能力上达到了良好的效果，特别是在处理奖励函数设置不合理的问题上，展现出了强大的鲁棒性。

1 基础算法

1.1 PPO 算法

在强化学习领域，基于策略梯度的方法在采用深度神经网络作为策略函数时如果单次更新的步长过长，可能导致策略在单次迭代中发生显著恶化，进而对整体训练效率与稳定性产生不利影响[15]。为了有效缓解这一问题，近端策略优化算法被提出，其核心思想在于通过精确控制策略更新过程中的步长，来限制新旧策略之间的差异，确保学习的稳健性[16]。PPO 算法的设计初衷在于，通过与环境进行交互收集数据后，利用这些数据来优化一个精心设计的“替代”目标函数，以此逐步改进策略的表现。该算法引入了一种创新的目标函数形式，该函数内置了一个截断的比率项（Clipping Ratio），这一机制巧妙地限制了新策略相对于旧策略变化的幅度。通过这种方式，PPO 算法能够在每一步策略更新时，既追求代价函数的减少，又确保策略性能的稳步提升[17]，从而降低因更新步长过大导致的性能骤降风险[18]。

1.2 最大熵学习

最大熵强化学习（MERL，maximum entropy reinforcement learning）是一种先进的策略优化方法，其核心在于通过向传统的强化学习优化目标中引入熵正则化项，以鼓励智能体在策略学习过程中保持一定的探索性[19-20]。熵作为信息论中的一个核心概念，在此被用作衡量智能体行为分布随机性的度量标准。较高的熵值表明智能体在选择行动时表现出更大的不确定性或随机性，而较低的熵值则反映了行为选择的确定性增加。在 MERL 框架中，通过在优化目标函数中加入策略熵项以求在平衡智能体在探索未知状态－动作空间与利用当前已知最优策略之间的需求[21]。这种平衡是强化学习中的一个核心挑战，因为过度的探索可能导致学习效率低下，而过度的利用则可能使智能体陷入局部最优解。

MERL 通过调整熵项在优化目标中的权重，实现了这一平衡的动态调整，使得智能体能够在不同的学习阶段灵活切换其探索与利用的策略。从数学角度来看，MERL 的优化目标通常可以表示为最大化一个结合了期望累积奖励和策略熵的加权和。这种设计不仅促使智能体寻求高奖励的行为，还鼓励其探索那些能够增加行为多样性的策略，从而避免了过早收敛到单一的最优解。

研究表明，MERL 在提高强化学习算法的学习速度、稳定性和泛化能力方面展现出了显著的优势。通过保持一定的探索性，MERL 能够帮助智能体在面临复杂、非线性和不确定性的环境时，更好地应对变化，发现更优的解决策略。此外，MERL 还具有一定的鲁棒性，能够在一定程度上缓解由模型误差、噪声干扰或环境变化引起的不稳定问题。

1.3　GAIL

模仿学习是一种为解决强化学习中奖励函数设置问题而产生的替代性策略[22]，其核心思想在于利用专家智能体的示范行为作为学习信号，而非依赖环境提供的奖励[23]。在模仿学习的框架下，专家智能体通过与环境交互生成一系列状态－动作对，这些数据作为训练样本，被用于指导模仿者（即待训练的智能体）学习如何执行特定任务[24]。模仿学习的优势在于它绕开了奖励函数设计的难题，直接利用专家经验来优化智能体的行为策略，从而提高了学习的效率和可靠性。

在众多模仿学习算法中，生成对抗模仿学习 GAIL 以其独特的生成对抗网络结构脱颖而出。GAIL 算法将模仿学习问题转化为一个生成对抗过程，其中生成器（对应于待训练的智能体）试图生成与专家行为相似的状态－动作对，而判别器则负责区分这些生成的数据与专家提供的真实数据。通过生成器与判别器之间的不断对抗与优化，GAIL 能够逐渐提升生成器模仿专家行为的能力，从而在无需显式奖励信号的情况下，使智能体学会执行复杂任务。

GAIL 算法的优势在于其能够有效地处理复杂环境和高维状态空间中的模仿学习任务。通过引入生成对抗机制，GAIL 不仅能够在数据层面上逼近专家行为，还能在策略层面上捕捉到专家策略的深层特征，从而实现更加精确和鲁棒的模仿。此外，GAIL 还具有良好的泛化能力，能够在不同场景和任务中迁移专家知识，为智能体的快速适应和灵活应用提供了有力支持[25]。

2　最大熵项构造

在马尔科夫奖励过程中，从一个状态出发的未来累计奖励的期望被称为这个状态的价值，所有的状态价值组成状态价值函数，其中表示状态。

在马尔科夫奖励过程中，从一个状态出发的未来累计奖励的期望被称为这个状态的价值，所有的状态价值组成状态价值函数 $V(s)$，其中 s 表示状态。

根据策略提升定理，定义旧策略 π 下的状态价值函数为 V^{π}，若存在另一个策略 π'，其状态价值函数为 $V^{\pi'}$，在任意状态下都满足：

$$Q^{\pi}[s,\pi'(s)] \geqslant V^{\pi}(s) \tag{1}$$

则有：

$$V^{\pi'}(s) \geqslant V^{\pi}(s) \tag{2}$$

其中：Q 为动作价值函数，$Q(s,a)$ 表示在遵循策略 π 时，对当前状态 s 执行动作 a 得到的期望回报。

定义 $J(\theta)=E_{s_0}[V^{\pi\theta}(S_0)]$，$\theta$ 为策略的参数，S_0 表示初始状态。作为基于策略的方法，PPO 的目标就是找到：

$$\theta^{*}=\operatorname{argmax}_{\theta} J(\theta) \tag{3}$$

熵表示对一个随机变量的随机程度的度量。具体而言，如果 X 是一个随机变量，且它的概率密度函数为 p，那么它的熵 H 就被定义为：

$$H(X)=E_{x\sim p}[-\log p(x)] \tag{4}$$

结合最大熵思想，为使策略更加随机，在目标函数中除了状态价值函数，还要引入最大熵正则项：

$$\alpha H[\pi(\cdot \mid s_t)] \tag{5}$$

该项中的 α 是一个正则化系数，用来控制熵的重要程度。

如何选择熵正则项的系数非常重要。在不同的状态下需要不同大小的熵：在最优动作不确定的某个状态下，熵的取值应该大一点；而在某个最优动作比较确定的状态下，熵的取值可以小一点。为了自动调整熵正则项，最大熵算法将强化学习的目标改写为一个带约束的优化问题，定义损失函数为：

$$L(\alpha)=E_{s_t\sim R,a_t\sim\pi(\cdot\mid s_t)}[-\alpha\log\pi(a_t \mid s_t)-\alpha H_0] \tag{6}$$

当策略的熵低于目标值 H_0 时，训练目标 $L(\alpha)$ 会使 α 的值增大，进而在最小化策略的损失函数的过程中增加了策略熵对应项的重要性；而当策略的熵高于目标值 H_0 时，训练目标 $L(\alpha)$ 会使 α 的值减小，进而使得策略训练时更专注于价值提升。定义引入最大熵正则项的状态价值函数 V_S^{π} 为：

$$V_S^{\pi}(s)=V^{\pi}(s)+\alpha H[\pi(\cdot \mid s_t)] \tag{7}$$

假设当前策略 π_θ 的参数为 θ，为借助当前参数找到更优参数 θ' 使得 $J(\theta') \geqslant J(\theta)$，有如下推导：

设衰减因子为 γ，由定义可知 $J(\theta)$ 也可以写为：

$$J(\theta)=E_{s_0}[V_S^{\pi_\theta}(S_0)]=$$

$$E_{\pi_\theta}\Big[\sum_{t=0}^{\infty}\gamma^t V_S^{\pi_\theta}(S_t)-\sum_{t=1}^{\infty}\gamma^t V_S^{\pi_\theta}(S_t)\Big]=$$

$$-E_{\pi_\theta}\Big[\sum_{t=0}^{\infty}\gamma^t[\gamma V_S^{\pi_\theta}(S_{t+1})-V_S^{\pi_\theta}(S_t)]\Big] \tag{8}$$

基于以上等式，可以推导新旧策略目标函数之间的差距：

$$J(\theta')-J(\theta)=E_{s_0}[V_S^{\pi_{\theta'}}(S_0)]-E_{s_0}[V_S^{\pi_\theta}(S_0)]=$$

$$E_{\pi_{\theta'}}\left[\sum_{t=0}^{\infty}\gamma^t[r(s_t,a_t)+\gamma V_S^{\pi_\theta}(S_{t+1}-V_S^{\pi_\theta}(S_t)]\right] \quad (9)$$

其中：r（s，a）为状态 s 下采取动作 a 获得的回报。

定义优势函数 A 为：

$$A_S^{\pi}(s_t,a_t)=r(s_t,a_t)+\gamma V_S^{\pi_\theta}(S_{t+1})-V_S^{\pi_\theta}(S_t) \quad (10)$$

则上式可写为：

$$J(\theta')-J(\theta)=E_{\pi_{\theta'}}[\sum_{t=0}^{\infty}\gamma^t A_S^{\pi_\theta}(s_t,a_t)] \quad (11)$$

根据状态访问分布的定义：

$$\nu^{\pi}(s)=(1-\gamma)\sum_{t=0}^{\infty}\gamma^t P_t^{\pi}(s) \quad (12)$$

其中：P_t^{π}（s）为策略 π 使智能体在 t 时刻状态为 s 的概率。

式（11）也可以写成：

$$J(\theta')-J(\theta)=$$

$$\frac{1}{1-\gamma}E_{s\sim\nu\pi_{\theta'}}E_{a\sim\pi_{\theta'}(\cdot|s)}[A_S^{\pi_\theta}(s,a)] \quad (13)$$

由于当新旧策略接近时，状态访问分布变化很小，为简化计算，忽略两个策略之间的状态访问分布变化，使用旧策略的状态分布，并使用重要性采样对动作分布进行处理. 优化目标可以写为：

$$\max_{\theta}E_{s\sim\nu^{\pi_\theta}}E_{a\sim\pi_{\theta_k}(\cdot|s)}\left[\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}A_S^{\pi_{\theta_k}}(s,a)\right] \quad (14)$$

为限制步长，本文采用 PPO 截断的方式，即在目标函数中进行限制，确保新旧参数差距不会过大。设置参数，目标函数改为：

$$\operatorname{argmax}_{\theta}E_{s\sim\nu^{\pi_{\theta_k}}}E_{a\sim\pi_{\theta_k}}(\cdot|s)$$

$$\left[\min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)}A_S^{\pi_{\theta_k}}(s,a),\operatorname{clip}\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)},1-\varepsilon,1+\varepsilon\right)A_S^{\pi_{\theta_k}}(s,a)\right] \quad (15)$$

其中：clip（a，b，c）$=\max(\min(a,c),b)$，作用为把 a 限制在［b，c］内。参数 ε 表示截断的范围，如图 1 所示。

对于优势函数 A，使用广义优势估计（GAE，generalized advantage estimation）进行估计。记：

$$\delta_t=r_t+\gamma V(s_{t+1})-V(s_t) \quad (16)$$

有 $A_t^{\mathrm{GAE}}=\sum_{l=0}^{\infty}(\gamma\lambda)^l\delta_{t+l}$ 其中 λ 是 GAE 中的一个超参数，$\lambda=0$ 时仅看第一步差分得到的优势，$\lambda=1$ 时看每一步差分得到的优势的平均值。

3 引入 GAIL

GAIL 的核心理念在于通过模拟专家智能体的占用度量（occupancy measure）。为了达成这个目标，策略网络需要与环境进行动态交互，通过不断试错与调整，

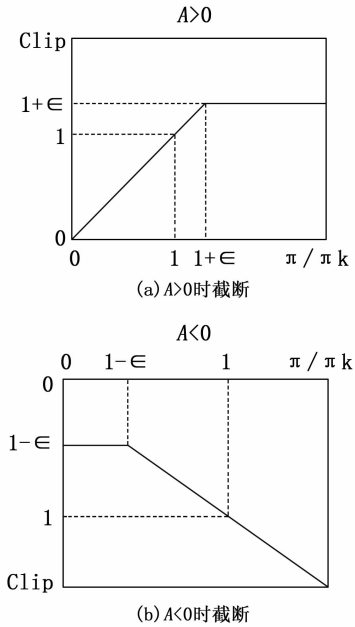

图 1 PPO 截断示意图

收集后续状态信息并据此作出合理的动作选择。在 GAIL 的架构中，策略网络扮演了生成式对抗网络中生成器（generator）的角色，其核心功能是根据当前环境状态生成相应的动作决策，给定任一环境状态，策略网络能够输出一个动作分布，从中采样即可得到该状态下应执行的动作。这一过程体现了策略网络对专家策略行为模式的学习与模仿。与此同时，GAIL 引入了一个高度专业化的判别器（discriminator），其职责是区分输入的状态－动作对是源自专家策略还是当前策略（即模仿者）。判别器接收状态－动作对作为输入，并输出一个介于 0 和 1 之间的标量值，该值代表了判别器将输入数据判定为非专家数据（即模仿者策略生成）的置信度。判别器的训练目标在于最小化将专家数据误判为模仿者数据的概率，同时最大化将模仿者数据正确识别的概率，即追求将专家数据的输出尽可能推向 0，而将模仿者数据的输出推向 1。

定义判别器的损失函数如下：

$$L(\phi)=$$

$$-E_{\rho_E}[\log D_\phi(s,a)]-E_{\rho_\pi}[\log(1-D_\phi(s,a))] \quad (17)$$

其中：ϕ 为判别器的参数。

基于此判别器 D，模仿者策略的目标就是其交互产生的轨迹能被判别器误认为专家轨迹。为达到这一目的，使强化学习采用判别器 D 的输出作为奖励函数来训练模仿者策略。若模仿者策略在环境中采样到状态 s，并且采取动作 a，此时该状态动作对（s，a）会输入到判别器 D 中，输出 D（s，a）的值，然后将奖励设置为 r，使用 r 替换原先强化学习算法中的奖励环节。通过使用专家数据对策略进行指导，在对抗不断进行后，

模仿者（即策略）生成的数据分布将接近真实的专家数据分布，达到训练的目的。

GAIL 具体的执行步骤如下。

1）收集专家经验：使用已经熟练掌握某个任务的专家策略，收集一批样本轨迹或动作的数据。

2）初始生成器和判别器。

3）训练判别器：在当前的生成器策略下，利用专家样本和生成器生成的样本进行训练，使判别器能够准确地区分生成器和专家的样本。

4）训练生成器：固定判别器，将生成器的策略更新为使得判别器无法区分生成器生成的样本和专家样本。

5）重复训练：反复迭代执行步骤 3）和步骤 4），直到生成器的策略接近专家样本的策略，同时判别器无法准确区分生成器和专家的样本。

至此，算法的整体架构如图 2 所示。

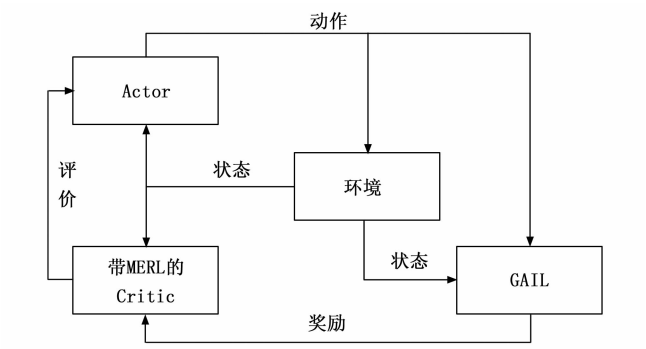

图 2　算法结构示意图

4　实验结果与分析

为验证本文提出的在 PPO 框架中引入最大熵项和模仿学习的有效性，设计仿真实验在 GYM 提供的车杆环境（如图 3 所示）下测试其性能[26]，并在使用相同超参数的情况下与其他强化学习的性能进行对比。

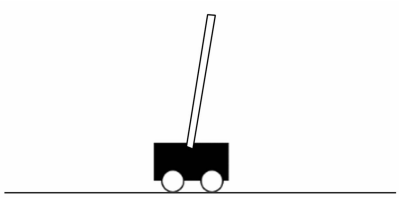

图 3　车杆环境示意图

考虑到模仿学习需要专家数据进行指导，本文在实验中预先训练好一个专家智能体，通过使用该专家智能体在同一环境中与环境交互进行采样得到专家数据，该数据由多个状态－动作对组成，表示专家智能体在该状态下会采取何种行为。

在车杆环境中，存在一个动态系统，其核心组件为一部小车，其上竖立一根杆。此任务要求智能体（Agent）通过执行离散动作（仅限于左右移动小车），以维持杆处于竖直状态，避免其倾斜度超过预设阈值。同时，还需控制小车在水平方向上的位移，防止其偏离初始位置过远。游戏终止条件设定为杆的倾斜度数超限、小车位置偏移过大或游戏持续达到预设的最大帧数（200 帧）。

智能体所处的状态空间被精确定义为一个四维连续向量，每一维度均承载了关于当前环境状态的连续信息；而智能体的动作空间则限制为离散且有限，具体包含两个可能的动作。环境的动作空间和状态空间如表 1 和表 2 所示。

表 1　车杆环境动作空间表

编号	动作
0	小车向左移动
1	小车向右移动

表 2　车杆环境状态空间表

维度	含义	最小值	最大值
0	小车位置	－2.4	2.4
1	小车速度	－Inf	Inf
2	杆角度	～－41.8°	～41.8°
3	杆顶端速度	－Inf	Inf

奖励机制为每成功维持一帧的稳定状态，智能体即获得 1 分的即时奖励，旨在鼓励智能体延长稳定维持的时间，从而达到累积最高分数的目标（即坚持至游戏结束时的 200 帧，可获得满额奖励）。

在测试中，策略网络学习率为 0.001，价值网络学习率为 0.01，未来回报的折扣因子 γ 为 0.98，GAE 中 λ 取 0.95，截断参数 ε 取 0.2，熵的正则化系数 α 的学习率为 0.01，目标熵为－1，训练轮数 500 轮。

PPO 采用 Actor－Critic 架构，本文中策略网络和价值网络均采用包含两个全连接层，一个隐藏层，激活函数为 ReLU 函数的神经网络，其中策略网络使用 softmax 函数将输出转换为概率分布。

首先是该方法与 PPO 方法的对比，结果如图 4 和图 5 所示。

图 4 和图 5 分别是改进前和改进后的总奖励变化曲线。横轴表示训练轮数，纵轴表示获得的回报。

可以看到，优化后的算法相较于原 PPO 算法，在收敛速度上保持高效，同时在稳定性方面展现出显著提升，波动明显减少。

如前所述，引入生成式对抗学习的目标是改善奖励函数设置不合理导致的性能损失，现对环境中的奖励函数做一定修改以模拟奖励函数设置不合理的情况（本例

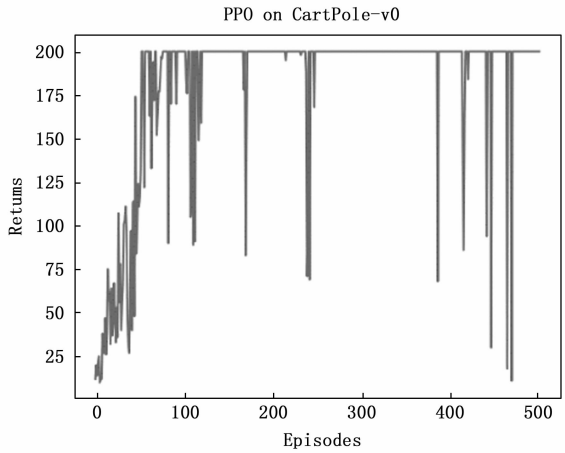

图 4 PPO 算法获得回报随轮次的变化

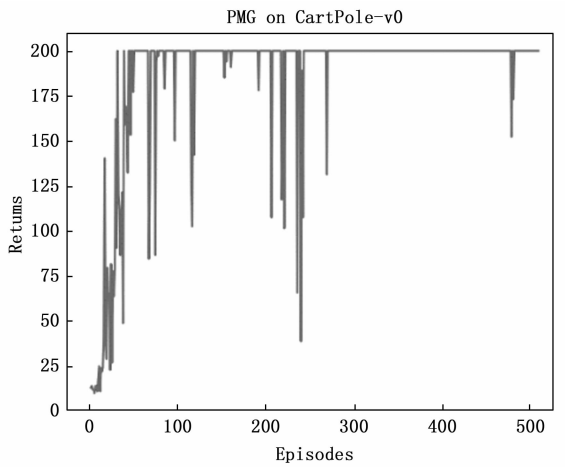

图 5 改进后的 PPO 算法获得回报随轮次的变化

中模拟奖励函数设置过大和过小的情况），其他参数不变。结果如图 6 至图 8 所示，可以看出，此时传统 PPO 算法的效果受到了明显负面影响，波动显著增加，而改进的 PPO 算法则没有影响。这是因为本文提出的算法不再使用环境中预设的奖励函数，而是根据 GAIL 中判别器的结果进行奖励。

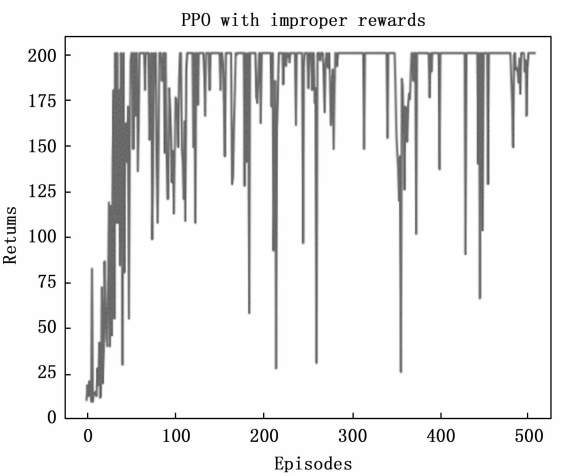

图 6 PPO 算法在奖励函数过大时的训练结果

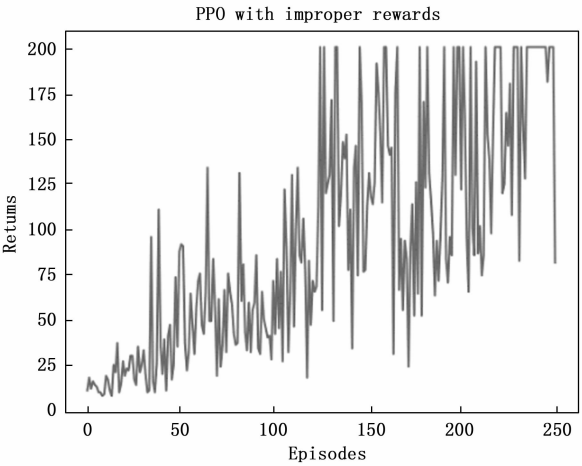

图 7 PPO 算法在奖励函数过小时的训练结果

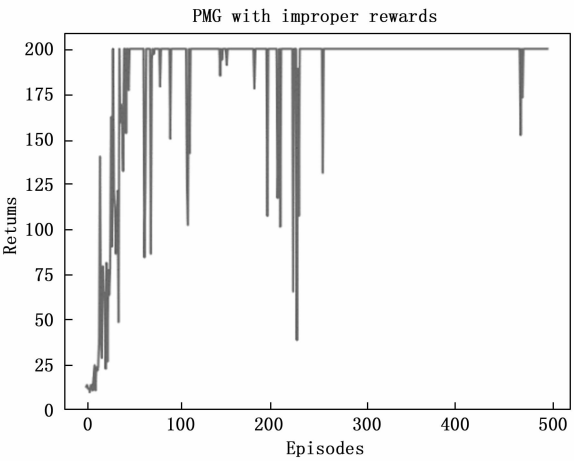

图 8 改进后的 PPO 算法在奖励函数不理想时的训练结果

5 结束语

在探索强化学习领域的应用中，本文提出了基于 PPO 框架，并融合最大熵思想与 GAIL 的强化学习算法。该算法通过引入最大熵正则项，不仅增强了策略的探索能力，使得智能体能够在面对不确定性时做出更为灵活多样的决策，还通过 GAIL 机制，利用专家数据来指导策略学习，从而在一定程度上缓解了奖励函数设计不当可能导致的训练偏差和性能瓶颈。

实验结果显示，该算法在基准测试任务中展现出了卓越的收敛速度和稳定的性能表现，证明了其在实际应用中的潜力和价值。特别地，通过 GAIL 的引入，算法有效降低了对精确奖励函数的依赖，使得智能体能够在缺乏完美定义奖励信号的环境中依然能够学习到高质量的行为策略。这一特性对于许多现实世界的复杂场景尤为关键，因为在实际应用中，设计一个全面且准确的奖励函数往往是一项极具挑战性的任务。

然而，本算法的一个显著缺点是训练效果对专家数据的质量和数量有着较高的依赖。这意味着，在没有足

够高质量专家数据支持的情况下，算法的性能可能会受到限制。因此，未来的研究方向可以聚焦于如何降低对专家数据的依赖，比如通过开发更加高效的数据增强技术、探索无监督学习或自监督学习的方法来提升智能体的自主学习能力，以及结合其他先进的机器学习技术来增强算法的鲁棒性和泛化能力。

参考文献：

[1] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal Policy Optimization Algorithms [J]. ArXiv Preprint ArXiv: 1707.06347, 2017.

[2] WANG Y, HE H, TAN X, et al. Trust region-guided proximal policy optimization [C] // Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 626 - 636.

[3] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor [C] // International conference on machine learning, PML-R, 2018: 1861 - 1870.

[4] HAARNOJA T, ZHOU A, HARTIKAINEN K, et al. Soft Actor-Critic Algorithms and Applicatio-ns [J]. ArXiv Preprint ArXiv: 1812.05905, 2018.

[5] HAARNOJA T, TANG H, ABBEEL P, et al. Reinforcement learning with deep energy-based policies [C] // International Conference on Machine Learning, PMLR, 2017: 1352 - 1361.

[6] SCHULMAN J, CHEN X, ABBEEL P. Equivalence Between Policy Gradients and Soft Q-Learning [J]. ArXiv Preprint ArXiv: 1704.06440, 2017.

[7] SYED U, BOWLING M, SCHAPIRE R E. Apprenticeship learning using linear programming [C] // Proceedings of the 25th International Conference on Machine Learning, 2008: 1032 - 1039.

[8] KABELA L. Experience Replay Methods in Soft Actor-Critic [D/OL]. University of Texas at Austin. 2023. https: //www. cs. utexas. edu/~ yukez/cs391r _ reports/files/Fall-2020/LK. pdf

[9] SHI L, LI S, ZHENG Q, et al. Maximum entropy reinforcement learning with evolution strategies [C] // 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020: 1 - 8.

[10] ABBEEL P, NG A Y. Apprenticeship learning via inverse reinforcement learning [C] // Proceedings of the twenty-first international conference on Machine learning. 2004: 1.

[11] HO J, ERMON S. Generative adversarial imitation learning [J]. Advances in Neural Information Processing Systems, 2016, 29: 4565 - 4573.

[12] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning [J]. ArXiv Preprint ArXiv: 1509.02971, 2015.

[13] TSURUMINE Y, MATSUBARA T. Goal-aware generative adversarial imitation learning from imperfect demonstration for robotic cloth manipulation [J]. Robotics and Autonomous Systems, 2022, 158: 104264.

[14] COUTO G C K, ANTONELO E A. Generative adversarial imitation learning for end-to-end autonomous driving on urban environments [C] // 2021 IEEE Symposium Series on Computational Intelligence (SSCI) . IEEE, 2021: 1 - 7.

[15] SCHULMAN J, LEVINE S, ABBEEL P, et al. Trust region policy optimization [C] // International Conference on Machine Learning, PMLR, 2015: 1889 - 1897.

[16] KAKADE S M. A natural policy gradient [C] // Advances in Neural Information Processing Systems 14, 2001: 1531 - 1538.

[17] SCHULMAN J, MORITA P, LEVINE S, et al. High-dimensional continuous control using generalized advantage estimation [J]. ArXiv Preprint ArXiv: 1506.02438, 2015.

[18] ZHU H, GUPTA A, RAJESWARAN A, et al. Dexterous manipulation with deep reinforcement learning: Efficient, general, and low-cost [C] // 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 3651 - 3657.

[19] ZIEBART, MAAS, BAGNELL, et al. Maximum entropy inverse reinforcement learning [C] //AAAI, 2008: 1433 - 1438.

[20] WANG Z, BAPST V, HEESS N, et al. Sample efficient actor-critic with experience replay [J]. ArXiv Preprint ArXiv: 1611.01224, 2016.

[21] VOLODYMYR M, KORAY K, DAVID S, et al. Human-level control through deep reinforce-ment learning [J]. Nature, 2015, 518 (7540) : 529 - 533.

[22] RATLIFF D N, SILVER D, BAGNELL A J. Learning to search: Functional gradient techniques for imitation learning [J]. Auton. Robots, 2009, 27 (1): 25 - 53.

[23] ROSS S, BAGNELL D. Efficient reductions for imitation learning [C] // Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, 2010: 661 - 668.

[24] ZHANG Q, WU C, TIAN H, et al. Safety reinforcement learning control via transfer learning [J]. Automatica, 2024, 166: 111714.

[25] LI J, HUANG S, XU X, et al. Generative Adversarial Imitation Learning from Human Behavior with Reward Shaping [C] // 2022 34th Chinese Control and Decision Conference (CCDC). IEEE, 2022: 6254 - 6259.

[26] BROCKMAN G, CHEUNG V, PETTERSSON L, et al. Openai gym [J]. ArXiv Preprint ArXiv: 1606.01540, 2016.