

不确定度增强下基于模型的探索-学习策略 联合优化*

肖士湘¹, 黄文振², 焦建彬^{1†}

(1 中国科学院大学电子电气与通信工程学院, 北京 100049; 2 清华大学信息科学技术学院, 北京 100084)

(2023 年 12 月 4 日收稿; 2024 年 9 月 5 日收修稿稿)

Xiao S X, Huang W Z, Jiao J B. Model-based exploration-learning joint optimization via uncertainty augmentation [J]. Journal of University of Chinese Academy of Sciences, 2026, 43(5): 667-676. DOI: 10. 7523/j.ucas. 2024. 072.

摘要 在现有基于模型的强化学习方法中, 智能体采用同一策略进行与真实环境 and 环境模型的交互, 难以兼顾探索环境的效率和策略更新的稳定性。为解决此问题, 提出一种不确定度增强下基于模型的探索-学习策略联合优化方法。该方法联合优化 2 个策略, 分别是用于与真实环境交互的探索策略和与环境模型交互的学习策略。在优化探索策略的过程中, 引入基于模型不确定度的内在激励以提升探索真实环境的效率, 同时在优化学习策略时将模型不确定度作为约束条件, 保证学习策略优化的稳定性。多个连续控制任务的实验结果表明, 该方法相比于现有的代表性方法在渐进性能和样本效率上有明显的优势。

关键词 深度强化学习; 基于模型的强化学习; 样本效率; 强化学习探索; 不确定度; 模型误差
中图分类号: TP183 **文献标志码**: A **DOI**: 10. 7523/j.ucas. 2024. 072

Model-based exploration-learning joint optimization via uncertainty augmentation

XIAO Shixiang¹, HUANG Wenzhen², JIAO Jianbin¹

(1 School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China; 2 School of Information Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract In existing model-based reinforcement learning methods, a single policy is adopted to interact with the real environment and the environment model, which makes it hard for the agent to balance the efficiency of exploring the environment and the stability of policy updating. To address this issue, this paper proposes a model-based explorer-learner joint optimization via uncertainty augmentation method (MELO-UA). MELO-UA simultaneously optimizes a pair of policies, namely the explorer policy for interacting with the real environment, and the learner policy for interacting with the environment model. During the optimization of the explorer, an implicit bonus based on model uncertainty is introduced to enhance the efficiency of exploring the real environment. At the

* 中国科学院先导科技专项(XDA27010300)资助

† 通信作者, E-mail: jiaojb@ucas.ac.cn

same time, during the optimization of the learner, the model uncertainty is used as a constraint to ensure the stability of the policy optimization. Experimental results on multiple continuous control tasks show that the proposed method has significant advantages in both asymptotic performance and sample efficiency compared to state-of-the-art methods.

Keywords deep reinforcement learning; model-based reinforcement learning; sample efficiency; exploration in reinforcement learning; uncertainty; model errors

强化学习是一种解决序列决策问题的机器学习范式,其核心思想是让智能体在环境中不断试错来优化智能体策略。强化学习已经在许多复杂序列决策任务中取得了超越人类的表现,例如围棋^[1-2]、高维复杂控制^[3]、电子游戏等^[4]。这些研究工作所使用的强化学习算法为无模型强化学习:智能体直接从交互过程中进行策略的学习,而不直接构建额外的环境模型。这类算法的样本效率较为低下,需要智能体与环境进行百万次量级的交互才能获得不错的性能。高昂的交互开销使得无模型算法多在仿真环境中实现,难以迁移到真实环境。

基于模型的强化学习被认为是一类能有效提升样本效率的方法,这类方法利用智能体与环境的交互,同时进行环境模型的学习和策略优化。基于模型的方法相对于无模型方法具有明显的样本效率优势^[5],使其更有可能应用于真实场景。然而,由于基于模型的方法在策略优化过程中会利用环境模型进行虚拟交互,这使得环境模型的准确性变得十分重要。一个有偏的环境模型会输出与真实环境分布不符的状态预测,误导智能体选择价值虚高的动作,即某些状态下该动作在有偏环境模型中有较高价值,但是在真实环境下价值较低。有偏模型对策略的误导会使得策略在模型被训练得完全准确之前表现不佳,或是使策略陷入全局次优甚至不收敛,从而限制了基于模型方法的性能和样本效率。

现有的方法主要从2个角度来降低模型误差的影响:在模型学习中,学习一个更准确的模型^[6-7];或是在模型利用时,减少对模型误差的错误利用^[8-9]。第1类方法无法完全避免环境模型在虚拟交互中产生误差,这是因为环境模型是以历史真实状态-动作对作为训练样本进行监督学习获得的,尽管环境模型在历史样本区域上具有较高的精度,但在区域外却有较高的分布外误差;同时,在强化学习过程中,智能体会在交互时不断探索新的区域,即智能体在与环境模型交互时,会

不断访问环境模型的高误差区域,从而在虚拟交互时产生有偏样本。第2类方法则利用模型不确定度的估计等方法来判断虚拟样本是否可信,降低置信度较低的虚拟样本在策略优化中的权重,并通过保守策略优化的方式降低模型误差对策略优化的影响。但是,这类方法往往会导致智能体在真实交互时也较为保守,使其在与真实环境交互时的探索效率下降,影响智能体的样本效率。

以上分析表明,在基于模型的强化学习中,智能体需要在真实环境中进行有效的分布外探索的同时,避免在与环境模型交互时访问分布外状态-动作。但是,现有方法中,使用同一策略分别进行与真实环境和环境模型的交互难以实现上述需求。为此,我们认为可以使用一对不同的策略,分别与真实环境和环境模型进行交互。根据这一思想,提出一种不确定度增强下基于模型的探索-学习策略联合优化方法(model-based explorer-learner joint optimization via uncertainty augmentation, MELO-UA)。算法的核心思想是训练一组探索策略-学习策略,探索策略负责探索真实环境,学习策略负责利用模型进行优化。在优化2个策略时,MELO-UA分别采用2种不同的基于不确定度的数据增强方法:对于探索策略,将不确定度作为内在激励,鼓励探索策略探索环境;对于学习策略,利用不确定度对模拟数据进行筛选,减少有偏数据对策略优化的影响。最后通过实验,验证了MELO-UA在样本效率和性能上相对于现有代表性算法均有较大优势,且该性能优势在高状态-动作复杂度和探索难度较大的任务上更为明显。

1 相关工作

1.1 基于模型的强化学习

基于模型的方法交替地进行两阶段学习:1)模型学习:智能体与真实环境进行交互,获取真实样本,并使用这些真实样本以监督学习的方式训练环境状态转移模型(下简称环境模型);2)模型利

用:智能体与环境模型进行大量虚拟交互,利用虚拟样本来优化策略。环境模型的构建使真实交互样本中的环境信息被编码,这些环境信息会在智能体与环境模型大量交互的过程中被反复利用,提升真实样本的利用效率。从模型利用的角度,基于模型的强化学习可分为4类:1)基于模型的样本生成,例如 Dyna^[10]、基于模型的策略优化(model-based policy optimization, MBPO)^[5]等,这类方法利用环境模型与智能体进行模拟交互,生成大量交互样本作为无模型强化学习方法训练的额外数据,也称类 Dyna 方法。这类方法容易在当前先进的无模型基础算法上扩展,且对性能有良好的理论保证^[11]。2)基于模型的值函数扩展,例如基于模型的价值扩展(model-based value expansion, MVE)^[9]、随机集成价值扩展(stochastic ensemble value expansion, STEVE)^[12]等,这些算法利用环境模型生成更精确的值函数估计,应用于无模型值迭代算法中,这类算法渐进性能较对应的无模型算法差距较小。3)基于模型的梯度生成,例如学习控制的概率推理(probabilistic inference for learning control, PILCO)^[13]、随机价值梯度(stochastic value gradient, SVG)^[14]、Dreamer^[15-16]等,这些算法利用可微环境模型进行白盒模拟,传递策略梯度,并利用梯度下降直接对环境模型和策略模型进行联合优化。这类算法在一些简单任务上有极高的样本效率,但是容易在长序列和复杂环境中出现梯度的消失和爆炸,导致算法不收敛。4)基于模型的动作序列规划,例如概率集成轨迹采样法(probabilistic ensembles with trajectory sampling, PETS)^[8]、深度规划网络(deep planning network, PLaNet)^[17]、MuZero^[1]等,这类方法利用环境模型结合模型预测控制或蒙特卡洛树搜索等规划算法,在每个时间步规划动作序列并执行。这类方法在决策复杂度较高的任务上有较好的性能,但是样本效率大多不如类 Dyna 算法。本文提出的探索策略-学习策略优化框架是对类 Dyna 算法中交互方式的改进,可作为现有 Dyna 算法的拓展模块。

1.2 基于不确定度的模型利用

模型误差会导致产生的模拟交互样本与真实样本有差异,影响策略在真实环境中的性能。模型误差主要有2个来源:1)分布内误差:由于模型容量不足,刻画复杂环境时存在泛化误差;2)分布

外误差:随着策略不断优化,其在环境中交互的动作-状态空间分布会发生偏移,这一分布偏移会使得环境模型需要预测训练分布外的状态转移,从而产生分布外误差。

当前研究中,为降低模型误差的影响,思路可分为2种:1)在模型学习过程中学习一个更准确的模型,例如选用表达能力更强的模型结构如深度神经网络^[6-7],或是构建随机性模型^[8]、集成模型^[8,18]等;2)在策略学习过程中估计模型误差,根据估计调整策略优化过程,避免策略优化时对有偏模型的错误利用。

在深度学习中,对模型误差的估计通常会转变为对预测不确定度的估计。预测不确定度可分为2类:统计不确定度(又作感知不确定度)和参数不确定度(又作认知不确定度)^[19-20]。其中统计不确定度是指数据样本本身的统计随机性造成的不确定度,来源于数据样本的噪声扰动和环境的随机性,这一不确定度也与认知偏差相关^[21];参数不确定度是给定模型在学习过程中参数数值的不确定度,主要来源于训练样本的不充分,这一不确定度与模型误差有一定相关性^[22]。为评估模拟样本的统计不确定度,通常会使用随机性神经网络作为环境模型^[8];在评估模拟样本的参数不确定度时,则会使用贝叶斯推断、集成神经网络等方法,其中基于集成神经网络的方法容易在深度神经网络模型上部署,是当前的主流环境模型结构。在利用上述结构进行环境模拟时,通过随机采样等方式得到的模拟数据样本中会隐式地包含对不确定度的估计,使得模拟交互与真实交互更加贴近,降低了策略优化过程中的方差,例如 PETS 和 MBPO 等。此外,近期的一些研究会利用启发式算法^[23]计算集成模型间偏差以直接量化不确定度,再利用不确定度对模拟数据进行筛选或重加权,例如掩码式基于模型的演员-评论家算法(masked model-based actor-critic, M2AC)^[24]、基于模型的离线策略优化(model-based offline policy optimization, MOPO)^[25]、STEVE^[12]和 Curious iLQR^[26]等方法,其中 M2AC 在类 Dyna 框架下滤除高不确定度的模拟样本,MOPO 则针对离线强化学习提出,避免产生分布外样本,STEVE 依据不确定度控制值函数扩展的步长,Curious iLQR 则利用不确定度激励环境探索。

1.3 基于不确定度的探索

在强化学习探索^[27]的研究中,面向不确定度的乐观探索^[28-30]是一种常见的方法。其核心思想是基于不确定度设计无监督内在奖励,鼓励智能体访问之前较少被访问的状态,也被称为新颖状态。乐观演员-评论家算法(optimistic actor-critic, OAC)^[31]利用Q函数的自举过程估计参数不确定度,以激励柔性演员-评论家算法(soft actor-critic, SAC)^[32]算法的探索;随机网络蒸馏(random network distillation, RND)^[33]则通过随机网络蒸馏获取不确定度。乐观值分布探索器(optimistic value distribution explorer, OVD-explorer)^[34]和动力学瓶颈(dynamics bottleneck)^[35]方法关注于在进行乐观探索的同时,估计统计不确定度,以消除环境白噪声的影响。

利用环境模型估计参数不确定度,从而进行乐观探索是一种常见的方法^[36-39],内在奖励演员-评论家算法(intrinsic actor-critic, IAC)^[36]等方法利用前向模型直接预测不确定度作为内在奖励,而内在好奇心模块(intrinsic curiosity module, ICM)^[37]采用更加稳定的反向模型进行预测,分歧驱动探索(exploration via disagreement, EvD)^[38]则利用集成模型间的不一致性作为内在奖励。本文方法参照面向不确定度的乐观探索思想,采用与EvD类似的方法,通过衡量集成模型间的不一致性并将其构建为内在激励,鼓励探索策略在真实环境中的探索,以针对性地提升环境模型的性能,减少环境模型对分布外样本的预测偏差。

2 背景知识

2.1 问题定义

考虑马尔可夫决策过程(Markov decision process, MDP),定义为六元组 $\mathbf{M}=\{\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho_0\}$,其中 $\mathcal{S}$ 和 $\mathcal{A}$ 分别为环境状态集合和智能体动作集合;环境的状态转移过程被记为条件密度函数 $P(s'|s, a): \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, +\infty)$; $r: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [-R, R]$ 为环境的有界奖励函数; $\gamma \in [0, 1]$ 为折扣因子; ρ_0 为初始状态分布。智能体学习一个策略 $\pi_\theta(\cdot|s)$,在每个时间步 t 根据观测状态 s_t 依据策略输出的动作分布来采样动作 a_t 与环境进行交互,再从环境中获得奖励 r_t 和下一时间步的状态 s_{t+1} 。强化学习的目标便是优化策略函数以最大化MDP中的累积折损奖励估计 $J(\pi, P)$,记为

$$J(\pi, P) = \mathbb{E}_{s_0 \sim \rho_0, s' \sim P(s, a), a \sim \pi(s)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]. \quad (1)$$

在基于模型的强化学习中,我们会通过监督学习,利用智能体在真实环境中交互收集的数据,训练出环境模型 $P_\psi(s'|s, a)$ 。

2.2 基础算法

本文将MBPO^[5]作为基础算法。MBPO是一种类Dyna算法,轮流执行以下3个步骤:模型学习、模型利用和策略优化。MBPO的整体算法如下。

算法1 MBPO

初始化: SAC策略 $\{\pi_\theta, Q_\phi\}$, 环境模型 $\{P_\psi^{(i)}\}_{i=1}^N$, 环境回放池 D_{env} , 模型回放池 D_{model}

```

1 for  $N$  epochs do
2   利用 $D_{\text{env}}$ 训练环境模型 $\{P_\psi^{(i)}\}_{i=1}^N$ 
3   for  $E$  steps do
4     在真实环境中根据 $\pi_\theta$ 进行交互,放入 $D_{\text{env}}$ 
5     for  $M$  model rollouts do
6       从 $D_{\text{env}}$ 均匀采样 $s_t$ 
7       以 $s_t$ 为起点,利用 $\{P_\psi^{(i)}\}_{i=1}^N$ 和 $\pi_\theta$ 进行 $k$ 步模拟交互,
          放入 $D_{\text{model}}$ 
8     end
9     for  $G$  gradient updates do
10      利用 $D_{\text{model}}$ 对 $\{\pi_\theta, Q_\phi\}$ 进行SAC更新
11    end
12  end
13 end

```

1) 模型学习 MBPO学习了一组集成概率性神经网络 $\mathcal{P}_\psi = \{P_\psi^{(1)}, \dots, P_\psi^{(N)}\}$ 以建模环境,其中每个概率性神经网络 $P_\psi^{(i)}$ 为网络结构完全一致的高斯随机性神经网络,接收状态动作对 (s, a) 作为输入,输出对角高斯分布 $P_\psi^{(i)}(s'|s, a) = \mathcal{N}(\mu_\psi^{(i)}(s, a), \Sigma_\psi^{(i)}(s, a))$,其中 $\mu_\psi^{(i)}(s, a)$ 是与 s 相同维度的均值向量, $\Sigma_\psi^{(i)}(s, a)$ 为对角协方差矩阵。每个神经网络将负对数似然度作为损失函数进行监督学习,方差项用于刻画环境的统计不确定性,随机神经网络中的方差项蕴含统计不确定度信息。在训练集成模型时,MBPO独立地对每个模型进行随机参数初始化和数据样本采样使得模型参数产生差异性,从而获得参数不确定度。

2) 模型利用 在利用模型时,MBPO采用平均采样,随机地从集成神经网络中抽取模型与策略进行交互,生成大量模拟样本。通过平均采样机制和高斯神经网络结构,模拟样本蕴含了2种

不确定度信息。利用这些样本进行策略优化,会在统计角度上减少策略对模型误差的错误利用,使得策略具备对2种不确定性的鲁棒性。同时,为降低多步模拟时随步长不断增加的累积误差,MBPO采用大批次小步长的模拟策略,在保证模拟样本数据量的同时降低数据误差。

3)策略优化 MBPO将SAC^[32]作为无模型基础算法。SAC基于最大熵强化学习框架^[40]优化演员-评论家策略,在许多单智能体决策场景下有优异性能。SAC包含1对状态-价值函数(下简称 Q 函数) Q_{ϕ_1}, Q_{ϕ_2} 、1对目标函数 $\bar{Q}_{\phi_1}, \bar{Q}_{\phi_2}$ 和1个策略函数 π_{θ} 。

在策略更新时,先从经验池 D 中抽取小批次转移样本 $B = \{(s, a, r, s', d)\}$,其中 d 为回合结束信号,当回合结束时为1,否则为0。为更新 Q 函数,计算软- Q 目标,如以下公式

$$y(r, s', d) = r + \gamma(1 - d) \left(\min_{i=1,2} \bar{Q}_{\phi_i}(s', a') - \alpha \log \pi_{\theta}(a'|s') \right), \quad (2)$$

$$a' \sim \pi_{\theta}(\cdot|s'),$$

其中, α 为控制策略随机性的参数。接着用软- Q 目标来计算 Q 函数损失,如公式(2);再使用梯度下降进行 Q 函数参数更新。最后利用软策略损失迭代更新策略函数,如下式

$$\mathcal{L}(\pi_{\theta}) = \frac{1}{|B|} \sum_B \left(\min_{i=1,2} Q_{\phi_i}(s, a') - \alpha \log \pi_{\theta}(a|s) \right), \quad (3)$$

$$a' \sim \pi_{\theta}(\cdot|s).$$

3 MELO-UA 算法

MELO-UA 算法的模型结构如图1所示。MELO-UA 优化一对结构相同的演员-评论家策略,分别记为探索策略 π_{θ_E} 和学习策略 π_{θ_L} 。与其他类Dyna算法一样,MELO-UA 在每个循环周期依次进行模型学习、模型利用和策略优化,结合探索策略-学习策略架构和基于不确定性的数据增强,算法的流程概括如下:1)探索策略与环境交互,收集真实交互样本,再利用真实样本进行环境模型的监督学习;2)学习策略与环境模型交互,生成大量模拟样本;3)基于不确定性估计对模拟样本进行不同形式的数据增强,并用于探索策略和学习策略的联合优化。在本方法中,步骤1)和2)与MBPO相似,现就步骤3)中的数据增强方法和策略联合优化的方式进行详细说明。

3.1 基于模型不确定度的数据增强

引入探索策略-学习策略后,“与真实环境交互”和“与环境模型交互”由探索策略和学习策略分别执行,使得智能体可以同时进行基于不确定度的探索和基于不确定度的模型利用。本文使用了2种目的相反的基于不确定度的数据增强方式:基于不确定度的内在奖励和基于不确定度的数据筛选,分别用于构建探索策略与学习策略训练时所需的模拟样本。

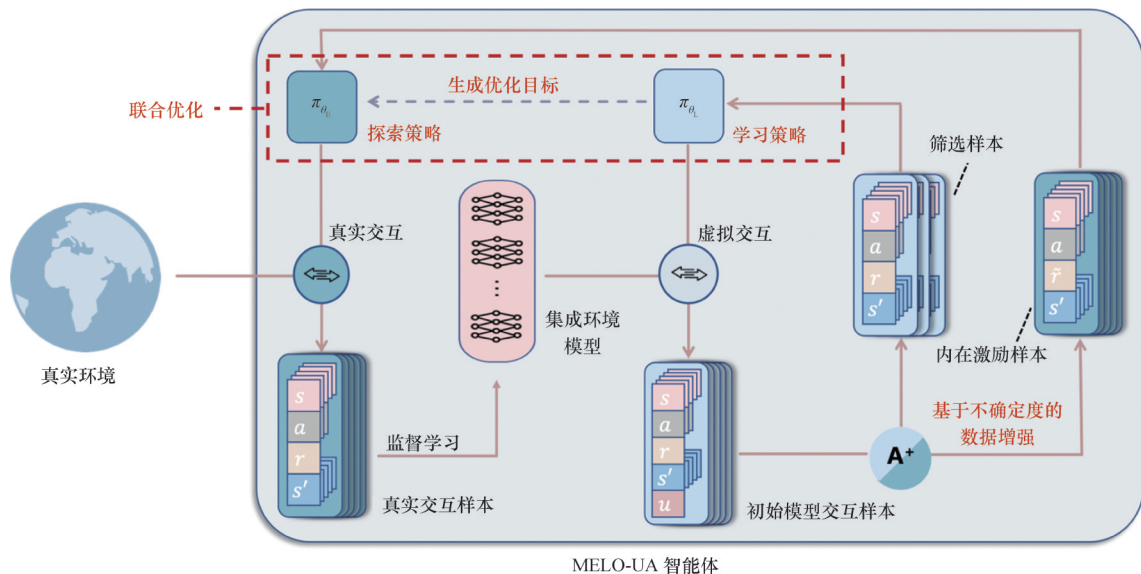


图1 MELO-UA 模型结构

Fig. 1 Model structure of MELO-UA

$$\begin{aligned} \text{Evar}(s, a) = \Sigma^*(s, a) \\ \frac{1}{N} \sum_{i=1}^N [\Sigma_{\psi}^{(i)}(s, a) + (\mu_{\psi}^{(i)}(s, a))(\mu_{\psi}^{(i)}(s, a))^T] - \\ \left(\frac{1}{N} \sum_{i=1}^N \mu_{\psi}^{(i)}(s, a) \right) \left(\frac{1}{N} \sum_{i=1}^N \mu_{\psi}^{(i)}(s, a) \right)^T. \quad (4) \end{aligned}$$

采用集成方差^[19]和集成标准差作为不确定性的量化指标。对于输入 (s, a) ,集成方差Evar定义为公式(4),集成标准差Estd定义为集成方差的平方根 $\sqrt{\text{Evar}}$ 。

MELO-UA 在利用集成网络环境模型生成模拟交互样本时,学习策略会与模型进行 $H(H \geq 1)$ 步交互。在第1步交互时,从真实环境经验池中抽取 B_r 个状态 $\{s_1^{(j)}, j = 1, \dots, B_r\}$,其中 B_r 为模拟交互的 batch size。然后利用学习策略 π_{θ_L} 对每个状态生成动作 $a_1^{(j)}$,再由集成环境模型经过平均采样生成 B_r 组转移样本 $\{(s_1^{(j)}, a_1^{(j)}, r_1^{(j)}, s_2^{(j)}, d_1^{(j)}), j = 1, \dots, B_r\}$ 和对应的不确定度估计值 $u_i^{(j)} = U(s_1^{(j)}, a_1^{(j)})$,其中 $U(s, a)$ 为任意不确定度量化函数。在后续第 $i(1 < i \leq H)$ 步的模拟交互时,选用由前一步模拟预测得到的当前步状态作为输入状态,进行新一轮模拟。经过 H 轮迭代,一共可以生成 $B_r \times H$ 组带有不确定度量化的模拟样本。

采用如下2种不同数据增强方法,分别对模拟样本进行数据增强,并利用不同的增强数据进行探索策略和学习策略的训练:

1)基于不确定度的内在奖励 为了鼓励探索策略进行基于不确定度估计的乐观探索,在将样本放入探索策略的经验池时,MELO-UA将前向多步不确定度作为转移对的内在激励,将内在激励与原始奖励叠加后作为总奖励,如以下公式

$$\begin{aligned} \tilde{r} &= r + \eta b, \\ b &= \sum_{i=1}^{K_b} \gamma^{k-1} u_i, \\ u_i &= U(s_i, a_i), \end{aligned} \quad (5)$$

其中: γ 为MDP六元组中的折扣因子, η 为激励系数, K_b 为前向视野因子。考虑到当前步不确定度较高的模拟样本往往具有较高的预测偏差,若直接对其给予较高的激励,会对探索策略的优化带来较大的不稳定性。因此,将内在激励 b 构建为多步不确定度的加权和。当 $K_b = 1$ 时,激励退化为单步不确定度。本文在实验中针对不同的 K_b 对模型性能的影响进行了测试。

2)基于不确定度的数据筛选 为避免学习策

略优化受有误差的数据样本影响,采用与M2AC类似的筛选机制。根据不确定度对每步模拟交互生成的 B_r 组样本进行排序,只保留其中不确定度较低的 wB_r 组样本用于学习策略的策略优化,其中 $w \in (0, 1]$ 为筛选率。

3.2 探索策略-学习策略联合优化

通过利用2种不同的方式进行数据增强,探索策略和学习策略会产生一定差异性。这一差异性使得智能体在与真实环境交互时,更容易访问探索不充分的状态,但也有可能出现冗余探索,即过多地探索低奖励区域,而不去探索学习策略价值函数较高的区域。为此,需要对探索策略和学习策略的策略进行联合优化,而非独立优化,从而在保留差异性的同时,让探索策略能够利用学习策略中的高价值状态-动作区域信息。

探索策略-学习策略联合优化的核心思想是利用学习策略的价值函数引导探索策略的价值函数学习。在探索策略-学习策略的策略优化过程中,学习策略的优化过程与SAC算法一致。不同点在于,在探索策略的 Q 函数更新过程中,其目标函数 y_E 引入了学习策略的 Q 值项,如以下公式

$$\begin{aligned} y_E &= \zeta y_{\text{Learner}} + (1 - \zeta) y_{\text{Bellman}}, \\ y_{\text{Learner}} &= \min_{i=1,2} Q_{\phi_{L_i}}(s, a), \\ y_{\text{Bellman}} &= r + \gamma(1 - d) \left(\min_{i=1,2} \bar{Q}_{\phi_{E_i}}(s, \tilde{a}) - \right. \\ &\quad \left. \alpha \log \pi_{\theta}(a'|s) \right), \end{aligned} \quad (6)$$

其中

$$\begin{aligned} \zeta &= \frac{\delta(Q_{\phi_L}, Q_{\phi_E})}{\delta(Q_{\phi_L}, Q_{\phi_E}) + \delta(Q_{\phi_{L_1}}, Q_{\phi_{L_2}}) + \delta(Q_{\phi_{E_1}}, Q_{\phi_{E_2}})}, \\ Q_{\phi} &= \min_{i=1,2} Q_{\phi}(s, a), \\ \delta(a, b) &= \frac{|a - b|}{|a + b|}. \end{aligned} \quad (7)$$

由探索策略 Q 函数自身归一化差值、学习策略 Q 函数自身归一化差值以及探索-学习 Q 函数间归一化差值决定: Q 函数自身差值越小,其对价值的估计就越准确,则智能体在此状态-动作区域的探索激励越小,因此包含内在激励项的 y_{Bellman} 所占的权重也越小,需要较大的 ζ ;当2个策略间的 Q 函数差值越大,说明该状态-动作其余的价值探索不够准确,需要智能体进一步探索, y_{Bellman} 所占权重也越大,所需 ζ 也较小。联合优化使得探索策略-学习策略的差异性受到了约束,减少了冗余探

索的概率。

MELO-UA的完整算法伪代码如算法2所示。

算法2 MELO-UA

```

初始化: 探索者策略  $\{\pi_{\theta_E}, Q_{\phi_E}\}$ , 学习者策略  $\{\pi_{\theta_L}, Q_{\phi_L}\}$  环境
模型  $\{P_{\psi}^{(i)}\}_{i=1}^N$ , 环境回放池  $D_{\text{env}}$ , 探索者回放池
 $D_{\text{exp}}$ , 学习者回放池  $D_{\text{learn}}$ 
1 for  $N$  epochs do
2   利用  $D_{\text{env}}$  训练环境模型  $\{P_{\psi}^{(i)}\}_{i=1}^N$ 
3   for  $E$  steps do
4     在真实环境中根据  $\pi_{\theta_E}$  进行交互, 放入  $D_{\text{env}}$ 
5     for  $M$  model rollouts do
6        $\mathcal{D}_{\text{cache}} \leftarrow \emptyset$ 
7       从  $D_{\text{env}}$  均匀采样  $s_t$ 
8       以  $s_t$  为起点, 利用  $\{P_{\psi}^{(i)}\}_{i=1}^N$  和  $\pi_{\theta_L}$  进行  $k$  步引入不确
          定度量化的模拟交互, 放入  $D_{\text{cache}}$ 
9       参照公式(5), 利用不确定度对  $D_{\text{cache}}$  进行数据增
          强, 放入  $D_{\text{exp}}$ 
10      按不确定度筛选  $D_{\text{cache}}$  中不确定度较低样本, 放入  $D_{\text{learn}}$ 
11    end
12    for  $G$  gradient updates do
13      参照公式(7), 利用  $D_{\text{exp}}$  和  $D_{\text{learn}}$  对  $\{\pi_{\theta_E}, Q_{\phi_E}\}$ 、
           $\{\pi_{\theta_L}, Q_{\phi_L}\}$  进行探索者-学习者联合更新
14    end
15  end
16 end

```

4 实验设计与结果分析

本文实验验证以下问题: 1) 在典型的强化学习验证环境下, 验证 MELO-UA 相比当前代表性基于模型的方法和无模型方法, 在最终性能和样本效率上有优势; 2) 通过消融实验, 验证 MELO-UA 新增模块的有效性。

4.1 实验环境与超参数设置

在连续复杂控制环境 MuJoCo-Gym 中的 HalfCheetah、Walker、Ant 和 Humanoid 任务上进行

对比实验和消融实验。

在 MBRL-Lib 算法库^[41]实现的 MBPO 算法基础上实现了 MELO-UA, 对于两者共有的结构和超参数, MELO-UA 的设置与 MBPO^[5]完全一致。在 MELO-UA 的额外结构上, 在对比实验中使用如下设置: 探索策略额外奖励的激励系数 $\eta = 1$, 多步不确定度的前向视野 $K_b = 2$, 学习策略的筛选率 $w = \frac{H - k + 1}{H + 1}$, 其中 k 为当前模拟步数, H 为总模拟前向视野, 在各任务中的设置各不相同, 本文与 MBPO^[5]保持相同设置。

4.2 对比实验

将 MELO-UA 与若干代表性的基于模型/无模型的强化学习算法进行对比。基于模型的算法包括 MBPO 和 M2AC, 其中 M2AC 在模型利用过程中利用了不确定度估计进行数据的惩罚和筛选, 是目前性能最优的算法之一。无模型算法包括 SAC 和 PPO。每个算法会在每个任务中进行 4 次独立实验, 以计算平均值和方差。

对比实验的结果如图 2 中的收敛曲线所示。在收敛曲线图中, 实线代表算法在多组实验中的平均值, 阴影带表示实验标准差, 水平虚线表示 SAC 算法的收敛性能, 其中 HalfCheetah、Walker2D、Ant 的收敛性能在 SAC 交互 3M 步后取得, 而 Humanoid 的收敛性能在 SAC 交互 8M 步后取得。

如收敛曲线显示, MELO-UA 在样本效率和最终性能上均表现优异, 其样本效率超过了所有对比算法。且这一性能优势在环境较为复杂的 Ant、Walker2D 和 Humanoid 任务中尤为明显; 在 Walker2D 任务中, MELO-UA 只需要 100 k 步迭代就可以达到其他基于模型算法 200k 步时的性能, 与 SAC 在 3 M 步时的收敛性能接近; 在 Ant 任务中, 算法在样本效率和性能上均比现有算法有明显的提升。而在最复杂的 Humanoid 任务中,

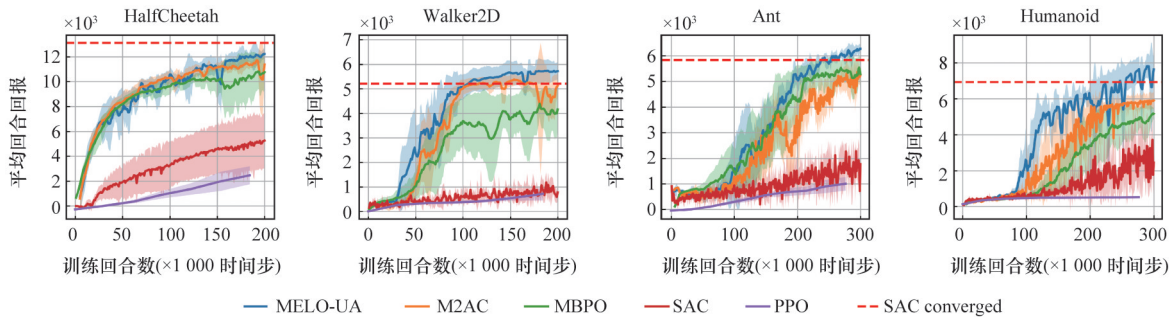


图2 MuJoCo对比实验结果

Fig. 2 Performance of MELO-UA and baselines on MuJoCo

MELO-UA 在 200 k 步时的性能已经超过了其他所有基于模型算法的 300 k 步性能,且达到了 SAC 的收敛性能,最终渐进性能相对于 SAC 甚至取得了提升,相比其他基于模型的方法有 30% 以上的提升。上述实验结果说明本文所提方法在上述复杂控制任务中表现出优异的样本效率和渐进性能,验证了 MELO-UA 可以有效地缓解基于模型的强化学习中的探索利用困境。反观在环境较为简单的 HalfCheetah 上,MELO-UA 的性能提升就不太明显。造成这一现象的原因可能是由于该环境较为简单,探索复杂度不高,MBPO 原生的基于随机性策略的探索机制就已经足够。

4.3 消融实验

在消融实验中,对 MELO-UA 的新增模块进行不同程度的删减,进行模块有效性实验;同时,针对基于不确定度的探索过程中,多步不确定度计算的前向视野超参数 K_b 的不同设置进行实验。待比较消融模型及其说明如表 1 所示。实验在 Humanoid 和 Walker2D 任务上进行。

消融实验结果如图 3 所示。其中,MBPO-B 在 Humanoid 任务上无法收敛,而在 Walker2D 任务上造成了性能下降,这印证了在利用模型时对高不确定度样本的错误利用确实会影响算法性能。在引入探索策略-学习策略架构后,独立优化探索策略-学习策略的 MELO-UA-I 相比于只利用惩罚

表 1 消融模型说明

Table 1 Algorithm information in ablation study	
模型名称	说明
MBPO	—
MBPO-B	在 MBPO 上加入基于不确定度的奖励
MBPO-P	在 MBPO 上加入基于不确定度的约束
MELO-UA	—
MELO-UA-I	消去联合优化的 MELO-UA
MELO-UA-s1	多步不确定度的前向视野为 1
MELO-UA-s5	多步不确定度的前向视野为 5

机制的 MBPO-P 算法性能更低,加上联合优化之后则体现了优异的性能,这说明探索策略和学习策略之间的差异确实会影响算法的效果,本文提出的联合优化方法能缓解这一差异的影响。此外,将多步不确定度前向视野 K_b 设置为 1 时,MELO-UA-s1 的性能相比 $K_b = 2$ 的 MELO-UA 有着较为明显的性能降低,而 MELO-UA-s5 的性能则要好于 MELO-UA-s1,但也低于 MELO-UA。这证明本文提出的多步不确定度激励是有效的,但步长过长时,算法的性能同样会受到影响。此外,注意到在 HalfCheetah 上的实验中,几种消融算法与 MBPO 的性能相差不大,与对比实验结果相似,这验证了在对比实验中的猜想,即 HalfCheetah 是一个较为简单的任务,MBPO 就可以获得足够好的样本效率和渐进性能。

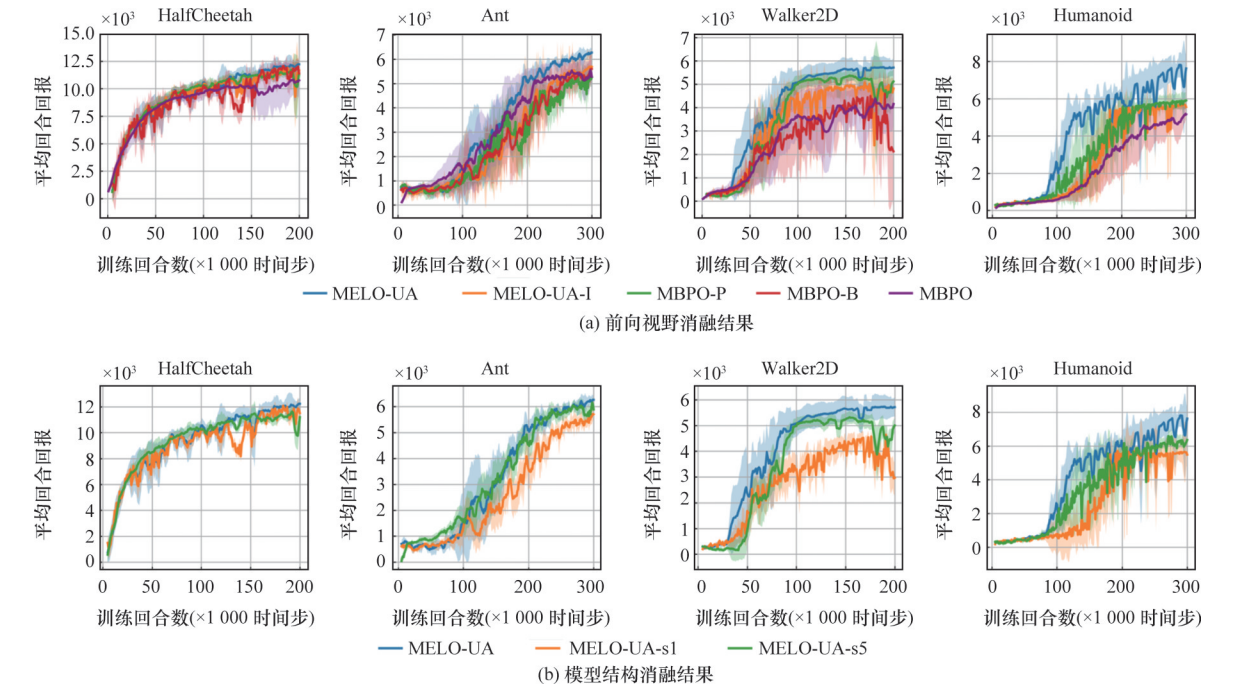


图 3 消融实验结果

Fig. 3 Result of ablation study

该部分实验有效地验证了本文所提出的数据增强模块和联合优化算法的有效性。

5 总结与讨论

本文提出不确定度增强下基于模型的探索-学习策略联合优化方法 MELO-UA, 使用2种不同的基于不确定度的数据增强方法缓解基于模型的强化学习中探索-利用之间的冲突。实验验证了所提方法的有效性, 结果表明该方法在样本效率和性能上都超过了现有的最优方法。

参考文献

- [1] Schrittwieser J, Antonoglou I, Hubert T, et al. Mastering Atari, Go, chess and shogi by planning with a learned model [J]. *Nature*, 2020, 588(7839): 604-609. DOI: 10.1038/s41586-020-03051-4.
- [2] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge [J]. *Nature*, 2017, 550(7676): 354-359. DOI: 10.1038/nature24270.
- [3] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms [EB/OL]. arXiv 2017: 1707.06347. (2017-07-20) [2024-07-23]. <http://arxiv.org/abs/1707.06347v2>.
- [4] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning [J]. *Nature*, 2015, 518(7540): 529-533. DOI: 10.1038/nature14236.
- [5] Janner M, Fu J, Zhang M, et al. When to trust your model: model-based policy optimization [EB/OL]. arXiv 2019: 1906.08253. (2019-06-19) [2024-07-23]. <http://arxiv.org/abs/1906.08253v3>.
- [6] Nagabandi A, Kahn G, Fearing R S, et al. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning [C]//2018 IEEE International Conference on Robotics and Automation (ICRA). Brisbane, QLD, Australia. IEEE, 2018: 7559-7566. DOI: 10.1109/ICRA.2018.8463189.
- [7] Higuera J C G, Meger D, Dudek G. Synthesizing neural network controllers with probabilistic model-based reinforcement learning [C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018: 2538-2544. DOI: 10.1109/IROS.2018.8594018.
- [8] Chua K, Calandra R, McAllister R, et al. Deep reinforcement learning in a handful of trials using probabilistic dynamics models [C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. December 3 - 8, 2018, Montréal, Canada. ACM, 2018: 4759-4770. DOI: 10.5555/3327345.3327385.
- [9] Feinberg V, Wan A, Stoica I, et al. Model-based value estimation for efficient model-free reinforcement learning [EB/OL]. arXiv 2018: 1803.00101. (2018-02-28) [2024-07-23]. <http://arxiv.org/abs/1803.00101v1>.
- [10] Sutton R S. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming [M]//Machine Learning Proceedings 1990. Amsterdam: Elsevier, 1990: 216-224. DOI: 10.1016/b978-1-55860-141-3.50030-4.
- [11] Luo Y, Xu H, Li Y, et al. Algorithmic Framework for model-based deep reinforcement learning with theoretical guarantees [C/OL]//International Conference on Learning Representations. 2019. (2018-12-21) [2024-07-23]. <https://openreview.net/forum?id=BJe1E2R5KX>.
- [12] Buckman J, Hafner D, Tucker G, et al. Sample-efficient reinforcement learning with stochastic ensemble value expansion [C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. December 3 - 8, 2018, Montréal, Canada. ACM, 2018: 8234-8244. DOI: 10.5555/3327757.3327916.
- [13] Deisenroth M P, Rasmussen C E. PILCO: a model-based and data-efficient approach to policy search [C]//Proceedings of the 28th International Conference on International Conference on Machine Learning. 28 June 2011, Bellevue, Washington, USA. ACM, 2011: 465-472. DOI: 10.5555/3104482.3104541.
- [14] Heess N, Wayne G, Silver D, et al. Learning continuous control policies by stochastic value gradients [EB/OL]. arXiv 2015: 1510.09142. (2015-10-30) [2024-07-23]. <http://arxiv.org/abs/1510.09142v1>.
- [15] Hafner D, Lillicrap T, Ba J, et al. Dream to control: Learning behaviors by latent imagination [EB/OL]. arXiv 2019: 1912.01603. (2019-12-03) [2024-07-24]. <http://arxiv.org/abs/1912.01603v3>.
- [16] Hafner D, Pasukonis J, Ba J, et al. Mastering diverse domains through world models [EB/OL]. arXiv 2023: 2301.04104. (2023-01-10) [2024-07-23]. <http://arxiv.org/abs/2301.04104v2>.
- [17] Hafner D, Lillicrap T, Fischer I, et al. Learning latent dynamics for planning from pixels [EB/OL]. arXiv 2018: 1811.04551. (2018-11-12) [2024-07-23]. <http://arxiv.org/abs/1811.04551v5>.
- [18] Kurutach T, Clavera I, Duan Y, et al. Model-ensemble trust-region policy optimization [EB/OL]. arXiv 2018: 1802.10592. (2018-02-28) [2024-07-23]. <http://arxiv.org/abs/1802.10592v2>.
- [19] He W, Jiang Z. A Comprehensive Survey on Uncertainty Quantification for Deep Learning [EB/OL]. arXiv 2023: 2302.13425. (2023-02-26) [2024-07-23]. <https://arxiv.org/abs/2302.13425>.
- [20] Psaros A F, Meng X H, Zou Z R, et al. Uncertainty quantification in scientific machine learning: methods, metrics, and comparisons [J]. *Journal of Computational Physics*, 2023, 477: 111902. DOI: 10.1016/j.jcp.2022.111902.
- [21] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep

- Ensembles [EB/OL]. arXiv 2016: 1612.01474. (2016-12-05)[2024-07-23]. <https://arxiv.org/abs/1612.01474>.
- [22] Wang Z, Jusup M, Shi L, et al. Exploiting a cognitive bias promotes cooperation in social dilemma experiments [J]. *Nature Communications*, 2018, 9 (1): 2954. DOI: 10.1038/s41467-018-05259-5.
- [23] Lu C, Ball P, Parker-Holder J, et al. Revisiting design choices in offline model-based reinforcement learning [EB/OL]. arXiv 2021:2110.04135. (2021-10-08)[2024-07-23]. <https://arxiv.org/abs/2110.04135>.
- [24] Pan F, He J, Tu D, et al. Trust the model when it is confident: masked model-based actor-critic [EB/OL]. arXiv 2020: 2020.04893. (2020-10-10)[2024-07-23]. <https://arxiv.org/abs/2010.04893>.
- [25] Yu T H, Thomas G, Yu L T, et al. MOPO: model-based offline policy optimization [EB/OL]. arXiv 2020: 2005.13239. (2020-05-27)[2024-07-23]. <https://arxiv.org/abs/2005.13239>.
- [26] Bechtle S, Lin Y X, Rai A, et al. Curious iLQR: resolving uncertainty in model-based RL [EB/OL]. arXiv 2019:1904.06786. (2019-04-15)[2024-07-23]. <http://arxiv.org/abs/1904.06786v2>.
- [27] Hao J Y, Yang T P, Tang H Y, et al. Exploration in deep reinforcement learning: from single-agent to multiagent domain [J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. DOI: 10.1109/TNNLS.2023.3236361.
- [28] Abbasi-Yadkori Y, Pál D, Szepesvári C. Improved algorithms for linear stochastic bandits [C]//*Proceedings of the 24th International Conference on Neural Information Processing Systems*. December 12 - 15, 2011, Granada, Spain. ACM, 2011: 2312-2320. DOI: 10.5555/2986459.2986717&widgetKey=ux3-publicationContent-widget_f69d88a8-b404-4aae-83a9-9acea4426d78_1498_en.
- [29] Bai C J, Wang L X, Han L, et al. Principled exploration via optimistic bootstrapping and backward induction [EB/OL]. arXiv 2021: 2105.06022. (2021-05-13)[2024-07-23]. <http://arxiv.org/abs/2105.06022v2>.
- [30] Zheng Y, Liu Y, Xie X F, et al. Automatic web testing using curiosity-driven reinforcement learning [C]//2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE). Madrid, ES. IEEE, 2021: 423-435. DOI: 10.1109/ICSE43902.2021.00048.
- [31] Ciosek K, Vuong Q, Loftin R, et al. Better exploration with optimistic actor-critic [EB/OL]. arXiv 2019: 1910.12807. (2019-10-28)[2024-07-23]. <http://arxiv.org/abs/1910.12807v1>.
- [32] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor [EB/OL]. arXiv 2018: 1801.01290. (2018-01-04)[2024-07-23]. <http://arxiv.org/abs/1801.01290v2>.
- [33] Burda Y, Edwards H, Storkey A, et al. Exploration by random network distillation [EB/OL]. arXiv 2018: 1810.12894. (2018-10-30)[2024-07-23]. <http://arxiv.org/abs/1810.12894v1>.
- [34] Liu J Y, Wang Z, Zheng Y, et al. OVD-explorer: optimism should not be the sole pursuit of exploration in noisy environments [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 38(12): 13954-13962. DOI: 10.1609/aaai.v38i12.29303.
- [35] Bai C J, Wang L X, Han L, et al. Dynamic bottleneck for robust self-supervised exploration [EB/OL]. arXiv 2021: 2110.10735. (2021-10-20)[2024-07-23]. <http://arxiv.org/abs/2110.10735v2>.
- [36] Oudeyer P Y, Kaplan F, Hafner V V. Intrinsic motivation systems for autonomous mental development [J]. *IEEE Transactions on Evolutionary Computation*, 2007, 11(2): 265-286. DOI: 10.1109/TEVC.2006.890271.
- [37] Pathak D, Agrawal P, Efros A A, et al. Curiosity-driven exploration by self-supervised prediction [C/OL]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2017: 488-489. DOI: 10.1109/CVPRW.2017.70.
- [38] Pathak D, Gandhi D, Gupta A. Self-supervised exploration via disagreement [EB/OL]. arXiv 2019: 1906.04161. (2019-01-10)[2024-07-23]. <https://arxiv.org/abs/1906.04161>.
- [39] Yuan Y F, Hao J Y, Ni F, et al. EUCLID: towards efficient unsupervised reinforcement learning with multi-choice dynamics model [EB/OL]. arXiv 2022: 2210.00498. (2022-10-02)[2024-07-23]. <http://arxiv.org/abs/2210.00498v2>.
- [40] Haarnoja T, Tang H R, Abbeel P, et al. Reinforcement learning with deep energy-based policies [EB/OL]. arXiv 2017: 1702.08165. (2017-02-27)[2024-07-23]. <http://arxiv.org/abs/1702.08165v2>.
- [41] Pineda L, Amos B, Zhang A, et al. MBRL-lib: a modular library for model-based reinforcement learning [EB/OL]. arXiv 2021: 2104.10159. (2021-04-20)[2024-07-23]. <http://arxiv.org/abs/2104.10159v1>.