

伍家威, 郝志峰. 基于技能发现的元强化学习[J]. 广东工业大学学报, 2025, 42(3): 52–61. doi: 10.12052/gdutxb.240041.
Wu Jiawei, Hao Zhifeng. Meta reinforcement learning based on skill discovery[J]. Journal of Guangdong University of Technology, 2025, 42(3): 52–61. doi: 10.12052/gdutxb.240041.

基于技能发现的元强化学习

伍家威¹, 郝志峰^{1,2}

(1. 广东工业大学 计算机学院, 广东 广州 510006; 2. 汕头大学 理学院, 广东 汕头 515063)

摘要: 在机器人控制任务的复杂环境中, 元强化学习(Meta Reinforcement Learning, Meta-RL) 作为关键组成部分, 它可以利用先前的经验来解决未见过的、长期的以及奖励稀疏的复杂任务。基于技能的元强化学习(Skill-based Meta-RL) 方法旨在利用任务上下文提取出有用的技能, 这些技能可以帮助智能体在元测试阶段快速适应新环境。然而现有的方法所学到的技能缺乏泛用性与适应性, 从而限制了其在元测试任务集中的表现。为了克服这一限制, 本文提出了一种基于技能发现的元强化学习(Skill Discovery Meta-RL, SDMRL) 方法, 通过此方法可以在没有奖励函数的情况下学习到更加有用的技能。本文将目标形式化为利用最大熵策略最大化信息论目标, 这个简单而有效的探索目标能让智能体从非结构化数据中无监督地学习到有用的技能和技能先验。在迷宫导航(Maze Navigation) 连续控制任务中的实验表明, SDMRL方法比以往的元强化学习方法更有效, 学习到的技能可以解决长期的复杂稀疏奖励任务。

关键词: 元强化学习; 强化学习; 技能发现

中图分类号: TP181

文献标志码: A

文章编号: 1007-7162(2025)03-0052-10

Meta Reinforcement Learning Based on Skill Discovery

Wu Jiawei¹, Hao Zhifeng^{1,2}

(1. School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China; 2. College of Science, Shantou University, Shantou 515063, China)

Abstract: In the realm of complex environments for robot control tasks, Meta Reinforcement Learning has emerged as a pivotal component, leveraging prior experiences to tackle unseen, long-term, and sparsely rewarded intricate tasks. Skill-based Meta-RL methods aim to extract useful skills from task contexts, aiding agents in swiftly adapting to new environments during meta-testing. However, the skills learned by existing methods lack generality and adaptability, limiting their performance across meta-testing task sets. To address this, a Skill Discovery Meta-RL (SDMRL) approach is proposed in this paper, by learning more useful skills in the absence of reward functions. The objective of the SDMRL is formalized as maximizing an information-theoretic objective using maximum entropy policies, enabling agents to learn valuable skills and skill priors from unstructured data in an unsupervised manner. Experimental results in continuous control tasks such as Maze Navigation demonstrate the effectiveness of the SDMRL approach over previous meta reinforcement learning methods, and the learned skills proficiently address the long-term complex sparse reward tasks.

Key words: meta-reinforcement learning; reinforcement learning; skill discovery

目前, 元强化学习(Meta Reinforcement Learning, Meta-RL) 备受关注, 在各个领域广泛应用, 尤其在机器人控制任务中展现出显著的成效^[1-2]。这些元强化学习方法在元训练阶段从一系列任务中获得先验经验, 然后在元测试阶段将先验经验应用于未知的目

标任务, 且目标任务通常与训练任务存在某种隐含关系。这些先验经验经常使用的形式有两种: 任务背景和技能。任务背景与任务的关键统计信息关联, 这些信息可通过智能体生成的过去的轨迹中获得。从轨迹中提取出来的任务背景能够让智能体适应未知

收稿日期: 2024-03-20 录用日期: 2024-05-06

基金项目: 科技创新2030——“新一代人工智能”重大项目(2021ZD0111501); 国家优秀青年科学基金资助项目(62122022); 国家自然科学基金资助项目(61876043, 61976052, 62206064)

作者简介: 伍家威(1998–), 男, 硕士研究生, 主要研究方向为机器学习、强化学习, E-mail: 289873439@qq.com

通信作者: 郝志峰(1968–), 男, 教授, 博士, 主要研究方向为机器学习、数据挖掘, E-mail: zfhao@gdut.edu.cn

的目标任务。另一方面,技能则代表可以解决子任务的有用行为。这些技能可以从过去的轨迹中学习,并且可以应用于新任务甚至新的环境当中^[3]。通过学习和积累解决不同子任务的一系列技能,可以将这些技能组合起来,从而实现复杂任务的解决方案^[4-5]。这种技能的组合和应用方式为智能体在处理多样化、复杂性任务时提供了更加灵活和高效的方法。

然而,目前大多数元强化学习方法在面对长期且奖励稀疏的任务时存在较差的泛化性和样本效率低下问题,对于元测试阶段里未知的目标任务不能做出很好的适应^[6]。这一问题的关键根源在于最初阶段的探索不足,导致对环境和任务的理解不够深入,进而影响到所学到的技能的广度和深度。探索和学习是相互依存的过程,若在最初阶段缺乏有效的探索,将会限制对环境特征和任务结构的全面理解,使得学得技能范围受限。缺乏充分的探索将进一步导致从潜在空间中提取的先验经验存在局限性。这种局限性将使得智能体在探索和学习过程中仅能涉及潜在空间的一小部分,从而导致产生次优的策略^[7]。

此外,智能体在最初阶段的探索时需要大量的数据,数据不足也会导致探索的不充分。而在自动驾驶、室内导航或者机器人操作等领域,有大量收集到的智能体数据可用。但这些数据大多是非结构化的,没有明确的任务或奖励定义,因此很难用于学习新任务。在这项工作中,本文的目标是设计一种能利用非结构化的数据来加速技能学习的方法。在没有任何任务或奖励信息的情况下,从数据中学习技能,并且学到的技能可以用于解决未知的目标任务。无需奖励信息学习技能有几个实际用途^[8]。一方面,稀疏奖励的环境在智能体到目标状态之前都是没有奖励的。在这些环境中,无监督学习到的技能可能有助于解决探索中的挑战^[9]。另一方面,在许多实际环境中,与环境的交互基本上是免费的,但评估奖励需要人类反馈^[10]。虽然可以通过设计奖励函数将人类排除在外,但是设计一个能够让智能体做出期望行为的奖励函数是具有挑战性的^[11]。无需奖励的技能发现部分回答了这个问题。

在奖励函数未知的情况下无监督学会技能是一个极具挑战性的问题^[12]。过去的方法如基于技能的元策略学习(Skill-based Meta-Policy Learning, SiMPL)^[6]主要依赖于行为克隆与熵正则化,但其局限性在于探索与泛化能力不足。因此,解决这一问题的关键在于指定一个明确的学习目标,以确保每个技能都具有差异性,并且这些技能可以有效探索绝

大部分的状态空间。因此在元测试阶段面对从未见过的目标任务时,能够有效且快速地探索环境并实现策略收敛。

针对此挑战,本文提出了一种基于技能发现的元强化学习算法(Skill Discovery Meta Reinforcement Learning, SDMRL)。SDMRL通过基于互信息的目标来鼓励模型无监督发现可区分的有用技能。通过学习尽可能随机的技能,可以将技能之间的距离推开,使每种技能都能够适应扰动并有效地探索环境。从本质上讲,SDMRL利用该目标增强了技能发现阶段获得的技能的可识别性、泛化性和多样性。本文的主要贡献有以下几点:

(1) 提出了一种通用的元强化学习框架,旨在学习有用的技能而无需任何奖励信息。SDMRL能让智能体从非结构化数据中无监督地学习到有用的技能和技能先验。尽管从未收到真正的任务奖励,所学到的技能能够快速适应目标任务。

(2) 拓展了最大熵强化学习(Soft Actor Critic, SAC),提出了基于技能的SAC,将目标形式化为利用最大熵策略最大化信息论目标,能够通过离线数据集学习到更有用的技能。

(3) 展示了在元测试阶段如何将学习到的技能和技能先验应用到元测试任务上,并且能有效加速对于未知的目标任务的学习。

1 相关工作

1.1 离线元强化学习

离线元强化学习(Offline Meta-RL)的主要目标是从离线数据中学习并获取元知识,使得智能体能够在遇到新任务时,利用先前的离线数据快速适应,并做出有效的决策^[13-15]。但是在线数据和离线数据之间存在分布差异,因此,在学习过程中如何获得泛化良好的鲁棒任务表示是十分重要的。以前的方法如贝叶斯离线强化学习(Bayesian Offline Reinforcement Learning, BOREL)^[16]、优势加权的元执行者-评论家方法(Meta-Actor Critic with Advantage Weighting, MACAW)^[17]通过有标注的数据来进行元学习,例如奖励和任务标注,并仅用少量新数据适应新任务。然而由于缺乏相应的先验任务经验,一些元训练任务难以标注。也有利用离线和在线数据的半监督学习来进行分布转移的方法,如半监督元执行者-评论家方法(Semi-supervised Meta Actor-Critic, SMAC)^[18]利用带有奖励的离线数据元训练好的自适应策略,可以无需奖励信息弥合与无监督在线数据的分布转

移,但严重依赖于离线数据的注释。相比之下,SDMRL能利用跨多个任务的大型离线数据库,无需奖励和任务注释就能学习到有用的技能。

1.2 基于上下文的元强化学习

基于上下文的元强化学习(Context-based Meta-Reinforcement Learning)元训练阶段将先前的经验作为特定于任务的上下文的形式^[19-20]。以前的方法如基于概率嵌入的演员-评论家强化学习方法(Probabilistic Embeddings for Actor-critic RL, PEARL)^[2]、潜在动态的元强化学习(Meta-RL with Latent Dynamics, MELD)^[21]被提出用于元学习动态模型和策略,以实现未知目标任务的快速适应。PEARL通过将环境状态和元学习任务嵌入到一个共享的潜在空间中,可以在不同的环境上下文下学习到通用的策略表示。这种通用表示可以帮助智能体在新的环境中快速适应,并且能够推广到不同的任务和环境设置中。SDMRL也是一种基于上下文的方法,在元训练阶段利用上下文学习有用的技能,从而提高智能体的泛化能力,并在元测试阶段仅需少量上下文即可实现对未知任务的适应。

1.3 基于技能的学习

技能发现(Skill Discovery)是一种通过没有奖励和任务注释的离线数据提取出可重用的短期行为的方法^[22-23]。基于技能的学习方法如基于离线原始行为发现的离线强化学习加速方法(Offline Primitive discovery for Accelerating offline RL, OPAL)^[24]、多样性即所求的无监督技能发现方法(Diversity Is All You Need: Learning Skills without a Reward Function, DIAYN)^[9]、基于利普希茨约束的技能发现方法

(Lipschitz-constrained Skill Discovery, LSD)^[25]重新组合这些技能来解决未知的目标任务,并且比从头开始学习的方法收敛得更快。利用不同的数据集进行训练,这些方法可以提取到广泛的技能库,并且可以学习复杂的、长期的任务^[26]。但是,它们仍然需要大量的环境交互去学习新任务。SiMPL^[6]将基于技能的学习与元学习相结合,利用元学习提高了样本效率,又通过技能提高了完成元测试任务的成功率。SDMRL也是一种基于技能的学习方法,将技能发现与元强化学习相结合,在元训练阶段从离线数据中提取更具有泛化性的技能,并在元测试阶段快速适应新任务。

2 方法

如图1所示,SDMRL包括3个阶段:技能发现(Skill Discovery)、元训练(Meta Train)、元测试(Meta Test)。技能发现阶段:通过一个技能提取器 $q(z|s)$ 和一个基于技能的低级策略 $\pi(a|s, z)$ 从离线数据集 D 中学习有用的技能,并训练一个技能先验 $p(z|s_0)$;元训练阶段:通过元训练任务 T 训练一个任务编码器 $q(e|c)$ 和一个高级技能策略 $\pi(z|s, e)$ 分别用于学习任务编码和技能,而预训练的基于技能的低级策略 $\pi(a|s, z)$ 则保持固定;元测试阶段:利用元训练模块快速适应未见过的目标任务 T^* ,任务上下文由它的几个转移(Transitions)产生,高级技能策略 $\pi(z|s, e)$ 保持微调状态,而任务编码器 $q(e|c)$ 和低级策略 $\pi(a|s, z)$ 则保持固定。

2.1 问题定义

元强化学习任务通常形式化为完全可观察的马尔可夫决策过程(Markov Decision Process, MDP),定

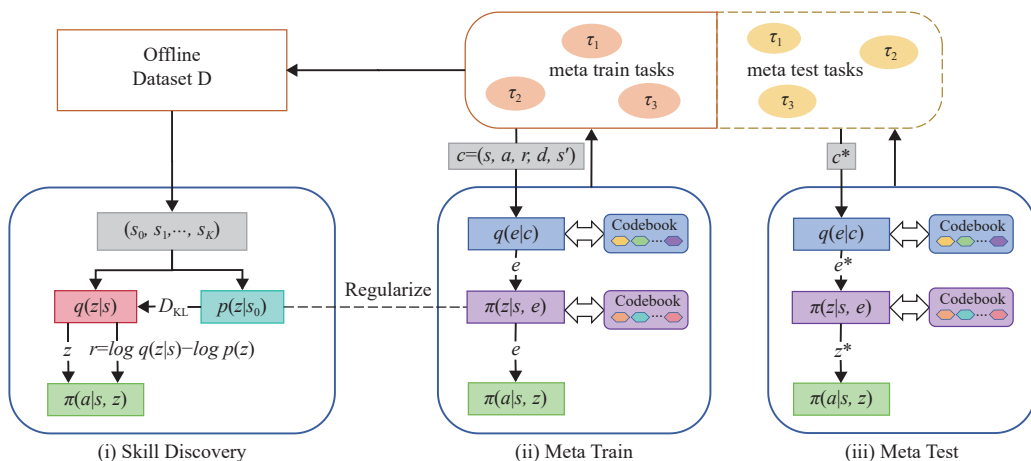


图1 SDMRL框架图

Fig.1 The overview of SDMRL

义为一个元组 $\langle S, A, p, r, \gamma, \rho_0 \rangle$ 。其中 S 是状态空间, A 是动作空间, $p(s'|s, a)$ 是状态转移, $r(S, A)$ 是奖励函数, ρ_0 是初始状态分布, γ 是未来奖励的折扣因子。策略是动作上的一个分布 $\pi(a|s)$ 。在MDP中, 智能体在给定的状态下初始化, 并通过策略 π 采样, 在每个时间步选择一个动作。同时, 环境通过使用状态转移概率 p 来更新状态并提供奖励 r 和完成标志 d 来做出响应。此外, 时间步长 t 的边际状态被定义为 $\mu'_\pi(s)$, 智能体的目标是最大化期望累积奖励 $\max_\pi \mathcal{J}_M(\pi) = \mathbb{E}_{s_t \sim \mu'_\pi, a_t \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]$, 其中 $\mathbb{E}$ 表示数学期望, $\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$ 表示累计奖励。

作为一种离线元强化学习方法, 本文假设可以访问状态-动作轨迹 $D = (s_0, a_0, \dots, s_t, a_t)$, 这些轨迹来自从各种未知任务或未标记的数据收集来的数据集。由于行为的多样性, D 能够加速不同任务的学习。本文还假设能够访问 n_{task} 个元训练任务 $T = \{\mathcal{T}_1, \dots, \mathcal{T}_{n_{\text{task}}}\}$, 同时有一组未知的目标任务 T^* , 并将每个任务分布表示为一个MDP。并且假设 D 与 T^* 或 T 与 T^* 之间不存在直接关系, 即离线数据集 D 中不包含任何在 T^* 中解决任务的演示, 并且 T 与 T^* 不存在交集。

SDMRL旨在从离线数据集 D 中进行技能发现, 使智能体学习到有用的技能; 在元训练任务集合 T 上执行元训练阶段; 在元测试阶段处理未知的目标任务集合 T^* 。本文将任务上下文和技能的分布表示为高斯分布, 用 C 和 Z 表示。而 c 和 z 分别是 C 和 Z 中采样得到的任务上下文嵌入和技能嵌入。

2.2 技能发现

2.2.1 3个目标

SDMRL使用信息论符号来构建优化目标, 其中 $I(\cdot; \cdot)$ 表示互信息, $H[\cdot]$ 表示熵。首先, 为了使学到的技能有用, 技能应该能够影响智能体所处的状态。不同的技能应该访问不同的状态, 这样的技能才是可区分的。通过最大化技能和状态之间的互信息 $I(S; Z)$, 可以确保学到的技能能够指导智能体在不同状态之间进行转换, 即使得技能对状态的控制能力更强。

其次, 应该利用状态而不是动作来区别技能, 因为如果某些动作并未对环境作出影响, 那么对于外部观察者来说这些动作是不可见的。通过最小化在给定状态下技能和动作之间的互信息 $I(A; Z|S)$, 可以减小技能与动作之间的关联性, 从而使得技能更加独立于具体的动作选择, 而更加准确地反映对状态

的控制能力。以抓起物块举例, 施加不同的力就是不同的动作, 如果物块没有被移动或抬起, 外部观察者无法判断机械臂在抓取时所施加的力, 一个有用且可识别的技能应该是改变物块的状态, 即移动物块。

最后, 应该鼓励探索, 并旨在学习尽可能随机的技能来使得学到的技能尽可能多样化。一个具有高熵值的可区分的技能, 应当远离其他技能的状态空间的一部分, 以免动作的随机性导致陷入无法区分的状态。本文将所有技能与 $p(z)$ 一起视为策略的混合, 最大化该混合策略的熵 $H[A|S]$ 鼓励模型学习多样化的技能, 从而促进探索与学习的过程。

2.2.2 优化目标

根据上述3个目标: 最大化技能和状态之间的互信息 $I(S; Z)$, 最大化该混合策略的熵 $H[A|S]$, 最小化在给定状态下技能和动作之间的互信息 $I(A; Z|S)$, 最终可以总结成一个最大化目标:

$$\begin{aligned} \mathcal{F}(\theta) &\triangleq I(S; Z) + H[A|S] - I(A; Z|S) = \\ &(H[Z] - H[Z|S]) + H[A|S] - (H[A|S] - H[A|S, Z]) = \\ &H[Z] - H[Z|S] + H[A|S, Z] \end{aligned} \quad (1)$$

式(1)第2行第1项鼓励在 $p(z)$ 上的先验分布具有高熵。在SDMRL中, 本文将 $p(z)$ 固定为均匀分布, 确保它具有最大熵。第2项表明从当前状态判断技能 z 是容易的。第3项表明每个技能应该尽可能随机地行动。由于无法对所有状态和技能进行准确计算 $p(z|s)$, 所以用可学习的技能提取器 $q_\phi(z|s)$ 来近似这个后验。根据Jensen不等式, 用 $q_\phi(z|s)$ 替换 $p(z|s)$ 可以给目标 $\mathcal{F}(\theta)$ 提供一个变分下界 $\mathcal{G}(\theta, \phi)$ ^[27]。为了更直观地了解如何优化该目标, 本文重新排列了式(1):

$$\begin{aligned} \mathcal{F}(\theta) &\triangleq H[A|S, Z] - H[Z|S] + H[Z] = \\ &H[A|S, Z] + \mathbb{E}_{z \sim p(z), s \sim \pi(z)} [\ln p(z|s)] - \\ &\mathbb{E}_{z \sim p(z)} [\ln p(z)] \geq H[A|S, Z] + \\ &\mathbb{E}_{z \sim p(z), s \sim \pi(z)} [\ln p(z|s) - \ln p(z)] \triangleq \mathcal{G}(\theta, \phi) \end{aligned} \quad (2)$$

2.2.3 基于技能的SAC

本文拓展了最大熵强化学习 (Soft Actor Critic, SAC) 算法^[28], 实现了一个以技能 z 为条件的低级策略 $\pi_\theta(a|s, z)$ 。SAC算法策略的目标函数为同时最大化累计回报和策略的熵。最大化策略的熵对应式(2)的第1项, 并且将熵项按 α 缩放。然后本文通过将任务奖励替换为式(2)后两项作为伪奖励来最大化 $\mathcal{G}$ 中的期望。最后可以得到基于技能的SAC的低级策略的目标函数:

$$\mathcal{J}(\pi_\theta) = \arg\max \mathbb{E} \left[\sum r_z(s, a) + \alpha H(\pi_\theta(\cdot|s)) \right],$$

$$\text{where } r_z(s, a) \triangleq \ln q_\varphi(z|s) - \ln p(z) \quad (3)$$

智能体访问那些容易区分的状态时会获得高奖励,同时技能提取器也会被更新来更好地从访问状态中推断出技能 z 。根据经验可知 $\alpha = 0.1$ 可以在探索性和可辨别性之间提供一个良好的平衡。

此外,本文通过交叉熵训练一个技能提取器 $q(z|s)$,对于从离线数据集中随机绘制的 K 步轨迹,技能提取器可以从 K 步的连续的状态中提取出技能 z 。本文还训练了一个技能先验 $p(z|s_0)$ 。利用KL散度使技能先验的输出与技能提取器相匹配: $\min_p D_{\text{KL}}([q(z|s)], p(z|s_0))$,其中 $[q(z|s)]$ 表示梯度流禁止进入技能编码器。

2.3 元训练

2.3.1 任务上下文对比学习

任务上下文作为一种先验知识,能够指导智能体应该专注于分布中的哪个任务,并提供智能体在遇到新任务时调整策略所需的信息^[29]。本文通过将任务特征加以区分,增强任务上下文表征。

在元训练任务的轨迹上,本文分为两个阶段进行采样:首先,对当前任务的较长轨迹进行采样,形成一个迷你数据集;然后,从长轨迹中随机采样两段长度相同的不同轨迹作为锚点(Anchor)和正样本(Positive Samples),并保持元组的相对顺序。负样本(Negative Samples)则从其他任务中随机采样。

本文实现了一个任务编码器 $q(e|c)$ 处理这些轨迹,用于生成任务上下文的表征 $\tilde{e}$, $\tilde{e}^+$ 和 $\tilde{e}^-$ 。将锚点、正样本、负样本用于经典的对比学习损失函数:三元组损失。三元组损失的目标是最小化 $\tilde{e}$ 和 $\tilde{e}^-$ 之间的相似性,最大化 $\tilde{e}$ 和 $\tilde{e}^+$ 之间的相似性。

$$\mathcal{L}_{\text{triplet}}(\tilde{e}, \tilde{e}^+, \tilde{e}^-) = \sum_{\tilde{e} \in e} \max(0, \text{sim}(\tilde{e}, \tilde{e}^-) - \text{sim}(\tilde{e}, \tilde{e}^+) + \varepsilon) \quad (4)$$

2.3.2 任务编码器学习

本文用高斯量化变分自编码器(Gaussian Quantization Variational Autoencoder, GQ-VAE)^[30]实现了任务编码器学习阶段,包括3个模块:任务编码器 $q(e|c)$ 、可学习的码本(Code Book) $\text{CB}_e = \{e^1, \dots, e^{K_e}\}$ 和相应的解码器。采样一批上下文 $c = \{c_1, \dots, c_N\}$,其中 N 是采样的任务数量, $c_i = \{(s_j^i, a_j^i, r_j^i, d_j^i, s_j^i)\}_{j=0}^{n_e}$ 是长度为 n_e 的任务上下文输入,任务编码器 $q(e|c)$ 输出任务上下文表征 $\tilde{e}$ 。

本文用 $\mathcal{L}_e = \mathcal{L}_{\text{BC}} + \lambda \mathcal{L}_{\text{GQ}}$ 作为目标,其中 $\mathcal{L}_{\text{BC}}$ 是基于技能的低级策略产生的行为克隆损失, λ 是损失的权重系数, $\mathcal{L}_{\text{GQ}}$ 的具体公式为

$$\mathcal{L}_{\text{GQ}} = \mathbf{m}(\tilde{e}) - \tilde{e}_2 + \eta \mathbf{m}(\tilde{e}) - \tilde{e}_2 + \tilde{c} - c_2,$$

$$\text{where } \mathbf{m}(\tilde{e}) := \arg\min_{e^k \in \text{CB}_e, 1 \leq k \leq K_e} \tilde{e} - e_2^k \quad (5)$$

2.3.3 高级技能策略学习

本文用GQ-VAE实现了高级技能策略学习阶段,包括3个模块:1个高级技能策略 $\pi(z|s, e)$ 、1个可学习的码本 $\text{CB}_z = \{z^1, \dots, z^{K_z}\}$ 和1个相应的解码器。给定另一批采样的上下文 $c' = \{c'_1, \dots, c'_N\}$,它与任务上下文学习阶段使用的不同,采样上下文的长度 n_z 与 n_e 也不同。高级技能策略 $\pi(z|s, e)$ 输入一个采样上下文 c'_i 和任务表征 e ,输出一个技能 z 。

本文用 $\mathcal{L}_z = \mathcal{L}_{\text{BC}} + \gamma \mathcal{L}_{\text{GQ}}$ 作为目标, γ 是权重系数, $\mathcal{L}_{\text{GQ}}$ 的具体公式为

$$\mathcal{L}_{\text{GQ}} = \mathbf{m}(\tilde{z}) - \tilde{z}_2 + \iota \mathbf{m}(\tilde{z}) - \tilde{z}_2 + \tilde{c}' - c'_2 \quad (6)$$

式中: $\mathbf{m}(\tilde{z}) := \arg\min_{z^k \in \text{CB}_z, 1 \leq k \leq K_z} \tilde{z} - z_2^k$, ι 是权重系数。总之,

GQ-VAE解耦了对任务表征和技能两个连续潜在空间的探索和相应的两个码本的学习。

除了应用GQ-VAE之外,本文还利用技能发现阶段的技能先验 $p(z|s_0)$ 更新高级技能策略^[6],公式为

$$\max_{\pi} \mathbb{E}_{e \sim q(c)} \left[\sum_t \mathbb{E}_{(s_t, z) \sim p_{\pi|e}} [r_\tau(s_t, z) - \beta D_{\text{KL}}(\pi(z|s, e), p(z|s_0))] \right] \quad (7)$$

最后,基于技能的低级策略 $\pi(a|s, z)$ 用高级技能策略 $\pi(z|s, e)$ 采样的技能 z 来选择相应的动作。

2.4 元测试

在元测试阶段,其首要目标是利用在元训练阶段获得的策略,快速而有效地适应元测试任务集中未见过的目标任务。这意味着模型需要利用之前学到的技能,以最快的速度适应新的任务环境,从而实现对未知任务的快速解决。

从直觉上来看,一个有效的策略应该首先利用各种不同的技能来对新任务的环境进行探索。通过多样化的技能应用,智能体能够更全面地了解当前任务的要求和挑战,从而为后续的决策和行动提供更为丰富的参考和可能性。这种广泛的探索过程可以帮助智能体建立起对任务环境的整体认知,为其后续的行为选择提供更为准确和有效的指导。随后,一旦智能体对任务环境有了初步的了解和探索,策

略的优化重点将转向将输出分布缩小到解决任务所需的核心技能上。这意味着智能体需要通过不断的学习和调整,将其行为和决策逐渐专注于最为有效和相关的技能上,以最大程度地提高任务执行的效率和成功率。

在元测试阶段,模型会面对未见过的目标任务。为了完成任务,首先模型会与新环境进行交互收集一些任务上下文 c^* ,并通过任务编码器 $q(e|c)$ 提取任务表征 e^* 。然后高级技能策略 $\pi(z|s, e)$ 利用任务与环境的信息 e^* 和当前的状态 s 选择当前任务需要的技能 z^* 。最后基于技能的低级策略 $\pi(a|s, z)$ 用高级技能策略 $\pi(z|s, e)$ 采样的技能 z^* 来做出一系列连续动作以完成未见过的任务。本文在技能发现阶段训练了基于技能的低级策略 $\pi(a|s, z)$,并在元训练任务集上训练了任务编码器 $q(e|c)$ 和高级技能策略 $\pi(z|s, e)$ 。为了适应未知任务,本文通过式(6)和式(7)继续优化高级技能策略,使其能够对未见过的目标任务做出及时调整,选择合适的技能完成任务,而任务编码器和低级策略则保持固定。

3 实验结果及讨论

本文的实验主要针对长期且奖励稀疏的任务,重点评估两个关键问题:(1)是否可以学习到更有用的技能;(2)是否可以实现对未见过的目标任务的快速适应。

3.1 实验设置

3.1.1 迷宫导航

本文在具有挑战性的连续控制环境迷宫导航(Maze Navigation)^[31]中评估SDMRL,如图2所示。相对于过去的元强化学习算法常见的评估方式,即跨越几十个时间步并提供密集奖励的任务^[32],迷宫导航要求智能体从迷宫中心出发,到达位于设置在迷宫各处的目标点,完成任务通常需要数百个时间步,并且只有在成功到达目标时才会获得稀疏奖励。智能体的状态空间由其在二维平面上的位置坐标和速度坐标组成,智能体的动作空间是连续的速度命令。这种任务设定给元强化学习算法带来了全新的挑战,要求算法具备长期规划、持久性学习和稀疏奖励的处理能力。

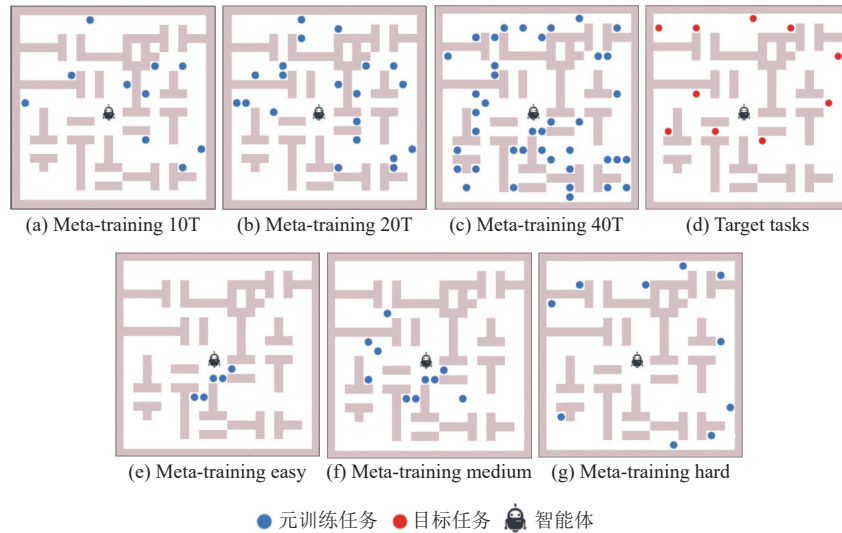


图2 迷宫导航

Fig.2 Maze navigation

3.1.2 离线数据和元训练/元测试任务

本文遵循Fu等^[31]的方法,通过在迷宫域中随机采样成对的起点和目标位置,并利用规划算法生成从起点到终点的可穿越轨迹,构建了一个没有奖励信息和任务标签的离线数据集。根据目标位置数量的不同,将元训练任务分为10T、20T和40T三种规模,如图2(a)~(c)所示。本文将采用40T规模作为元训练任务集。此外,本文元训练任务分为3种不同的难度用

于比较不同的元强化学习算法,如图2(e)~(g)所示,其中简单难度为目标位置相邻的短期任务;中等难度为目标位置方差更大的短期任务;困难难度为目标位置方差更大的长期任务。

3.2 对比方法

解耦元强化学习(Decoupled Meta Reinforcement Learning, DCMRL)^[30]是一种基于技能的元强化学习

方法,专注于解耦任务上下文和技能空间的探索和学习过程。它通过聚类方法分别对任务上下文和技能进行聚类,并将聚类中心存储于码本中,增强了它们的泛化能力。设置学习率为0.000 3, $\beta_1 = 0.9$, $\beta_2 = 0.999$ 。而经验回放池(Replay Buffer)的大小设置为20 000,并使用20个随机采样任务的梯度平均值,其中每个任务梯度通过从每个任务的经验回放池中批量采样转换(Transitions)来计算。其他的对比方法也采用同样的设置以保证对比的公平性。

基于技能的元策略学习(Skill-based Meta-Policy Learning, SiMPL)^[6]是一种基于技能的元强化学习方法,能在没有额外奖励或任务注释的情况下利用大型离线数据集的先验经验。它从离线数据集中提取可重用的技能和技能先验,并在元训练阶段利用这些技能来促进对目标任务的学习。

基于技能先验的强化学习(Skill Prior Reinforcement Learning, SPiRL)^[5]是一种迁移学习方法,旨在通过增强从这些数据集中获得的技能和技能先验来利用离线数据的先验知识。虽然它确实利用了以前从离线数据集中学习的技能,但它不涉及元训练任务的利用。因此这种方法旨在确定利用元训练任务进行进一步优化的有效性。本文将高级策略初始化为技能先验,同时为SPiRL的目标任务学习固定低级技能策略,并使用与SDMRL相同的超参。

多任务强化学习(Multi-task RL, MTRL)^[33]是一种高效的多任务强化学习基准,它通过将这些方法提炼成一个共享策略来进行学习,从而实现更高效的多任务学习。这种方法核心理念在于通过共享策略,让智能体能够在多个任务之间共享经验和知识,从而提升学习效率和泛化能力。本文对 α 使用与SDMRL相同的设置,其中 α 被自动调整以满足目标KL散度, $\delta = 0.1$ 。

基于概率嵌入的参与者-评论家强化学习(Probabilistic Embeddings for Actor-critic Reinforcement Learning, PEARL)^[2]是一种结合了概率嵌入和演员-评论家方法的强化学习算法,专注于有效解决连续动作空间下的强化学习问题。通过在元训练阶段学习到任务-潜在变量的分布,PEARL能够在元测试阶段通过推断潜在变量来适应新任务,从而提高在未见过环境下的性能表现。它主要从元训练任务中学习,而不扩展到离线数据集。因此这种方法旨在确定利用离线数据集学习技能的有效性。初始化 $\alpha = 0.1$,并设置目标熵为 $H = -\dim(A)$ 。

PEARL-ft是在PEARL基础上用SAC继续在目标

任务上微调的模型。本文将Q函数和熵系数初始化为元训练阶段学习到的值。

最大熵强化学习(Soft Actor Critic, SAC)^[28]是经典的强化学习方法,学习从头开始解决目标任务,而不利用离线数据或元训练任务。专为连续动作空间设计,通过引入策略的随机性(熵)优化探索与利用的平衡。本文初始化 $\alpha = 0.1$,并设置目标熵为 $H = -\dim(A)$ 。

3.3 对比实验

在对比实验中,本文提供了两个方面的实验结果:定量性能和定性适应。具体来说,通过图3,本文展示了对未见过的目标任务的适应能力,并对其性能进行了定量评估。而图4则对各个方法在复杂迷宫导航领域进行了定性分析。此外,通过定性分析,本文还展示了SDMRL在复杂迷宫导航任务的适应能力。

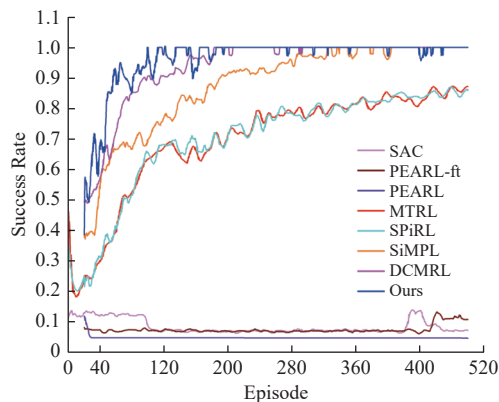


图3 样本效率的比较

Fig.3 Comparison of sample efficiency

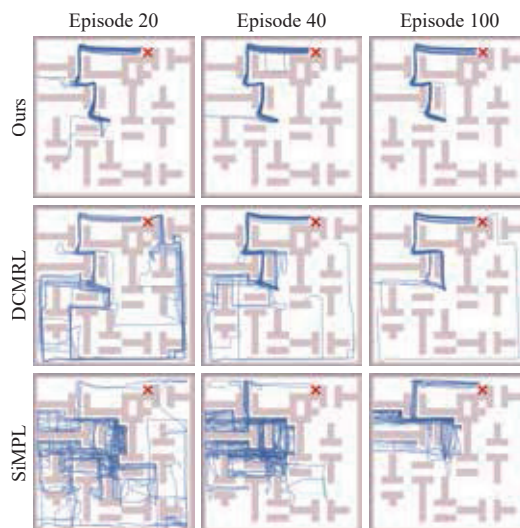


图4 适应过程的可视化

Fig.4 Visualization of the adaptation process

本文深入探讨了利用先验知识在强化学习方法

中对未见过的目标任务进行调整的影响。实验证明,若缺乏先验知识,SAC在适应性方面将受到较大限制,几乎无法适应元测试阶段中未见过的目标任务。尽管PEARL和PEARL-ft从元训练任务中学习任务上下文作为先验知识,但它们没有采用基于技能的方法,即使经过微调,仍然难以有效地适应到未见过的目标任务。相比之下,SPiRL利用从离线数据集中提取的一系列连续技能作为先验知识,为适应性提供了一定的帮助。SPiRL并不涉及元训练任务,与SDMRL的对比也能得知利用元训练任务进行进一步优化的有效性是十分显著的。另外,MTRL通过元训练任务集训练了一个共享策略,该策略能够一定程度上适应未见过的目标任务,展示了与SPiRL相媲美的适应性表现。SiMPL利用从离线数据集和元训练任务中分别提取的技能和任务上下文作为先验知识。然而,由于先验知识的泛化性受到限制,它的适应性仍然有限。DCMRL利用对比学习和VQ-VAE解耦了元训练阶段对任务上下文和技能的探索和学习,增强了对目标任务适应速度。然而,技能学习阶段利用含有行为克隆项的目标函数训练了低级策略,使得低级策略对于未见过的目标任务无法做出很好的调整。

总体来说,SDMRL在面对未见过的目标任务时表现出了明显的优势,尤其在适应速度方面。短短几十轮次就实现了策略的快速收敛,成功解决了迷宫环境中近95%的未见过的目标任务。进一步的适应性实验显示,通过继续训练和优化,SDMRL可以成功解决几乎100%的未见过的目标任务,彰显了其在面对未知挑战时的卓越表现。

SDMRL、DCMRL和SiMPL 3种方法均从离线数据集中学习到了有用的技能。如图4所示的可视化表明,3种方法均可以利用所学到的技能快速适应新任务,找到一条通往迷宫终点的最短路径。SDMRL在收敛速度上明显优于DCMRL和SiMPL,实现了更高的样本效率。

3.4 不同难度的对比实验

在本节中,本文比较了不同算法在不同难度的元训练任务上的学习效率。假设SDMRL能够从离线数据中提取到可重用的技能,相比于其他方法应该更适合学习长期任务。本文在迷宫导航环境中复杂度不断增加的3个元训练任务分布上进行比较:简单难度为方差较小的短期任务,中等难度为方差较大的短期任务,困难难度为方差较大的长期任务。通过增加每个任务分布中任务的方差和完成任务所需的时间步长,可以研究不同算法的学习能力。

如图5所示,在简单难度任务上,所有的方法都可以学习如何解决任务,本文的方法在样本效率方面优于其他对比方法。在中等难度任务上,除PEARL外的其他方法均能有效解决任务,且使用了离线数据的3种方法均优于MTRL,说明了利用离线数据学习技能的有效性。在困难难度任务上,只有DCMRL、SiMPL和本文方法可以有效解决任务。并且本文的方法能够更快速更有效地学习,突出了通过技能发现从离线数据中学习技能的重要性,而不是简单的行为克隆。该实验结果印证了本文的观点,即利用状态区分技能能够学习到更有用的技能,并对长期任务特别有益。

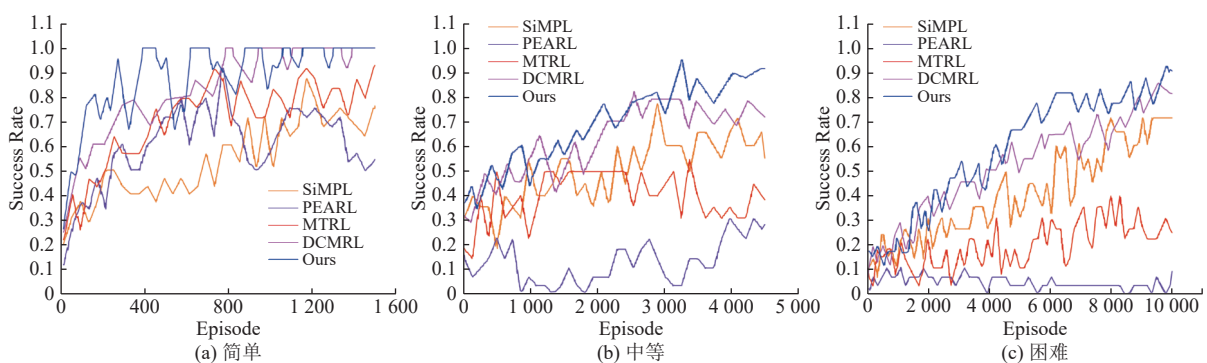


图5 不同难度的比较

Fig.5 Comparison of different difficulties

3.5 消融实验

在消融实验中,本文主要关注技能发现阶段。为了验证基于技能的SAC的有效性,本文在保持其他参数不变的情况下,基于完整的SDMRL方法构建了一个变体。这个变体在技能发现阶段不再采用基于技能的SAC训练,而是像DCMRL和SiMPL一样,通过行为克隆和技能嵌入正则化联合训练低级策略和技能提取器。

一个变体。这个变体在技能发现阶段不再采用基于技能的SAC训练,而是像DCMRL和SiMPL一样,通过行为克隆和技能嵌入正则化联合训练低级策略和技能提取器。

如图6所示,根据消融实验的结果显示,SDMRL和这个变体在迷宫环境中表现出相近的性能水平,给予足够轮次的情况下,均能够实现对未见过的目标任务的适应。然而,不使用基于技能SAC会导致样本效率的下降。SDMRL适应速度明显优于变体。具体而言,SDMRL仅需40~50步就能适应近90%的目标任务,而变体则需要90~100步才能达到相似的性能水平。这表明了基于技能SAC的有效性,使得SDMRL在学习效率和适应速度上具有显著提升,能够利用技能更好更快速地适应元测试阶段中未见过的目标任务。

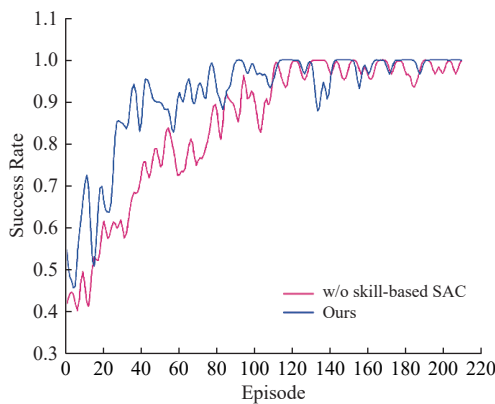


图6 消融实验

Fig.6 Ablation experiments

4 结论

本文提出了一种在没有任何奖励的情况下学习有用技能的元强化学习框架,它可以获得可识别的、有用的技能和更具有泛化性的技能先验,以实现对未见过的目标任务的有效适应。特别地,本文通过使用最大熵策略最大化信息论目标来学习技能,并实现了一个低级策略用于根据当前状态并利用技能做出一系列连续动作。在迷宫导航环境中,本文证明该目标可以无监督地学到能够完成任务的技能,尽管从未收到真正的任务奖励。研究表明,该方法可以作为一种有效的技能发现方法,以克服元强化学习中探索和数据效率的挑战。

参考文献:

- [1] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//International Conference on Machine Learning. Sydney: PMLR, 2017: 1126-1135.
- [2] RAKELLY K, ZHOU A, FINN C, *et al.* Efficient off-policy meta-reinforcement learning via probabilistic context variables[C]//International Conference on Machine Learning. California: PMLR, 2019: 5331-5340.
- [3] ZHAO T Z, LUO J, SUSHKOV O, *et al.* Offline meta-reinforcement learning for industrial insertion[C]//2022 International Conference on Robotics and Automation (ICRA). Philadelphia: IEEE, 2022: 6386-6393.
- [4] BAUMLI K, WARDE-FARLEY D, HANSEN S, *et al.* Relative variational intrinsic control[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Online: AAAI, 2021, 35(8): 6732-6740.
- [5] PERTSCH K, LEE Y, LIM J. Accelerating reinforcement learning with learned skill priors[C]//Conference on Robot Learning. Online: PMLR, 2021: 188-204.
- [6] NAM T, SUN S H, PERTSCH K, *et al.* Skill-based meta-reinforcement learning[C]//International Conference on Learning Representations. Vancouver: ICLR, 2022: 1-22.
- [7] GUO Y, WU Q, LEE H. Learning action translator for meta reinforcement learning on sparse-reward tasks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(6): 6792-6800.
- [8] ZHANG J, WANG J, HU H, *et al.* Metacure: meta reinforcement learning with empowerment-driven exploration[C]//International Conference on Machine Learning. Online: PMLR, 2021: 12600-12610.
- [9] EYSENBACH B, GUPTA A, IBARZ J, *et al.* Diversity is all you need: learning skills without a reward function[EB/OL]. arXiv: 1802.06070(2018-10-09) [2024-03-19]. <https://doi.org/10.48550/arXiv.1802.06070>.
- [10] CHRISTIANO P F, LEIKE J, BROWN T, *et al.* Deep reinforcement learning from human preferences[J]. Advances in Neural Information Processing Systems, 2017, 30: 4302-4310.
- [11] DYLAN H, SMITH A, PIETER A, *et al.* Inverse reward design[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach: Curran Associates, Inc., 2017: 6768-6777.
- [12] CAMPOS V, TROTT A, XIONG C M, *et al.* Explore, Discover and learn: unsupervised discovery of state-covering skills[C]//International Conference on Machine Learning. Long Beach: 2019: 1317-1327.
- [13] LIN S, WAN J L, XU T Y, *et al.* Model-based offline meta-reinforcement learning with regularization[C]//International Conference on Learning Representations. Kigali: ICLR, 2022: 1-22.
- [14] SIEGEL N Y, SPRINGENBERG J T, BERKENKAMP F, *et al.* Keep doing what worked: behavioral modelling priors for offline reinforcement learning[C]//International Conference on Learning Representations. Online: ICLR, 2020: 1-21.
- [15] SEO Y, LEE K, CLAVERA GILABERTE I, *et al.* Trajectory-wise multiple choice learning for dynamics generalization in reinforcement learning[J]. Advances in Neural In-

- formation Processing Systems, 2020, 33: 12968-12979.
- [16] DORFMAN R, SHENFELD I, TAMAR A. Offline meta reinforcement learning—identifiability challenges and effective data collection strategies[J]. Advances in Neural Information Processing Systems, 2021, 34: 4607-4618.
- [17] Mitchell E, Rafailov R, Peng X B, *et al.* Offline meta-reinforcement learning with advantage weighting[C]//International Conference on Machine Learning. Online: PMLR, 2021: 7780-7791.
- [18] PONG V H, NAIR A V, SMITH L M, *et al.* Offline meta-reinforcement learning with online self-supervision[C]//International Conference on Machine Learning. Baltimore: PMLR, 2022: 17811-17829.
- [19] LIU E Z, RAGHUNATHAN A, LIANG P, *et al.* Decoupling exploration and exploitation for meta-reinforcement learning without sacrifices[C]//International Conference on Machine Learning. Online: PMLR, 2021: 6925-6935.
- [20] LEE K, SEO Y, LEE S, *et al.* Context-aware dynamics model for generalization in model-based reinforcement learning[C]//International Conference on Machine Learning. Online: PMLR, 2020: 5757-5766.
- [21] ZHAO T Z, ANUSHA N, KATE R, *et al.* MELD - meta-reinforcement learning from images via latent state models[C]//Conference on Robot Learning. Online: PMLR, 2020: 1246-1261.
- [22] PERTSCH K, LEE Y, WU Y, *et al.* Guided reinforcement learning with learned skills[C]//Conference on Robot Learning. Auckland: PMLR, 2022: 729-739.
- [23] MEREL J, TUNYASUVUNAKOOL S, AHUJA A, *et al.* Catch & carry: reusable neural controllers for vision-guided whole-body tasks[J]. ACM Transactions on Graphics (TOG), 2020, 39(4): 39: 1-39: 12.
- [24] AJAY A, KUMAR A, AGRAWAL P, *et al.* OPAL: offline primitive discovery for accelerating offline reinforcement learning[C]//International Conference on Learning Representations. Online: ICLR, . 2021: 1-23.
- [25] PARK S, CHOI J, KIM J, *et al.* Lipschitz-constrained unsupervised skill discovery[C]//International Conference on Learning Representations. Online: ICLR, 2021: 1-24.
- [26] CHEBOTAR Y, HAUSMAN K, LU Y, *et al.* Actionable models: unsupervised offline reinforcement learning of robotic skills[C]//38th International Conference on Machine Learning. Online: ICML, 2021: 1518-1528.
- [27] BARBER D, AGAKOV F. The im algorithm: a variational approach to information maximization[J]. Advances in Neural Information Processing Systems, 2004, 16(320): 201.
- [28] 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. stockholm: PMLR, 2018: 1861-1870.
- [29] YUAN H, LU Z. Robust task representations for offline meta-reinforcement learning via contrastive learning[C]//International Conference on Machine Learning. Baltimore: PMLR, 2022: 25747-25759.
- [30] HE H C, ZHU A J, LIANG S, *et al.* Decoupling meta-reinforcement learning with Gaussian task contexts and skills[EB/OL]. arXiv: 2312.06518(2023-12-11) [2024-03-19]. <https://doi.org/10.48550/arXiv.2312.06518>.
- [31] FU J, KUMAR A, NACHUM O, *et al.* D4RL: datasets for deep data-driven reinforcement learning[EB/OL]. arXiv: 2004.07219(2021-02-06) [2024-04-25]. <https://doi.org/10.48550/arXiv.2004.07219>.
- [32] LUISA Z, KYRIACOS S, MAXIMILIAN I, *et al.* VariBAD: a very good method for bayes-adaptive deep RL via meta-learning[C]//Proceedings of the Eighth International Conference on Learning Representations. Online: ICLR, 2020: 1-20.
- [33] TEH Y, BAPST V, CZARNECKI W M, *et al.* Distral: robust multitask reinforcement learning[J]. Conference on Neural Information Processing Systems, 2017, 30: 4496-4506.

(责任编辑: 杨耀辉 英文审核: 费伦科)