计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(5):14–23 [doi: 10.15888/j.cnki.csa.009845] [CSTR: 32024.14.csa.009845]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

结合动态缓冲池和时间递减约束的离线到在线强化学习①

闫雷鸣[1,2], 朱永昕[1,2], 刘　健[1,2]

[1](南京信息工程大学 数字取证教育部工程研究中心, 南京 210044)
[2](南京信息工程大学 计算机、网络空间安全学院, 南京 210044)
通信作者: 闫雷鸣, E-mail: yan_leiming@163.com

摘　要: 离线到在线强化学习中, 虽然智能体能够通过预先收集的离线数据进行初步策略学习, 但在线微调阶段, 早期过程常常表现出不稳定性, 且微调结束后, 性能提升幅度较小. 针对这一问题, 提出了两种关键设计: 1) 模拟退火的动态离线-在线缓冲池; 2) 模拟退火的行为约束衰减. 第 1 种设计在训练过程中利用模拟退火思想动态选择离线数据或者在线交互经验, 获得优化的更新策略, 动态平衡在线训练的稳定性和微调性能; 第 2 种设计通过带降温机制的行为克隆约束, 改善微调早期使用在线经验更新导致的性能突降, 在微调后期逐渐放松约束, 促进模型性能提升. 实验结果表明, 所提出的结合动态缓冲池和时间递减约束的离线到在线强化学习 (dynamic replay buffer and time decaying constraints, DRB-TDC) 算法在 Halfcheetah、Hopper、Walker2d 这 3 个经典 MuJoCo 测试任务中, 在线微调训练后性能分别提升 45%、65%、21%, 所有任务的平均归一化得分比最优基线算法提升 10%.

关键词: 深度强化学习; 离线到在线强化学习; 模拟退火; 动态缓冲池; 行为克隆约束

引用格式: 闫雷鸣,朱永昕,刘健.结合动态缓冲池和时间递减约束的离线到在线强化学习.计算机系统应用,2025,34(5):14–23. http://www.c-s-a.org.cn/1003-3254/9845.html

Offline to Online Reinforcement Learning Combining Dynamic Replay Buffer and Time Decaying Constraint

YAN Lei-Ming[1,2], ZHU Yong-Xin[1,2], LIU Jian[1,2]

[1](Engineering Research Center of Digital Forensics Ministry of Education, Nanjing University of Information Science &Technology, Nanjing 210044, China)
[2](School of Computer Science and Cyber Science and Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: In offline-to-online reinforcement learning, though the agent can leverage pre-collected offline data for initial policy learning, the online fine-tuning phase often exhibits instability in the early stages, and the performance improvement after fine-tuning is relatively small. To address this issue, two key designs are proposed: 1) a simulated annealing-based dynamic offline-online replay buffer and 2) simulated annealing-based behavior constraint attenuation. The first design dynamically selects offline data or online interaction experiences during training using the simulated annealing concept to obtain an optimized update strategy, dynamically balancing the stability of online training and fine-tuning performance. The second design introduces a behavior cloning constraint with a cooling mechanism to mitigate the sharp performance drop caused by using online experience updates in the early fine-tuning stage, gradually relaxing the constraint in the later stage to enhance model performance. Experimental results demonstrate that the proposed dynamic replay buffer and time decaying constraints (DRB-TDC) algorithm improves performance by 45%, 65%, and 21% on the

① 基金项目: 国家自然科学基金 (62172292, 42375147)
收稿时间: 2024-10-22; 修改时间: 2024-11-19; 采用时间: 2024-12-04; csa 在线出版时间: 2025-03-24
CNKI 网络首发时间: 2025-03-25

Halfcheetah, Hopper, and Walker2d tasks from the MuJoCo benchmark after online fine-tuning, respectively. The average normalization score of all tasks exceeds the best baseline algorithm by 10%.

Key words: deep reinforcement learning; offline to online reinforcement learning; simulated annealing; dynamic replay buffer; behavior cloning constraint

近年来, 强化学习 (reinforcement learning, RL) 领域取得了显著进展[1,2], 并被广泛应用于各种复杂任务中, 例如机器人控制[3]、游戏 AI[4]、自动驾驶[5]以及医疗健康[6]等. 尤其在机器人控制中, 强化学习被用来优化机器人的行为策略, 提升整体的决策效率. 然而传统在线强化学习需要与环境进行大量交互才能取得良好效果, 但这会产生高昂的时间成本, 且在训练过程中可能存在一定的风险[7]. 例如, 在自动驾驶训练中, 智能体在学习早期可能因决策失误导致交通事故发生[8]; 在机器人控制中, 机器人在训练过程中可能会因为错误操作导致设备损坏[9].

而与在线强化学习不同, 离线强化学习[10]使用预先收集的离线数据进行训练, 避免智能体与环境实时交互. 但策略性能依赖于离线数据的质量, 当离线数据质量不高时, 直接将离线策略应用于在线环境时, 可能会导致策略性能大幅下降甚至失效[11]. 因此研究者们进一步提出一种结合离线训练和在线微调的范式, 即离线到在线强化学习 (offline to online reinforcement learning, O2O RL). 在这种范式下, 将离线策略应用于在线环境之前, 需要与在线环境交互进行微调训练, 使策略更加适用于在线环境. 然而, 离线到在线强化学习仍然面临巨大的挑战[12]: (1) 由于离线数据分布与在线数据分布存在差异, 离线策略在微调早期面临性能大幅度下降的问题; (2) 在微调过程中对策略更新施加强制约束, 可以部分缓解性能大幅下降, 但这些约束在一定程度上也限制了智能体探索更优策略, 使得性能提升幅度较小.

针对上述问题, 本文提出了模拟退火的动态离线-在线缓冲池和模拟退火的约束衰减. 利用模拟退火的优化思想, 引入能量函数来动态调整离线数据和在线经验的使用. 通过计算能量差值, 当策略性能提升时, 使用在线交互经验更新策略; 当策略表现不佳时, 以较高的概率选择离线数据更新策略, 避免策略性能继续下降. 随着微调过程的推进, 模拟退火算法中的温度参数逐步降低, 离线缓冲池的选择概率相应减少, 使得策略在后期更多依赖在线经验, 确保最终能够适应在线环境, 提升策略的整体性能.

同时针对微调早期直接使用在线经验导致的性能抖动, 将离线强化学习中的行为克隆约束与模拟退火中的降温机制相结合. 在微调初期, 设置较强的行为克隆约束, 防止价值估计出现较大偏差. 随着微调的进行, 逐步减弱行为克隆约束, 鼓励智能体探索更优策略, 避免微调后期策略性能难以提升.

本文的主要贡献如下.

(1) 提出了模拟退火的动态离线-在线缓冲池策略, 利用模拟退火思想动态调整离线数据和在线经验的使用概率, 提升微调后的策略性能.

(2) 提出了模拟退火的行为约束衰减, 将降温机制与行为克隆约束结合, 提升微调早期训练的稳定性, 防止后期性能提升停滞.

为了验证方法的有效性, 在连续控制任务 MuJoCo 中进行实验, 结果表明, 本文算法 DRB-TDC 在 Half-cheetah、Hopper、Walker2d 任务中, 微调结束后性能分别提升 45%、65%、21%, 所有任务的平均归一化得分比最优基线算法提升 10%.

本文第 1 节介绍离线到在线强化学习的相关研究. 第 2 节介绍本文提出的结合动态缓冲池和时间递减约束的离线到在线强化学习算法 DRB-TDC. 第 3 节通过对比实验验证了所提方法的有效性.第 4 节对本文工作进行总结.

1 相关研究

1.1 在线强化学习

在线强化学习的核心在于智能体与环境的持续交互, 智能体通过环境反馈不断更新策略以优化决策. Mnih 等[13]在 2015 年提出算法 DQN (deep Q network), 首次将深度卷积神经网络与 Q 学习相结合, 成功训练出在 Atari 游戏中超越人类专家水平的智能体. 在此之后许多经典强化学习算法相继涌现, 例如 TD3 (twin delayed deep deterministic policy gradient)[14]、SAC

(soft Actor-Critic)[15]、PPG (phasic policy gradient)[16]等. 此外, 通过引入蒙特卡洛方法增强 Actor-Critic 算法[17]、调整神经网络中的噪声类型[18]、引入好奇心机制[19]、动态平衡探索与利用[20]、定期重置神经网络[21]等技术, 强化学习方法得到了进一步的发展与完善.

在线强化学习虽然通过持续的交互训练能够取得良好的效果, 但由于训练过程中需要大量的交互, 这使其在一些存在交互风险的实际应用场景 (如机械臂控制) 中难以推广和应用.

1.2 离线强化学习

与在线强化学习不同, 离线强化学习旨在从固定的离线数据集中训练强化学习智能体. Fujimoto 等[22]首次提出了可以从任意固定批次数据中有效学习的离线强化学习算法 BQL (batch constrained Q-learning), 通过限制动作空间, 迫使智能体在给定数据的子集上学习近似于策略内的行为. Kostrikov 等[23]提出算法 IQL (implicit Q-learning), 不再直接评估最新策略中的未见动作, 而是将状态值函数视为一个随机变量, 采用该随机变量的状态条件上分位值来估计该状态下最佳动作的价值, 从而避免查询 Q 函数中的未见动作. Chen 等[24]提出了一种基于 Transformer 的方法, 通过对期望回报、过去状态和动作进行条件化, 从而直接生成最优动作, 克服了误差传播和价值过高估计导致的次优行为问题. Fujimoto 等[25]提出了方法 TD3BC, 在 TD3 算法的基础上添加行为克隆项作为策略更新时的约束, 并对数据进行归一化操作, 在降低训练时间的同时保证了算法的性能. Wu 等[26]在密度支持约束理论形式化的基础上提出了一种新的算法 SPOT, 利用变分自动编码器进行密度估计, 从而显式建模策略的支持集, 并引入基于密度的正则化项. Wang 等[27]在离线强化学习中引入了扩散 Q 学习, 利用条件扩散模型表示策略, 并在条件扩散模型的训练损失中添加最大化动作-价值的决策值函数.

尽管离线强化学习在利用数据集进行智能体训练方面已经取得显著进展, 但将训练好的策略从离线环境迁移到在线环境时, 可能出现性能大幅度下降甚至策略失效的问题, 这为离线强化学习的推广和应用提出了巨大的挑战.

1.3 离线到在线强化学习

在离线强化学习中, 智能体不能通过与环境交互迭代更新策略, 只能依赖给定数据集进行训练, 这导致离线训练形成的策略性能将受到离线数据集质量的限制. 此外, 虽然离线强化学习方法减少了在线交互的风险, 但离线训练出的策略在应用于真实环境时, 通常并不能取得令人满意的效果. 为了解决这一问题, 研究者们尝试将离线强化学习与在线强化学习进行结合.

Nair 等[28]提出了算法 AWAC (advantage weighted Actor-Critic), 通过使用优势函数回归的方法, 避免对行为策略的估计. Lee 等[29]提出了一种方法, 通过平衡重放和悲观 Q 集成, 在在线微调阶段优先考虑在线样本, 并结合离线数据集中接近策略的样本进行训练, 同时引入悲观 Q 集成方案, 从而缓解 Q 值高估的问题. Zhang 等[30]提出了算法 PEX (policy expansion), 在离线策略学习完后, 冻结该策略将其作为策略集中的一个候选策略. 同时通过在策略集中增加一个新的策略, 该策略通过复制离线策略获得. 在训练过程中只对新的策略进行微调, 在保证原有策略行为的同时能够学习新的能力. Nakamoto 等[31]通过学习一个保守的值函数初始化来实现从现有数据集中获得策略初始化, 该初始化低估离线数据中学习到的策略的价值, 同时确保学习到的 Q 值处于合理的范围内. Zheng 等[32]提出了算法 ODT (online decision Transformer), 一种基于序列建模的离线到在线强化学习算法, 将离线预训练与在线微调结合在一个统一的框架中, 在该框架中使用序列级熵正则化器, 并结合自回归建模目标, 以实现样本高效的探索和微调. Zheng 等[33]提出了一种自适应策略学习框架, 通过在离线数据集上采用悲观更新策略和在在线数据集上采用乐观更新策略的方式, 实现了离线到在线强化学习的最佳结合.

为了直观地理解在线、离线及离线到在线强化学习的优缺点, 如表 1 所示对三者进行了对比总结.

表 1 各类强化学习方法的优缺点

类别	优点	缺点
在线强化学习	通过交互实时更新策略, 适应环境变化	需要大量交互, 成本高, 同时存在安全风险
离线强化学习	能够从固定数据集中训练, 避免交互风险	性能受离线数据质量限制
离线到在线强化学习	既能利用离线数据又能通过在线交互微调策略	离线与在线部分结合不当可能导致策略退化甚至失效

2 结合动态缓冲池与约束衰减的在线微调

强化学习智能体在利用离线数据进行训练时, 往往存在性能上的局限, 因此需要通过与在线环境的交互进行微调来获得更好的效果. 然而, 在线微调过程中, 从离线数据直接过渡到在线环境时, 会面临显著的状态-动作分布差异.

为了更直观地揭示两者在状态分布上的差异, 利用离线强化学习算法 TD3BC 在数据集 Halfcheetah-medium-v2 上训练一个离线策略, 训练完成后, 利用该离线策略与环境交互并保存在线交互经验. 分别从离线数据与在线交互经验中采集相同数量的样本, 使用 t-SNE (t-distributed stochastic neighbor embedding) 将样本降维到二维空间进行可视化, 并观察离线数据和在线交互经验在状态分布上的差异.

如图 1 所示, 其中蓝点代表在线交互经验, 红点代表离线数据. 可以看到, 在图的右上角、左下角以及中间部分, 蓝点和红点的分布并未完全重叠, 这表明即使离线训练的策略与环境交互后产生的经验, 仍然与离线数据在状态分布上存在一定差异. 这种差异反映了离线数据在覆盖状态-动作空间时的局限性, 尤其是在某些状态区域, 离线数据无法有效涵盖. 而在线交互经验的引入能够弥补这些不足, 在一定程度上扩展离线数据的覆盖范围, 使得策略在更新过程中能够接触到更多的状态-动作组合.

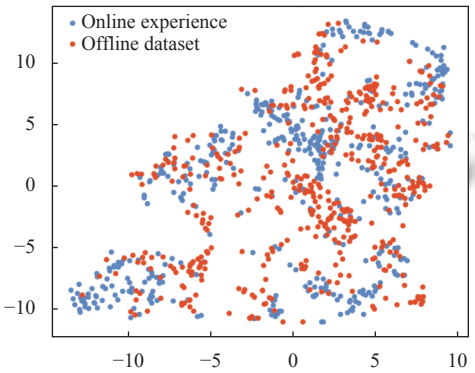


图 1 离线数据与在线交互经验分布

而强化学习在从离线数据训练到在线环境微调的过程中, 状态-动作分布的显著差异是导致性能不稳定甚至下降的根源. 这一问题源于离线数据覆盖范围的局限性, 离线策略在覆盖较少的状态区域时容易发生价值估计偏差, 在线交互经验的引入虽然能够扩展状态分布覆盖, 但其分布差异也可能进一步导致训练不稳定.

为了避免因分布差异导致原本离线训练良好的智能体出现性能大幅度下降的问题, 可以在微调时继续沿用离线训练的约束方式. 但这种做法会使策略逐渐趋向于行为策略, 导致微调过程中性能提升缓慢, 最终在微调结束时, 相较于离线训练的性能改进幅度较小. 因此针对上述问题, 提出了结合动态缓冲池和时间递减约束的离线到在线强化学习算法 DRB-TDC (dynamic replay buffer and time decaying constraints), 利用模拟退火的思想动态选择缓冲池并调节约束衰减的幅度.

2.1 模拟退火的动态离线-在线缓冲池

由于离线数据覆盖范围的局限性, 离线策略往往并不是最优策略. 为进一步提升离线策略的性能并有效利用离线数据和在线交互经验, 本文提出了一种基于模拟退火思想的动态离线-在线缓冲池策略. 模拟退火通过模仿物理系统中的退火过程, 在策略优化时逐渐降低系统温度减少随机性, 帮助策略从次优解逐步逼近全局最优解. 这样做的好处体现在以下两个方面.

(1) 离线数据与在线交互经验的动态平衡. 在微调早期使用在线交互经验更新策略导致性能下降时, 利用模拟退火方法, 使得策略更新时偏向使用离线数据, 从而提升训练的稳定性. 同时在后期逐渐偏向使用在线交互经验, 使策略更好适应环境, 突破离线数据的限制, 提升最终性能.

(2) 避免局部最优. 当策略性能下降时, 仍然存在一定的可能性继续选择在线交互经验更新策略, 避免陷入局部最优, 而随着温度的降低, 策略更新逐步收敛到更优解, 从而改善性能.

因此为了动态调整从离线数据与在线交互经验中采样的概率, 从而更加平衡地使用两种数据更新策略, 进一步提升策略性能. 利用模拟退火的思想, 通过能量函数的变化来控制离线缓冲池 B_{off} 与在线缓冲池 B_{on} 的选择, 能量函数 E 的定义如下:

$$E=\begin{cases} Q(s,\pi(s)), & B=B_{\text{on}} \\ Q(s,\pi(s))-(\pi(s)-a)^2, & B=B_{\text{off}} \end{cases} \tag{1}$$

其中, $\pi(s)$ 为策略 π, $Q(s,\pi(s))$ 为 Q 值函数, 表示策略 π 所选动作所得到的期望累积回报, $(\pi(s)-a)^2$ 为行为克隆约束项, 表示策略 π 所选动作与原数据集中实际动作 a 间的差异性.

在线微调开始时, 先从在线缓冲池 B_{on} 中采样更新

策略, 策略更新的公式如下:

$$\pi = \arg\max_{\pi} E_{(s,a)\sim B_{\text{on}}} Q(s,\pi(s)) \tag{2}$$

在策略更新时, 通过计算当前时刻与上一时刻的能量差值ΔE, 来决定下一步是从在线缓冲池还是离线缓冲池中采样:

$$\Delta E = E_t - E_{t-1} \tag{3}$$

如果ΔE大于 0, 即当前策略表现优于上一时刻, 则继续从在线缓冲池B_{on}中采样更新策略; 如果ΔE小于等于 0, 即当前策略表现不佳, 则按照的P_{off}概率选择离线缓冲池B_{off}进行策略更新, 策略更新的公式如下:

$$\pi = \arg\max_{\pi} E_{(s,a)\sim B_{\text{off}}} \left[Q(s,\pi(s)) - \alpha(\pi(s)-a)^2 \right] \tag{4}$$

其中, α为权重因子, 用于衡量 Q 值和动作选择差异的权衡.

选择离线缓冲池的概率P_{off}的计算公式如下:

$$P_{\text{off}} = \begin{cases} 0, & t = 0 \,\|\, \Delta E > 0 \\ \mathrm{e}^{-\frac{1}{\frac{|\Delta E|}{T(i)}}}, & t > 0 \,\|\, \Delta E \leqslant 0 \end{cases} \tag{5}$$

其中, 温度T随时间t递减, 公式如下:

$$T(t) = \beta T(t-1) \tag{6}$$

其中, 初始温度为$T(0)$, 取值为 1, β为温度衰减系数, 取值为 0.999 98.

ΔE越小, 即策略表现越差, 选择离线缓冲池B_{off}的概率P_{off}就越大. 这样做的好处在于, 在线微调的早期阶段, 由于离线数据与在线交互经验之间存在一定分布差异, 直接使用在线交互经验进行策略更新可能导致价值估计过高, 从而影响策略更新的稳定性. 当策略性能出现下降时, 通过使用离线数据进行策略更新, 可以在一定程度上减缓这种偏差对策略更新的影响.

随着微调过程的进行, 温度T逐渐降低, 此时选择从离线缓冲池B_{off}中采样的概率P_{off}会越来越低, 策略会更加依赖在线交互经验, 并逐步减少对离线数据的使用, 这样策略能够更好地通过与在线环境交互不断优化, 从而突破离线数据的局限性. 最终, 在微调结束时策略完全从在线缓冲池中采样更新, 确保策略充分适应在线环境.

2.2 模拟退火的行为约束衰减

尽管动态离线-在线缓冲池策略利用模拟退火想法在离线数据和在线交互经验之间实现了动态平衡, 显著提升了最终训练出的策略性能, 同时避免了微调过程中的性能出现大幅下降导致无法恢复的情况. 但在微调的早期阶段, 由于在线交互经验的引入, 训练过程仍然可能出现不稳定现象. 这是因为在线交互经验中往往包含与离线数据分布差异较大的状态-动作组合, 直接将这些经验用于策略更新时容易导致价值估计过高, 进而影响策略更新的稳定性.

为了提升使用在线交互经验更新策略的稳定性, 将离线强化学习算法 TD3BC 中的行为克隆约束与模拟退火中的降温机制相结合, 提出了模拟退火的约束衰减. 在微调早期阶段, 行为克隆约束较大, 以确保在面对新的在线交互经验时不会产生过高的价值估计. 随着策略逐渐适应在线环境, 逐步减小约束, 允许策略在后期更加灵活地利用在线交互经验, 从而提升整体策略性能. 从在线缓冲池B_{on}中抽取样本更新策略的公式如下:

$$\pi = \arg\max_{\pi} E_{(s,a)\sim B_{\text{on}}} \left[Q(s,\pi(s)) - f(t)(\pi(s)-a)^2 \right] \tag{7}$$

其中, 时间递减项$f(t)$为:

$$f(t) = \mathrm{e}^{-0.0001t} \tag{8}$$

此时, 能量函数E的公式如下:

$$E = \begin{cases} Q(s,\pi(s)) - f(t)(\pi(s)-a)^2, & B = B_{\text{on}} \\ Q(s,\pi(s)) - (\pi(s)-a)^2, & B = B_{\text{off}} \end{cases} \tag{9}$$

在微调的早期, 通过较强的约束可以有效缓解因在线交互经验带来的价值估计偏差, 确保训练稳定性. 随着训练过程的进行, 结合降温机制的约束逐渐减弱, 策略能够充分利用在线交互经验, 从而更好地适应在线环境.

2.3 DRB-TDC 伪代码

DRB-TDC 算法的主要流程如下.

(1) 输入参数初始化: 包括离线数据集、训练步数、批量大小等参数;

(2) 缓冲池初始化;

(3) 离线策略初始化;

(4) 在线训练: 包括与环境交互、样本选择、网络参数更新等.

DRB-TDC 的伪代码如算法 1 所示.

算法 1. 结合动态缓冲池和时间递减约束的离线到在线强化学习

输入: 离线数据集D (数据为(s,a,r,s')), 初始训练步数T_{initial}, 在线训练步数T_{on}, 批量大小B, 策略网络更新频率d, 离线策略π_{off}.

1) 初始化在线缓冲池 B_{on} 为空值, 利用 D 初始化离线缓冲池 B_{off}
2) 利用 π_{off} 初始化价值网络参数 θ_1、θ_2, 策略网络参数 φ 以及目标网络参数 θ_1'、θ_2'、φ'
3) for t=0 to $T_{on}+T_{initial}$ do
4) 与在线环境交互, 其中 $a\sim\pi_\varphi(s)+N(0,\sigma)$
5) 将收集的经验 (s,a,r,s') 存储至 B_{on}
6) if $t>T_{initial}$
7) if $\Delta E>0$
8) 从在线缓冲池 B_{on} 中随机选取批量大小为 B 的样本集合 (s,a,r,s')
9) else
10) 按照 $e^{-\frac{1}{\frac{|\Delta E|}{T(t)}}}$ 的概率从离线缓冲池 B_{off} 中随机选取批量大小为 B 的样本集合 (s,a,r,s')
11) end if
12) $a'=clip\left(\pi_{\varphi'}(s')+\varepsilon,-0.5,0.5\right)$
13) 使用 $\nabla_{\theta_i}\frac{1}{|B|}\sum_{(s,a)\sim B}\left(Q_{\theta_i}-r-\gamma\min_{i=1,2}Q_{\theta_i}(s',a')\right)^2$ 更新价值网络参数
14) if $t\%d$
15) if $\Delta E>0$
16) 使用 $\nabla_\varphi\frac{1}{|B|}\sum_{(s,a)\sim B}\left[Q_{\theta_1}(s,\pi_\varphi(s))-e^{-0.0001(t-T_{initial})}(\pi(s)-a)^2\right]$ 更新策略网络参数 φ
17) else
18) 使用 $\nabla_\varphi\frac{1}{|B|}\sum_{(s,a)\sim B}\left[Q_{\theta_1}(s,\pi_\varphi(s))-(\pi(s)-a)^2\right]$ 更新策略网络参数 φ
19) end if
20) 使用 $\theta_i'\leftarrow\tau\theta_i+(1-\tau)\theta_i'$、$\varphi_i'\leftarrow\tau\varphi_i+(1-\tau)\varphi_i'$ 更新目标网络的参数 θ_i、φ_i
21) end if
22) end if
23) end for

3 实验分析

3.1 实验环境介绍

OpenAI gym[34]作为 OpenAI 团队针对强化学习研发以及算法对比开发的工具包, 它提供了标准化的环境接口以及预定义的环境, 这些环境覆盖了从简单的控制问题到复杂的机器人模拟任务等多个方面, 使得开发者和研究人员能够直接进行仿真实验和测试. 如图 2 所示, 本文采用 OpenAI gym 的 MuJoCo 物理引擎中一系列连续控制任务作为实验环境.

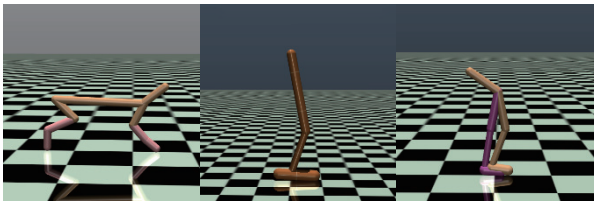


图 2 MuJoCo 环境

为了合理评估算法的性能, 本文选取了 MuJoCo 物理引擎中的 Halfcheetah-v2、Walker2d-v2 及 Hopper-v2 进行实验验证, 其相关介绍列于表 2 中.

表 2 实验任务介绍

任务	动作维度	状态维度	任务目标
Halfcheetah-v2	6	17	训练双足智能体学会行走
Walker2d-v2	6	17	训练单足智能体向前跳跃
Hopper-v2	3	11	训练双足智能体尽可能快地向前走

离线数据集来自流行的离线 RL 基准 D4RL[35], 该基准包含广泛的任务和数据集, 旨在测试智能体从各种环境中学习有效策略的能力. 本文主要选取的数据集主要包含以下内容.

(1) Medium: 数据来自中等策略, 其性能约为专家的 1/3.

(2) Medium-expert: 混合等量的专家策略和中等策略收集的数据.

(3) Medium-replay: 在线训练策略达到中等性能时, 训练期间缓冲区中的所有样本.

3.2 实验设置

为了确保算法的公平性和实验评估的一致性, 本文重新运行算法 AWAC、IQL 以及 PEX.

AWAC[28]: 使用优势加权的 Actor-Critic 的方法, 将最大化 Q 值的目标替换为最大化优势函数, 使策略模仿数据集中具有高优势估计的动作, 避免直接估计对行为策略, 同时通过限制策略分布保持接近更新时最新的数据来避免误差累计.

IQL[23]: 将状态价值函数视为随机变量来隐式近似策略改进步骤, 避免利用最新策略评估未见动作, 同时使用随机变量的上期望值来估计该状态下的最优动作价值, 是一种具有代表性质的离线强化学习算法, 同时可以直接进行在线微调.

PEX[30]: 引入策略扩展的方法初始化在线学习, 在学会离线策略后, 将其作为候选策略放入策略集中, 同时将复制离线策略得到一个新的策略, 在训练过程中只对新的策略进行微调, 保证原有离线策略不变, 同时这两个策略将以自适应的方式组合, 与环境进行交互.

每个算法均进行 100 万个时间步长的离线训练, 离线策略训练完成后, 再使用已训练完成的离线策略进行在线训练. 在线训练时, 采用 5 个不同的随机种子, 每个种子独立运行 25 万个时间步长, 每 5 000 个时间步长评估一次策略, 每次评估均包含 10 个不同的情节,

记录每个情节的累积奖赏均值作为该策略在当前时间步的性能指标. 所有实验均在配置为 Intel Xeon Silver 4214 CPU、NVIDIA GeForce RTX 3090 GPU 和 128 GB 内存的服务器上运行.

DRB-TDC 采用 3 层全连接层的线性神经网络作为行动者网络和评论家网络, 并使用 Adam 优化器更新神经网络参数. 行动者网络的输入为状态向量, 隐藏层每层包含 256 个神经元, 激活函数为 ReLU 函数, 最终输出动作向量并使用 tanh 函数作为激活函数. 评论家网络与行动者网络类似, 但不同的是输入为连接后的状态向量与动作向量, 输出是 Q 值. 其他超参数设置如表 3 所示.

3.3 实验结果

图 3 展示了 DRB-TDC 与基线方法在 MuJoCo 任务中微调过程中策略性能, 实线表示每个算法独立运行 5 次后在当前时间步评估所得的平均性能, 阴影部分代表 5 次独立运行的性能波动范围 (阴影上界为该时间步的最高性能, 阴影下界为最低性能). DRB-TDC 在微调结束时, 策略性能在 Halfcheetah 中提升 45%, 在 Hopper 任务中提升 65%, 在 Walker2d 任务中提升 21%. 在 Halfcheetah-medium 和 Halfcheetah-medium-replay 两个任务中, DRB-TDC 的性能提升远远优于其他基线算法, 分别优于最好基线算法性能的 20% 和 35%. DRB-TDC 仅在 Halfcheetah-medium-expert 以及 Hopper-medium-expert 两个任务中, 性能在微调结束时略低于 AWAC, 但也取得了与基线算法相似的性能.

表 3 DRB-TDC 超参数设置

超参数	取值	参数描述
γ	0.99	折扣因子
l	0.000 3	网络学习率
o	0.2	动作噪声
B	256	训练策略样本批量大小
τ	0.005	软更新常数

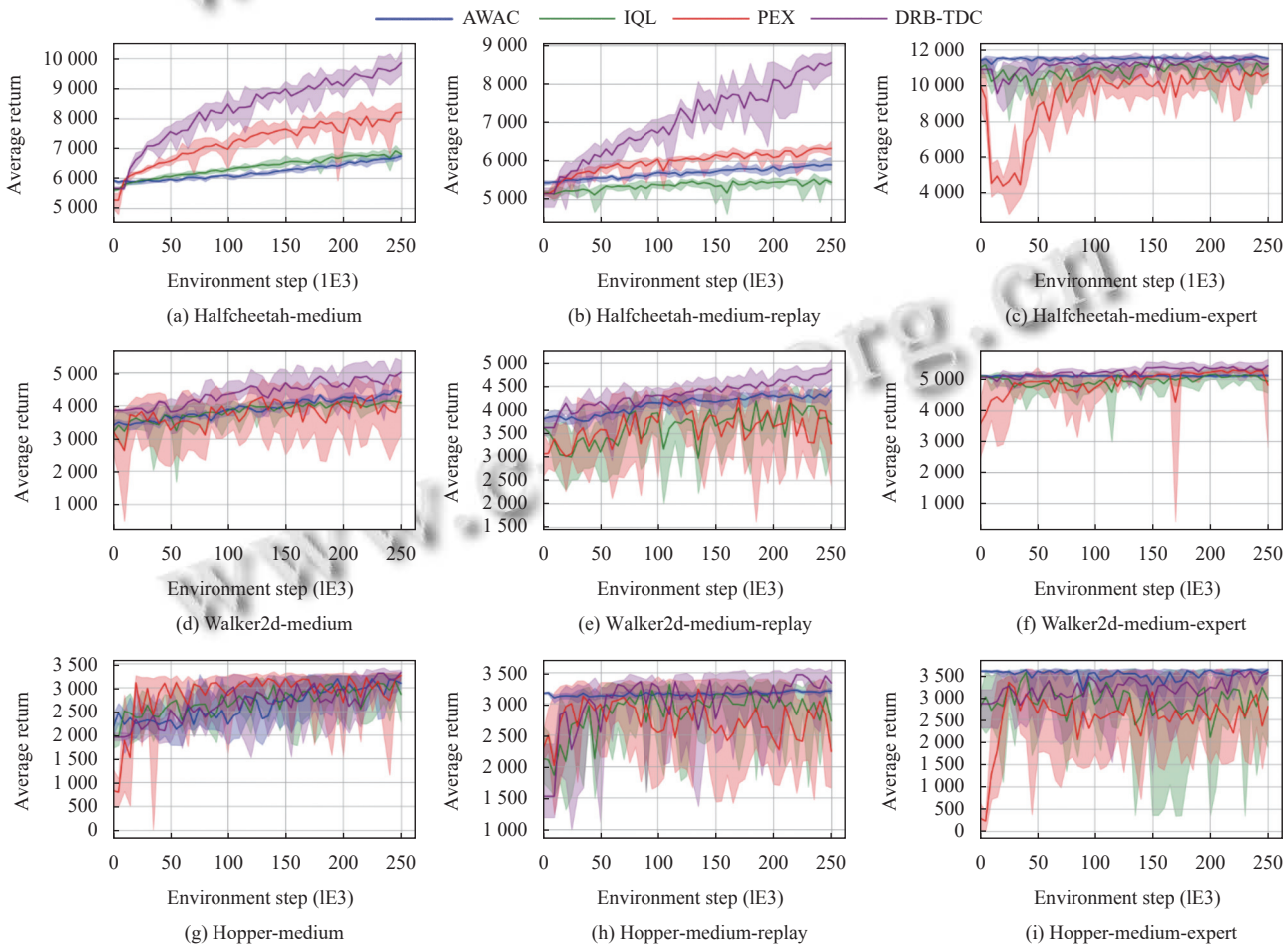


图 3 在线微调过程的策略性能评估

此外在图 3 中可以看到, DRB-TDC 的训练过程比 IQL 以及 PEX 更加稳定, 仅次于 AWAC. 因为 AWAC 算法本身是使策略模仿数据集中具有高优势估计的动作, 不会做出偏离状态-动作空间较大的行为, 因此训练过程中相对稳定, 但是也可以看到 AWAC 在微调过程中策略性能提升缓慢, 微调结束时在 Halfcheetah 任务中提升 7%, 在 Hopper 任务中提升 16%, 在 Walker2d 任务中提升 15%. DRB-TDC 的性能提升远高于 AWAC, 在保持训练过程相对稳定的情况下, 能够有效提升微调过程中的策略性能.

为了对比训练结束时策略表现的稳定性, 表 4 给出了 DRB-TDC 与 3 个基线算法在 5 次独立训练结束时, 在 9 个不同任务上策略经过 10 次评估所得的平均归一化分数及其标准偏差. 结果显示, 9 个对比任务中, DRB-TDC 获得了 7 个任务的最高分. 在 Halfcheetah-medium-expert 任务中比 AWAC 低 2.4, 以及在 Hopper-medium-expert 任务中比 AWAC 低 1.9, 在其余 7 个任务中性能均优于基线算法, 且 10 次评估的标准差也相对较小, 策略性能稳定.

表 4 策略评估的平均归一化得分和标准差

数据集	AWAC[28]	IQL[23]	PEX[30]	DRB-TDC (本文)
Halfcheetah-medium	56.6±0.4	57.1±0.7	68.4±1.7	**81.7±2.0**
Halfcheetah-medium-replay	49.8±0.9	46.1±0.3	53.2±1.0	**71.1±1.8**
Halfcheetah-medium-expert	**95.1±0.5**	91.8±2.5	88.3±3.0	92.7±1.1
Walker2d-medium	96.1±1.2	89.1±3.0	93.8±13.2	**109.3±4.0**
Walker2d-medium-replay	96.0±2.0	80.5±7.3	71.5±19.8	**105.7±3.7**
Walker2d-medium-expert	111.5±0.5	107.6±5.3	104.9±22.0	**118.5±3.7**
Hopper-medium	95.5±3.3	88.4±12.4	100.5±1.3	**102.5±1.4**
Hopper-medium-replay	99.1±1.0	84.3±10.6	69.4±9.2	**103.1±6.0**
Hopper-medium-expert	**112.0±0.5**	91.5±18.5	86.6±15.9	110.1±1.6

图 4 为各算法在不同数据集上的平均归一化得分累计结果. 可以看到 DRB-TDC 整体累计得分显著高于其他算法, 相较于其他算法的得分提升约为 10%. 进一步证明了 DRB-TDC 在离线到在线强化学习中的优势.

3.4 消融实验

为了验证 DRB-TDC 算法中各个组件对最终策略性能的贡献, 分别对 DRB 和 TDC 进行消融验证. 通过对比完整算法与消融版本的性能, 可以直观地看到每个组件的作用和影响.

当删除 DRB 组件时, 策略更新时随机从在线或离线缓冲池等概率抽取样本进行训练, 保持时间递减约束机制不变. 通过这种设置, 如图 5, 可以观察到由于保留了约束, 在微调早期策略性能不再出现明显下降, 但直至微调结束时, 由于策略更新时仍然会选择从离线缓冲池采样, 并没有充分利用在线交互经验, 导致训练结束时, 相较于 DRB-TDC, 策略性能提升幅度相对较小.

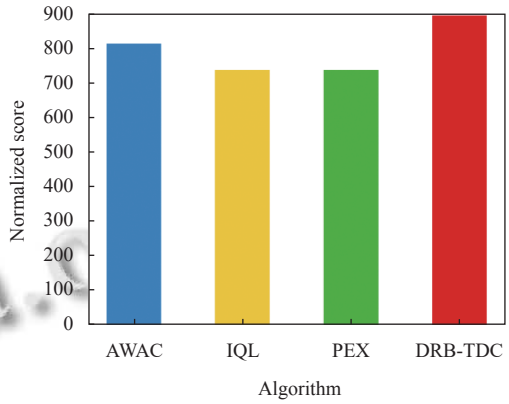
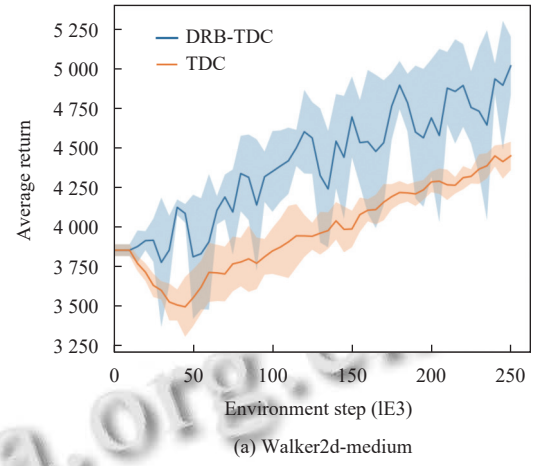
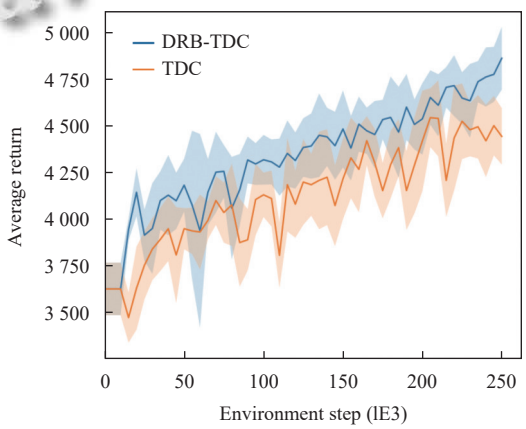


图 4 所有任务的平均归一化得分总和



(a) Walker2d-medium

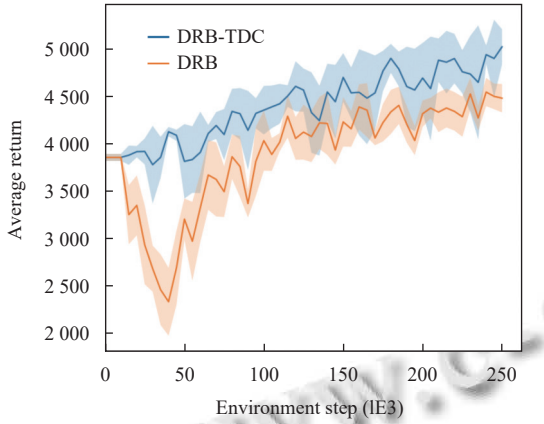


(b) Walker2d-medium-replay

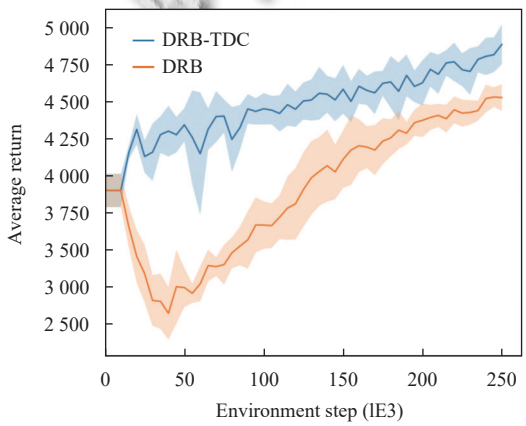
图 5 DRB 组件的消融实验

当删除 TDC 组件时, 在线微调过程中, 从在线缓冲池采样更新策略时不再添加约束条件. 通过这种设

置, 如图 6, 可以观察到尽管 DRB 组件能够帮助策略更好地利用在线交互经验, 保证微调后期能够尽可能提升策略性能. 但由于在微调早期使用在线交互经验时缺乏约束, 策略性能会出现较大波动, 导致训练结束时, 相较于 DRB-TDC 而言策略性能提升幅度相对较小.



(a) Walker2d-medium



(b) Walker2d-medium-replay

图 6 TDC 组件的消融实验

4 结论与展望

本文提出了结合动态缓冲池和时间递减约束的离线到在线强化学习算法 DRB-TDC. 通过引入模拟退火算法, 动态调整离线数据与在线交互经验在策略微调过程中的采样概率, 从而更加有效地利用两种数据. 同时在微调早期, 当策略性能不佳时, 及时增加使用离线数据更新策略的概率, 在一定程度上提升了训练过程的稳定性. 随着微调过程的进行, 逐渐降低对离线数据的依赖, 增大对在线交互经验的使用, 使得策略在微调结束时更加适应在线环境, 性能能够获得较大提升. 同时结合降温机制的行为克隆约束的引入, 使得在线微调早期策略训练更加稳定, 同时随着微调的进行, 逐渐降低约束幅度, 避免策略性能提升出现停滞. 本文选取 3 个连续控制任务 (Halfcheetah、Walker2d、Hopper) 中的不同数据集 (medium、medium-replay、medium-expert) 来验证算法的有效性. 实验结果表明, 本文提出的算法在微调结束时, 分别提升 45%、65%、21%. 在 7 个任务中表现均优于基线算法, 仅在 Halfcheetah-medium 和 Halfcheetah-medium-replay 两个任务中略低于基线方法. 而策略评估的平均归一化分数相较于基线方法提升约为 10%. 同时通过消融实验, 证实了本文所提出的两个组件对于 DRB-TDC 的有效性.

尽管 DRB-TDC 在一定程度上提升了策略微调后的性能同时避免微调早期性能的大幅度下降, 但是在微调过程中尤其是微调后期, 在增大使用在线交互经验的同时大幅减少约束幅度, 会导致在微调过程出现训练波动的情况. 在未来的工作中, 将会把研究重心着眼于该问题上, 尝试采用集成的方法改善训练波动的情况, 在不过度增加训练开销的前提下, 尽可能提升微调后期的稳定性, 确保策略性能稳定提升.

参考文献

1 Wang X, Wang S, Liang XX, *et al*. Deep reinforcement learning: A survey. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(4): 5064–5078.
2 曹宏业, 刘潇, 董绍康, 等. 面向强化学习的可解释性研究综述. 计算机学报, 2024, 47(8): 1853–1882.
3 赵静, 裴子楠, 姜斌, 等. 基于深度强化学习的无人机虚拟管道视觉避障. 自动化学报, 2024, 50(11): 2245–2258.
4 Wei H, Chen JX, Ji XY, *et al*. Honor of kings arena: An environment for generalization in competitive reinforcement learning. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 863.
5 Al-Sharman M, Dempster R, Daoud MA, *et al*. Self-learned autonomous driving at unsignalized intersections: A hierarchical reinforced learning approach for feasible decision-making. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(11): 12345–12356.
6 Lee HY, Chung S, Hyeon D, *et al*. Reinforcement learning model for optimizing dexmedetomidine dosing to prevent delirium in critically ill patients. npj Digital Medicine, 2024, 7(1): 325. [doi: 10.1038/s41746-024-01335-x]
7 王雪松, 王荣荣, 程玉虎. 基于表征学习的离线强化学习方法研究综述. 自动化学报, 2024, 50(6): 1104–1128.
8 何逸煦, 林泓熠, 刘洋, 等. 强化学习在自动驾驶技术中的应用与挑战. 同济大学学报 (自然科学版), 2024, 52(4):

520–531.
9 王雪松, 王荣荣, 程玉虎. 安全强化学习综述. 自动化学报, 2023, 49(9): 1813–1835.
10 Prudencio RF, Maximo MROA, Colombini EL. A survey on offline reinforcement learning: Taxonomy, review, and open problems. IEEE Transactions on Neural Networks and Learning Systems, 2023, 35(8): 10237–10257.
11 乌兰, 刘全, 黄志刚, 等. 离线强化学习研究综述. 计算机学报, 2025, 48(1): 156–187.
12 Zhao K, Ma Y, Liu J, *et al*. Improving offline-to-online reinforcement learning with Q-ensembles. Proceedings of the 2023 ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems. OpenReview.net, 2023.
13 Mnih V, Kavukcuoglu K, Silver D, *et al*. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529–533. [doi: 10.1038/nature14236]
14 Fujimoto S, Hoof H, Meger D. Addressing function approximation error in Actor-Critic methods. Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018. 1587–1596.
15 Haarnoja T, Zhou A, Abbeel P, *et al*. Soft Actor-Critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018. 1861–1870.
16 Cobbe KW, Hilton J, Klimov O, *et al*. Phasic policy gradient. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 2020–2027.
17 Wilcox A, Balakrishna A, Dedieu J, *et al*. Monte Carlo augmented Actor-Critic for sparse reward deep reinforcement learning from suboptimal demonstrations. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 164.
18 Eberhard O, Hollenstein J, Pinneri C, *et al*. Pink noise is all you need: Colored noise exploration in deep reinforcement learning. Proceedings of the 11th International Conference on Learning Representations. OpenReview.net, 2023.
19 Pathak D, Agrawal P, Efros AA, *et al*. Curiosity-driven exploration by self-supervised prediction. Proceedings of the 34th International Conference on MachineLearning, Sydney: PMLR, 2017. 2778–2787.
20 Moskovitz T, Parker-Holder J, Pacchiano A, *et al*. Tactical optimism and pessimism for deep reinforcement learning. Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021. 984.
21 Kim W, Shin Y, Park J, *et al*. Sample-efficient and safe deep reinforcement learning via reset deep ensemble agents. Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 2317.
22 Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 2052–2062.
23 Kostrikov I, Nair A, Levine S. Offline reinforcement learning with implicit Q-learning. arXiv:2110.06169, 2021
24 Chen LL, Lu K, Rajeswaran A, *et al*. Decision Transformer: Reinforcement learning via sequence modeling. Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021. 1156.
25 Fujimoto S, Gu SS. A minimalist approach to offline reinforcement learning. Proceedings of the 35th Conference on Neural Information Processing Systems. NeurIPS, 2021. 20132–20145.
26 Wu JL, Wu HX, Qiu ZH, *et al*. Supported policy optimization for offline reinforcement learning. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 2268.
27 Wang Z, Hunt JJ, Zhou M. Diffusion policies as an expressive policy class for offline reinforcement learning. arXiv:2208.06193, 2022.
28 Nair A, Gupta A, Dalal M, *et al*. AWAC: Accelerating online reinforcement learning with offline datasets. arXiv:2006.09359, 2020.
29 Lee S, Seo Y, Lee K, *et al*. Offline-to-online reinforcement learning via balanced replay and pessimistic Q-ensemble. Proceedings of the 5th Conference on Robot Learning. London: PMLR, 2021. 1702–1712.
30 Zhang H, Xu W, Yu H. Policy expansion for bridging offline-to-online reinforcement learning. arXiv:2302.00935, 2023.
31 Nakamoto M, Zhai YX, Singh A, *et al*. Cal-QL: Calibrated offline RL pre-training for efficient online fine-tuning. Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 2719.
32 Zheng QQ, Zhang A, Grover A. Online decision Transformer. Proceedings of the 39th International Conference on Machine Learning. Baltimore: PMLR, 2022. 27042–27059.
33 Zheng H, Luo XF, Wei PF, *et al*. Adaptive policy learning for offline-to-online reinforcement learning. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI, 2023. 11372–11380.
34 Brockman G, Cheung V, Pettersson L, *et al*. OpenAI gym. arXiv:1606.01540, 2016.
35 Fu J, Kumar A, Nachum O, *et al*. D4RL: Datasets for deep data-driven reinforcement learning. arXiv:2004.07219, 2020.

(校对责编: 王欣欣)