基于优先采样模型的离线强化学习

顾 扬 [1] 程玉虎 [1] 王雪松 [1]

摘 要 离线强化学习通过减小分布偏移实现了习得策略向行为策略的逼近, 但离线经验缓存的数据分布往往会直接影响习得策略的质量. 通过优化采样模型来改善强化学习智能体的训练效果, 提出两种离线优先采样模型: 基于时序差分误差的采样模型和基于鞅的采样模型. 基于时序差分误差的采样模型可以使智能体更多地学习值估计不准确的经验数据, 通过估计更准确的值函数来应对可能出现的分布外状态. 基于鞅的采样模型可以使智能体更多地学习对策略优化有利的正样本, 减少负样本对值函数迭代的影响. 进一步, 将所提离线优先采样模型分别与批约束深度 Q 学习 (Batch-constrained deep Q-learning, BCQ) 相结合, 提出基于时序差分误差的优先 BCQ 和基于鞅的优先 BCQ. D4RL 和 Torcs 数据集上的实验结果表明: 所提离线优先采样模型可以有针对性地选择有利于值函数估计或策略优化的经验数据, 获得更高的回报.

关键词 离线强化学习, 优先采样模型, 时序差分误差, 鞅, 批约束深度 Q 学习

引用格式 顾扬, 程玉虎, 王雪松. 基于优先采样模型的离线强化学习. 自动化学报, 2024, **50**(1): 143−153

DOI 10.16383/j.aas.c230019

Offline Reinforcement Learning Based on Prioritized Sampling Model

GU Yang[1] CHENG Yu-Hu[1] WANG Xue-Song[1]

Abstract Offline reinforcement learning algorithms realize the approximation of learned policy to behavior policy by reducing the distribution shift, but the data distribution of offline experience buffer often directly affects the quality of learned policy. In this paper, two offline prioritized sampling models including temporal difference error-based and martingale-based are proposed to improve the training effect of reinforcement learning agent. The temporal difference error-based sampling model enables agents to learn more experience data with inaccurate value estimation, thus deals with possible out-of-distribution states by estimating more accurate value functions. The martingale-based sampling model enables agents to learn more positive samples beneficial to policy optimization and reduces the impact of negative samples on value function iteration. Furthermore, the proposed offline prioritized sampling models are combined with the batch-constrained deep Q-learning (BCQ) respectively, to propose temporal difference error-based prioritized BCQ and martingale-based prioritized BCQ. Experimental results on D4RL and Torcs datasets show that the proposed two offline prioritized sampling models can be targeted to select the experience data that are conducive to value function estimation or policy optimization, so as to obtain higher rewards.

Key words Offline reinforcement learning, prioritized sampling model, temporal difference error, martingale, batch-constrained deep Q-learning (BCQ)

Citation Gu Yang, Cheng Yu-Hu, Wang Xue-Song. Offline reinforcement learning based on prioritized sampling model. *Acta Automatica Sinica*, 2024, **50**(1): 143−153

由于兼具了强化学习优良的决策能力以及深度学习强大的表征能力和泛化性能, 深度强化学习已成为解决复杂环境下感知决策问题的一个可行方案[1]. 近年来, 深度强化学习已经在机器人控制[2]、电力系统优化[3]、网络安全[4]、视频游戏[5−6]、医疗健康[7]、自动驾驶[8−9] 等领域取得了成功应用.

随着深度强化学习理论和方法的发展, 学者们尝试开发智能体去处理一些数据采集困难, 对硬件设备安全构成威胁的学习任务[10]. 2020 年之前, 参考机器学习中批量学习的方法, 学者们提出了一种无需进行探索、经验缓存固定的深度强化学习, 并命名为批强化学习[11]. 2020 年后, 随着批强化学习热度的提升, Levine 等[10] 将此类算法重新命名为离线强化学习. 离线强化学习有着行为策略下固定大小的经验缓存, 可以避免在线探索带来的环境噪声和危险行为[12]. 一方面, 离线强化学习可以从在线强化学习的经典算法中汲取灵感[13], 有较长远的发展前景. 另一方面, 离线强化学习中, 大部分算法通

收稿日期 2023-01-13 录用日期 2023-04-04
Manuscript received January 13, 2023; accepted April 4, 2023
国家自然科学基金 (62176259, 62373364), 江苏省重点研发计划项目 (BE2022095) 资助
Supported by National Natural Science Foundation of China (62176259, 62373364) and Key Research and Development Program of Jiangsu Province (BE2022095)
本文责任编委 杨涛
Recommended by Associate Editor YANG Tao
1. 中国矿业大学信息与控制工程学院 徐州 221116
1. School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116

过引入模仿学习[14] 来减小分布偏移, 降低了强化学习与其他机器学习方法之间的壁垒. 但一个值得关注的问题是: 习得策略下, 智能体对离线经验缓存分布之外的 (Out-of-distribution, OOD) 状态评估会包含误差, 从而表现并不理想.

针对这一问题, 研究者们提出了许多解决方案. Fujimoto 等[15] 率先提出了第一个能够从任意批数据 (离线数据) 中学习而无需探索的批约束深度 Q 学习 (Batch-constrained deep Q-learning, BCQ). BCQ 采用 Q 学习技术, 在选取最大化 Q 值对应的动作时, 希望只考虑实际出现在离线数据集中的状态–动作对, 而不考虑分布外的动作. 为此, Kumar 等[16] 利用变分自编码器来生成与离线数据集分布相近的动作, 并结合一个扰动网络模型对生成的动作进行调优, 从而使动作具有多样性. 测试阶段, 在生成的动作空间中选择使 Q 值最大的那些动作. 由于 BCQ 不涉及对未知状态–动作对的考虑, 因此不会在策略与值函数上引入额外的偏差, 同时, 动作与值函数分开学习, 也避免了误差累积. 然而, Kumar 等[16] 指出: 由于 BCQ 对策略施加的约束较强, 因此当离线数据集质量较差时, BCQ 只能有限地改善策略性能. 进一步, Kumar 等[16] 分析了分布偏移导致的自举误差, 提出了使用两个独立值函数结构的自举误差累积消减算法 (Bootstrapping error accumulation reduction, BEAR), 利用支持集匹配的思想来防止自举误差累积. 此外, BEAR 通过约束当前策略与行为策略之间的最大均值差异 (Maximum mean discrepancy, MMD)[17] 来使习得策略尽可能接近行为策略以缓解分布偏移问题. 然而, 由于需要计算 MMD 距离, BEAR 的计算代价较大. Jaques 等[18] 通过减小习得策略和行为策略之间的 KL 散度, 使学习到的策略逼近行为策略. 与之类似, Maran 等[19] 使用 Wasserstein 距离来描述策略间差异, 将减小策略分布间的 Wasserstein 距离作为正则化项添加到优化目标中. 为评估不同行为策略正则化项的重要性, Wu 等[20] 引入一个通用的算法框架, 称为行为正则化 Actor-Critic. 该框架涵盖了 BCQ、BEAR 等, 同时提供了多种实际选择方案, 使研究人员能够以模块化的方式比较不同变体的性能. 进一步, Wu 等[20] 提出两类正则化方法: BRAC-v 与 BRAC-p, 前者是对值函数进行正则化, 后者则是对策略进行正则化. 值得注意的是, 值函数正则化虽然可以提高 OOD 状态评估的准确程度, 但也会在值函数更新过程中增加噪声, 使习得策略难以收敛. 策略正则化虽然能有效降低分布偏移且提高习得策略的稳定性, 但会增大习得策略陷入局部最优的概率.

上述离线强化学习方法都倾向于通过降低分布偏移来提高习得策略的质量, 但忽视了离线数据集质量对离线强化学习性能的影响. 类似的, 在在线强化学习方法中, 经验的好坏对智能体的训练起到非常重要的作用. 因此, 如何让智能体高效地选择样本也是提高强化学习算法性能的一个有效措施. Schaul 等[21] 在在线强化学习 (深度 Q 网络) 中采用了优先经验回放技术, 主要思路为: 通过时序差分 (Temporal difference, TD) 误差估计经验池 (经验缓存区) 中样本的重要程度并赋予样本不同的优先级, 使那些在训练过程中对智能体更加重要的样本更容易被选择. Horgan 等[22] 在优先经验回放技术的基础上提出了分布式经验池的思想, 进一步提升了强化学习智能体在复杂环境中的表现.

离线经验缓存的质量主要会通过以下两个方面来影响离线强化学习的训练: 1) 行为策略下生成的离线经验缓存中会包含折扣回报低于平均水平的失误经验, 这些经验所占比例往往不高. 因此, 训练过程中智能体容易忽视失误经验, 无法在对应的场景下做出最优的行为. 2) 离线经验缓存中的样本根据其是否有利于策略优化可以分为正样本与负样本, 负样本更多的存在于失误经验集合中, 过多采样负样本进行训练会导致习得策略的质量不理想. 于是, 参考在线强化学习采用的优先经验回放技术, 离线强化学习也需要通过优化采样模型来改善强化学习智能体的训练效果, 从而提高习得策略的质量. 为此, 本文提出两种离线优先采样模型: 1) 基于时序差分误差的采样模型, 可以提高值函数的估计精度, 有效地应对可能出现的 OOD 状态. 2) 基于鞅的采样模型, 可以对经验数据进行筛选, 使智能体自主地优先学习对策略优化有利的正样本. 进一步, 将这两种采样模型与 BCQ 相结合, 提出基于时序差分误差的优先 BCQ (TD-PBCQ) 和基于鞅的优先 BCQ (M-PBCQ). D4RL 和 Torcs 数据集上的实验结果表明: 1) TD-PBCQ 适用于行为策略基本收敛, 且离线经验缓存中包含少量失误经验的离线强化学习任务. 2) M-PBCQ 适用于离线经验缓存中包含较多失误经验的离线强化学习任务.

1 批约束深度 Q 学习

为提高离策略深度强化学习算法在离线强化学习场景下的工作效果, Fujimoto 等[15] 通过构建编码器网络和扰动网络来生成更好的策略, 提出了批约束深度 Q 学习. 在 BCQ 中, 编码器网络和扰动网络输出的动作可表示为状态到动作的映射 μ^{BCQ}:

$$\mu^{\mathrm{BCQ}}(s)=\underset{a_i+\xi_\phi(s,a_i,\Phi)}{\arg\max}\ Q_\theta\left(s,\ a_i+\xi_\phi\left(s,\ a_i,\ \Phi\right)\right) \tag{1}$$

其中, ξ_ϕ 为扰动网络, ϕ 为扰动网络的参数, Φ 为最大扰动参数, $\{a_i\sim VAE_\omega(s)\}_{i=1}^n$. VAE_ω 为变分自编码器 (Variational auto-encoder, VAE), 它可以根据固定的经验缓存来建模潜在的状态−动作空间, 在离线经验动作邻域内生成 n 个随机动作, 尽可能地最大化累积回报. 变分自编码器模型由 $E_{\omega 1}$ 和 $D_{\omega 2}$ 组成, 前者用于估计状态−动作服从的分布参数 $\{\hat{\mu},\ \hat{\sigma}\}=E_{\omega 1}(s,\ a)$, 后者用于估计期望的动作 $\tilde{a}=D_{\omega 2}(s,\ z),\ z\sim\mathcal{N}(\hat{\mu},\ \hat{\sigma})$. 变分自编码器模型的目标函数为:

$$\begin{aligned}\mathcal{L}_\omega=&\ \arg\min_\omega\sum\left(a-\tilde{a}\right)^2+\\&D_{\mathrm{KL}}\left(\mathcal{N}\left(\hat{\mu},\ \hat{\sigma}\right)\|\mathcal{N}\left(0,\ 1\right)\right)\end{aligned} \tag{2}$$

通过变分自编码器 VAE_ω 和扰动网络 ξ_ϕ, BCQ 可以在不与环境进行交互的限制条件下, 遍历到一个受限域区间内的多个动作, 因此 BCQ 有概率学习到比行为策略更好的策略. 在值函数更新部分, BCQ 使用了两个 Q 值网络 Q_{θ_1} 和 Q_{θ_2} 来降低过估计误差, 其目标值的计算方法为:

$$\begin{aligned}y_{\mathcal{B}}=r+\gamma\max_{a_i}\Big[&\lambda\min_{j=1,\,2}Q_{\theta_j'}\left(s',\ a_i\right)+\\&(1-\lambda)\max_{j=1,\,2}Q_{\theta_j'}\left(s',\ a_i\right)\Big]\end{aligned} \tag{3}$$

其中, λ 为在区间 $(0,\ 1)$ 取值的参数, 可以通过选择不同的 λ 来调节未来时间步不确定性给值函数更新带来的影响. 当 $\Phi=0$ 且 $n=1$ 时, BCQ 会退化为行为克隆算法, 机械地学习所有离线经验数据. 当 Φ 趋向于动作的上下限且 $n\to\infty$ 时, BCQ 等价于在线 Q 学习, 会产生较大的外推误差. BCQ 通过在线强化学习使值函数估计逼近最优值函数, 通过行为克隆算法减小测试时 OOD 状态出现的概率.

BCQ 算法定义了外推误差, 主要用于描述强化学习算法由于经验数据不足导致的估计误差. 在离线强化学习场景下应用离策略算法, 离线值函数 $Q_{\mathcal{B}}^\pi$ 和在线值函数 Q^π 之间的差异 $\delta_{\mathrm{MDP}}^\pi$ 为:

$$\begin{aligned}\delta_{\mathrm{MDP}}^\pi:=&\sum_s P_\pi(s)\times\\&\sum_a\pi\left(a\,|s\right)|Q^\pi\left(s,\ a\right)-Q_{\mathcal{B}}^\pi\left(s,\ a\right)|\end{aligned} \tag{4}$$

其中, $P_\pi(s)$ 为策略 π 下遍历到状态 s 的概率.

2　基于时序差分误差的采样模型

假设离线经验缓存为 $\mathcal{B}$, 其中包含的样本数为 $\mathcal{M}$, 对应的行为策略为 $\pi_{\mathcal{B}}$. 行为克隆 (Behavior clone, BC) 可以高效地学习 $\mathcal{B}$ 中状态到动作的映射, 但 $\mathcal{B}$ 中经验数据相关性较高, BC 的训练很容易过拟合, 因此训练得到的策略鲁棒性很差. 与行为克隆算法相比, 离线强化学习算法的样本效率虽然不高, 但会根据经验数据学习状态值等指标来评价状态和动作的好坏. 这些指标可以帮助智能体在访问 OOD 状态时做出合理的动作, 因此离线深度强化学习习得策略的鲁棒性更高. 但是, 离线深度强化学习仍面临着这样一个问题: 经验数据分布不理想会导致学习过程中产生累积误差.

假设离线数据集中存在两类状态 s^+ 和 s^-, 其中状态 s^- 对应的经验即为失误经验. 离线经验缓存 $\mathcal{B}$ 中 s^+ 被采样的概率越大, 意味着 s^+ 有更高的概率被采样, 由 s^+ 计算得到的损失会主导模型的训练, 离线强化学习算法对 s^+ 的状态值的估计越准确. 如果 s^- 被采样的概率很小, 由失误经验计算得到的梯度很容易被忽略, 进而导致智能体无法在状态 s^- 做出正确的行为. 因此, 增强对状态 s^- 的学习有利于逼近真实的策略评价指标.

对于优先经验回放 (Prioritized experience replay, PER) 来说, 样本的采样概率定义为[21]:

$$P(\upsilon):=\frac{p^o(\upsilon)}{\sum\limits_{i=0}^{N}p^o(i)} \tag{5}$$

其中, υ 为 $(s_\upsilon,\ a_\upsilon,\ r\left(s_\upsilon,\ a_\upsilon\right),\ s_\upsilon')$ 对应的经验数据, $p(\upsilon)$ 为经验数据 υ 对应的优先级. o 为指数参数, 用于决定优先级使用的程度. 如果取 $o=0$, 则采样模型在 $\mathcal{B}$ 中均匀采样. 我们考虑将优先经验回放引入离线强化学习算法中, 并命名为基于时序差分误差的采样模型.

在基于时序差分误差的采样模型中, $p(\upsilon)=|\delta_\upsilon|+\sigma$, σ 为优先级修正系数, 用来避免优先级为 0 的经验被采样的概率为 0. 如果使用一步更新的 Q 学习算法, 则 $\mathcal{B}$ 中经验数据 υ 对应的 TD 误差 δ_υ 为:

$$\begin{aligned}\delta_\upsilon=&\ r\left(s_\upsilon,\ a_\upsilon\right)+\\&\gamma\max_{\left(s_\upsilon',\,a'\right)\sim\mathcal{B}}Q\left(s_\upsilon',\ a'\right)-Q\left(s_\upsilon,\ a_\upsilon\right)\end{aligned} \tag{6}$$

由于离线经验缓存的数据分布是固定的, 离线经验优先级的计算比在线场景下的确定性更强. 离线训练中, PER 会使智能体更多地关注失误经验, 减少信息的浪费. 然而, 如果失误经验中包含较多的负样本, PER 反而会增大负样本的采样概率, 阻碍策略的优化.

3 基于鞅的采样模型

3.1 基于鞅的经验数据评估

鞅论是现代概率论的一个重要内容, 也是随机过程和数理统计研究的重要工具. 实际上, 在强化学习算法的发展过程中, 鞅论和强化学习之间一直存在着很深的联系, 很多鞅论的方法被用于理论证明强化学习算法的有效性. 例如, Mandl[23] 找到了有限控制 Markov 过程中存在的鞅过程. Hernández-Lerma 和 Ozak[24] 研究了离散 Markov 过程, 并给出了策略优化的等价命题, 其中研究的很多值迭代过程与鞅有关. Even-Dar 和 Mansour[25] 使用 Azuma 不等式来约束鞅的变化偏差, 估计值函数在某更新步完成优化的概率, 进而估计策略优化所需的时间. Hu 等[26] 使用杜布分解来简化下鞅过程, 使得复杂系统更容易被智能体学习. Chow 等[27] 利用上鞅收敛性来确保 Lyapunov 函数的收敛, 并用于求解约束 MDP 问题. 为此, 本文尝试通过分析采样数据对应的轨迹是否为下鞅来推断经验数据是否有利于策略优化.

定理 1. 假设存在一个持续性任务 (不会终止), 如果 $\mathrm{E}[r(s_{t+1})|s_t] = r(s_t)$, 则有 $\mathrm{E}[V(s_{t+1})|s_t] = V(s_t)$.

证明. 由强化学习值函数的迭代式可以得出 $V(i) = r(i) + \gamma \sum_{j\in S} P(j|i)V(j)$. 一般地, 由于 $|V(i)| \le V_{\max} < \infty$, 则有:

$$\begin{aligned}\mathrm{E}[V(s_{t+1})|s_t] &= \mathrm{E}[r(s_{t+1})|s_t] + \\ &\quad \gamma \sum_{j\in S}\sum_{s_{t+1}\in S} P(j|s_{t+1})P(s_{t+1}|s_t)V(j) = \\ &\quad \mathrm{E}[r(s_{t+1})|s_t] + \gamma \sum_{j\in S} P(j|s_t)V(j) = \\ &\quad \mathrm{E}[r(s_{t+1})|s_t] - r(s_t) + V(s_t) \end{aligned} \tag{7}$$

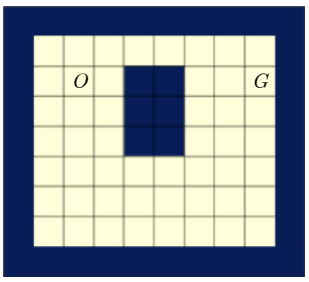
进一步, 可以得出

$$\begin{aligned}&\mathrm{E}[V(s_{t+1})|s_t] - V(s_t) = \\ &\quad \mathrm{E}[r(s_{t+1})|s_t] - r(s_t)\end{aligned} \tag{8}$$

由此可以得出: 如果 $\mathrm{E}[r(s_{t+1})|s_t] = r(s_t)$, 则有 $\mathrm{E}[V(s_{t+1})|s_t] = V(s_t)$. □

通过定理 1 可以看出: 如果回报函数为鞅, 即 $\mathrm{E}[r(s_{t+1})|s_t] = r(s_t)$, 说明经验对应的路径和值函数更新过程都为鞅. 由停时定理和鞅的一致收敛性可知, 对任意停时 $T < \infty$, 总有 $\mathrm{E}[V_T] = \mathrm{E}[V_0]$. 也就是说, 此时无论训练多少步, 值函数的期望都不会发生变化. 当且仅当 $\mathrm{E}[V(s_{t+1})|s_t] > V(s_t)$ 时, 值函数更新才满足强化学习的策略优化条件. 于是, 可以通过估计 $\mathrm{E}[V(s_{t+1})|s_t]$ 与 $V(s_t)$ 之间的大小差异来评估经验数据对策略优化的有利程度.

为了更好地解释鞅与策略优化之间的关系, 以格子世界环境为例加以阐述. 如图 1(a) 所示环境示意图, 智能体从 O 出发, 到达目标 G 终止一个情节. 如图 1(b) 所示最优值函数热图, 由于到达 G 点情节被终止, 因此 G 点的状态值并不会迭代更新, 导致其数值较小. 本次实验使用基于线性函数逼近的 Q 学习在迷宫中训练 300 个迭代步, 每隔 50 次迭代绘制一张值函数热图. 共进行了两个批次的训练, 值函数迭代更新过程如图 2 所示. 图 2 中, 相比于训练批次 2, 训练批次 1 的值函数明显更趋近于最优值函数. 于是, 可以得出如下观点:



(a) 环境示意图
(a) Schematic diagram of the environment

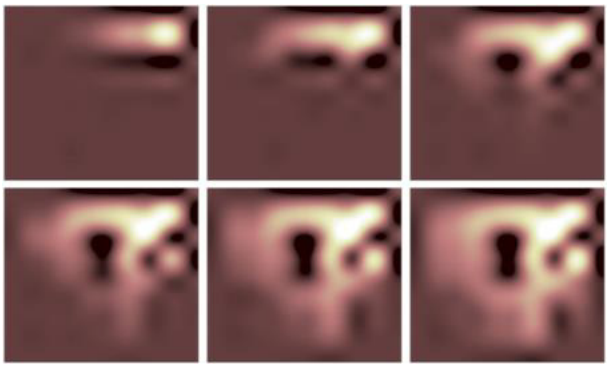
(b) 最优值函数热图
(b) Heatmap of the optimal value function

图 1 格子世界实验图
Fig. 1 Experimental diagram of grid-world

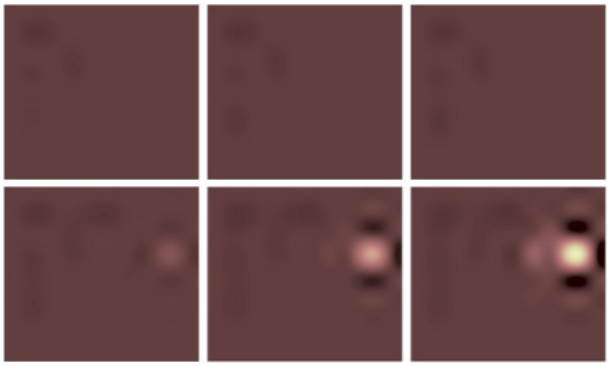
1) 图 2 中每一个像素点 s 的亮度用于描述对应状态值 $V(s)$ 的大小. 如果热图中像素点 s' 比 s 的亮度高, 则说明 $V(s') > V(s)$.

2) 值函数的更新会按照被访问的先后顺序 $s \to s'$, 从亮点逐级反向传播, 即有效的值函数更新从满足 $\mathrm{E}[V(s')|s] > V(s)$ 的状态 s 开始. 如图 2 所示, 批次 1 中满足 $\mathrm{E}[V(s')|s] > V(s)$ 的状态数量明显高于批次 2 中的状态数量. 因此, 经验缓存中, 满足 $\mathrm{E}[V(s')|s] > V(s)$ 的经验数据占比越高, 越有利于值函数的学习.

3) 如图 2(b) 所示, 前 150 次迭代没有亮点出现, 值函数热图维持不变. 因此, 如果状态值满足



(a) 训练批次 1
(a) Training batch 1



(b) 训练批次 2
(b) Training batch 2

图 2　值函数更新热图

Fig. 2　Heatmap of value function updating

$\mathrm{E}[V(s')|s] \leq V(s)$, 值函数优化效率很低. 150 次迭代后, 批次 2 的热图中虽然出现了亮点, 但亮度十分有限. 说明训练批次 2 的经验缓存中, 满足 $\mathrm{E}[V(s')|s] \leq V(s)$ 的经验数据占比较高, 从而会产生累积误差, 不利于值函数的学习.

综上所述, 我们认为经验缓存中包含越多符合 $\mathrm{E}[V(s')|s] > V(s)$ 的经验数据越有利于值函数和策略的优化, 这一观点在离线强化学习场景中同样适用.

3.2　基于鞅的采样模型

由于负样本会一直存在于离线经验缓存 $\mathcal{B}$ 中, 其对习得策略的不良影响会随着重复采样而增强. 于是, 为减少对负样本的采样频率, 提出基于鞅的采样模型. 设策略 π 下状态–动作对 (s, a) 被采样的概率为 $P^{\pi}_{\mathcal{B}}(s, a, \mathcal{B}, \pi) = P_{\pi}(s)\,\pi(a|s)$. 由于强化学习会贪心地选择动作, 因此状态 s 下选择不同动作的概率差异一般会较大, 可以得到推论 1.

推论 1. 在离线强化学习场景下, 均匀采样学习得到的策略有概率不为离线经验中的最优策略.

证明. 离线强化学习场景下, 将折扣回报值记为 $\mathcal{R}_{\mathcal{B}}(s, a)$, 根据 $\mathcal{B}$ 中数据统计得到的平均值为 $\bar{\mathcal{R}}_{\mathcal{B}}(s, a)$. 某个状态–动作对 (s, a) 被采样的概率为 $P^{\pi}_{\mathcal{B}}(s, a, \mathcal{B}, \pi)$. 设在相同状态 s 下, 存在两个不同的状态–动作对 (s, a_m) 和 (s, a_n) 满足 $\bar{\mathcal{R}}_{\mathcal{B}}(s, a_m) > \bar{\mathcal{R}}_{\mathcal{B}}(s, a_n)$, 即从 $\mathcal{B}$ 中观测的结果看 (s, a_m) 所在的路径是优于 (s, a_n) 的. 最终, 训练得到的 Q 值函数为:

$$Q_{\mathcal{B}}(s, a) = \mathrm{E}[\mathcal{R}_{\mathcal{B}}(s, a)] = P^{\pi}_{\mathcal{B}}(s, a, \mathcal{B}, \pi)\,\bar{\mathcal{R}}_{\mathcal{B}}(s, a) \tag{9}$$

于是可以得出: 如果 $P^{\pi}_{\mathcal{B}}(s, a_m, \mathcal{B}, \pi) < P^{\pi}_{\mathcal{B}}(s, a_n, \mathcal{B}, \pi)\bar{\mathcal{R}}_{\mathcal{B}}(s, a_n)/\bar{\mathcal{R}}_{\mathcal{B}}(s, a_m)$, 习得策略反而会倾向于选择 a_n. □

推论 1 说明: 离线经验缓存中如果折扣回报低的经验数据占比很高, 则离线强化学习算法就有高概率陷入局部最优.

根据第 3.1 节的描述可知, 如果 (s, a, r, s') 对应的轨迹为下鞅, 则认为 (s, a, r, s') 更有利于策略的优化. 如果 (s, a, r, s') 对应的轨迹为鞅或上鞅, 则频繁地采样 (s, a, r, s') 以更新网络参数反而会出现如图 2(b) 一样的误差累积状况, 从而阻碍值函数的优化. 为此, 可以考虑基于 $\mathrm{E}[V(s_{t+1})|s_t]$ 与 $V(s_t)$ 之间的数值差异来设计一种样本评估方法, 得到下述推论.

推论 2. 经验数据有利于值函数优化的程度与鞅差 $\mathrm{E}[V(s')|s] - V(s)$ 正相关.

证明. 设在第 k 个迭代步, 值函数优化的幅度为 $\Delta V_k := V_{k+1}(s) - V_k(s)$, 使用期望状态值来计算目标值, 则有:

$$\Delta V_k = r(s) + \gamma \mathrm{E}[V_k(s')|s] - V_k(s) \tag{10}$$

由于同一状态下即时回报 $r(s)$ 是一个常数, 且 γ 大于 0, 因此得到:

$$\Delta V_k \propto \mathrm{E}[V_k(s')|s] - V_k(s) \tag{11}$$

如果 ΔV_k 很大, 则说明当前的状态值过于低估了数据 (s, a, r, s'), 优先学习这个数据可以让值函数找到优化的方向, 并可在此基础上更准确地判断其他数据的 ΔV, 使得整个策略向着一个好的方向发展. 反之, 如果 ΔV_k 很小, 则说明数据 (s, a, r, s') 所在的过程更可能是上鞅, 此时状态值会随更新迭代变小或维持原样, 不利于策略的优化. □

推论 2 表明在值函数的优化过程中, 应当着重学习 $\mathrm{E}[V(s')|s] - V(s)$ 数值较高的数据 (s, a, r, s'), 并降低对数值过低数据的采样频率. 在实际训练过程中, 鉴于增加额外的网络用于学习 $\mathrm{E}[V(s')|s]$ 和 $V(s)$ 会比较耗时, 此处考虑使用一种近似的简便计算方法来求取基于鞅的优先级.

推论 3. 对于数据 $\upsilon(s_\upsilon, a_\upsilon, r(s_\upsilon, a_\upsilon), s'_\upsilon)$, 基于鞅的优先级为:

$$u_{\upsilon} := \mathrm{E}_{\mathcal{B}}\left[V\left(s_{\upsilon}'\right)\right] - r\left(s_{\upsilon}, a_{\upsilon}\right) \tag{12}$$

证明. 对于离线强化学习来说, 其状态值迭代公式为:

$$V_{\mathcal{B}}(s) = \sum_{s,\,a} P\left(a \,|s\,, \mathcal{B}\right) Q(s,\,a) = r(s,\,a) + \gamma \mathrm{E}_{\mathcal{B}}\left[V\left(s'\right)\right] \tag{13}$$

对应地, 有利于值函数优化的程度可以表征为:

$$\tilde{u} = \mathrm{E}_{\mathcal{B}}\left[V\left(s'\right)\right] - V_{\mathcal{B}}(s) = \frac{1}{\beta}\mathrm{E}_{\mathcal{B}}\left[V\left(s'\right)\right] - r(s,\,a) \tag{14}$$

其中, $\beta = 1/(1-\gamma)$ 为大于0、小于1 的常数, 离线经验数据对训练的有利程度与 $\mathrm{E}_{\mathcal{B}}[V(s')]/\beta - r(s,\,a)$ 的大小正相关. 当值函数估计存在误差时, $\mathrm{E}_{\mathcal{B}}[V(s')]/\beta - r(s,\,a)$ 的值会很小, 使得优先级差异不大, 难以区分. 可以进一步推导出:

$$u = \mathrm{E}_{\mathcal{B}}\left[V\left(s'\right)\right] - r(s,\,a) \propto \tilde{u} \tag{15}$$

因此, 推论 3 成立. □

综上所述, 基于鞅的采样模型使用基于鞅的优先级来决定数据 υ 被采样的概率:

$$p(\upsilon) = \begin{cases} -\sigma, & u_{\upsilon} = 0 \\ u_{\upsilon}, & \text{其他} \end{cases} \tag{16}$$

其中, σ 为优先级修正系数, 用于避免样本的采样概率完全为 0.

4 基于离线优先采样模型的 BCQ

将基于时序差分误差的采样模型和基于鞅的采样模型分别与 BCQ 相结合, 得到两种离线强化学习方法: TD-PBCQ 和 M-PBCQ. 为表述方便, 算法 1 给出 BCQ 的伪代码.

算法 1. BCQ

输入. 采样批次大小 m, 上限训练次数 T, 软更新参数 τ, 步长 η, 最大扰动参数 Φ, 目标值计算参数 λ, 采样动作个数 n, 离线经验回放 $\mathcal{B}$.

初始化优先回放经验 $\mathcal{H} = \emptyset$, $\Delta_{\theta} = 0, \Delta_{\phi} = 0$, $p_1 = 1$, $\mathcal{B}$ 中共有 $\mathcal{M}$ 个经验数据. 将离线经验 $\mathcal{B}$ 中的数据存入 $\mathcal{H}$, 并赋予最大优先级 $p_t = \max_{i<t} p_i$. 初始化 Q 网络 Q_{θ_1}、Q_{θ_2}, 扰动网络 ξ_{ϕ}, VAE 网络 $VAE_{\omega} = \{E_{\omega_1}, D_{\omega_2}\}$, 目标网络 $Q_{\theta_1'}$、$Q_{\theta_2'}$、$\xi_{\phi'}$, 参数更新 $\theta_1' \leftarrow \theta_1$, $\theta_2' \leftarrow \theta_2$, $\phi' \leftarrow \phi$.

for 迭代次数 $= 1 : T$ **do**

从经验池 $\mathcal{B}$ 随机采样 m 个数据

$\{\hat{\mu}, \hat{\sigma}\} = E_{\omega_1}(s,\,a)$, $\tilde{a} = D_{\omega_2}(s,\,z)$, $z \sim \mathcal{N}(\hat{\mu}, \hat{\sigma})$

$\omega \leftarrow \arg\min_{\omega} \sum (a - \tilde{a})^2 + D_{\mathrm{KL}}\left(\mathcal{N}(\hat{\mu}, \hat{\sigma}) \| \mathcal{N}(0,\,1)\right)$

for $j = 1 : \mathcal{Y}$ **do**

从经验池 $\mathcal{H}$ 采样数据 $(s,\,a,\,r,\,s') \sim P(j) = \frac{p_j}{\sum_i p_i}$

采样 n 个动作 $\left\{a_{j,\,k}' \sim G_{\omega}\left(s'\right)\right\}_{k=1}^{n}$

扰动动作 $\left\{a_{j,\,k}' = a_{j,\,k}' + \xi_{\phi}\left(s',\,a_{j,\,k}',\,\Phi\right)\right\}_{k=1}^{n}$

计算目标值

$y_{\mathcal{B}} = r\left(s_j,\,a_j\right) + \gamma \max_{a_{j,\,k}'}[\lambda \min_{i=1,\,2} Q_{\theta_i}\left(s_j',\,a_{j,\,k}'\right) + (1-\lambda) \max_{i=1,\,2} Q_{\theta_i}(s_j',\,a_{j,\,k}')]$

1) 计算重要性采样权重

2) 更新优先级

3) 累积 Q 值网络参数变化 Δ_{θ} 和扰动网络参数变化 Δ_{ϕ}

end for

更新 Q 值网络参数并清空权重 Δ_{θ}

$\theta_i' \leftarrow \tau\left(\theta + \eta \cdot \Delta_{\theta}\right) + (1-\tau)\,\theta_i'\,|_{i=1,\,2},\ \Delta_{\theta} = 0$

更新扰动网络参数并清空权重 Δ_{ϕ}

$\phi' \leftarrow \tau\left(\phi + \eta \cdot \Delta_{\phi}\right) + (1-\tau)\,\phi',\ \Delta_{\phi} = 0$

end for

4.1 基于 TD 误差的优先批约束 Q 学习

TD-PBCQ 通过变分自编码器生成 n 个动作, 并根据这些动作进行目标值的计算和网络的优化. 考虑到目标值中会包含一定的扰动, 优先级 δ 改写为:

$$\delta_j = y_{\mathcal{B}} - \frac{1}{2}\left(Q_{\theta_1}(s,\,a) + Q_{\theta_2}(s,\,a)\right) \tag{17}$$

将 BCQ 伪代码中的步骤 1)、2)、3) 替换为算法 2 中的步骤, 即可得到 TD-PBCQ 的伪代码.

算法 2. TD-PBCQ

1) 计算重要性采样权重: $w_j = \left(\frac{1}{\mathcal{M}} \cdot \frac{1}{P(j)} \cdot \frac{1}{\max_i w_i}\right)$

2) 更新优先级: 按照式 (17) 计算 TD 误差 δ_j, 更新数据的优先级 $p_j \leftarrow |\delta_j| + \sigma$

3) 累积 Q 值网络参数变化 Δ_{θ} 和扰动网络参数变化 Δ_{ϕ}:

累积 Q 值网络参数变化

$\Delta_{\theta} \leftarrow \Delta_{\theta} + w_j \nabla_{\theta}(y_{\mathcal{B}} - Q_{\theta}\left(s_j,\,a_j\right))^2$

累积扰动网络参数变化

$\Delta_{\phi} \leftarrow \Delta_{\phi} - w_j \nabla_{\phi} Q_{\theta_1}\left(s_j,\,a_j + \xi_{\phi}\left(s_j,\,a_j,\,\Phi\right)\right)$,

$a_j \sim G_{\omega}(s)$

4.2 基于鞅的优先批约束 Q 学习

由式 (12) 可以看出, 基于鞅的采样模型需要计算 $\mathrm{E}[V(s')]$. 由于扰动网络会生成置信区间内的 n 个动作, 如果使用贪心策略, 则可以认为 $\mathrm{E}[V(s')] = \max_{a_i} Q(s',\,a_i)$. 因此, 将这些状态−动作对应的 Q 值取平均作为期望状态值 $\mathrm{E}[V(s')]$, 使得对基于鞅的优先级评估更加保守. 为此, 基于鞅的优先级可改写为:

$$u_j = \frac{1}{n}\sum_{i=1}^{n}\min\left(Q_{\theta_1}\left(s', a_i\right), Q_{\theta_2}\left(s', a_i\right)\right) - r\left(s, a\right) \tag{18}$$

将 BCQ 伪代码中的步骤 1)、2)、3) 替换为算法 3 中的步骤, 即可得到 M-PBCQ 的伪代码.

算法 3. M-PBCQ

1) 计算重要性采样权重: M-PBCQ 不计算重要性采样权重

2) 更新优先级: 根据式 (18) 计算优先级 u_j, 根据式 (16) 更新经验数据的优先级

3) 累积 Q 值网络参数变化 Δ_θ 和扰动网络参数变化 Δ_ϕ:

累积 Q 值网络参数变化

$\Delta_\theta \leftarrow \Delta_\theta + \nabla_\theta(y_{\mathcal{B}} - Q_\theta\left(s_j, a_j\right))^2$

累积扰动网络参数变化

$\Delta_\phi \leftarrow \Delta_\phi - \nabla_\phi Q_{\theta_1}\left(s_j, a_j + \xi_\phi\left(s_j, a_j, \Phi\right)\right)$,

$a_j \sim G_\omega\left(s\right)$

5　实验结果与分析

首先, 将 TD-PBCQ、M-PBCQ 和 BCQ 在 D4RL 提供的公用离线数据集上, 针对 Ant、HalfCheetah、Hopper、Walker2d 等任务在中等 (medium) 和专家 (expert) 数据集上进行实验. 然后, 将 TD-PBCQ、M-PBCQ 和 BCQ 在 Torcs 任务的离线经验缓存上进行实验. 实验中, 具体的参数设置如表 1 所示.

表 1　参数设置

Table 1　Parameter settings

参数名称	参数数值
扰动网络各层神经元个数	400、300
两个 Q 值网络各层神经元个数	400、300
$E_{\omega 1}$ 网络各层神经元个数	750、750
$D_{\omega 2}$ 网络各层神经元个数	750、750
优先级修正系数 σ	10^{-7}
折扣因子 γ	0.99
软更新参数 τ	0.5
步长 η	2.5×10^{-4}

5.1　medium 经验数据

当离线数据集中的经验数据为 medium 等级时, TD-PBCQ、M-PBCQ 和 BCQ 在 D4RL 任务上取得的回报曲线如图 3 所示, 其中实线为平均回报曲线, 阴影区域为平均奖励的标准差. 图 4 给出了 medium 离线数据集中各路径所对应总回报的统计直方图. 由图 3、4 可以得出如下结论:

1) 在 Ant、HalfCheetah 和 Walker2d 中 TD-PBCQ 取得了最高的回报. 这是由于: Ant、HalfChee-

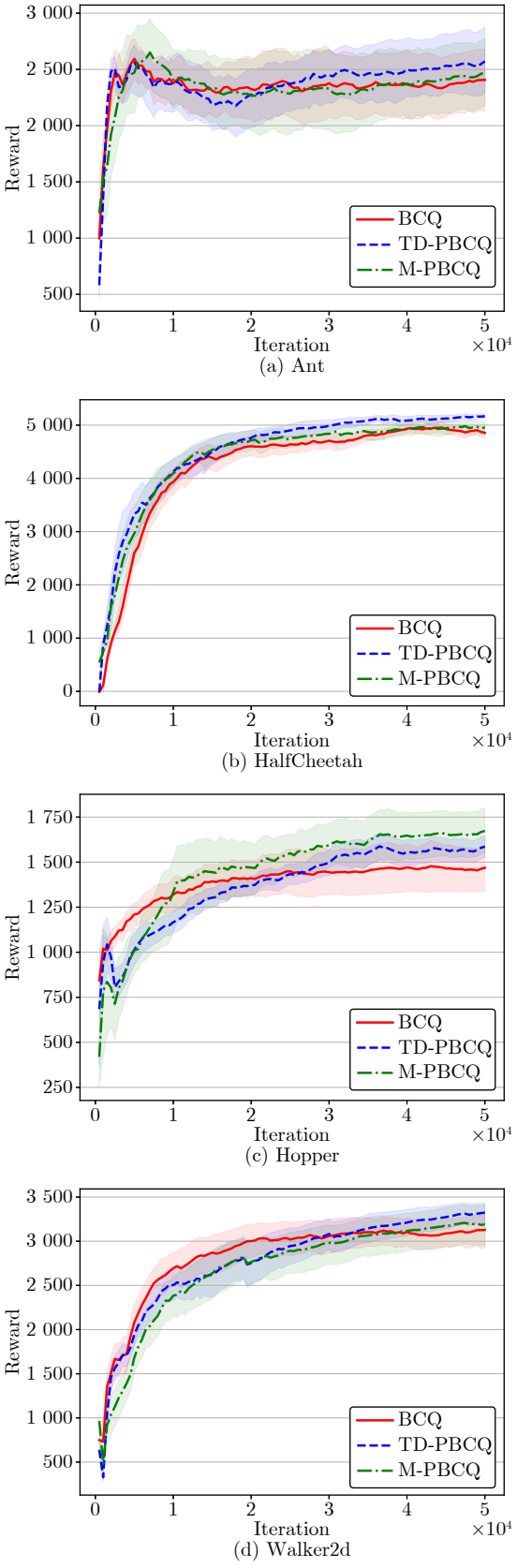


图 3　平均回报曲线对比 (medium 经验数据)

Fig. 3　Comparison of average reward curves (medium experience data)

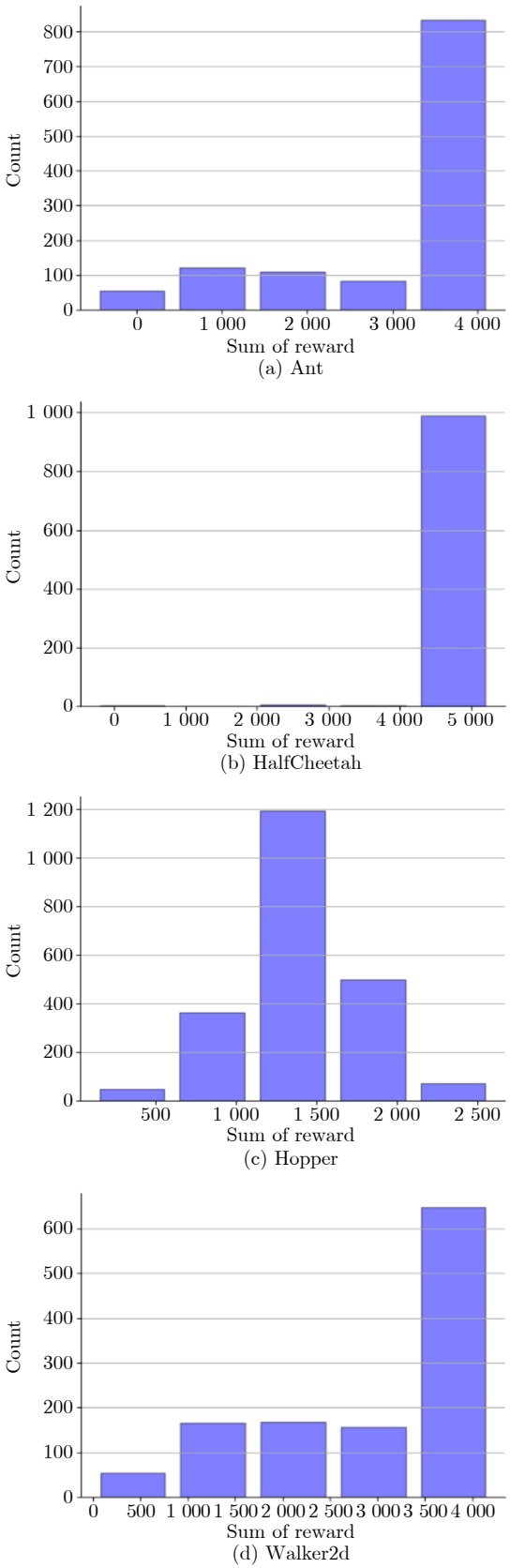


图 4 回报的统计直方图 (medium 经验数据)

Fig. 4 Statistical histogram of reward (medium experience data)

tah 和 Walker2d 任务中 medium 离线经验数据的回报统计直方图是右偏的, 且最高峰在最右侧. 此种情况下, TD-PBCQ 通过降低时序差分误差, 得到了更准确的值函数; BCQ 和 M-PBCQ 均是更倾向于最优路径的学习, 值函数估计误差的累积使得其最终性能不如 TD-PBCQ. 也就是说, 如果策略没有收敛, 且离线经验都分布在缓存中最优路径周围, TD-PBCQ 可以取得更好的实验效果.

2) 在 Hopper 任务中, M-PBCQ 的平均回报收敛到 1 600 以上, 而 BCQ 和 TD-PBCQ 的平均回报均在 1 600 以下. 由图 4(c) 可以看出, 与其他 3 个任务不同, Hopper 任务中 medium 离线经验缓存中的路径总回报大都分布在 1 100 ~ 1 600 的中等水平区间内. 因此, Hopper 任务中 medium 的离线经验缓存中有较多负样本, 导致 BCQ 和 TD-PBCQ 陷入局部最优. 但是, M-PBCQ 能够减弱负样本对策略优化的负面影响, 使得习得策略明显优于离线经验缓存中的平均水平.

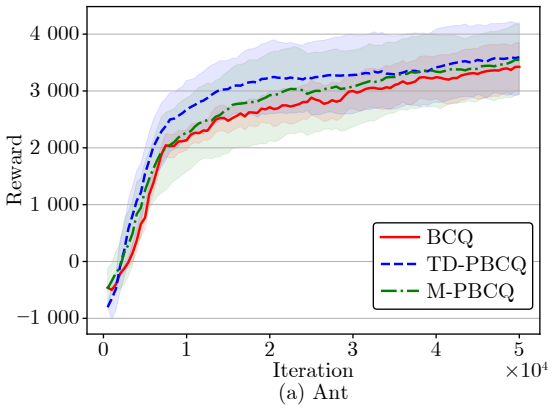
3) 在所有 4 个测试任务上, TD-PBCQ 和 M-PBCQ 的平均回报曲线都要高于 BCQ. 这是由于: medium 策略并不是最优策略, 如果使用均匀采样, 正、负样本有相同的概率被选择, 因此 BCQ 的性能被抑制. 也就是说, 改变采样模型可以有效降低离线强化学习中的误差累积, 提高算法的学习性能.

5.2 expert 经验数据

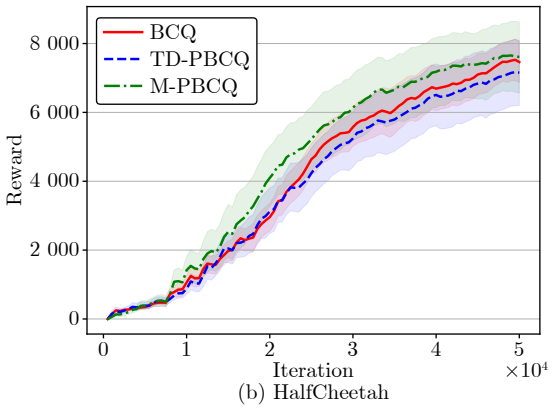
expert 策略等价于最优策略, 收集得到的经验数据集也基本上都分布在最优路径的周围. 当离线数据集中的经验数据为 expert 等级时, TD-PBCQ、M-PBCQ 和 BCQ 在 D4RL 任务上取得的回报曲线如图 5 所示. 图 6 给出了 expert 离线数据集中各路径所对应总回报的统计直方图. 由图 5、6 可以看出:

1) TD-PBCQ 在 Ant 和 Hopper 任务上取得了最高的回报. 这是由于: Ant 和 Hopper 任务中 expert 离线经验数据的回报统计直方图是右偏的且最高的峰在最右侧. 另外, 这两个任务中的 expert 行为策略并没有完全收敛, 都有一定概率访问远离主要路径的经验数据. 在此类离线强化学习任务中, TD-PBCQ 有效减小了值估计误差, 因此取得了最好的实验效果.

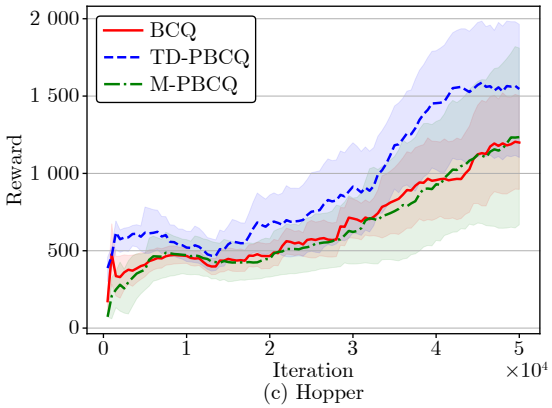
2) 如果策略完全收敛到最优策略, TD-PBCQ 的训练会过拟合, 影响实验效果. 从图 6(b) 可以看出, HalfCheetah 任务中回报统计直方图左侧的数据非常少. TD-PBCQ 由于过度采样左侧的数据导致值函数的训练过拟合, 算法性能受到抑制, 最终表现不如 BCQ.



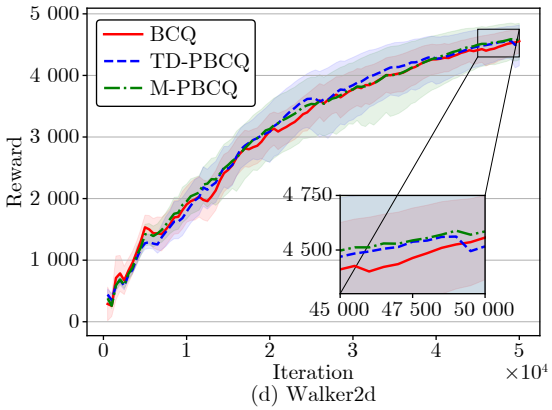
(a) Ant



(b) HalfCheetah



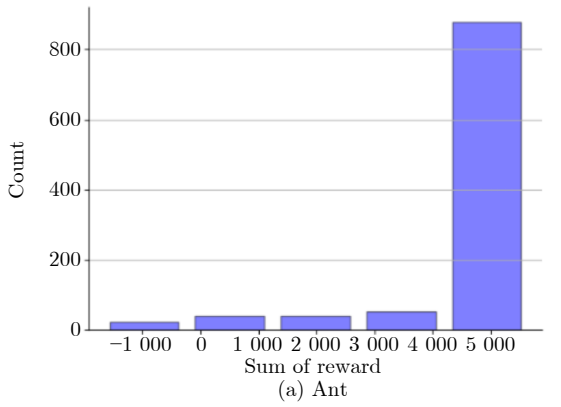
(c) Hopper



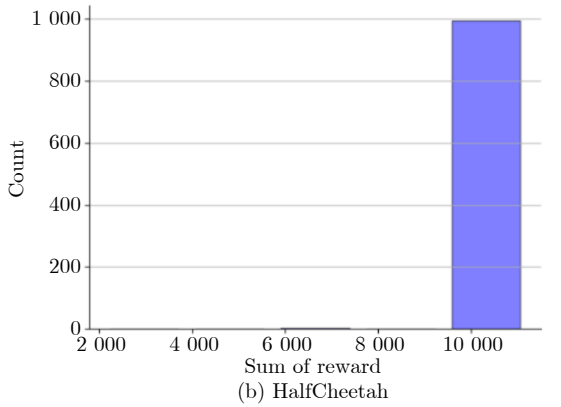
(d) Walker2d

图 5　平均回报曲线对比 (expert 经验数据)

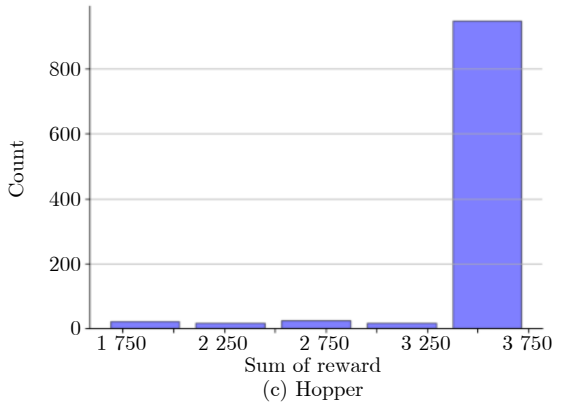
Fig. 5　Comparison of average reward curves (expert experience data)



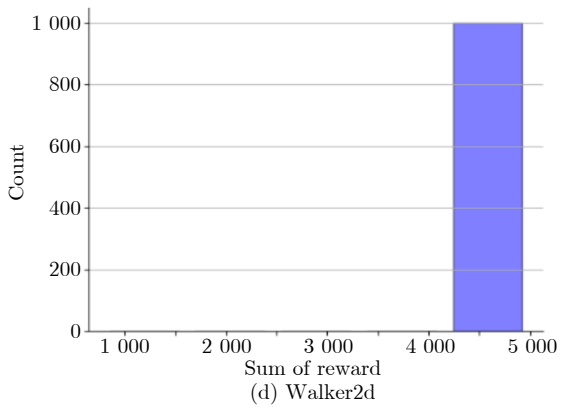
(a) Ant



(b) HalfCheetah



(c) Hopper



(d) Walker2d

图 6　回报的统计直方图 (expert 经验数据)

Fig. 6　Statistical histogram of reward (expert experience data)

3) 从图 6(d) 可以看出, Walker2d 任务中的离线经验数据基本都分布在最优路径上. BCQ、TD-PBCQ 和 M-PBCQ 的平均回报曲线较为相似, 最后都取得了超过 4 500 的平均回报. 这是因为在经过多次迭代后, BCQ 和 TD-PBCQ 的采样模型均为均匀采样, 抑制了回报的上升趋势. 然而, M-PBCQ 可以一直降低对负样本的采样频率, 因而以较小的优势强于 BCQ 和 TD-PBCQ.

5.3 自动驾驶离线数据

Torcs 是一款开源 3D 赛车模拟游戏, 其赛道较长、路况复杂且没有公开的经验数据集. 在实验过程中, 使用中等行为策略收集了平均回报为 7 820 的离线数据. 表 2 和图 7 给出了 BCQ、TD-PBCQ 和 M-PBCQ 在 Torcs 任务上的实验结果, 可以得出:

表 2 Torcs 任务上平均回报对比

Table 2 Comparison of average reward on Torcs task

算法	平均回报
BCQ	8 304.852
TD-PBCQ	7 121.107
M-PBCQ	11 097.551

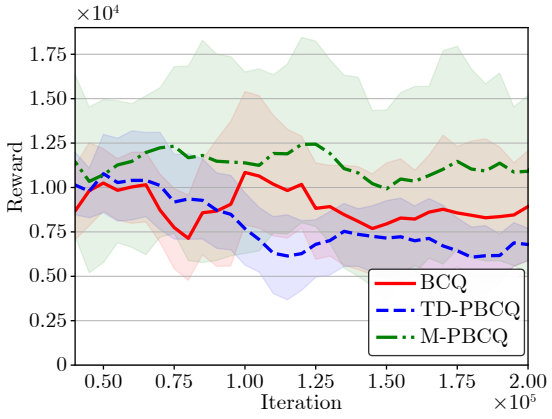


图 7 平均回报曲线对比 (Torcs)

Fig. 7 Comparison of average reward curves (Torcs)

1) 如图 7 所示, 50 000 步之前, TD-PBCQ 学习到了优于 BCQ 的策略. 然而, TD-PBCQ 习得策略的稳定性并不高. 在 93 000 个训练步后, 随着 TD 误差的降低, 基于 TD 误差的采样模型会退化为均匀采样. 因此, 负样本对算法训练的不良影响逐渐变强, 使得习得策略发生了退化.

2) 相比于 TD-PBCQ 和 BCQ, M-PBCQ 的习得策略有着明显的优势, 这是因为离线优先采样模型可以降低负样本对训练的影响, 使智能体学习到更好的策略. 另一方面, 与基于 TD 误差的采样模型不同, 基于鞅的采样模型不会退化为均匀采样, 一些不利于策略优化的经验数据在整个训练过程中被采样的频率都会受到限制, 因此 M-PBCQ 的稳定性更好.

6 总结

强化学习通过智能体与环境在线交互来学习最优策略, 近年来已成为求解复杂环境下感知决策问题的重要手段. 然而, 在线收集数据的方式可能会引发安全、时间或成本等问题, 极大限制了强化学习在实际中的应用. 幸运的是, 离线强化学习能够仅从历史经验数据中学习策略, 而无需与环境产生交互, 这种数据驱动的方式为实现通用人工智能提供了新契机. 然而, 离线数据集的质量将影响算法的学习性能, 想要从离线数据集中学到一个好的策略并非易事. 为此, 本文围绕如何从离线数据集中高效地选择有价值的样本展开研究, 利用时序差分误差和鞅来构造样本优先级, 提出两种离线优先采样模型: 基于时序差分误差的采样模型和基于鞅的采样模型. 在智能体训练过程中, 这两种采样模型可以有针对性地选择经验数据, 引导值函数估计和策略优化. 进一步, 将所提两种采样模型与 BCQ 相结合, 提出基于时序差分误差的优先 BCQ 和基于鞅的优先 BCQ. 需要指出的是, 所提离线优先采样模型具有通用性, 可以方便地与其他离线强化学习方法相结合.

References

1 Sun Chang-Yin, Mu Chao-Xu. Important scientific problems of multi-agent deep reinforcement learning. *Acta Automatica Sinica*, 2020, **46**(7): 1301−1312
(孙长银, 穆朝絮. 多智能体深度强化学习的若干关键科学问题. 自动化学报, 2020, **46**(7): 1301−1312)

2 Wu Xiao-Guang, Liu Shao-Wei, Yang Lei, Deng Wen-Qiang, Jia Zhe-Heng. A gait control method for biped robot on slope based on deep reinforcement learning. *Acta Automatica Sinica*, 2021, **47**(8): 1976−1987
(吴晓光, 刘绍维, 杨磊, 邓文强, 贾哲恒. 基于深度强化学习的双足机器人斜坡步态控制方法. 自动化学报, 2021, **47**(8): 1976−1987)

3 Yin Lin-Fei, Chen Lv-Peng, Yu Tao, Zhang Xiao-Shun. Lazy reinforcement learning through parallel systems and social system for real-time economic generation dispatch and control. *Acta Automatica Sinica*, 2019, **45**(4): 706−719
(殷林飞, 陈吕鹏, 余涛, 张孝顺. 基于 CPSS 平行系统懒惰强化学习算法的实时发电调控. 自动化学报, 2019, **45**(4): 706−719)

4 Chen Jin-Yin, Zhang Yan, Wang Xue-Ke, Cai Hong-Bin, Wang Jue, Ji Shou-Ling. A survey of attack, defense and related security analysis for deep reinforcement learning. *Acta Automatica Sinica*, 2022, **48**(1): 21−39
(陈晋音, 章燕, 王雪柯, 蔡鸿斌, 王珏, 纪守领. 深度强化学习的攻防与安全性分析综述. 自动化学报, 2022, **48**(1): 21−39)

5 Tang Zhen-Tao, Liang Rong-Qin, Zhu Yuan-Heng, Zhao Dong-Bin. Intelligent decision making approaches for real time fighting game. *Control Theory & Applications*, 2022, **39**(6): 969−985
(唐振韬, 梁荣钦, 朱圆恒, 赵冬斌. 实时格斗游戏的智能决策方法. 控制理论与应用, 2022, **39**(6): 969−985)

6 Sun Y X. Performance of reinforcement learning on traditional video games. In: Proceedings of the 3rd International Conference on Artificial Intelligence and Advanced Manufacture (AIAM). Manchester, United Kingdom: IEEE, 2021. 276−279

7 Liu Jian, Gu Yang, Cheng Yu-Hu, Wang Xue-Song. Prediction of breast cancer pathogenic genes based on multi-agent reinforcement learning. *Acta Automatica Sinica*, 2022, **48**(5): 1246−1258
(刘健, 顾扬, 程玉虎, 王雪松. 基于多智能体强化学习的乳腺癌致病基因预测. 自动化学报, 2022, **48**(5): 1246−1258)

8 Kiran B R, Sobh I, Talpaert V, Mannion P, Al Sallab A A, Yogamani S, et al. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2022, **23**(6): 4909−4926

9 Zhang Xing-Long, Lu Yang, Li Wen-Zhang, Xu Xin. Receding horizon reinforcement learning algorithm for lateral control of intelligent vehicles. *Acta Automatica Sinica*, 2023, **49**(12): 2481−2492
(张兴龙, 陆阳, 李文璋, 徐昕. 基于滚动时域强化学习的智能车辆侧向控制算法. 自动化学报, 2023, **49**(12): 2481−2492)

10 Levine S, Kumar A, Tucker G, Fu J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprint arXiv: 2005.01643, 2020.

11 Huang Z H, Xu X, He H B, Tan J, Sun Z P. Parameterized batch reinforcement learning for longitudinal control of autonomous land vehicles. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019, **49**(4): 730−741

12 Agarwal R, Schuurmans D, Norouzi M. An optimistic perspective on offline reinforcement learning. arXiv preprint arXiv: 1907.04543, 2020.

13 Fujimoto S, Gu S S. A minimalist approach to offline reinforcement learning. arXiv preprint arXiv: 2106.06860, 2021.

14 Rashidinejad P, Zhu B H, Ma C, Jiao J T, Russell S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *IEEE Transactions on Information Theory*, 2022, **68**(12): 8156−8196

15 Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR, 2019. 2052−2062

16 Kumar A, Fu J, Soh M, Tucker G, Levine S. Stabilizing off-policy Q-learning via bootstrapping error reduction. In: Proceedings of the 33rd Annual Conference on Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2019. 11784−11794

17 Gretton A, Borgwardt K M, Rasch M, Scholköpf B, Smola A J. A kernel approach to comparing distributions. In: Proceedings of the 22nd National Conference on Artificial Intelligence. Vancouver British Columbia, Canada: AAAI Press, 2007. 1637−1641

18 Jaques N, Ghandeharioun A, Shen J H, Ferguson C, Lapedriza G, Jones N, et al. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. arXiv preprint arXiv: 1907.00456, 2019.

19 Maran D, Metelli A M, Restelli M. Tight performance guarantees of imitator policies with continuous actions. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press, 2023. 9073−9080

20 Wu Y F, Tucker G, Nachum O. Behavior regularized offline reinforcement learning. arXiv preprint arXiv: 1911.11361, 2019.

21 Schaul T, Quan J, Antonoglou I, Silver D. Prioritized experience replay. arXiv preprint arXiv: 1511.05952, 2016.

22 Horgan D, Quan J, Budden D, Barth-Maron G, Hessel M, van Hasselt H, et al. Distributed prioritized experience replay. In: Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net, 2018.

23 Mandl P. A connection between controlled Markov chains and martingales. *Kybernetika*, 1973, **9**(4): 237−241

24 Hernández-Lerma O, de Ozak M M. Discrete-time Markov control processes with discounted unbounded costs: Optimality criteria. *Kybernetika*, 1992, **28**(3): 191−212

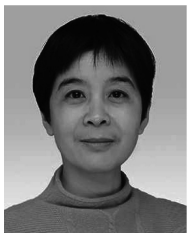
25 Even-Dar E, Mansour Y. Learning rates for Q-learning. In: Proceedings of the 14th Annual Conference on Computational Learning Theory. Berlin, Germany: Springer Verlag, 2001. 589−604

26 Hu Y L, Skyrms B, Tarrès P. Reinforcement learning in signaling game. arXiv preprint arXiv: 1103.5818, 2011.

27 Chow Y, Nachum O, Duenez-Guzman E, Ghavamzadeh M. A Lyapunov-based approach to safe reinforcement learning. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc., 2018. 8092−8101

顾 扬 2022 年获中国矿业大学博士学位. 主要研究方向为深度强化学习. E-mail: guyang@cumt.edu.cn
(**GU Yang** Received his Ph.D. degree from China University of Mining and Technology in 2022. His main research interest is deep reinforcement learning.)

程玉虎 中国矿业大学教授. 2005 年获中国科学院自动化研究所博士学位. 主要研究方向为机器学习, 智能系统. E-mail: chengyuhu@163.com
(**CHENG Yu-Hu** Professor at China University of Mining and Technology. He received his Ph.D. degree from Institute of Automation, Chinese Academy of Sciences in 2005. His research interest covers machine learning and intelligent system.)



王雪松 中国矿业大学教授. 2002 年获中国矿业大学博士学位. 主要研究方向为机器学习, 模式识别. 本文通信作者.
E-mail: wangxuesongcumt@163.com
(**WANG Xue-Song** Professor at China University of Mining and Technology. She received her Ph.D. degree from China University of Mining and Technology in 2002. Her research interest covers machine learning and pattern recognition. Corresponding author of this paper.)