基于双重值修正的离线强化学习

杨 露，王雪松，金 可，程玉虎†

(中国矿业大学 信息与控制工程学院，江苏 徐州 221116)

摘 要: 离线强化学习 (ORL) 依赖固定数据集进行动态决策学习, 常因分布外动作引发外推误差. 现有方法通常通过约束策略分布或采用保守的Q值估计来缓解该问题, 但由此带来的悲观性会导致习得的策略次优. 为此, 从提升值函数估计准确性的角度出发, 构造了一种Q值修正 (QVC) 贝尔曼算子, 其以习得Q函数与行为Q函数之间的差异作为方向性信号, 对Q函数更新目标进行有界修正. 在此基础上, 将 QVC 贝尔曼算子与分布内贝尔曼算子相结合, 提出平衡贝尔曼算子以更好地利用分布内外数据. 理论结果表明, 通过平衡贝尔曼算子迭代得到的Q函数具有收敛性, 且其与真实Q函数之间的误差是有界的. 进一步, 将平衡贝尔曼算子集成至隐式Q学习中, 并在V函数更新过程中引入针对Q值高估与低估的自适应修正机制, 提出基于双重值修正的离线强化学习 (ORL-DVC) 方法. 实验结果表明, 在 D4RL 基准的 Gym-Mujoco 移动控制和 AntMaze 导航控制任务中, ORL-DVC 的平均归一化得分达到 80.9 和 62.7, 整体性能优于现有主流 ORL 方法, 体现出更优的泛化性能.

关键词: 离线强化学习；分布外动作；贝尔曼算子；Q函数；V函数；值修正

中图分类号: TP18 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.1154

引用格式: 杨露, 王雪松, 金可, 等. 基于双重值修正的离线强化学习 [J]. 控制与决策, xxxx, x(x): xxxx-xxxx.

Offline reinforcement learning based on dual value correction

YANG Lu，*WANG Xue-song*，*JIN Ke*，*CHENG Yu-hu*†

(School of Information and Control Engineering，China University of Mining and Technology，Xuzhou 221116，China)

Abstract: Offline reinforcement learning (ORL) leverages fixed datasets for dynamic decision-making, but is often prone to extrapolation errors due to out-of-distribution actions. Existing approaches typically mitigate this issue by constraining the policy distribution or applying conservative Q-value estimation, but the induced pessimism often leads to suboptimal policies. To address this issue, this paper introduces a Q-value correction (QVC) Bellman operator from the perspective of improving the accuracy of value function estimation. The QVC Bellman operator uses the difference between the learned and behavior Q-functions as a directional signal, applying bounded adjustments to the Q-value update target. Building on this, we combine the QVC Bellman operator with the in-distribution Bellman operator to form a balanced Bellman operator, enabling more effective utilization of both in-distribution and out-of-distribution data. Theoretical analysis confirms that the Q-function derived from iterative application of the balanced Bellman operator is convergent, and its deviation from the true Q-function is bounded. Furthermore, we integrate the balanced Bellman operator into implicit Q-learning and incorporate an adaptive correction mechanism in V-function update to jointly address Q-value overestimation and underestimation, thus propose a novel ORL method based on dual value correction (ORL-DVC). Experimental results on the D4RL benchmark, including Gym-Mujoco locomotion and AntMaze navigation tasks, demonstrate that the ORL-DVC achieves an average normalized score of 80.9 and 62.7, respectively, outperforming existing state-of-the-art ORL methods with superior generalization capability.

Keywords: offline reinforcement learning； out-of-distribution action； Bellman operator； Q-function； V-function；value correction

0 引 言

近年来, 强化学习 (Reinforcement Learning, RL) 在各类决策任务中的应用受到广泛关注, 涵盖了无人机控制[1]、机器人领域[2-3]、路径规划[4-5] 等多个领域. 然而, 强化学习的发展面临诸多挑战, 尤其是探索过程带来的潜在风险以及与环境交互产生的高昂

收稿日期: 2025-11-06；**录用日期:** 2026-03-03.

基金项目: 国家自然科学基金项目 (62573416, 62373364).

责任编委: 卢剑权.

†通信作者. E-mail: chengyuhu@163.com

本文附带电子附录文件, 可登录本刊官网该文“资源附件”区自行下载阅览.

成本[6-7]. 离线强化学习 (Offline RL, ORL) 致力于在无需与环境进行额外交互的前提下, 从预先收集的静态数据集中学习有效策略, 从而突破了强化学习在高风险或高成本场景下的应用瓶颈[8-9]. 然而, 离线强化学习在实践中仍面临显著挑战. 虽然理论上可直接将传统在线强化学习方法迁移至离线场景, 但实际效果往往不佳, 其根本原因在于分布偏移及其引发的外推误差[10-11]. 当学习到的策略偏离行为策略时, 极易导致 Q 函数在自举更新过程中产生乐观高估[12]. 具体而言, 在时序差分更新中, Q 函数通过自举方式迭代更新, 即使用当前估计值来更新目标值. 这种估计误差会在自举更新过程中被逐步放大, 并进一步影响其他样本的价值估计. 为缓解这一问题, 现有研究主要围绕如何减少选择分布外 (Out-of-Distribution, OOD) 动作展开[13].

一类典型方法是策略约束. Fujimoto 等[14] 提出的批约束深度 Q 学习通过条件变分自编码器生成与数据集中相似的动作, 从而将学习到的策略限制在行为策略的支持集内. Wu 等[15] 通过在策略更新中引入行为策略正则项, 提出了行为正则 Actor-Critic 方法. Ran 等[16] 进一步提出具有数据集约束的正则化 (Policy Regularization with Dataset Constraint, PRDC) 方法, 使得策略在相似状态下产生的动作与数据集中的动作保持一致. Fujimoto 等[17] 提出一种基于行为克隆的双延迟深度确定性策略梯度 (Twin Delayed Deep Deterministic Policy Gradient with Behavior Cloning, TD3BC), 显式约束策略不偏离数据分布. 此外, Peng 等[18] 提出加权行为克隆来提取策略, 能偏向性地模仿高价值的动作. 但是, 这些方法往往导致策略过于保守, 从而难以显著超越行为策略.

另一类方法是值约束. 代表性工作如 Kumar 等[19] 提出的保守 Q 学习 (Conservative Q-Learning, CQL), 通过强制学习到的 Q 函数期望值低于真实值, 抑制策略对 OOD 动作的过度偏好. Kostrikov 等[20] 和 Huang 等[21] 采用悲观的 Q 值估计来降低风险, Chen 等[22] 从状态值函数的角度设计惩罚. 除此以外, Deng 等[23] 利用集成方法估计不确定性以缓解 OOD 动作带来的影响, 提出了虚假相关减少 (False Correlation Reduction, SCORE) 方法. 虽然这些方法在一定程度上缓解了 Q 值高估, 但由于普遍采用悲观估计, 它们不可避免地限制了策略的改进能力.

近期研究开始关注样本内学习. Kostrikov 等[24] 提出仅基于数据集内动作进行价值更新的隐式 Q 学习 (Implicit Q-Learning, IQL), 从根本上避免 OOD 动作查询. IQL 能够对分布内的动作进行更准确的价值估计与利用, 但这也会限制 Q 函数与策略的泛化能力. 为此, Mao 等[25] 提出的双重温和泛化 (Doubly Mild Generalization, DMG) 通过结合分布内和分布外数据的价值估计, 有效提升了 Q 函数和策略的泛化能力. 与此同时, Ma 等[26] 将样本内学习方法应用于真实机器人任务中, 进一步验证了此类方法在现实场景中应用的有效性与可行性.

综上所述, 现有方法普遍采取了对 OOD 动作的悲观态度, 虽然保障了学习稳定性, 但也可能舍弃其中潜在的高价值动作. 如何通过更准确的价值估计, 在确保稳定性的前提下使策略向高价值区域泛化, 是离线强化学习面临的关键问题. 针对上述挑战, 本文提出一种基于双重值修正的离线强化学习 (ORL Based on Dual Value Correction, ORL-DVC) 方法, 主要贡献如下:

1) 针对 Q 值估计的准确性问题, 提出 Q 值修正 (Q-Value Correction, QVC) 贝尔曼算子. QVC 贝尔曼算子通过引入习得 Q 函数与行为 Q 函数之间的差值对 Q 函数更新目标进行修正, 可以在安全范围内对其进行有界调整, 从而使策略能够合理利用分布外的高价值动作.

2) 在 Q 函数更新中, 将 QVC 贝尔曼算子与分布内贝尔曼算子加权融合, 提出平衡贝尔曼算子以有效提升 Q 函数的泛化能力. 在 V 函数更新中, 引入自适应 Q 值修正机制, 根据高估或低估风险动态调整 V 函数的学习目标, 从而为优势函数的计算提供一个稳健的基线, 最终提高策略更新的可靠性.

3) 通过理论分析, 证明了习得 Q 函数与行为 Q 函数差值的有界性. 在此基础上, 进一步证明了通过平衡贝尔曼算子迭代得到的 Q 函数不仅具有收敛性, 且该 Q 函数与真实 Q 函数之间的估计误差有界.

1 基于双重值修正的离线强化学习

为解决离线强化学习泛化能力不足与价值估计偏差问题, 本文提出 ORL-DVC 方法. 如图 1 所示, ORL-DVC 由策略评估与策略提升构成.

在策略评估阶段, 引入 Q 值修正贝尔曼算子, 将当前 Q 值与行为 Q 值之间的偏差作为修正项, 并借助修正强度参数α对 Q 函数更新目标进行有界调整. 随后, Q 网络通过加权的回归损失, 在修正贝尔曼目标与分布内贝尔曼目标之间进行权衡更新. 与此同时, V 网络以自适应修正 Q 值作为基准, 利用期望分位数回归进行更新.

在策略提升阶段, 采用优势加权行为克隆与价

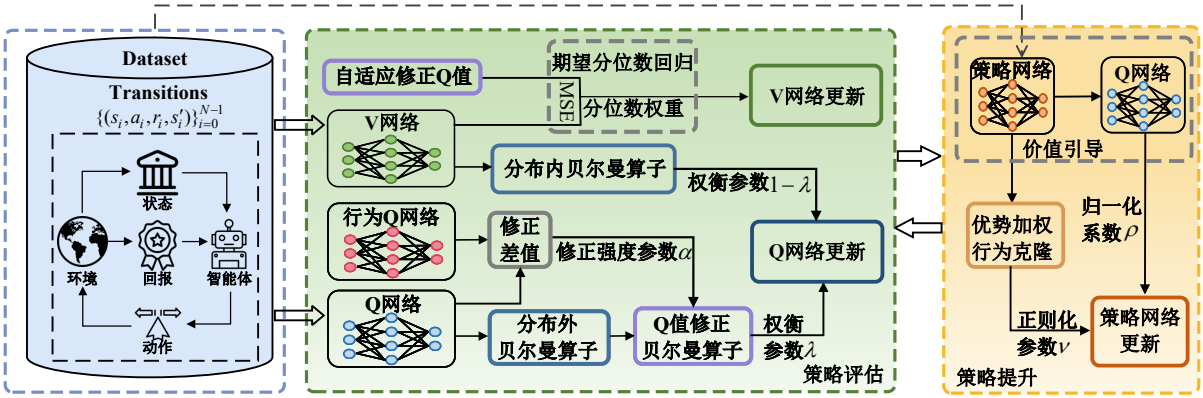


图1　ORL-DVC 框图

值引导相结合的方式进行策略更新, 通过归一化系数ρ与正则化平衡因子ν的联合调节, 在模仿行为策略与探索高价值动作之间实现平衡, 从而缓解策略过度保守的问题并提升其泛化能力.

通过策略评估与策略提升的交替迭代, 可靠的价值估计能提升习得策略的稳定性, 而更加稳定的策略又反过来进一步降低价值估计误差. 二者形成协同闭环, 从而系统性地提升值函数估计的准确性与整体算法性能.

1.1　问题描述

强化学习问题一般被建模为马尔可夫决策过程$\mathcal{M}=\{\mathcal{S},\ \mathcal{A}, P, r,\ \gamma, \eta_0\}$, 其中$\mathcal{S}$是有限状态空间, $\mathcal{A}$是动作空间, $P:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\Delta(\mathcal{S})$是动态转移概率, $r:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$是奖励函数, $\gamma\in[0,1)$是折扣因子, η_0表示初始状态分布. 不失一般性, 假设奖励函数r的取值范围在$[0,\ r_{\max}]$之间. 确定性策略 $\pi:\mathcal{S}\to\mathcal{A}$表示在与环境交互过程中, 智能体在观测到状态$s$后采取的动作$a=\pi(s)$, 接着智能体会根据$s'\sim P(\cdot|s,a)$转移到下一个状态并得到即时奖励$r=r(s,a)$. RL 的目标是寻找一个策略$\pi$, 以最大化期望累计折扣回报$\mathbb{E}_{s_0\sim\eta_0, a_t\sim\pi(s_t), s_{t+1}\sim P(\cdot|s_t,a_t)}[\sum_{t=0}^{\infty} r(s_t,a_t)]$.

对于一个给定的策略π, 其状态值函数 (简称 V 函数) 定义为$V_\pi(s)=\mathbb{E}_\pi[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)|s_0=s]$, 且状态-动作值函数 (简称 Q 函数) 定义为$Q_\pi(s,a)=\mathbb{E}_\pi[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)|s_0=s,a_0=a]$. 这两个值函数在确定性策略$\pi$下的贝尔曼算子分别定义为:

$$\mathcal{B}^\pi Q_\pi(s,a)=r+\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[Q_\pi(s',\pi(s'))]. \tag{1}$$

$$\mathcal{B}^\pi V_\pi(s)=r+\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V_\pi(s')]. \tag{2}$$

式 (1) 和式 (2) 可以分别重写为$Q_\pi(s,a)=\mathcal{B}^\pi Q_\pi(s,a)$及$V_\pi(s)=\mathcal{B}^\pi V_\pi(s)$.

在离线强化学习中, 训练基于一个由行为策略π_ω预先收集的数据集$\mathcal{D}=\{(s_i,a_i,r_i,s'_i)\}_{i=0}^{N-1}$. 由于数据集$\mathcal{D}$的有限性, 环境无法被完全覆盖, 因此在实际学习中通常采用经验贝尔曼算子$\hat{\mathcal{B}}^\pi$来近似真实的贝尔曼算子$\mathcal{B}^\pi$. 然而, 受限于数据分布, 传统的离线强化学习方法在更新过程中可能访问 OOD 动作, 从而引入外推误差并导致 Q 值高估. 为避免该问题, IQL[24] 提出了一种仅在分布内动作上进行更新的思想, 其核心在于以$V(s)$替代$Q(s,a)$的最大化操作, 从而避免对 OOD 动作的显式查询. 具体而言, IQL 通过期望分位数回归学习 V 函数:

$$\min_V \mathbb{E}_{(s,a)\sim\mathcal{D}}[|\tau-\mathbf{1}_{\{Q(s,a)-V(s)<0\}}| (Q(s,a)-V(s))^2,]. \tag{3}$$

其中, τ为期望回归系数. 通过最小化式 (3), IQL 学习 V 函数以近似动作支持集上的最大 Q 值, 实现对高价值动作的保守估计. 随后, IQL 通过最小化贝尔曼误差进行 Q 函数更新:

$$\min_Q \mathbb{E}_{(s,a,s')\sim D}[(\hat{\mathcal{B}}^\pi\hat{V}_\pi(s)-Q(s,a))^2], \tag{4}$$

其中, 目标值$\hat{\mathcal{B}}^\pi\hat{V}_\pi(s)=r+\gamma\mathbb{E}_{s'\sim\hat{P}(\cdot|s,a)}[V_\pi(s')]$. 由于该目标不依赖于$\max_{a'} Q(s,a')$, 而是依赖于 V 函数, 因此 IQL 提升了分布内数据上的学习稳定性.

1.2　修正贝尔曼算子

尽管 IQL 能有效抑制 OOD 动作带来的高估偏差, 但也限制了策略发掘分布外高价值动作的潜力. 为在维持稳定性的同时提高值函数的泛化能力, 考虑将自举更新扩展至 OOD 动作区域. 基于此, 提出 Q 函数更新规则如下:

$$\hat{Q}_{k+1}= \arg\min_Q \lambda\mathbb{E}_{(s,a)\sim\mathcal{D}}[Q(s,a)-\hat{\mathcal{B}}^\pi_{\text{out}}\hat{Q}_k(s,a)]^2+ (1-\lambda)\mathbb{E}_{(s,a)\sim\mathcal{D}}[Q(s,a)-\hat{\mathcal{B}}^\pi_{\text{in}}\hat{V}_k(s)]^2, \tag{5}$$

其中, $\lambda\in[0,1]$为权衡参数, $\hat{\mathcal{B}}^\pi_{\text{out}}$为分布外贝尔曼算子, 定义为:

$$\hat{\mathcal{B}}^\pi_{\text{out}}\hat{Q}_k(s,a)=r+\gamma\mathbb{E}_{s'\sim\hat{P}(\cdot|s,a)}[Q_k(s',\pi(s'))], \tag{6}$$

$\hat{\mathcal{B}}^{\pi}_{\text{in}}$为分布内贝尔曼算子, 定义为:

$$\hat{\mathcal{B}}^{\pi}_{\text{in}}\hat{V}_k(s)=r+\gamma\mathbb{E}_{s'\sim\hat{P}(\cdot|s,a)}[V_k(s')]. \tag{7}$$

考虑到分布内数据的价值估计更为可靠, 而直接对 OOD 动作进行自举更新可能放大估计误差, 因此对不同更新情形采取差异化处理策略:对分布内更新, 使用分布内贝尔曼算子以保持稳定;对分布外更新, 则在分布外贝尔曼算子的基础上引入修正差值以抑制过度外推. 具体地, 在分布外贝尔曼算子$\hat{\mathcal{B}}^{\pi}_{\text{out}}$的基础上引入修正差值项, 构建 Q 值修正贝尔曼算子$\hat{\mathcal{G}}^{\pi}_{\text{QVC}}$, 将其定义为:

$$\begin{aligned}\hat{\mathcal{G}}^{\pi}_{\text{QVC}}\hat{Q}_k(s,a)=&\hat{\mathcal{B}}^{\pi}_{\text{out}}\hat{Q}_k(s,a)+\\&\alpha(\hat{Q}_k(s,a)-\hat{Q}_{\pi_\omega}(s,a)),\end{aligned} \tag{8}$$

其中, $\alpha<1-\gamma$是修正强度参数, $\hat{Q}_{\pi_\omega}$为行为 Q 函数. 该修正差值项的核心作用是提供一个方向性的引导信号, 通过当前 Q 值$\hat{Q}_k(s,a)$与行为 Q 值$\hat{Q}_{\pi_\omega}(s,a)$的差值对更新目标进行调整. 具体来说:当$\hat{Q}_k(s,a)>\hat{Q}_{\pi_\omega}(s,a)$时, 说明当前策略在该状态-动作对上的表现优于行为策略, 此时差值项为正, 能使目标 Q 值$\hat{\mathcal{G}}^{\pi}_{\text{QVC}}\hat{Q}_k(s,a)$相应上调, 从而促进对潜在高价值动作的泛化;当$\hat{Q}_k(s,a)<\hat{Q}_{\pi_\omega}(s,a)$时, 情况相反, 差值项为负, 目标 Q 值$\hat{\mathcal{G}}^{\pi}_{\text{QVC}}\hat{Q}_k(s,a)$相应降低, 从而抑制对低质量动作的过度估计.

选择$\hat{Q}_{\pi_\omega}(s,a)$作为修正差值项的基准有以下两点考虑:

1) 行为策略π_ω生成的数据构成训练数据集的经验分布, 其对应的 Q 估计值$\hat{Q}_{\pi_\omega}$在分布内区域更稳定且方差较小;在分布外动作上的当前 Q 估计值$\hat{Q}_k$容易偏离真实值. 因此, 将行为 Q 函数作为参考锚点, 可以在分布外区域中提供一种校准信号, 使得更新方向既不会盲目悲观地削弱潜在优质动作的价值, 也能有效约束不可靠的高估行为.

2) 修正差值项针对当前状态-动作对(s,a), 而非在下一状态-动作对(s',a')上进行, 是为了直接调整当前动作估计的偏移量. 由于自举更新的传播特性, 若在下一状态上进行修正, 误差可能在多步回溯中被放大并引入额外不确定性;在(s,a)处进行修正差值, 能够即时抑制分布外样本的过度外推, 使得值函数在每次迭代中逐步回归到经验分布的可信区间内, 实现更加稳健的更新.

为支持后续理论分析, 引入以下关于系统动态连续性的标准假设.

假设 1 $\forall s\in\mathcal{S},\forall a_1,a_2\in\mathcal{A}$, 存在一个常数$\iota_r$使$|r(s,a_1)-r(s,a_2)|\leqslant\iota_r\|a_1-a_2\|$成立.

假设 2 $\forall s,s'\in\mathcal{S},\forall a_1,a_2\in\mathcal{A}$, 存在一个常数$\iota_p$使$|P(s'|s,a_1)-P(s'|s,a_2)|\leqslant\iota_p\|a_1-a_2\|$成立.

假设 1 与假设 2 分别要求奖励函数与状态转移概率在动作维度上满足 Lipschitz 连续性. 在强化学习的理论体系中, 此类连续性条件被广泛采纳为一种标准设置[16][27]. 它们反映了系统动态随动作变化的平滑特性, 为值函数估计的收敛性和准确性分析奠定了必要基础.

引理 1 在满足γ-收缩的累积回报条件下, 通过 QVC 贝尔曼算子迭代得到的 Q 函数可在L_∞范数下收敛至唯一不动点.

定理 对于任意的$(s,a)\in\mathcal{S}\times\mathcal{A}$, 习得策略$\pi(s)$和行为策略$\pi_\omega(s)$的 Q 值差距满足:

$$\begin{aligned}&\|Q_\pi(s,a)-Q_{\pi_\omega}(s,a)\|_\infty\leqslant\\&\quad\frac{\gamma(1-\gamma)\iota_r+\gamma^2r_{\max}|\mathcal{S}|\iota_p}{(1-\gamma^2)(1-\gamma)}\Delta_\pi,\end{aligned} \tag{9}$$

其中, $\Delta_\pi=\max\limits_{s\in S}\|\pi(s)-\pi_\omega(s)\|$表示习得策略与行为策略在动作分布上的最大差异, $|\mathcal{S}|$表示马尔可夫决策过程中状态空间的大小.

根据引理 1, 可以知道 QVC 贝尔曼算子具有收敛性. 并且, 定理 1 说明习得策略的 Q 值不会无界地偏离行为策略的 Q 值, 其偏差受到折扣因子γ、常数$|\mathcal{S}|$、Lipschitz 常数ι_r、ι_p以及策略差距Δ_π的共同约束. 该结论进一步保证了以行为 Q 函数$\hat{Q}_{\pi_\omega}$作为参照的合理性, 即通过差值项$\hat{Q}_\pi-\hat{Q}_{\pi_\omega}$进行修正是可控且稳定的.

1.3 平衡贝尔曼算子

基于第 1.2 节提出的 QVC 贝尔曼算子, 此时的 Q 函数更新规则变为:

$$\begin{aligned}\hat{Q}_{k+1}=\arg\min_Q\lambda\mathbb{E}_{(s,a)\sim\mathcal{D}}&[Q(s,a)-\\&\hat{\mathcal{G}}^{\pi}_{\text{QVC}}\hat{Q}_k(s,a)]^2+\\&(1-\lambda)\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[Q(s,a)-\hat{\mathcal{B}}^{\pi}_{\text{in}}\hat{V}_k(s)\right]^2.\end{aligned} \tag{10}$$

其中, 参数λ在基于分布内与分布外的自举更新之间进行权衡: 当λ较小时, 学习主要依赖于经验数据, 训练过程更为稳定; 当λ较大时, 修正后的分布外自举更新占比提升, 使得策略能在定理 1 所保障的可控范围内, 向分布外动作方向进行可靠泛化.

对式 (10) 中$\lambda\mathbb{E}_{(s,a)\sim\mathcal{D}}[Q(s,a)-\hat{\mathcal{G}}^{\pi}_{\text{QVC}}\hat{Q}_k(s,a)]^2+(1-\lambda)\mathbb{E}_{(s,a)\sim\mathcal{D}}[Q(s,a)-\hat{\mathcal{B}}^{\pi}_{\text{in}}\hat{V}_k(s)]^2$关于$Q(s,a)$求导, 并令其为 0, 可得:

$$\hat{Q}_{k+1}=\lambda\hat{\mathcal{G}}^{\pi}_{\text{QVC}}\hat{Q}_k(s,a)+(1-\lambda)\hat{\mathcal{B}}^{\pi}_{\text{in}}\hat{V}_k(s), \tag{11}$$

用平衡贝尔曼算子$\hat{\mathcal{K}}^{\pi}_{\text{Bal}}$来更新 Q 函数, 即:

$$\hat{Q}_{k+1}(s,a)=\hat{\mathcal{K}}^{\pi}_{\mathrm{Bal}}\hat{Q}_k(s,a). \tag{12}$$

为分析在该更新过程中平衡贝尔曼算子的收敛性及其得到的 Q 函数估计误差, 引入如下关于经验模型估计误差的假设.

假设 3 对于任意的状态-动作对$(s,a)\in\mathcal{D}$, 以下不等式成立的概率至少为$1-\xi$:

$$\sum_{s'\in\mathcal{S}}|\hat{P}(s'|s,a)-P(s'|s,a)|\leqslant c_P\sqrt{\frac{1}{\mathcal{D}_c}}, \tag{13}$$

其中, $\xi\in(0,1)$为置信水平, $\mathcal{D}_c\neq 0$表示数据集中包含的状态-动作对(s,a)的出现次数, $c_P>0$为与转移概率分布集中性质相关的常数.

经验状态转移概率分布$\hat{P}$和真实状态转移概率分布P之间的偏差不会在训练过程中自动消失, 因此该假设在离线强化学习理论分析中是标准设置[28]. 当数据集$\mathcal{D}$包含足够多的样本时, 该偏差会逐渐减小, 并在极限情况下趋近于 0.

定理 2 给定任意初始 Q 函数$\hat{Q}_0(s,a)$, 通过迭代更新规则 (11) 对 Q 函数进行连续迭代, 生成的序列$\{\hat{Q}_0(s,a),\hat{Q}_1(s,a),\ldots,\hat{Q}_k(s,a)\}$在$k\to\infty$时收敛到唯一的不动点$\hat{Q}^{\pi}_{\mathrm{Bal}}(s,a)$, 满足$\hat{Q}^{\pi}_{\mathrm{Bal}}(s,a)=\hat{\mathcal{K}}^{\pi}_{\mathrm{Bal}}\hat{Q}^{\pi}_{\mathrm{Bal}}(s,a)$.

定理 3 给定一个策略π, 平衡贝尔曼算子的固定点$\hat{Q}^{\pi}_{\mathrm{Bal}}$与真实 Q 函数$Q_\pi$之间的误差满足:

$$\begin{aligned}\|\hat{Q}^{\pi}_{\mathrm{Bal}}-Q_\pi\|_\infty\leqslant\\ \frac{\lambda\alpha(\gamma(1-\gamma)\iota_r+\gamma^2 r_{\max}|\mathcal{S}|\iota_p)\Delta_\pi}{(1-\lambda\alpha-\gamma)(1-\gamma)(1-\gamma^2)}+\\ \frac{(\lambda\alpha+\gamma)c_P r_{\max}}{(1-\lambda\alpha-\gamma)(1-\gamma)\sqrt{\mathcal{D}_c}}.\end{aligned} \tag{14}$$

由定理 2 可知, 平衡贝尔曼算子具有收敛性. 在此基础上, 定理 3 表明通过平衡贝尔曼算子学习到的 Q 函数与真实 Q 函数之间的差距是有界的. 该误差由以下关键因素影响: 1) 策略差异与系统动态的 Lipschitz 性质, 且λ与α过大时, $\hat{Q}^{\pi}_{\mathrm{Bal}}$与$Q_\pi$之间的差距会显著增大; 2) 增大数据集样本$\mathcal{D}_c$能减小$\hat{Q}^{\pi}_{\mathrm{Bal}}$与$Q_\pi$之间的差距. 该结果指出, 尽管适度的分布外更新有助于策略提升, 但过大的λ和α将导致估计误差显著放大. 因此, 在实际应用中需对λ和α进行合理选择, 以在稳定性与泛化性之间取得平衡.

1.4 ORL-DVC

将平衡贝尔曼算子与 IQL 相结合, 提出了 ORL-DVC 方法. ORL-DVC 由一个策略网络π_{ϕ_a}, 一个行为策略网络π_{ϕ_ω}, 两个行为 Q 网络$Q_{\theta_{\omega 1}}$, $Q_{\theta_{\omega 2}}$, 两个 Q 网络Q_{θ_1}, Q_{θ_2}以及一个 V 网络V_φ组成, 其中ϕ_a, ϕ_ω, $\theta_{\omega 1}$, $\theta_{\omega 2}$, θ_1, θ_2和φ是对应的神经网络参数.

1) 行为策略训练

ORL-DVC 采用确定性神经网络π_{ϕ_ω}来近似行为策略, 其优化目标是使网络输出的动作尽可能接近数据集中的动作. 具体地, 行为策略网络通过标准的行为克隆方法进行学习, 其损失函数定义为:

$$L_{\pi_\omega}(\phi_\omega)=\mathbb{E}_{(s,a)\sim\mathcal{D}}[\|\pi_{\phi_\omega}(s)-a\|^2]. \tag{15}$$

行为 Q 网络通过标准的时序差分学习, 并引入双 Q 结构与目标网络机制以缓解自举带来的估计不稳定性, 其损失函数定义为:

$$L_{\pi_\omega}(\theta_\omega)=\mathbb{E}_{(s,a)\sim\mathcal{D}}[(y_\omega-Q_{\theta_{\omega 1}}(s,a))^2] \tag{16}$$

其中, $y_\omega=r+\gamma\min\limits_{j=1,2}\{Q_{\theta'_{\omega j}}(s',\pi_{\phi'_\omega}(s')+\varepsilon)\}$, $\varepsilon\sim\mathrm{clip}(\mathcal{N}(0,\sigma),-c,c)$为方差是$\sigma$的高斯噪声, clip(·)和$c$分别是截断函数和范围, $Q_{\theta'_{\omega 1}}$和$Q_{\theta'_{\omega 2}}$分别代表行为 Q 网络$Q_{\theta_{\omega 1}}$和$Q_{\theta_{\omega 2}}$对应的目标网络, 而$\pi_{\phi'_\omega}$代表行为策略网络π_{ϕ_ω}对应的目标网络.

2) 策略评估

根据式 (5), Q 网络的损失函数定义为:

$$\begin{aligned}L_{\pi_a}(\theta_i)=\lambda\mathbb{E}_{(s,a,s')\sim\mathcal{D}}[(Q_{\theta_i}(s,a)-y_1)^2]+(1-\lambda)\\ \mathbb{E}_{(s,a,s')\sim\mathcal{D}}[(Q_{\theta_i}(s,a)-y_2)^2],\ i=1,2\end{aligned} \tag{17}$$

其中, 目标值y_1定义为:

$$\begin{aligned}y_1=r+\gamma\cdot\min_{i=1,2}\{Q_{\theta'_i}(s',\pi_{\phi'_a}(s')+\varepsilon)\}+\alpha\\ (\min_{i=1,2}\{Q_{\theta'_i}(s,a)\}-\max_{j=1,2}\{Q_{\theta'_{\omega j}}(s,a)\}),\end{aligned} \tag{18}$$

其中, $Q_{\theta'_1}$和$Q_{\theta'_2}$是Q_{θ_1}和Q_{θ_2}对应的目标网络. 通过引入最小学习 Q 值与最大行为 Q 值之差, 可以抑制价值高估, 提升训练稳定性. 目标值y_2定义为:

$$y_2=r+\gamma V_\varphi(s'). \tag{19}$$

当 Q 函数出现局部异常偏差时, 直接采用 IQL 的回归目标可能导致 $V_\varphi(s)$ 被错误牵引, 从而引发训练的不稳定. 因此, 为增强 V 函数的稳定性, 通过以下期望分位数回归更新 V 函数:

$$\begin{aligned}L_V(\varphi)=\mathbb{E}_{(s,a)\sim\mathcal{D}}[|\tau-\mathbf{1}_{\{Q^{\mathrm{sel}}_\theta(s,a)-V_\varphi(s)<0\}}|\\ (Q^{\mathrm{sel}}_\theta(s,a)-V_\varphi(s))^2].\end{aligned} \tag{20}$$

其中, Q^{sel}_θ是自适应修正 Q 函数, 定义为:

$$\begin{aligned}Q^{\mathrm{sel}}_\theta(s,a)=\\ \begin{cases}\min\limits_i Q_{\theta_i}(s,a),\ \text{if}\max\limits_i\ Q_{\theta_i}(s,a)-V_\varphi(s)>0,\\ \max\limits_i Q_{\theta_i}(s,a),\ \text{if}\max\limits_i\ Q_{\theta_i}(s,a)-V_\varphi(s)\leqslant 0.\end{cases}\end{aligned} \tag{21}$$

当最大 Q 值高于当前 V 值估计时, 代表 Q 值存在高估风险, 这时选择保守的最小 Q 值;当 Q 值整体偏低时, 采用相对乐观的最大 Q 值. 这有效缓解了因 Q 函数估计偏差导致的 V 值不稳定问题.

3) 策略提升

类似 IQL, 本文采用优势加权回归来学习策略:

$$L_\pi(\phi)=\mathbb{E}_{(s,a)\sim D}[\exp(\beta(Q(s,a)-V(s)))(\pi_\phi-a)^2], \tag{22}$$

其中, β为温度系数. 该目标由两部分构成:指数优势加权与行为克隆. 尽管优势加权行为克隆能够保正策略学习的稳定性, 其本质仍属于加权模仿学习, 难以突破行为策略的性能上限. 因此, 进一步考虑结合价值引导的策略提升目标, 定义为:

$$J_{\pi_a}(\phi_a)=\rho\mathbb{E}_{(s,a)\sim\mathcal{D}}[\min_{i=1,2}Q_{\theta_i}(s,\pi_{\phi_a}(s))-\nu\cdot \exp(\beta(\min_{i=1,2}Q_{\theta'_i}(s,a)-V_\varphi(s)))(\pi_{\phi_a}(s)-a)^2], \tag{23}$$

其中, $\rho=1/\mathbb{E}_{s\sim\mathcal{D}}[|\min_{i=1,2}Q_{\theta_i}(s,\pi_{\phi_a}(s))|]$为来自TD3BC的归一化系数, ν为正则化平衡因子. 当$\rho\gg\nu$时, 策略更新偏向于价值最大化, 表现出更强的泛化能力;当$\rho\ll\nu$时, 策略趋向于模仿行为策略, 从而保证稳定性. 值得注意的是, 经过修正的 Q 值为策略更新提供了更准确的价值引导, 使策略能够超越行为策略的限制, 实现有效泛化. 与此同时, 更稳定的 V 值提高了优势加权项的准确性, 从而保障了策略更新过程的稳定性. ORL-DVC 的完整流程如算法 1 所示.

2 实 验

2.1 实验细节

在离线数据集 D4RL Gym-Mujoco 运动控制任务上评估 ORL-DVC 的性能, 该数据集提供三种任务分别是 halfcheetah、hopper 和 walker2d. 每一种任务有五种不同质量的数据集, 分别是 r=random, m-r=medium-replay, m=medium, m-e=medium-expert, e=expert. 为进一步验证 ORL-DVC 在稀疏奖励场景下的有效性, 在 AntMaze 导航任务上进行了实验. 该任务要求智能体在复杂迷宫环境中, 从起始位置到达目标区域. 由于奖励信号高度稀疏且决策序列较长, 该任务对值函数估计的准确性提出了较高要求.

为公平起见, 在相同的实验方案下重新运行作者的开源代码并记录得到的数据. 整个训练过程被设置为 100 万步, 并且每 5000 步运行 10 个回合进行一次评估, 一共 200 个 epoch. 策略网络和值网络学习率均设为3×10^{-4}, 目标网络软更新系数τ_s设为 0.005. 行为策略训练总步数T_r为3×10^5, 策略评估与策略提升的训练总步数T_t为1×10^6, 折扣因子γ设为 0.99. 本文采用 D4RL 提供的归一化得分作为评估结果, 计算公式为:

$$\text{归一化得分}=\frac{\text{平均回报-参考随机回报}}{\text{参考专家回报-参考随机回报}}\times100, \tag{24}$$

其中, 平均回报指的是 10 个回合的平均回报.

算法1 ORL-DVC算法

输入: 目标网络软更新系数τ_s, 离线数据集 $\mathcal{D}$, 训练总步数T_r, T_t, 策略网络更新频率 m.

初始化: 策略网络参数 ϕ_a, 行为策略网络参数 ϕ_ω, Q网络参数 θ_1,θ_2, V网络参数 φ, 行为Q网络参数$\theta_{\omega1},\theta_{\omega2}$.

行为策略训练

for training step $e=1,2,\ldots,T_r$ **do**.

  从 $\mathcal{D}$ 中采样小批次$\{(s,a,r,s')\}$.

  通过最小化式(16)更新 θ_ω.

  if $e \bmod m=0$ **then**

    通过最小化式(15)更新 ϕ_ω.

  end if

end for

策略评估与策略提升

for training step $e=1,2,\ldots,T_t$ **do**

  从 $\mathcal{D}$ 中采样小批次$\{(s,a,r,s')\}$.

  对于$i=1,2$, 通过最小化式(17)更新 θ_i.

  通过最小化式(20)更新 φ.

  if $e \bmod m=0$ **then**

    通过最小化式(23)更新 ϕ_a.

  end if

  更新目标网络:

  $\theta'_i\leftarrow(1-\tau_s)\theta'_i+\tau_s\theta_i,\ \theta'_{\omega j}\leftarrow(1-\tau_s)\theta'_{\omega j}+\tau_s\theta_{\omega j},$

  $\phi'_a\leftarrow(1-\tau_s)\phi'_a+\tau_s\phi_a,\ \phi'_\omega\leftarrow(1-\tau_s)\phi'_\omega+\tau_s\phi_\omega.$

end for

2.2 对比结果与分析

为评估 ORL-DVC 性能, 将其与两大类方法进行对比:一类是基于时序差分的典型方法, 包括 TD3BC、CQL、PRDC 和 SCORE;另一类是样本内学习方法, 包括 IQL 和 DMG. 图 2 展示了 ORL-DVC 与各对比方法在 Gym-Mujoco 任务上的归一化得分曲线. 此外, 对于每一个数据集, 均运行 5 个独立的随机种子, 并记录每个种子最后 10 个 epoch 的评估结果, 最终取其平均值±标准差. 表 1 展示了 ORL-DVC 在 Gym-Mujoco 任务上的平均归一化得分, 其中最高分加粗显示.

从图 2 和表 1 可以看出, ORL-DVC 在大多数任务上的表现都优于对比方法, 并取得最高的总分. 具体来说, ORL-DVC 的平均得分为 80.9, 较基线方法 (68.4) 提高约 18.4%, 相较于表现最好的对比方法 (76.5) 仍提升约 5.7%. 根据表 1 和图 2, 对 Gym-

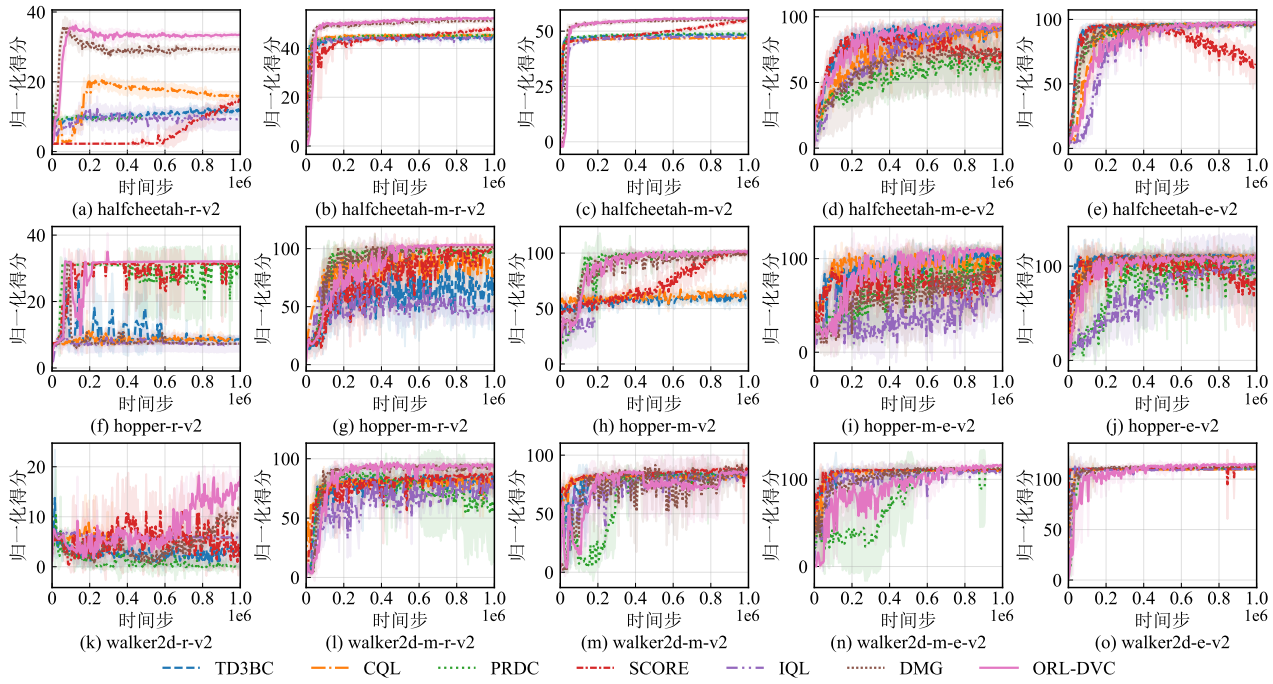


图2　不同任务上的归一化得分曲线

表1　在 Gym-Mujoco 任务上的平均归一化得分对比

任务	TD3BC	CQL	PRDC	SCORE	IQL	DMG	ORL-DVC
halfcheetah-r	11.5±1	16±1	9.9±3.7	14.3±12.6	9.9±3.7	30.2±1.6	**33.6±0.7**
hopper-r	8.6±0.7	8.4±1	31.1±0.3	31.4±0.3	8.1±0.4	7.8±1.1	**32.2±0.2**
walker2d-r	3.1±2.8	4.7±1.7	0.0±0.6	3.4±4.8	6±0.8	9.1±2.1	**15.6±2.7**
halfcheetah-m-r	44.5±0.9	45.3±0.7	45.4±0.5	47.8±0.8	44±0.7	50.9±0.7	**52.1±0.6**
hopper-m-r	68.6±23.8	85.3±17.1	96.6±2.7	98.2±3.2	44±4	101.7±0.4	**103.3±0.6**
walker2d-m-r	82±6.9	79±6.2	61.8±36.3	78.5±8.5	73.2±9.4	93.6±2.9	**94.9±3.4**
halfcheetah-m	48.4±0.3	46.9±0.3	48.8±0.2	54.7±0.9	47.5±0.2	55±0.3	**56.7±0.4**
hopper-m	59.3±4.9	62.6±6.3	100.3±1.1	99.8±3.1	56.4±5	96.8±5.8	**101.8±1.4**
walker2d-m	83.5±2.4	81.4±2.2	85±1.6	85.9±11.6	82.4±2.7	**87.2±6.9**	84.9±3.4
halfcheetah-m-e	90.9±4.9	89.9±6.8	64.7±12.3	71.4±10.4	87.1±6.7	93.2±2.4	**94±1.9**
hopper-m-e	100.3±10	93.8±15.3	94.8±16.8	86.8±19.9	60.8±13.4	95.8±17.7	**110.7±2.7**
walker2d-m-e	110.3±0.6	109.7±0.4	111.7±0.9	110.9±1	108.3±2.2	111.5±0.3	**115±4.6**
halfcheetah-e	96.6±1.7	96.3±1.2	96.5±3.7	65.9±12.6	95.4±0.4	93.3±1.1	**97.5±1.5**
hopper-e	107.6±15.1	**110.5±3.2**	91.7±24.6	81.1±28.9	107.8±3.5	109.2±3.8	109.2±3.8
walker2d-e	110.2±0.4	109.5±0.3	111±0.5	111.9±0.9	109.7±0.4	111.6±0.1	**115±0.8**
Average	68.4±5.1	69.3±4.2	70.0±7.1	69.5±8	62.7±3.6	76.5±3.1	**80.9±2.0**

Mujoco 任务中不同数据集类型的结果进行分析, 可以得出如下结论:

1) 在 random 数据集上, ORL-DVC 相较于施加强策略约束的方法具有显著优势. 这是因为过强的策略约束会迫使习得策略过度接近行为策略, 而当数据质量较差时, 这种限制会导致效果不佳;

2) 在 medium-replay 和 medium 数据集上, ORL-DVC 在策略约束与 Q 值泛化之间取得了良好的平衡, 从而在这些任务中同样获得了较高的得分.

3) 在 medium-expert 和 expert 数据集上, ORL-DVC 相较于策略约束或值约束方法同样表现出明显提升. 这表明在具有专家演示数据的条件下, ORL-DVC 能通过双重值修正机制, 有效利用具有潜力的分布外动作, 从而习得超越原始数据集的、具有更高价值的策略;

在 AntMaze 任务中, 选取时序差分方法中的 TD3BC、CQL 与 SCORE, 以及样本内学习方法中的 IQL 作为对比方法. 表 2 汇总了各方法在不同数据集上的平均归一化得分. 结果表明, ORL-DVC 在 antmaze-umaze、antmaze-umaze-diverse 以及 antmaze-medium-play 数据集上均取得了较优性能. 进一步地, 即使在奖励信号更加稀疏的 antmaze-large-play

表2 在 AntMaze 任务上的平均归一化得分对比

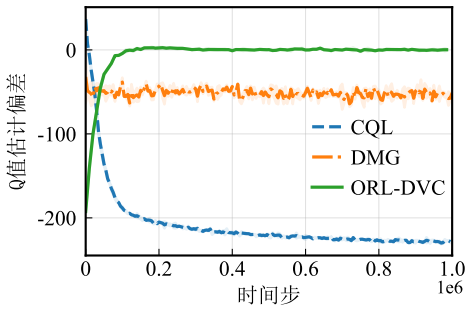
任务	TD3BC	CQL	SCORE	IQL	ORL-DVC
antmaze-umaze	35.7±59.8	63.4±12.5	85.2±3.8	75.9±5.7	**89.1±5.7**
antmaze-umaze-diverse	24.3±56.7	53.4±21.4	4.3±9.4	60.6±10.9	**68.6±8.2**
antmaze-medium-play	2.2±4.3	0.3±0.1	59.6±14.4	66.2±10.7	**68.3±4.5**
antmaze-medium-diverse	2.5±5.0	0.4±0.8	60.4±15.8	**63.1±3.2**	62.3±4.6
antmaze-large-play	0.1±0.4	0.0±0.0	7.4±6.8	37.8±6.7	**44.9±5.8**
antmaze-large-diverse	0.0±0.0	0.0±0.0	13.7±8.2	25.6±8.4	**42.9±4.7**
Average	10.8±21.0	19.6±5.8	38.4±9.7	54.9±7.6	**62.7±5.6**

与 antmaze-large-diverse 数据集上, ORL-DVC 仍显著优于在稀疏奖励任务中表现突出的 IQL 方法.

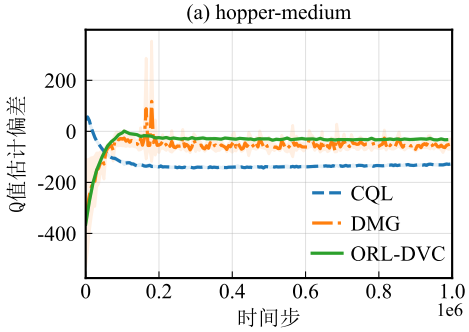
综上所述, ORL-DVC 在不同数据集与任务中均展现出了卓越的性能. 这说明 ORL-DVC 能在策略约束与价值泛化之间取得较好的平衡, 从而在各种数据质量条件下均能保持稳定且出色的性能.

2.3 ORL-DVC 的稳定性分析

为了验证 ORL-DVC 所学 Q 函数的稳定性, 图 3 以 hopper-medium 和 halfcheetah-medium 任务为例, 展示了不同离线强化学习方法的 Q 值估计偏差. 这里, Q 值估计偏差定义为估计 Q 值与真实 Q 值之差.



(a) hopper-medium



(b) halfcheetah-medium

图3 不同方法的 Q 值估计偏差对比

真实 Q 值通过蒙特卡罗方法进行近似估计, 即对同一初始状态-动作对(s_0, a_0)采样 M 条轨迹, 并计算其累计回报的平均值:

$$Q_{\mathrm{MC}}(s_0, a_0) = 1/M \sum_{i=1}^{M} r_i, \tag{25}$$

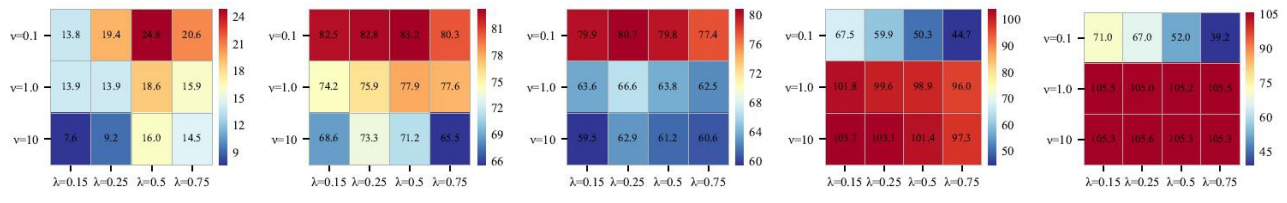
其中, r_i表示第i条轨迹的累计回报. 从图 3 可以看出, 在整个训练过程中, ORL-DVC 的 Q 值估计偏差曲线始终负向接近 0, 且负向偏差明显小于 CQL 和 DMG. 这意味着 ORL-DVC 所学 Q 函数既没有出现高估发散情况, 也不存在 CQL 和 DMG 的过度保守问题, 即 ORL-DVC 能显著提升 Q 函数的估计准确性与稳定性.

2.4 参数分析

为了分析正则化平衡因子ν和权衡参数λ的影响, 针对不同数据集类型, 对 ORL-DVC 在 halfcheetah、hopper、walker2d 任务上获得的归一化回报取平均, 在不同超参数 $\nu = \{0.1, 1.0, 10\}$ 与 $\lambda = \{0.15, 0.25, 0.5, 0.75\}$设置下, 结果如图 4 所示. 实验结果表明:

1) 在 random、medium-replay 和 medium 这类次优的数据集上, ORL-DVC 对ν和λ的变化较为敏感. 当选取较小的$\nu = 0.1$并将λ控制在 [0.25, 0.5] 区间内时, ORL-DVC 表现较为稳定且具有竞争力. 这一现象说明, 较小的ν能够减轻策略对已知数据的依赖, 而适中的λ可在保持保守更新的同时引入有效修正, 从而共同缓解了参数敏感性问题.

2) 在 medium-expert 与 expert 这类接近专家的数据集上, 当ν在 [1.0, 10] 范围内时, ORL-DVC 对λ的变化不敏感, 且整体性能普遍较好. 这是由于在



(a) random (b) medium-replay (c) medium (d) medium-expert (e) expert

图4 在不同的λ和ν下, ORL-DVC 在不同数据类型上的平均归一化得分

此类高质量数据中, 优势加权行为克隆项主导了策略更新, 使得 ORL-DVC 能够有效利用数据集中的高价值动作, 降低了对价值修正项的依赖.

此外, 为了探究修正强度参数α对 ORL-DVC 性能的影响, 设置$\alpha = \{0.001, 0.005, 0.009\}$. 图 5 展示了在不同$\alpha$设置下, 针对每一种任务 (halfcheetah、hopper、walker2d), ORL-DVC 在不同数据集类型上获得的归一化得分总和. 当$\alpha=0.009$时, 在这三种任务上均能得到最高的回报总和. 值得注意的是, $\alpha = 0.009$接近理论约束 ($\alpha<1-\gamma=0.01$) 所允许的上界. 这表明, 在满足理论条件的前提下, 选取较大的α值能更有效地抑制分布外自举误差, 从而带来更准确的值估计.

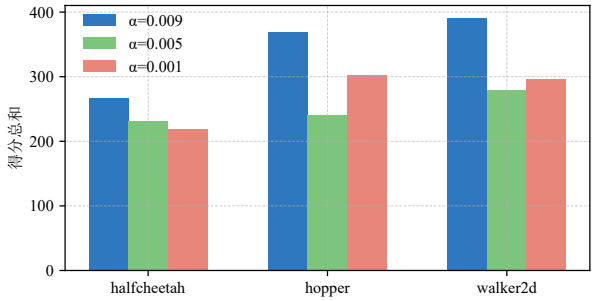


图5　在不同的修正强度参数α下, ORL-DVC 在不同任务上归一化得分总和

根据上述分析可以观察到, 当数据质量由次优逐步提升至接近专家水平时, ν 的取值呈现出增大的趋势, 而 ORL-DVC 对λ 的敏感性逐渐降低. 同时, 在不同任务和数据集类型上, $\alpha=0.009$ 均展现出较为一致的性能优势, 因此可作为默认的参数设置.

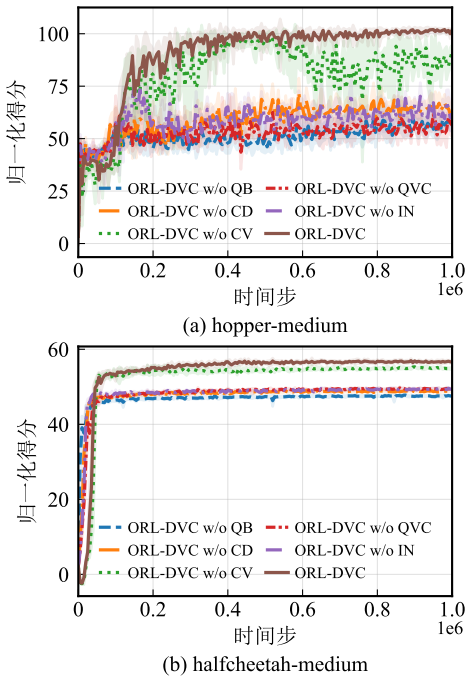


图6　消融实验的归一化得分曲线

2.5　消融实验

为评估 ORL-DVC 中各组件的重要性, 通过逐一移除特定组件的方式进行消融实验, 结果见图 6. 实验过程中, 构造了 ORL-DVC 的消融变体, 包括:

1) ORL-DVC w/o CD: 移除 QVC 贝尔曼算子$\hat{\mathcal{G}}^{\pi}_{\mathrm{QVC}}$中的修正差值项;

2) ORL-DVC w/o QVC: 移除 QVC 贝尔曼算子$\hat{\mathcal{G}}^{\pi}_{\mathrm{QVC}}$, 此时平衡贝尔曼算子仅保留分布内贝尔曼算子;

3) ORL-DVC w/o IN: 移除分布内贝尔曼算子$\hat{\mathcal{B}}^{\pi}_{\mathrm{in}}$, 此时平衡贝尔曼算子仅保留 QVC 贝尔曼算子;

4) ORL-DVC w/o CV: 移除 V 网络更新中的自适应 Q 值修正机制;

5) ORL-DVC w/o QB: 移除策略提升中的价值引导项.

如图 6 所示, V 函数更新过程中引入的自适应 Q 值修正机制对算法的稳定性具有显著影响. 修正差值项和价值引导项对性能提升起到了至关重要的作用, 当缺少这两个组件时, 算法难以获得最优的归一化得分. 与此同时, 平衡贝尔曼算子同样至关重要, 当仅保留分布内贝尔曼算子或 QVC 贝尔曼算子时, ORL-DVC 会因为过于乐观或过于保守取得较低的归一化得分. 总体而言, 缺失其中任意一个组件均会导致性能下降.

3　结　论

针对离线强化学习中的值函数估计不准确问题, 本文首先设计了 QVC 贝尔曼算子, 通过在自举目标中引入基于相对行为 Q 函数的修正差值项, 有效调控对 OOD 动作的价值估计. 进一步, 将分布内贝尔曼算子与 QVC 贝尔曼算子进行加权融合, 提出平衡贝尔曼算子, 以充分利用分布内外样本信息, 在保障训练稳定性的同时增强策略的泛化能力. 在此基础上, 将平衡贝尔曼算子嵌入 IQL 方法中, 构建了 ORL-DVC 方法. 该方法在 Q 函数更新中采用平衡贝尔曼算子, 在策略更新中融合价值引导与加权优势回归, 在 V 函数更新中则通过引入自适应修正 Q 函数以增强期望分位数回归的稳定性, 从而系统性地提升方法性能. 理论分析表明, ORL-DVC 在保持算法收敛的同时, 能够有效控制价值估计误差. 在多个基准任务上的实验结果表明, ORL-DVC 的性能显著优于现有基线方法, 验证了其有效性与优越性.

参考文献 (References)

[1]　倪浩, 章胜, 刘福炜, 等. 基于域随机化增强 EfficientZero

的无人机空战智能决策[J]. 控制与决策, 2025, 40(11): 3273-3286.
(Ni H, Zhang S, Liu F W, et al. UAV air combat intelligent decision-making based on domain randomization enhanced EfficientZero[J]. Control and Decision, 2025, 40(11): 3273-3286.)

[2] 肖云发, 韩芳, 王青云. 基于脉冲强化学习和 CPG 的四足机器人分层运动控制[J]. 控制与决策, 2025, 40(7): 2070-2078.
(Xiao Y F, Han F, Wang Q Y. Hierarchical motion control of quadruped robot based on spiking reinforcement learning and CPG[J]. Control and Decision, 2025, 40(7): 2070-2078.)

[3] He Z M, Chen P Y, Shi H B, et al. Robotic locomotion skill learning using unsupervised reinforcement learning with controllable latent space partition[J]. IEEE Transactions on Industrial Informatics, 2025, 21(1): 902-911.

[4] 于绍琪, 田玉平. 基于 Petri 网与多智能体深度强化学习的 AGV 路径规划[J]. 控制与决策, 2025, 40(5): 1438-1446.
(Yu S Q, Tian Y P. AGV path planning based on Petri net and multi-agent deep reinforcement learning[J]. Control and Decision, 2025, 40(5): 1438-1446.)

[5] 李佩哲, 张文彪. 基于改进经验回放策略的路径规划算法[J]. 控制与决策, 2025, 40(8): 2545-2552.
(Li P Z, Zhang W B. A path planning algorithm based on improved experience replay strategy[J]. Control and Decision, 2025, 40(8): 2545-2552.)

[6] Feng X Y, Jiang L, Yu X D, et al. Curriculum goal-conditioned imitation for offline reinforcement learning[J]. IEEE Transactions on Games, 2024, 16(1): 102-112.

[7] 王雪松, 张恒瑞, 张佳志, 等. 基于优势约束扩散策略的离线强化学习[J]. 控制与决策, 2025, 40(6): 1903-1912.
(Wang X S, Zhang H R, Zhang J Z, et al. Offline reinforcement learning based on advantage-constrained diffusion policy[J]. Control and Decision, 2025, 40(6): 1903-1912.)

[8] Figueiredo Prudencio R, Maximo M R O A, Colombini E L. A survey on offline reinforcement learning: Taxonomy, review, and open problems[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(8): 10237-10257.

[9] Wu K, Zhao Y N, Xu Z Y, et al. ACL-QL: Adaptive conservative level in Q-learning for offline reinforcement learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(6): 11399-11413.

[10] Rezaeifar S, Dadashi R, Vieillard N, et al. Offline reinforcement learning as anti-exploration[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(7): 8106-8114.

[11] 顾扬, 程玉虎, 王雪松. 基于优先采样模型的离线强化学习[J]. 自动化学报, 2024, 50(1): 143-153.
(Gu Y, Cheng Y H, Wang X S. Offline reinforcement learning based on prioritized sampling model[J]. Acta Automatica Sinica, 2024, 50(1): 143-153.)

[12] Cheng Y H, Huang L Y, Philip Chen C L, et al. Robust actor-critic with relative entropy regulating actor[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(11): 9054-9063.

[13] Chen C, Ji X Y, Mao Y X, et al. Offline reinforcement learning with OOD state correction and OOD action suppression[C]. Advances in Neural Information Processing Systems 37. Vancouver, Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 93568-93601.

[14] Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration[C]. International Conference on Machine Learning., 2018.

[15] Wu Y F, Tucker G, Nachum O. Behavior regularized offline reinforcement learning[J/OL]. 2019, arXiv: 1911.11361.

[16] Ran Y H, Li Y C, Zhang F X, et al. Policy regularization with dataset constraint for offline reinforcement learning[C]. International Conference on Machine Learning., 2023.

[17] Fujimoto S, Gu S S. A minimalist approach to offline reinforcement learning[J/OL]. 2021, arXiv: 2106.06860.

[18] Peng X B, Kumar A, Zhang G, et al. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning[J/OL]. 2019, arXiv: 1910.00177.

[19] Kumar A, Zhou A, Tucker G, et al. Conservative Q-learning for offline reinforcement learning[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, New York: ACM, 2020: 1179-1191.

[20] Kostrikov I, Tompson J, Fergus R, et al. Offline reinforcement learning with fisher divergence critic regularization[J/OL]. 2021, arXiv: 2103.08050.

[21] Huang L Y, Dong B T, Zhang W D. Efficient offline reinforcement learning with relaxed conservatism[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(8): 5260-5272.

[22] Chen L T, Yan J, Shao Z D, et al. Conservative state value estimation for offline reinforcement learning[J/OL]. 2023, arXiv: 2302.06884.

[23] Deng Z H, Fu Z Y, Wang L X, et al. False correlation reduction for offline reinforcement learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(2): 1199-1211.

[24] Kostrikov I, Nair A, Levine S. Offline reinforcement learning with implicit Q-learning[J/OL]. 2021, arXiv: 2110.06169.

[25] Mao Y X, Wang Q, Qu Y, et al. Doubly mild generalization for offline reinforcement learning[C]. Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, 2024: 51436-51473.

[26] Ma C Z, Yang D Y, Wu T Y, et al. Improving offline

reinforcement learning with in-sample advantage regularization for robot manipulation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(6): 11215-11227.

[27] Xiong H Q, Xu T Y, Zhao L, et al. Deterministic policy gradient: convergence analysis[C]. Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence. Eindhoven, 2022: 2159-2169.

[28] Jin Y, Yang Z R, Wang Z R. Is pessimism provably efficient for offline RL[C]. Proceedings of International Conference on Machine Learning. Virtual, 2021: 5084-5096.

作者简介

杨露 (2001−), 女, 博士生, 主要研究方向为强化学习, E-mail: yanglu010607@163.com;

王雪松 (1974−), 女, 教授, 博士生导师, 主要研究方向为机器学习、具身智能, E-mail: wangxuesongcumt@163.com;

金可 (1997−), 男, 博士生, 主要研究方向为离线强化学习, E-mail: jkpopo97@163.com;

程玉虎 (1973−), 男, 教授, 博士生导师, 主要研究方向为机器学习、智能系统, E-mail: chengyuhu@163.com.