广义行为正则化离线Actor-Critic

程玉虎[1)] 黄龙阳[1)] 侯棣元[1)] 张佳志[1)] 陈俊龙[2)] 王雪松[1)]

[1)](中国矿业大学信息与控制工程学院 江苏 徐州 221116)

[2)](华南理工大学计算机科学与工程学院 广州 510006)

摘 要 行为正则化Actor-Critic(BRAC)是一种离线强化学习算法，通过将当前策略与行为策略之间的Kullback-Leibler(KL)散度作为策略目标函数的正则化项来缓解分布偏移问题.但是，由于KL散度是一种无界的分布差异度量，在策略差异过大时，策略目标函数中的累积期望回报项将仅对策略改进发挥有限的作用，从而导致最终学到的策略性能较差.针对该问题，将当前策略与行为策略之间的斜对称Jensen-Shannon(JS)散度作为策略目标函数的正则化项，提出了一种广义行为正则化离线Actor-Critic(GOACBR)算法.理论分析表明：由于斜对称JS散度有界，将其作为正则化项有助于降低策略性能差异.进一步，针对行为策略未知导致难以直接计算当前策略和行为策略间斜对称JS散度的问题，设计了一个辅助网络来对其进行间接估计.最后，给出了GOACBR的收敛性理论证明.在D4RL基准数据集上的评估结果表明：相较于BRAC，GOACBR在所有测试任务上获得的平均累积回报总和提升了289.8%.相关代码公布在https://github.com/houge1996/GOAC.

关键词 离线Actor-Critic；行为正则化；斜对称JS散度；分布偏移

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2023.00843

Generalized Offline Actor-Critic with Behavior Regularization

CHENG Yu-Hu[1)] HUANG Long-Yang[1)] HOU Di-Yuan[1)] ZHANG Jia-Zhi[1)]
CHEN Jun-Long[2)] WANG Xue-Song[1)]

[1)](*School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, Jiangsu* 221116)

[2)](*School Computer Science and Engineering, South China University of Technology, Guangzhou* 510006)

Abstract The behavior regularized actor-critic (BRAC) is an offline reinforcement learning algorithm. It alleviates the distribution shift problem by taking the Kullback-Leibler (KL) divergence between the current and behavior policies as the regularization term in the policy objective function. However, KL divergence is an unbounded measure of distribution difference. When the policy difference is too large, the expected cumulative return in the policy objective function will only play a limited role in the policy improvement, resulting in poor performance of the learned policy. To address the issue, we take the skew-symmetric Jensen-Shannon (JS) divergence between the current and behavior policies as the regularization term in the policy objective function and propose a generalized offline actor-critic with behavior regularization (GOACBR) algorithm. The theoretical analysis shows that since the skew-symmetric JS divergence is bounded, it's helpful to reduce the difference in policy performance by taking it as the regularization term. Furthermore, since it's difficult to directly calculate the skew-symmetric JS divergence between policies due that the behavior policy is unknown, an auxiliary neural

收稿日期：2022-08-01；在线发布日期：2022-12-29.本课题得到国家自然科学基金项目(62176259，61976215)、江苏省重点研发计划项目(BE2022095)资助.**程玉虎**，博士，教授，中国计算机学会(CCF)会员，主要研究领域为机器学习和智能系统.E-mail：chengyuhu@163.com.**黄龙阳**，硕士研究生，主要研究方向为强化学习.**侯棣元**，硕士研究生，主要研究方向为强化学习.**张佳志**，硕士研究生，主要研究方向为强化学习.**陈俊龙**，博士，教授，主要研究领域为数据科学和计算智能.**王雪松**(通信作者)，博士，教授，中国计算机学会(CCF)会员，主要研究领域为机器学习和模式识别.E-mail：wangxuesongcumt@163.com.

network is designed to indirectly estimate it. Finally, the convergence of GOACBR is proved theoretically. The performance of GOACBR is evaluated on the D4RL benchmark dataset. Compared with BRAC, the total average cumulative return achieved by GOACBR on all testing tasks has increased by 289.8%. The source code is available at https://github.com/houge1996/GOAC.

Keywords offline actor-critic; behavior regularization; skew-symmetric JS divergence; distribution shift

1 引 言

众所周知，强化学习是机器学习的一个分支[1-2]. 在强化学习中，智能体利用与环境交互所得的评价性反馈信号实现决策的优化[3-4]. 根据智能体在优化策略时能否与环境交互，可将强化学习分为在线强化学习与离线强化学习. 在在线强化学习中，智能体可以与环境或模拟器交互，并使用收集的经验数据优化策略[5-6]. 这意味着试验时需要从头开始收集大量数据，因此需要极高的经济成本，并且智能体的安全性很难得到保证[7-8]，限制了在线强化学习在复杂现实世界的应用. 离线强化学习，也称为批量强化学习，关注的是智能体如何根据固定的离线数据集学习到最优策略，而不需要与环境进行交互[9]. 因此，离线强化学习在现实世界中具有广泛的应用前景[10]. 然而，离线强化学习存在一个重大的挑战：分布偏移[11]. 分布偏移指的是在优化策略的过程中，更新当前策略时所使用的数据集中的数据分布与当前策略下的访问分布之间存在差异. 由于存在分布偏移，在估计Q函数时，那些未出现在数据集中的状态动作对的Q值会被错误地估计，进一步导致无法学习到最优策略[12]. 因此，如何缓解分布偏移对策略改进过程的影响成为离线强化学习领域的研究热点.

为了缓解分布偏移对策略改进过程的影响，Fujimoto等[13]率先提出了批约束深度Q学习(Batch-Constrained Deep Q-Learning，BCQ). BCQ使用条件变分自动编码器[14]估计行为策略，并利用估计得到的行为策略修正值函数网络的更新以降低值函数估计偏差. 然而，Kumar等[15]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[15]分析了分布偏移导致的自举误差，提出了一种名为自举累积误差消减(Bootstrapping Error Accumulation Reduction，BEAR)的算法来控制自举误差，该算法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[16](Maximum Mean Discrepancy，MMD)来使当前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. 行为正则化Actor-Critic[5](Behavior Regularized Actor-Critic，BRAC)将Kullback-Leibler(KL)散度[17]作为当前策略和行为策略之间的差异度量，有效降低了计算量，而且保证了策略改进过程的稳定性. 上述讨论中提及的离线强化学习算法克服分布偏移的思想为：在优化策略目标函数时，最小化当前策略与行为策略的差异[18]. 因此，上述算法需要准确地估计行为策略. 为了解决这个问题，Kumar等[19]提出了保守Q学习(Conservative Q-Learning，CQL). CQL不需要估计行为策略，而是通过添加额外的值函数网络损失函数正则化项来克服分布偏移的影响. 具体来说，通过设计合适的Q函数网络损失函数正则化项，使那些在行为策略分布之外的动作分配到更低的Q值. Kostrikov等[20]提出了隐式Q学习(Implicit Q-Learning，IQL). IQL使用期望回归拟合值函数并使用它来计算时序差分目标，而不需要在值函数训练过程中执行显式策略差异约束. Wu等[21]提出了不确定性加权Actor-Critic(Uncertainty Weighted Actor-Critic，UWAC)，该算法可以检测那些处于行为策略分布之外的状态动作对，并降低它们对策略优化过程的贡献. 然而，由于在值函数网络损失函数中添加了额外的正则化项，此类离线强化学习算法无法学习到最优值函数.

尽管上述方法能够在不同程度上缓解分布偏移对策略改进过程的影响，但是仍然存在下述问题. 一方面，当在值函数网络损失函数中添加额外的正则化项来降低分布偏移的影响时，往往不能学习到最优值函数，且不能保证策略改进过程的收敛性. 另一方面，当通过约束当前策略与行为策略之间的差异来降低分布偏移的影响时，尽管通过设计合适的策略目标函数能保证策略的收敛性，但策略仅能收敛到一个次优策略. 此外，尽管使用当前策略与行为策略之间的KL散度作为策略目标函数的正则化项

能保证策略改进过程的收敛性. 但是，由于KL散度是一种无界的分布差异度量，因此在策略改进的过程中，若当前策略与行为策略之间的差异过大，则策略目标函数中的累积期望回报项对策略性能的贡献就会很小，从而导致最终学到的策略性能较差.

受上述讨论启发，本文提出了广义行为正则化离线Actor-Critic，主要贡献包括：(1)将当前策略与行为策略之间的斜对称Jensen-Shannon(JS)散度[22]作为策略目标函数的正则化项，提出一种广义行为正则化Actor-Critic(Generalized Offline Actor-Critic with Behavior Regularization，GOACBR)；(2)设计了一个辅助网络来估计当前策略与行为策略之间的斜对称JS散度；(3)利用斜对称JS散度的有界性，在理论上证明了GOACBR的收敛性.

2 马尔科夫决策过程

在强化学习中，智能体与环境交互的过程可以用一个马尔科夫决策过程 $\mathcal{M}=(\mathcal{S},\mathcal{A},p,R,\gamma,d_0)$ 来表征，其中，$\mathcal{S}$和$\mathcal{A}$分别表示状态空间和动作空间，R是奖励函数，p是状态转移概率，$\gamma\in[0,1)$是折扣因子，d_0是初始状态分布. 智能体的目的是通过与环境在线交互学习一个最优策略$\pi^*_{\mathcal{M}}$使以下折扣累积期望回报最大：

$$J_{\mathcal{M}}(\pi)=\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_t|\boldsymbol{s}_0\sim d_0\right] \tag{1}$$

其中，$\pi\in\Theta$，Θ是一个封闭的策略空间，$r_t\triangleq R(\boldsymbol{s}_t,\boldsymbol{a}_t)$，$\tau$表示智能体和环境交互得到的轨迹. 在策略$\pi$下，状态值函数的定义为

$$V^{\pi}(\boldsymbol{s})=\mathbb{E}_{\tau\sim\pi}\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k}|\boldsymbol{s}_t=\boldsymbol{s}\right] \tag{2}$$

状态动作值函数的定义为

$$Q^{\pi}(\boldsymbol{s},\boldsymbol{a})=\mathbb{E}_{\tau\sim\pi}\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k}|\boldsymbol{s}_t=\boldsymbol{s},\boldsymbol{a}_t=\boldsymbol{a}\right] \tag{3}$$

由式(2)和(3)可知，$Q^{\pi}(\boldsymbol{s},\boldsymbol{a})$与$V^{\pi}(\boldsymbol{s})$满足：

$$Q^{\pi}(\boldsymbol{s},\boldsymbol{a})=R(\boldsymbol{s},\boldsymbol{a})+\gamma\mathbb{E}_{\boldsymbol{s}'\sim p}\left[V^{\pi}(\boldsymbol{s}')\right] \tag{4}$$

3 行为正则化Actor-Critic

在离线强化学习中，智能体不能与环境交互，其仅能使用一个数据集$\mathcal{B}=\{(\boldsymbol{s}_t,\boldsymbol{a}_t,r_t,\boldsymbol{s}_{t+1})\}$中的数据来学习使折扣累积期望回报(1)最大的策略. 数据集$\mathcal{B}$中的数据由行为策略生成，将行为策略记为$\pi_B(\cdot|\boldsymbol{s})$. 对于数据集$\mathcal{B}$中存储的任意一条轨迹$\tau=\{(\boldsymbol{s}_0,\boldsymbol{a}_0,r_0),(\boldsymbol{s}_1,\boldsymbol{a}_1,r_1),\cdots,(\boldsymbol{s}_T,\boldsymbol{a}_T,r_T)\}$，都有$\boldsymbol{a}_t\sim\pi_B(\cdot|\boldsymbol{s}_t)$，$\forall t\in\{0,1,2,\cdots,T\}$. 需要注意的是，在离线强化学习的设定中，行为策略$\pi_B(\cdot|\boldsymbol{s})$是未知的. 由于数据集$\mathcal{B}$是一个静态数据集，因此数据集$\mathcal{B}$中状态动作对的分布是固定的. 在策略学习的过程中，如果所学策略的访问分布与数据集$\mathcal{B}$中状态动作对的分布差异过大，将产生分布偏移问题，进而导致策略改进过程中Q函数估计不准确. 因此，在离线强化学习设定下，直接使用在线强化学习算法无法保证策略改进过程中策略性能的递增性. 为了克服分布偏移的影响，BRAC[5]考虑了如下代理策略目标函数：

$$J_{KL}(\pi)=\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t(r_t-\alpha_1\cdot KL\left(\pi(\cdot|\boldsymbol{s}_t)\|\,\pi_B(\cdot|\boldsymbol{s}_t)\right)|\boldsymbol{s}_0\sim d_0\right] \tag{5}$$

其中，$\alpha_1>0$，$KL\left(\pi(\cdot|\boldsymbol{s}_t)\|\,\pi_B(\cdot|\boldsymbol{s}_t)\right)=\mathbb{E}_{a\sim\pi(\cdot|\boldsymbol{s}_t)}\left[\ln\left(\frac{\pi(a|\boldsymbol{s}_t)}{\pi_B(a|\boldsymbol{s}_t)}\right)\right]$. 将最大化策略目标函数(5)的最优策略记为$\pi^*_{KL}$. 由于智能体的目标是学习使折扣累积期望回报最大的策略，因此，任给一个策略π，定义其性能指标为：

$$\eta(\pi)=\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_t|\boldsymbol{s}_0\sim d_0\right] \tag{6}$$

定理1. $\eta(\pi^*_{KL})$和$\eta(\pi^*_{\mathcal{M}})$满足$\eta(\pi^*_{KL})\leqslant\eta(\pi^*_{\mathcal{M}})$. 当且仅当$\pi^*_{KL}=\pi^*_{\mathcal{M}}$时，$\eta(\pi^*_{KL})=\eta(\pi^*_{\mathcal{M}})$.

证明. 根据式(1)和(6)可知$J_{\mathcal{M}}(\pi)=\eta(\pi)$，所以$\pi^*_{\mathcal{M}}=\mathrm{argmax}_{\pi\in\Theta}\eta(\pi)$. 因此，如果$\pi^*_{KL}=\pi^*_{\mathcal{M}}$，则$\eta(\pi^*_{KL})=\eta(\pi^*_{\mathcal{M}})$；若$\pi^*_{KL}\neq\pi^*_{\mathcal{M}}$，则$\eta(\pi^*_{KL})<\eta(\pi^*_{\mathcal{M}})$. 综上可得$\eta(\pi^*_{KL})\leqslant\eta(\pi^*_{\mathcal{M}})$. 证毕.

推论1. 当且仅当$\pi_B=\pi^*_{\mathcal{M}}$时，$\eta(\pi^*_{KL})=\eta(\pi^*_{\mathcal{M}})$成立.

证明. 根据式(1)、(6)和定理1容易证明该推论成立.

定理1和推论1表明：当且仅当$\pi_B=\pi^*_{\mathcal{M}}$时，策略$\pi^*_{KL}$是使折扣累积期望回报(1)最大的最优策略. 此外，直觉上，因为KL散度是无界的，若优化的策略π与行为策略π_B之间的差异过大，则优化策略目标函数(5)时KL散度正则化项对策略目标函数的贡献会很大，这会导致策略π^*_{KL}的性能值$\eta(\pi^*_{KL})$很小.

4 广义行为正则化离线Actor-Critic

4.1 广义行为正则化离线Actor-Critic的框架

为了缓解分布偏移对BRAC策略改进过程的影响，在GOACBR中考虑如下代理策略目标函数：

$$\bar{J}(\pi)=\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t(r_t-\alpha_2\cdot D_{JS}^{\beta}\left(\pi(\cdot|\boldsymbol{s}_t)\|\pi_B(\cdot|\boldsymbol{s}_t)\right))|\boldsymbol{s}_0\sim d_0\right] \tag{7}$$

其中，$\alpha_2>0$是权重参数，$D_{JS}^{\beta}\left(\pi(\cdot|\boldsymbol{s})\|\pi_B(\cdot|\boldsymbol{s})\right)$是斜对称JS散度，其定义为

$$D_{JS}^{\beta}\left(\pi(\cdot|\boldsymbol{s})\|\pi_B(\cdot|\boldsymbol{s})\right)=(1-\beta)\int_{a\in\mathcal{A}}\pi_B(\boldsymbol{a}|\boldsymbol{s})Y_{\pi}^{\pi_B}(\boldsymbol{s},\boldsymbol{a})+\beta\int_{a\in\mathcal{A}}\pi(\boldsymbol{a}|\boldsymbol{s})X_{\pi}^{\pi_B}(\boldsymbol{s},\boldsymbol{a}) \tag{8}$$

其中，$X_{\pi}^{\pi_B}(\boldsymbol{s},\boldsymbol{a})=\ln\left(\frac{\pi(\boldsymbol{a}|\boldsymbol{s})}{\beta\pi(\boldsymbol{a}|\boldsymbol{s})+(1-\beta)\pi_B(\boldsymbol{a}|\boldsymbol{s})}\right)$，$\beta\in(0,1)$，$Y_{\pi}^{\pi_B}(\boldsymbol{s},\boldsymbol{a})=\ln\left(\frac{\pi_B(\boldsymbol{a}|\boldsymbol{s})}{\beta\pi(\boldsymbol{a}|\boldsymbol{s})+(1-\beta)\pi_B(\boldsymbol{a}|\boldsymbol{s})}\right)$.

需要注意的是，斜对称JS散度$D_{JS}^{\beta}\left(\pi(\cdot|\boldsymbol{s})\|\pi_B(\cdot|\boldsymbol{s})\right)$是一种策略分布差异度量方式. 通过设定不同的参数$\beta$，可以得到不同形式的策略差异度量方式. 另外，注意到当$\beta=\frac{1}{2}$时，$D_{JS}^{\beta}\left(\pi(\cdot|\boldsymbol{s})\|\pi_B(\cdot|\boldsymbol{s})\right)$退化为JS散度. 因此，本文称基于策略目标函数(7)的Actor-Critic为广义行为正则化离线Actor-Critic.

定理2. 给定两个策略分布$\pi_1(\cdot|\boldsymbol{s})$和$\pi_2(\cdot|\boldsymbol{s})$，任取$\beta\in(0,1)$，以下关系成立：

$$0\leqslant D_{JS}^{\beta}\left(\pi_1(\cdot|\boldsymbol{s})\|\pi_2(\cdot|\boldsymbol{s})\right)\leqslant g(\beta) \tag{9}$$

其中，$g(\beta)=-\beta\ln\beta-(1-\beta)\ln(1-\beta)$.

证明. 根据KL散度的非负性可得：

$$D_{JS}^{\beta}\left(\pi_1(\cdot|\boldsymbol{s})\|\pi_2(\cdot|\boldsymbol{s})\right)=(1-\beta)KL\left(\pi_2(\cdot|\boldsymbol{s})\|\beta\pi_1(\cdot|\boldsymbol{s})+(1-\beta)\pi_2(\cdot|\boldsymbol{s})\right)+\beta KL\left(\pi_1(\cdot|\boldsymbol{s})\|\beta\pi_1(\cdot|\boldsymbol{s})+(1-\beta)\pi_2(\cdot|\boldsymbol{s})\right)\geqslant 0 \tag{10}$$

注意到$Y_{\pi_1}^{\pi_2}(\boldsymbol{s},\boldsymbol{a})\leqslant\ln\left(\frac{\pi_2(\boldsymbol{a}|\boldsymbol{s})}{(1-\beta)\pi_2(\boldsymbol{a}|\boldsymbol{s})}\right)=-\ln(1-\beta)$和$X_{\pi_1}^{\pi_2}(\boldsymbol{s},\boldsymbol{a})\leqslant\ln\left(\frac{\pi_1(\boldsymbol{a}|\boldsymbol{s})}{\beta\pi_1(\boldsymbol{a}|\boldsymbol{s})}\right)=-\ln\beta$，这表明：

$$\begin{aligned}D_{JS}^{\beta}\left(\pi_1\|\pi_2\right)&=\beta\sum_{a\in\mathcal{A}}\pi_1(\boldsymbol{a}|\boldsymbol{s})\cdot X_{\pi_1}^{\pi_2}(\boldsymbol{s},\boldsymbol{a})+(1-\beta)\sum_{a\in\mathcal{A}}\pi_2(\boldsymbol{a}|\boldsymbol{s})\cdot Y_{\pi_1}^{\pi_2}(\boldsymbol{s},\boldsymbol{a})\\&\leqslant-\beta\sum_{a\in\mathcal{A}}\pi_1(\boldsymbol{a}|\boldsymbol{s})\cdot\ln\beta-(1-\beta)\times\sum_{a\in\mathcal{A}}\pi_2(\boldsymbol{a}|\boldsymbol{s})\cdot\ln(1-\beta)\\&=-\beta\ln\beta-(1-\beta)\ln(1-\beta)\end{aligned} \tag{11}$$

证毕.

定理2表明：D_{JS}^{β}的上界为$g(\beta)=-\beta\ln\beta-(1-\beta)\ln(1-\beta)$. 因此，在优化策略时，$\mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty}\gamma^t D_{JS}^{\beta}(\pi(\cdot|\boldsymbol{s}_t)\|\pi_B(\cdot|\boldsymbol{s}_t))]$对策略目标函数(7)的贡献是有限的. 这保证了折扣累积期望回报项对策略目标函数(7)的贡献.

接下来，为求解使策略目标函数(7)最大的最优策略$\tilde{\pi}^*$，给出一种策略迭代算法. 首先，定义策略π下的状态值函数为：

$$\bar{V}^{\pi}(\boldsymbol{s})=\mathbb{E}_{a\sim\pi(\cdot|\boldsymbol{s})}\left[\bar{Q}^{\pi}(\boldsymbol{s},\boldsymbol{a})-\alpha_2\beta\cdot X_{\pi}^{\pi_B}(\boldsymbol{s},\boldsymbol{a})\right]-\alpha_2(1-\beta)\cdot\mathbb{E}_{a\sim\pi_B(\cdot|\boldsymbol{s})}\left[Y_{\pi}^{\pi_B}(\boldsymbol{s},\boldsymbol{a})\right] \tag{12}$$

其中，Q函数$\bar{Q}^{\pi}(\boldsymbol{s},\boldsymbol{a})$满足：

$$\bar{Q}^{\pi}(\boldsymbol{s},\boldsymbol{a})=R(\boldsymbol{s},\boldsymbol{a})+\gamma\mathbb{E}_{s'\sim p(\cdot|\boldsymbol{s},\boldsymbol{a})}\left[\bar{V}^{\pi}(\boldsymbol{s}')\right] \tag{13}$$

引理1. 考虑如下贝尔曼算子：

$$\mathcal{Z}^{\pi}Q(\boldsymbol{s},\boldsymbol{a})=R(\boldsymbol{s},\boldsymbol{a})+\gamma\mathbb{E}_{s'\sim p}\left[V(\boldsymbol{s}')\right] \tag{14}$$

可以得到$\lim_{k\to\infty}Q_k(\boldsymbol{s},\boldsymbol{a})=\bar{Q}^{\pi}(\boldsymbol{s},\boldsymbol{a})$，其中，$V(\boldsymbol{s})=\mathbb{E}_{a\sim\pi(\cdot|\boldsymbol{s})}\left[Q(\boldsymbol{s},\boldsymbol{a})-\alpha_2\beta\cdot X_{\pi}^{\pi_B}(\boldsymbol{s},\boldsymbol{a})\right]-\alpha_2(1-\beta)\times\mathbb{E}_{a\sim\pi_B(\cdot|\boldsymbol{s})}\left[Y_{\pi}^{\pi_B}(\boldsymbol{s},\boldsymbol{a})\right]$，$Q_k(\boldsymbol{s},\boldsymbol{a})=(\mathcal{Z}^{\pi})^kQ(\boldsymbol{s},\boldsymbol{a})$.

证明. 考虑两个不同的Q函数Q_{a}和Q_{b}，可以得到：

$$\begin{aligned}&\left|\mathcal{Z}^{\pi}Q_{\mathrm{a}}(\boldsymbol{s}_t,\boldsymbol{a}_t)-\mathcal{Z}^{\pi}Q_{\mathrm{b}}(\boldsymbol{s}_t,\boldsymbol{a}_t)\right|\\&=\gamma\left|\mathbb{E}_{s_{t+1}\sim p}\left[V_{\mathrm{a}}(\boldsymbol{s}_{t+1})\right]-\mathbb{E}_{s_{t+1}\sim p}\left[V_{\mathrm{b}}(\boldsymbol{s}_{t+1})\right]\right|\\&=\gamma\left|\mathbb{E}_{\substack{s_{t+1}\sim p\\a_{t+1}\sim\pi}}\left[Q_{\mathrm{a}}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})-\alpha_2\beta X_{\pi}^{\pi_B}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})\right]-\alpha_2(1-\beta)\cdot\mathbb{E}_{\substack{s_{t+1}\sim p\\a_{t+1}\sim\pi_B}}\left[Y_{\pi}^{\pi_B}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})\right]-\right.\\&\left.\left\{\mathbb{E}_{\substack{s_{t+1}\sim p\\a_{t+1}\sim\pi}}\left[Q_{\mathrm{b}}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})-\alpha_2\beta\cdot X_{\pi}^{\pi_B}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})\right]-\alpha_2(1-\beta)\mathbb{E}_{\substack{s_{t+1}\sim p\\a_{t+1}\sim\pi_B}}\left[Y_{\pi}^{\pi_B}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})\right]\right\}\right|\\&\leqslant\gamma\max_{s_{t+1},a_{t+1}}\left|Q_{\mathrm{a}}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})-Q_{\mathrm{b}}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})\right|\\&=\gamma|Q_{\mathrm{a}}-Q_{\mathrm{b}}|\end{aligned} \tag{15}$$

式(15)表明$\mathcal{Z}^{\pi}$是一个压缩算子，因此序列$\{(\mathcal{Z}^{\pi})^{i}Q(\boldsymbol{s},\boldsymbol{a})\}_{i=0}^{k}$收敛到唯一的不动点$\bar{Q}^{\pi}(\boldsymbol{s},\boldsymbol{a})$.

证毕.

引理1表明：在策略评估阶段，从初始$Q_0(\boldsymbol{s},\boldsymbol{a})$起，根据迭代规则$Q_{k+1}(\boldsymbol{s},\boldsymbol{a})=\mathcal{Z}^{\pi_{\text{old}}}Q_k(\boldsymbol{s},\boldsymbol{a})$重复迭代可以估计策略$\pi_{\text{old}}$的Q函数$\bar{Q}^{\pi_{\text{old}}}(\boldsymbol{s},\boldsymbol{a})$. 在策略改进阶段，根据以下策略更新规则将$\pi_{\text{old}}$更新为$\pi_{\text{new}}$：

$$\pi_{\text{new}}(\cdot|\boldsymbol{s})=\underset{\pi\in\Theta}{\operatorname{argmin}}\, L_{\pi_{\text{old}}}\big(\pi(\cdot|\boldsymbol{s})\big) \tag{16}$$

其中，

$$L_{\pi_{\text{old}}}\big(\pi(\cdot|\boldsymbol{s})\big)=\mathbb{E}_{a\sim\pi}\big[-\bar{Q}^{\pi_{\text{old}}}(\boldsymbol{s},\boldsymbol{a})+\alpha_2\beta\cdot X_{\pi}^{\pi_B}(\boldsymbol{s},\boldsymbol{a})\big]+\alpha_2(1-\beta)\mathbb{E}_{a\sim\pi_B}\big[Y_{\pi}^{\pi_B}(\boldsymbol{s},\boldsymbol{a})\big]$$

引理2. 任取$(\boldsymbol{s},\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}$，$\bar{Q}^{\pi_{\text{new}}}(\boldsymbol{s},\boldsymbol{a})\geqslant\bar{Q}^{\pi_{\text{old}}}(\boldsymbol{s},\boldsymbol{a})$成立.

证明. 根据式(16)可知$L_{\pi_{\text{old}}}(\pi_{\text{old}}(\cdot|\boldsymbol{s}))\geqslant L_{\pi_{\text{old}}}(\pi_{\text{new}}(\cdot|\boldsymbol{s}))$，这表明：

$$\begin{aligned}
&\mathbb{E}_{a\sim\pi_{\text{new}}}\big[\bar{Q}^{\pi_{\text{old}}}(\boldsymbol{s},\boldsymbol{a})-\alpha_2\beta X_{\pi_{\text{new}}}^{\pi_B}(\boldsymbol{s},\boldsymbol{a})\big]-\alpha_2(1-\beta)\mathbb{E}_{a\sim\pi_B}\big[Y_{\pi_{\text{new}}}^{\pi_B}(\boldsymbol{s},\boldsymbol{a})\big]\\
&\geqslant\mathbb{E}_{a\sim\pi_{\text{old}}}\big[\bar{Q}^{\pi_{\text{old}}}(\boldsymbol{s},\boldsymbol{a})-\alpha_2\beta X_{\pi_{\text{old}}}^{\pi_B}(\boldsymbol{s},\boldsymbol{a})\big]-\alpha_2(1-\beta)\mathbb{E}_{a\sim\pi_B}\big[Y_{\pi_{\text{old}}}^{\pi_B}(\boldsymbol{s},\boldsymbol{a})\big]\\
&=\bar{V}^{\pi_{\text{old}}}(\boldsymbol{s})
\end{aligned}\tag{17}$$

结合式(13)和(17)，可以得到：

$$\begin{aligned}
\bar{Q}^{\pi_{\text{old}}}(\boldsymbol{s}_t,\boldsymbol{a}_t)&=r_t+\gamma\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p}\big[\bar{V}^{\pi_{\text{old}}}(\boldsymbol{s}_{t+1})\big]\\
&=r_t+\gamma\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p}\Big\{\mathbb{E}_{\boldsymbol{a}_{t+1}\sim\pi_{\text{old}}}\big[\bar{Q}^{\pi_{\text{old}}}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})-\alpha_2\beta X_{\pi_{\text{old}}}^{\pi_B}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})\big]-\alpha_2(1-\beta)\mathbb{E}_{\boldsymbol{a}_{t+1}\sim\pi_B}\big[Y_{\pi_{\text{old}}}^{\pi_B}(\boldsymbol{s}_{t+1},\boldsymbol{a})\big]\Big\}\\
&\leqslant r_t+\gamma\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p}\Big\{\mathbb{E}_{\boldsymbol{a}_{t+1}\sim\pi_{\text{new}}}\big[\bar{Q}^{\pi_{\text{old}}}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})-\alpha_2\beta\cdot X_{\pi_{\text{new}}}^{\pi_B}(\boldsymbol{s}_{t+1},\boldsymbol{a})\big]-\alpha_2(1-\beta)\mathbb{E}_{a\sim\pi_B}\big[Y_{\pi_{\text{new}}}^{\pi_B}(\boldsymbol{s}_{t+1},\boldsymbol{a})\big]\Big\}\\
&\leqslant r_t+\gamma\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p}\big[\bar{V}^{\pi_{\text{new}}}(\boldsymbol{s}_{t+1})\big]\\
&=\bar{Q}^{\pi_{\text{new}}}(\boldsymbol{s}_t,\boldsymbol{a}_t)
\end{aligned}\tag{18}$$

证毕.

定理3. 令$\{\pi_i\}_{i=0}^{k}$表示重复执行策略评估和策略改进得到的一个策略序列，则当k趋于无穷时，该策略序列收敛到最优策略$\tilde{\pi}^*$，即$\bar{Q}^{\tilde{\pi}^*}(\boldsymbol{s},\boldsymbol{a})\geqslant\bar{Q}^{\pi}(\boldsymbol{s},\boldsymbol{a})$，$\forall\pi\in\Theta$.

证明. 根据引理2可知，序列$\{\bar{Q}^{\pi_i}\}_{i=0}^{k}$单调递增. 由于$R$和$D_{JS}^{\beta}(\pi\|\pi_B)$是有界的，因此$\bar{Q}^{\pi}$有界. 这个事实表明：序列$\{\pi_i\}_{i=0}^{k}$会收敛到一个策略$\pi^{\tau}$. 接下来，进一步表明$\pi^{\tau}$是最优策略$\tilde{\pi}^*$. 显然，在收敛时，任取$\pi\in\Theta$，下式一定成立：

$$\mathbb{E}_{a\sim\pi^{\tau}(\cdot|\boldsymbol{s})}\big[\bar{Q}^{\pi^{\tau}}(\boldsymbol{s},\boldsymbol{a})\big]-\alpha_2 D_{JS}^{\beta}\big(\pi^{\tau}(\cdot|\boldsymbol{s})\|\pi_B(\cdot|\boldsymbol{s})\big)\geqslant\mathbb{E}_{a\sim\pi(\cdot|\boldsymbol{s})}\big[\bar{Q}^{\pi}(\boldsymbol{s},\boldsymbol{a})\big]-\alpha_2 D_{JS}^{\beta}\big(\pi(\cdot|\boldsymbol{s})\|\pi_B(\cdot|\boldsymbol{s})\big) \tag{19}$$

这表明：

$$\begin{aligned}
\bar{Q}^{\pi^{\tau}}(\boldsymbol{s}_t,\boldsymbol{a}_t)&=r_t+\gamma\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p}\big[\bar{V}^{\pi^{\tau}}(\boldsymbol{s}_{t+1})\big]\\
&=r_t+\gamma\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p}\Big\{\mathbb{E}_{\boldsymbol{a}_{t+1}\sim\pi^{\tau}}\big[\bar{Q}^{\pi^{\tau}}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})-\alpha_2\beta X_{\pi^{\tau}}^{\pi_B}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})\big]-\alpha_2(1-\beta)\mathbb{E}_{\boldsymbol{a}_{t+1}\sim\pi_B}\big[Y_{\pi^{\tau}}^{\pi_B}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})\big]\Big\}\\
&\geqslant r_t+\gamma\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p}\Big\{\mathbb{E}_{\boldsymbol{a}_{t+1}\sim\pi^{\tau}}\big[\bar{Q}^{\pi_{\text{old}}}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})-\alpha_2\beta X_{\pi^{\tau}}^{\pi_B}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})\big]-\alpha_2(1-\beta)\mathbb{E}_{\boldsymbol{a}_{t+1}\sim\pi_B}\big[Y_{\pi^{\tau}}^{\pi_B}(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})\big]\Big\}\\
&\quad\vdots\\
&\geqslant r_t+\gamma\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p}\big[\bar{V}^{\pi}(\boldsymbol{s}_{t+1})\big]\\
&=\bar{Q}^{\pi}(\boldsymbol{s}_t,\boldsymbol{a}_t).
\end{aligned}\tag{20}$$

因此，π^{τ}是最优策略$\tilde{\pi}^*$.　　证毕.

4.2 广义行为正则化离线Actor-Critic的实现

当使用神经网络逼近方法去逼近Q函数和策略π时，分别记Q函数网络和策略网络的网络参数为$\boldsymbol{\omega}$和$\boldsymbol{\theta}$. 接下来，为了分析方便，令：

$$\Phi_{\pi_B}^{\pi_\theta}(\boldsymbol{s},\boldsymbol{a})=\frac{\beta\pi_\theta(\boldsymbol{a}|\boldsymbol{s})}{\beta\pi_\theta(\boldsymbol{a}|\boldsymbol{s})+(1-\beta)\pi_B(\boldsymbol{a}|\boldsymbol{s})} \tag{21}$$

$$\bar{\Phi}_{\pi_B}^{\pi_\theta}(\boldsymbol{s},\boldsymbol{a})=\frac{(1-\beta)\pi_B(\boldsymbol{a}|\boldsymbol{s})}{\beta\pi_\theta(\boldsymbol{a}|\boldsymbol{s})+(1-\beta)\pi_B(\boldsymbol{a}|\boldsymbol{s})} \tag{22}$$

根据式(8)、(21)与(22)，可得：

$$\begin{aligned}
D_{JS}^{\beta}\big(\pi_\theta(\cdot|\boldsymbol{s})\big\|\,\pi_B(\cdot|\boldsymbol{s})\big)=&\\
&\beta\cdot\mathbb{E}_{a\sim\pi}\big[\ln\big(\Phi_{\pi_B}^{\pi_\theta}(\boldsymbol{s},\boldsymbol{a})\big)\big]+\\
&(1-\beta)\mathbb{E}_{a\sim\pi_B}\big[\ln\big(\bar{\Phi}_{\pi_B}^{\pi_\theta}(\boldsymbol{s},\boldsymbol{a})\big)\big]-\\
&\beta\ln\beta-(1-\beta)\ln(1-\beta)
\end{aligned}\tag{23}$$

由式(23)可得策略网络的损失函数为

$$\begin{aligned}
L_\pi(\boldsymbol{\theta})=&\mathbb{E}_{a\sim\pi_\theta}\big[-Q_{\boldsymbol{\omega}}(\boldsymbol{s},\boldsymbol{a})+\alpha_2\beta\ln\big(\Phi_{\pi_B}^{\pi_\theta}(\boldsymbol{s},\boldsymbol{a})\big)\big]+\\
&\alpha_2(1-\beta)\mathbb{E}_{a\sim\pi_B}\big[\ln\big(\bar{\Phi}_{\pi_B}^{\pi_\theta}(\boldsymbol{s},\boldsymbol{a})\big)\big]+g(\beta)
\end{aligned}\tag{24}$$

显然，常数项$g(\beta)$不会影响策略网络参数的优化，故可将式(24)转化为

$$\begin{aligned}
L_\pi(\boldsymbol{\theta})=&\mathbb{E}_{a\sim\pi_\theta}\big[-Q_{\boldsymbol{\omega}}(\boldsymbol{s},\boldsymbol{a})+\alpha_2\beta\ln\big(\Phi_{\pi_B}^{\pi_\theta}(\boldsymbol{s},\boldsymbol{a})\big)\big]+\\
&\alpha_2(1-\beta)\mathbb{E}_{a\sim\pi_B}\big[\ln\big(\bar{\Phi}_{\pi_B}^{\pi_\theta}(\boldsymbol{s},\boldsymbol{a})\big)\big]
\end{aligned}\tag{25}$$

需要注意的是，由于行为策略$\pi_B(\cdot|s)$未知，因此无法直接通过最小化策略损失函数(25)来更新策略网络参数.接下来，为了解决这个问题，首先给出定理4.

定理4. 对于$\Phi_{\pi_B}^{\pi_\theta}(s,a)$和$\bar{\Phi}_{\pi_B}^{\pi_\theta}(s,a)$，在$\theta=\theta_{\text{old}}$处，以下关系成立：

$$\begin{aligned}&\nabla_\theta\Big\{\beta\mathbb{E}_{a\sim\pi_\theta}\Big[\ln\big(\Phi_{\pi_B}^{\pi_\theta}(s,a)\big)\Big]+\\&\quad(1-\beta)\mathbb{E}_{a\sim\pi_B}\Big[\ln\big(\bar{\Phi}_{\pi_B}^{\pi_\theta}(s,a)\big)\Big]\Big\}\Big|_{\theta=\theta_{\text{old}}}\\&=\beta\nabla_\theta\mathbb{E}_{a\sim\pi_\theta}\Big[\ln\big(\Phi_{\pi_B}^{\pi_{\theta_{\text{old}}}}(s,a)\big)\Big]\Big|_{\theta=\theta_{\text{old}}}\end{aligned}\tag{26}$$

其中，θ_{old}是式(16)中策略π_{old}的参数.

证明. 首先，根据式(21)和(22)可知：

$$\bar{\Phi}_{\pi_B}^{\pi_\theta}(s,a)=1-\Phi_{\pi_B}^{\pi_\theta}(s,a)\tag{27}$$

所以，可以得到：

$$\mathbb{E}_{a\sim\pi_B}\Big[\ln\big(\bar{\Phi}_{\pi_B}^{\pi_\theta}(s,a)\big)\Big]=\mathbb{E}_{a\sim\pi_B}\Big[\ln\big(1-\Phi_{\pi_B}^{\pi_\theta}(s,a)\big)\Big]\tag{28}$$

分别求式(28)等号两边对参数θ的梯度得：

$$\begin{aligned}&\nabla_\theta\mathbb{E}_{a\sim\pi_B}\Big[\ln\big(\bar{\Phi}_{\pi_B}^{\pi_\theta}(s,a)\big)\Big]\\&=\mathbb{E}_{a\sim\pi_B}\left[\frac{\nabla_\theta\big(1-\Phi_{\pi_B}^{\pi_\theta}(s,a)\big)}{1-\Phi_{\pi_B}^{\pi_\theta}(s,a)}\right]\\&=-\mathbb{E}_{a\sim\pi_B}\left[\frac{\beta\pi_\theta(a|s)+(1-\beta)\pi_B(a|s)}{(1-\beta)\pi_B(a|s)}\right]\times\\&\nabla_\theta\Phi_{\pi_B}^{\pi_\theta}(s,a)\Big]\end{aligned}\tag{29}$$

使用重要性采样方法，进一步可以得到：

$$\begin{aligned}&\nabla_\theta\mathbb{E}_{a\sim\pi_B}\Big[\ln\big(\bar{\Phi}_{\pi_B}^{\pi_\theta}(s,a)\big)\Big]=\\&\quad-\mathbb{E}_{a\sim\pi_\theta}\left[\frac{\beta\pi_\theta(a|s)+(1-\beta)\pi_B(a|s)}{(1-\beta)\pi_\theta(a|s)}\times\right.\\&\quad\left.\nabla_\theta\Phi_{\pi_B}^{\pi_\theta}(s,a)\right]\\&\quad=-\frac{\beta}{1-\beta}\mathbb{E}_{a\sim\pi_\theta}\left[\frac{\nabla_\theta\Phi_{\pi_B}^{\pi_\theta}(s,a)}{\Phi_{\pi_B}^{\pi_\theta}(s,a)}\right]\end{aligned}\tag{30}$$

对于$\mathbb{E}_{a\sim\pi_\theta}\Big[\ln\big(\Phi_{\pi_B}^{\pi_\theta}(s,a)\big)\Big]$，容易得到：

$$\begin{aligned}&\nabla_\theta\mathbb{E}_{a\sim\pi_\theta}\Big[\ln\big(\Phi_{\pi_B}^{\pi_\theta}(s,a)\big)\Big]\\&=\mathbb{E}_{a\sim\pi_\theta}\Big[\ln\big(\Phi_{\pi_B}^{\pi_\theta}(s,a)\big)\cdot\nabla_\theta\ln\big(\pi_\theta(a|s)\big)\Big]+\\&\mathbb{E}_{a\sim\pi_\theta}\Big[\nabla_\theta\ln\big(\Phi_{\pi_B}^{\pi_\theta}(s,a)\big)\Big]\end{aligned}\tag{31}$$

式(31)表明：当$\theta=\theta_{\text{old}}$时，有：

$$\begin{aligned}&\nabla_\theta\mathbb{E}_{a\sim\pi_\theta}\Big[\ln\big(\Phi_{\pi_B}^{\pi_\theta}(s,a)\big)\Big]\\&=\nabla_\theta\mathbb{E}_{a\sim\pi_\theta}\Big[\ln\big(\Phi_{\pi_B}^{\pi_{\theta_{\text{old}}}}(s,a)\big)\Big]+\\&\mathbb{E}_{a\sim\pi_\theta}\Big[\nabla_\theta\ln\big(\Phi_{\pi_B}^{\pi_\theta}(s,a)\big)\Big]\end{aligned}\tag{32}$$

综上可得：

$$\begin{aligned}&\nabla_\theta\Big\{\beta\mathbb{E}_{a\sim\pi_\theta}\Big[\ln\big(\Phi_{\pi_B}^{\pi_\theta}(s,a)\big)\Big]+\\&\quad(1-\beta)\mathbb{E}_{a\sim\pi_B}\Big[\ln\big(\bar{\Phi}_{\pi_B}^{\pi_\theta}(s,a)\big)\Big]\Big\}\\&=\beta\nabla_\theta\mathbb{E}_{a\sim\pi_\theta}\Big[\ln\big(\Phi_{\pi_B}^{\pi_{\theta_{\text{old}}}}(s,a)\big)\Big]+\\&\beta\mathbb{E}_{a\sim\pi_\theta}\Big[\nabla_\theta\ln\big(\Phi_{\pi_B}^{\pi_\theta}(s,a)\big)\Big]-\\&\beta\mathbb{E}_{a\sim\pi_\theta}\left[\frac{\nabla_\theta\Phi_{\pi_B}^{\pi_\theta}(s,a)}{\Phi_{\pi_B}^{\pi_\theta}(s,a)}\right]\\&=\beta\nabla_\theta\mathbb{E}_{a\sim\pi_\theta}\Big[\ln\big(\Phi_{\pi_B}^{\pi_{\theta_{\text{old}}}}(s,a)\big)\Big]\end{aligned}\tag{33}$$

证毕.

考虑如下代理策略网络损失函数：

$$J(\theta)=\mathbb{E}_{a\sim\pi_\theta}\Big[-Q_\omega(s,a)+\alpha_2\beta\ln\big(\Phi_{\pi_B}^{\pi_{\theta_{\text{old}}}}(s,a)\big)\Big]\tag{34}$$

根据定理4可知，策略损失函数(25)与(34)在$\theta=\theta_{\text{old}}$处梯度相同.因此，在优化策略网络参数时通过最小化代理策略网络损失函数(34)来更新策略网络的网络参数.由于行为策略$\pi_B(a|s)$未知，故需要去估计$\Phi_{\pi_B}^{\pi_\theta}(s,a)$.为了推导出估计$\Phi_{\pi_B}^{\pi_\theta}(s,a)$的方法，首先给出以下定理.

定理5. 考虑函数$f(x)=\beta a\cdot\ln x+(1-\beta)b\cdot\ln(1-x)$，其中，$a,b,x,\beta\in(0,1)$，则$f(x)$在$x=\dfrac{\beta a}{\beta a+(1-\beta)b}$处取得最大值$\beta a\cdot\ln\left(\dfrac{\beta a}{\beta a+(1-\beta)b}\right)+(1-\beta)b\cdot\ln\left(\dfrac{(1-\beta)b}{\beta a+(1-\beta)b}\right)$.

证明. 求$f(x)$对x的导数得：

$$\frac{df(x)}{dx}=\frac{\beta a-\big((1-\beta)b+\beta a\big)x}{x(1-x)}\tag{35}$$

由于$a,b,x,\beta\in(0,1)$，故容易得到$f(x)$在$x=\dfrac{\beta a}{\beta a+(1-\beta)b}$处取得最大值：

$$\begin{aligned}f_{\max}(x)=&\beta a\cdot\ln\left(\frac{\beta a}{\beta a+(1-\beta)b}\right)+\\&(1-\beta)b\cdot\ln\left(\frac{(1-\beta)b}{\beta a+(1-\beta)b}\right)\end{aligned}$$

证毕.

接下来，考虑以下函数：

$$\begin{aligned}g\big(\Phi(s_t,\cdot)\big)=&(1-\beta)\mathbb{E}_{a\sim\pi_B}\Big[\ln\big(1-\Phi(s_t,a)\big)\Big]+\\&\beta\mathbb{E}_{a\sim\pi}\Big[\ln\big(\Phi(s_t,a)\big)\Big]\end{aligned}\tag{36}$$

进一步，将式(36)展开可得到：

$$g\left(\Phi(\boldsymbol{s}_t,\cdot)\right)=\int_{\boldsymbol{a}}\Big[\beta\pi(\boldsymbol{a}|\boldsymbol{s}_t)\cdot\ln\left(\Phi(\boldsymbol{s}_t,\boldsymbol{a})\right)+(1-\beta)\pi_B(\boldsymbol{a}|\boldsymbol{s}_t)\cdot\ln\left(1-\Phi(\boldsymbol{s}_t,\boldsymbol{a})\right)\Big]d\boldsymbol{a}\quad(37)$$

根据定理5可知，当$\Phi(\boldsymbol{s}_t,\boldsymbol{a})=\dfrac{\beta\pi(\boldsymbol{a}|\boldsymbol{s}_t)}{\beta\pi(\boldsymbol{a}|\boldsymbol{s}_t)+(1-\beta)\pi_B(\boldsymbol{a}|\boldsymbol{s}_t)}$时，被积函数$\beta\pi(a|s_t)\cdot\ln\left(\Phi(\boldsymbol{s}_t,\boldsymbol{a})\right)+(1-\beta)\pi_B(\boldsymbol{a}|\boldsymbol{s}_t)\cdot\ln\left(1-\Phi(\boldsymbol{s}_t,\boldsymbol{a})\right)$取最大值.因此，对于任意的$\boldsymbol{a}\in\mathcal{A}$，当且仅当$\Phi(\boldsymbol{s}_t,\boldsymbol{a})=\dfrac{\beta\pi(\boldsymbol{a}|\boldsymbol{s}_t)}{\beta\pi(\boldsymbol{a}|\boldsymbol{s}_t)+(1-\beta)\pi_B(\boldsymbol{a}|\boldsymbol{s}_t)}$时，$g\left(\Phi(\boldsymbol{s}_t,\cdot)\right)$取最大值$g^*\left(\Phi(\boldsymbol{s}_t,\cdot)\right)$.值得注意的是：

$$D_{JS}^{\beta}\left(\pi(\cdot|\boldsymbol{s}_t)\|\pi_B(\cdot|\boldsymbol{s}_t)\right)=g^*\left(\Phi(\boldsymbol{s}_t,\cdot)\right)\quad(38)$$

这表明可以通过最大化$g\left(\Phi(\boldsymbol{s}_t,\cdot)\right)$来估计$\Phi_{\pi_B}^{\pi_\theta}$.

在GOACBR中，使用一个神经网络来估计$\Phi_{\pi_B}^{\pi_\theta}$函数，并称其为$\Phi$函数网络，该神经网络的网络参数为$\boldsymbol{\varphi}$.$\Phi$函数网络的损失函数为

$$L_\Phi(\boldsymbol{\varphi})=-\mathbb{E}_{(\boldsymbol{s}_t,\boldsymbol{a}_t)\sim\mathcal{B}}\Big[\beta\cdot\ln\left(\Phi_\varphi(\boldsymbol{s}_t,\bar{\boldsymbol{a}}_t)\right)+(1-\beta)\cdot\ln\left(1-\Phi_\varphi(\boldsymbol{s}_t,\boldsymbol{a}_t)\right)\Big],\bar{\boldsymbol{a}}_t\sim\pi_\theta.\quad(39)$$

根据代理策略网络损失函数(34)可知，策略网络的损失函数为

$$L_\pi(\boldsymbol{\theta})=\mathbb{E}_{\boldsymbol{s}_t\sim\mathcal{B}}\Big[-Q_{\boldsymbol{\omega}}(\boldsymbol{s}_t,\bar{\boldsymbol{a}}_t)+\alpha_2\beta\cdot\ln\left(\Phi_\varphi(\boldsymbol{s}_t,\bar{\boldsymbol{a}}_t)\right)\Big],\bar{\boldsymbol{a}}_t\sim\pi_\theta.\quad(40)$$

为了保证最终习得策略的性能，在更新网络参数的过程中，GOACBR逐步降低正则化项$\alpha_2\beta\ln\left(\Phi_\varphi(\boldsymbol{s}_t,\bar{\boldsymbol{a}}_t)\right)$对策略网络训练过程的影响.具体来说，在更新网络参数的过程中，β值随策略网络训练步数周期性降低.在估计Q函数时，为了缓解Q函数高估问题，GOACBR引入两个Q函数网络.Q函数网络的损失函数为

$$L_Q(\boldsymbol{\omega}_i)=\mathbb{E}_{(\boldsymbol{s}_t,\boldsymbol{a}_t)\sim\mathcal{B}}\Big[\left(y_t-Q_{\boldsymbol{\omega}_i}(\boldsymbol{s}_t,\boldsymbol{a}_t)\right)^2\Big]\quad(41)$$

其中，

$$y(\boldsymbol{s}_t,\boldsymbol{a}_t)=r_t+\gamma\Big[\min\left\{Q_{\bar{\boldsymbol{\omega}}_1}(\boldsymbol{s}_{t+1},\bar{\boldsymbol{a}}_{t+1}),Q_{\bar{\boldsymbol{\omega}}_2}(\boldsymbol{s}_{t+1},\bar{\boldsymbol{a}}_{t+1})\right\}-\alpha_2(1-\beta)\cdot\ln\left(1-\Phi_\varphi\left((\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1})\sim\mathcal{B}\right)\right)-\bar{\beta}\cdot\ln\left(\Phi_\varphi(\boldsymbol{s}_{t+1},\bar{\boldsymbol{a}}_{t+1})\right)\Big],$$

$\bar{\boldsymbol{a}}_{t+1}\sim\pi_\theta$，$\boldsymbol{\omega}_i(i=1,2)$表示第$i$个Q函数网络的网络参数，$\bar{\boldsymbol{\omega}}_i$表示第$i$个目标Q函数网络的网络参数，$\bar{\boldsymbol{\theta}}$表示目标策略网络的网络参数.算法1给出了GOACBR的伪代码.

算法1. GOACBR.

输入：数据集$\mathcal{B}=\left\{(\boldsymbol{s}_t,\boldsymbol{a}_t,r_t,\boldsymbol{s}_{t+1})\right\}_N$

输出：π_θ.

初始化：Q函数网络参数$\boldsymbol{\omega}_1$、$\boldsymbol{\omega}_2$，策略网络参数$\boldsymbol{\theta}$，Φ函数网络参数$\boldsymbol{\varphi}$，目标Q函数网络参数$\bar{\boldsymbol{\omega}}_1$、$\bar{\boldsymbol{\omega}}_2$，目标策略网络参数$\bar{\boldsymbol{\theta}}$，衰减系数$\lambda\in(0,1)$，权重参数$\alpha_2$，目标网络平滑系数$w$，策略更新频率$d_1$，权重参数的衰减频率$d_2$，参数更新总次数$L$.

1. 循环参数更新步数$l=1,2,\cdots,L$.
2. 从数据集$\mathcal{B}$采样一个小批次$\mathcal{D}$;
3. 最小化损失函数(39)更新参数$\boldsymbol{\varphi}$;
4. 最小化损失函数(41)更新参数$\boldsymbol{\omega}_i$;
5. 判断$l \bmod d_1==0$
6. 最小化损失函数(40)更新参数$\boldsymbol{\theta}$;
7. 更新目标网络参数

$$\bar{\boldsymbol{\omega}}_1=(1-w)\bar{\boldsymbol{\omega}}_1+w\boldsymbol{\omega}_1$$
$$\bar{\boldsymbol{\omega}}_2=(1-w)\bar{\boldsymbol{\omega}}_2+w\boldsymbol{\omega}_2$$
$$\bar{\boldsymbol{\theta}}=(1-w)\bar{\boldsymbol{\theta}}+w\boldsymbol{\theta}$$

8. 结束判断
9. 判断$l \bmod d_2==0$

$$\beta=\beta\cdot\lambda$$

10. 结束判断
11. 结束循环

5 实验结果与分析

5.1 实验设置

在D4RL基准测试数据集中，使用不同性能的行为策略收集数据，以模拟不同质量的离线数据集.根据用于收集数据的行为策略的性能，将离线数据集的质量分为三个等级：随机、中等和专家.需要注意的是，行为策略的性能决定了数据集的质量.因此，它会影响所评估的离线强化学习算法的性能.

首先，为了说明在离线强化学习中，如果不能有效缓解分布偏移对策略学习过程的影响，直接使用在线强化学习算法训练策略将无法提升策略性能，测试了TD3[23]在D4RL数据集上的性能表现.需要注意的是，TD3通过降低值函数估计偏差来改善Actor-Critic算法的性能，其性能优于A3C[24]、PPO[25]、DDPG[26]等Actor-Critic算法.在实验中，使用了TD3原文的实验参数设置.然后，为了验证GOACBR的性能优越性，考虑了以下对比算法：BC[5]、BCQ[13]、BRAC[5]和BEAR[15]，其中，BRAC

和BEAR分别是基于KL散度正则化和MMD距离正则化的离线Actor-Critic算法. 为确保公平对比，使用作者提供的代码重新运行了所有对比算法. 实验过程中，所有算法均训练50万步，策略性能评估周期为5000步. 此外，在对比不同算法的性能时，所有算法的折扣因子都设置为0.99. 评估算法性能时，使用当前策略与环境交互5个回合，并将平均累积回报作为性能评估指标. 表1列出了GOACBR的参数设置. 注意到由式(8)可知：当$\beta=0.5$时，D_{JS}^{β}的上界最大. 由于β随策略训练过程逐渐降低，因此将β的初始值设置为0.5以保证D_{JS}^{β}在策略训练早期阶段对策略目标函数的贡献. 对于权重参数α_2，我们测试了以下几组值：0.2、0.5、1、2. 当$\alpha_2=1$时，GOACBR在测试任务上的表现最佳，其它对比算法的参数设置如下：

表1 GOACBR参数设置

设置	名称	值
GOACBR超参数	优化器	Adam
	Q函数网络学习率	0.0003
	策略网络学习率	0.0003
	小批次大小	256
	折扣因子	0.99
	权重参数α_2	1
	目标网络平滑系数w	0.005
	JS散度参数β	0.5
	策略更新频率d_1	2
	衰减频率d_2	10 000
	衰减系数λ	0.9
网络结构	Q函数网络隐藏层数量	2
	Q函数网络各个隐藏层神经元数量	256
	Q函数网络激活函数	ReLU
	策略网络隐藏层数量	2
	策略网络各个隐藏层神经元数量	256
	策略网络激活函数	ReLU
	Φ函数网络隐藏层数量	2
	Φ函数网络各个隐藏层神经元数量	256
	Φ函数网络激活函数	ReLU

(1)在BC中，策略网络是一个含有两个隐藏层的全连接神经网络，其中，每个隐藏层神经元的数量均为256，激活函数为ReLU. 在训练策略网络参数时，学习率设置为0.0003，优化器为Adam，小批次大小为256.

(2)BCQ的编码网络、解码网络、扰动网络和Q函数网络都是两层全连接神经网络. 在编码网络和解码网络中，各个隐藏层均有750个神经元，扰动网络和Q函数网络中两个隐藏层的神经元数量依次为400和300. 所有网络使用的激活函数均为ReLU. 在更新网络参数时，使用Adagrad优化器，学习率设置为0.001，小批次大小为100.

(3)在BRAC中，权重参数$\alpha_1=0.1$，Q函数网络的网络结构与BCQ的Q函数网络结构相同，策略网络和行为策略网络有三个隐藏层，隐藏层的神经元数量依次为400、300和300. 更新网络参数时，使用的优化器和学习率与BCQ相同，小批次大小为100.

(4)在BEAR中，所有网络的网络结构，更新网络参数时使用的优化器、学习率和小批次大小与BCQ相同.

5.2 实验结果

图1给出了在不同等级质量的三个基准数据集下TD3的性能表现，其中，实线表示5次运行下的平均累积回报，阴影表示回报的方差. 如图1所示，在walker2d和hopper任务上，由于在策略训练过程中智能体不能与环境交互，智能体无法充分探索状态动作空间，因此TD3在专家、中等和随机三种不同质量数据集下均无法改善策略性能. 此外，由于智能体不能从环境中接收到任何纠正性反馈，因此Q函数网络更新过程中的误差会不断累积并影响后续的参数更新，这使得Q函数估计偏差不能被有效纠正.

另一方面，策略训练过程中的分布偏移现象会使那些未包含在离线数据集中的状态动作对的Q值被错误估计，因此智能体无法学习到最优策略. 在halfcheetah任务上，TD3在随机数据集和中等数据集中均可以改善策略性能. 这是由于这两种数据集中数据更富多样性，因此TD3能在一定程度上改善策略性能.

为了探究β值对GOACBR性能的影响，我们在halfcheetah任务上测试了固定β值与周期性降低β值两种情形下GOACBR的性能. 根据图2可知，相较于固定β值，在周期性降低β值的情形下，GOACBR的策略学习过程更加稳定. 同时，在训练的后期，GOACBR获得的回报更高. 这是因为周期性降低β值使得正则化项对策略网络训练过程的影响被逐渐降低. 在固定β值的情形下，我们发现GOACBR在一些种子上获得的回报较低，这使得其在专家数据集和中等数据集上的回报方差较大. 这可能是因为在这些随机种子下，当前策略与行为策

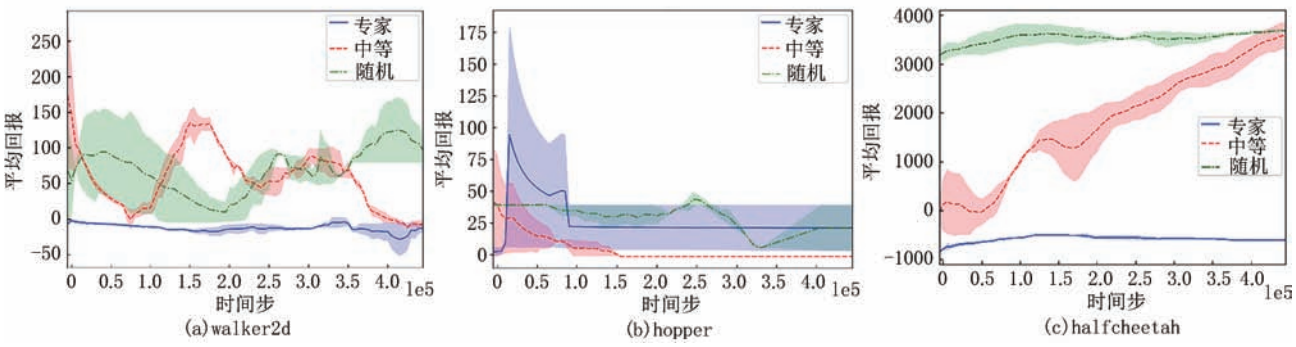

图1　TD3在不同质量数据集下的性能

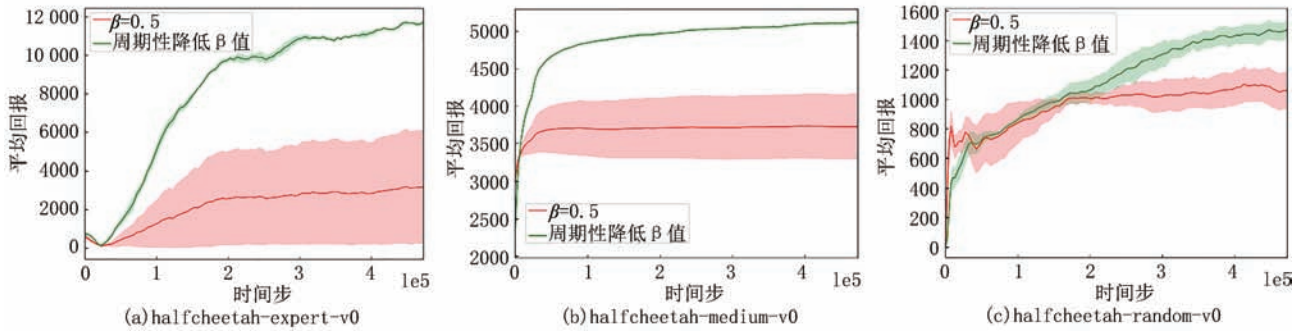

图2　固定β值与周期性降低β值两种情形下GOACBR的平均回报曲线

略之间的差异不能被很好地约束，使得分布偏移问题较为显著，从而无法改善策略性能. 图3给出了GOACBR中斜对称JS散度和BRAC中KL散度随训练步的变化曲线. 图3表明：相较于BRAC中KL散度的变化，GOACBR中的斜对称JS散度随训练步稳定变化，且其值保持在0附近的一个很小的区域内. 事实上，根据定理2可知，对于任意的$\beta\in(0,1)$，斜对称JS散度$D_{JS}^{\beta}(\pi_1(\cdot|s)\|\pi_2(\cdot|s))\leqslant\ln 2$. 另一方面，由于$\beta$随训练步逐渐减小，因此斜对称JS散度的上界$g(\beta)$逐渐减小，这使得累积期望回报能有效地对GOACBR的策略目标函数(7)产生贡献.

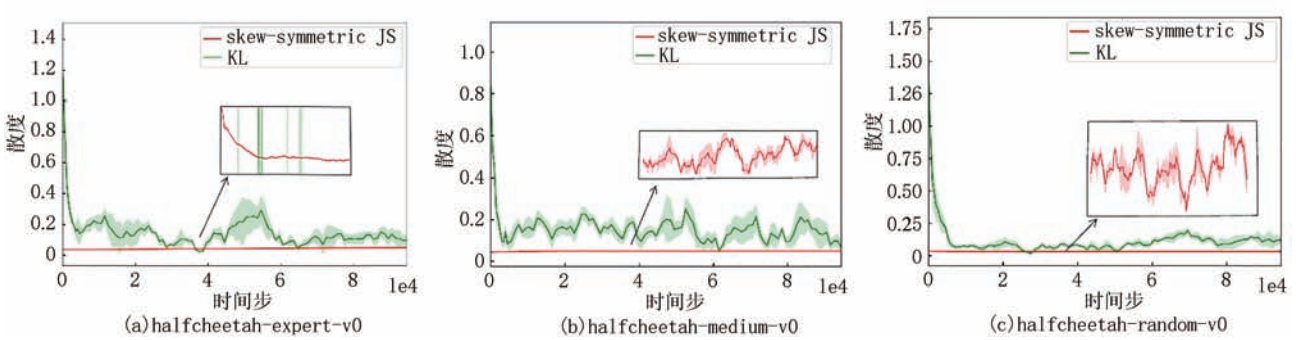

图3　GOACBR的斜对称JS散度和BRAC的KL散度随训练步的变化曲线

图4和表2给出了GOACBR和对比算法的性能比较. 在图4中，实线表示5次运行下的平均累积回报，阴影表示回报的方差. 在表2中，加粗的数据表示在相应的任务上，5个离线强化学习算法在最后10%时间步上获得的最大累积平均回报，加下划线的数据则表示仅次于最大累积平均回报的累积回报值. 平均累积回报总和提升比例根据以下式子计算：

$$\text{回报总和比例提升}=\frac{\text{GOACBR回报总和}-\text{对比算法回报总和}}{\text{对比算法回报总和}}\times 100\% \tag{42}$$

由图4和表2可以看出：

(1)当使用专家数据训练策略时，GOACBR在

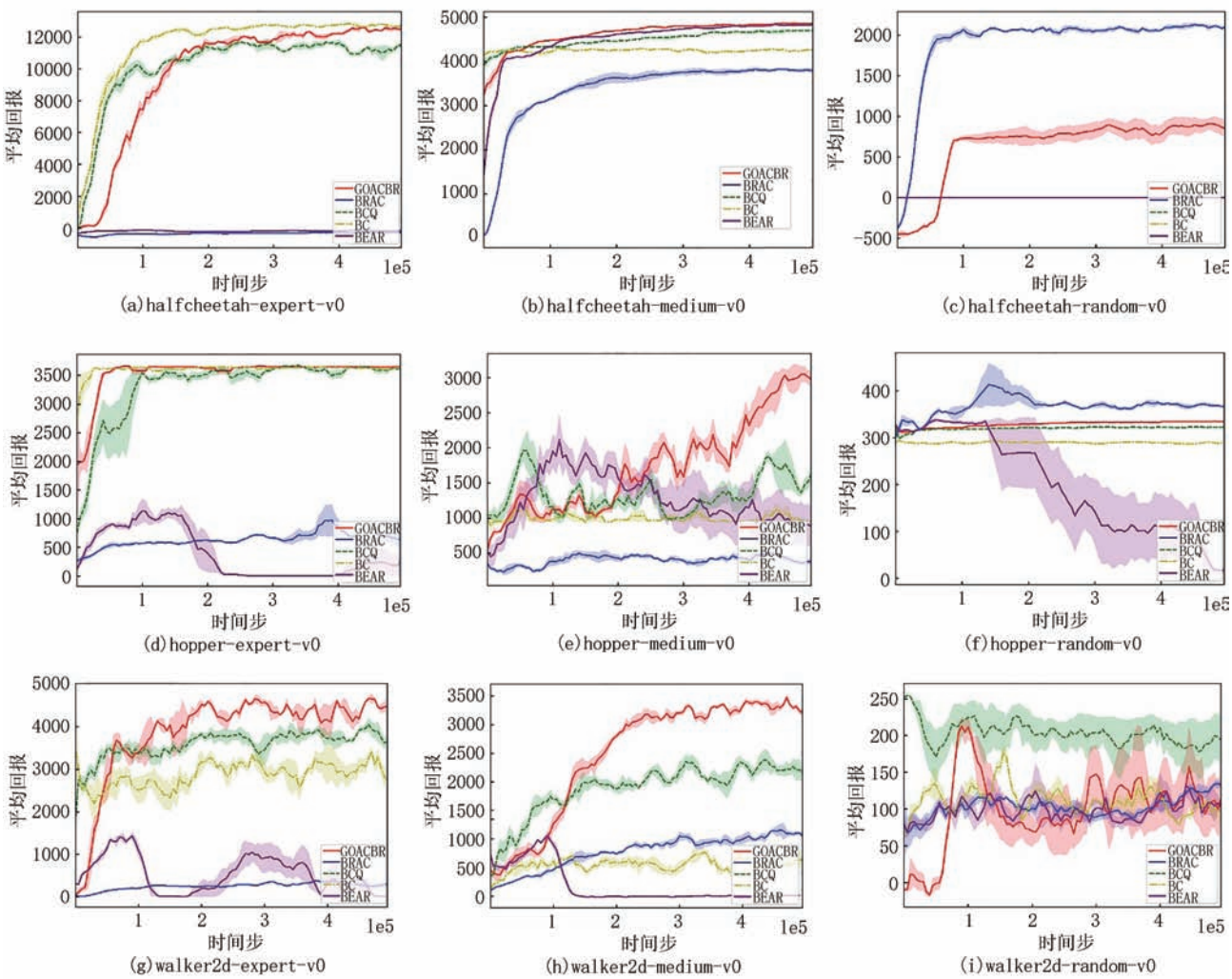

图4 离线强化学习算法在不同质量数据集下的性能

表2 离线强化学习算法在最后10%训练步上的平均回报

测试任务	BCQ[13]	BC[5]	BRAC[5]	BEAR[15]	GOACBR
halfcheetah-expert-v0	11604.70	**12747.04**	-275.95	-109.11	12549.20
hopper-expert-v0	3377.36	3600.61	663.01	122.32	**3626.22**
walker2d-expert-v0	3687.11	3080.48	337.43	456.21	**4574.61**
halfcheetah-medium-v0	4787.89	4252.63	3797.66	4932.11	**4990.98**
hopper-medium-v0	1776.45	1074.09	383.67	983.60	**2980.68**
walker2d-medium-v0	2278.97	593.77	1075.50	99.45	**3348.65**
halfcheetah-random-v0	-0.77	-0.942	**2097.44**	-0.46	956.66
hopper-random-v0	323.74	289.60	**368.69**	41.45	329.51
walker2d-random-v0	**189.44**	97.34	144.54	100.50	136.34
回报总和	28 024.89	25 734.62	8591.99	20 234.44	**33 492.85**
回报总和提升比例	19.5%	30.1%	289.8%	404.71%	

hopper和walker2d任务上均能获得最大平均累积回报，在halfcheetah任务上获得的回报与BC相当. 这是因为离线专家数据是由专家策略生成的，而GOACBR在优化策略时约束了当前策略和行为策略之间的差异，因此在策略学习过程中学到的策略在一定程度上接近专家策略. 需要注意的是，BC通过最小化当前策略与专家策略的差异来更新策略，因此它在专家数据集中也能学习到好的策略. 但是，由于BC不能有效利用Q函数信息，故其在hopper和walker任务上的表现不如GOACBR. 虽然

BRAC也在策略优化过程中约束当前策略与行为策略的差异，但是BRAC在整个策略优化过程中未有效调节策略损失函数中KL散度正则化项与累积期望回报项所占的权重，故其在专家数据中表现较差. 由于无法保证策略改进过程的收敛性，BCQ和BEAR在专家数据集上的性能表现均不如GOACBR. 以上结果充分表明了GOACBR在专家数据集上的性能优异性.

(2)当使用中等数据集训练策略时，GOACBR在三个任务上的平均累积回报均高于对比算法，这显示了GOACBR有效利用中等离线数据学习策略的能力. 事实上，GOACBR在策略训练过程中周期性降低行为正则化项在策略损失函数中所占的权重，这使得其在初始策略训练阶段可以有效地从离线数据集中存放的质量较高的轨迹中学习. 然后，在策略训练过程中，随着行为正则化项对策略损失函数的影响逐步减小，GOACBR允许习得策略偏离行为策略，这在一定程度上增大了智能体学习到更好策略的可能性.

(3)当使用随机数据集训练策略时，GOACBR在三个任务上的表现均不如BRAC. 这是因为GOACBR在策略训练的初始阶段过度约束当前策略与行为策略之间的差异使得其学习到的策略更加接近随机策略. 此外，GOACBR在所有任务上的表现均优于BC和BEAR，原因是：BC直接通过最小化当前策略与行为策略之间的差异来更新策略，因此其学习到的策略接近随机策略；BEAR不能保证策略改进过程的收敛性. 此外，注意到GOACBR同时考虑克服分布偏移对策略学习过程的影响和降低当前策略与行为策略之间的性能差异，因此尽管BCQ在walker2d-random-v0上的平均累积回报大于GOACBR，但是在其余任务上的表现均不如GOACBR.

(4)尽管GOACBR在部分任务上的表现不是最优，然而，其在所有测试任务上的回报总和远高于所有对比算法. 值得一提的是，相较于BRAC，GOACBR在所有测试任务上获得的回报总和提升了289.8%. 为展示计算代价，表3给出了算法的平均运行时间. 实验过程中，所有算法的总策略训练步数均设为500 000，硬件配置为GeForce GTX 2080 GPU和4.00 GHz的Intel Core i7-6700K CPU. 根据表3可知，BC的计算代价最小. 这是因为BC仅需训练策略网络. 此外，GOACBR的计算代价与其余三个对比算法相当.

表3 离线强化学习算法的平均运行时间

性能指标	BCQ[13]	BC[5]	BRAC[5]	BEAR[15]	GOACBR
平均运行时间/h	1.4	0.7	1.6	1.5	1.5

综上所述，由于有效地缓解了分布偏移和降低当前策略与行为策略之间的策略性能差异，GOACBR在随机、中等和专家三类不同等级质量的离线数据集中拥有最佳的综合性能表现.

6 结 论

离线强化学习通常面临所学习的策略与从离线数据集中观测到的行为策略分布不同而造成的分布偏移问题. 为此，本文将当前策略与行为策略之间的斜对称JS散度作为代理策略目标函数的正则化项，同时在策略训练过程中逐渐降低正则化项对策略损失函数的影响. 在GOACBR的具体实现过程中，考虑到由于行为策略未知导致难以直接计算JS散度的问题，本文设计了一个辅助网络来对其进行间接估计. 在D4RL基准测试数据集上的实验结果表明：与最先进的离线强化学习算法相比，GOACBR更具有性能竞争力.

未来的研究工作包括探究斜对称JS散度中参数β的自适应律以适应不同任务场景，设计能保证策略改进过程收敛性的隐式策略约束和隐式值函数约束情形下的离线Actor-Critic.

参 考 文 献

[1] Barto A G, Sutton R S. Reinforcement learning. Journal of Cognitive Neuroscience, 1999, 11(1): 126-134

[2] Chai Lai, Zhang Ting-Ting, Dong Hui, Wang Nan. Multi-agent deep reinforcement learning algorithm based on partitioned buffer replay and multiple process interaction. Chinese Journal of Computers, 2021, 44(6): 1140-1152.(in Chinese)
(柴来，张婷婷，董会，王楠. 基于分区缓存区重放与多线程交互的多智能体深度强化学习算法. 计算机学报，2021，44(6)：1140-1152)

[3] Liu Quan, Zhai Jian-Wei, Zhang Zong-Zhang, Zhong Shan, Zhou Qian, Zhang Peng, Xu Jin. A survey on deep reinforcement learning. Chinese Journal of Computers, 2018, 41(1): 1-27 (in Chinese)
(刘全，翟建伟，章宗长，钟珊，周倩，章鹏，徐进. 深度强化学习综述. 计算机学报，2018，41(1)：1-27)

[4] Cheng Y, Huang L, Chen C L P, Wang X. Robust actor-critic with relative entropy regulating actor. IEEE Transactions on Neural Networks and Learning Systems, 2022, doi: 10.1109/

TNNLS.2022.3155483

[5] Wu Y, Tucker G, Nachum O. Behavior regularized offline reinforcement learning. ArXiv preprint arXiv:1911.11361, 2019.

[6] 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 (in Chinese)
(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)

[7] Dulac-Arnold G, Mankowitz D, Hester T. Challenges of real-world reinforcement learning. ArXiv preprint arXiv: 1904.12901, 2019

[8] Zhu Fei, Wu Wen, Fu Yu-Chen, Liu Quan. A dual deep network based secure deep reinforcement learning method. Chinese Journal of Computers, 2019, 42(08): 1812-1826 (in Chinese)
(朱斐, 吴文, 伏玉琛, 刘全. 基于双深度网络的安全深度强化学习方法. 计算机学报, 2019, 42(8): 1812-1826)

[9] Dadashi R, Rezaeifar S, Vieillard N, Hussenot L, Pietquin O, Geist M. Offline reinforcement learning with pseudometric learning//Proceedings of the 38th International Conference on Machine Learning. Virtual, Online, 2021: 2307-2318

[10] Xu H, Dominguez A D, Sauer P W. Optimal tap setting of voltage regulation transformers using batch reinforcement learning. IEEE Transactions on Power Systems, 2020, 35(3): 1990-2001

[11] Levine S, Kumar A, Tucker G, Fu J. Offline reinforcement learning: tutorial, review, and perspectives on open problems. ArXiv preprint arXiv:2005.01643, 2020

[12] Agarwal R, Schuurmans D, Norouzi M. An optimistic perspective on offline reinforcement learning//Proceedings of the 37th International Conference on Machine Learning. Virtual, Online, 2020: 92-102

[13] Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration//Proceedings of the 36th International Conference on Machine Learning. California, USA, 2019: 3599-3609

[14] Yang Z L, Zhang S Y, Hu Y T, Hu Z W, Huang Y F. VAE-stega: linguistic steganography based on variational auto-encoder. IEEE Transactions on Information Forensics and Security, 2021, 16: 880-895

[15] Kumar A, Fu J, Tucker G, Levine S. Stabilizing off-policy q-learning via bootstrapping error reduction//Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada, 2019

[16] Wang W, Li H J, Ding Z M, Nie F P, Chen J Y, Dong X, Wang Z H. Rethinking maximum mean discrepancy for visual domain adaptation. IEEE Transactions on Neural Networks and Learning Systems, 2021, doi: 10.1109/TNNLS.2021.3093468.

[17] Joe F. Relative entropy measures of multivariate dependence. Journal of the American Statistical Association, 1989, 84(405): 157-164

[18] Xu H, Zhan X, Li J, Yin H. Offline reinforcement learning with soft behavior regularization. ArXiv preprint arXiv: 2110.07395, 2021

[19] Kumar A, Zhou A, Tucker G, Levine S. Conservative q-learning for offline reinforcement learning//Proceedings of the 34th Conference on Neural Information Processing Systems. Virtual, Online, 2020: 1179-1191

[20] Kostrikov I, Nair A, Levine S. Offline reinforcement learning with implicit q-learning. ArXiv preprint arXiv: 2110.06169, 2021

[21] Wu Y, Zhai S, Srivastava N, Susskind J M, Zhang J, Salakhutdinov R, Goh H. Uncertainty weighted actor-critic for offline reinforcement learning//Proceedings of the 38th International Conference on Machine Learning. Virtual, Online, 2021: 11319-11328

[22] Nielsen F. On the jensen-shannon symmetrization of distances relying on abstract means. Entropy, 2019, 21(5): 485

[23] Fujimoto S, Van H, Meger D. Addressing function approximation error in actor-critic methods//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 2587-2601

[24] Babaeizadeh M, Frosio I, Tyree S, Clemons J, Kautz J. Reinforcement learning through asynchronous advantage actor-critic on a gpu//Proceedings of the 5th International Conference on Learning Representations. Toulon, France, 2017

[25] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. ArXiv preprint arXiv: 1707.06347, 2017

[26] Lillicrap T P, Hunt J J, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D. Continuous control with deep reinforcement learning. //Proceedings of the 4th International Conference on Learning Representations, San Juan, Puerto Rico, 2016

CHENG Yu-Hu, Ph.D., professor. His main research interests include machine learning and intelligent systems.

HUANG Long-Yang, M. S. candidate. His research interest is reinforcement learning.

HOU Di-Yuan, M. S. candidate. His research interest is reinforcement learning.

ZHANG Jia-Zhi, M.S. candidate. His research interest is reinforcement learning.

CHEN Jun-Long, Ph. D., professor. His main research interests include data science and computational intelligence.

WANG Xue-Song, Ph. D., professor. Her main research interests include machine learning and pattern recognition.

Background

Reinforcement learning (RL) is an important branch of machine learning, which can be divided into online reinforcement learning (online RL) and offline reinforcement learning (offline RL). In online RL, the interaction cost between agent and environment is high and the security of agent is difficult to be guaranteed, which limits its application to the real world. Offline RL is concerned with how an agent can learn the optimal policy by using a fixed offline dataset without interacting with the environment. Therefore, the offline RL has wide potential applications in the real world. However, there is a major challenge in offline RL: distribution shift. Due to the distribution shift, Q values of the state-action pairs that are not included in the offline dataset will be incorrectly estimated. In order to alleviate the distribution shift problem, this paper proposes a novel generalized offline actor-critic with behavior regularization (GOACBR). To alleviate the effect of distribution shift on policy improvement process, the skew-symmetric JS divergence between the current and behavior policies is taken as the regularization term in the policy objective function. The theoretical analysis shows that the convergence of GOACBR can be guaranteed. An auxiliary neural network is designed to indirectly estimate the skew-symmetric Jensen-Shannon (JS) divergence between the current and behavior policies. Experimental results on the D4RL benchmark dataset show that GOACBR outperforms various SOTA offline RL algorithms.

This work was supported by the National Natural Science Foundation of China under Grant 62176259 and Grant 61976215, and the Key Research and Development Program of Jiangsu Province under Grant BE2022095.