智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

动态距离约束的离线到在线强化学习

赵若涵, 魏巍, 王达, 黄利

引用本文:
赵若涵, 魏巍, 王达, 等. 动态距离约束的离线到在线强化学习[J]. 智能系统学报, 2026, 21(4): 1055-1065.
ZHAO Ruohan, WEI Wei, WANG Da, et al. Dynamic distance-constrained offline-to-online reinforcement learning[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(4): 1055-1065.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202510017

DOI: 10.11992/tis.202510017
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20260403.1329.013

动态距离约束的离线到在线强化学习

赵若涵, 魏巍, 王达, 黄利
(山西大学 计算机与信息技术学院, 山西 太原 030006)

摘　要: 离线到在线强化学习能够在有限的在线交互下快速提升策略性能, 但在微调初期, 策略分布往往偏离离线数据支持区域, 导致估值不稳并限制性能提升。现有方法多依赖统一的保守约束, 将离线数据视作边界条件, 却忽视了样本质量与分布差异, 从而在更新时既过度约束, 又难以避免分布外的不稳定估值。为此, 本文提出一种动态距离约束的离线到在线强化学习算法。该方法基于状态条件距离函数, 将策略约束在离线数据支持区域内, 并随着在线交互的进行, 动态地将约束放宽至高质量数据区域, 从而实现离线到在线的平滑迁移。实验结果表明, 与现有主流方法相比, 所提出的方法可以在多个模拟任务上实现稳定有效的性能改进。

关键词: 机器学习; 强化学习; 离线到在线强化学习; 马尔可夫过程; 策略优化; 距离约束; 策略迁移; 约束优化
中图分类号: TP181　　**文献标志码**: A　　**文章编号**: 1673−4785(2026)04−1055−11

中文引用格式: 赵若涵, 魏巍, 王达, 等. 动态距离约束的离线到在线强化学习 [J]. 智能系统学报, 2026, 21(4): 1055–1065.
英文引用格式: ZHAO Ruohan, WEI Wei, WANG Da, et al. Dynamic distance-constrained offline-to-online reinforcement learning[J]. CAAI transactions on intelligent systems, 2026, 21(4): 1055–1065.

Dynamic distance-constrained offline-to-online reinforcement learning

ZHAO Ruohan, WEI Wei, WANG Da, HUANG Li
(School of Computer Science and Information Technology, Shanxi University, Taiyuan 030006, China)

Abstract: Offline-to-online reinforcement learning enables rapid policy improvement with limited online interactions. However, in the early stage of fine-tuning, the policy distribution often deviates from the support region of offline data, leading to unstable value estimation and restricting performance gains. Existing methods typically rely on uniform conservative constraints, treating offline data as boundary conditions, but they overlook sample quality and distributional differences. As a result, policy updates are simultaneously over-constrained and prone to unstable estimation outside the offline distribution. To address this issue, we propose a dynamic distance-constrained algorithm for offline-to-online reinforcement learning. Our method employs a state-conditioned distance function to constrain the policy within the support region of offline data and gradually relaxes the constraint toward high-quality data regions as online interactions progress, thereby achieving a smooth transition from offline to online learning. Experimental results demonstrate that, compared with mainstream baselines, the proposed method achieves stable and effective performance improvements across multiple simulated tasks.

Keywords: machine learning; reinforcement learning; offline-to-online reinforcement learning; Markov processes; policy optimization; distance constraint; policy transfer; constrained optimization

离线强化学习[1-2]允许智能体仅通过预收集的静态数据集进行策略学习, 而无需与环境进行实时交互。这一特性使其在交互成本高昂或环境不可复现的场景中(如医疗决策[3-4]、机器操作[5]、推荐系统[6-7]等)展现出巨大潜力。然而, 由于离线数据集通常由特定行为策略生成, 存在数据分布受限、状态−动作空间覆盖不全等问题, 导致离线学习所得策略往往表现出明显的次优性。为克服这一局限性, 近年来研究者开始关注将离线策略作为初始化, 在后续通过与环境的在线交互进行微调, 以进一步提升策略性能[8]。这种结合离线预训练与在线微调的范式, 即离线到在线强化学习。

收稿日期: 2025−10−16.　**网络出版日期**: 2026−04−03.
基金项目: 国家自然科学基金项目(62276160); 山西省基础研究计划(202203021211294).
通信作者: 魏巍. E-mail: weiwei@sxu.edu.cn.

在线微调过程中面临的核心挑战之一是分布偏移问题，即离线阶段和在线阶段的状态分布与动作分布存在显著差异。这种分布偏移导致在线策略在遇到离线数据未覆盖或罕见的状态−动作对时，容易产生估计误差和策略崩溃[9]，影响在线微调的稳定性和效率。在实际应用场景中（如医疗决策[3-4]、机器操作[5]、推荐系统[6-7]等），在线交互代价高昂甚至存在安全风险，因此，如何在保障安全性的同时充分利用在线数据，是保证离线到在线强化学习落地的关键瓶颈。

尽管已有研究通过保留部分[10]或全部离线策略约束[11]，或针对不同任务精细调节约束强度[12]，以及引入替代性正则化机制来缓解策略崩溃[13-14]，近期也有工作尝试从贝叶斯建模视角出发联合估计策略与数据的不确定性[15]，或引入周期性的策略“复兴”机制以保持行为稳定性[16]。然而，这些方法存在两个显著局限：1）约束强度与真实分布偏移程度之间缺乏直接联系，难以反映当前策略行为是否真正处于“危险”的分布外区域；2）约束调节通常依赖训练步数或人工经验，无法根据策略与数据分布的关系进行自适应调整，从而在微调过程中容易出现“过度保守”或“过早放松”的问题。为此，本文提出动态距离约束的离线到在线强化学习算法 (dynamic distance-constrained offline-to-online reinforcement learning, D2O2O)，该方法通过状态条件距离函数，显式量化样本与离线数据的几何关系，使其不偏离离线数据的支持区域，并随着在线微调的进行，动态调整约束强度，使其逐步扩展到在线数据中具有较高质量的区域，平滑地完成离线到在线的迁移。

在此基础上，本文进一步设计了一种对偶形式的动态约束调节机制，通过对拉格朗日乘子进行自适应更新，使约束强度能够根据当前策略的偏离程度动态变化。当策略倾向于访问高风险的分布外区域时，该机制会自动增强约束强度，以抑制潜在的价值高估；而随着策略逐步积累高质量的在线经验、其偏离风险相应降低，约束强度则被逐步放松，从而提升策略的探索能力。与现有采用静态约束或手动调节的方法不同，该机制不依赖预设的时间表，而是直接由策略行为与数据分布之间的关系驱动。因此，该机制能够在在线微调早期提供必要的稳定性约束，并在后期实现对探索能力的提升,从而在稳定性与性能提升之间取得更合理的平衡。

本文的主要贡献如下：

1) 引入数据距离感知的策略约束机制，防止策略在微调初期进入高风险区域，实现了离线到在线阶段的平稳过渡。

2) 提出适应性惩罚机制调节策略，采用对偶优化方法动态调节策略约束强度，在保障稳定性的同时逐步提升策略优化能力。

3) 在标准 D4RL 基准环境上的实验结果表明，所提方法可轻易与现有算法结合，并在多个任务和数据集设置中均优于现有基线算法。

1　预备知识

1.1　马尔可夫决策过程

强化学习问题通常建模为一个马尔可夫决策过程（Markov decision process, MDP）[17]，记为一个五元组(S,A,P,r,γ)。其中S表示状态空间；A表示动作空间；$P(s'|s,a)$是状态转移概率，表示在状态s下采取动作a后转移至下一状态s'的概率；$r(s,a)$是即时奖励函数，表示在状态s执行动作所获得的即时奖励；$\gamma[0,1)$为折扣因子，用于对未来奖励进行衰减。策略$\pi(a|s)$是在给定状态s下采取动作a的条件分布。强化学习的目标是学习一个最优策略$\pi*$，使得期望累积折扣奖励最大化。相应地，可以定义状态−动作值函数（Q 函数）为

$$Q^{\pi}(s,a)=\mathrm{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t}r(s_t,a_t)|s_0=s,a_0=a\right]$$

式中：$s_{t+1}\sim P(\cdot|s_t,a_t)$，$a_t\sim\pi(\cdot|s_t)$，E 表示求期望。

1.2　离线强化学习

离线强化学习是指智能体在不可与环境交互的前提下，利用固定的数据集$D=\left\{\left(s_i,a_i,r_i,s_i'\right)\right\}_{i-1}^{N}$来进行策略学习[18]。该数据集通常由某个行为策略在与环境交互过程中收集得到。

在离线强化学习中，Q 函数通常通过最小化时序差分（temporal difference, TD）误差进行训练：

$$\min_{\theta}\mathrm{E}_{(s,a,r,s')}\left[\left(r+\gamma\mathrm{E}_{a'\sim\pi(\cdot|s')}\left[Q_{\theta'}(s',a')\right]-Q_{\theta}(s,a)\right)^{2}\right]$$

式中：θ为当前 Q 网络的参数，θ'是目标网络的延迟参数副本。

离线强化学习面临两个核心挑战：

1) 分布偏移问题：训练策略π所生成的状态−动作分布可能远离数据集D的分布，从而使函数在未见区域的估计产生严重偏差。

2) 外推误差：当策略π查询Q值时，其动作可能落在数据支持之外（out-of-distribution, OOD），这会导致 Q 网络进行不可信的外推，严重误导策略更新。

为缓解这些问题，许多离线强化学习方法会引入保守 Q 估计、行为克隆或策略约束，以限制

策略偏离已知数据支持区域。

1.3　离线到在线强化学习

离线到在线强化学习旨在结合离线强化学习的高样本效率与在线强化学习的探索能力, 构建更加实用的强化学习算法。在该框架中, 训练过程分为两个阶段:

1) 离线预训练阶段: 使用离线数据集 D_{offline} 对策略和 Q 函数进行初始化;

2) 在线微调阶段: 基于当前策略与环境交互, 采集新的在线数据 D_{online}, 用于进一步更新策略和 Q 函数。

此框架的核心目标是: 在不需要大量环境交互的前提下, 学习得到具有泛化能力的高性能策略。

2　相关工作

离线到在线强化学习将学习过程划分为两个阶段: 首先在离线数据集上进行预训练以获得初始策略或价值函数, 然后通过有限的在线交互进一步微调, 以提升策略性能。

在离线阶段, 为保证策略安全性, 常采用基于行为约束 (如 Kullback-Leibler 散度限制或行为克隆[19]) 或悲观估计 (如保守 Q 学习 (conservative Q-learning, CQL)[20])的方法。然而, 这类方法强调对策略的保守性, 在实际在线微调时常常会限制策略的探索能力, 导致性能提升缓慢甚至出现性能退化的问题。为了应对上述问题, 已提出多种离线到在线强化学习方法。

为缓解离线到在线阶段由分布偏移引发的价值高估与策略崩溃问题, 一类代表性方法通过在策略优化或价值估计过程中引入显式或隐式约束, 限制策略对离线数据分布的偏离程度。Nair 等[8] 通过在策略更新中引入优势加权的行为克隆目标, 使策略在利用在线样本进行改进的同时, 仍然对离线数据保持隐式对齐, 从而在一定程度上缓解了分布外动作带来的价值误导问题。隐式 Q 学习 (implicit Q-learning, IQL)[21] 从另一角度出发, 避免显式评估分布外动作。Lee等[22] 提出的平衡回放与悲观 Q 集成方法, 在离线阶段引入悲观的 Q 函数初始化, 有效抑制了微调初期由分布偏移导致的自举误差。类似地, 校准 Q 学习 (Cal-QL)[23] 在 CQL 的基础上进一步强调价值函数的“校准性”, 通过构造保守但尺度合理的 Q 函数初始化, 为后续在线微调提供更安全的起点。总体而言, 约束类方法的核心思想是在离线到在线过渡阶段维持策略的保守性, 以防止分布外动作引发的价值高估和性能骤降。然而, 这类方法通常采用固定或弱自适应的约束形式, 约束强度与策略当前的实际偏离风险之间缺乏直接、连续的关联。当约束过强时, 在线探索受限, 难以充分利用环境交互; 而约束不足时, 又可能重新暴露于价值误导与策略崩溃风险之中。值得注意的是, 近期的研究开始探索通过增加在线预训练阶段[24]、将在线微调划分为 3 种不同阶段[25] 或悲观 Q 函数重构[26] 等更精细的方式来稳定在线微调过程, 但如何在不同质量的数据混合场景下自适应地调节保守程度, 仍是该方向面临的核心难题。

为克服静态约束在探索–利用权衡上的局限, 近年来的研究开始关注如何在在线微调过程中对策略改进幅度或约束强度进行更精细、自适应的调节。Chen 等[27] 提出了一种状态自适应的策略改进框架, 实现了对策略更新幅度的细粒度控制。策略扩展 (policy expansion, PEX)[28] 则通过在在线阶段引入可扩展的策略集合, 使离线策略与新策略以组合形式参与决策, 在保留离线行为的同时逐步引入新的动作模式。Guo 等[29] 从探索角度出发, 引入乐观探索与元学习机制, 以加速在线阶段的适应过程。Liu 等[30] 进一步利用扩散模型与能量函数, 从离线数据中提取结构化先验, 用于生成更符合在线需求的高质量样本, 从而缓解直接复用离线数据带来的分布不匹配问题。这种生成式方法已成为当前的研究热点, 例如, 最新的研究通过无分类器引导的扩散生成[31] 来弥合离线与在线数据之间的分布差距, 或将离线到在线微调范式引入自动谈判领域[32]。此外, 一些工作从贝叶斯原理[15] 或元学习[29] 视角出发, 设计更高效的离线到在线适应机制, 而 Chaudhary 等[33] 则通过元策略统一离线与在线轨迹, 在降低计算复杂度的同时实现了高效探索。这类方法已经明确认识到“单一、固定约束不足以支撑高效的在线微调”, 并通过策略组合、结构扩展或样本生成等方式增强调节能力。然而, 多数方法的调节依据仍然是策略结构、训练阶段或经验性规则, 并未显式刻画当前策略行为与离线数据支持区域之间的关系, 因此难以直接反映策略是否正在进入高风险的分布外区域。

综上所述, 现有离线到在线强化学习方法大致可分为基于约束的安全型方法与基于调节的探索增强型方法。前者在稳定性方面表现可靠, 但往往缺乏对约束强度的精细控制; 后者提升了探索灵活性, 但对策略偏离风险的刻画仍然较为间接。本文提出的动态距离约束方法在两类思路之间建立了统一联系: 通过状态条件距离函数显式

量化策略行为与离线数据分布之间的偏离程度，并借助对偶优化机制实现约束强度的连续、自适应调节，从而在保证稳定性的同时逐步释放探索能力，实现更加平滑和可控的离线到在线迁移。

3 本文方法

在离线到在线强化学习框架中，策略首先在离线数据上进行预训练，再通过与环境交互进行在线微调[8]。尽管离线阶段能够提供较高质量的初始策略，但在线微调初期由于分布偏移和外推误差，策略很容易发生不稳定更新，甚至导致性能退化。这种外推误差的根本原因是：策略与环境交互过程中不断引入新的在线数据，使得状态−动作分布快速偏离离线数据支持区域，Q 函数在这些未见区域难以准确估值，从而误导策略学习方向。因此，如何在微调初期控制策略偏离分布、实现稳定过渡，同时又不过分抑制策略提升，是设计离线到在线强化学习方法的关键。

为此，本文提出一种基于状态条件距离函数的策略正则化方法 (D2O2O)，并设计了配套的动态惩罚机制，以实现对在线微调早期策略偏移的自适应控制。

3.1 状态条件距离函数

在离线到在线强化学习中，预训练策略生成的状态−动作分布最初与离线数据一致，但随着策略在环境中开始交互，状态−动作分布会发生显著变化。这种分布偏移使得策略探索可能触及离线数据未覆盖的区域，导致 Q 函数估计处于外推状态，产生严重误差。

为应对这一挑战，有效的离线到在线强化学习方法应当具备两方面能力：一是能够在微调早期阶段抑制策略快速偏离训练数据；二是在策略积累更多经验之后，不应强行将其限制在离线数据分布中，避免策略过度保守。因此，需要一种既能约束早期偏离，又不限制策略后期提升的机制。

已有研究和理论分析表明[34]，深度 Q 网络在连续状态−动作空间中，更擅长在插值区域学习，即那些可由离线数据点构成的凸包内部或边界附近的状态−动作对上，Q 函数通常具有更低的近似误差；而对于落在凸包之外的外推点，误差显著增加。因此，若能根据数据分布评估当前策略生成状态−动作对的几何偏离程度，将有助于为策略优化过程提供明确的稳定性评估依据。

受此启发，本文引入一个状态条件距离函数 $g(s,a)$，用于刻画在给定状态 s 下，动作 a 与训练数据几何中心之间的相对偏离程度。该函数能够估计当前动作是否处于数据支持区域，进而在策略优化时作为行为约束。具体地，通过监督方式训练 $g(s,a)$：对于离线数据集中采样的状态−动作对，从动作空间中均匀采样多个噪声动作 $\hat{a}\sim \mathrm{Unif}(A)$，并以其与参考动作之间的距离作为监督信号，优化目标函数：

$$\min_{g} \mathrm{E}_{(s,a)\sim D_{\mathrm{offline}}}\left[\mathrm{E}_{\hat{a}\sim \mathrm{Unif}(A)}(\|a-\hat{a}\|-g(s,\hat{a}))^2\right] \tag{1}$$

对于在线阶段，策略不断采集新的状态−动作对，这些数据被存入在线重放缓冲区 D_{online}。每次训练时，从 D_{offline} 与 D_{online} 中按照一定比例混合采样构建训练批次，并持续利用 $g(s,a)$ 评估策略当前行为的几何偏离程度，从而有效约束策略更新方向。

该距离函数具备两方面的优势：一方面，能够逼近训练集凸包的边界结构，形成对数据支持区域的近似建模；另一方面，其输出在面对分布外状态时依然具有合理的解释性，可用于策略优化中的正则化控制，从而在微调初期有效限制策略的不合理偏移，维持策略行为的可控性与稳定性。相比传统行为克隆距离或均方误差，状态条件距离函数提供了更强的数据几何感知能力，更适合离线到在线强化学习的结构性约束需求。

3.2 动态策略约束

尽管距离约束提供了评估策略行为是否偏离数据支持区域的能力，但在策略逐渐积累更多高质量在线经验之后，过强的惩罚会阻碍策略的进一步优化。为此，本文引入了一个动态调节的拉格朗日乘子 λ，其作用是控制几何约束项在策略优化中的权重，从而在策略更新过程中实现动态约束。

具体而言，本文使用对偶梯度下降方法来优化 λ。该方法将原始的带约束优化问题转化为无约束的鞍点问题，在每轮策略更新时，固定策略参数，使用对偶更新公式对 λ 进行梯度上升：

$$\lambda \leftarrow \underset{\lambda_{\min},\lambda_{\max}}{\mathrm{clip}}\left(\lambda+\eta\cdot\left(g(s,a)-g\left(s,a_{\mathrm{expert}}\right)-\delta\right),\right. \tag{2}$$

式中：η 是学习率，δ 为允许的距离偏差范围，$\lambda_{\min}$ 与 $\lambda_{\max}$ 为约束强度上下界，$g\left(s,a_{\mathrm{expert}}\right)$ 是以离线数据中动作为参考的基准距离。在对偶优化过程中，拉格朗日乘子初始化为 $\lambda_0=5.0$，以在训练初期对策略施加相对较强的几何约束，防止策略快速偏离离线数据支持区域，从而缓解由分布偏移引起的估值不稳定问题。对偶变量的学习率设置为 $\eta=3\times10^{-4}$，该取值与策略网络和价值网络的优化步长保持一致，从而确保约束权重的更新过程平稳，不会引入额外的训练不稳定性。距离偏差阈值设置为 $\delta=0$，即仅当策略生成的状态−动

作对在状态条件距离意义下相对于离线参考动作产生实际偏离时, 才增强约束强度。该设定避免了引入额外的人为容忍区间, 使约束调节完全由策略与离线数据之间的几何偏移所驱动。更新目标是将策略生成动作a与参考动作a_{expert}的距离差控制在阈值δ附近, λ根据策略生成状态–动作对的几何偏离程度来自适应增强或减弱正则约束。

对于在线微调初期, 当策略容易偏离数据支持区域时, λ会增强约束, 从而有效稳定策略更新; 在策略逐步积累更多在线经验后, 来自D_{online}的样本比例逐渐上升, 此时λ自适应下降, 逐步减弱对策略行为的几何限制。这种逐步放松的过程使策略能够跳出原有离线分布, 探索新的高价值区域, 从而在稳定性与优化能力之间实现动态平衡。

对于离线到在线强化学习而言, 核心挑战在于如何在缓解分布偏移风险的同时, 避免对策略探索能力的过度抑制。离线阶段训练得到的 Q 函数通常仅在离线数据分布所覆盖的状态–动作区域内具有较高估计精度, 一旦在线微调过程中策略生成分布外动作, Q 函数可能产生高估, 从而误导策略更新并导致性能骤降。另一方面, 若始终强制策略严格约束在离线数据分布内, 虽然能够保证安全性, 但在线交互难以带来新的有效信息, 在线微调阶段的优化潜力受到显著限制。

本文方法通过状态条件距离函数与动态约束调节机制的设计, 在机制层面缓解了上述矛盾。具体而言, 状态条件距离函数$g(s,a)$显式刻画了当前策略动作相对于离线数据支持区域的偏离程度。当策略在在线微调初期倾向于访问分布外或高风险区域时, 该距离度量能够直接反映潜在的分布偏移风险, 从而通过约束项抑制由 Q 函数外推误差引起的过度乐观更新。在此基础上, 引入的拉格朗日乘子λ通过对偶优化进行自适应更新。当策略偏离离线数据支持区域的程度增大时, λ会相应增大, 从而增强约束强度, 限制策略对不可靠Q值的利用; 反之, 随着在线交互的进行, 策略逐渐积累高质量在线经验, 状态–动作分布向可靠区域扩展, 距离偏差减小, λ将自动下降, 使约束逐步放松, 允许策略更充分地探索并利用在线数据。

这种由策略行为与数据几何关系驱动的动态调节机制, 在无需人工设定时间表或经验规则的情况下, 实现了对分布偏移风险与探索需求的自适应平衡: 在线微调早期以稳定性和安全性为主, 后期则逐步释放探索能力, 从而避免“过度保守”与“盲目探索”这两种极端情形。

为进一步验证所提出动态调节机制的优势, 本文额外设计了两种对比策略。

1) 标准的严格指数衰减方法(exponential decay, ExpDecay), 其更新公式为$\lambda=\lambda_{\text{start}}\cdot(\gamma_{\text{base}})^{t}$, 其中$\gamma_{\text{base}}=\exp\left(\dfrac{\ln(\lambda_{\text{end}}/\lambda_{\text{start}})}{d}\right)$。衰减率$d$在训练开始时预设, 确保在指定步数内从$\lambda_{\text{start}}$平滑衰减至$\lambda_{\text{end}}$。

2) 基于性能反馈的自适应指数衰减方法(performance-informed exponential decay, PerfDecay), 其核心公式仍保持严格的指数衰减结构$\lambda_t=\lambda_{\text{start}}\cdot\geqslant(\gamma_t)^{\min(t,\text{decay})}$, 其中, 动态衰减因子$\gamma_t$定义为

$$\gamma_t=\begin{cases}\gamma_{\text{base}}\cdot(1-\alpha), & R_t>R_{\text{th}}\\ \gamma_{\text{base}}, & \text{其他}\end{cases}$$

$$\gamma_{\text{base}}=\exp\left(\frac{\ln(\lambda_{\text{end}}/\lambda_{\text{start}})}{d}\right)$$

严格指数衰减方法方法能够提供一个稳定的从强约束到弱约束的过渡, 但缺乏灵活性, 无法根据策略演化或任务难度动态调整。自适应指数衰减方法通过引入性能监测机制: 当策略评估得分超过阈值时, 动态调整衰减速度, 以便更快地减弱约束强度; 反之则保持原始速度。该方法通过性能反馈实现自适应调节, 但缺乏对策略偏离几何信息的建模。

在具体实现中, 本文将状态条件距离函数与动态约束机制集成至 Cal-QL[23] 的策略更新流程中。预训练阶段使用固定的离线数据集对策略网络、Q 网络与距离网络$g(s,a)$同步训练, 完成策略初始化; 随后进入在线微调阶段, 策略开始与环境交互并持续采集新数据, 形成在线回放缓冲区。每轮训练时, 从离线数据和在线数据中分别按比例混合采样, 用于更新策略与 Q 函数。本文的整体架构如图 1 所示。

策略更新采用最大化目标Q值的方式进行, 同时加入几何约束项形成优化目标:

$$\mathcal{L}_{\text{policy}}=\mathrm{E}_{(s,a)\sim\pi}\left[\alpha\cdot\log\pi(a|s)-\beta\cdot Q(s,a)\right]+\lambda\cdot\left(g(s,a)-\mathrm{g}\left(s,a_{\text{expert}}\right)\right)\tag{3}$$

式中: 前两项来自策略熵正则化与 Q 函数加权项, $\lambda\cdot\left(g(s,a)-g\left(s,a_{\text{expert}}\right)\right)$为引入的几何约束正则项。$g\left(s,a_{\text{expert}}\right)$使用离线数据中的动作作为参考, 其作用是提供一个稳定的“数据中心”或“参考点”, 从而使得策略产生的行为能够在相对距离上保持靠近离线分布而非绝对位置。这样的构造可有效避免策略盲目朝任意方向偏移, 提升几何稳定性。

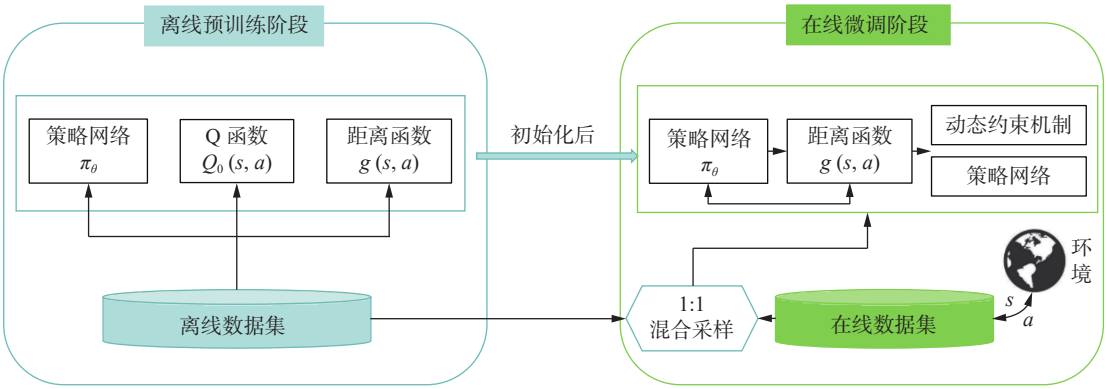

图 1　本文方法整体架构

Fig. 1　Overall framework of our method

在算法实现中，采用双 Q 网络结构，并对 Q 网络进行标准的时序差分 (temporal-difference, TD) 目标训练。具体而言，Q 网络通过最小化损失函数进行更新：

$$\mathcal{L}_Q = \mathrm{E}_{(s,a,r,s')\sim D}\left[(Q_\theta(s,a)-y)^2\right] \tag{4}$$

式中：$y = r+\gamma\cdot\min\limits_{i=1,2} Q_{\theta_i'}\left(s',\pi_{\phi'}(s')\right)$，$\theta$表示当前 Q 网络参数，$\theta_i'$表示两个目标 Q 网络的参数，$\pi_{\phi'}$是目标策略网络，$\gamma$为折扣因子。目标网络采用软更新机制：

$$\theta' \leftarrow \tau\cdot\theta_i+(1-\tau)\cdot\theta_i' \tag{5}$$

同时，对状态条件距离函数$g_\phi(s,a)$进行同步更新，使其能够适应不断变化的数据分布。在训练过程中，ϕ通过最小化监督回归目标进行学习：

$$\mathcal{L}_{\text{dist}} = \mathrm{E}_{(s,a)\sim D}\left[\mathrm{E}_{\hat{a}\sim\text{Unif}(A)}\left(\|a-\hat{a}\|-g_\phi(s,\hat{a})\right)^2\right] \tag{6}$$

该结构可有效缓解值函数的过高估计问题，并与状态条件距离函数联合训练，提升在线微调早期策略更新的鲁棒性。具体的算法伪代码如算法 1 所示。

算法 1　动态距离约束的离线到在线强化学习算法 (D2O2O)

输入　离线数据集D_{offline}，策略网络π_θ，距离网络g_ϕ，Q网络$\{Q_{\theta_1},Q_{\theta_2}\}$，目标网络$\{Q'_{\theta_1},Q'_{\theta_2}\}$，动态约束参数$\lambda$

输出　调整后的策略π_θ

1) 使用D_{offline}对π_θ、g_ϕ、Q进行训练，完成初始化

2) 初始化在线回放缓冲区$D_{\text{online}} \leftarrow \varnothing$

3) for t=1 to T_{online} do

4) 在环境中采样$a\sim\pi_\theta(s)$得到(s,a,r,s')

5) 将(s,a,r,s')存入D_{online}

6) 从D_{offline}与D_{online}中按比例混合采样批次

7) $y \leftarrow r+\gamma\min Q'_{\theta_i}(s',\pi_\theta(s'))$

8) $\mathcal{L}_Q = \sum\limits_i (Q_\theta(s,a)-y)^2$

9) 更新 Q 网络参数θ_i以最小化$\mathcal{L}_Q$

10) 对于 batch 中的每一个(s,a)，采样$\hat{a}\sim\text{Unif}(A)$

11) $\mathcal{L}_g = \left(\|a-\hat{a}\|-g_\phi(s,\hat{a})\right)^2$

12) 更新g_ϕ以最小化$\mathcal{L}_g$

13) $\mathcal{L}_\pi = -Q_{\theta_1}(s,\pi_\theta(s))+\lambda g_\phi(s,\pi_\theta(s))$

14) 更新π_θ以最小化$\mathcal{L}_\pi$

15) $\lambda \leftarrow \text{clip}(\lambda+\eta(g(s,a)-g(s,a_{\text{expert}})-\delta),\lambda_{\min},\lambda_{\max})$

16) endfor

17) $\theta' \leftarrow \tau\cdot\theta_i+(1-\tau)\cdot\theta_i'$

4　实验及结果分析

本节旨在验证所提出方法在离线到在线强化学习任务中的有效性与稳定性。实验设计为分两部分：首先在 D4RL 基准环境上与多个代表性方法进行性能对比，随后通过消融实验进一步探究状态条件距离函数与动态参数调节机制对最终性能的贡献。为更深入评估方法在策略偏离与分布外场景中的适应能力，额外设计了一个验证性实验，分析离线数据覆盖度不足时各方法的策略微调表现，以突出状态条件距离函数在早期策略偏移场景下提供结构化数据引导的重要作用。

4.1　与基础算法的对比实验

本文在 MuJoCo 物理引擎下的 D4RL[35] 基准环境中，选取了 Walker2d、HalfCheetah 和 Hopper 等 3 个任务，涵盖不同数据质量（medium、medium-re-play、medium-expert）与难度设置，共计 9 个子任务。对比方法包括在线决策转换器 (online decision transformer, ODT)[36]、IQL[21]、PEX[28]、Cal-QL[23] 以及乐观探索与元适应 (optimistic exploration and meta adaptation, OEMA)[29]，其中 OEMA 的结果直接引用自其原论文，在线微调阶段为 30 万步，其余方法统一使用 100 万步离线预

训练 + 20 万步在线微调的设定。

实验结果如表 1 所示, 在绝大多数环境中, 所提方法均表现出领先的性能, 特别是在 Walker2d-medium-expert 等具有较强策略退化风险的任务上, 显著优于现有方法。需要指出的是, 尽管所提算法在在线阶段仅使用了 20 万步的交互数据, 其总体性能依然优于 OEMA 的 30 万步结果, 展示了方法在训练效率上的潜在优势。这表明状态条件距离函数与动态参数调节机制在早期提供稳定约束的同时, 能保证策略在有限交互中快速提升。同时对比发现, 不同算法在任务间表现差异明显。例如 IQL 在 medium 任务上与 Cal-QL 接近, 但在 expert 数据集上出现退化, 说明单纯的悲观值函数约束难以兼顾不同数据质量。PEX 在 medium-replay 下表现较好, 但在 expert 环境中不稳定, 可能与策略集冗余性相关。相比之下, 本文方法在不同数据设置下表现更为稳健, 体现了状态条件距离函数与动态调节结合的优势。

表 1　本文方法和其他离线到在线强化学习方法在 D4RL 数据集上的性能比较

Table 1　Comparison of the proposed method and other offline-to-online reinforcement learning methods on the D4RL dataset

任务	ODT	IQL	Cal-QL	PEX	OEMA	D2O2O
Walker2d-medium-expert	108.7	108.9	111.1	117.9	102.8	**111.3**
Walker2d-medium-replay	76.8	83.6	87.3	89.3	96.2	**96.6**
Walker2d-medium	76.7	83.6	84.2	**94.8**	89.3	85.6
Halfcheetah-medium-expert	94.1	87.9	91.0	91.0	86.1	**95.9**
Halfcheetah-medium-replay	40.4	45.6	59.5	55.2	**68.7**	65.4
Halfcheetah-medium	42.1	48.6	72.3	67.8	75.8	**77.4**
Hopper-medium-expert	111.1	90.0	107.9	107.8	109.3	**112.1**
Hopper-medium-replay	88.8	99.0	102.2	**103.3**	101.4	101.8
Hopper-medium	97.5	59.2	97.6	78.6	100.5	**100.6**
总计	736.2	706.4	813.1	805.7	829.1	**832.3**

注: 加粗表示在该任务中结果最好。

此外, 表 2 给出了参数调节方式的对比结果。将本文提出的方法与两种基线方法进行比较, 其一是基于性能反馈的自适应指数衰减(PerfDecay), 其二是标准指数衰减(ExpDecay)。结果表明, 尽管 ExpDecay 能够在后期放松约束促进探索, 但在早期阶段约束过弱, 导致策略偏离数据支持区域而带来不稳定性; PerfDecay 通过性能反馈自适应调整衰减速率, 在一定程度上改善了早期稳定性, 但整体性能仍逊于我们的方法。相较之下, 提出的方法能够在约束与策略优化之间实现更细致的自适应平衡, 因此在面对复杂分布偏移时往往难以兼顾稳定性与性能提升。

表 2　本文方法与不同参数调节方式的对比结果

Table 2　Comparison results of our method with different parameter adjustment strategies

任务	ExpDecay	PerfDecay	D2O2O
Walker2d-medium-expert	110.3	111.1	**111.3**
Walker2d-medium-replay	95.3	95.4	**96.6**
Walker2d-medium	85.0	85.3	**85.6**
Halfcheetah-medium-expert	95.6	95.9	**95.9**
Halfcheetah-medium-replay	52.9	57.1	**65.4**
Halfcheetah-medium	62.8	72.7	**77.4**
Hopper-medium-expert	107.6	108.3	**112.1**
Hopper-medium-replay	97.9	101.6	**101.8**
Hopper-medium	96.8	100.2	**100.6**
总计	804.2	827.6	**832.3**

注: 加粗表示在该任务中结果最好。

上述实验结果验证了本文提出的状态条件距离函数与动态参数调整机制在微调早期提供稳定约束的有效性，同时也说明在约束逐步放松的过程中，策略能够有效吸收新经验并实现性能提升。

4.2　消融实验

为进一步分析方法中关键模块对整体性能的贡献，在标准 D4RL 环境上设计了两组消融实验。第 1 组实验移除了状态条件距离函数项（绿色曲线），以验证该模块在在线微调早期的稳定性作用；第 2 组实验则将策略正则强度 Q 固定 (Fixed) 为一个常数（橙色曲线），以评估动态调节机制在策略演化过程中的自适应能力。

如图 2 所示，本文方法（蓝色曲线）在微调初期即展现出显著的性能提升，并在整个训练过程中保持稳定领先。相比之下，去除状态条件距离函数的方法在微调早期表现不稳定，收敛速度明显减慢，表明状态条件距离函数在策略偏离严重时提供了有效的边界引导；而固定参数策略在中后期表现受限，说明 Q 的动态调节机制有助于策略从保守探索过渡到性能提升阶段。

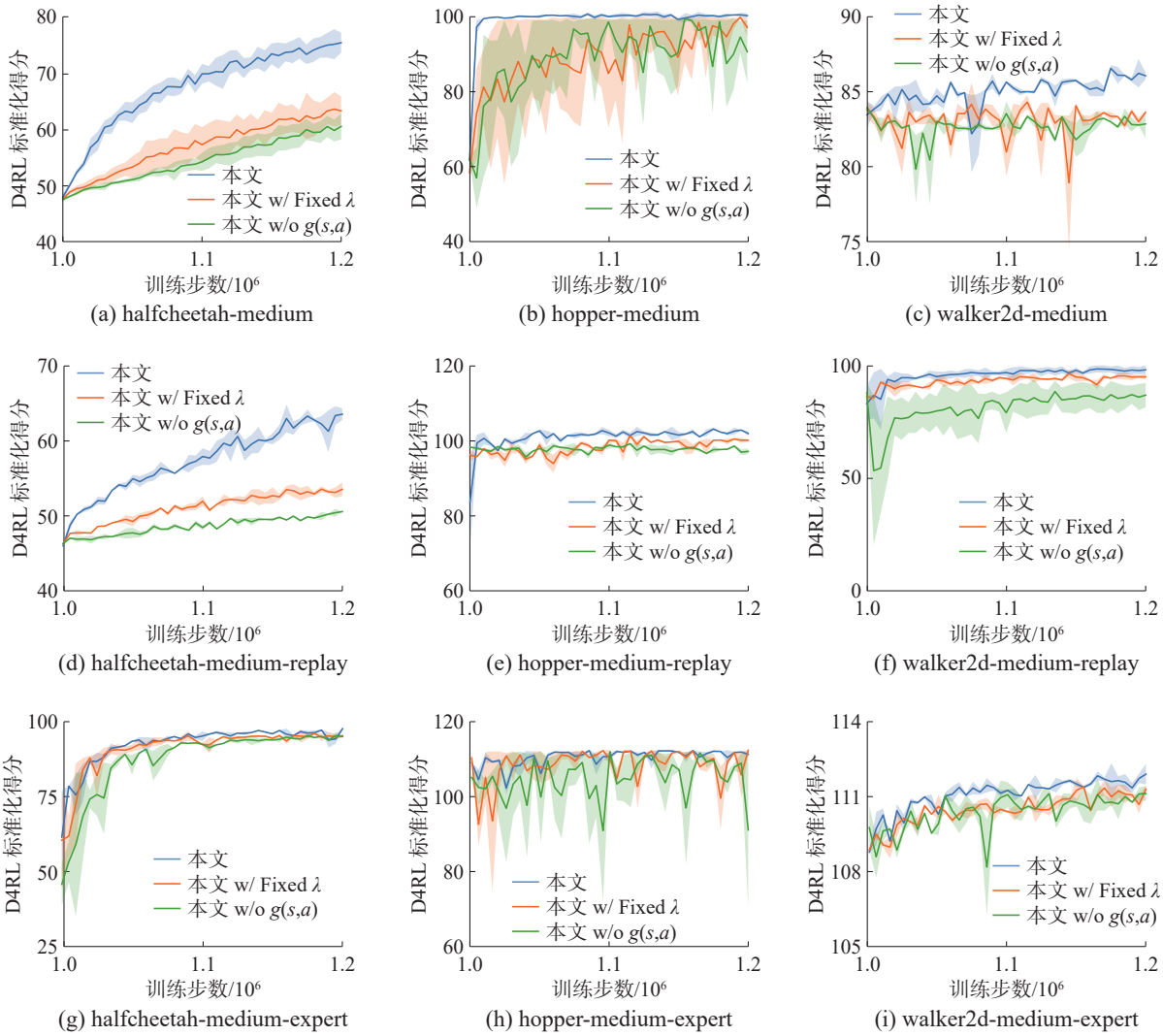

图 2　消融实验结果

Fig. 2　Results of ablation study

4.3　离线数据覆盖度对策略微调性能的影响

为验证本文方法在策略初始分布偏离数据集支持区域时的有效性与鲁棒性，设计了一个针对离线数据覆盖度的验证性实验。该实验旨在模拟一种更具挑战性的场景：初始策略产生的状态–动作分布与离线数据存在明显偏移，从而检验状态条件距离函数与动态正则机制在缓解策略崩溃与提升泛化能力方面的作用。

具体地，基于原始的初始策略，设计了一种策略搅动机制，在训练初期（即进入在线微调阶段前），对策略网络加入扰动项，打乱其在状态空间中的分布，从而人为增加策略与离线数据之间的偏移程度。该机制模拟了现实中某些离线策略存在偏差或目标任务变动导致策略失效的情形。在

此设置下，分别比较了本文方法与其变体（去除状态条件距离函数）在策略微调过程中的表现，如图 3 所示。

从图 3 可以观察到，对比方法在微调初期由于策略偏移，导致 Q 函数估计出现剧烈震荡，最终性能停滞在较低水平。而所提方法在引入状态条件距离函数后，能够有效感知策略与离线数据支持集之间的偏离程度，并通过动态惩罚调节机制稳步引导策略向更可靠的区域收敛，最终在微调过程中持续提升性能。该实验进一步验证了状态条件距离函数在预训练不足或策略初始偏移场景下的鲁棒性优势，说明所提出的正则项机制在策略安全性与适应性之间达成了更优的平衡。

离线到在线强化学习中，微小的最终性能差异往往掩盖了训练过程中显著不同的动态行为。如图 3 所示，在引入策略搅动后，去除状态条件距离函数的对比方法在微调初期即出现明显的性能震荡甚至长期停滞，而本文方法能够在显著分布偏移条件下保持稳定上升趋势。这一现象表明，本文方法的优势不仅体现在最终回报数值上，更体现在避免策略早期崩溃、缩短恢复时间、提升训练可控性等方面。对于离线到在线强化学习场景而言，这种稳定性提升直接决定了方法是否具备实际可用性。

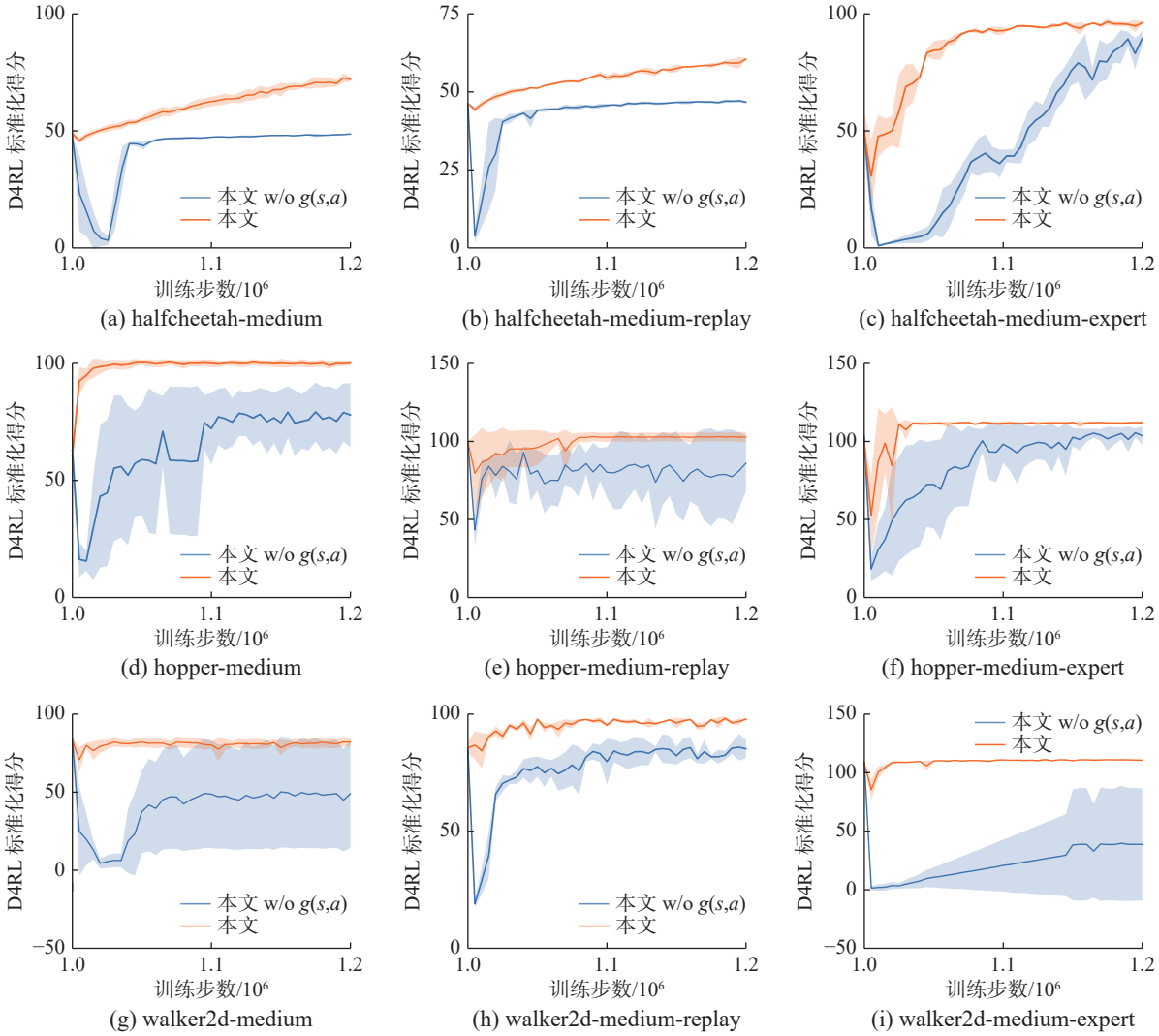

图 3　所提方法与其变体在策略微调过程中的表现

Fig. 3　Policy fine-tuning performance of the proposed method compared to its variant

4.4　运行时间与计算开销分析

为评估本文方法在实际训练过程中的计算开销，本文进一步对不同算法的运行时间进行了对比分析。考虑到 D2O2O 是在 Cal-QL 的基础上引入状态条件距离约束与动态调节机制而形成的改进方法，而 Cal-QL 本身又是对 CQL[20] 的校准性改进，因此本节重点比较了 CQL、Cal-QL 以及本文提出的 D2O2O 等 3 种方法在单回合运行时间上的差异，实验结果如表 3 所示。所有实验均在相同的硬件与软件环境下进行，以保证对比的公

平性。实验在一台配备多核 CPU、充足内存以及 NVIDIA GPU 的计算服务器上完成，训练过程采用 GPU 加速，所有方法均在单 GPU 设置下运行，CPU 仅用于环境交互与数据调度。

表 3　单回合运行时间对比
Table 3　Per-episode runtime comparison　s

任务	Walker2d medium-replay	Halfcheetah medium-replay	Hopper medium-replay
CQL	38.4	37.1	34.8
Cal-QL	34.5	38.7	39.5
D2O2O	36.0	38.5	45.2

从表 3 可以观察到，D2O2O 的单回合运行时间与 CQL 及 Cal-QL 处于同一数量级。在 Walker2d 与 HalfCheetah 任务中，D2O2O 的运行时间与 Cal-QL 基本相当，仅存在轻微差异；在 Hopper 任务中，D2O2O 的运行时间有所增加，但整体仍保持在可接受范围内。需要指出的是，这一开销增长并非源于额外的环境交互或复杂优化过程，而主要来自状态条件距离网络在策略更新阶段引入的有限前向与反向计算。尤其是 D2O2O 中引入的状态条件距离网络结构较为轻量，其计算开销相对于 Q 网络与策略网络而言占比有限，因此不会成为训练过程中的主要瓶颈。

上述结果表明，尽管本文方法在算法结构上引入了额外的距离评估模块，但其计算复杂度并未随之显著上升。这表明本文方法并非通过大幅增加计算成本来换取有限的性能提升，而是在保持训练效率基本不变的前提下，提高了在线微调阶段的稳定性与鲁棒性，从而在性能收益与计算代价之间实现了合理权衡。

5　结束语

本文提出了一种面向离线到在线强化学习场景的新方法 (D2O2O)，核心思想是在策略优化过程中引入状态条件距离函数作为正则项，并通过对偶优化机制动态调整约束强度。该方法能够有效感知策略与训练数据的偏离程度，从而在微调早期提供稳定性保障，并在后期逐步放松约束以释放策略的性能潜力。实验结果表明，本文方法在多个 D4RL 基准任务上表现出优于现有方法的性能，尤其在训练资源更少的设定下仍保持了领先的在线微调效果。消融实验进一步验证了状态条件距离函数与动态调节机制在策略稳定性与最终性能方面的关键作用。本文的研究为如何在策略迁移过程中平衡安全性与优化能力提供了一种可行的思路。

参考文献：

[1] 刘全, 翟建伟, 章宗长, 等. 深度强化学习综述[J]. 计算机学报, 2018, 41(1): 1−27.
LIU Quan, ZHAI Jianwei, ZHANG Zongchang, et al. A survey on deep reinforcement learning[J]. Chinese journal of computers, 2018, 41(1): 1−27.

[2] 王硕汝, 牛温佳, 童恩栋, 等. 强化学习离线策略评估研究综述[J]. 计算机学报, 2022, 45(9): 1926−1945.
WANG Shuoru, NIU Wenjia, TONG Endong, et al. Research on off-policy evaluation in reinforcement learning: a survey[J]. Chinese journal of computers, 2022, 45(9): 1926−1945.

[3] TANG Shengpu, WIENS J. Model selection for offline reinforcement learning: practical considerations for healthcare settings[EB/OL]. (2021−07−23)[2025−10−16]. https://arxiv.org/abs/2107.11003.

[4] 刘健, 顾扬, 程玉虎, 等. 基于多智能体强化学习的乳腺癌致病基因预测[J]. 自动化学报, 2022, 48(5): 1246−1258.
LIU Jian, GU Yang, CHENG Yuhu, et al. Prediction of breast cancer pathogenic genes based on multi-agent reinforcement learning[J]. Acta automatica sinica, 2022, 48(5): 1246−1258.

[5] 吴晓光, 刘绍维, 杨磊, 等. 基于深度强化学习的双足机器人斜坡步态控制方法[J]. 自动化学报, 2021, 47(8): 1976−1987.
WU Xiaoguang, LIU Shaowei, YANG Lei, et al. A gait control method for biped robot on slope based on deep reinforcement learning[J]. Acta automatica sinica, 2021, 47(8): 1976−1987.

[6] XIAO Teng, WANG Donglin. A general offline reinforcement learning framework for interactive recommendation[J]. Proceedings of the AAAI conference on artificial intelligence, 2021, 35(5): 4512−4520.

[7] GAO Chongming, WANG Shiqi, LI Shijun, et al. CIRS: bursting filter bubbles by counterfactual interactive recommender system[J]. ACM transactions on information systems, 2024, 42(1): 1−27.

[8] NAIR A, GUPTA A, DALAL M, et al. AWAC: accelerating online reinforcement learning with offline datasets [EB/OL]. (2020−06−17)[2025−09−11]. https://arxiv.org/abs/2006.09359.

[9] MAO Yihuan, WANG Chao, WANG Bin, et al. MOORe: model-based offline-to-online reinforcement learning [EB/OL]. (2022−01−23)[2025−09−11]. https://arxiv.org/abs/2201.10070.

[10] BEESON A, MONTANA G. Improving TD3-BC: relaxed policy constraint for offline learning and stable online fine-tuning[EB/OL]. (2022−11−21)[2025−09−11]. https://arxiv.org/abs/2211.11802.

[11] GUO Siyuan, SUN Yanchao, HU Jifeng, et al. A simple unified uncertainty-guided framework for offline-to-online reinforcement learning[EB/OL]. (2023−06−13)[2025−09−11]. https://arxiv.org/abs/2306.07541.

[12] ZHENG Han, LUO Xufang, WEI Pengfei, et al. Adaptive policy learning for offline-to-online reinforcement learning[J]. Proceedings of the AAAI conference on artificial intelligence, 2023, 37(9): 11372−11380.

[13] LUO Yicheng, KAY J, GREFENSTETTE E, et al. Finetuning from offline reinforcement learning: challenges, trade-offs and practical solutions[EB/OL]. (2023−03−30)[2025−09−11]. https://arxiv.org/abs/2303.17396.

[14] LI Jianxiong, HU Xiao, XU Haoran, et al. PROTO: iterat-

ive policy regularized offline-to-online reinforcement learning[EB/OL]. (2023−05−25)[2025−09−11]. https://arxiv.org/abs/2305.15669.
[15] HU Hao, YANG Yiqin, YE Jianing, et al. Bayesian design principles for offline-to-online reinforcement learning [EB/OL]. (2024−05−31)[2025−09−11]. https://arxiv.org/abs/2405.20984.
[16] KONG Rui, WU Chenyang, GAO Chenxiao, et al. Efficient and stable offline-to-online reinforcement learning via continual policy revitalization[C]//International Joint Conference on Artificial Intelligence. Jeju: IJCAI, 2024.
[17] SUTTON R S, BARTO A G. Reinforcement Learning [M]. Cambridge: MIT Press, 1998: 9−11.
[18] KUMAR A, ZHOU A, TUCKER G, et al. Conservative Q-learning for offline reinforcement learning[C]//Advances in Neural Information Processing Systems. New York: Curran Associates Inc. , 2020: 1179−1191.
[19] FUJIMOTO S, MEGER D, PRECUP D. Off-policy deep reinforcement learning without exploration[C]//Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019: 2052−2062.
[20] 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.
[21] LI Jianxiong, ZHAN Xianyuan, XU Haoran, et al. When data geometry meets deep function: generalizing offline reinforcement learning[EB/OL]. (2022−05−23)[2025−09−16]. https://arxiv.org/abs/2205.11027.
[22] FUJIMOTO S, VAN HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[C]//International Conference on Machine Learning. Stockholm: PMLR, 2018: 1587−1596.
[23] ZHENG Qinqing, ZHANG A, GROVER A. Online decision Transformer[C]//Proceedings of the 39th International Conference on Machine Learning. Baltimore: PMLR, 2022: 27042−27059.
[24] KOSTRIKOV I, NAIR A, LEVINE S. Offline reinforcement learning with implicit Q-learning[EB/OL]. (2021−10−12)[2025−09−16]. https://arxiv.org/abs/2110.06169.
[25] ZHANG Haichao, XU W, YU Haonan. Policy expansion for bridging offline-to-online reinforcement learning [EB/OL]. (2023−02−02)[2025−09−16]. https://arxiv.org/abs/2302.00935.
[26] FINN C, KUMAR A, LEVINE S, et al. Cal-QL: calibrated offline RL pre-training for efficient online fine-tuning[C]//Advances in Neural Information Processing Systems 36. New Orleans: Neural Information Processing Systems Foundation, Inc, 2023: 62244−62269.
[27] GUO Siyuan, ZOU Lixin, CHEN Hechang, et al. Sample efficient offline-to-online reinforcement learning[J]. IEEE transactions on knowledge and data engineering, 2024, 36(3): 1299−1310.
[28] FU J, KUMAR A, NACHUM O, et al. D4RL: datasets for deep data-driven reinforcement learning[EB/OL]. (2020−04−15)[2025−09−16]. https://arxiv.org/abs/2004.07219.
[29] FUJIMOTO S, GU S S. A minimalist approach to offlinereinforcement learning[C]//Advances in Neural Information Processing Systems. Virtual Conference: CurranAssociates Inc. , 2021: 20132−20145.
[30] LEE S, SEO Y, LEE K, et al. Offline-to-online reinforcement learning via balanced replay and pessimistic Q-ensemble[EB/OL]. (2021−07−15)[2025−09−16]. https://arxiv.org/abs/2107.00591.
[31] CHEN Hao, GAO Jiawei, HUANG Gao, et al. Train once, get a family: state-adaptive balances for offline-to-online reinforcement learning[C]//Advances in Neural Information Processing Systems 36. New Orleans: Neural Information Processing Systems Foundation, Inc. , 2023: 47081−47104.
[32] LIU Xuhui, LIU Tianshuo, JIANG Shengyi, et al. Energy-guided diffusion sampling for offline-to-online reinforcement learning[EB/OL]. (2024−07−17)[2025−09−16]. https://arxiv.org/abs/2407.12448.
[33] HE Longxiang, YE Deheng, TAN Junbo, et al. Robust policy expansion for offline-to-online RL under diverse data corruption[EB/OL]. (2025−09−29)[2025−10−03]. https://arxiv.org/abs/2509.24748.
[34] HUANG Xiao, LIU Xu, ZHANG Enze, et al. Offline-to-online reinforcement learning with classifier-free diffusion generation[EB/OL]. (2025−08−09)[2025−09−16]. https://arxiv.org/abs/2508.06806.
[35] HUANG Shengjun, LUO Qinwen, WANG Yewen, et al. Optimistic critic reconstruction and constrained fine-tuning for general offline-to-online RL[C]//Advances in Neural Information Processing Systems 37. Vancouver: Neural Information Processing Systems Foundation, Inc. , 2024: 108167−108207.
[36] CHEN Siqi, ZHAO Jianing, ZHAO Kai, et al. ANOTO: improving automated negotiation via offline-to-online reinforcement learning[C]//Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems. London: ACM, 2024: 2195−2197.

作者简介：

赵若涵，硕士研究生，主要研究方向为强化学习。E-mail：zruohan2023@163.com。

魏巍，教授，博士生导师，山西大学计算机与信息技术学院（大数据学院）副院长，主要研究方向为数据挖掘、机器学习与具身智能。主持和参与国家重点研发计划项目、国家自然科学基金重点项目、国家自然科学基金面上项目、山西省自然科学基金项目 20 余项。发表学术论文 40 余篇。E-mail：weiwei@sxu.edu.cn。

王达，讲师，博士，主要研究方向为强化学习和具身智能，获国家发明专利授权 2 项，发表学术论文 11 篇。E-mail：wanda@sxu.edu.cn。