基于 Q 网络集成下置信域引导的高效扩散策略

张 旭，曾玉婷，张 峰†

(河北大学 数学与信息科学学院, 河北 保定 071000)

摘 要: 离线强化学习中的数据常由多种策略混合采集, 导致动作空间呈现复杂多模分布. 现有扩散策略虽能有效刻画多模分布, 但由于其动作生成依赖多步逆向推理, 效率较低. 此外, 生成的动作可能接近分布外区域, 但并不完全超出数据支持范围, 这种不确定性容易引起 Q 值高估, 导致策略不稳定或性能退化. 鉴于此, 提出一种基于 Q 集成下置信域引导的高效扩散策略 (E2DP). 该方法通过设计两步逆向推理机制, 在显著降低推理开销的同时保留对多模的动作分布的建模能力. 为解决 Q 值高估问题, 引入集成 Q 网络下置信域估计, 利用独立目标函数和随机权重系数的方差正则化增强网络多样性, 使 E2DP 仅需少量 Q 网络即可在分布内动作与数据支持附近的潜在高价值候选之间形成有效权衡, 有效提升策略的性能及鲁棒性. 在 Bandit任务和 D4RL 基准任务上的实验结果表明, E2DP 在保持与现有扩散策略相当分布表达能力的同时, 推理速度提升约 2.5 倍, 并在多个任务中取得最优归一化得分.

关键词: 离线强化学习；扩散策略；两步逆向推理；集成 Q 网络；下置信域引导

中图分类号: TP18 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.1127

引用格式: 张旭, 曾玉婷, 张峰. 基于 Q 网络集成下置信域引导的高效扩散策略 [J]. 控制与决策, xxxx, x(x): xxxx-xxxx.

Efficient Q-ensembles diffusion policy for offline reinforcement learning

ZHANG XU，*ZENG Yu-ting*，*ZHANG Feng*†

(College of Mathematics and Information Science，Hebei University，Baoding 071000，China)

Abstract: Offline reinforcement learning datasets are often collected by mixtures of different behaviour policies, resulting in complex multimodal action distributions. Diffusion-based policy methods can effectively model such offline action distributions; however, their action generation relies on multi-step reverse diffusion, which results in high inference latency, becoming a significant bottleneck for practical deployment. Additionally, while the generated actions may still fall within the distribution's support, the inherent uncertainty can lead to overestimation of Q-values, thereby compromising policy stability. To address the challenges, we propose an efficient Q-ensemble diffusion policy (E2DP). The E2DP significantly reduces the computational cost of action generation through a two-step reverse diffusion mechanism. Meanwhile, ensemble Q networks with variance regularization are employed to improve uncertainty estimation under a small ensemble size, and a lower confidence bound constraint is incorporated during policy improvement to balance in-distribution actions and high-risk candidates near the data support. Experimental results on Bandit and D4RL benchmark demonstrate that the E2DP achieves an inference speedup of approximately 2.5× while maintaining distribution modeling capability comparable to existing diffusion-based policies, and obtain improved normalized performance across multiple tasks.

Keywords: offline reinforcement learning； diffusion policy； two-step inference； Q-ensemble； lower confidence bound

0 引 言

强化学习 (reinforcement learning, RL) 是一种通过智能体与环境动态交互持续优化决策策略的学习范式, 目标是获得长期回报最大的最优策略.目前 RL 已在机器人控制[1-2]、游戏博弈[3]、自动驾驶[4-5] 等领域取得了显著进展, 为复杂场景下的决策优化提

收稿日期: 2025-10-29；**录用日期:** 2026-02-26.

基金项目: 科技部重点研发项目 (2022YFE0196100)；教育部春晖合作项目 (HZKY20220256-202200417).

责任编委: 侯忠生.

†通信作者. E-mail: fengzhang@hbu.cn.

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.

供了有效解决方案.然而, 在许多现实高风险场景中, 在线 RL 依赖持续探索获取数据, 其交互成本高昂且伴随不可忽视的安全风险. 此外, 在线模式下产生的大量交互轨迹往往难以高效复用, 训练效率与数据利用率都受到限制.

离线强化学习 (Offline Reinforcement Learning, ORL)[6-10] 通过仅依赖预先收集的静态数据集进行策略学习, 在无需与真实环境交互的条件下完成优化, 从而有效规避探索风险并提升数据利用效率. 然而, 离线 RL 面临的核心挑战在于分布偏移[9-13] 问题: 学习策略在优化过程中可能生成超出数据支持范围的状态–动作对 (Out-of-Distribution, OOD). 由于这些样本缺乏真实回报约束, 其价值估计依赖函数外推结果, 并在时序差分 (TD) 更新的自举机制下被逐步放大, 最终导致 Q 值系统性高估与策略性能退化.

现有研究主要从策略约束[9-12]、价值约束[13-14] 等方向缓解分布偏移问题. 其中, 策略约束方法因原理直观、工程实现难度低, 且与主流在线 RL 算法 (如 SAC[15]、TD3[16]) 兼容性强, 成为当前离线 RL 领域的研究热点. 然而, 多数策略约束方法隐含假设离线数据的动作分布为单峰或近似单峰分布[10], 当离线数据由多种不同性能、不同目标的行为策略混合采集时, 动作分布通常呈现多峰、高维且非对称的复杂结构. 此时, 传统策略约束方法中常用的简单生成模型 (如 CVAE[17]) 由于其表达能力不足, 容易出现“模式平均”现象, 即生成的动作集中于多个分布模态的均值区域, 而该区域往往属于离线数据分布的低密度区域, 反而进一步加剧分布偏移风险.

扩散策略 (Diffusion Policy)[10], 将扩散模型[18] 作为策略函数的参数化形式, 扩散策略能有效拟合由混合行为策略产生的复杂动作结构, 在 D4RL 基准任务上取得了显著优于条件变分自编码器 (CVAE) 及确定性策略等传统参数化方式的性能[10]. 然而, 扩散策略的动作生成依赖于多步逆向扩散推理过程, 导致推理延迟高, 难以满足实时应用需求; 最近有部分工作在结合一步生成模型[19-20] 来提升动作生成的速度, 然而, 这些模型将动作生成过程变换为单步确定性的映射, 在处理离线复杂多模分布时, 过于确定的生成路径可能限制了学习策略的性能.

此外, 即便采用上述模型在一定程度上增强了对离线动作分布的拟合能力, 扩散策略在推理阶段生成的动作仍可能落入数据支持不足区域. 本文将这类动作称为 Semi-OOD 动作, 它通常位于离线数据动作分布的邻域, 由于离线数据集规模有限, 其中部分动作可能实际上属于真实可行的行为分布, 相较于完全偏离数据分布的 OOD 动作更具潜在价值. 然而, Q 网络往往对这些数据支持不足的动作产生过度乐观的价值估计.

这种价值高估会误导策略优化方向, 使策略倾向于选择真实价值较低的动作, 最终导致策略收敛至次优解, 甚至在稀疏奖励场景下出现性能崩溃. 因此, 有必要引导 Q 网络进行更加审慎且真实的价值权衡, 从而提升离线强化学习中策略优化的稳定性与可靠性

为解决这些问题, 本文提出一种简单而高效的改进方法, 兼顾扩散策略的效率与表现. 主要贡献可概括如下:

1) 提出一种两步逆向推理机制, 在显著提升扩散策略推理效率的同时, 有效保持对多模动作分布的宏观结构建模能力.

2) 设计了基于集成 Q 网络的下置信域 (lower confidence bound, LCB) 引导机制, 在 Q 网络训练中引入随机方差正则化, 增强集成 Q 网络的多样性. 在较小规模的集成 Q 网络中, 该机制能够构造出低不确定性的下置信域值, 从而在分布内动作与潜在高价值候选之间形成有效权衡.

3) 在 Bandit 和 D4RL 基准任务上的实验表明, E2DP 显著提升了推理效率, 并在稀疏奖励、长程规划的 Antmaze 任务中获得最优归一化得分, 展现了在复杂场景下的稳定性和应用潜力

1 相关工作

本节回顾离线强化学习的基本概念, 分析其主要挑战及代表性解决方案, 并重点介绍扩散策略在离线强化学习中的典型工作.

1.1 离线强化学习

离线强化学习是强化学习的重要分支. 与在线强化学习不同, 离线强化学习不依赖实时环境交互, 而是直接从一个预先收集的静态数据集进行学习. 该数据集由状态、动作、即时奖励以及转移后的下一状态组成: $D = \{s, a, r, s'\}$, 离线数据可由任意单一策略或混合策略收集. 由于现实中的数据受采集成本、环境复杂性或策略偏好的影响, 几乎不可能覆盖环境中所有可能的“状态-动作”组合及对应的转移关系, 因此传统在线 RL 中基于“真实环境动态”的贝尔曼算子 (Bellman Operator) 无法直接使用, 离线强化学习通常采用经验贝尔曼算子[9,13] (Empirical Bellman Operator, 记为Γ) 来近似更新 Q 函数, 其数学表达式为:

$$(\mathcal{T}^{\pi}Q)(s,a)=r+\gamma E_{s'\sim\hat{p}(\cdot|s,a),a'\sim\pi(\cdot|s')}[Q(s',a')]. \quad (1)$$

其中, π为待学习的目标策略, γ是折扣因子, 用于权衡即时奖励与未来奖励的重要性, $Q(s,a)$表示在状态s执行动作a后, 并在后续决策过程中遵循策略π的期望累积折扣奖励, 其中折扣因子为γ, $\hat{p}(\cdot|s,a)$为经验状态转移概率. 当前主流离线 RL 算法多采用演员-批评家 (Actor-Critic) 架构, 其中 Critic 网络通常遵循 Q-learning 的自举 (bootstrapping) 更新机制, 即利用当前价值估计构造下一步的学习目标. 演员-批评家标准更新形式如 (2) 式.

$$Q_{k+1}=\arg\min_{Q}E_{(s,a,r,s')\sim D}[((\mathcal{T}^{\pi}Q_k)(s',a')-Q(s,a))^2]\pi_{k+1}=\arg\max_{\pi}E_{s\sim D,a\sim\pi(\cdot|s)}[Q_{k+1}(s,a)]. \quad (2)$$

其中$(\mathcal{T}^{\pi}Q_{\hat{\psi}})(s',a')$为目标 Q 网络估计的经验贝尔曼算子, $k>0$表示迭代轮数.

在离线设置下, 当学习得到的目标策略 (学习策略) 在优化过程中生成超出数据集支持范围的动作时, 即产生分布外 (out-of-distribution, OOD) 状态–动作对. 由于这些动作在离线数据集中缺乏真实回报约束, 其价值估计主要依赖函数逼近器的外推结果. 在 Actor–Critic 框架下, 这类外推估计会通过基于自举的时序差分 (temporal difference, TD) 更新被反复用作学习目标, 从而逐步放大估计误差, 最终导致 Critic 在分布外区域产生系统性的价值高估 (外推误差). 该高估进一步误导策略更新方向, 使学习策略偏向于选择数据支持不足但被错误评估为高价值的动作, 进而引发性能退化甚至训练不稳定加剧分布偏移问题. 为缓解上述问题, 现有研究主要从策略约束 (Actor Regularization)[9-12] 和价值约束 (Critic Regularization)[13-14] 两个方向展开.

1.2　策略约束

策略约束方法通过限制学习策略接近行为策略分布来减少 OOD 动作. 典型例子包括行为克隆 (Behavioral Cloning, BC)[21], BC 通过监督学习直接模仿数据中的动作分布, 然而, BC 完全放弃价值优化, 缺乏泛化能力, 在复杂任务中性能有限. 为兼顾模仿与优化, BCQ[9] 引入条件变分自编码器 (Conditional Variational Autoencoder, CVAE) 刻画行为动作分布并结合 Q 值筛选, 但在混合行为策略导致的复杂多模分布下, CVAE 等弱生成模型易出现“模式平均”, 生成动作落入低密度区域, 导致策略选择低置信度动作, 反而加剧分布偏移风险. 为摆脱对显式动作生成模型的依赖, BEAR[11] 采用最大均值差异 (Maximum Mean Discrepancy, MMD) 度量策略分布距离, 并通过动态调整惩罚权重约束策略接近行为策略的数据分布. 但 MMD 的计算复杂度随动作维度显著增长带来显著的硬件开销. TD3+BC[12] 则采取更简洁的思路: 在 TD3 算法[16] 的策略更新目标中加入行为克隆正则项, 使策略在价值最大化与模仿行为数据之间取得平衡. 由于该方法实现简单、效果稳健, 已成为离线强化学习的经典基线, 但在面对高度多模或稀疏奖励的数据集时, 其策略表达能力仍显不足.

近年来, 随着生成模型的发展, 研究者开始探索更强的分布建模工具. Diffusion-QL[10] 等工作以扩散模型作为策略网络, 通过逐步去噪推理建模复杂的动作分布, 并在 D4RL 基准任务中取得了显著优于传统生成模型 (如 VAE[17]、RealNVP[22]) 的表现, 展示了扩散模型在策略约束类方法中的潜力.

1.3　价值约束

价值约束方法通过构建保守的 Q 函数估计, 对 OOD 动作赋予偏低的价值, 间接引导策略避免选择不可靠动作. CQL[13] 通过引入正则项来最小化 OOD 动作的 Q 值, 使得估计值接近真实 Q 值的下界. 但其易导致“过度保守”问题, 策略为规避 OOD 动作风险, 会过度局限于离线数据的高密度区域, 最终削弱策略泛化能力. 对此, IQL[14] 通过期望回归训练状态价值函数$V(s)$, 并在策略更新阶段基于优势加权的行为克隆进行优化, 仅对数据分布内的动作计算优势, 通过优势权重控制策略对不同动作的偏好. 这种设计既避免 OOD 动作的价值高估, 也避免了对分布内动作的显式惩罚. 另一类策略约束方法是基于集成 Q 网络的不确定性引导方法[23-26]: 通过多个 Critic 网络之间的差异来量化价值估计的不确定性, 从而实现保守引导. EDAC[24] 在 SAC 算法中引入多样化的集 Q 网络, 利用梯度相似性惩罚来提升网络独立性, 并结合 Clip-Q 技术[16] 提高价值函数的估计可靠性, 从而提升策略训练的稳定性. 针对多 Q 网络共享目标易导致过度乐观的问题, MSG[25] 进一步强调“独立目标”以降低集成同质化导致的乐观偏差. SCORE[26] 在集成 Q 网络的目标函数中引入与方差相关的惩罚项, 削弱 OOD 动作与高 Q 值的虚假相关性. 然而, SCORE 对退火行为克隆 (BC) 的依赖使其在使其在低质量数据集上的适用性受到限制, 并且固定超参数的方差正则化无法有效调制 Q 网络的多样性, 使得集成 Q 网络对价值不确定性的量化效果仍然依赖集成数量.

1.4 扩散策略

扩散策略作为近年来兴起的策略约束方法, 利用扩散模型强大的分布表达能力对离线数据集中复杂的动作分布进行建模[10], 将动作选择约束在数据集支持范围内, 有效缓解因学习策略与行为策略的分布偏移引发的外推误差问题. 代表性工作有Diffusion-QL[10], EDPQ[27], EDP [28], IDQL[29].

Diffusion-QL 将扩散模型作为策略网络引入离线 RL, 结合行为克隆损失与 Q 值引导, 在 D4RL 基准任务上表现优异. 在训练阶段优化行为克隆损失, 推理阶段通过逆向扩散链推理得到候选动作集, 并依据 Q 值进行多项式采样最终的动作. 该方法采用 DDPM 的扩散架构进行策略参数化, 为平衡分布拟合能力与推理效率[10], 其实验中取T=5 逆向推理, 使扩散策略能够捕捉动作分布的宏观多模结构, 但难以精细复原细粒度的密度信息. 尽管如此, 其算法效率仍显不足. 在实际应用中, 扩散策略的推理效率是影响其可用性的关键因素. EDP(efficient diffusion policy)[28] 通过学习前向加噪空间中的条件映射, 以单步重参数化近似生成动作, 显著降低推理成本. 但在高度多模场景下, 单步映射难以逐步修正生成轨迹, 可能导致不同模态在噪声空间重叠, 从而削弱对宏观结构的刻画但其在中低复杂度任务中仍具效率优势.

当扩散步数为 5 时, 扩散策略虽能捕捉宏观多模结构, 但可能生成相对数据支持发生有限偏移、仍位于邻近区域的动作. 本文将其称为“半分布外动作 (Semi-OOD)”, 用于描述扩散策略在连续动作空间中的典型生成特征. Semi-OOD 并非是显式识别类别, 而是用于解释离线强化学习中潜在的分布偏移风险与收益. 传统扩散策略未对该类动作加以处理, 在数据稀缺或长程规划任务中可能加剧偏移. EDPQ 引入基于均值-随机微分方程 (SDE) 的扩散建模方式, 并结合集成 Q 网络构造 Q 值的 LCB 筛除不可靠动作. 该方法在复杂稀疏奖励任务中取得了显著性能提升, 但其效果在很大程度上依赖于集成 Q 网络的数量. IDQL[29] 将 IQL 的隐式 Q 约束机制嵌入扩散策略框架, 利用优势加权的行为克隆抑制低价值的 Semi-OOD 动作, 但其算法开销并没明显降低. EDPQ 和 IDQL 的实验结果表明, 在数据稀缺或长程规划场景下, 通过降低动作价值的的不确定性或优势加权等机制, 对扩散策略生成的动作进行合理筛选, 有助于提升扩散策略在复杂任务中的整体表现.

为此, 本文提出一种基于 Q 集成下置信域引导的高效扩散策略 (Efficient Q-Ensemble Diffusion Policy for Offline Reinforcement Learning, E2DP), 该方法以理论推导结果为基础, 设计通过两步推理得到动作, 既能保持传统扩散策略在复杂动作分布中的宏观多模建模能力, 又能显著压缩推理时间; 引入基于随机权重系数的方差正则化作为不确定性惩罚, 增加集成 Q 网络的多样性, 并基于此构造更可靠的 LCB 值, 用于优化策略更新和引导动作选择. 从而降低对大规模集成的依赖. 实验结果验证了所提的 E2DP 方法仅需 4 个 Q 网络, 即可在分布内动作与数据支持附近的潜在高价值候选之间形成有效权衡, 在性能和计算效率之间实现更优平衡. Bandit 可视化实验验证了两步推理对多模高斯结构的建模能力. 在 D4RL 基准任务上, E2DP 在保持与主流扩散策略相当甚至更优的性能水平下, 显著降低了计算开销.

2 基于 Q 网络集成下置信域引导的高效扩散策略

本节介绍我们提出的基于 Q 网络集成下置信域引导的高效扩散策略 (E2DP). 该方法在降低扩散策略推理开销的同时, 通过更有效的价值引导机制, 对高效推理条件下生成的 Semi-OOD 动作进行合理约束与利用, 从而提升扩散策略在稀疏奖励、长程规划等复杂任务中的整体表现.

2.1 两步逆向动作推理

现有扩散策略通过对原始动作数据执行一个马尔科夫加噪过程, 逐步将动作数据变换为服从标准高斯分布$a_T \sim \mathcal{N}(0, I)$, 该过程被称为前向扩散过程的分布; 然后从标准高斯分布中随机采样噪声动作数据, 通过条件逆向扩散链式来推理得到动作 (称为推理过程). 前向扩散过程可被形式化地描述为:

$$q(a_{1:T}|a_0) := \prod_{t=1}^{T} q(a_t|a_{t-1}),$$
$$q(a_t|a_{t-1}) = N(\sqrt{1-\beta_t}a_{t-1}, \beta_t I) \tag{3}$$

$t \in \{1, \dots T\}$表示扩散过程时间步, a_t表示扩散过程中带噪声的动作, $a_0 \sim p(a_0)$, β_t表示预定义的参数, 其值随扩散步数t变化[30].

推理过程则通过最大化原始数据的对数似然的证据下界 (ELBO), 推导出形如 (5) 式的逆扩散链, 利用重参数化技巧, 从$t = T$到$t = 0$逐步推理, 能够近似还原出原始数据的分布. ELBO 的形式化表达如式 (4), 详细推导参考文献 [30].

$$\begin{aligned}\text{ELBO} = \mathrm{E}_{q(a_{1:\,T}|a_0)}[\log\frac{p(a_{0:\,T})}{q(a_{1:\,T}|a_0)}] = \\ \mathrm{E}_{q(a_1|a_0)}[\log p_\theta(a_0|a_1)] - \\ D_{KL}(q(a_T|a_0)\parallel p(a_T)) - \\ \sum_{t=2}^{T} E_{q(a_t|a_0)}[D_{KL}(q(a_{t-1}|a_t,a_0)\parallel p_\theta(a_{t-1}|a_t))],\end{aligned} \tag{4}$$

$$p_\theta(a_{0:\,T}) = N(a_T;0,I)\prod_{t=1}^{N} p_\theta(a_{t-1}\mid a_t). \tag{5}$$

扩散策略[10] 从标准高斯分布中随机采样噪声动作数据, $a_T \sim N(0,I)$, 通过条件逆向扩散链式 (6) 来推理得到动作.

$$\begin{aligned}\pi_\theta(a\mid s) = p_\theta(a_{0:\,T}\mid s) = \\ N(a_T;0,I)\prod_{t=1}^{N} p_\theta(a_{t-1}\mid a_t,s).\end{aligned} \tag{6}$$

其中 $p_\theta(a_{t-1}\mid a_t,s)\sim N(a_{t-1};u_\theta(a_t,s,t),\sum_\theta(a_t,s,t))$, 方差为$\sum_\theta(a_t,s,t)=\beta_t\mathrm{I}$, 均值部分通过 (3) 式推导获得.

现有扩散策略的推理过程分析表明, 虽然式 (5) 在理论上展现出优越的分布重构精度, 但其推理复杂度却与扩散步数T呈线性关系, 即推理时间随扩散步数T增长而增加, 当T过大时, 尽管能够实现较高精度的动作分布恢复, 但也可能伴随不必要的噪声引入、泛化能力下降、计算开销显著增加, 甚至影响有效的学习与探索; 而目前的扩散策略在T取较小的步数时, 会陷入与 BCQ(使用 CVAE) 一样的模式平均困境. 因此, 相较于图像生成任务中对像素级逐步重构的要求, 离线强化学习更应关注动作分布的宏观多模结构建模能力.

基于上述观察, 本文提出一种两步逆向推理机制, 将原有多步扩散推理过程压缩为两步生成. 该机制在显著降低计算开销的同时, 保留了动作分布的关键结构特征, 使扩散策略在较小扩散步数设置下仍能有效刻画复杂多模的动作分布, 从而在效率与表达能力之间实现更优平衡.

扩散模型通常通过最大化证据下界 (Evidence Lower Bound, ELBO) 逼近原始数据分布, 但在离线强化学习场景中, 多步逆向扩散带来的推理开销成为实际应用瓶颈. 受 DDIM[31] 与 Consistency Model [20] 等工作的启发, 本文在保持动作整体分布一致性的前提下, 直接刻画关键时间步之间的映射关系. 为此, 我们对扩散模型的 ELBO 形式进行了重新分析, 以构建更高效的动作生成过程.

式 (4) 中, KL 散度项用以控制动作分布逆向推理过程的分布以匹配真实的去噪分布, 通过对公式 (4) 推导过程的分析可以发现, KL 散度项仅依赖于前向扩散过程的马尔可夫性假设. 这意味着在保持这一设定的情况下, 我们构造逆向推理过程时, 即便不遵循逐步的去噪匹配过程, a_0依然可以作为条件使用. 通过弱化T逐步到 2 中间部分的 KL 项, 我们构造出从a_T到a_1的条件分布$q(a_1|a_T,a_0)$, 从该分布中重参数化得到a_1, 此时 (4) 式可得到以下近似:

$$\begin{aligned}\text{ELBO} \approx E_{q(\boldsymbol{a_1}|\boldsymbol{a_0})}[\log p_{\boldsymbol{\theta}}(a_0|a_1)] - \\ D_{\mathrm{KL}}(q(a_{_T}|a_0))\parallel p(a_T)) - \\ \mathrm{E}_{q(a_T|a_0)}[D_{KL}(q(a_1|a_T,a_0)||p_\theta(a_1|a_T))].\end{aligned} \tag{7}$$

条件分布$q(a_1|a_t,a_0)$, 根据贝叶斯规则, 有$q(a_1|a_t,a_0)=\dfrac{q(a_t|a_1,a_0)q(a_1|a_0)}{q(a_t|a_0)}$, 根据 (3) 式, 我们推导出$q(a_1|a_t,a_0)$服从如下的高斯分布:

$$\begin{aligned}q(a_1|a_t,a_0)=\mathcal{N}(a_1;\sigma_{1|t}^2(\frac{\sqrt{\overline{\alpha}_{1,t}}}{1-\overline{\alpha}_{1,t}}a_t+\frac{\sqrt{\overline{\alpha}_1}}{\beta_1}a_0), \\ \sigma_{1|t}^2 I),\sigma_{1|t}^2=(\frac{\overline{\alpha}_{1,t}}{1-\overline{\alpha}_{1,t}}+\frac{1}{1-\overline{\alpha}})^{-1}.\end{aligned} \tag{8}$$

采用重参数化的技巧得:

$$\begin{aligned}a_1=\sigma_{1|t}^2(\frac{\sqrt{\overline{\alpha}_{1,t}}}{1-\overline{\alpha}_{1,t}}a_t+\frac{\sqrt{\alpha_1}}{\beta_1}f_\phi(a_t,t;s))+ \\ \sqrt{\sigma_{1|t}^2}I.\end{aligned} \tag{9}$$

其中$\overline{a}_{1,t}=\prod_{i=2}^{t}a_i$, a_0用$f_\phi(a_t,t;s)$构建, $f_\phi(a_t,t;s)$目标函数为:

$$\begin{aligned}L_{\text{diff}}(\phi)=E_{t\sim u,\epsilon\sim\mathcal{N}(0,\boldsymbol{I}),(s,a_0)\sim D} \\ [||a_0-f_{\boldsymbol{\phi}}(\sqrt{\overline{\alpha}_t}a_0+\sqrt{1-\overline{\alpha}_t}\epsilon,t,s)||^2].\end{aligned} \tag{10}$$

在获得中间含噪态a_1后, 我们需对 ELBO 中其余项进行优化, 以避免因近似引入的偏差使模型表达能力发生过大退化. 对于 (11) 式中第二项, 通过控制前向扩散的噪声调度参数 (α_t,β_t) 依照式 (3) 前向扩散式, 保证从a_0到a_T的前向扩散过程仍服从预设的高斯分布. 最后对第一项$E_{q(\boldsymbol{a_1}|\boldsymbol{a_0})}[\log p_{\boldsymbol{\theta}}(a_0|a_1)]$的优化, 和 DDPM 中做法相同, 采用参数化近似的方式将a_1与状态s输入$f_\phi(a_t,t;s)$, 得到最终动作a_0.

$$a_0=f_\phi(a_1,1;s). \tag{11}$$

在上述评估设置下, E2DP 相比标准扩散策略在理论层面平均推理时间降低显著, 对应推理速度提升约 2.5 倍. 式 (12) 的具体理论推导见附录 A.

2.2 基于集成 Q 网络的下置信域引导

在扩散策略采用小步数或高效近似推理的设置下, 生成动作整体仍位于离线数据分布的高密度区域, 但部分动作具有较高不确定性, 即 Semi-OOD 动

作. 由于该区域缺乏充分数据支撑, 其价值估计主要依赖 Q 网络外推结果. 在 Actor–Critic 框架中, 该外推误差易通过 TD 自举更新被放大, 进而引发价值高估并加剧策略分布偏移风险.

为在抑制高估的同时保留有效探索能力, 本文采用基于集成 Q 网络的 LCB 方法对动作价值进行评估. 相较于直接施加保守惩罚, LCB 利用集成成员间的估计差异刻画价值不确定性, 为策略更新提供更可靠的引导依据. 具体设计如下:

本文对每个 Q 网络采用随机参数初始化, 并为其配置独立的目标 Q 网络 (Target Q). 确保集成成员对同一动作的价值估计在初始化阶段存在合理差异, 可一定程度缓解同质化的集成导致的乐观估计偏差[25], 进一步地, 为削弱 OOD 动作与高 Q 值的虚假相关性[26], 并进一步提升集成 Q 网络的多样性, 本文在 Q 网络训练阶段引入基于集成方差的不确定性惩罚作为正则化项, 其损失函数定义如式 (17) 所示. 该正则项作用于不同 Q 网络的 TD 更新阶段, 不确定性惩罚超参数采用每次更新时从均匀分布区间 [0,1] 内随机采样的值. 该设计可在小规模 Q 集成条件下持续引入更新差异, 避免各 Q 网络过早收敛到高度相似的解, 从而提升不确定性估计的有效性.

$$\begin{aligned} J_Q(\psi^i) &= \mathrm{E}_{\boldsymbol{s}_i,\boldsymbol{a}_i,r_i,\boldsymbol{s}_{i+1}\sim D}[Q_{\psi^m}(s_i,a_i)- \\ &\quad y^m(r_i,s_{i+1},\pi_{\boldsymbol{\phi}})] \\ y^m &= r_i+\gamma \mathrm{E}_{\boldsymbol{a}_{i+1}\sim\pi_{\boldsymbol{\phi}}} Q_{\overline{\psi}^m}(s_{i+1},a_{i+1})- \\ &\quad \kappa\sqrt{\mathrm{V}_{\mathbf{ens}}[Q_{\boldsymbol{\psi}^m}(s,a)]}. \end{aligned} \tag{12}$$

其中, ψ^m和$\overline{\psi}^m$分别为第 m 个 Q 网络及其目标 Q 网络的参数, κ控制悲观惩罚强度, $\mathrm{V}_{\mathbf{ens}}[Q_{\boldsymbol{\psi}^m}]$表示方差. 进一步利用集成 Q 网络的均值与标准差构造 Q 的下置信域估计, 并将其用于策略改进与动作选择, 其形式如式 (14) 所示.

$$Q_{\psi}^{\mathrm{LCB}} = E_{ens}[Q_{\psi^m}(s,a)] - \xi\sqrt{\mathrm{V}_{ens}[Q_{\psi^M}(s,a)]}. \tag{13}$$

其中均值项$E_{ens}[Q_{\psi^m}(s,a)]$反映集成网络对动作价值的整体评估, 标准差项量化估计的不确定性幅度, 置信系数$\xi \geqslant 0$. 最终, 结合式 (17) 构造 E2DP 的整体损失:

$$L_\pi(\phi) = L_{\mathrm{diff}}(\phi) - \lambda \mathrm{E}_{s\sim D, a_0\sim\pi_\phi} Q_{\psi}^{\mathrm{LCB}}. \tag{14}$$

其中, $\lambda = \eta/\mathrm{E}_{(s,a)\sim\mathcal{D}}[|Q_{\psi}^{\mathrm{LCB}}|]$对$Q_{\psi}^{\mathrm{LCB}}$进行归一化用于平衡损失项的尺度.

通过以上设计, 集成 Q 网络的多样性差异被方差转化为可量化的不确定性估计, 通过该集成的 LCB 值引导策略更新与策略动作选择, 即可在分布内动作与数据支持附近的潜在高价值候选之间形成有效权衡. 具体 E2DP 的算法流程如算法 1 所示.

算法 1 E2DP

输入: 拉回矩阵F, 软更新系数τ, 扩散时间步T, 置信度参数β, 不确定性惩罚κ

初始化: $\pi_\phi, \pi_{\tilde{\phi}}, \{Q_{\psi^m}, Q_{\tilde{\psi}^m}\}_{m=1}^M$; 初始化经验参数

for $i = 1,\ldots,N$ **do**

采样小批量轨迹数据$\{(s,a,r,s')\} \sim \mathcal{D}_{init}$

#集成Q网络训练

根据(10)及(12) Actor输出动作$a_0' \leftarrow \pi_\phi(|s')$

根据(13)更新集成Critic网络参数$\{\psi^m\}_{m=1}^M$

#策略网络训练

从$1,\ldots,M$中采样t, 根据(3)式, 将a_0加到噪到a_t, 根据(11)训练$f_\phi(a_t,t;s)$

采样$a_t^* \sim \mathcal{N}(0,I)$

根据(10)(12) Actor输出动作$a_0' \leftarrow \pi_\phi(|s')$, 根据(14)式计算$Q^{LCB}$, 并通过(15)更新扩散策略的参数$\phi$

#更新目标网络

$\tilde{\phi} \leftarrow \tau\phi + (1-\tau)\tilde{\phi}$

$\tilde{\psi}^m \leftarrow \tau\psi^m + (1-\tau)\tilde{\psi}^m, \quad m \in \{1,\ldots,M\}$

end

3 实 验

3.1 实验初始化

实验基于 D4RL 基准中的 Gym-MuJoCo 连续控制任务与 AntMaze 稀疏奖励导航任务开展. 均在 Ubuntu20.04 系统、NVIDIA RTX 3090 GPU 的服务器平台上完成. 实验所用超参数设置详见表 1.

表1 超参数设置

超参数	取值
网络隐藏层	2
网络隐藏层维度	256
优化器	Adam
激活函数	Mish
学习率	3e-4
不确定性惩罚κ	[0,0,5]
置信度超参数ξ	2
训练总迭代数	1e6
批次大小(Batch-size)	256
折扣率γ	0.99
软更新系数τ	0.005
扩散时间步T	10

Gym-MuJoCo 任务是离线强化学习中常用的连续控制评估基准, 基于物理仿真环境构建, 奖励函数平滑、动力学结构相对简单, 适用于评估算法在连续

控制场景下的学习能力与泛化性能. 本文选取 HalfCheetah、Hopper 和 Walker2d 三个典型任务进行实验, 并在每个任务上分别测试三种不同质量的数据集: medium、medium-replay(m-r) 和 medium-expert (m-e). 其中, medium 数据集由中等性能 SAC 策略训练过程中的重放缓冲区静态采样 100 万步构成; m-r 数据集从 SAC 由随机初始化至收敛全过程的重放缓冲区采样 100 万步, 行为分布多样性更高; m-e 数据集由 medium 数据与专家数据融合后重采样至 100 万步构成, 包含较高比例的近最优轨迹. 不同数据质量条件下的测试有助于评估算法的稳定性与鲁棒性.

AntMaze 任务为典型稀疏奖励导航基准, 仅在智能体到达目标位置时给予终止奖励, 其余时间步奖励为零, 对长时序决策能力要求较高. 实验采用 D4RL 中的六种 AntMaze 数据集, 按迷宫规模分为 umaze(u)、medium(m) 和 large(l) 三类, 环境复杂度依次递增; 按轨迹分布特性分为 play(p) 与 diverse(d) 两类. play 数据集包含从多个固定起点到固定目标的轨迹, 分布相对集中; diverse 数据集则包含随机起点与随机目标组合, 状态–动作空间覆盖更广. 由于奖励稀疏且轨迹覆盖受限, 该任务能够有效检验算法在复杂离线场景下的决策能力与稳定性.

为便于直观对比不同算法在各任务上的性能, 所有任务得分均归一化至 [0, 100], 计算方式为:

$$\text{归一化得分} = 100\times \frac{(\text{平均回报} - \text{参考随机回报})}{(\text{参考专家回报} - \text{参考随机回报})}. \tag{15}$$

其中参考专家/随机回报由 D4RL 基准任务的官方文档定义, 见表 2.

表2　D4RL 基准测试任务参考回报说明

任务	参考专家回报	参考随机回报
Walker2d	4 592.3	1.63
Hopper	3 234.3	−20.2
Halfcheetah	12 135.0	−280.1
Antmaze	1.0	0.0

3.2　bandit 任务上的实验结果与分析

为分析不同策略在多模动作分布建模及价值引导条件下的动作选择特性, 本文在二维连续动作 bandit 场景中构造可视化实验. 该实验固定状态 s, 仅比较不同方法在相同条件下的动作输出分布. 离线动作数据由四个等权高斯分布混合采样得到, 形成典型的多模结构 (图 1(a)).

图 1 上半部分展示了不同方法在仅进行行为克隆时对离线动作分布的拟合结果. BC-MLE 由于单峰近似难以刻画多模结构; 扩散 BC 在较小扩散步数 T 下分布刻画较粗糙, 随 T 增大可逐步恢复多模形态; 该现象源于有限推理步数带来的近似误差, 而非训练未收敛

图 1 下半部分在上述动作空间中引入回报信息, 比较不同方法在价值引导下的策略输出分布. 可以观察到, TD3+BC 的策略输出较为分散; Diffusion-QL 在 $T = 2$ 时能够一定程度上稳定选取最优模态但还是受分布建模影响, 部分值选取趋于各模态的中间值, $T = 5$ 仍对部分次优模态分配一定选取概率; 即使在 $T = 50$ 的情况下, 该现象仍然存在. 相比之下, E2DP 在仅采用两步推理的条件下, 通过集成 Q 网络的下置信界 (LCB) 引导, 使策略输出稳定集中于最高回报模态, 并显著减少对次优动作的误选.

实验表明, 在离线强化学习场景中, 更精细的动作分布拟合并不必然带来更优的策略决策, 而结合价值引导对候选动作进行筛选对于提升决策稳定性更为关键

3.3　D4RL 基准测试结果与分析

为比较不同扩散策略的计算效率, 本文在 AntMaze-medium-v0 环境上, 对 Diffusion-QL、EDPQ 与 E2DP 的训练与推理时间进行评估. 训练时间以每 epoch(1 000 次更新) 的平均 wall-clock 时间统计; 推理时间通过测试阶段 100 个 episode 的平均执行时间衡量. 由于不同方法在相同环境与评估协议下运行, 环境交互带来的时间开销基本一致, 因此评估时间的差异主要反映了策略动作生成 (即推理阶段) 的计算复杂度差异.

表 3 的结果表明, , 在相同扩散步数下, EDPQ 由于采用 SDE 形式建模, 单步计算结构较离散反向扩散更为紧凑, 因此推理时间略低于 Diffusion-QL. E2DP 通过减少推理步数并优化推理结构, 在训练与推理阶段均显著降低计算开销, 实现更优的性能–效率权衡. 此外, 我们在 E2DP 中对比了在不用集成 Q 网络, 仅用与 Diffusion-QL 一样的双 Q 网络结构时的训练时间与评估时间. 结果表明, 使用集成 Q 网络为 4 时, 训练时间开销较之提升了 4.5%, 评估时间开销提升了 1.1%, 这表明, E2DP 只需较少数量的 Q 网络集成就能获得明显的性能提升, 但并不会带来较大的计算开销.

为了更全面评估 E2DP 的性能, 我们将其与离线强化学习领域的多种代表性方法进行对比, 包括经典方法 (BC、BCQ[9]、TD3+BC[12]、IQL[14]、CQL[[13]、

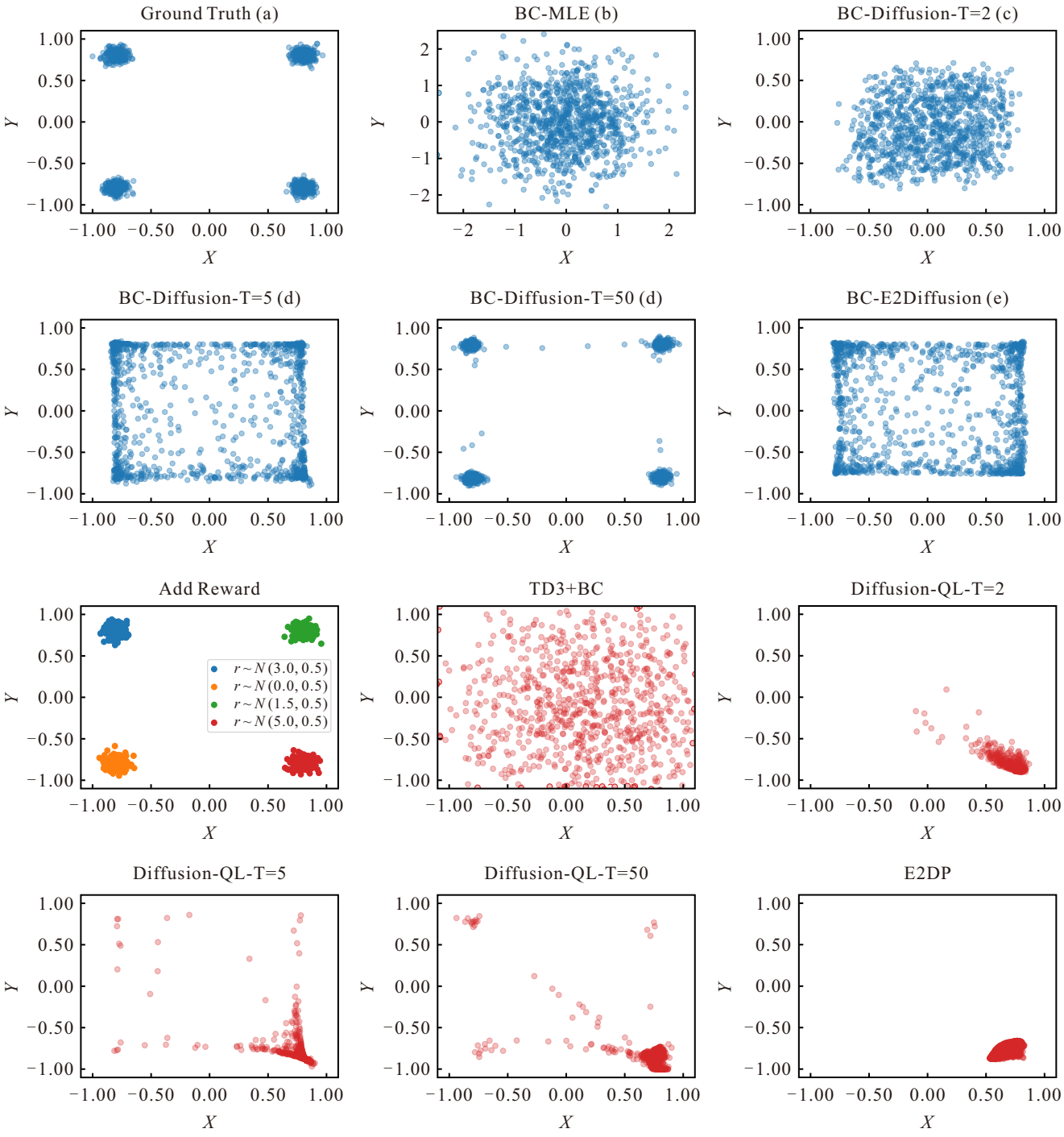

图1 bandit 任务可视化结果

表3 扩散策略类算法训练开销和评估开销

算法	扩散时间步T	训练时间	评估时间
Diffusion-QL	5	16.94s	3.61s
EDPQ	5	16.56s	3.06s
E2DP	10	12.04s	1.82s
E2DP W/O Q-集成	10	11.52s	1.80s

MSG[25])、扩散策略类方法 (Diffusion-QL[10]、EDP[28]、ACDP[32])、扩散模型预训练与蒸馏的方法 DTQL[33]、以及基于一步生成模型 Consistency model 的方法 CPIQL[34].

大部分方法的实验结果均严格取自其原始论文在 D4RL 基准上的报告值 (部分方法参考 IDQL[29] 论文中报告的最终结果). 实验过程中 E2DP 采用 5 个随机种子独立运行, 最终结果取最后 10 步评估的均值, 具体数据计入表 4 和表 5, 图 2 展示了 E2DP 与经典扩散策略方法 Diffusion-QL、采用集成 Q 技术的扩散策略方法 EDPQ 在 Gym-Mujoco 任务 medium 质量数据集上的得分曲线对比, 综合表 4-5 与图 2 的结果, 分析如下:

1) 在 Gym-MuJoCo 运动任务中, 大多数方法在数据质量较高 medium-expert 和 medium-replay 数据集上表现良好, 但在包含大量次优轨迹的 medium

表4　Gym-MuJoCo 任务的平均归一化得分对比

Gym Tasks	BC	TD3+BC	CQL	IQL	Diffusion-QL	EDP	CPIQL	IDQL	DTQ	ACDP	E2DP
walker2d-m-v2	75.3	83.7	72.5	78.3	87.0	86.9	88.4	82.5	**89.4**	75.86	88.7± 1.3
walker2d-m-r-v2	26.0	81.8	77.2	73.9	**95.5**	94.9	95.0	85.1	88.5	93.18	92.7 ± 1.3
walker2d-m-e-v2	107.5	110.1	108.8	109.6	110.1 3	110.2	112.3	**112.7**	110	110.31	111.5 ± 0.2
hopper-m-v2	52.9	59.3	58.5	66.3	90.5	81.9	**101.5**	65.4	99.6	99.39	100.7±1.5
hopper-m-r-v2	18.1	60.9	95.0	94.7	101.3	101.0	**101.7**	92.1	100.0	100.78	101.3 ± 0.3
hopper-m-e-v2	52.5	98.0	105.4	91.5	111.1	97.4	**113.4**	108.6	109.3	101.02	112.3 ± 0.6
halfcheetah-m-v2	42.6	48.3	44.0	47.4	51.1	52.1	55.3	51.0	57.9	53.8	**58.5**±0.7
halfcheetah-m-r-v2	36.6	44.6	45.5	44.2	47.8	49.4	49.8	45.9	50.9	48.68	**53.4**±1.4
halfcheetah-m-e-v2	55.2	90.7	91.6	86.7	**96.8**	95.5	90.2	95.9	92.7	96.59	95.5 ± 0.5
Average	51.9	75.3	77.6	77.0	88.0	85.4	89.7	82.13	88.7	86.6	**90.5**

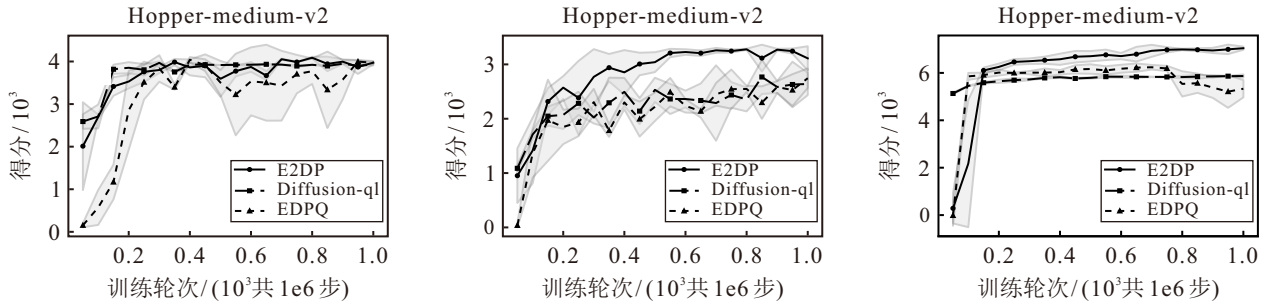

图2　扩散策略类方法在 Gym-MuJoCo 中 medium 质量数据集任务上得分曲线

数据集上性能明显下降. 扩散策略类方法 (如 Diffusion-QL 和 ACDP) 由于能够建模动作的多模分布, 因此在复杂分布场景下表现相对优异. E2DP 在所有 medium 任务中均取得最佳性能 (平均 82.6), 显著优于 Diffusion-QL(76.2). 这一优势主要源于两步推理机制在建模复杂动作分布时能够保留宏观形态, 同时结合集成 Q 网络的下置信域 (LCB) 估计, 引导扩散策略在分布内与 Semi-OOD 动作之间实现价值的权衡, 从而生成更优的动作. EPQE 增加了额外的熵探索机制, 虽然在某些任务上表现优异, 但由于扩散策略的特性, 容易造成过度探索影响策略性能.

ACDP 通过优势约束聚焦高回报轨迹以提升性能, 但该机制较依赖优势函数的超参数设定, 相比之下, E2DP 通过基于不确定性估计的动态价值引导, 在探索 Semi-OOD 动作与利用高价值分布内动作之间实现更合理的平衡, 尤其在高质量数据稀缺的 medium 数据集上优势更为显著.

2) 在 AntMaze 任务中, BC 在 umaze 上尚可获得部分得分, 但在 medium 和 large 场景中性能降至近零, 表明纯模仿方法难以应对长程稀疏奖励下的路径拼接问题. E2DP 在六项任务中的四项上取得最佳结果 (平均 83.78), 尤其在最具挑战性的 antmaze-large-diverse-v0 任务中, 其平均归一化得分达 74.4, 显著优于 Diffusion-QL(56.6) 和 MSG(71.4). Diffusion-QL 虽依托扩散模型的生成特性具备建模宏观动作分布的潜力, 理论上可通过该能力弥补 AntMaze 数据集的覆盖缺陷, 但因未结合有效的高价值动作筛选机制, 未能将宏观分布建模的优势转化为实际性

表5　AntMaze 平均归一化得分对比

AntMaze Tasks	BC	TD3+BC	CQL	IQL	MSG	Diffusion-QL	EDP	IDQL	ACDP	DTQ	EDPQ	E2DP
antmze-u-v0	54.6	78.6	74	87.5	97.8	93.4	96.6	94.0	97.28	94.8	99.0	**99±0.70**
antmaze-u-d-v0	45.6	71.6	84.0	62.2	81.8	66.2	69.5	80.2	**87**	78.8	67.5	66.5±3.7
antmaze-m-p-v0	0.0	10.6	61.2	71.2	89.6	76.6	0.0	84.5	**89.64**	79.6	84.0	89.6±3.04
antmaze-m-d-v0	0.0	3.0	53.7	70.0	88.6	78.6	6.4	84.8	86.80	82.2	85.4	**89.8±4.08**
antmaze-l-p-v0	0.0	0.2	15.8	39.6	72.6	46.4	1.6	63.5	46.28	52.0	72.6	**83.4±7.92**
antmaze-l-d-v0	0.0	0.0	14.9	47.5	71.4	56.6	4.4	67.9	35.28	54.0	65.9	**74.4±7.09**
Average	16.7	27.3	50.6	63.0	83.6	69.6	29.75	79.1	73.71	73.6	79.0	**83.78**

能提升, 反而因无效探索影响了长程路径拼接效果. MSG 通过集成技术增强了价值估计的可靠性, 在稀疏奖励场景中展现出一定有效性, 但受限于 AntMaze 数据集本身的次优特性 (多数轨迹未达最优路径), 其性能提升被数据质量瓶颈所约束, 难以突破现有数据分布的局限. E2DP 通过两步推理与基于集成 Q 网络 LCB 的价值引导, 在探索 Semi-OOD 动作与利用分布内高价值动作之间实现有效权衡, 从而有效突破数据质量限制, 最终实现“由次优轨迹拼接近最优路径”的能力, 这与 AntMaze 任务对算法的关键要求高度契合.

3.4 消融实验

3.4.1 集成 Q 下置信域引导对于扩散策略的性能影响

集成 Q 通过多网络评估提升价值估计置信度. 不同于依赖 64 个 Q 网络构造 LCB 引导的 EDPQ, E2DP 在 Q 更新中引入基于随机权重的方差正则项作为不确定性惩罚, 增强 Q 网络在高不确定区域的差异, 从而提升集成多样性, 使得仅使用 4 个 Q 网络即可构造可靠的 LCB 估计. 此外, 我们还将 Diffusion-QL 与本文所设计集成 Q 下置信域引导结合, 显著提升 Diffusion-QL 在复杂稀疏任务 antmaze-large-play-v0 的性能, 验证了所设计集成 Q 方法对扩散策略的增益.

图 3 表明, 与 EDPQ 相比, E2DP 在更小集成规模下取得更高性能且波动更小. 同时, 我们还设计了方差正则化的消融, 移除方差正则项后, E2DP 中所用的集成 Q 的不确定性刻画能力下降, 导致 LCB 引导效果减弱, 性能与稳定性均明显下降.

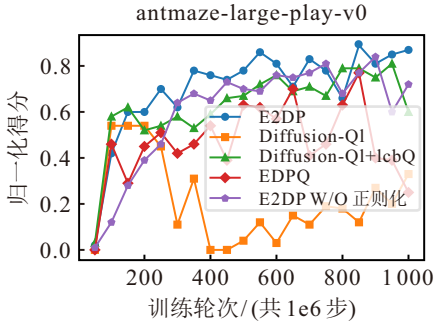

图3 集成 Q 下置信域引导对于扩散策略的性能影响

最后, 将本文所设计集成 Q 下置信域引导应用于 Diffusion-QL, 在 antmaze-large-play-v0 上显著提升其表现, 验证了方法的有效性与可迁移性.

3.4.2 集成 Q 网络数量的依赖

为分析 E2DP 对集成 Q 网络数量的依赖性, 我们在 antmaze-large-play-v0 环境中对比了不同 Q 网络规模 (4、16、32、64) 下的算法表现, 实验结果如图 4 所示.

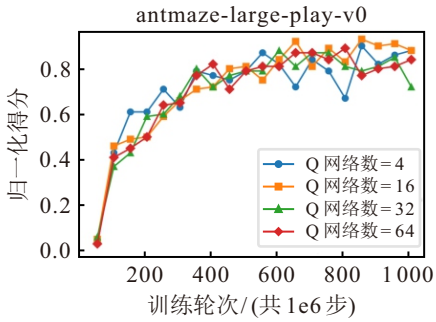

图4 集成网络数量对 E2DP 算法的影响

实验结果表明, 随着 Q 网络数量的增加, 策略训练过程整体趋于更加平稳, 更大的集成规模有助于降低价值估计的随机波动. 然而, 当集成规模超过一定数量后, 性能提升幅度趋于饱和, 归一化得分未呈现显著增长趋势. 仅使用 4 个 Q 网络的情况下, E2DP 已能够获得与更大规模集成相当的平均性能与稳定性.

该结果与前一小节的分析一致: 在引入随机权重的方差正则化以增强集成 Q 的多样性的前提下, 小规模 Q 集成已能够提供具有判别力的不确定性估计, 从而支撑稳定有效的 LCB 构造. 综合考虑性能表现与计算开销, 本文在实验中统一采用 4 个 Q 网络作为默认集成规模. 消融实验详见附录 B.

4 结 论

本文提出了一种面向离线强化学习的高效集成扩散策略 (E2DP). 我们设计了两步逆向推理机制, 使得 E2DP 在保留扩散策略对原始动作分布宏观结构建模能力的同时, 显著降低了推理计算开销. Bandit 实验结果验证, E2DP 仅需两步推理便能达到标准扩散模型在复杂多模动作分布上的表达效果.

在 D4RL 基准测试中的实验结果进一步验证了 E2DP 在处理复杂动作空间时的高效建模优势. 此外, 通过将集成 Q 技术与扩散策略深度结合, E2DP 在 Q 网络训练中引入不确定性惩罚, 增强集成 Q 网络的多样性, 使得集成 Q 的 LCB 值具备更低不确定性, 基于此引导策略的更新以及高价值动作的筛选. 实现了更为稳健的探索–利用权衡. 在仅使用 4 个 Q 网络的条件下, E2DP 性能优于依赖 64 个 Q 网络进行 LCB 构造的 EDPQ 方法, 体现了较高的效率优势. 然而, E2DP 仍然存在一定局限性, 两步推理的近似机制可能削弱对分布细节的刻画能力. 这在一些场景中可能受限, 未来工作将进一步平衡推理效率与精细建模能力, 以适应更复杂的应用场景.

参考文献 (References)

[1] 户高铭, 蔡克卫, 王芳, 等. 基于深度强化学习的无地图移动机器人导航[J]. 控制与决策, 2024, 39(3): 985-993.
(Hu G M, Cai K W, Wang F, et al. Mapless navigation based on deep reinforcement learning for mobile robots[J]. Control and Decision, 2024, 39(3): 985-993.)

[2] 孙辉辉, 胡春鹤, 张军国. 基于主动风险防御机制的多机器人强化学习协同对抗策略[J]. 控制与决策, 2023, 38(5): 1420-1429.
(Sun H H, Hu C H, Zhang J G. Cooperative countermeasure strategy based on active risk defense multi-agent reinforcement learning[J]. Control and Decision, 2023, 38(5): 1420-1429.)

[3] 闫超, 相晓嘉, 徐昕, 等. 多智能体深度强化学习及其可扩展性与可迁移性研究综述[J]. 控制与决策, 2022, 37(12): 3083-3102.
(Yan C, Xiang X J, Xu X, et al. A survey on scalability and transferability of multi-agent deep reinforcement learning[J]. Control and Decision, 2022, 37(12): 3083-3102.)

[4] Wang X S, Zhang J Z, Hou D Y, et al. Autonomous driving based on approximate safe action[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(12): 14320-14328.

[5] 王云泽, 孙宇, 骆中斌, 等. 基于深度强化学习的自动驾驶行为决策研究综述[J]. 控制与决策, 2026, 41(2): 305-328.
(Wang Y Z, Sun Y, Luo Z B, et al. Review of autonomous driving behavior decision-making based on deep reinforcement learning[J]. Control and Decision, 2026, 41(2): 305-328.)

[6] 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.)

[7] Levine S, Kumar A, Tucker G, et al. Offline reinforcement learning: Tutorial, review, and perspectives on open problems[J/OL]. 2020, arXiv: 2005.01643. https://arxiv.org/abs/2005.01643.

[8] 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.

[9] Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration[C]. International Conference on Machine Learning. Piscataway: IEEE, 2019: 2052-2062.

[10] Wang Z D, Hunt J J, Zhou M Y. Diffusion policies as an expressive policy class for offline reinforcement learning[J/OL]. 2022, arXiv: 2208.06193.

[11] Kumar A, Fu J, Soh M, et al. Stabilizing off-policy Q-learning via bootstrapping error reduction[J]. Advances in Neural Information Processing Systems, 2019, 32: 11761-11771.

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

[13] 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, 2020: 1179-1191.

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

[15] Haarnoja T, Zhou A, Hartikainen K, et al. Soft actor-critic algorithms and applications[J/OL]. 2018, arXiv: 1812.05905.

[16] Fujimoto S, van Hoof H, Meger D. Addressing function approximation error in actor-critic methods[C]. International Conference on Machine Learning. Piscataway: IEEE, 2018: 1587-1596.

[17] Kingma D P, Welling M. Auto-encoding variational Bayes[J/OL]. 2013, arXiv: 1312.6114.

[18] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[C]. Proceedings of Advances in Neural Information Processing Systems. Virtual, 2020: 6840-6851.

[19] Geng Z Y, Deng M Y, Bai X J, et al. Mean flows for one-step generative modeling[J/OL]. 2025, arXiv: 2505.13447.

[20] Song Y, Dhariwal P, Chen M, et al. Consistency models[C]. Proceedings of the 40th International Conference on Machine Learning. Honolulu, 2023: 32211-32252.

[21] Pomerleau D A. ALVINN: An autonomous land vehicle in a neural network[C]. Proceedings of the 2nd International Conference on Neural Information Processing Systems. Piscataway: IEEE, 1988: 305-313.

[22] Dinh L, Sohl-Dickstein J, Bengio S. Density estimation using real NVP[J/OL]. 2016, arXiv: 1605.08803.

[23] Agarwal R, Schuurmans D, Norouzi M. An optimistic perspective on offline reinforcement learning[J/OL]. 2019, arXiv: 1907.04543.

[24] An G, Moon S, Kim J H, et al. Uncertainty-based offline reinforcement learning with diversified Q-ensemble[C]. Proceedings of the 35th International Conference on Neural Information Processing Systems. New York, 2021: 7436-7447.

[25] Ghasemipour K, Gu S S, Nachum O. Why so pessimistic estimating uncertainties for offline RL through ensembles, and why their independence matters[C]. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, 2022: 18267-18281.

[26] 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.

[27] Zhang R Q, Luo Z W, Sjölund J, et al. Entropy-

regularized diffusion policy with *Q*-ensembles for offline reinforcement learning[C]. Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, 2024: 98871-98897.

[28] Kang B Y, Ma X, Du C, et al. Efficient diffusion policies for offline reinforcement learning[C]. Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, 2023: 67195-67212.

[29] Hansen-Estruch P, Kostrikov I, Janner M, et al. IDQL: Implicit *Q*-learning as an actor-critic method with diffusion policies[J/OL]. 2023, arXiv: 2304.10573.

[30] Luo C. Understanding diffusion models: A unified perspective[J/OL]. 2022, arXiv: 2208.11970.

[31] Song J M, Meng C L, Ermon S. Denoising diffusion implicit models[J/OL]. 2020, arXiv: 2010.02502.

[32] 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.)

[33] Chen T Y, Wang Z D, Zhou M Y. Diffusion policies creating a trust region for offline reinforcement learning[C]. Advances in Neural Information Processing Systems. Piscataway: IEEE, 2024: 50098-50125.

[34] Chen Y H, Li H R, Zhao D B. Boosting continuous control with consistency policy[J/OL]. 2023, arXiv: 2310.06343.

作者简介

张旭 (2000−), 男, 硕士生, 主要研究方向为强化学习, E-mail: 1183530127@qq.com;

曾玉婷 (2000−), 女, 硕士生, 主要研究方向为强化学习, E-mail: 2743913158@qq.com;

张峰 (1976−), 女, 博士, 教授, 博士生导师、主要研究方向为机器学习、深度学习、强化学习、智能决策等, E-mail: fengzhang@hbu.cn.