arXiv:2602.01260v1 [cs.LG] 1 Feb 2026

样本 离线强化学习的高效主动算法 g

Soumyadeep Roy [*1] **Shashwat Kushwaha** [*1] **Ambedkar Dukkipati** [1]

摘要

离线强化学习（RL）使得能够基于静态数据学习策略，但常常面临状态-动作空间覆盖不足和分布偏移问题。允许有限的在线交互，有选择地细化已学习价值函数中不确定的区域，即主动强化学习（ActiveRL），可以应对这一问题。尽管该方法在实证方面取得了良好成效，现有文献中却尚无理论分析。我们从高斯过程（GP）不确定性建模的视角出发，开展严谨的主动强化学习样本复杂度分析，以填补这一空白。为此，我们提出一种算法，并利用高斯过程集中不等式和信息增益界，推导出高概率保证，表明只需 ϵ 个主动转移即可学习到一个 $\mathcal{O}(1/\epsilon^2)$ 最优策略，优于纯离线方法的 $\Omega(1/\epsilon^2(1-\gamma)^4)$ 速率。我们的结果揭示，主动强化学习能够实现近乎最优的信息效率：有针对性的不确定性降低可以在仅使用少量在线数据的情况下，加速价值函数收敛。我们的分析基于高斯过程集中不等式和信息增益界，构建了贝叶斯非参数回归理论与强化学习理论之间的桥梁。我们开展了多项实验，以验证该算法和理论发现。

1. 引言

强化学习（RL）在序贯决策任务中取得了显著的实证成果。然而，这类成果通常依赖于与环境进行大量在线交互。在许多现实应用中，例如导航（Dukkipati 等，2022）、自动驾驶（Delavari 等，2025）、医疗保健（Yu 等，2021年）或机器人控制（Tang 等，2025），不受限制的探索代价高昂、风险大，甚至完全不可行。在非平稳环境中（Ayyagari 等，2025），这些问题可能更加棘手。离线强化学习提供了一种颇具吸引力的替代方案，即仅利用日志数据学习策略。然而，其有效性取决于离线数据集的多样性和覆盖范围。覆盖不足时，学习得到的策略在数据集支持集之外可能泛化不佳，从而导致任意程度的次优或不安全决策。数据效率与可靠泛化之间的这种权衡引出了一个根本问题： 我们能否仅通过少量精心挑选的在线交互，改进离线训练得到的策略？

近期的实证研究主动离线强化学习通过允许有限的在线交互预算，并将其有选择地分配到状态空间中的高不确定性区域（Dukkipati 等人，2025），对这一问题进行了深入探索。这些方法相较于纯离线方法，显著提升了数据效率。然而，我们对其理论的理解仍然有限。经典的强化学习样本复杂度结果——无论是在表格设定（Wang 等人，2023）、线性设定还是核化设置（Chowdhury & Gopalan，2017年；Yeh 等人，2023）中——都无法直接推广到离线与在线相结合的混合场景。此外，深度或参数化强化学习方法中的不确定性估计往往依赖启发式方法，缺乏将不确定性降低幅度与策略改进联系起来所需的统计校准。因此，关于需要多少主动交互才能达到 ϵ最优性这一根本问题仍未得到解答。

本研究旨在通过高斯过程（GP）的视角分析主动离线强化学习的样本复杂度，为其建立一套有原则且定量的理论。本文探讨以下问题——需要多少额外的主动转移，才能保证获得一个 ϵ——最优策略，其中认知不确定性通过价值函数上的高斯过程先验建模？

我们为基于高斯过程的主动强化学习推导了一个 PAC 风格的高概率次优性差距界。在标准的光滑性、利普希茨性和覆盖假设下，我们证明，不确定性引导的主动采样能够以 ϵ 最优性，仅需 $\mathcal{O}(1/\epsilon^2)$ 额外的状态转移，与基于高斯过程的老虎机优化已知的最优速率相匹配。我们的证明综合了：(i) 高斯过程集中不等式（Srinivas 等人，2012），(ii) 基于信息增益的方差缩减

[*]共同贡献 [1]印度科学研究所计算机科学与自动化系，印度班加罗尔。通信作者：Ambedkar Dukkipati <ambedkar@iisc.ac.in>。

预印本。*2026*年*2*月*3*日。

(iii) 贝尔曼算子的利普希茨收缩 tor.

贡献。 本研究的贡献如下。

(1) 我们使用高斯过程先验对价值估计中的认知不确定性进行建模，并分析其在引导主动探索中的作用。

(2) 我们证明， $\mathcal{O}(1/\epsilon^2)$ 次主动交互足以在标准光滑性假设下学习一个 ϵ-最优策略。**(3)**与纯离线强化学习相比，主动设定将 $(1-\gamma)^{-4}$ 的规模提升为 $(1-\gamma)^{-2}$，实现了二次改进。
(4) 我们提供了大量实验结果，以验证理论发现并展示所提方法的实用性。

2. 相关工作

离线强化学习旨在从固定数据集中学习策略，而无需进一步与环境交互（Levine 等人，2020年；Uehara & Sun，2021年）。早期算法引入了保守或悲观正则化，以避免在数据集支持集之外进行外推。然而，当数据覆盖稀疏时，这些方法往往表现不佳。为此，新兴的主动离线强化学习（Dukkipati 等人，2025）提出了混合方法，用数量有限的主动选择的在线交互补充离线数据。这些算法在实证中表明，通过针对高不确定性区域或关键决策边界进行采样，可以提升样本效率。然而，对其理论的理解仍然有限。现有分析通常依赖表格抽象或启发式不确定性度量。相比之下，我们的工作首次对 ActiveRL 进行了非参数、基于函数空间的分析，并利用高斯过程建立了 PAC 型样本复杂度保证。

高斯过程（GP）为量化认知不确定性提供了严谨的贝叶斯框架。它们已成功应用于价值函数近似和基于模型的强化学习，尤其是 GP-SARSA（Engel 等人，2005）及相关方法（Kuss& Rasmussen，2003）。高斯过程也被用于贝叶斯优化和赌博机设定，以推导遗憾界和置信界（Srinivas 等人，2009年）。然而，大多数既有的基于高斯过程的强化学习研究，要么聚焦于纯在线或完全离线学习，并未涉及混合设定，即智能体可以选择少量额外在线样本的情形。我们的分析通过将高斯过程集中不等式和信息增益论证应用于 ActiveRL 设定，拓展了这一理论基础。因此，我们量化了不确定性引导的探索如何降低样本复杂度。

强化学习中的经典结果在表格型设置中建立了多项式样本复杂度界（Kearns& Singh、2002年；Azar 等人，2017年）。后续工作将这些结果推广至使用线性模型、再生核希尔伯特空间（RKHS）和高斯过程的函数近似（Chowdhury & Gopalan，2017年）。这些分析通常假设进行充分的在线交互并能够访问生成模型，因此不适用于离线强化学习或部分交互机制。此外，尽管 GP-UCB 分析在监督学习或多臂老虎机场景中提供了紧致的信息论界（Srinivas 等人，2009年），但将其应用于强化学习需要控制贝尔曼误差传播和价值迭代稳定性。我们的工作通过结合基于高斯过程的信息增益与贝尔曼收缩性质，弥合了这些差距，建立了一个统一框架，将高斯过程理论与离线—在线混合强化学习的样本复杂度联系起来。

总之，先前的研究推动了离线强化学习算法、不确定性感知探索以及基于高斯过程的学习理论界的发展。然而，这些研究都没有针对混合式主动离线强化学习设定提供正式的样本复杂度分析。我们的工作通过构建基于高斯过程的理论框架填补了这一空白，该框架：(1) 在函数空间中显式建模不确定性；(2) 通过信息增益量化学习进展；(3) 推导出关于 ϵ–最优性的 PAC 式保证，所需主动样本数为 $\tilde{\mathcal{O}}(1/\epsilon^2)$。因此，我们的工作在经验性的 ActiveRL 方法与高斯过程学习理论的正式统计基础之间架起了桥梁。

3. 基于高斯过程的主动轨迹收集

3.1. 预备知识

考虑一个折扣马尔可夫决策过程（MDP），$\mathcal{M}=(\mathcal{S},\ \mathcal{A},\ P,\ r,\ \rho,\ \gamma)$，其中 $\mathcal{S}$ 是状态空间，$\mathcal{A}$ 是动作空间，$P(\cdot\mid s,\ a)$ 是转移核，用于指定给定 $(s,\ a)$ 时的下一状态分布；$r\colon \mathcal{S}\times\mathcal{A}\to[0,R_{\max}]$ 是有界奖励函数，ρ 是初始状态分布，$\gamma\in(0,\ 1)$ 是折扣因子。策略 $\pi:\ \mathcal{S}\to\Delta(\mathcal{A})$ 定义了给定状态 S 时动作的分布。对于策略 π，状态-动作价值函数和价值函数定义如下：

$$Q^{\pi}(s,a)\triangleq\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^{t}r(s_t,a_t)\right], \tag{1}$$

$$V^{\pi}(s)\triangleq\mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^{\pi}(s,a)\right]. \tag{2}$$

智能体的目标是找到一个策略 π，使期望折扣回报最大化，如下所示：

$$J(\pi)=\mathbb{E}_{s_0\sim\rho}[V^{\pi}(s_0)],\qquad \pi^{*}=\arg\max_{\pi\in\Pi}J(\pi), \tag{3}$$

其中 Π 表示所有可容许策略构成的空间。

我们考虑主动离线强化学习的混合离线-在线学习设定。学习器获得了一个

离线数据集 $D_{\text{off}} = \{(s_i,\ a_i,\ r_i,\ s_i')\}_{i=1}^{N}$，该数据集由一个未知行为策略 π_b 收集，并配有有限的主动交互预算 M，表示可以从环境中查询的额外状态转移总数。

在最终进行策略优化之前，智能体可以策略性地使用这些 M 在线样本，以减少状态空间关键区域中的认知不确定性。主动查询后的完整数据集为 $D_{\text{all}} = D_{\text{off}} \cup D_{\text{act}}$，其中$D_{\text{act}} = \{(s_t、\ a_t、\ r_t、\ s_t')\}_{t=1}^{M}$。

3.2. 价值函数的高斯过程建模

我们在最优价值函数 $V(。) \sim \mathcal{GP}(0、\ k(、)$上置以高斯过程（GP）先验，其中 $k:\ \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ 是一个正定核，用于编码价值函数结构的光滑性与相似性。直观地说，$k(s、\ s')$量化了状态 s处的价值与状态 s'处的价值之间的相关程度，从而定义了跨状态空间的泛化几何。

给定训练数据 $\mathcal{D}_t = \{(s_i、\ y_i)\}_{i=1}^{N_t}$，其中 N_t 表示截至时间 t所见样本（离线+主动）的总数，记$\mathbf{y}_t = [y_1, y_2, \ldots, y_{N_t}]^\top$ 为观测目标值（可能含噪声）组成的向量，其中每个目标值 y_i 的生成方式为

$$y_i = V^*(s_i) + \eta_i,\quad \eta_i \sim \mathcal{N}(0, \sigma^2),$$

其中，σ^2 表示观测噪声方差。标量 $\sigma > 0$ 量化了每个目标观测中的随机不确定性。例如，这可能是由环境随机性导致的。

在第 t次迭代时，高斯过程以迄今收集的所有数据 $\mathcal{D}_t = \{(s_i,\ y_i)\}_{i=1}^{N_t}$为条件，其中 N_t 表示学习器可用的样本累积总数。离线数据集提供 N 个初始点，每次主动交互都会增加一个样本，因此 $N_t = N + t$，$0 \le t \le M$。

因此，N_t 从 $N_0 = N$（纯离线阶段）单调增长至 $N_M = N + M$（主动探索预算耗尽后）。核矩阵$\mathbf{K}_t \in \mathbb{R}^{N_t \times N_t}$。因此，后验（$\mu_t$，$\sigma_t$）是根据截至第 t 轮的累积数据定义的。

测试状态 s 处的高斯过程后验均值和方差为

$$\mu_t(s) = \mathbf{k}_t(s)^\top (\mathbf{K}_t + \sigma^2 I)^{-1} \mathbf{y}_t,\quad \text{and} \tag{4}$$

$$\sigma_t^2(s) = k(s,s) - \mathbf{k}_t(s)^\top (\mathbf{K}_t + \sigma^2 I)^{-1} \mathbf{k}_t(s), \tag{5}$$

其中 $\mathbf{K}_t$ 是核（Gram）矩阵，其元素为

$$[\mathbf{K}_t]_{ij} = k(s_i, s_j)\ \ \text{and} \tag{6}$$

$$\mathbf{k}_t(s) = [k(s,s_1), \ldots, k(s,s_{N_t})]^\top \tag{7}$$

是测试点 s 与所有训练输入之间的核向量。后验均值 $\mu_t(s)$作为当前价值函数的估计，而后验标准差 $\sigma_t(s)$量化了该估计的认知不确定性。在我们的主动离线强化学习框架中，这一不确定性度量 $\sigma_t(s)$用于指导主动探索中新样本的选择。

3.3. 所提算法

在这一混合式主动离线强化学习框架中，我们拟合一个高斯过程（GP）模型来近似最优价值函数 $V^*(s)$。高斯过程作为一种非参数贝叶斯回归器，既能提供均值预测，也能给出经过校准的认知不确定性度量。高斯过程先验由核函数 $k(.，.)$以及固定的观测噪声方差 σ^2指定。训练数据$\mathcal{D}_t = \{(s_i,\ y_i)\}_{i=1}^{N_t}$——包括来自离线数据集的样本，以及之后主动选取的状态转移——也已提供。

根据这些输入，高斯过程推断出价值函数的后验分布。其后验均值

$$\mu_t(s) = \mathbf{k}_t(s)^\top (\mathbf{K}_t + \sigma^2 I)^{-1} \mathbf{y}_t$$

是对 $V^*(s)$ 的当前估计，后验方差

$$\sigma_t^2(s) = k(s,s') - \mathbf{k}_t(s)^\top (\mathbf{K}_t + \sigma^2 I)^{-1} \mathbf{k}_t(s)$$

则量化了剩余的不确定性。

每当智能体收集到新的主动样本，后验均值和方差便会根据贝叶斯法则确定性地更新，并通过核函数所编码的光滑性，在全局范围内降低 $\sigma_t(s)$ 的不确定性。因此，随着信息不断积累，高斯过程会持续改进对 $V^*(s)$ 的估计。

在离线阶段，高斯过程使用固定数据集 D_{off} 进行训练，以初始化 $(\mu_0,\ \sigma_0)$。在随后的主动阶段，智能体可以查询此前未见过且后验不确定性较高的状态。这些新观测扩大了状态空间的覆盖范围，并收窄了高斯过程的置信带。因此，主动阶段能够精准地针对离线数据薄弱的区域。

总之，高斯过程模型提供了一种有理论依据的方法，将给定的先验结构（核函数和数据）与学习得到的后验预测相结合，从而在主动离线强化学习中实现考虑不确定性的价值估计和选择性探索。该过程列于算法1。

4. 样本复杂度

4.1. 设定

令 $J(\pi^*)$ 表示最优策略的回报，$J(\pi_T)$ 表示经过 T 个主动轮次后学习得到的策略。性能

算法 1 基于高斯过程的主动离线强化学习 输入：离线数据集 D_{off}、核函数 k、噪声 σ^2、主动预算 M。初始化：使用 D_{off} 拟合高斯过程先验 $V(s) \sim \mathcal{GP}(0, k)$。计算初始后验均值 $\mu_0(s)$ 和方差 $\sigma_0(s)$。对于 $t = 1$ 到 T：选择查询状态 $s_t = \arg\max_{s \in \mathcal{S}} \sigma_{t-1}(s)$；选择动作 $a_t \sim \epsilon$-贪心(μ_{t-1})；观察状态转移 (s_t, a_t, r_t, s'_t)；构造目标值 $y_t = r_t + \gamma \mu_{t-1}(s'_t) + \eta_t$；使用 (s_t, y_t) 更新高斯过程后验，得到 (μ_t, σ_t)；使用增强数据集 $D_t = D_{\text{off}} \cup$ 更新策略 $\pi_t \{(s_i, y_i)\}_{i=1}^{t}$。结束循环。输出：最终策略 π_T。

差距定义为 $\Delta_T \triangleq J(\pi^*) - J(\pi_T)$，我们希望满足 $\Delta_T \leq \epsilon$。学习的目标是达到 ϵ–最优性，即 $\Delta_T \leq \epsilon$，其概率至少为 $1 - \delta.$。该算法的样本复杂度是达到这一目标所需的最少主动交互次数 M（或等价地，主动轮次 T）。

4.2. 假设

我们考虑一个折扣马尔可夫决策过程，并作出以下假设；其解释和技术作用将在附录B中说明。

假设 A1 奖励函数一致有界如下：$r(s,\ a) \in [0, R_{\max}]$ 对所有 $(s,\ a) \in \mathcal{S} \times \mathcal{A}$均成立。

假设 A2 最优价值函数 V^* 属于由核函数 k诱导的再生核希尔伯特空间（RKHS）$\mathcal{H}_k$，且其范数有界为 $\|V^*\|_k \leq B$。

假设 A3 转移核 $P(\cdot \mid s, a)$ 关于状态-动作对在总变差距离下是 L_p-利普希茨的，即：

$$\|P(\cdot \mid s, a) - P(\cdot \mid s', a')\|_1 \leq L_p\big(\|s - s'\| + \|a - a'\|\big),$$

对所有 (s,a),(s',a') 均成立。

假设 A4 观测到的贝尔曼目标满足 $y_t = r_t + \gamma \mu_{t-1}(s'_t) + \eta_t$，其中 $\eta_t \sim \mathcal{N}(0, \sigma^2)$。

假设 A5 令 $\sigma_0(s)$ 表示在离线数据集 D_{off} 上进行条件化后的高斯过程后验标准差。我们假设离线覆盖半径 $\sigma_{\max} := \max_{s \in \mathcal{S}} \sigma_0(s)$ 有限。

信息增益 设 $f: \mathcal{S} \to \mathbb{R}$为一个潜在函数，$\mathcal{D}_A = \{(s_i, y_i)\}_{i=1}^{|A|}$ 为含噪观测，其中$y_i = f(s_i) + \eta_i$且 $\eta_i \sim \mathcal{N}(0,\ \sigma^2)$。关于 f观测$\mathbf{y}_A$所获得的信息增益定义为 $I(\mathbf{y}_A;\ f)$，而经过 T次观测后的最大信息增益 条件为

$$\gamma_T = \max_{A \subseteq \mathcal{S},\, |A| = T} I(\mathbf{y}_A; f).$$

对于高斯过程先验 $V \sim \mathcal{GP}(0, k)$，互信息具有如下闭式表达式：$I(\mathbf{y}_A; V) = \frac{1}{2} \log \left| I + \sigma^{-2} \mathbf{K}_A \right|$ 其中，$\mathbf{K}_A$ 是 A 上的核矩阵。量 γ_T 仅取决于核函数 k 和输入维度。

信息增益的增长 对于常用的平稳核函数，最大信息增益满足 $\gamma_T = \mathcal{O}(g_k(T,\ d))$，其中 $g_k(T,\ d)$ 取决于核函数的平滑度和维度 d。具体而言，对于径向基函数核，$\gamma_T = \mathcal{O}((\log T)^{d+1})$；对于 Mat´ern 核，$\gamma_T = \mathcal{O}(T^{\frac{d}{2\nu+d}} \log T)$；对于线性核，$\gamma_T = \mathcal{O}(d \log T)$。

4.3. 主要结果

现在，我们给出主要理论结果，即基于高斯过程（GP）不确定性建模的离线强化学习主动算法 1的样本复杂度界。全文中，我们假设 (A1)–(A5) 成立

定理 4.1(算法 1 的样本复杂度）。设$\epsilon > 0$ 且 $\delta \in (0,1)$。假设环境满足假设 *(A1)-(A5)*。那么，在基于高斯过程的主动采样策略按照最大后验不确定性选择状态的条件下，学习得到的策略 π_T 经过 T个主动轮次后，以至少 $1 - \delta$ 的概率满足：

$$J(\pi^*) - J(\pi_T) \leq C \left(\beta_T + \frac{L}{1 - \gamma} \right) \sqrt{\frac{\gamma_T}{T}}, \qquad (8)$$

其中，$C > 0$是一个通用常数，

$$\beta_T = B + \sigma \sqrt{2(\gamma_T + \log(1/\delta))},$$

γ_T 表示经过 T次查询后的最大信息增益，常数 $L = \gamma\ (1 + L_p)$ 是贝尔曼算子 $\mathcal{T}^\pi$的诱导利普希茨常数，由状态转移光滑性参数 L_p推导得出，而 B 是真实价值函数的*RKHS* 范数上界。

主要困难在于，如何在贝尔曼动态下将高斯过程后验不确定性与价值函数误差严格联系起来。与老虎机或监督学习场景不同——在那些场景中，高斯过程先验直接对目标函数建模——此处的价值函数通过贝尔曼算子隐式定义，而该算子将不确定性在各状态与动作之间耦合起来。我们的分析通过引入利普希茨传播论证克服了这一挑战，该论证控制了高斯过程不确定性在重复贝尔曼备份下的演化。这一步对于确保高斯过程置信界在整个策略评估过程中始终有效至关重要。该界（8）表明，次优性差距与 $\sqrt{\gamma_T / T}$成比例下降，其中 γ_T刻画了

高斯过程后验随信息性样本的获取而收缩的速度。这将认知不确定性的降低与有限在线探索下的策略改进直接联系起来。在主动离线强化学习设定中，若新的在线样本被选择在最不确定的区域（依据高斯过程后验），则策略的改进速率取决于不确定性下降的速率，而该速率由高斯过程的信息增益量化。

经典表格型强化学习的结果（Kearns & Singh, 2002年; Azar 等人, 2017年）实现了关于 ϵ 的 $\tilde{\mathcal{O}}(1/\epsilon^2)$ 依赖。然而，随着有效时域增加，其规模表现不佳，依赖关系为 $(1-\gamma)^{-4}$。核化及基于高斯过程的在线强化学习分析（Chowdhury & Gopalan, 2017年）恢复了类似的 $\tilde{\mathcal{O}}(1/\epsilon^2)$ 依赖，但假设可以进行完整的在线交互。相比之下，我们的结果表明，主动离线设置在有效时域因子降至 $(1-\gamma)^{-2}$ 的情况下，仍能实现相同的最优 ϵ 依赖；与纯离线强化学习的界（Uehara & Sun, 2021年）相比，这是二次改进。这一改进源于高斯过程引导的主动查询将在线样本分配到由不确定性主导的区域，从而更快地降低后验方差。

现在，我们给出主定理 4.1 的一些推论。

推论 4.2（平方指数核）。对于 $\mathbb{R}^d$ 中的平方指数核，信息增益满足 $\gamma_T = \mathcal{O}((\log T)^{d+1})$。因此，在定理*4.1*的条件下，$\epsilon$-最优性可通过以下方式实现：

$$T = \tilde{\mathcal{O}}\left(\frac{(\log T)^{d+1}}{\epsilon^2}\right) \quad \textit{and} \quad M = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2(1-\gamma)^2}\right).$$

推论4.3（与纯离线强化学习的比较）。纯离线强化学习算法需要

$$N_{\text{offline}} = \Omega\left(\frac{1}{\epsilon^2(1-\gamma)^4}\right)$$

个样本才能达到 ϵ-最优性（*Uehara 与 Sun，2021年*）。相比之下，在高斯过程不确定性下，主动离线强化学习能以

$$N_{\text{offline}} + M = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2(1-\gamma)^2}\right).$$

由于采用了有针对性的主动探索，这使得时域依赖性实现了二次方的改进。

定理4.1 中的界源于三个相互作用的原理。(i)光滑性：价值函数属于再生核希尔伯特空间，这会产生宽度为 $\beta_T\sigma_T(s)$的高斯过程集中界。(ii)信息增益：后验方差按 $\mathcal{O}(\sqrt{\gamma_T/T})$ 衰减；(iii)贝尔曼收缩：Lipschitz 连续性确保 V中的不确定性沿贝尔曼算子呈几何级数传播。这些因素共同使认知不确定性降低与策略改进紧密耦合，并首次给出了主动离线强化学习的非参数样本复杂度刻画。所得 $\tilde{\mathcal{O}}(1/\epsilon^2)$ 速率与高斯过程老虎机的已知最优尺度相匹配（Srinivas 等人，2009年），也与核化强化学习相匹配，同时将离线—在线混合学习动态纳入模型。

5. 实验

5.1. 环境与数据集

我们在 D4RL 的连续控制基准上进行评估，涵盖稀疏奖励导航和密集奖励运动控制。导航实验在 Maze2D 和 AntMaze环境中进行，这些环境需要在稀疏奖励下进行长时域规划。运动控制实验则在 HalfCheetah、Hopper 和 Walker2d 环境中进行，用于检验方法在高维状态空间和复杂动力学下的可扩展性。

对于所有环境，我们使用标准的 D4RL 离线数据集，其中包含状态转移（s_t，a_t，r_t，s_{t+1}，d_t）。性能采用 D4RL 归一化分数进行报告，该分数取 20 个评估回合的平均值。

5.2. 离线数据集剪枝

为研究数据覆盖范围有限或存在偏差时的鲁棒性，我们对数据集进行了受控剪枝。对于导航任务，我们采用基于区域的剪枝，移除或截断进入状态空间预定义区域的轨迹。截断的回合被视为超时，而非终止状态转移，以确保贝尔曼备份的有效性。此外，对于所有环境，我们还进行随机回合子采样，仅保留原始数据集的一部分。

这些剪枝策略有助于模拟现实中的离线数据限制，同时保留稳定价值传播所需的轨迹结构。完整的剪枝细节和区域定义见附录 F。

5.3. 使用稀疏高斯过程进行价值函数建模

我们使用高斯过程（GP）对状态价值函数 $V(s)$ 进行建模，用于不确定性感知状态选择和贝尔曼目标值估计，且绝不将其用作控制策略。

精确的高斯过程推断随数据点数量呈立方增长，对于大规模离线数据集不可行。因此，我们采用了稀疏变分高斯过程（SVGP），它使用固定的一组诱导点来近似完整的高斯过程。这产生了可处理的推断，同时保留

校准的认知不确定性，这对于指导主动数据收集至关重要。

具体而言，我们为价值函数 $V(s)$ 设置高斯过程先验，其均值为零，核函数为平稳核。我们考虑平方指数（RBF）核和 Mat´ern 核，并默认使用平滑参数为 $\nu = 2.5$ 的 Mat´ern 核，因为对于强化学习中常见的非平滑价值景观，它能提供更好的归纳偏置。我们通过在诱导变量上的变分分布来近似高斯过程后验，并联合学习诱导位置和核函数超参数。

所有高斯过程模型均使用 GPyTorch 库实现。架构和优化细节见附录F

5.4. 拟合值迭代初始化

使用离线数据集上的拟合值迭代（FVI）初始化高斯过程。从目标值 $y^{(0)} = r$ 开始，我们迭代应用贝尔曼更新

$$y^{(k+1)} = r + \gamma(1-d)\,\mu^{(k)}(s'),$$

其中 $\mu^{(k)}(s')$ 表示高斯过程后验均值。该过程会在主动数据收集之前，将稀疏奖励传播到整个状态空间，并随着新数据的获取定期重复。

5.5. 控制策略学习

控制策略使用标准离线强化学习算法独立于高斯过程进行训练。对于 Maze2D 和运动任务，我们使用 TD3+BC；对于 AntMaze环境，则使用 IQL。该策略作为稳定的行为先验，并定期使用离线轨迹与主动轨迹的混合数据重新训练。主动数据通过完整的环境展开收集，而非孤立的状态转移。训练计划和网络架构详见附录F。

5.6. 主动数据集收集

主动数据收集从标准的环境重置开始。在每一步中，智能体要么遵循控制策略，要么执行不确定性引导探索。进行不确定性引导探索时，我们围绕策略给出的动作采样多个候选动作，并执行单步前瞻展开。随后执行使高斯过程预测的下一状态价值的后验标准差最大的动作。这种动作条件准则将探索引向认知不确定性较高的区域，同时确保动作在动态上可行。

训练在小规模和大规模再训练阶段之间交替进行。小规模更新使用最新的主动数据进行轻量级策略再训练大规模更新则使用所有累积数据，对策略和高斯过程进行再训练。这种双时间尺度调度兼顾了计算成本与稳定性。

5.7. 基线与评估指标

我们将结果与以下基线进行了比较。其中包括：(i) 行为克隆（BC），即在离线数据集上通过监督学习训练策略；(ii) 离线强化学习（Offline），即仅使用离线数据集训练策略；以及 (iii) 随机探索（Random），即使用均匀随机动作收集主动数据。

表1. 不同环境下性能与样本效率的比较。回报为归一化的 D4RL 分数。降低幅度（%）是根据离线数据集中的所有样本计算的。

环境	BC	离线强化学习	随机探索	我们的方法（高斯过程）降低幅度（%）	
halfcheetah-random-v2	2.3	13.5	38.2	**39.1**	41.3
hopper-random-v2	4.2	8.2	21.5	**22.6**	30.1
walker2d-random-v2	2.0	7.9	10.7	**10.8**	34.4
halfcheetah-medium-v2	42.8	48.3	54.3	**57.7**	30.1
hopper-medium-v2	54.0	68.1	83.1	**85.1**	32.4
walker2d-medium-v2	73.1	83.6	84.9	**85.2**	42.2
maze2d-umaze-hard-v1	-8.1	-11.5	100.3	**152.12**	78.1
maze2d-medium-easy-v1	-4.5	-4.2	59.6	**156.71**	75.4
maze2d-medium-hard-v1	-4.4	-3.7	41.8	**150.71**	61.9
maze2d-medium-dense-hard-v1	-6.1	-12.3	78.3	**119.2**	89.2
maze2d-large-easy-v1	1.7	-3.6	15.6	**120.5**	54.0
maze2d-large-hard-v1	-2.3	-2.0	5.3	**109.1**	40.4
antmaze-umaze-v0	62.0	74.7	79.9	**83.5**	40.1
antmaze-umaze-diverse-v0	54.0	56.0	38.6	**80.6**	51.6
antmaze-medium-play-v0	0.0	58.1	51.3	**71.8**	40.3
antmaze-medium-diverse-v0	1.3	62.1	61.2	**69.5**	30.4

所有基线均使用相同的网络架构、训练预算和数据集混合流程。性能通过 D4RL归一化分数以及随主动环境步数变化的学习曲线进行评估。对于导航任务，我们还会绘制价值和不确定性热力图。定量结果汇总于表 1，更多分析见附录 F。

5.8. 实现概述

为确保结果可复现，我们在此列出关键超参数；完整细节见附录 F。 **架构：** 控制策略（Actor/Critic）采用具有 256 个隐藏单元和 ReLU 激活函数的 3 层多层感知机进行参数化。为确保广泛覆盖状态空间，稀疏变分高斯过程的诱导点通过对离线数据集进行 K-Means 聚类初始化。**主动循环：** 我们采用 ϵ- 贪心探索策略，$\epsilon = 0.2$。在不确定性引导步骤中，我们采样 $N = 32$ 个候选动作，并选择预测的下一状态方差最大的动作。为兼顾稳定性与速度，我们每经过 10^3 个环境步数便进行轻量级策略更新，同时每经过 5×10^3 步进行完整的高斯过程重新训练（通过 FVI）和策略整合。

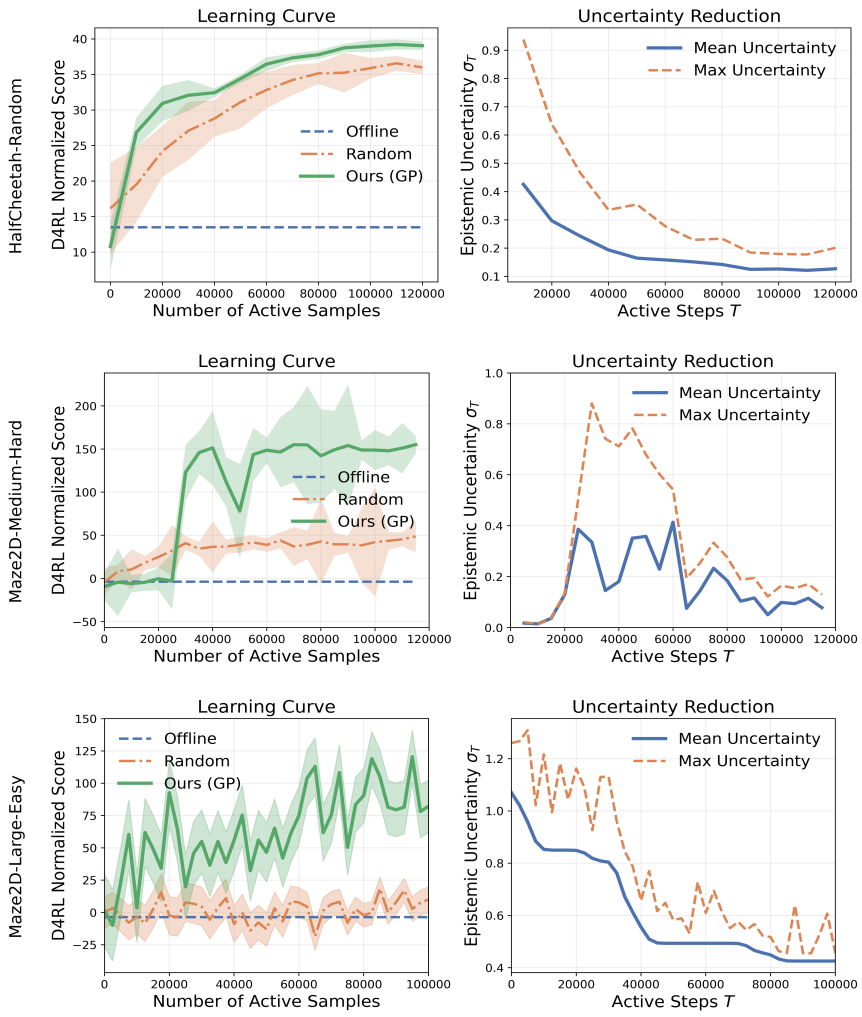

图*1*。 不同环境下的学习性能与不确定性降低。**上方**：HalfCheetah-Random。**中间**：Maze2D- 中等-困难。**下方**：Maze2D-Large-Easy。

6. 结果与讨论

表1 总结了所有基准测试的性能。在密集奖励运动控制和稀疏奖励导航任务中，不确定性感知主动数据收集始终优于行为克隆、纯离线强化学习和随机探索基线，同时所需的环境交互显著更少。

6.1. 性能与样本效率

在各类运动环境中，我们的方法达到了或超过了最强基线的表现，同时所需样本大幅减少。在 halfcheetah-medium-v2 和 hopper-medium-v2 上，我们分别取得了 57.7 和 85.1 的归一化分数，所需样本约减少 30–32%。在 walker2d-medium-v2 上也观察到了类似表现，所需样本减少超过 40%，同时实现了相当的性能。这些结果表明，即使在密集奖励设定中，不确定性引导探索仍然有效。

在稀疏奖励导航任务中，性能提升幅度显著更大。在 Maze2D 上，我们的方法始终优于所有基线方法，且优势通常十分明显。例如，在 maze2d-umaze-hard-v1 和 maze2d-medium-easy-v1 中，相较于随机探索，我们观察到归一化分数提升超过 50 分，同时将样本总数减少了 75% 以上。这凸显了无引导探索在长时域目标条件任务中的失败，以及针对认知不确定区域进行探索的重要性。

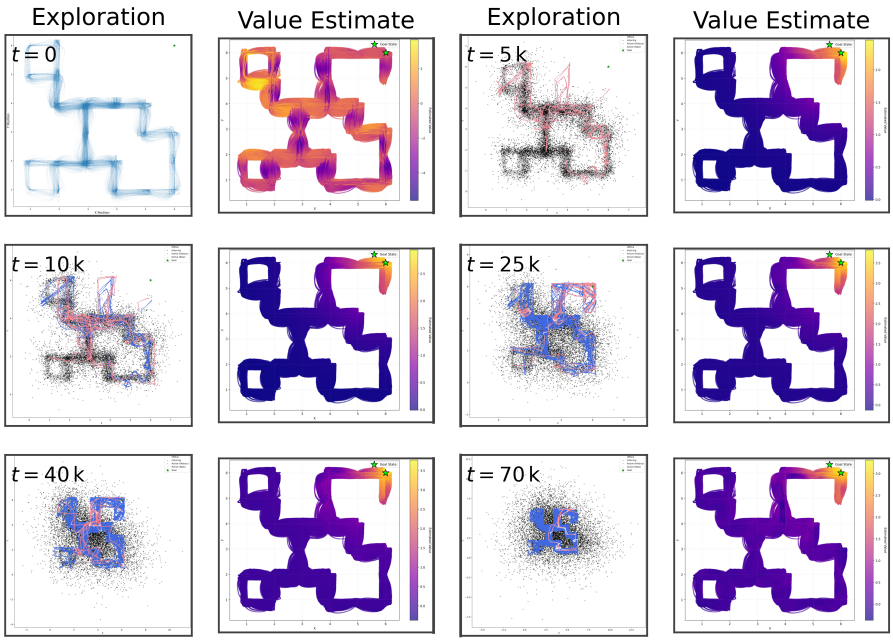

图2. maze2d-medium-hard 中的不确定性驱动探索与价值估计。对于每个时间步 t，展示探索图（左）以及对应的价值函数估计（右）。价值估计使用完整的离线数据集计算。

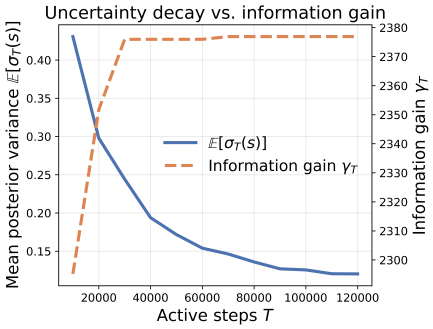

图3. **不确定性衰减与信息增益的关系（halfcheetah-random-v2）**。随着主动步数 T 增加，平均后验方差降低，而累积信息增益增加，这与高斯过程集中界 $\sigma_T = O(\gamma_T/T)$ 一致。

6.2. 不确定性驱动的探索与价值传播

图 2 展示了不确定性如何引导 `maze2d-medium-hard`中的探索。训练初期，较高的后验方差集中在未探索的瓶颈和死胡同区域附近。随着主动数据不断收集，不确定性在局部区域逐渐降低，价值函数估计也会在迷宫中连贯地传播。这一行为与贝尔曼压缩论证一致：一旦关键状态转移上的不确定性得到降低，价值信息就会沿长时域快速传播。

6.3. 不确定性衰减与学习动态

图1 比较了密集奖励任务（`HalfCheetah-Random`）和稀疏奖励任务（`Maze2D-Medium-Hard`）中的学习曲线与不确定性降低幅度。在这两种情况下，性能提升都伴随着高斯过程后验方差持续下降。值得注意的是，在密集奖励环境中，不确定性衰减得更快，因为频繁的奖励反馈加速了价值估计；而在稀疏奖励任务中，不确定性降低较慢，但结构更为明显，并集中在与任务相关的区域。

6.4. 信息增益与理论验证

为直接验证定理4.1背后的信息论论证，我们追踪了平均后验方差 $\mathbb{E}[\sigma_T(s)]$ 以及 γ_T在主动探索过程中的累积信息增益。正如预测的那样，不确定性会随着主动步数的增加而降低，而信息增益则随 T次线性增长。这种实证关系证实，不确定性最大化查询能够有效降低后验方差，其变化速率与$\mathcal{O}(\sqrt{\gamma_T/T})$相符，从而验证了信息增益在提升样本效率方面的作用。图3，4。

6.5. 可扩展性与实践考量 s

高斯过程模型并非天然具备可扩展性，即使采用稀疏变分近似，也会引入不容忽视的计算开销。在我们的实验中，我们使用了 5,000 到 25,000 个诱导点来完成这些任务。

尽管这一设定足以产生可靠的不确定性估计并实现稳定的价值学习，但训练成本仍高于纯神经方法。因此，我们的目标并非声称其可扩展性已达到深度方法的水平，而是通过实验验证我们在分析中推导出的不确定性与样本复杂度保证。

7. 总结

基于高斯过程的主动离线强化学习算法应被视为深度 ActiveRL 方法的连续核化极限。该算法通过假设贝叶斯后验精确，理想化了不确定性估计过程，从而能够严格分析不确定性降低幅度和价值函数的收敛性。在实践中，基于集成的不确定性估计近似体现了同一原理。认知不确定性较高的区域对应较高的高斯过程方差。因此，我们的理论结果为主动离线强化学习中不确定性引导探索的经验成功提供了形式化依据。

参考文献

Ayyagari, R. S.、Ega, R. R.与Dukkipati, A. 外部时间过程下的马尔可夫决策过程。*arXiv:2305.16056*，2025。Azar, M. G.、Osband, I.与Munos, R. 强化学习的极小极大遗憾界。见国际机器学习会议，第263–272页。PMLR，2017年。Chowdhury, S. R.与Gopalan, A. 论核化多臂老虎机，2017年。Delavari, E.、Khanzada, F. K.与Kwon, J. CARLA模拟器中自动驾驶强化学习的综合综述。*arXiv:2509.08221*，2025。Dukkipati, A.、Banerjee, R.、Ayyagari, R. S.与Udaybhai, D. P. 无需主控的导航技能学习：一种序贯多策略强化学习算法。见*IEEE/RSJ*智能机器人与系统国际会议（*IROS*）论文集，2022。Dukkipati, A.、Ayyagari, R. S.、Dasgupta, B.、Dutta, P.与Onteru, P. R. 面向离线策略改进的主动强化学习策略。见第*39*届人工智能促进协会人工智能会议论文集（*AAAI*），2025。Engel, Y.、Mannor, S.与Meir, R. 基于高斯过程的强化学习。见第*22*届国际机器学习会议论文集，第201–208页，2005。

Kearns, M.与Singh, S. 多项式时间内的近最优强化学习。机器学习，49(2):209–232，2002年。

Kuss, M. 和 Rasmussen, C. 强化学习中的高斯过程。神经信息处理系统进展，16，2003。

Levine, S.、Kumar, A.、Tucker, G. 和 Fu, J. 离线强化学习：教程、综述及开放问题展望。*arXiv* 预印本 *arXiv:2005.01643*，2020。

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. 多臂老虎机场景中的高斯过程优化：无遗憾与实验设计。*arXiv*预印本 *arXiv:0912.3995*，2009年。

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. 多臂老虎机场景中高斯过程优化的信息论遗憾界。*IEEE* 信息论汇刊，58(5):3250–3265，2012年5月。ISSN 1557-9654。doi: 10.1109/tit.2011.2182033。URL
h
t
tp://dx.doi.org/10.1109/TIT.2011.2182033。

Tang, C., Abbatematteo, B., Hu, J., Chandra, R., Mart´ın-Mart´ın, R., and Stone, P. 面向机器人学的深度强化学习：真实世界成功案例综述。载于第*39*届人工智能 *AAAI* 会议论文集，2025年。Uehara, M. and Sun, W. 部分覆盖下的悲观模型型离线强化学习。*arXiv*预印本 *arXiv:2107.06226*，2021年。Wang, S., Blanchet, J., and Glynn, P. 混合折扣马尔可夫决策过程强化学习的最优样本复杂度，2023年。URLhttps://arxiv.org/abs/2302.07477。

Yeh, S.-Y., Chang, F.-C., Yueh, C.-W., Wu, P.-Y., Bernac- chia, A., and Vakili, S. 基于核函数的 Q-learning 的样本复杂度，2023年。URLhttps://arxiv.org/abs/2302.00727。Yu, C., Liu, J., Nemati, S., and Yin, G. 医疗保健中的强化学习：综述。*ACM*计算调查，55 (5):1–36，2021年。

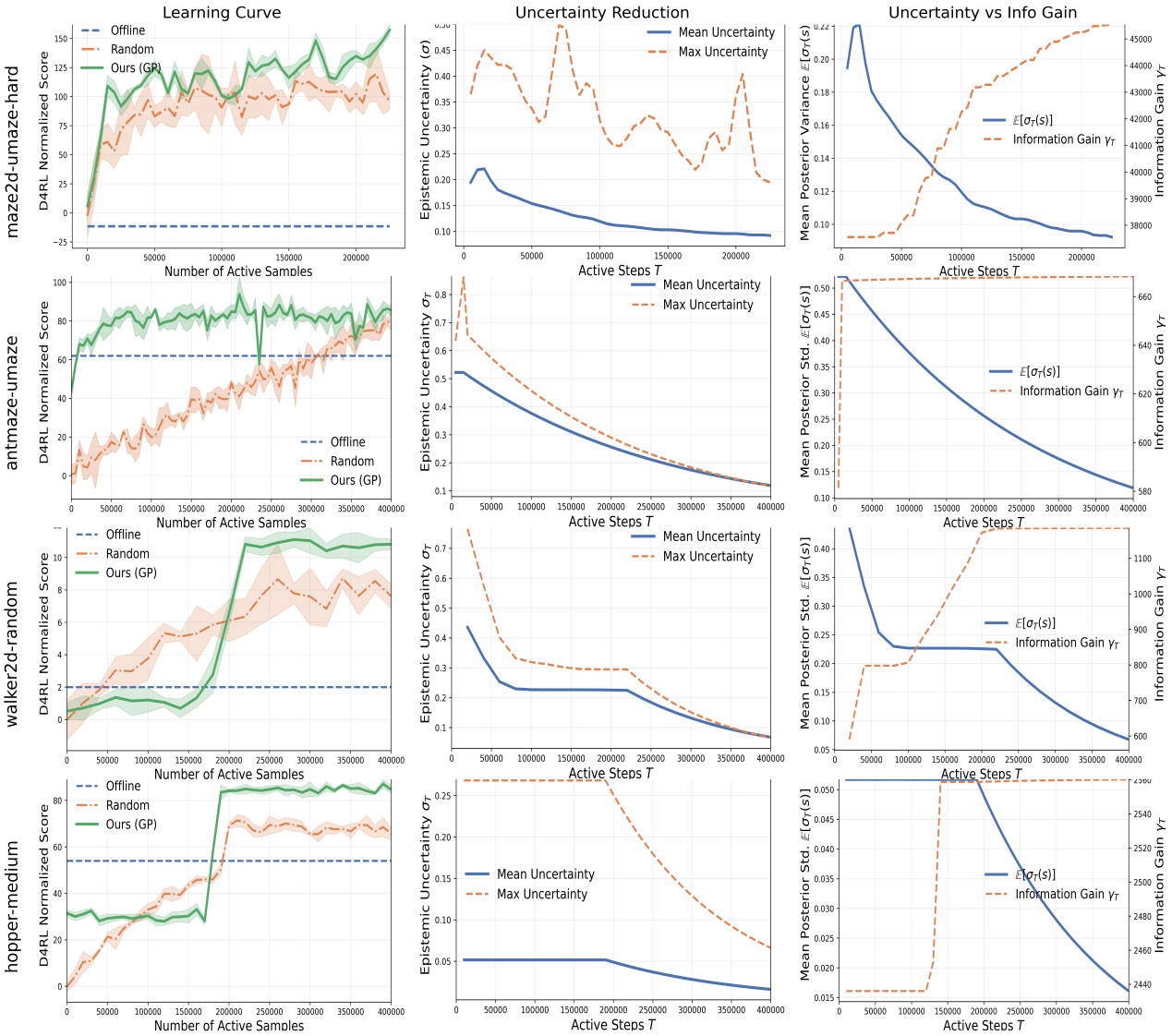

图*4*。**不同环境中的学习动态**。每一行对应一个不同的环境，各列分别展示（**a**）主动数据收集下的学习曲线、（**b**）主动步数中高斯过程认知不确定性的降低幅度，以及（**c**）后验不确定性与累积信息增益之间的关系。

离线强化学习的样本高效主动算法附录

A. 符号说明

我们在此总结定理 4.1 中及整个分析过程中出现的关键量和常数。

主动离线强化学习符号说明：

- $J(\pi)$：策略 π 的期望折扣回报。
- $\gamma \in (0, 1)$：马尔可夫决策过程的折扣因子，也决定了贝尔曼算子的收缩率。
- T：主动采样轮次的数量（每轮新增一次状态转移）。
- C：用于汇总归一化常数和不等式常数的通用数值常数。
- δ：高概率保证的置信水平（$1-\delta$）。

利普希茨常数和光滑性常数：

- L_p：转移核的利普希茨常数 $P(\cdot|s,\ a)$，定义见假设 B.3：

$$\|P(\cdot|s,a) - P(\cdot|s',a')\|_1 \le L_p(\|s-s'\| + \|a-a'\|).$$

  它衡量下一状态分布对（$s,\ a$）微小变化的敏感程度。

- L：贝尔曼算子 $\mathcal{T}^\pi$ 的诱导利普希茨常数，定义如下

$$L = \gamma(1+L_p),$$

  其中 它量化了价值函数中的估计误差在贝尔曼更新中的传播程度 .

高斯过程记号：

- $\boldsymbol{B}$：真实价值函数 $\|V^*\|_k \le B$ 的 RKHS 范数上界，见假设 B.2。它刻画了V^* 相对于高斯过程（GP）核函数的光滑性。
- σ：高斯观测噪声的标准差 高斯观测噪声的偏差 $\eta_t \sim \mathcal{N}(0,\sigma^2)$，出现在 GP 回归模式中 l.
- β_T：高斯过程置信宽度参数，

$$\beta_T = B + \sigma\sqrt{2(\gamma_T + \log(1/\delta))},$$

  它决定了高斯过程后验围绕真实函数集中的程度。

- γ_T： 最大信息增益在 T 个样本后，

$$\gamma_T = \max_{|A|=T} I(\mathbf{y}_A; V) = \frac{1}{2}\log\left|I + \sigma^{-2}\mathbf{K}_A\right|.$$

  它衡量了在获得 T 个观测后，后验不确定性能够降低的程度。

B. 详细假设

假设 B.1 （有界奖励）。 奖励函数一致有界： $r(s,a) \in [0, R_{\max}]$ 对所有 $(s,a) \in \mathcal{S}\times\mathcal{A}$均成立。**假设 B.2** （再生核希尔伯特空间平滑性）。 最优价值函数满足 $V^* \in \mathcal{H}_k$，其中 $\|V^*\|_k \le B$，且 $\mathcal{H}_k$ 是由核函数 k诱导的再生核希尔伯特空间。

Assu设 B.3 （利普希茨转移）。 转移核 $P(\cdot \mid s,a)$ 关于 (s,a) 是 L_p-利普希茨的 ：

$\|P(\cdot \mid s,a) - P(\cdot \mid s',a')\|_1 \le L_p\big(\|s-s'\| + \|a-a'\|\big), \quad \forall (s,a),(s',a').$

假设 B.4 （有界观测噪声）。 观测到的贝尔曼目标满足 $y_t = r_t + \gamma\mu_{t-1}(s'_t) + \eta_t$，其中$\eta_t \sim \mathcal{N}(0,\ \sigma^2)$ 为独立同分布，且已知 $\sigma > 0$。

定义 B.1 （离线覆盖半径）。 令 $\sigma_0(s)$ 表示在离线数据集 D_{off} 上条件化后的高斯过程后验标准差。定义 $\sigma_{\max} := \max_{s\in\mathcal{S}} \sigma_0(s)$。

假设 B.5 （有界离线覆盖）。 离线覆盖半径是有限的： $\sigma_{\max} < \infty$。

对于任意固定策略 π，作用于有界价值函数的贝尔曼算子 $\mathcal{T}^\pi$ 由下式给出

$$(\mathcal{T}^\pi V)(s) = r(s,\pi(s)) + \gamma\, \mathbb{E}_{s'\sim P(\cdot|s,\pi(s))}[V(s')].$$

此外， $\mathcal{T}^\pi$ 在上确界范数下是一个 γ-压缩映射。

定义 B.2 （信息增益）。设 f： $\mathcal{S} \to \mathbb{R}$ 为一个潜在函数， $\mathcal{D}_A = \{(s_i,\ y_i)\}_{i=1}^{|A|}$ 为输入位置 $A \subseteq \mathcal{S}$ 处的含噪观测，其中 $y_i = f(s_i) + \eta_i$，且噪声 $\eta_i \sim \mathcal{N}(0,\ \sigma^2)$ 相互独立。观测$\mathbf{y}_A$关于 f 的信息增益定义为互信息： $I(\mathbf{y}_A;\ f) = H(\mathbf{y}_A) - H(\mathbf{y}_A \mid f)$。

其中 $H(\cdot)$ 表示香农微分熵。经过 T 次观测后，最大信息增益为 $\gamma_T = \max_{A\subseteq\mathcal{S},\ |A|=T} I(\mathbf{y}_A;\ f)$。该量衡量观测 T 个含噪样本能够在多大程度上降低对潜在函数 f 的不确定性。 γ_T 的增长率取决于底层函数类或核函数的光滑性。

定义 B.3（高斯过程回归中的信息增益）。设 $V \sim \mathcal{GP}(0,k)$ 为核函数为 k 的高斯过程先验， $\mathbf{y}_A = [y_1,\dots,y_T]^\top$ 表示在点 $A = \{s_1,\dots,s_T\}$ 处对 V 的含噪观测，其中 $y_i = V(s_i) + \eta_i$,$\eta_i \sim \mathcal{N}(0,\sigma^2)$。那么，函数 V 与观测 $\mathbf{y}_A$ 之间的互信息具有如下闭式表达式： $I(\mathbf{y}_A;V) = \frac{1}{2}\log\left|I + \sigma^{-2}\mathbf{K}_A\right|$，其中 $\mathbf{K}_A$ 为核矩阵，其元素为 $[\mathbf{K}_A]_{ij} = k(s_i,s_j)$。经过 T 个样本后，最大信息增益为 $\gamma_T = \max_{A\subseteq\mathcal{S},|A|=T} \frac{1}{2}\log\left|I + \sigma^{-2}\mathbf{K}_A\right|$.。项 γ_T 量化了通过 T 个观测所能实现的高斯过程后验不确定性的最大降低幅度。它仅取决于核函数 k 的选择和输入维度 d，并且在控制累积后验方差和样本复杂度方面发挥着核心作用。

引理 B.4 （光滑核的有界信息增益）。 对于常用的平稳核，最大信息增益 γ_T满足 $\gamma_T = \mathcal{O}(g_k(T,d))$，其中 $g_k(T,d)$ 取决于核光滑度和输入维度d。具体而言，对于径向基函数核， $\gamma_T = \mathcal{O}\left((\log T)^{d+1}\right)$；对于 *Mat´ern*， $\gamma_T = \mathcal{O}\left(T^{\frac{d}{2\nu+d}}\log T\right)$；对于线性核， $\gamma_T = \mathcal{O}(d\log T)$。

这一结果源于核积分算子的谱衰减。对于平方指数核，特征值以超多项式速度衰减，从而得到 $\gamma_T = \mathcal{O}((\log T)^{d+1})$，并确保不确定性缩减呈次线性。

这些假设共同构成了一个光滑且性质良好的环境，在其中高斯过程后验经过校准，贝尔曼更新保持稳定。这些假设较为宽松，也是高斯过程学习理论和基于核函数的强化学习中的标准假设（Chowdhury &Gopalan，2017年），但足以推导出我们的主要 $\mathcal{O}(1/\epsilon^2)$ 样本复杂度结果。

C. 关键引理

本节介绍用于证明定理4.1的关键中间结果。每个引理分别剖析分析中的一个独立部分——集中性、方差控制、贝尔曼误差传播和性能改进。完整证明见附录 E。

引理 C.1 (高斯过程集中界)。 在假设 *B.2* - 假设 *B.4*下，以至少 $1-\delta$的概率，对所有$t \ge 1$和 $s \in \mathcal{S}$,

$$|V^*(s) - \mu_{t-1}(s)| \le \left(B + \sigma\sqrt{2(\gamma_{t-1} + \log(1/\delta))}\right)\sigma_{t-1}(s).$$

角色： 该引理根据预测方差界定高斯过程后验均值的估计误差，并为 V^* 建立了一个高概率置信包络，这是整个不确定性驱动的探索论证的基础。 证明概要： 直接将 *GP-UCB* 置信界（*Srinivas* 等人，*2009*年，定理 *2*）应用于价值函数V^*即可。完整证明见附录 *E*。

引理 C.2（方差–信息增益关系）。 对于任意被查询状态序列 $\{s_t\}_{t=1}^{T}$，累积预测方差满足

$$\sum_{t=1}^{T} \sigma_{t-1}^2(s_t) \le C_\sigma \, \gamma_T,$$

其中 γ_T 是在假设*B.5*中定义的最大信息增益，且 C_σ 仅取决于核函数 k 和噪声方差 σ^2。作用：该引理将后验不确定性降低幅度与信息增益联系起来，表明在高方差区域进行主动查询可以实现次线性累积不确定性。证明概要：由*Srinivas 等人（2009年，引理 5）*中的方差求和不等式推出，并针对我们的强化学习设定进行了重述。

引理 C.3（贝尔曼误差的 Lipschitz 传播）。在假设*B.3*下，贝尔曼算子 $\mathcal{T}^\pi$ 满足

$$|\mathcal{T}^\pi V(s) - \mathcal{T}^\pi \tilde{V}(s)| \le \gamma(1 + L_p)\|V - \tilde{V}\|_\infty, \quad \forall V, \tilde{V} : \mathcal{S} \to \mathbb{R}.$$

作用：该引理界定了价值函数中的估计误差如何通过贝尔曼更新传播，并对迭代过程中的误差放大提供几何控制。证明概要：利用$\mathcal{T}^\pi$ 的压缩性质以及 L_p–*Lipschitz* 连续性推出，其中 $P(\cdot|s,\ a)$；完整证明见附录*E*。

引理 C.4（期望性能差）。令 π_T表示经过 T轮主动查询后的策略，且 μ_T为相应的高斯过程均值估计。然后，在假设*B.1*下，

$$J(\pi^*) - J(\pi_T) \le \left(\beta_T + \frac{L}{1-\gamma}\right) \mathbb{E}_{s\sim d_\rho^{\pi_T}}[\sigma_T(s)],$$

其中 $d_\rho^{\pi_T}$表示 π_T下的折扣状态分布。 作用： 将高斯过程不确定性 σ_T直接与策略次优性联系起来，贯通分析中的统计与控制两个方面。 证明思路： 结合高斯过程集中界（定理 *C.1*）与贝尔曼利普希茨性质（定理*C.3*）；详见附录 *E*。

定理 C.1 给出了高斯过程后验的高概率置信区域；定理 C.2确保通过信息增益实现次线性不确定性降低；定理 C.3 保证误差通过贝尔曼算子传播时受到控制；定理 C.4 则将这些结果转化为策略性能界。综合以上结果，可得定理 4.1中的主要样本复杂度界。

D. 主要定理的证明

本节给出主要理论结果（定理 4.1）的详细证明。每个证明都遵循模块化结构，并调用第 C 节中建立的引理以及第B节中陈述的假设。为便于理解，我们先简要概述论证思路。

证明分为四个步骤。

1. **误差分解：** 将次优性差距表示为估计误差与传播误差之和。

2. **引理应用：** 利用高斯过程集中性（定理 C.1）和贝尔曼利普希茨传播（定理 C.3）来界定这些误差。

3. **方差控制：** 应用方差–信息增益关系（定理 C.2），将预期不确定性降低幅度表示为主动查询数量 T 的函数。

4. **结 合结果：** 汇总所有界，得到最终速率，并完成定理 4 的证明 .1.

定理证明 *4.1*。 **步骤 1：次优性分解**。 根据期望回报的定义（公式（3）），我们将次优性差距表示为

$$J(\pi^*) - J(\pi_T) = \mathbb{E}_{s\sim\rho}[V^*(s) - V^{\pi_T}(s)].$$

按照标准方法，我们将差距分解为两项：

$$V^*(s) - V^{\pi_T}(s) = [V^*(s) - \mu_T(s)] + [\mu_T(s) - V^{\pi_T}(s)].$$

第一项对应于真实值与高斯过程估计值之间的估计误差，第二项表示在 π_T 下因不完美的贝尔曼更新而产生的传播误差。

步骤 2：界定估计项与传播项。 根据高斯过程集中不等式（定理 C.1）以及假设 B.2，我们以至少 $1-\delta$ 的概率得到，

$$|V^*(s) - \mu_T(s)| \le \beta_T \sigma_T(s), \quad \beta_T = B + \sigma\sqrt{2(\gamma_T + \log(1/\delta))}.$$

接下来，调用引理C.3，我们得到

$$|\mu_T(s) - V^{\pi_T}(s)| \le \frac{L}{1-\gamma}\,\sigma_T(s),$$

其中 $L = \gamma(1+L_p)$ 由假设 B.3。对 $s \sim d_\rho^{\pi_T}$取期望得到

$$J(\pi^*) - J(\pi_T) \le \left(\beta_T + \frac{L}{1-\gamma}\right)\mathbb{E}_{s\sim d_\rho^{\pi_T}}[\sigma_T(s)]. \tag{9}$$

步骤 3：方差–信息增益关系。 根据方差–信息增益引理（定理 C.2）以及有界信息增益假设（假设 B.5），累积后验方差满足

$$\sum_{t=1}^{T} \sigma_{t-1}^2(s_t) \le C_\sigma\,\gamma_T.$$

根据詹森不等式，我们有

$$\mathbb{E}[\sigma_T(s)] \le \sqrt{\mathbb{E}[\sigma_T^2(s)]} \le \sqrt{\frac{C_\sigma \gamma_T}{T}}.$$

将其代入方程（9）可得

$$J(\pi^*) - J(\pi_T) \le C'\left(\beta_T + \frac{L}{1-\gamma}\right)\sqrt{\frac{\gamma_T}{T}}, \tag{10}$$

其中 C' 吸收了来自 C_σ 以及对数据分布取期望所产生的常数。

步骤 4：得出界限。重新整理方程（10以确保 ϵ-最优性（$J(\pi^*) - J(\pi_T) \le \epsilon$），可得所需的主动轮次数：

$$T \ge \tilde{\mathcal{O}}\left(\frac{(\beta_T + \frac{L}{1-\gamma})^2\,\gamma_T}{\epsilon^2}\right).$$

由于每个主动轮次对应至多 $\bar{H} \le (1-\gamma)^{-1}$ 次状态转移，因此所需主动转移的总数为 s

$$M = T \cdot \bar{H} = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2(1-\gamma)^2}\right),$$

从而证明了定理4.1中的结论。 □

该证明展示了高斯过程不确定性如何随主动查询而收缩，以及这种收缩如何转化为更高的策略价值。其关键新颖之处在于，通过利普希茨传播，将基于信息增益的后验方差衰减（定理 C.2）与贝尔曼稳定性联系起来（见定理 C.3）。这种耦合使我们能够达到 $\tilde{\mathcal{O}}(1/\epsilon^2)$ 的速率，同时将时域依赖从 $(1-\gamma)^{-4}$（离线强化学习）改进为 $(1-\gamma)^{-2}$（主动离线强化学习）。该论证在非参数设定下统一了高斯过程学习理论与强化学习样本复杂度分析。

E. 完整证明

本附录给出第 C 节和 D节中引理和主定理的详细证明。所有符号均沿用第 A节，假设见第B节。

性能差异引理。 对于任意两个平稳策略 π 和 π'，它们的期望折扣回报之差满足

$$J(\pi') - J(\pi) = \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d_\rho^{\pi'}}\Big[(\mathcal{T}^{\pi'}V^{\pi})(s) - V^{\pi}(s)\Big], \tag{11}$$

其中，$d_\rho^{\pi'}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P(s_t = s \mid s_0 \sim \rho,\ \pi')$ 表示 π' 的折扣状态访问分布，而 $\mathcal{T}^{\pi'}$ 是与 π' 相关联的贝尔曼算子。

直觉： 该引理用新策略 π' 下 V^{π} 的贝尔曼残差，表示两个策略之间的期望回报差。这是策略评估中的标准结论，也是强化学习中许多遗憾和样本复杂度分析的基础。

E.1. 引理证明C.1（高斯过程集中界）

为便于说明，我们重述该引理。

引理 E.1 （高斯过程集中界）。在假设 *B.2* –假设 *B.4*成立时，以至少 $1-\delta$ 的概率，对所有$t \geq 1$和 $s \in \mathcal{S}$,

$$|V^*(s) - \mu_{t-1}(s)| \leq \left(B + \sigma\sqrt{2(\gamma_{t-1} + \log(1/\delta))}\right)\sigma_{t-1}(s).$$

证明。 该证明直接由 GP-UCB 集中不等式（Srinivas 等人，2009年）。设 $V^* \in \mathcal{H}_k$ 表示真实价值函数，且 $\|V^*\|_k \leq B$ 由假设 B.2。对于观测 $\{(s_i,\ y_i)\}_{i=1}^{t-1}$ 满足高斯噪声模型（假设 B.4），高斯过程的后验均值与方差满足：

$$\mu_{t-1}(s) = \mathbf{k}_{t-1}(s)^\top(\mathbf{K}_{t-1} + \sigma^2 I)^{-1}\mathbf{y}_{t-1},\quad \sigma_{t-1}^2(s) = k(s,s) - \mathbf{k}_{t-1}(s)^\top(\mathbf{K}_{t-1} + \sigma^2 I)^{-1}\mathbf{k}_{t-1}(s).$$

根据高斯过程的标准集中性，以至少 $1-\delta$ 的概率，对所有 s,

$$|V^*(s) - \mu_{t-1}(s)| \leq \beta_t\sigma_{t-1}(s),\quad \beta_t = B + \sigma\sqrt{2(\gamma_{t-1} + \log(1/\delta))}.$$

该结果源于在再生核希尔伯特空间范数下界定后验均值与真实函数的偏差。该论证利用再生性质 $\langle V^*,\ k(s,\ \cdot)\rangle_k = V^*(s)$ 以及互信息 γ_{t-1} 作为模型复杂度的度量。

□

E.2. 引理 C.2 的证明（方差–信息增益关系）

引理 E.2 (方差–信息增益关系)。 对于任意一组被查询状态 $\{s_t\}_{t=1}^T$,

$$\sum_{t=1}^{T}\sigma_{t-1}^2(s_t) \leq C_\sigma\,\gamma_T.$$

证明。 这可由高斯过程的信息增益恒等式得出（Srinivas 等人，2009年）。定义 $\gamma_T = \max_{A\subset\mathcal{S},\ |A|=T} I(\mathbf{y}_A; V)$, 其中 $I(\mathbf{y}_A; V)$ 是潜在函数 V与带噪观测值 $\mathbf{y}_A$之间的互信息。对于观测噪声为 σ^2 的高斯过程模型，我们有

$$I(\mathbf{y}_A; V) = \frac{1}{2}\log\left|I + \sigma^{-2}\mathbf{K}_A\right|.$$

根据互信息的性质，在 s_t 处的每个新观测最多贡献 $\frac{1}{2}\log(1 + \sigma_{t-1}^2(s_t)/\sigma^2)$ 个信息比特。对 T 个轮次求和，并利用不等式 $\log(1+x) \geq x/(1+x)$，可得

$$\sum_{t=1}^{T}\min\{1, \sigma_{t-1}^2(s_t)\} \leq 2\log\left|I + \sigma^{-2}\mathbf{K}_T\right| = 2\gamma_T.$$

设置ng $C_\sigma = 2$ 并利用 $\sigma_t(s)$的有界性去掉截断 ≤ 1，即完成证明。 □

E.3. 引理 C.3 （贝尔曼误差的 Lipschitz 传播）

引理 E.3 (贝尔曼误差的 Lipschitz 传播)。 在假设 *B.3*下，

$$|\mathcal{T}^\pi V(s) - \mathcal{T}^\pi \tilde{V}(s)| \le \gamma(1+L_p)\|V-\tilde{V}\|_\infty.$$

证明。 回顾 $\mathcal{T}^\pi V(s) = \mathbb{E}_{a\sim\pi(s)}[r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')]]$。根据三角不等式和假设 B.3,

$$\begin{aligned}|\mathcal{T}^\pi V(s) - \mathcal{T}^\pi \tilde{V}(s)| &= \gamma\left|\mathbb{E}_{a\sim\pi(s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s') - \tilde{V}(s')]\right| \\ &\le \gamma \mathbb{E}_{a\sim\pi(s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}|V(s') - \tilde{V}(s')| \\ &\le \gamma(1+L_p)\|V-\tilde{V}\|_\infty,\end{aligned}$$

其中 $(1+L_p)$ 说明了在相近状态和动作下， $P(\cdot|s,a)$中的扰动。这为 $\mathcal{T}^\pi$确立了所需的 Lipschitz 常数。

□

E.4. 引理 C.4（期望性能差）

引理 E.4（期望性能差）。 令 π_T 表示经过 T 个主动轮次后的策略， μ_T 表示相应的高斯过程均值估计。则，

$$J(\pi^*) - J(\pi_T) \le \left(\beta_T + \frac{L}{1-\gamma}\right)\mathbb{E}_{s\sim d_\rho^{\pi_T}}[\sigma_T(s)].$$

证明。利用性能差异引理（公式（11））,

$$J(\pi^*) - J(\pi_T) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi_T}}[(\mathcal{T}^{\pi^*}V^{\pi_T})(s) - V^{\pi_T}(s)].$$

我们将被积函数分解为

$$(\mathcal{T}^{\pi^*}V^{\pi_T})(s) - V^{\pi_T}(s) = (\mathcal{T}^{\pi^*}V^*)(s) - \mathcal{T}^{\pi^*}\mu_T(s) + \mathcal{T}^{\pi^*}\mu_T(s) - \mathcal{T}^{\pi_T}\mu_T(s) + \mathcal{T}^{\pi_T}\mu_T(s) - V^{\pi_T}(s).$$

第一项差值可由高斯过程集中界（定理 C.1）界定为 $\beta_T\sigma_T(s)$。其余两项由 $\mathcal{T}^\pi$ 的 Lipschitz 连续性（定理 C.3）控制，额外产生 $\frac{L}{1-\gamma}\sigma_T(s)$。考虑期望即可完成证明。

□

E.5. 定理 4.1 的证明

为方便起见，我们将主定理重述如下:

定理 E.5. 算法 *1* 的样本复杂度。在假设*B.1*–假设 *B.5*下，经过 π_T 个 T主动轮次后得到的策略满足：其成立概率至少为 $1-\delta$,

$$J(\pi^*) - J(\pi_T) \le C\left(\beta_T + \frac{L}{1-\gamma}\right)\sqrt{\frac{\gamma_T}{T}}.$$

证明。 结合定理 C.4与定理 C.2 中的方差界，我们得到

$$J(\pi^*) - J(\pi_T) \le \left(\beta_T + \frac{L}{1-\gamma}\right)\mathbb{E}[\sigma_T(s)] \le \left(\beta_T + \frac{L}{1-\gamma}\right)\sqrt{\frac{C_\sigma\gamma_T}{T}}.$$

为确保 ϵ-最优性（$J(\pi^*) - J(\pi_T) \le \epsilon$），只需

$$T = \tilde{\mathcal{O}}\left(\frac{(\beta_T + \frac{L}{1-\gamma})^2\gamma_T}{\epsilon^2}\right).$$

将 $\gamma_T = \mathcal{O}((\log T)^{d+1})$ 代入平方指数核，得到 $M = T\cdot\bar{H} = \tilde{\mathcal{O}}(1/[\epsilon^2(1-\gamma)^2])$，如定理4.1所述。

□

F. 实验

F.1. 离线数据集剪枝

为了系统地研究离线数据覆盖度不同情况下的性能，我们构建了每个环境的剪枝版本，并控制其难度等级。

导航任务（Maze2D、AntMaze）
对于导航环境，我们通过逐步加强基于区域的剪枝，创建了简单、中等和困难变体。在二维状态空间中定义轴对齐的矩形区域，进入这些区域的轨迹要么被移除，要么在首次进入时被截断。被截断的轨迹被标记为超时，以保持贝尔曼目标的有效性。

在区域剪枝之后，我们应用了回合级随机剪枝，以在保持时间连贯性的同时进一步控制数据集规模。在所有导航任务中，所得离线数据集保留了约原始数据的7–10%，较低比例对应于更困难的变体。

运动任务（HalfCheetah、Hopper、Walker2D）
对于运动环境，由于空间结构没有实际意义，我们采用均匀随机回合剪枝。对于所有运动任务，我们保留了原始数据集的 20%，以维持轨迹一致性和行为多样性。在所有设定中，均未引入奖励重标记、轨迹拼接或合成数据。所有主动数据均在线从环境中收集。

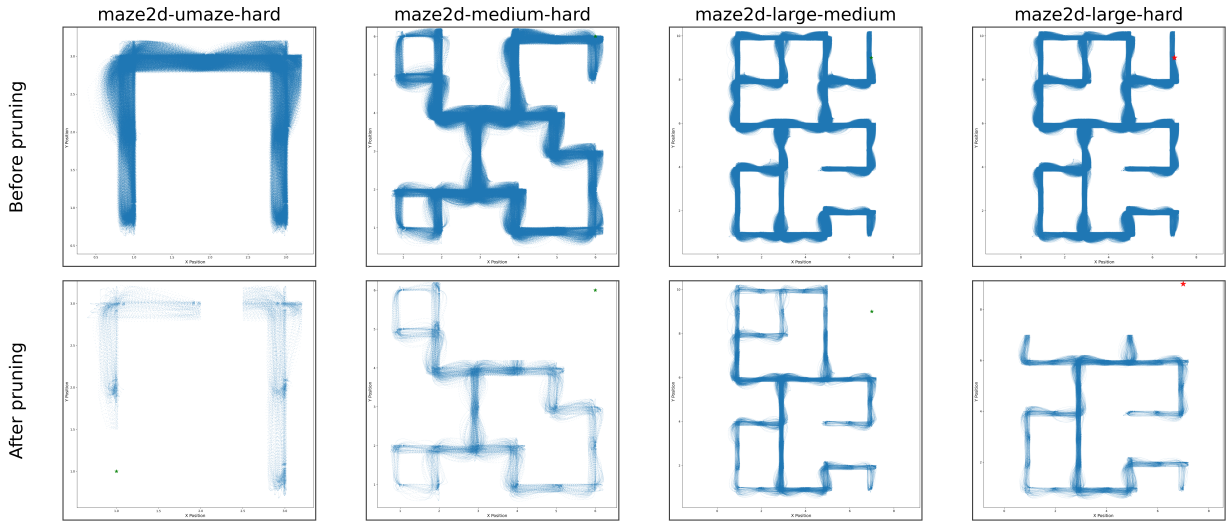

图5. 四种 Maze2D 环境在数据集剪枝前（上）后的（下）状态访问分布：maze2d-umaze-hard、maze2d-medium-hard、maze2d-large-medium 和 maze2d-large-hard。尽管剪枝后仅保留少量轨迹，但任务相关区域的覆盖范围得以保留，同时去除了冗余和低效用状态。

F.2. 高斯过程架构与训练

F.2.1. 模型架构

我们使用具有零均值先验的稀疏变分高斯过程对价值函数 $V(s)$ 进行建模。核函数选用平方指数（径向基函数）族或 Mat´ern 族，并设置光滑度参数 $\nu = 2.5$。默认使用 Mat´ern 核函数，因为它对非光滑价值景观具有更好的归纳偏置。令 $\mathcal{Z} = z_m{}_{m=1}^{M}$表示诱导变量。高斯过程后验通过诱导变量上的变分分布进行近似，

$$q(U) = N(m, S)$$

其中 $\mathbf{u} = V(\mathcal{Z})$。所有诱导位置均与核超参数联合学习 .

F.2.2. 诱导点初始化

诱导点已初始化 使用小批量 K 均值聚类对离线数据集中的归一化状态观测进行初始化在人线数据集。这种初始化与随机初始化相比，可以加快训练早期的收敛速度。

诱导点的数量取决于环境，并经过选择以平衡近似质量和计算成本。导航任务的诱导点数量从数千个到运动任务的最多一万个不等。

F.2.3. 变分推断与优化

后验推断通过最大化变分证据下界（证据下界（ELBO ），

$$\mathcal{L}_{\mathrm{ELBO}} = \mathbb{E}_{q(f)}\big[\log p(y \mid f)\big] - \mathrm{KL}(q(u) \,\|\, p(u))\,.$$

我们使用自然梯度下降（NGD）优化变分参数，该方法能够在高斯分布空间中提供稳定且高效的更新。核超参数和观测噪声使用 Adam 进行优化。这两种优化器在每个训练步骤中联合应用。

训练使用状态-价值对的小批量进行。这使得方法能够扩展到大型数据集，同时保持不确定性估计的一致性。

F.2.4. 使用高斯过程目标值的拟合值迭代

在拟合值迭代过程中，高斯过程使用通过贝尔曼备份迭代更新的目标值进行训练。在第k次迭代中，状态转移（s， a, r， s'， d）的目标值为

$$y^{(k+1)} = r + \gamma(1-d)\,\mu^{(k)}(s'),$$

首先仅使用离线数据应用 FVI，以获得价值函数的稳定初始化。在每个主动数据收集阶段之后，使用增强数据集恢复 FVI，使价值信息能够传播到状态空间中新探索的区域。

F.2.5. 不确定性的作用

高斯过程后验方差被解释为由数据覆盖有限而产生的认知不确定性。这种不确定性用于指导主动探索，优先选择价值估计最不确定的状态转移。重要的是，不确定性绝不会直接用于控制目标；它只会影响所收集的数据。

表 2. 稀疏变分高斯过程（SVGP）训练超参数

超参数	值
高斯过程模型	稀疏变分高斯过程（SVGP）
核函数	径向基函数（平方指数）与马特恩
均值函数	零均值 / 常数均值
似然	高斯
变分推断	证据下界优化
变分分布	自然变分分布
变分优化器	自然梯度下降（NGD）
NGD 学习率	0.01
核超参数优化器	Adam
Adam 学习率	1×10^{-3}
批次大小	2048–4096
初始高斯过程训练轮数	10–20
主动阶段高斯过程训练轮数	5
诱导点初始化	小批量 K 均值
学习诱导位置	Yes

表 3. 稀疏变分高斯过程中使用的诱导点数量

环境	诱导点数量
Maze2D / AntMaze	5,000 – 25,000
HalfCheetah	5,000–10,000
Hopper	5,000–10,000
Walker2d	5,000–10,000

表 4. 拟合值迭代与主动高斯过程更新配置。

参数	值
离线拟合值迭代次数	10–20
主动拟合值迭代次数	5
折扣因子 γ	0.99
每批主动转移次数	5,000
每次离线 FVI 的高斯过程训练轮数	3–5
每次主动 FVI 的高斯过程训练轮数	2

表5. 控制策略训练计划。

设定	导航	运动
离线强化学习算法	TD3+行为克隆 / IQL	TD3+行为克隆 / IQL
离线训练步数	$\lvert\mathcal{D}_{\text{offline}}\rvert$ (TD3+行为克隆) $2 \times \lvert\mathcal{D}_{\text{offline}}\rvert$ (IQL)	$\lvert\mathcal{D}_{\text{offline}}\rvert$ $2 \times \lvert\mathcal{D}_{\text{offline}}\rvert$ (IQL)
较小的主动更新间隔	1k 步	2k 步数
较小的策略更新步数	5k（TD3+BC）/ 10k（IQL）	10k / 20k
大主动更新间隔	5k 步数	10k 步数
较大的策略更新步数	25k / 50k	50k / 100k
批次大小	256	256
网络架构	两层多层感知机（256 个单元）	两层多层感知机（256 个单元）

F.3. 控制策略训练细节

我们根据环境使用 TD3+行为克隆和 IQL 训练控制策略。所有优化器设置和算法特定超参数均遵循既有工作中报告的标准配置，并在各项实验中保持不变。

每个策略首先在剪枝数据集上进行离线训练，直至收敛。在主动学习过程中，策略会定期使用混合数据集重新训练，该数据集结合了离线数据与新收集的转移数据。我们采用了两种重新训练计划。**小规模再训练阶段** 触发频繁，仅使用最新批次的主动数据，从而能够快速适应新探索的区域。**大规模再训练阶段** 触发频率较低，使用完整的混合数据集，以确保稳定性并防止先前学到的行为发生灾难性遗忘。

对于导航任务，我们在 Maze2D环境中使用 TD3+BC，在 AntMaze环境中使用 IQL。离线训练和每个主动再训练阶段中的梯度更新次数会根据数据集大小进行调整，并且在导航任务与运动任务之间有所不同，详见表 5

G. 补充结果

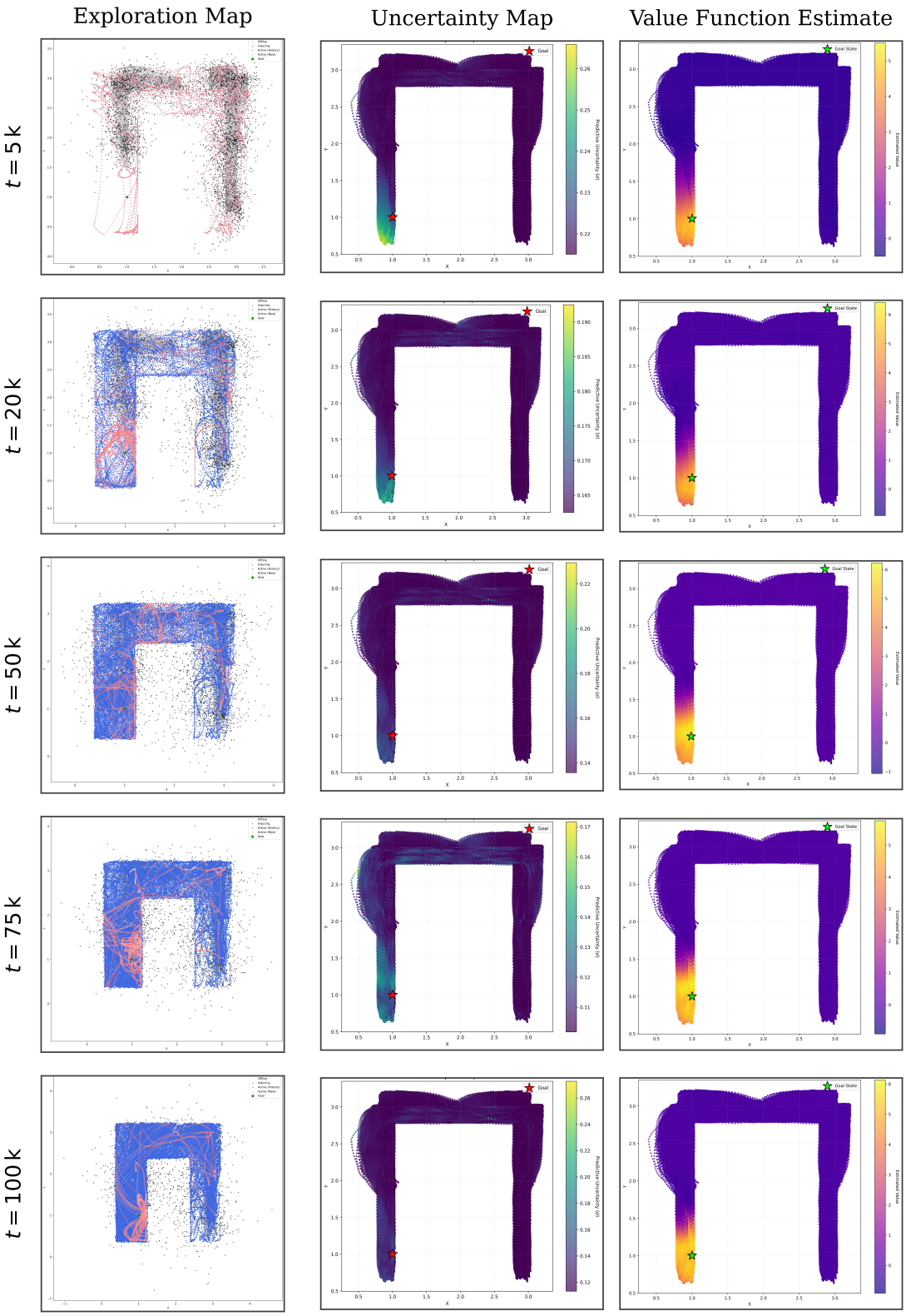

图*6.* 在 maze2d-umaze-hard 中，主动学习期间探索、不确定性和价值估计的演变。各列分别展示**(a)** 采样得到的探索轨迹、**(b)** 高斯过程认知不确定性 $\sigma_T(s)$，以及**(c)** 学习得到的价值函数估计 $V(s)$。各行对应逐渐增加的主动交互预算。随着探索不断进行，不确定性逐渐集中，并在频繁访问的区域内降低；与此同时，价值函数在到达目标的轨迹周围变得更加清晰，展现了不确定性驱动的数据收集及其对策略学习的影响。

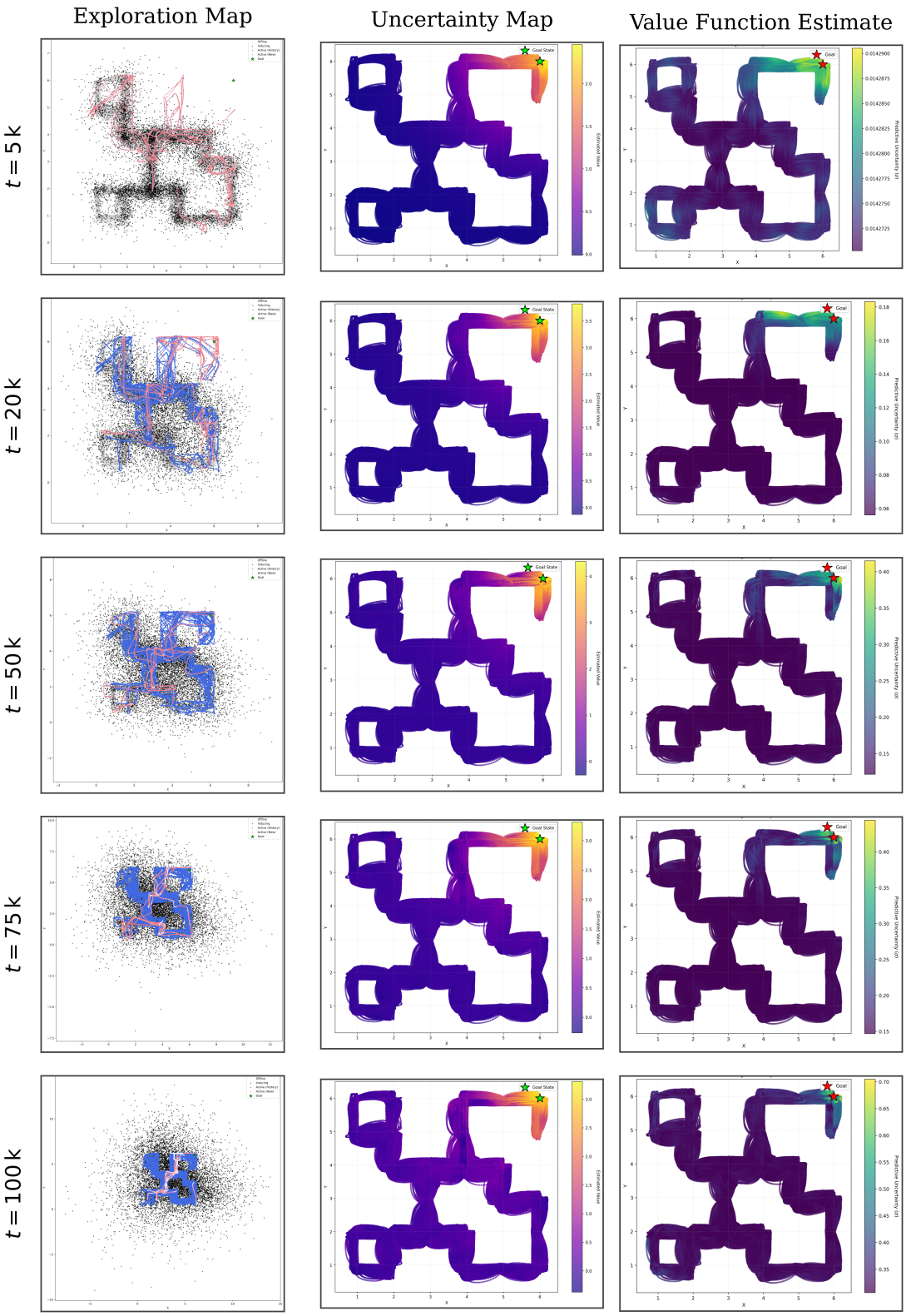

图 7。 探索的演变过程、不确定性以及在 maze2d-medium-hard 中主动学习期间的价值估计。各列展示 w
(a)采样得到的探索轨迹，**(b)** 高斯过程认知不确定性 $\sigma_T(s)$，以及**(c)** 学习得到的价值函数估计 $V(s)$。