基于信念空间度量的离线POMDP学习覆盖框架

Youheng Zhu
西北大学

Yiping Lu
西北大学

arXiv:2603.03191v1 [stat.ML] 3 Mar 2026

摘要

在部分可观测马尔可夫决策过程（POMDP）的离策略评估（OPE）中，智能体必须根据过去的观测推断隐状态，这加剧了现有离策略评估方法中的时域诅咒和记忆诅咒。本文提出一种新颖的覆盖分析框架，利用信念空间（潜在状态上的分布）的内在度量结构，放宽传统的覆盖假设。通过假设与价值相关的函数在信念空间中满足利普希茨连续性，我们推导出误差界，从而缓解时域和记忆长度呈指数级增长的问题。我们的统一分析技术适用于广泛的离策略评估算法，并给出具体的误差界和覆盖要求；这些要求以信念空间度量而非原始历史覆盖来表述。我们通过案例研究展示该框架在样本效率上的提升：双重采样贝尔曼误差最小化算法，以及基于记忆的未来依赖价值函数（FDVF）。在这两种情况下，基于信念空间度量的覆盖定义都能得到更紧的误差界。

1 引言

离策略评估（OPE）是离线强化学习中的一个核心问题，旨在利用在不同的行为策略 π_b下收集的数据，估计目标策略 π_e 的预期累积奖励。这一设定自然出现在现实世界的应用中，因为交互式数据收集通常不可行或不安全，学习必须完全依赖预先收集的离线轨迹。在本文中，我们考虑一种更现实但也更具挑战性的设定：只能获得底层潜在状态的部分观测。这会导致非马尔可夫的观测动态，其可能依赖于动作-观测对的完整历史。此类场景可由部分可观测马尔可夫决策过程（POMDP）建模，相较于完全可观测的马尔可夫决策过程，它们能够更充分地表达现实世界问题 [Atrash 等人，2009年、Lauri 等人， 2022]但也引入了显著的复杂性。

虽然对于部分可观测马尔可夫决策过程，将历史轨迹视为状态后，马尔可夫性得以恢复，此时该问题可简化为马尔可夫决策过程问题，但直接应用传统的马尔可夫决策过程方法（如重要性采样和贝尔曼残差最小化），必然会导致误差界随时域 H 呈指数级增长，这种现象称为时域诅咒。例如，在重要性采样中，序贯重要性权重会随时域呈指数级增长，导致估计方差难以处理。为缓解这一问题，有人提出了一种称为未来依赖价值函数（FDVF）的方法，适用于无记忆策略，但在引入基于记忆的策略后便会失效；在这种情况下，覆盖会随记忆长度呈指数级增长，面临记忆诅咒 [Zhang 和 Jiang， 2024]。

为克服历史即状态马尔可夫决策过程中的时域与记忆诅咒，我们将问题重新表述于信念空间中。信念空间是部分可观测马尔可夫决策过程中的核心概念，定义为：给定观测到的动作与观测历史后，潜在状态上的概率分布所构成的空间。信念空间中的每个元素——称为信念状态——都可作为历史轨迹的替代表示。由于显式计算信念状态需要智能体能够获知潜在动态过程，因此它最常见于部分可观测马尔可夫决策过程规划文献中。利用信念空间的度量结构，基于点的价值迭代（PBVI）等规划方法通过稀疏覆盖信念子空间 [Shani等人，2013，Lee等人，2007，Zhang

29^{th}人工智能与统计国际会议（AISTATS）2026年论文集，摩洛哥丹吉尔。PMLR：第 300 卷。版权所有 2026年，作者所有。

等人， 2014]。尽管信念空间结构已在部分可观测马尔可夫决策过程规划 [Lee等人，2007] 和模型学习 [Zhang 等人， 2012], 中得到广泛研究，但其在离策略评估（ OPE）中的作用仍未得到充分探索。值得注意的是，当前的离线学习方法通常忽略这种度量结构，显式处理历史空间，导致对时域长度呈指数级依赖。这引出了一个关键问题：

尽管信念空间的度量结构已被证明能有效刻画部分可观测马尔可夫决策过程规划中的计算复杂度，它能否同样刻画离线学习中的样本复杂度？更具体地说，我们能否利用这种信念空间结构，以规避离线*POMDP*学习中的指数级复杂度？

我们的贡献 受这一问题启发，我们的工作探讨了信念空间度量结构的思想，并借助信念度量研究一些常见无模型离策略评估算法的理论保证。我们框架的核心思想类似于状态抽象 [Li等人， 2006], 即信念空间的复杂度可以通过一种将相似状态收缩在一起的抽象而显著降低。也就是说，如果两条历史轨迹具有相似的信念状态，则在分析中应将它们视为相似。为此，我们将自身限制在一个**策略子集**上，即具有稳定性的策略。这一关于策略类的结构性假设足够丰富，能够包含我们感兴趣的所有可能策略，并具备有助于实现更紧覆盖的良好性质。我们的分析总体结果，以及与受视野/记忆诅咒影响的现有结果的比较，呈现在下方的表 1 中。总体而言，我们的结果缓解了覆盖的指数性，尤其是在信念空间具有光滑性结构时，如例 1 和例 2 所示。为具体说明我们的贡献：

- 我们提出了一种分析框架，利用由 ε-覆盖诱导的状态抽象，在抽象空间上获得覆盖，以适用于离策略评估问题中的广泛场景。该框架很容易推广到其他算法，甚至其他强化学习任务。

- 我们在表 1、定理 4 和定理 5 中证明，使用信念空间覆盖得到的覆盖不劣于原始覆盖。我们还在示例 1 和示例 2 中说明，在 POMDP模型满足特定光滑性性质时，我们的覆盖能够解决维度灾难或时域／记忆灾难。

- 在第 5.1 章中，我们以双重采样算法为例，对贝尔曼误差最小化算法进行了详尽分析。在第5.2章中，我们还介绍了未来依赖价值函数的流程，其中利用了部分可观测马尔可夫决策过程和策略的快速遗忘特性。我们随后证明，FDVF 的分析更为简单，无需对部分可观测马尔可夫决策过程系统本身作出任何假设。**这表明，与“视界诅咒”相比，“记忆诅咒”要容易处理得多**。此外，这也回答了 [Zhang 和 Jiang，2024], 提出的一个悬而未决的问题：在对策略作出结构性假设的情况下，我们能够缓解“记忆诅咒”。

2 相关工作

部分可观测马尔可夫决策过程规划。 在部分可观测马尔可夫决策过程规划文献中，基于点的价值迭代（PBVI）的思想 [Shani 等人，2013，Kurniawati 等人，2008，Poupart 等人，2011，Shani 等人，2008，Smith 和 Simmons，2012，Spaan 和 Vlassis， 2005]是在整个信念空间的有限子集上进行计算，以求得最优策略。值得注意的是，PBVI 的一个重要特点是，其信念子空间的选择利用了信念空间中的度量结构：算法每次扩展信念子集时，都会在一步可达的范围内，寻找相对于当前信念子集距离最远的信念点，从而尽可能稀疏地构建可达信念子集。此外，Lee 等人，2007，Zhang 等人， 2014], 揭示了复杂度与信念空间度量之间的联系，并证明存在一种近似算法，其复杂度关于可达信念空间的覆盖数呈多项式关系。

离策略评估中的时域诅咒与记忆诅咒。针对完全可观测马尔可夫决策过程中的离策略评估（OPE），已有许多算法，例如重要性采样 [Precup 等人，2000，Jiang 和 Li，2016，Jiang 等人，2019，Hu 和 Wager， 2023], 拟合 Q 迭代（FQE） [Ernst 等人，2005，Munos 和 Szepesvári，2008，Le 等人， 2019], 基于双重采样的 Bellman 残差最小化 [Baird 等人，1995], 极小极大估计量 [Antos 等人，2008，Chen 和 Jiang，2019，Feng 等人，2019，Nachum 等人，2019，Uehara 等人，2021，Zanette 和 Wainwright， 2022], 以及边缘化重要性采样 [Uehara 等人， 2020]。然而，若将每条轨迹历史视为一个独立状态，并直接将这些方法应用于部分可观测马尔可夫决策过程（POMDP），便会遇到一个根本性挑战，即时域诅咒：随着时域增长，误差界会呈指数级恶化，因为覆盖假设会随着呈指数级增长的历史空间而扩大。另一种方法是采用近期提出的未来依赖价值函数（FDVF） [Uehara 等人，2023，Zhang 和 Jiang，2024]将覆盖要求转移到隐状态上，从而为无记忆策略提供多项式级保证。然而，该方法受到记忆诅咒的限制，其复杂度

扩展到基于记忆的策略时会重新变为指数级，这是因为必须捕捉未来观测与历史记忆状态之间的依赖关系，从而大幅增加覆盖复杂度。

3 预备知识

无限时域折扣部分可观测马尔可夫决策过程：无限时域折扣部分可观测马尔可夫决策过程可表示为一个 7 元组： $\mathcal{M}=\langle\mathcal{S},\ \mathcal{A},\ \mathcal{O},\ r,\ \gamma,\ \mathbb{O},\ \mathbb{T}\rangle$ 其中，$\gamma\in[0,1)$ 是折扣因子，$\mathcal{S}$ 是隐状态空间，$\mathcal{A}$ 是动作空间，$\mathcal{O}$ 是观测空间，$r\colon \mathcal{S}\times\mathcal{A}\to[0,R_{\max}]$ 是有界奖励函数，$\mathbb{O}:\ \mathcal{S}\ \to\ \Delta(\mathcal{O})$ 是发射核（即给定状态时观测的条件分布），$\mathbb{T}\colon \mathcal{S}\times\mathcal{A}\to\Delta(\mathcal{S})$ 是转移核（即给定当前状态-动作对时下一状态的条件分布）。我们用 $\Delta(\cdot)$ 表示给定空间上的概率分布，用 $|\cdot|$ 表示集合的基数。为简单起见，且不失一般性，我们假设空间 $\mathcal{S}$，$\mathcal{A}$，$\mathcal{O}$是离散且有限的，其基数可以很大。

部分可观测马尔可夫决策过程的演化如下：从初始隐状态 $s_1\sim d_0(s)$ 开始，在每个步骤 h，隐状态s_h 会发出一个从 $\mathbb{O}(s_h)$ 中抽取的观测 o_h，环境则根据当前状态-动作对 $(s_h,\ a_h)$ 生成奖励 r_h。随后，状态按照 $s_{h+1}\sim\mathbb{T}(s_h,\ a_h)$ 转移。至关重要的是，在一般的部分可观测马尔可夫决策过程中，学习者无法访问隐状态空间 $\mathcal{S}$；可用的只有通过离线行为策略收集的轨迹。

我们还广泛考虑第5.2章中讨论的有限时域部分可观测马尔可夫决策过程设置。在有限时域场景中，我们令折扣因子 $\gamma=1$，智能体与环境交互有限个步骤 H。

离线数据：离线数据集 $\mathcal{D}$ 由行为策略 π_b采集。采集过程包括从部分可观测马尔可夫决策过程独立采集 n 条样本轨迹（$o_1,\ a_1,\ \cdots$）。从每条轨迹中截取前 h 个元素组成前缀，形成元组（$o_1,\ a_1,\ r_1,\ o_2,\ a_2,\ r_2,\ \cdots,\ o_h,\ a_h,\ r_h,\ o_{h+1}$），其中 h 是随机选取的。最后，数据集的形式为 $\mathcal{D}_1$，如下所示。在第6章中，对于未来依赖价值函数（FDVF），离线数据的定义略有不同。在 FDVF 设置中，我们考虑长度为 H 的有限时域部分可观测马尔可夫决策过程。同样，使用行为策略 π_b 与环境交互并采集数据。这次将整条轨迹视为单个数据点，如 $\mathcal{D}_2$所示。

$$\mathcal{D}_1=\{(o_1^{[i]},a_1^{[i]},r_1^{[i]},\cdots,o_{h_i}^{[i]},a_{h_i}^{[i]},r_{h_i}^{[i]},o_{h_i+1}^{[i]})\}_{i=1}^n,$$

$$\mathcal{D}_2=\{((o_1^{[i]},a_1^{[i]},r_1^{[i]},\cdots,o_H^{[i]},a_H^{[i]},r_H^{[i]})\}_{i=1}^n$$

状态抽象：对于一个马尔可夫决策过程（$\mathcal{S},\mathcal{A},r,\gamma,P$），其中 $P\colon\ \mathcal{S}\times\mathcal{A}\to\Delta(\mathcal{S})$ 表示转移核，抽象 ϕ 是从 $\mathcal{S}$ 到抽象状态空间 $\mathcal{S}_\phi$的映射，并将该马尔可夫决策过程转换为抽象马尔可夫决策过程（$\mathcal{S}_\phi,\mathcal{A},r_\phi,\gamma,P_\phi$），其中 $r_\phi(\phi(s),\ a):=\mathbb{E}_{s\sim p_{\phi(s)}}[r(s,\ s_d\ \ s\ \ a\ =\mathbb{E}_{s'\sim p_{\phi(s)}}[\sum_{\phi\ s''=\phi\ s_d}P\ s''|s'\ a\ ]\ \ \{p_x\}_{x\in\mathcal{S}_\phi}\ )]$，且 $P_\phi(\phi(\ )|\phi(\ )\ ):(\ )\ (\ )\ (\ ,\ )$。这里，$p_x$是任意一个分布族，其分布的支撑集位于 $\phi^{-1}(x)$。对于定义在抽象系统 f_{bin} 上的任意函数 $\mathcal{S}_\phi\to\mathbb{R}$，我们将其提升版本定义为 $[f_{\text{bin}}]_{\text{true}}(\cdot):=f_{\text{bin}}(\phi(\cdot))$。对于抽象策略 $\pi_\phi:\mathcal{S}_\phi\to\Delta(\mathcal{A})$，其提升版本 $[\pi_\phi]_{\text{true}}(\cdot):=\pi_\phi(\phi(\cdot))$ 的定义也类似。在下文中，ϕ 通常由 ε选取，并被视为等价。通常也使用带上标或下标 ϕ 的记号来表示定义在抽象系统上的函数；每当我们说 $f_\phi\in\mathcal{F}$，其中 $\mathcal{F}$ 是真实系统上定义的函数类时，我们指的是 $\exists f\in\mathcal{F},f(\phi(\cdot))=f_\phi(\phi(\cdot))$。

其他符号：我们将状态-动作占用率记为 $d^\pi(s,\ a):=(1-\gamma)\sum_{k=1}^\infty\Pr_\pi(S_k=s,\ A_k=a)$。$J(\pi)$ 表示策略 π 的期望奖励，而$J_{\hat{Q}}(\pi)$ 则是使用近似函数 $\hat{}\ Q$得到的 π 估计奖励。

4 统一分析概述

本节简要说明信念状态空间的几何结构如何帮助刻画部分可观测马尔可夫决策过程离策略评估的样本复杂度，并概述我们的总体结果。我们还介绍信念空间的基础知识、由 ε-覆盖诱导的信念空间抽象，以及与信念度量相关的假设。

信念状态空间与光滑性条件：由于无法直接观测隐状态，我们可以利用观测和动作的完整历史信息来预测当前状态。我们将时间步 h 的历史记为 $\tau_h=(o_1,a_1,o_2,a_2,\cdots,o_{h-1},a_{h-1})\in\mathcal{H}_h\subset\mathcal{H}$ 和$\tau_h^+:=(\tau_h,o_h)\in\mathcal{H}_h^+\subset\mathcal{H}^+$。因此，信念状态 $\mathsf{b}(\tau_h^+):\ \subset\mathbb{R}^{|\mathcal{S}|}$ $\mathcal{B}$

$\Pr(s_h|\tau_h^+)$ 在 $|\mathcal{S}|<\infty$ 时是 $\Delta(\mathcal{S})$ 的一个元素。我们用 来表示信念状态空间，使得 $\mathcal{B}=\{b\colon\ \exists h\in\mathbb{N}\ \exists\tau_h^+,\mathsf{b}(\tau_h^+)=b\}$。考虑一种常见情况：此类 b 是双射，此时 $\mathcal{B}$ 就成为 $\mathcal{H}$ 的完美代理，其基数随时域呈指数增长。在无限时域情况下，$|\mathcal{B}|=\infty$，但考虑到 $\mathbb{R}^{|\mathcal{S}|}$ 的有界子集具有紧性，$\mathcal{B}$ 必定存在聚点。为简单起见，我们为具有相同信念状态的历史分配不同的信念副本

表 1：与现有覆盖准则的比较

准则	现有覆盖与 诅咒 视野/记忆		基于信念的覆盖 空间平滑性
贝尔曼误差最小化（例如 双重采样 ）			
覆盖定义 [Jiang 和 Xie， 2024]	$\left\|\frac{d^{\pi_e}(\tau_h,a)}{d^D(\tau_h,\ a)}\right\|_\infty$		$\left\|\frac{d^{\pi_e^\phi}(\phi(b)\ a)}{d^D(\phi(b)\ a)}\right\|_\infty$
覆盖最坏情况[1] 尺度	$\lvert\mathcal{B}\rvert = \Theta((\lvert\mathcal{O}\rvert\lvert\mathcal{A}\rvert)^H)$	$>$	覆盖$(\mathcal{B}, \Theta(n^{-1/2}))$[2]
处理能力 $H \to \infty$	✘: 无限		✓: 多项式保证，见例 1
未来依赖价值函数			
覆盖定义 [Zhang 和 Jiang, 2024]	上确界$_{h,V}$ $\sqrt{\frac{\mathbb{E}_{\pi_e}[(\mathcal{B}^{(\mathcal{S},\mathcal{H}_H)}V)(s_h,\ \tau_h)^2]}{\mathbb{E}_{\pi_b}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2]}}$		上确界$_{h,V}$ $\sqrt{\frac{\mathbb{E}_{\pi_e^\phi}[\ (\mathcal{B}^{(\mathcal{S},\mathcal{H}_T)}V)\ \ (s_h,\ \tau_{[h-T+1:h]})\ ^2]}{\mathbb{E}_{\pi_b^\phi}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2]}}$
L_2 信念覆盖 (独热信念) [Zhang 和 Jiang, 2024]	$\mathbb{E}_{\pi_b}\left[\left(\frac{d^{\pi_e}(s_h,\tau_h)}{d^{\pi_b}(s_h,\tau_h)}\right)^2\right]$	$\overset{>}{\text{定理 } 4}$	$\mathbb{E}_{\pi_b^\phi}\left[\left(\frac{d_\phi^{\pi_e^\phi}(s_h,\ \tau_{[h-T+1:h]})}{d_\phi^{\pi_b^\phi}(s_h,\ \tau_{[h-T+1:h]})}\right)^2\right]$
L_∞ 信念覆盖 (独热信念) [Zhang 和 Jiang， 2024]	$\left\|\frac{d^{\pi_e}(s_h,\tau_h)}{d^{\pi_b}(s_h,\ \tau_h)}\right\|_\infty$	$\overset{>}{\text{定理 } 5}$	$\left\|\frac{d_\phi^{\pi_e^\phi}(s_h,\tau_{[h-T+1:h]})}{d_\phi^{\pi_b^\phi}(s_h,\ \tau_{[h-T+1:h]})}\right\|_\infty$
L_∞ 最坏情况[1] (独热信念)	$\Theta((\lvert\mathcal{O}\rvert\lvert\mathcal{A}\rvert)^H)$	$>$	$\Theta((\lvert\mathcal{O}\rvert\lvert\mathcal{A}\rvert)^T)$
处理能力 $H \to \infty$[3]	✘: 无限		✓: 多项式保证，参见示例 2

分布，因此信念空间度量是一个伪度量。我们将关注的策略记为 $\tilde{\pi}(\tau_h^+)$ $=$ π $\mathcal{H}^+ \to \Delta\ \mathcal{A}$ $(\mathbf{b}(\tau_h^+))$: ( )，给定历史时，该策略用于采样一个动作。价值函数 $\tilde{V}$ (τ_h^+) $= V(\mathbf{b}(\tau_h^+))$ 也类似。由于 $\tilde{V}, \tilde{\pi}, \tau_h^+ \in \mathcal{H}^+$ 与 $V, \pi,$ $b \in \mathcal{B}$ 一一对应，我们稍微滥用一下记号，在本文余下部分将它们视为等价；也就是说，每当我们提到 $b \in \mathcal{B}$，也指相应的 $\mathbf{b}^{-1}(b) \in \mathcal{H}^+$，尤其是在描述算法时，因为算法只能看到数据轨迹，而非实际信念。

然后，我们介绍利用信念空间度量来降低可能呈指数级的信念空间复杂度的核心思想，即通过覆盖。通过引入一个 ε-覆盖作为原始信念空间的一种抽象，我们可以将相近的信念状态视为同一个，从而使该空间更为简单。下文将以与 [François-Lavet 等人，2019, Subramanian 等人， 2022] 中的 ε-充分统计量 类似的思想对此进行形式化。

[1]“最坏情况覆盖”指的是探索性最强的数据采集分布下的最坏情况。[2]覆盖$(\mathcal{B},\ \varepsilon)$ 表示 $\mathcal{B}$ 的 L_1 覆盖数。[3]对于 $H \to \infty$，我们假设最坏情况覆盖以次多项式幂 $\alpha_0 \le 1$ 增长（而非对数增长；对数增长会轻易消除时域诅咒）。在 FDVF 情况下，可能需要特定的遗忘率。

由覆盖诱导的抽象。 考虑信念空间 $\mathcal{B}$，对于任意 $\varepsilon > 0$ 以及一个 ε-覆盖 $\mathcal{C}_\varepsilon \subset \mathcal{B}$（定义见附录 B）。存在一个抽象 $\phi : \mathcal{B} \to \mathcal{C}_\varepsilon$，使得 $\forall b \in \mathcal{B}$, $\|\phi(b) - b\|_1 \le \varepsilon$。任选这样一个 ϕ，以及第 3 节中提到的一族测度 $\{p_x\}_{x \in \mathcal{C}_\varepsilon}$，即可定义一个抽象信念马尔可夫决策过程，我们将其称为抽象系统。

为了获得有意义的结果，我们需要将注意力限制在所有可能策略的一个子集上，即具有稳定性的策略。这由我们感兴趣的策略的两个核心结构性假设刻画，这些假设最初主要见于利普希茨MDP文献 [Pirotta等人, 2015，Gelada等人， 2019]:

假设 1（局部稳定性）。 $\forall b_1,\ b_2 \in \mathcal{B},\ \|\pi(b_1) - \pi(b_2)\|_1 \le L_\pi \|b_1 - b_2\|_1$。

假设 2（价值稳定性）。 上确界$_{b_1,b_2 \in \mathcal{B}} |V^{[\pi_{\phi_\varepsilon}]_{\text{true}}}{}_{\varepsilon \ge 0, \phi_\varepsilon}$ $(b_1) - V^{[\pi_{\phi_\varepsilon}]_{\text{真}}}(b_2)| / \|b_1 - b_2\|_1 \le L_V < \infty.$

注1. 假设 *1* 的提出基于这样的直觉：良好的信念状态策略应当以相似的方式处理两个相似的信念状态，因此策略本身也应具有一定的局部稳定性。假设 *2* 衡量的是策略长期回报的稳定性。如下文定理*1*所示，它也可以视为衡量策略与最优策略有多相似的一个替代指标。

定理1（[Lee 等人， 2007]中的引理1）. 对于任意$b_1, b_2 \in \mathcal{B}$， $|V^*(b_1) - V^*(b_2)| \leq \frac{R_{\max}}{1-\gamma}\|b_1 - b_2\|_1$.

这表明，最优价值函数是 $\frac{R_{\max}}{1-\gamma}$-价值稳定的。除了最优价值固有的稳定性之外，人们还研究了这种稳定性性质，不仅针对部分可观测马尔可夫决策过程，也针对连续状态马尔可夫决策过程。这一系列研究，例如 [Pirotta 等人，2015、Asadi等人，2018、Gelada等人，2019], 探讨了各种情境，但在这些情形下，价值稳定性并不像在部分可观测马尔可夫决策过程中那样自然，因为与信念空间不同，连续状态马尔可夫决策过程中的系统动态相对于其内在度量可能并不平滑。

一般而言，这两个假设对某些有限常数 L_π 和 L_V成立，但在最坏情况下，它们的规模可能呈指数级增长。然而，具有恶性稳定性的策略通常表现不佳，也缺乏研究价值，因此我们只研究那些良好策略的行为会更有效率。尽管如此，我们的分析适用于这两种情况，而且稳定性常数越小，我们的界就越容易处理。无论如何，我们给出的界不会比原来的界更差。

4.1 统一分析概览

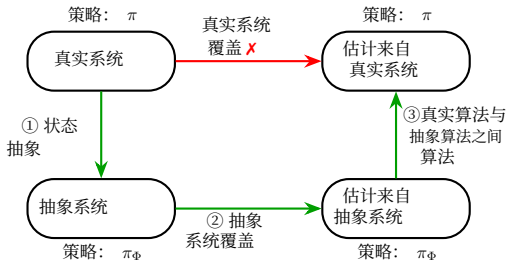

图1：分析流程

具体而言，如图1所示，在步骤1中，我们将真实信念空间马尔可夫决策过程系统（以及策略 π）降阶为抽象系统（以及抽象策略 π_ϕ）。借鉴状态抽象的类似思路，我们利用分箱大小 ε 控制抽象误差。在步骤2中，我们在抽象系统上运行算法，并采用针对抽象信念空间的覆盖假设。由于时域诅咒，这一假设可能比真实系统的覆盖假设更易处理。我们还通过定理4和定理5表明，抽象覆盖不会比原始覆盖更差。最后，在步骤3中，我们再次利用价值函数的稳定性，控制在同一份离线数据上实际运行的算法与虚拟运行的算法之间的差异。综合以上分析，我们得到了一个无需纳入传统覆盖假设的估计误差界。

本文利用信念空间的一个 aε-覆盖 $\mathcal{C}_\varepsilon$ 来构造抽象，其定义见附录B。为控制抽象误差，我们给出以下有用引理。

引理1. $\forall a \in \mathcal{A}, b_1, b_2 \in \mathcal{B}$, $\mathbb{E}_{o\sim P(\cdot|b_1,a)}[\|b_1^{o,a} - b^{o,a}_2\|_1] \leq 2\|b_1 - b_2\|_1$。这里， $b^{o,a}$ 表示采取动作 a 并观测到 o之后， b 的更新后下一个信念。

注2. 直观而言，在经历一组相同的动作和观测(a, o)后，两个信念状态共享同一状态的可能性会增大，从而使两个后续信念状态更加接近，即应满足数据处理不等式（*DPI*）。然而，正如例*3*所示，这种收缩性质通常并非逐点成立，而且 *Lipschitz* 值也可能趋于无穷大。其原因在于，信念更新是后验分布，而非马尔可夫核；数据处理不等式仅适用于后者。不过，引理表明，取期望后，该值有界于2。证明见附录 *A*。**命题 1.** 若对任意 $\tau_h \in \mathcal{H}$， $\mathbf{b}(\tau_h)$ 均为独热向量，则引理*1*中的2可替换为1。

定理 2. 在假设 *1* 和 *2* 下，对于一个依赖于 ϕ_ε ε 的抽象，我们有 $\|V^\pi_{\text{true}} - [V^{\pi_\phi}_{\text{分箱}}]_{\text{true}}\|_\infty \leq L^{[1]}_\phi \varepsilon$，其中 $L^{[1]}_\phi := \frac{(L_\pi+1)R_{\max}+2L_V}{1-\gamma} + \frac{\gamma R_{\max} L_\pi + R_{\max}(1-\gamma)^2}{(1-\gamma)^2}$，且 $\mathcal{V}$ 是用于函数估计的函数类。证明见附录 *B*。

备注 3. 对于有限时域部分可观测马尔可夫决策过程，只需将$(1-\gamma)^{-1}$ 替换为 H。

因此，我们之前的假设能够以原则性方式，将规模呈指数级增长的信念马尔可夫决策过程简化为一个抽象信念马尔可夫决策过程，同时保持近似误差可处理。抽象状态空间的基数与覆盖数同阶，这可能缓解时域诅咒，如例 1 和例 2 所示。形式上，我们有如下元定理，证明见附录 A。

定理 3 (元定理)。 对于部分可观测马尔可夫决策过程 $\mathcal{M}$、策略 π 和离线策略评估算法 $\text{Alg} := \{\mathfrak{alg}, \mathfrak{est}\}$， 其中 $\hat{}\mathfrak{alg}$: $\mathcal{D} \to \mathcal{V}$ 利用大小为 n 的离线数据集 $\mathcal{D}$ 学习函数 $Q^\pi \in \mathcal{V}$， 而 $\mathfrak{est} : \mathcal{V} \to \mathbb{R}$ 则根据 Q^π估计 π 的期望奖励。为简洁起见，记号中省略了Alg对 π 的依赖，以及 $\mathfrak{est}$对 $\mathcal{M}$的依赖。

那么，对于任意 $\varepsilon \geq 0$ 和一个抽象 $\phi : \mathcal{B} \to \mathcal{B}$，使得 $\forall\, b_1, b_2 \in \mathcal{B}$, $\phi(b_1) = \phi(b_2) \Rightarrow \|b_1 - b_2\|_1 \leq \varepsilon$，我们将抽象系统上执行的算法记为 $Alg^\phi := \{\mathfrak{alg}^\phi, \mathfrak{est}^\phi\}$。如果假设 *1*、*2* 成立，并且存在一个 $L^{[2]}_\phi$， 使得对于所有 $Q \in \mathcal{V}$，均有 $|\mathfrak{est}(Q) - \mathfrak{est}^\phi(Q)| \leq L^{[2]}_\phi \varepsilon$ ，我们还考虑 *Alg* 在抽象系统上的有限样本估计误差满足如下形式的情况： $|\mathfrak{est}^\phi(\hat{}Q^\pi) - \mathfrak{est}^\phi(Q^\pi_\phi)| \leq C^\phi_\pi \cdot \sqrt{\|\mathcal{V}\|_\infty \cdot \left(\frac{1}{n}\log\frac{|\mathcal{V}|}{\delta}\right)^\alpha} + L_\mathcal{E}\varepsilon$, $w.p. > 1-\delta$, *where* C^ϕ_π

是一个常数，$\|\mathcal{V}\|_\infty$, $|V|$ 分别表示用于函数ˆ近似的函数类的有界性和基数。那么我们有 $|\mathfrak{est}(Q^\pi) - \mathfrak{est}(Q^\pi)| \leq L_\phi + C_\pi^\phi \cdot \sqrt{\|\mathcal{V}\|_\infty \cdot \left(\frac{1}{n}\text{对数}\frac{|\mathcal{V}|}{\delta}\right)} + L_\mathcal{E}\varepsilon$,

$w.p. > 1-\delta$。这里，$L_\phi := L_\phi^{[1]} + L_\phi^{[2]}$，其中 $L_\phi^{[1]}$ 在定理 *2* 中定义；Q^π 和 Q_ϕ^π 表示真实估计量。

4.2 为什么覆盖比原始覆盖更好?

在接下来的部分中，我们通过给出两个定理作为表 1 的补充，展示为何我们的覆盖效果不逊于原始覆盖。由于直接比较π_e 的占用度与 π_e^ϕ 的抽象占用度较为困难，因此我们转而比较 $[\pi_e^\phi]_{\text{true}} := \tau_h \mapsto \pi_e^\phi(\phi(\tau_h))$的占用度，其缩放通常与 π_e相同。证明这些定理（见附录 E ）采用了一种信息论思路：在更粗的 σ- 代数上，两个概率测度之间的散度会变小，并使用 f-散度的变分表示。

定理 4. 考虑独热情形下的 L_2 信念覆盖。那么，对于任意行为策略 π_b和截断抽象 ϕ_T，都存在一个 $d_\phi^D \in \Delta(\mathcal{S}^\phi \times \mathcal{H}_T)$，使得对于任意 π_e，连同其抽象策略 π_e^ϕ 及对应的提升版本 $[\pi_e^\phi]_{\text{true}}$，我们有 $\mathbb{E}_{d_\phi^D}\left[\left(\frac{d_\phi^{\pi_e^\phi}}{d_\phi^D}(s_h, \tau_{[h-T+1:h]})\right)^2\right] \leq \mathbb{E}_{\pi_b}\left[\left(\frac{d^{[\pi_e^\phi]\text{真}}(s_h, \tau_h)}{d^{\pi_b}(s_h, \tau_h)}\right)^2\right]$.**定理 5.** 对于 L_∞ 信念覆盖，结论相同

$$that \left\|\frac{d_\phi^{\pi_e^\phi}(s_h, \tau_{[h-T+1:h]})}{d_\phi^D(s_h, \tau_{[h-T+1:h]})}\right\|_\infty \leq \left\|\frac{d^{[\pi_e^\phi]_{\text{true}}}(s_h, \tau_h)}{d^{\pi_b}(s_h, \tau_h)}\right\|_\infty.$$

接下来，我们提供说明性示例，以展示我们的结果在某些结构下的优越性。

例 1. 考虑具有光滑性结构的信念空间 *[*详细定义见附录 *E]*。在覆盖相对于最坏情况为次线性多项式时，我们得到有限样本保证 $O((C|\mathcal{S}|L_\mathcal{E} m R^2_{\max})^{\frac{1}{4}}(1-\gamma)^{\frac{3}{2}} \cdot (\frac{1}{n}\log\frac{|\mathcal{F}|}{\delta})^{\frac{1}{8}})$，其中 C、m 是与光滑性性质相关的常数。

示例 2. 考虑遗忘速度为 $T(\varepsilon) = O(\log\frac{1}{\varepsilon})$的快速遗忘策略，则在覆盖相对于最坏情况为次线性多项式时，我们可以获得有限样本保证 $O(\frac{\max\{\|\mathcal{V}\|_\infty, \|\Theta\|_\infty\}}{(1-\gamma)^2} \cdot (\frac{1}{n}\text{对数}\frac{|\mathcal{V}||\Theta|}{\delta})^{\frac{1}{4}})$。如果我们采用比记忆随对数增长更强的假设，即严格的短期记忆，那么结果就回到了 *[Uehara* 等人，*2023*，*Zhang* 和 *Jiang,2024]* 所讨论的内容。

5 应用示例

本节将统一分析应用于两种不同类型的离线策略评估算法，即双重采样贝尔曼误差最小化算法和未来依赖价值函数，以期获得更具样本效率的保证。

5.1 贝尔曼误差最小化算法分析

双重采样。考虑一种使用双重采样的贝尔曼误差最小化算法，每条离线数据包含两个元组(b, a, r, b'_A)和(b, a, r, b'_B)，后者在系统重置到信念 b后独立采样。相应的估计量可以ˆ写为 $Q^\pi = \arg\min_{f\in\mathcal{F}} \mathcal{E}(f, \pi)$，其中 $\mathcal{E}(f, \pi) = \mathbb{E}_\mathcal{D}[(f(b, a) - (r + \gamma f(b'_A, \pi)))(f(b, a) - (r + \gamma f(b'_B, \pi)))]$

我们不再假设真实系统满足标准覆盖，而是在抽象系统上采用以下抽象覆盖假设。

假设 3（抽象策略覆盖）。$\|d^{\pi_\phi}/d^D\|_\infty \leq C_\pi(\phi) < \infty$

注 4。值得注意的是，此处的覆盖 $C_\pi(\phi)$取决于特定的抽象映射 ϕ。在最具探索性的探索性数据收集分布 d^D下，$C_\pi(\phi)$的最坏情况增长率大致与 $|\mathcal{C}_\varepsilon|$一致，后者表示 ε-覆盖数。信念策略覆盖假设 *3* 的优势在于，它有可能优于原始空间中的覆盖假设。采用抽象信念空间，可以将规模呈指数级增长的历史空间缩减为大小为 ε-覆盖数的空间。

此外，还采用标准的可实现性假设。

假设 4（抽象可实现性）。$Q_\phi^{\pi_\phi} \in \mathcal{F}$，按照我们的记号，这是 $\exists f \in \mathcal{F}$、$f(\phi(\cdot)) = Q_\phi^{\pi_\phi}(\phi(\cdot))$的简写，因为 $Q_\phi^{\pi_\phi}$是在抽象系统上定义的。

注意到，我们此前假设了价值函数的稳定性，而很容易证明，该假设等价于 Q-函数在动作 a 处的利普希茨连续性。现在我们进一步假设，用于近似 Q-函数的函数类 $\mathcal{F}$ 关于信念状态也是利普希茨的。

假设5（利普希茨函数类）。$\forall f \in \mathcal{F}$，$\forall a \in \mathcal{A}$，$|f(b_1, a) - f(b_2, a)| \leq L_Q\|b_1 - b_2\|_1$。

然后，我们可以给出此特殊情形下定理 3 中定义的 $L_\phi^{[2]}$ 的取值，并进一步给出基于上述假设和分析方法的双重采样算法的最终保证。证明见附录C。

定理 6. 若假设*5*成立，则对于定理 *3* 中定义的 $L_\phi^{[2]}$，$L_\phi^{[2]} = \frac{R_{\max}}{1-\gamma} + L_Q$。

定理 7. 如果假设 *4*、*5*、*1* 和 *2* 均成立，则我们有：

$$|J_{\hat{Q}^\pi}(\pi) - J(\pi)| \le \inf_{\substack{\varepsilon \ge 0 \\ D(\varepsilon)}} \left(\frac{\sqrt{C_\pi(\varepsilon)}}{1-\gamma} \cdot \sqrt{\sqrt{\frac{32R_{\max}^4}{n(1-\gamma)^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}} + L_{\mathcal{E}}\varepsilon} + L_\phi \varepsilon \right)$$

其中 $L_{\mathcal{E}} = \frac{8R_{\max}}{1-\gamma}\big((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\big)L_\phi$ 的定义与定理 *3* 中相同，而 $D(\varepsilon)$ 表示满足可实现性（假设 *4*）的 ε。

推论 1 (有限样本保证)**.** 如果假设 *5*、*1* 和 *2* 均成立，则对于所有满足以下条件的 n

$n \ge 8R_{\max}^4 (L_\phi / L_{\mathcal{E}})^4 \log(2|\mathcal{F}|/\delta)$，并且由 $\varepsilon = \frac{1}{L_{\mathcal{E}}}\sqrt{\frac{32R_{\max}^4}{n\,1-\gamma\,()^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}}$ 的 ε-覆盖诱导的抽象 ϕ 满足假设 *4*，我们有 $|J_{\hat{Q}^\pi}(\pi) - J(\pi)| \le \frac{2\sqrt{C_\pi^n}}{1-\gamma} \cdot \big(\frac{128R^4}{n\,1-\gamma^{\max}()^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}\big)^{\frac{1}{4}}$，其中 $C_\pi^n := C_\pi\big(\frac{1}{L_{\mathcal{E}}}\sqrt{\frac{32R_{\max}^4}{n\,1-\gamma()^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}}\big)$。

注 5. 我们的方法所获得的保证依赖于抽象系统上的覆盖，正如注 *4* 和表 *1* 中所讨论的，这比原始覆盖更易于处理。此外，在适当的信念空间平滑性条件下（例 *1*），我们的结果提供了多项式有限样本保证，而原始界则趋于无穷大。

5.2 未来依赖价值函数 on.

FDVF 最初是针对无记忆策略提出的。这里我们介绍 FDVF 的基于记忆的版本，正如 [Zhang 和 Jiang, 2024] 中所讨论的，它受到“记忆诅咒”的影响。我们首先介绍未来空间的相应定义 $\mathcal{F}'$ 如下$f_h' := (o_h, a_h, o_{h+1}, a_{h+1}, \cdots, o_H, a_H) \in \mathcal{F}_h' \subset \mathcal{F}'$.

从这里开始，为方便起见，我们将 (f_h', τ_h) 简记为 f_h。类似地，我们将 $\mathcal{F}'$ 视为原始未来空间，并定义 $\mathcal{F} := \mathcal{F}' \times \mathcal{H}$为“（未来-历史）对”的新空间。这是因为 τ_h 可以视为扩展未来的一部分；或者等价地，对于每个历史序列，未来都会被单独复制。未来依赖价值函数 $V_{\mathcal{F}}$ 是满足$\mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_h, \tau_h)|s_h, \tau_h] = V_{\mathcal{S}}^{\pi_e}(s_h, \tau_h)$ 的任意此类函数，其中右侧为 π_e的价值函数；它是以下两个贝尔曼残差算子的零点。

定义 1（基于记忆的贝尔曼残差算子）。我们定义 $(\mathcal{B}^{(\mathcal{S},\mathcal{H}_T)}V)(s_h, \tau_{[h-T+1:h]}) := \mathbb{E}_{\substack{a_{1:h}\sim\pi_e \\ a_{h+1:H}\sim\pi_b}}[r_h + V(f_{h+1})|s_h, \tau_{[h-T+1:h]}] - \mathbb{E}_{\substack{a_{1:H-1}\sim\pi_e \\ a_{h:H}\sim\pi_b}}[V(f_h)|s_h, \tau_{[h-T+1:h]}]$，以及 $(\mathcal{B}^{\mathcal{H}}V)(\tau_h) := \mathbb{E}_{\substack{a_{1:h}\sim\pi_e \\ a_{h+1:H}\sim\pi_b}}[r_h + V(f_{h+1})|\tau_h] - \mathbb{E}_{\substack{a_{1:h-1}\sim\pi_e \\ a_{h:H}\sim\pi_b}}[V(f_h)|\tau_h]$。

基于记忆的算法。对于基于记忆的策略，我们定义 $\mu(a_h, \tau_h^+) := \frac{\pi_e(a_h|\tau_h^+)}{\pi_b(a_h|\tau_h^+)}$，然后极小极大算法定义如下：

$$\hat{V}_{\mathcal{F}} = \arg\min_{V\in\mathcal{V}} \max_{\theta\in\Theta} \sum_{h=1}^{H} \mathbb{E}_{\mathcal{D}}[\{\mu(a_h, \tau_h^+)(r_h + V(f_{h+1})) - V(f_h)\}\theta(\tau_h) - \frac{1}{2}\theta(\tau_h)^2]$$

FDVF 分析流程。FDVF 的分析遵循一个结构化框架，该框架使用先前介绍的方法；其示意图见图 2，可在附录 D 中找到。所有证明也可在附录 D 中找到。

截断映射诱导的抽象。该方法的第一步是引入抽象映射 $\tilde{\phi}$： $\mathcal{H}^+ \to \mathcal{H}_T^+$，其中 T 是时间窗口，$\mathcal{H}_T^+ := \bigcup_{t=1}^{T}(\mathcal{O}\times\mathcal{A})^{t-1}\times\mathcal{O}$ 表示受窗口 T 约束的历史序列集合。

$$\tilde{\phi}(o_1, a_1, \ldots, o_h) := \begin{cases} (o_{h-T+1}, a_{h-T+1}, \ldots, o_h), & h \ge T \\ \mathrm{id}, & h < T \end{cases}$$

为了复用先前的分析，我们还引入一个直接作用于信念状态的抽象映射 ϕ： $\mathcal{B} \to \mathcal{B}$。映射 ϕ 满足 $\phi(\mathbf{b}(\tau_h^+)) = \mathbf{b}(\tilde{\phi}(\tau_h^+))$。由于该映射 ϕ 依赖于时间窗口长度 T，我们将其记为 ϕ_T。注意到 ϕ_T 与 ϕ_T 是一一对应的，我们将它们视为等价。现在我们给出部分可观测马尔可夫决策过程与策略的快速遗忘假设。

假设 6 (快速遗忘 POMDP)。对于上述抽象映射 ϕ_T，以下条件成立：对于所有 $\varepsilon > 0$，存在 $T \in \mathbb{N}^+$，使得对于所有 $b_1, b_2 \in \mathcal{B}$，若 $\phi_T(b_1) = \phi_T(b_2)$，则 $\|b_1 - b_2\|_1 \le \varepsilon$。满足此条件的 T 的取值构成 ε的一个函数，记为 $T_1(\varepsilon)$。

假设 7 (快速遗忘策略)。对于抽象映射 ϕ_T，以下条件成立：对于所有 $\varepsilon > 0$，存在一个 $T \in \mathbb{N}^+$，使得对于所有 $\tau_h^{[1]+}, \tau_h^{[2]+} \in \mathcal{H}^+$以及所有 $\pi \in \pi_e, \pi_b$，若 $\tilde{\phi}_T(\tau_h^{[1]+}) = \tilde{\phi}_T(\tau_h^{[2]+})$，则$\|\pi(\tau_h^{[1]+}) - \pi(\tau_h^{[2]+})\|_1 \le L_\pi\varepsilon$。我们将 T 对 ε的依赖关系记为 $T_1(\varepsilon)$。

引理 2（稳定性蕴含快速遗忘）。如果假设 *6* 和假设 *1* 成立，那么假设 *7* 将自动成立，其中 $T_1 = T_0$。

条件：控制真实算法与抽象算法之间的差异。由于我们的分析基于这样一项要求：虚拟执行的算法与实际算法之间的差异很小，因此我们首先提出一些条件，以限制 ε 过大。

定义 2。我们定义 $\|\mathcal{V}\|_\infty := \max_{V\in\mathcal{V}} \|V\|_\infty$（$\Theta$ 类似），$C_\mathcal{V} := \max\{\|\mathcal{V}\|_\infty + 1, \|\Theta\|_\infty\}$, $C_\mu := \max_h \max_{a_h,\tau_h^+} \mu(a_h,\tau_h^+)$, 以及 $L_\mathcal{E} := 3\big(\frac{2H(C_\mu+1)L_\pi\|\mathcal{V}\|_\infty\|\Theta\|_\infty}{\min_h \min^{a_h,\tau_h^+}\pi_b(a_h|\tau_h^+)} + HC_\mu\|\mathcal{V}\|_\infty\|\Theta\|_\infty + \frac{3H^2\max\{C_\mu\|\mathcal{V}\|_\infty\|\Theta\|_\infty,\frac{1}{2}\|\Theta\|_\infty^2\}}{\min\{\min_h \min^{o_h,\tau_h} P(o_h|\tau_h), \min_h \min^{a_h,\tau_h^+}\pi_b(a_h|\tau_h^+)/L_\pi\}}\big)$.

条件 1. ε 足够小，使得 $L_\pi\varepsilon / \min_h \min^{a_h,\tau_h^+}\pi_b(a_h|\tau_h^+) \le \frac{1}{2}$。

条件2. ε 足够小，使得 $H\varepsilon \min_h \min^{o_h,\tau_h} P(o_h|\tau_h), \min_h \min^{a_h,\tau_h^+}\pi_b(a_h|\tau_h^+)/L_\pi\} \le 1$

条件 1 假设行为策略 π_b 的各项均非零，该策略为已知且由学习者选择。文献中也采用了这一假设，例如 [Zhang 和 Jiang, 2025], 我们沿用了相同的思路。在条件 2 中，概率 $P(o_h|\tau_h)$ 始终大于零是一个非平凡假设，但我们稍后会在一个更简单的流程中证明，实际上可以去掉这一条件。

接下来，我们陈述关于 $L_\phi^{[2]}$ 的以下定理。

定理 8。 若假设 *6*、*7* 和 *9* 成立，则对任意 $\varepsilon > 0$, 当 $T \ge \max\{T_0(\varepsilon)\ T_1(\varepsilon)\ T_2(\varepsilon)\}$ 时，我们有 $L_\phi^{[2]} = \|\mathcal{V}\|_\infty$，其中 $L_\phi^{[2]}$ 由定理 *3* 定义。

FDVF 的理论保证。 下述定理展示了统一分析下 FDVF 的保证；给定条件表明，我们选择的 ε 通常应与有限样本误差项具有相同的渐近阶。

条件 3。 对于某个一致常数 C，给定 ε、n、δ、$L_\mathcal{E}\varepsilon \le \frac{eCHC_\mathcal{V}^2C_\mu}{2n}\cdot\log\frac{4|\mathcal{V}||\Theta|}{\delta}$。

定理 9（FDVF 的理论保证）。 假设抽象可实现性条件 $V_\mathcal{F}^\phi \in \mathcal{V}$ 和贝尔曼完备性条件 $\forall V\in\mathcal{V}$、$\mathcal{B}^\mathcal{H}V \in \Theta$（此处的 $\mathcal{B}^\mathcal{H}$ 指抽象系统上的算子）成立，且假设 *2*、*6*、*7* 和 *9* 均满足。对于满足条件 *1*、*2*、*3* 的任意 $\varepsilon > 0$，定义 $T = \max\{T_0(\varepsilon)\ T_1(\varepsilon)\ T_2(\varepsilon)\}$。那么，对于某个一致常数 c，以至少 $1-\delta$ 的概率，我们有：

$$|J(\pi_e) - \mathbb{E}_{\pi_b}[\hat{V}(f_1)]| \le L_\phi\varepsilon + \sqrt{H}\cdot \max_{h\in[H]}\sup_{V\in\mathcal{V}}\sqrt{\frac{\mathbb{E}_{\pi_e^\phi}[(\mathcal{B}^{(\mathcal{S},\mathcal{H}_T)}V)(s_h,\tau_{[h-T+1:h]})^2]}{\mathbb{E}_{\pi_b^\phi}[(\mathcal{B}^\mathcal{H}V)(\tau_h)^2]}}\cdot \sqrt{\frac{cHC_\mathcal{V}^2C_\mu}{n}\log\frac{|\mathcal{V}||\Theta|}{\delta} + L_\mathcal{E}\varepsilon}$$

推论 2（增强的有限样本保证）。 对于足够大的 n，在满足必要的可实现性和完备性条件时，我们可以得到有限样本保证：

$$|J(\pi_e) - \mathbb{E}_{\pi_b}[\hat{V}(f_1)]| \le \sqrt{H}\cdot\sqrt{\frac{cHC_\mathcal{V}^2C_\mu}{n}\log\frac{|\mathcal{V}||\Theta|}{\delta}}.$$

更简单的流程：仅对策略进行抽象。 重新审视上述分析会发现，其中有一步实际上并无必要，即将原始部分可观测马尔可夫决策过程抽象为短期记忆部分可观测马尔可夫决策过程。这是因为策略对记忆的依赖才是“记忆诅咒”的真正根源。值得注意的是，引入假设 2 和假设 6 都是为了界定部分可观测马尔可夫决策过程本身的抽象误差，因此对于 FDVF 可以省去这些假设。**这表明，与历史即状态马尔可夫决策过程相比，FDVF 具有显著优势：“记忆诅咒”比“时域诅咒”容易处理得多，因为对于后者，抽象部分可观测马尔可夫决策过程本身是不可避免的**。当我们只对策略进行抽象时，先前的条件 1 和 2 可以放宽为条件 $2'$，以适用于 $H > 1$。

条件2′。 ε 足够小，因此 $HL_\pi\varepsilon / \min_h \min^{a_h,\tau_h^+}\pi_b(a_h|\tau_h^+) \le 1$。

定理 10（FDVF 的更紧理论保证）。假设抽象可实现性条件 $V_\mathcal{F}^\phi \in \mathcal{V}$ 和贝尔曼完备性条件 $\forall V\in\mathcal{V}$, $\mathcal{B}^\mathcal{H}V \in \Theta$（$\mathcal{B}^\mathcal{H}$ 此处指抽象系统上的算子）成立，且假设 *7* 和 *9* 满足。对于满足条件 *2′*, *3* 的任意 $\varepsilon > 0$，定义 $T = \text{最大值}\{T_1(\varepsilon)\ T_2(\varepsilon)\}$。那么，存在某个一致常数 c, c_1, c_2，使得以至少 $1-\delta$ 的概率有：

$$|J(\pi_e) - \mathbb{E}_{\pi_b}[\hat{V}(f_1)]| \le L_\phi\varepsilon + \sqrt{H}\cdot \max_{h\in[H]}\sup_{V\in\mathcal{V}}\sqrt{\frac{\mathbb{E}_{\pi_e^\phi}[(\mathcal{B}^{(\mathcal{S},\mathcal{H}_T)}V)(s_h,\tau_{[h-T+1:h]})^2]}{\mathbb{E}_{\pi_b^\phi}[(\mathcal{B}^\mathcal{H}V)(\tau_h)^2]}}\cdot \sqrt{\frac{cHC_\mathcal{V}^2C_\mu}{n}\log\frac{|\mathcal{V}||\Theta|}{\delta} + L_\mathcal{E}\varepsilon}$$

其中 $C_\mathcal{V} := \text{最大值}\{\|\mathcal{V}\|_\infty + 1, \|\Theta\|_\infty\}$, $L_\phi = R_{\max}HL_\pi + R_{\max}H^2L_\pi + \|\mathcal{V}\|_\infty$ 且 $L_\varepsilon = 3\big(\frac{HL_\pi(c_1(C_\mu+1)\|\mathcal{V}\|_\infty\|\Theta\|_\infty + c_2H\text{最大值}\{C_\mu\|\mathcal{V}\|_\infty\|\Theta\|_\infty,\frac{1}{2}\|\Theta\|_\infty^2\})}{\text{最小值}_h \text{最小值}^{a_h,\tau_h^+}\pi_b(a_h|\tau_h^+)} + HC_\mu\|\mathcal{V}\|_\infty\|\Theta\|_\infty\big)$

备注 6. 我们的结果中的覆盖只考虑窗口 T 内的历史，而非整个时域 H；定理 *4*、*5* 证明，在信念独热情形下，L_2 和 L_∞ 的信念覆盖不劣于原始覆盖。示例 *2* 还表明，该方法具有多项式有限样本保证，而原始界限并不存在，从而有效缓解了记忆诅咒。尽管定理 *9* 采用了关于 *POMDP*模型的结构性假设，但可以通过采用更简单的流程（即定理 *10*）来避免这一假设；该流程能给出更好的结果，体现了将记忆长度转化为时域长度在可处理性方面的优势。

参考文献

[Antos 等人， 2008] Antos, A., Szepesvári, C., and Munos, R.（2008）。基于贝尔曼残差最小化的拟合策略迭代与单条样本路径学习近似最优策略。《机器学习》，71:89–129。

[Asadi 等人， 2018] Asadi, K.、Misra, D. 和 Littman, M.（2018年）。基于模型的强化学习中的 Lipschitz 连续性。载于国际机器学习会议，第 264–273 页。PMLR。

[Atrash 等人， 2009] Atrash, A., Kaplow, R., Villemure, J., West, R., Yamani, H., and Pineau, J. (2009年). 开发并验证一种鲁棒语音接口，以改善人机交互。《国际社会机器人学期刊》, 1:345–356.[Baird et al., 1995] Baird, L. et al. (1995年). 残差算法：采用函数逼近的强化学习。载于第十二届国际机器学习会议论文集, 第 30–37 页。[Chen and Jiang, 2019] Chen, J. and Jiang, N. (2019). 批量强化学习中的信息论考量。载于国际机器学习会议, 第 1042–1051 页。PMLR。[Ernst et al., 2005] Ernst, D., Geurts, P., and We- henkel, L. (2005年). 基于树的批量强化学习。《机器学习研究期刊》, 6.[Feng et al., 2019] Feng, Y., Li, L., and Liu, Q. (2019). 用于求解贝尔曼方程的核损失。神经信息处理系统进展, 32.[François-Lavet等人， 2019] François-Lavet, V., Rabusseau, G., Pineau, J., Ernst, D., and Fonteneau, R. (2019). 关于部分可观测条件下批量强化学习中的过拟合与渐近偏差。《人工智能研究期刊》，65:1–30.[Gelada等人， 2019] Gelada, C., Kumar, S., Buckman, J., Nachum, O., and Bellemare, M. G. (2019). Deep-mdp：学习用于表示学习的连续潜在空间模型。载于国际机器学习会议, 第 2170–2179 页。PMLR。[Hao 等人， 2024] Hao, M., Su, P., Hu, L., Szabó, Z., Zhao, Q., and Shi, C. (2024). 基于深度抽象状态的离策略评估。*arXiv* 预印本 *arXiv:2406.19531*。

[Hu and Wager，2023] Hu, Y. and Wager, S. (2023). 在序贯可忽略性假设下，部分可观测马尔可夫决策过程中的离策略评估。《统计年刊》, 51(4):1561–1585。

[Jiang 等人， 2019] Jiang, B.、Song, R.、Li, J. 和 Zeng, D.（2019）。动态过程治疗方案的熵学习。《中华统计学志》，29(4):1633。

[Jiang 和 Li， 2016] Jiang, N. 和 Li, L.（2016）。强化学习的双重稳健离策略价值评估。载于国际机器学习会议，第 652–661 页。PMLR。

[Jiang 和 Xie， 2024] Jiang, N. 和 Xie, T.（2024）。大状态空间中的离线强化学习：算法与保证。《统计科学》。

[Kurniawati 等人， 2008] Kurniawati、Hsu 和 Lee（2008）。SARSOP：通过近似最优可达信念空间进行高效的基于点的部分可观测马尔可夫决策过程规划。载于机器人学：科学与系统，第 2008 卷。Citeseer。

[Lauri 等人， 2022] Lauri、Hsu 和 Pajarinen（2022）。机器人学中的部分可观测马尔可夫决策过程：综述。《*IEEE*机器人学汇刊》，39(1)：21–40。

[Le 等人， 2019] Le、Voloshin 和 Yue（2019）。约束条件下的批量策略学习。载于国际机器学习会议，第 3703–3712 页。PMLR。

[Lee 等人， 2007] Lee、Rong 和 Hsu（2007）。是什么让某些部分可观测马尔可夫决策过程问题易于近似？神经信息处理系统进展，20。

[Li 等人， 2006] Li, L., Walsh, T. J., 和 Littman, M. L. (2006). 面向马尔可夫决策过程状态抽象的统一理论。*AI&M*, 1(2):3.

[Munos 和 Szepesvári，2008] Munos, R. 和 Szepesvári, C. (2008). 拟合价值迭代的有限时间界。《机器学习研究期刊》, 9(5).

[Nachum 等人， 2019] Nachum, O., Chow, Y., Dai, B., 和 Li, L. (2019). Dualdice：与行为无关的折扣平稳分布校正估计。神经信息处理系统进展, 32.

[Pirotta等人， 2015] Pirotta、M. Restelli 和 L. Bascetta（2015）。Lipschitz 马尔可夫决策过程中的策略梯度。《机器学习》，100(2)：255–283。

[Poupart等人， 2011] P. Poupart、K.-E. Kim 和 D. Kim（2011）。缩小差距：改进部分可观测马尔可夫决策过程最优解的界。在国际自动规划与调度会议论文集，第21卷，第194–201页。

[Precup 等人， 2000] Precup, D., Sutton, R. S., 和 Singh, S. (2000). 用于离策略策略评估的资格迹。载于国际机器学习会议，第 2000 卷，第 759–766 页。Citeseer。

[Shani 等人， 2008] Shani, G., Brafman, R. I., 和 Shi- mony, S. E. (2008). 基于点的部分可观测马尔可夫决策过程求解器的优先级排序。*IEEE*系统、人和控制论汇刊*B*辑（控制论），38(6):1592–1605。[Shani 等人， 2013] Shani, G., Pineau, J., 和 Kaplow, R. (2013). 基于点的部分可观测马尔可夫决策过程求解器综述。自主智能体与多智能体系统，27:1–51。[Smith 和 Simmons，2012] Smith, T. 和 Simmons, R. (2012). 基于点的部分可观测马尔可夫决策过程算法：改进的分析与实现。*arXiv*预印本 *arXiv:1207.1412*。[Spaan 和 Vlassis， 2005] Spaan, M. T. 和 Vlassis, N. (2005). Perseus：针对部分可观测马尔可夫决策过程的随机化基于点的价值迭代。《人工智能研究期刊》，24:195–220。[Subramanian 等人， 2022] Subramanian, J., Sinha, A., Seraj, R., 和 Mahajan, A. (2022). 部分可观测系统中用于近似规划和强化学习的近似信息状态。《机器学习研究期刊》，23(12):1–83。[Uehara 等人， 2020] Uehara, M., Huang, J., 和 Jiang, N. (2020). 用于离策略评估的极小极大权重与Q函数学习。发表于国际机器学习会议，第9659–9668页。PMLR。[Uehara 等人， 2021] Uehara, M., Imaizumi, M., Jiang, N., Kallus, N., Sun, W., 和 Xie, T. (2021). 极小极大离线强化学习的有限样本分析：完备性、快速收敛率与一阶效率。*arXiv*预印本 *arXiv:2102.02981*。[Uehara 等人， 2023] Uehara, M., Kiyohara, H., Ben- nett, A., Chernozhukov, V., Jiang, N., Kallus, N., Shi, C., 和 Sun, W. (2023). 部分可观测马尔可夫决策过程中依赖未来的基于价值的离策略评估。神经信息处理系统进展，36:15991–16008。[Zanette 和 Wainwright， 2022] Zanette, A. 和 Wainwright, M. J. (2022). 用于离线强化学习的贝尔曼残差正交化。神经信息处理系统进展，35:3137–3151。

[Zhang 和 Jiang， 2024] Zhang, Y. 和 Jiang, N. (2024). 关于未来诅咒以及用于离策略评估的未来与历史依赖价值函数的历史诅咒。*arXiv* 预印本 *arXiv:2402.14703*。

[Zhang 和 Jiang，2025] Zhang, Y. 和 Jiang, N. (2025). 部分可观测马尔可夫决策过程（POMDP）中历史依赖策略离策略评估的统计可处理性。*arXiv* 预印本 *arXiv:2503.01134*。

[Zhang 等人， 2014] Zhang, Z., Hsu, D., 和 Lee, W.S. (2014). 用于高效基于启发式的部分可观测马尔可夫决策过程规划的覆盖数。载于国际机器学习会议，第 28–36 页。PMLR。

[Zhang 等人， 2012] Zhang, Z.、Littman, M. 和 Chen, X.（2012）。将覆盖数作为部分可观测马尔可夫决策过程规划与学习的复杂度度量。载于*AAAI*人工智能大会论文集，第 26 卷，第 1853–1859 页。

6 检查清单

1. 对于所介绍的所有模型和算法，请检查是否包含：

   (a) 对数学设定、假设、算法和/或模型的清晰描述。[是](b) 对任何算法的性质与复杂度（时间、空间、样本量）的分析。[是](c) （可选）匿名源代码，并说明所有依赖项，包括外部库。 [不适用]

2. 对于任何理论性主张，请检查是否包含：

   (a) 所有理论结果的全部假设陈述。 [是](b) 所有理论结果的完整证明。[是](c) 对任何假设的清晰解释。 [是]

3. 对于呈现实证结果的所有图和表，请检查是否包含：

   (a) 复现主要实验结果所需的代码、数据和说明（位于补充材料中或以 URL 形式提供）。[不适用](b) 所有训练细节（例如，数据划分、超参数及其选择方式）。 [不适用](c) 对具体指标或统计数据以及误差棒作出清晰定义（例如，多次运行实验后随机种子对结果的影响）。 [不适用](d) 描述所使用的计算基础设施（例如 GPU 类型、内部集群或云服务提供商）。 [不适用]

4. 如果使用了现有资产（例如代码、数据、模型），或整理/发布了新资产，请检查是否包含以下内容：

   (a) 若你的工作使用了现有资产，则注明创作者的引用。 [不适用](b) 资产的许可信息（如适用）。 [不适用](c) 新资产（如适用），可放在补充材料中或以 URL 形式提供。 [不适用](d) 关于数据提供者/管理者同意的信息。 [不适用](e) 如适用，讨论敏感内容，例如个人身份信息或冒犯性内容。 [不适用]

5. 如果你使用了众包或开展了涉及人类受试者的研究，请检查是否包含：

   (a) 提供给参与者的完整说明文本和屏幕截图。 [不适用](b) 对参与者潜在风险的说明，以及适用时机构审查委员会（IRB）批准的链接。 [不适用](c) 支付给参与者的预估时薪，以及用于参与者报酬的总金额。 [不适用]

A 统一分析

元定理3的证明

证明。 首先，由定理 2 可得，

$$\begin{aligned}|\mathfrak{est}(Q^\pi) - \mathfrak{est}^\phi(Q^\pi_\phi)| &= |\mathbb{E}_{b\sim d_0}[V^{\pi_\phi}_{\text{bin}}(\phi(b)) - V^\pi_{\text{true}}(b)]| \\ &\le \|[V^{\pi_\phi}_{\text{bin}}]_{\text{true}} - V^\pi_{\text{true}}\|_\infty \\ &\le L^{[1]}_\phi \varepsilon.\end{aligned}$$

然后利用三角不等式，我们得到

$$\begin{aligned}|\mathfrak{est}(\hat{Q}^\pi) - \mathfrak{est}^\phi(Q^\pi)| &\le |\mathfrak{est}(\hat{Q}^\pi) - \mathfrak{est}^\phi(\hat{Q}^\pi)| + |\mathfrak{est}^\phi(\hat{Q}^\pi) - \mathfrak{est}^\phi(Q^\pi_\phi)| + |\mathfrak{est}(Q^\pi) - \mathfrak{est}^\phi(Q^\pi_\phi)| \\ &\le L^{[1]}_\phi + L^{[2]}_\phi + C^\phi_\pi \cdot \sqrt{\|\mathcal{V}\|_\infty \cdot \left(\frac{1}{n}\log\frac{|\mathcal{V}|}{\delta}\right)^\alpha + L_\mathcal{E}\varepsilon}, \;\; w.p. > 1-\delta\end{aligned}$$

至此证明完成。 □

引理3. 如果对任意满足 $D(\varepsilon)$的 ε ， 下列结论成立

$$|J(\pi) - \hat{J}(\pi)| \le L_\phi\varepsilon + C^\phi_\pi \cdot \sqrt{\|\mathcal{V}\|_\infty \cdot \left(\frac{1}{n}\log\frac{|\mathcal{V}|}{\delta}\right)^\alpha + L_\mathcal{E}\varepsilon} \quad w.p. \ge 1-\delta. \tag{1}$$

Then

$$|J(\pi) - \hat{J}(\pi)| \le \inf_{\substack{\varepsilon\ge 0 \\ D(\varepsilon)}} \left(L_\phi\varepsilon + C^\phi_\pi \cdot \sqrt{\|\mathcal{V}\|_\infty \cdot \left(\frac{1}{n}\log\frac{|\mathcal{V}|}{\delta}\right)^\alpha + L_\mathcal{E}\varepsilon} \right) \quad w.p. \ge 1-\delta. \tag{2}$$

证明。令 $\beta(\varepsilon) := L_\phi\varepsilon + C^\phi_\pi \cdot \sqrt{\|\mathcal{V}\|_\infty \cdot \left(\frac{1}{n}\log\frac{|\mathcal{V}|}{\delta}\right)^\alpha + L_\mathcal{E}\varepsilon}$，且 $\beta^* := \inf_{D(\varepsilon)}^{\varepsilon\ge 0} \beta(\varepsilon)$。那么存在一个满足 $\varepsilon_i \ge 0$ 和 $D(\varepsilon_i)$ 的 $\{\varepsilon_i\}_{i=1}^\infty$ 序列，使得 $\beta(\varepsilon_i) \downarrow \beta^*$。于是，事件族 $\{E_i := \{\omega \in \Omega : |J(\pi) - J(\pi)|(\omega) \le \beta(\varepsilon_i)\}\}_{i=1}^\infty$ 是递减的，其极限为 $E_\infty := \{\omega \in \Omega : |J(\pi) - J(\pi)|(\omega) \le \beta^*\}$。因此，只需应用测度的单调收敛定理即可证明该结果；该定理表明 $\Pr(E_\infty) = \lim_{i\to\infty}\Pr(E_i) \ge 1-\delta$。

□

命题 1 的证明

证明。 这是因为对于任意 $b_1, b_2 \in \mathcal{B}, a \in \mathcal{A}, o \in \mathcal{O}$，若 $b_1 \ne b_2$, $b^o_{,1}{}^a$ 和 $b^o_{,2}{}^a$ 相同，则 $\frac{\|b^o_{,1}{}^a - b^o_{,2}{}^a\|_1}{\|b_1 - b_2\|_1} = 0$；若不同，则 $\frac{\|b^o_{,1}{}^a - b^o_{,2}{}^a\|_1}{\|b_1 - b_2\|_1} = 1$。 □

例子说明

例*3*。考虑一个只有一个动作 a、两个状态 s_1, s_2 和三个观测 o_1, o_2, o_3, o_4的潜在马尔可夫决策过程。初始状态在 s_1 和 s_2 上的分布均匀， s_1 的发射概率为 *(0.5, 0,* $0.5-\xi, \xi$*)*， s_2 的发射概率为 *(0, 0.5,* $\xi, 0.5-\xi$*)*。于是，对于两个信念状态 b_1 *:=* $b(o_2)$ *= (0, 1)* 和 b_2 *:=* $b(o_4)$ *=(*$2\xi, 1-2\xi$*)*，同时采取动作 a 并观测到 o_3 会使后续信念分别变为 $b^{o_3}_{,1}{}^a$ *= (0,1)* 和 $b^{o_3}_{,2}{}^a$ *= (0.5,0.5)*。这违反了收缩性质，因为对于任意 $\xi \le \frac{1}{4}$， $\|b^{o_3}_{,1}{}^a - b^{o_3}_{,2}{}^a\|_1 \ge \frac{1}{4\xi}\|b_1 - b_2\|_1$；这也表明利普希茨参数可以任意大，因为 $\xi \to 0$。

引理1的证明。

证明。 固定一个动作 $a \in \mathcal{A}$。对于信念 $b \in \mathcal{B}$，定义状态与观测上的联合分布：

$$P_b(s, o|a) := b(s)\, P(o|s, a), \qquad P_b(o|a) = \sum_s P_b(s, o|a).$$

当前状态的后验分布为

$$\tilde{b}^{o,a}(s) := P(s|b,o,a) = \frac{P_b(s,o|a)}{P_b(o|a)} = P_b(s|o,a).$$

观察到 o 后的下一个信念为

$$b^{o,a}(s') = \sum_s P(s'|s,a)\,\tilde{b}^{o,a}(s).$$

对于每个 $o \in \mathcal{O}$，全变差距离在状态转移核下会收缩 l:

$$\|b_1^{o,a} - b_2^{o,a}\|_1 = \|\tilde{b}_1^{o,a} P(\cdot|\cdot,a) - \tilde{b}_2^{o,a} P(\cdot|\cdot,a)\|_1 \le \|\tilde{b}_1^{o,a} - \tilde{b}_2^{o,a}\|_1.$$

对 $P(\cdot|b_1,\ a)$ 求期望，得到

$$\begin{aligned}\mathbb{E}_{o\sim P(\cdot|b_1,a)}[\|b_1^{o,a} - b_2^{o,a}\|_1] &\le \mathbb{E}_{o\sim P(\cdot|b_1,a)}[\|\tilde{b}_1^{o,a} - \tilde{b}_2^{o,a}\|_1] \\ &= \sum_o P_{b_1}(o|a) \sum_s \left|P(s|b_1,o,a) - P(s|b_2,o,a)\right| \\ &= \sum_{s,o} \left|P_{b_1}(s,o|a) - P_{b_1}(o|a)P_{b_2}(s|o,a)\right|.\end{aligned}$$

加上并减去 $P_{b_2}(s,\ o|a) = P_{b_2}(o|a)P_{b_2}(s|o,\ a)$，然后应用三角不等式 :

$$\le \sum_{s,o} \left|P_{b_1}(s,o|a) - P_{b_2}(s,o|a)\right| + \sum_{s,o} \left|P_{b_1}(o|a)P_{b_2}(s|o,a) - P_{b_2}(s,o|a)\right|.$$

这三个映射

$$k_1^a((o',s')|s) = \mathbb{I}(s=s')P(o'|s,a), \quad k_2^a((o',s')|o) = \mathbb{I}(o=o')P_{b_2}(s'|o,a), \quad k_3^a(o|s) = P(o|s,a)$$

其中， $\mathbb{I}(x=x')$ 是 $\mathcal{X}$ 上的示性函数，且它们是马尔可夫核。根据总变差的数据处理不等式，

$$\sum_{s,o} \left|P_{b_1}(s,o|a) - P_{b_2}(s,o|a)\right| = \left\| \sum_s k_1^a((\cdot,\cdot)|s)b_1(s) - \sum_s k_1^a((\cdot,\cdot)|s)b_2(s) \right\|_1 \le \|b_1 - b_2\|_1,$$

$$\begin{aligned}\sum_{s,o} \left|P_{b_1}(o|a)P_{b_2}(s|o,a) - P_{b_2}(s,o|a)\right| &= \left\| \sum_o k_2^a((\cdot,\cdot)|o)P_{b_1}(o|a) - \sum_o k_2^a((\cdot,\cdot)|o)P_{b_2}(o|a) \right\|_1 \\ &\le \|P_{b_1}(\cdot|a) - P_{b_2}(\cdot|a)\|_1 \\ &= \left\| \sum_s b_1(s)P(\cdot|s,a) - \sum_s b_2(s)P(\cdot|s,a) \right\|_1 \\ &= \left\| \sum_s k_3^a(\cdot|s)b_1(s) - \sum_s k_3^a(\cdot|s)b_2(s) \right\|_1 \\ &\le \|b_1 - b_2\|_1\end{aligned}$$

结合上述不等式可得

$$\mathbb{E}_{o\sim P(\cdot|b_1,a)}[\|b_1^{o,a} - b_2^{o,a}\|_1] \le 2\,\|b_1 - b_2\|_1.$$

根据对称性，当期望是相对于 $P(\cdot|b_2,\ a)$ 而非 $P(\cdot|b_1,\ a)$ 取值时，同样的界也成立。证明至此完成。

□

B 覆盖下的抽象

关于信念抽象误差上界的证明，即定理 11，采用了与 [Subramanian 等人， 2022] 中的定理 9 和命题 48 类似的思路。

定义 3. *A* ε-覆盖 $\mathcal{C}_\varepsilon$ 是信念状态空间的一个子空间，满足：

$$\mathcal{B} \subset \bigcup_{c \in \mathcal{C}_\varepsilon} \mathbf{B}(c, \varepsilon) \tag{3}$$

其中，$\mathbf{B}(c, \varepsilon)$ 表示以 c 为中心、半径为 ε的开球。$\mathcal{C}_\varepsilon$ 的基数称为 ε-覆盖数。对于每个 ε-覆盖 $\mathcal{C}_\varepsilon$，都存在信念状态空间的一个划分，其中每个 $c \in \mathcal{C}_\varepsilon$ 都作为该分箱的表示元素。

基于此，我们可以尝试刻画当两个信念状态足够接近时，某些重要量的表现。首先，以下引理给出了信念状态接近时，期望奖励之差的界。

引理 4. 对于两个信念状态 b_1 和 b_2， $\forall a \in \mathcal{A}$，有：

$$|r(b_1, a) - r(b_2, a)| \leq R_{\max} \|b_1 - b_2\|_1. \tag{4}$$

证明。 这可由以下内容直接得到：

$$\begin{aligned} |r(b_1, a) - r(b_2, a)| &= |\mathbb{E}_{s \sim b_1}[r(s, a)] - \mathbb{E}_{s \sim b_2}[r(s, a)]| \\ &= |\langle r(\cdot, a), b_1 - b_2 \rangle| \\ &\leq R_{\max} \|b_1 - b_2\|_1. \end{aligned}$$

这表明，当将部分可观测马尔可夫决策过程视为信念空间马尔可夫决策过程时，动态过程中存在内在平滑性。 □

引理 5。 （ *[Zhang* 等人的 *2012]* 中的引理 *2*）对于满足 $\|b_1 - b_2\|_1 \leq \varepsilon$，$\sum_o |P(o|b_1, a) - P(o|b_2, a)| \leq \|b_1 - b_2\|_1 \leq \varepsilon$ 的任意两个信念点 b_1，b_2 。

因此，我们提出以下命题。

命题 2。 对于满足假设 *1* 的策略 π ，对于 $\forall o, a$，有

$$\sum_{o,a} |P(o|b_1, a)\pi(a|b_1) - P(o|b_1, a)\pi(a|b_2)| \leq L_\pi \|b_1 - b_2\|_1. \tag{5}$$

证明。 这是数据处理不等式的直接应用。注意， $k^{b_1}((o', a')|a) = P(o'|b_1, a)\mathbb{I}(a = a')$ 是一个马尔可夫核，因此

$$\begin{aligned} \sum_{o,a} |P(o|b_1, a)\pi(a|b_1) - P(o|b_1, a)\pi(a|b_2)| &= \sum_{o', a'} \left| \sum_a k^{b_1}((o', a')|a)\pi(a|b_1) - \sum_a k^{b_1}((o', a')|a)\pi(a|b_2) \right| \\ &\leq \|\pi(\cdot|b_1) - \pi(\cdot|b_2)\|_1 \\ &\leq L_\pi \|b_1 - b_2\|_1 \end{aligned}$$

证明至此完成。 □

单步误差很容易控制；然而，若没有双模拟等与模型无关的状态抽象假设，就极难控制无限多个步骤所导致的累积误差。幸运的是，价值函数的稳定性为我们提供了另一种方法。

在抽象马尔可夫决策过程中，元组 $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ 经由抽象 ϕ 映射为 $(\mathcal{S}_\phi, \mathcal{A}, P_\phi, R_\phi, \gamma)$，这意味着抽象马尔可夫决策过程中的转移动力学 P_ϕ 由原始马尔可夫决策过程诱导而来。

具体而言，诱导出的 P_ϕ 满足：存在一族概率测度 $\{p_x\}_{x\in\mathcal{S}_\phi}$，其中每个 p_x 都定义在 $\phi^{-1}(x)$ 上，使得在抽象马尔可夫决策过程中，在动作 a下从 $\phi(s)$ 转移到 $\phi(s')$ 的转移概率，即 $P_\phi(\phi(s')|\phi(s)\ a)$，可以写为：

$$P_\phi(\phi(s')|\phi(s),a)=\mathbb{E}_{s\sim p_{\phi(s)}}[P_\phi(\phi(s')|s,a)] \tag{6}$$

由于对 P_ϕ 的这一刻画依赖于此类概率测度族的存在，却未指定它们的确切性质，因此，任何涉及价值函数 $V^{\pi_\phi}{}_{\text{分箱}}$ 的证明都必须将 $\{p_x\}_{x\in\mathcal{S}_\phi}$ 视为任意的。
基于这一理解，我们现在给出以下定理，它为 $V^{[\pi_\phi]_{\text{true}}}{}_{\text{真实}}$ 值与来自抽象马尔可夫决策过程的提升价值函数 $[V^{\pi_\phi}{}_{\text{分箱}}]_{\text{true}}$ 之间的误差提供了一个上界。值得注意的是，该定理的证明并不依赖于测度 $\{p_x\}_{x\in\mathcal{S}_\phi}$ 的具体形式。

定理 11. 若假设 *2* 成立，则 $V^{[\pi_\phi]_{\text{true}}}{}_{\text{真实}}$ 值与提升后的抽象马尔可夫决策过程的真实价值函数 $[V^{\pi_\phi}{}_{\text{分箱}}]_{\text{true}}$ 之间的误差可界定如下：

$$\|V^{[\pi_\phi]_{\text{true}}}_{\text{true}}-[V^{\pi_\phi}_{\text{bin}}]_{\text{true}}\|_\infty\le\frac{(R_{\max}+2L_V)\varepsilon}{1-\gamma}+\frac{R_{\max}}{(1-\gamma)^2}\varepsilon \tag{7}$$

证明。 我们先澄清并确立证明中使用的记号。固定一个任意的族 $\{p_x\}_{x\in\mathcal{S}_\phi}$，并令 $b'\sim\text{bin}(\phi(b))$ 表示对以下采样过程取的期望：

1. 由于 $\phi(b)\in\mathcal{B}_\phi$ 是分箱后信念空间某个划分的代表元素，集合$\phi^{-1}(\phi(b))\subset\mathcal{B}$ 则是原始信念空间中对应的元素——即由所有与 b 被归入同一分箱的信念状态构成的子集。

2. 采样　　根据固定分布 $p_{\phi(b)}$，从 $\phi^{-1}(\phi(b))$ 中采样一个临时信念状态 b_{temp}　　.

3. 从 b_{temp} 出发，执行信念更新过程，其中动作 a 由策略 π 决定。更新完成后，得到的信念状态就是采样得到的 b'。

确立了这些符号后，我们可以继续证明该定理。证明的主要思路是构造一个基于链式法则的论证。首先，注意到

$$[V^{\pi_\phi}_{\text{bin}}]_{\text{true}}(b)=\mathbb{E}_{\substack{b_1\sim\text{bin}(\phi(b))\\ b_2\sim\text{bin}(\phi(b_1))\\ \cdots}}[r_\phi(\phi(b_1),a_1)+\gamma r_\phi(\phi(b_2),a_2)+\gamma^2 r_\phi(\phi(b_3),a_3)+\cdots] \tag{8}$$

将 $V^{[k]}$ 视为

$$V^{[k]}=\mathbb{E}_{\substack{b_1\sim\text{bin}(\phi(b))\\ b_2\sim\text{bin}(\phi(b_1))\\ \cdots\\ b_k\sim\text{bin}(\phi(b_{k-1}))\\ b_{k+1}\sim b_k\\ \cdots}}[r_\phi(\phi(b_1),a_1)+\gamma r_\phi(\phi(b_2),a_2)+\gamma^2 r_\phi(\phi(b_3),a_3)+\cdots] \tag{9}$$

那么 $V^{[0]}(b)=V^{[\pi_\phi]_{\text{true}}}{}_{\text{为真}}(b)$。接下来，对于 $\forall b$,

$$|V^{[k+1]}(b)-V^{[k]}(b)| \tag{10}$$

$$=\left|\mathbb{E}_{\substack{b_1\sim\text{bin}(\phi(b))\\ b_2\sim\text{bin}(\phi(b_1))\\ \cdots\\ b_k\sim\text{bin}(\phi(b_{k-1}))\\ b_{k+1}\sim b_k\\ b_{k+2}\sim b_{k+1}\\ \cdots}}[\gamma^k V^{[\pi_\phi]_{\text{true}}}_{\text{true}}(b_{k+1})]-\mathbb{E}_{\substack{b_1\sim\text{bin}(\phi(b))\\ b_2\sim\text{bin}(\phi(b_1))\\ \cdots\\ b_k\sim\text{bin}(\phi(b_{k-1}))\\ b_{k+1}\sim\text{bin}(\phi(b_k))\\ b_{k+2}\sim b_{k+1}\\ \cdots}}[\gamma^k r_\phi(\phi(b_{k+1}),a)+\gamma^{k+1}V^{[\pi_\phi]_{\text{true}}}_{\text{true}}(b_{k+2})]\right| \tag{11}$$

$$=\left|\mathbb{E}_{\substack{b_1\sim\text{bin}(\phi(b))\\ b_2\sim\text{bin}(\phi(b_1))\\ \cdots\\ b_k\sim\text{bin}(\phi(b_{k-1}))\\ b_{k+1}\sim b_k\\ b_{k+2}\sim b_{k+1}\\ \cdots}}[\gamma^k V^{[\pi_\phi]_{\text{true}}}_{\text{true}}(b_{k+1})]-\mathbb{E}_{\substack{b_1\sim\text{bin}(\phi(b))\\ b_2\sim\text{bin}(\phi(b_1))\\ \cdots\\ b_k\sim\text{bin}(\phi(b_{k-1}))\\ b_{k+1}\sim\text{bin}(\phi(b_k))\\ b_{k+2}\sim b_{k+1}\\ \cdots}}[\gamma^k r_\phi(\phi(b_{k+1}),a)-\gamma^k r(b_{k+1},a)+\gamma^k V^{[\pi_\phi]_{\text{true}}}_{\text{true}}(b_{k+1})]\right| \tag{12}$$

$$\leq \gamma^k R_{\max}\varepsilon + \gamma^k 2L_V\varepsilon + \gamma^k \frac{R_{\max}}{1-\gamma}\varepsilon \tag{13}$$

其中，最后一个不等式用到了价值函数的稳定性（假设 2）、引理1 和引理 4、5，因为下一个信念是从同一分箱的起点采样得到的，因此足够接近。具体而言，它用到了以下事实:

$$\begin{aligned}
&\left|\mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ o\sim P(\cdot|b_k,a)}}[V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_k^{o,a})] - \mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ b_t\sim p_{\phi(b_k)}\\ o\sim P(\cdot|b_t,a)}}[V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_t^{o,a})]\right| \\
&= \left|\mathbb{E}_{b_t\sim p_{\phi(b_k)}}\left[\mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ o\sim P(\cdot|b_k,a)}}[V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_k^{o,a})] - \mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ o\sim P(\cdot|b_t,a)}}[V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_t^{o,a})]\right]\right| \\
&\leq \left|\mathbb{E}_{b_t\sim p_{\phi(b_k)}}\left[\mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ o\sim P(\cdot|b_k,a)}}[V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_k^{o,a})] - \mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ o\sim P(\cdot|b_k,a)}}[V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_t^{o,a})]\right]\right| \\
&\qquad + \left|\mathbb{E}_{b_t\sim p_{\phi(b_k)}}\left[\mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ o\sim P(\cdot|b_k,a)}}[V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_t^{o,a})] - \mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ o\sim P(\cdot|b_t,a)}}[V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_t^{o,a})]\right]\right| \\
&\leq \left|\mathbb{E}_{b_t\sim p_{\phi(b_k)}}\left[\mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ o\sim P(\cdot|b_k,a)}}[L_V\|b_k^{o,a} - b_t^{o,a}\|_1]\right]\right| \\
&\qquad + \left|\mathbb{E}_{\substack{b_t\sim p_{\phi(b_k)}\\ a\sim\pi(\phi(b_k))}}\left[\langle P(\cdot|b_k,a) - P(\cdot|b_t,a), V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_t^{\cdot,a})\rangle\right]\right| \\
&\leq 2L_V\varepsilon + \frac{R_{\max}}{1-\gamma}\varepsilon
\end{aligned}$$

最后，我们进行望远镜求和，并将所有 $V^{[k+1]} - V^{[k]}$ 相加，得到 $\forall b$,

$$|V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b) - [V_{\text{bin}}^{\pi_\phi}]_{\text{true}}(b)| \tag{14}$$

$$= \left|\sum_{k=0}^{\infty}\left(V^{[k+1]}(b) - V^{[k]}(b)\right)\right| \tag{15}$$

$$\leq \sum_{k=0}^{\infty}\left|\gamma^k R_{\max}\varepsilon + \gamma^k 2L_V\varepsilon + \gamma^k\frac{R_{\max}}{1-\gamma}\varepsilon\right| \tag{16}$$

$$\leq \frac{(R_{\max}+2L_V)\varepsilon}{1-\gamma} + \frac{R_{\max}}{(1-\gamma)^2}\varepsilon \tag{17}$$

□

在结束这一部分之前，我们还需要弥合目标策略与目标策略所源自的抽象策略之间的差距。以下定理将处理这一问题，它不依赖于对 POMDP 模型本身的任何假设。

定理 12. 若假设 *1* 成立。

$$\|V_{\text{true}}^{\pi} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty \leq \frac{R_{\max}L_\pi\varepsilon}{1-\gamma} + \frac{\gamma R_{\max}}{(1-\gamma)^2}L_\pi\varepsilon \tag{18}$$

证明。 利用 $V^{[\pi_\phi]_{\text{true}}}{}_{\text{确实}} = \mathcal{T}^{[\pi_\phi]_{\text{true}}}V^{[\pi_\phi]_{\text{true}}}{}_{\text{确实}}$ 这一事实,

$$\begin{aligned}
\|V_{\text{true}}^{\pi} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty &= \|V_{\text{true}}^{\pi} - \mathcal{T}^{[\pi_\phi]_{\text{true}}}V_{\text{true}}^{\pi} + \mathcal{T}^{[\pi_\phi]_{\text{true}}}V_{\text{true}}^{\pi} - \mathcal{T}^{[\pi_\phi]_{\text{true}}}V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty \\
&\leq \|V_{\text{true}}^{\pi} - \mathcal{T}^{[\pi_\phi]_{\text{true}}}V_{\text{true}}^{\pi}\|_\infty + \gamma\|V_{\text{true}}^{\pi} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty.
\end{aligned} \tag{19}$$

对于一个 MDP（$\mathcal{S}$, $\mathcal{A}$, r, γ, P），$(\mathcal{T}^{\pi}V)(s) := \mathbb{E}_{a'\sim\pi(\cdot|s),\ s'\sim P(\cdot|s,a')}[r(s,a') + \gamma V(s')]$ 是贝尔曼算子，它是关于无穷范数的γ-Lipchitz 压缩算子。因此,

$$\|V_{\text{true}}^{\pi} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty \leq \frac{1}{1-\gamma}\|V_{\text{true}}^{\pi} - \mathcal{T}^{[\pi_\phi]_{\text{true}}}V_{\text{true}}^{\pi}\|_\infty. \tag{20}$$

对于任意 b，我们有

$$\begin{aligned}
&|(V^{\pi}_{\text{true}} - \mathcal{T}^{[\pi_\phi]_{\text{true}}} V^{\pi}_{\text{true}})(b)| \\
&= |(\mathcal{T}^{\pi} V^{\pi}_{\text{true}} - \mathcal{T}^{[\pi_\phi]_{\text{true}}} V^{\pi}_{\text{true}})(b)| \\
&= \left| \mathbb{E}_{\substack{a\sim\pi(b) \\ b^{+1}\sim P(\cdot|b)}} \left[r + \gamma V^{\pi}_{\text{true}}(b^{+1}) \right] - \mathbb{E}_{\substack{a\sim\pi_\phi(\phi(b)) \\ b^{+1}\sim P(\cdot|b)}} \left[r + \gamma V^{\pi}_{\text{true}}(b^{+1}) \right] \right|
\end{aligned} \tag{21}$$

我们首先考察 r,

$$\begin{aligned}
|\mathbb{E}_{a\sim\pi(b)}[r] - \mathbb{E}_{a\sim\pi_\phi(\phi(b))}[r]| &= |\mathbb{E}_{a\sim\pi(b)}[r] - \mathbb{E}_{a\sim\pi(\phi(b))}[r]| \\
&\le R_{\max} L_\pi \varepsilon
\end{aligned} \tag{22}$$

接着考察 V^{π}_{true},

$$\begin{aligned}
&\left| \mathbb{E}_{\substack{a\sim\pi(b) \\ b^{+1}\sim P(\cdot|b)}} \left[\gamma V^{\pi}_{\text{true}}(b^{+1}) \right] - \mathbb{E}_{\substack{a\sim\pi_\phi(\phi(b)) \\ b^{+1}\sim P(\cdot|b)}} \left[\gamma V^{\pi}_{\text{true}}(b^{+1}) \right] \right| \\
&= \gamma \left| \sum_{o\in\mathcal{O}} \sum_{a\in\mathcal{A}} \left[\big(P(o|b,a)\pi(a|b) - P(o|b,a)\pi(a|\phi(b)) \big) \cdot V^{\pi}(b^{o,a}) \right] \right| \\
&\le \frac{\gamma R_{\max}}{1-\gamma} L_\pi \varepsilon
\end{aligned} \tag{23}$$

其中最后一个不等式用到了命题 2。 □

定理 2 的证明。

证明。 结合定理 11 和 12，我们得到

$$\begin{aligned}
\|[V^{\pi_\phi}_{\text{bin}}]_{\text{true}} - V^{\pi}_{\text{true}}\|_\infty &\le \|[V^{\pi_\phi}_{\text{bin}}]_{\text{true}} - V^{[\pi_\phi]_{\text{true}}}_{\text{true}}\|_\infty + \|V^{[\pi_\phi]_{\text{true}}}_{\text{true}} - V^{\pi}_{\text{true}}\|_\infty \\
&\le \frac{(L_\pi + 1) R_{\max} + 2L_V}{1-\gamma} \varepsilon + \frac{\gamma R_{\max} L_\pi + R_{\max}}{(1-\gamma)^2} \varepsilon,
\end{aligned} \tag{24}$$

从而完成该证明。 □

C 第 5.1 章中的双重采样分析

定义 4（抽象算法）. 考虑使用双重采样的贝尔曼误差最小化算法，它不仅在真实世界（模拟器）中执行，而且也在抽象系统中虚拟执行，并使用同一份离线数据。抽象算法的优化目标可以写为

$$\hat{Q}^{\pi_\phi}_{\phi} = \arg\min_{f\in\mathcal{F}} \mathcal{E}_\phi(f, \pi) \tag{25}$$

其中

$$\mathcal{E}_\phi(f,\pi) = \mathbb{E}_{\mathcal{D}}[(f(\phi(b),a) - (r_\phi + \gamma f(\phi(b'_A), \pi_\phi)))(f(\phi(b),a) - (r_\phi + \gamma f(\phi(b'_B), \pi_\phi)))]. \tag{26}$$

引理 6（马尔可夫决策过程望远镜恒等式 [Jiang 和 Xie， 2024]）. 对于一个马尔可夫决策过程（$\mathcal{S}, \mathcal{A}, r, \gamma, P$）以及任意函数 $Q: \mathcal{S}\times\mathcal{A} \to \mathbb{R}^+$，我们有

$$J_Q(\pi) - J(\pi) = \frac{1}{1-\gamma} \mathbb{E}_{d^\pi}[Q - \mathcal{T}^\pi Q] \tag{27}$$

其中 $(\mathcal{T}^\pi Q)(s,a) := r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)\, a'\sim\pi(\cdot|s')}[Q(s',a')]$ 是贝尔曼算子。

引理 7。 在分箱系统中，我们有如下望远镜误差

$$|J_{\hat{Q}}(\pi_\phi) - J(\pi_\phi)| \le \frac{\sqrt{C_\pi(\phi)}}{1-\gamma} \cdot \sqrt{\mathbb{E}_{d^{\mathcal{D}}}[(\hat{Q} - \mathcal{T}^{\pi_\phi}\hat{Q})^2]} \tag{28}$$

证明。 回顾前面提到的引理 6。将其代入抽象信念马尔可夫决策过程的情形，得到：

$$J_{\hat{Q}}(\pi_\phi) - J(\pi_\phi) = \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_\phi}}[\hat{Q} - \mathcal{T}^{\pi_\phi}\hat{Q}] \tag{29}$$

因此，我们有：

$$\begin{aligned}
|J_{\hat{Q}}(\pi_\phi) - J(\pi_\phi)| &= \left|\frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_\phi}}[\hat{Q} - \mathcal{T}^{\pi_\phi}\hat{Q}]\right| \\
&\leq \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_\phi}}[|\hat{Q} - \mathcal{T}^{\pi_\phi}\hat{Q}|] &(30)\\
&\leq \frac{1}{1-\gamma}\sqrt{\mathbb{E}_{d^{\pi_\phi}}[(\hat{Q} - \mathcal{T}^{\pi_\phi}\hat{Q})^2]} &(31)\\
&\leq \frac{1}{1-\gamma}\sqrt{\mathbb{E}_{d^D}[\frac{d^\pi}{d^D}(\hat{Q} - \mathcal{T}^{\pi_\phi}\hat{Q})^2]} &(32)\\
&\leq \frac{\sqrt{C_\pi(\phi)}}{1-\gamma}\cdot\sqrt{\mathbb{E}_{d^D}[(\hat{Q} - \mathcal{T}^{\pi_\phi}\hat{Q})^2]} &(33)
\end{aligned}$$

□

显然，我们有

$$\mathbb{E}_{d^D}[\mathcal{E}_\phi(\hat{Q}, \pi)] = \mathbb{E}_{d^D}[(\hat{Q} - \mathcal{T}^{\pi_\phi}\hat{Q})^2] \tag{34}$$

随着独立样本数量增加，上述值的经验估计与真实期望之间的差距会逐渐缩小，其收敛速度可用霍夫丁不等式或伯恩斯坦不等式等集中不等式来刻画。利用霍夫丁不等式，可得到以下引理。

引理 8。 以至少 $1-\delta$ 的概率，对于 $\forall f \in \mathcal{F}$，

$$|\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^D}[\mathcal{E}_\phi(f,\pi)]| \leq \sqrt{\frac{8R_{\max}^4}{n(1-\gamma)^4}\cdot\log\frac{2|\mathcal{F}|}{\delta}} \tag{35}$$

证明。 对于 $\mathcal{E}_\phi(f,\pi)$，我们首先估计其绝对值的一个上界。由于 $f \in \mathcal{F}$ 用于近似价值函数，因此可以假设其上界不超过 $R\max/(1-\gamma)$，即价值函数的上界。因此，我们可以给出一个粗略的上界（允许存在常数级的松弛，因为这只是一个常数）：

$$0 \leq \mathcal{E}_\phi(f,\pi) \leq \frac{4R_{\max}^2}{(1-\gamma)^2}$$

因此，根据霍夫丁不等式，对于任意 $f \in \mathcal{F}$，我们有：

$$\begin{aligned}
&\Pr(|\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^D}[\mathcal{E}_\phi(f,\pi)]| \geq t) \leq 2\exp\left(-\frac{2t^2n(1-\gamma)^4}{16R_{\max}^4}\right)\\
\rightarrow\ &\Pr(|\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^D}[\mathcal{E}_\phi(f,\pi)]| > t) \leq 2\exp\left(-\frac{2t^2n(1-\gamma)^4}{16R_{\max}^4}\right) &(36)
\end{aligned}$$

然而，证明的目标实际上是：

$$\Pr(\forall f \in \mathcal{F}, |\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^D}[\mathcal{E}_\phi(f,\pi)]| \leq t) \tag{37}$$

对于这类问题，一种常见的方法是使用并集界。设概率空间为 $(\Omega\ \Sigma,\ \Pr)$，并定义事件：

$$\begin{aligned}
A &:= \{\omega \in \Omega : \forall f \in \mathcal{F}, |\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^D}[\mathcal{E}_\phi(f,\pi)]|(\omega) \leq t\}\\
B_f &:= \{\omega \in \Omega : |\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^D}[\mathcal{E}_\phi(f,\pi)]|(\omega) > t\}
\end{aligned}$$

则:

$$\begin{aligned}
&\Pr(\forall f \in \mathcal{F}, |\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^\mathcal{D}}[\mathcal{E}_\phi(f,\pi)]| \le t) \\
=\ &\Pr(A) = 1 - \Pr(\Omega \backslash A) = 1 - \Pr(\bigcup_{f\in\mathcal{F}} B_f) \\
\ge\ &1 - \sum_{f\in\mathcal{F}} \Pr(B_f) \ge 1 - 2|\mathcal{F}| \exp\left(-\frac{t^2 n(1-\gamma)^4}{8R_{\max}^4}\right)
\end{aligned} \tag{38}$$

倒数第二步利用了概率的次可加性（可数次可加性），最后一步应用了不等式 (36)。

令 $\delta = 2|\mathcal{F}| \exp\big(-t^2 n(1-\gamma)^4/8R_{\max}^4\big)$，则解出 t 得到 $t = \sqrt{\frac{8R_{\max}^4}{n\ 1-\gamma()^4} \cdot \log \frac{2|\mathcal{F}|}{\delta}}$

将其代入 (37) 即完成证明。 □

事实上，霍夫丁不等式仅利用了函数的有界性。然而，通过引入以下贝尔曼完备性假设，我们还可以考虑函数的方差，并应用伯恩斯坦不等式来获得更紧的收敛速率。

标准的贝尔曼完备性假设如下:
假设 8 （贝尔曼完备性）. $\forall f \in \mathcal{F}, \mathcal{T}^{\pi_\phi} f \in \mathcal{F}$.
注 7. 对于满足 $|\mathcal{F}| < \infty$ 的有限函数空间，贝尔曼完备性假设 *8* 蕴含可实现性假设 *4*。

命题 3. 在贝尔曼完备性假设 *8* 下，我们可以得到收敛速率为 $O(1/n)$ 的上界。具体而言，以至少 $1-\delta$ 的概率，对所有 $f \in \mathcal{F}$，下列结论成立:

$$|\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^\mathcal{D}}[\mathcal{E}_\phi(f,\pi)]| \lesssim \frac{R_{\max}^2}{n(1-\gamma)^2} \cdot \log \frac{|\mathcal{F}|}{\delta} \tag{39}$$

当然，就本次讨论的目的而言，贝尔曼完备性假设并非必要——只需下述可实现性假设即可达成目标。然而，在这种情况下，我们只能利用由霍夫丁不等式导出的引理 8 来刻画集中速率，而无法使用命题 3 所给出的更紧的集中速率。

引理 9. 若假设*5*成立，则

$$|\mathcal{E}(f,\pi) - \mathcal{E}_\phi(f,\pi)| \le \frac{4R_{\max}}{1-\gamma} \cdot \left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon \tag{40}$$

证明。

$$\begin{aligned}
&|\mathcal{E}(f,\pi) - \mathcal{E}_\phi(f,\pi)| \\
=\ &|\mathbb{E}_\mathcal{D}[(f(b,a) - (r + \gamma f(b'_A,\pi)))(f(b,a) - (r + \gamma f(b'_B,\pi)))] - \\
&\quad \mathbb{E}_\mathcal{D}[(f(\phi(b),a) - (r_\phi + \gamma f(\phi(b'_A),\pi_\phi)))(f(\phi(b),a) - (r_\phi + \gamma f(\phi(b'_B),\pi_\phi)))]| \\
\le\ &|\mathbb{E}_\mathcal{D}[\{(f(b,a) - f(\phi(b),a)) - (r(b,a) - r_\phi(\phi(b),a)) - \gamma(f(b'_A,\pi) - f(\phi(b'_A),\pi_\phi)\} \\
&\qquad \cdot (f(b,a) - (r + \gamma f(b'_B,\pi)))]| + \\
&|\mathbb{E}_\mathcal{D}[\{(f(b,a) - f(\phi(b),a)) - (r(b,a) - r_\phi(\phi(b),a)) - \gamma(f(b'_B,\pi) - f(\phi(b'_B),\pi_\phi)\} \\
&\qquad \cdot (f(b,a) - (r + \gamma f(b'_A,\pi)))]|.
\end{aligned} \tag{41}$$

利用这一事实:

$$\begin{aligned}
&|f(b,\pi) - f(\phi(b),\pi_\phi)| \\
=\ &|\mathbb{E}_{\pi(a|b)}[f(b,a)] - \mathbb{E}_{\pi(a|\phi(b))}[f(\phi(b),a)]| \\
\le\ &|\mathbb{E}_{\pi(a|b)}[f(b,a)] - \mathbb{E}_{\pi(a|\phi(b))}[f(b,a)]| + |\mathbb{E}_{\pi(a|\phi(b))}[f(b,a)] - \mathbb{E}_{\pi(a|\phi(b))}[f(\phi(b),a)]| \\
\le\ &\frac{R_{\max}}{1-\gamma}\varepsilon + L_Q\varepsilon
\end{aligned} \tag{42}$$

我们有

$$\begin{aligned}&|\mathcal{E}(f,\pi)-\mathcal{E}_\phi(f,\pi)|\\&\leq 2\cdot\frac{2R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q+\frac{R_{\max}}{1-\gamma}\right)\varepsilon\end{aligned}\tag{43}$$

□

命题 4. 若假设 *4*、*5* 成立，则

$$|\mathbb{E}_{d^D}[\mathcal{E}_\phi(\hat{Q}^\pi,\pi)]|\leq\sqrt{\frac{32R_{\max}^4}{n(1-\gamma)^4}\cdot\log\frac{2|\mathcal{F}|}{\delta}}+\frac{8R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q+\frac{R_{\max}}{1-\gamma}\right)\varepsilon\tag{44}$$

证明。 根据引理 8，我们有

$$|\mathcal{E}_\phi(\hat{Q},\pi)-\mathbb{E}_{d^D}[\mathcal{E}_\phi(\hat{Q},\pi)]|\leq\sqrt{\frac{8R_{\max}^4}{n(1-\gamma)^4}\cdot\log\frac{2|\mathcal{F}|}{\delta}}\tag{45}$$

and

$$|\mathcal{E}_\phi(Q_\phi^{\pi_\phi},\pi)-\mathbb{E}_{d^D}[\mathcal{E}_\phi(Q_\phi^{\pi_\phi},\pi)]|\leq\sqrt{\frac{8R_{\max}^4}{n(1-\gamma)^4}\cdot\log\frac{2|\mathcal{F}|}{\delta}}\tag{46}$$

其中 $\mathbb{E}_{d^D}[\mathcal{E}_\phi(Q_\phi^{\pi_\phi},\pi)]=0$。然后根据引理 9，以大于 $1-\delta$ 的概率，我们有

$$|\mathcal{E}(\hat{Q},\pi)-\mathbb{E}_{d^D}[\mathcal{E}_\phi(\hat{Q},\pi)]|\leq\sqrt{\frac{8R_{\max}^4}{n(1-\gamma)^4}\cdot\log\frac{2|\mathcal{F}|}{\delta}}+\frac{4R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q+\frac{R_{\max}}{1-\gamma}\right)\varepsilon\tag{47}$$

and

$$|\mathcal{E}(Q_\phi^{\pi_\phi},\pi)-\mathbb{E}_{d^D}[\mathcal{E}_\phi(Q_\phi^{\pi_\phi},\pi)]|\leq\sqrt{\frac{8R_{\max}^4}{n(1-\gamma)^4}\cdot\log\frac{2|\mathcal{F}|}{\delta}}+\frac{4R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q+\frac{R_{\max}}{1-\gamma}\right)\varepsilon\tag{48}$$

根据抽象可实现性假设 4，我们有 $\mathcal{E}(\hat{Q},\pi)=\min_{f\in\mathcal{F}}\mathcal{E}(f,\pi)\leq\mathcal{E}(Q_\phi^{\pi_\phi},\pi)$，因此

$$|\mathbb{E}_{d^D}[\mathcal{E}_\phi(\hat{Q}^\pi,\pi)]|\leq\sqrt{\frac{32R_{\max}^4}{n(1-\gamma)^4}\cdot\log\frac{2|\mathcal{F}|}{\delta}}+\frac{8R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q+\frac{R_{\max}}{1-\gamma}\right)\varepsilon.\tag{49}$$

□

因此，

定理 13. 若假设 *4* 和假设 *5* 成立，则

$$|J_{\hat{Q}^\pi}(\pi_\phi)-J(\pi_\phi)|\leq\frac{\sqrt{C_\pi(\phi)}}{1-\gamma}\cdot\sqrt{\sqrt{\frac{32R_{\max}^4}{n(1-\gamma)^4}\cdot\log\frac{2|\mathcal{F}|}{\delta}}+L_{\mathcal{E}}\varepsilon}\tag{50}$$

其中

$$L_{\mathcal{E}}:=\frac{8R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q+\frac{R_{\max}}{1-\gamma}\right)\tag{51}$$

证明. 该结果可直接由命题 4、引理 7 和式 (34) 得出。 □

以下一系列定理均为最终控制总体误差的预备步骤。

定理14. 若假设5成立，则

$$|J_{\hat{Q}^\pi}(\pi) - J_{\hat{Q}^\pi}(\pi_\phi)| \leq \frac{R_{\max}}{1-\gamma}\varepsilon + L_Q\varepsilon \tag{52}$$

证明。我们有

$$\begin{aligned}
&|J_{\hat{Q}^\pi}(\pi) - J_{\hat{Q}^\pi}(\pi_\phi)| \\
&= |\mathbb{E}_{b\sim d_0}[\hat{Q}^\pi(b,\pi)] - \mathbb{E}_{b\sim d_0}[\hat{Q}^\pi(\phi(b),\pi_\phi)]| \\
&= |\mathbb{E}_{b\sim d_0}[\hat{Q}^\pi(b,\pi) - \hat{Q}^\pi(\phi(b),\pi_\phi)]| \\
&\leq \frac{R_{\max}}{1-\gamma}\varepsilon + L_Q\varepsilon
\end{aligned} \tag{53}$$

□

定理 6 的证明。

证明。 根据定理14和定理 3 中 $L_\phi^{[2]}$ 的定义，我们有 $L_\phi^{[2]} = \frac{R_{\max}}{1-\gamma} + L_Q$ □

定理 7 的证明。

证明。 结合定理 13、定理 6 和元定理3，我们得到 $|J_{\hat{}Q^\pi}(\pi) - J(\pi)| \leq \frac{\sqrt{C_\pi}}{(\varepsilon)\,1-\gamma} \cdot \sqrt{\sqrt{\frac{32R_{\max}^4}{n\ 1-\gamma\ ()^4}} \cdot \log\frac{}{}}$ $\frac{}{2|\mathcal{F}|}{\delta} + L_{\mathcal{E}}\varepsilon + L_\phi\varepsilon$, 其中 $L_{\mathcal{E}} = \frac{8R_{\max}}{1-\gamma} \cdot \big((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\big)$。注意，应用元定理3时，我们指定 $\mathfrak{est}^\phi(\hat{}Q^\pi) = J_{\hat{}Q^\pi}(\pi_\phi)$ 和 $\mathfrak{est}^\phi(Q_\phi^\pi) = J(\pi_\phi)$。

然后，应用引理3的结果即可证明该结论。 □

推论 1 的证明。

证明。 注意，平方根内的 ε 始终大于外部的 ε，只要 ε 足够小。因此，将 $L_{\mathcal{E}}\varepsilon$ 替换为 $\sqrt{\frac{32R_{\max}^4}{n\ 1-\gamma\ ()^4} \cdot \log\ }$ $\frac{2|\mathcal{F}|}{\delta}$，再找出使外部的 $L_\phi\varepsilon$ 不超过平方根内项的条件，即可证明该推论。该条件可表示为 $\big(\frac{32R^4}{n}{}_{\max(1-\gamma)^4} \cdot \log\ \frac{2|\mathcal{F}|}{\delta}\big)^{\frac{1}{4}} \leq \frac{L_{\mathcal{E}}}{L_\phi} \cdot \frac{\sqrt{2C_\pi^n}}{1-\gamma}$。注意，覆盖项通常是递增的，且始终有下界 1，因此我们给出一个充分条件：$\big(\frac{32R^4}{n}{}_{\max(1-\gamma)^4} \cdot \log\ \frac{2|\mathcal{F}|}{\delta}\big)^{\frac{1}{4}} \leq \frac{L_{\mathcal{E}}}{L_\phi} \cdot \frac{\sqrt{2}}{1-\gamma}$。求解可得条件 $n \geq 8R_{\max}^4(L_\phi/L_{\mathcal{E}})^4\log(2|\mathcal{F}|/\delta)$，在此条件下，$|J_{\hat{}Q^\pi}(\pi) - J(\pi)| \leq \frac{2\sqrt{C_\pi^n}}{1-\gamma} \cdot \big(\frac{128R^4}{n\ 1-\gamma^{\max()^4}} \cdot \log\ \frac{2|\mathcal{F}|}{\delta}\big)^{\frac{1}{4}}$。 □

D 依赖未来的价值函数

定义5。 若马尔可夫决策过程 $M_1 = \{\mathcal{S}_1, \mathcal{A}, P_1, R_1, H\}$ 和 $M_2 = \{\mathcal{S}_2, \mathcal{A}, P_2, R_2, H\}$ 之间存在双射 $\varphi\colon \mathcal{S}_1 \to \mathcal{S}_2$，使得 $\varphi(M_1) := \{\varphi(\mathcal{S}_1)\ \mathcal{A}, P_1(\varphi(\cdot)\ \cdot)\ R_1(\varphi(\cdot)\ \cdot)\ H\} = M_2$，则称它们同构。

定理 15。 对于任意部分可观测马尔可夫决策过程 $\mathcal{P}$ 和 $T \in \mathbb{N}^+$，都存在一个记忆窗口为 T 的短期记忆部分可观测马尔可夫决策过程 $\mathcal{P}_T$，使得对 $\mathcal{P}$的信念马尔可夫决策过程进行抽象 ϕ_T后，与 $\mathcal{P}_T$的信念马尔可夫决策过程同构。

证明。 要证明其存在，只需构造一个短期记忆部分可观测马尔可夫决策过程 $\mathcal{P}_T$。考虑原部分可观测马尔可夫决策过程 $\mathcal{P}$ 的信念马尔可夫决策过程 $\mathcal{M}$，并令 $\mathcal{M}_T$ 为通过 ϕ_T对 $\mathcal{M}$ 进行的抽象。用 **信念** 表示原部分可观测马尔可夫决策过程中的信念映射。现在，按如下方式构造 $\mathcal{P}_T = (\mathcal{S}'$, $\mathcal{A}$, $\mathcal{O}$, r', H, $\mathbb{O}'$, $\mathbb{T}')$:
定义 $\mathcal{S}' := \bigcup_{i=0}^{T-1}(\mathcal{O}\times\mathcal{A})^i \times \mathcal{O}$，并将观测函数定义为 $\mathbb{O}'(o|s' = (o_1, a_1, \cdots, o_T)) := \mathbb{I}\{o = o_T\}$，它是一个独热向量。奖励函数定义为 $r'(s' = (o_1, a_1, \cdots, o_T)\ a) := R_\phi(\phi_T(\mathfrak{b}(s'))\ a)$，转移概率定义为 $\mathbb{T}(s_1'|s_0', a) := P_\phi(\phi_T(\mathfrak{b}(s_0'))|\phi_T(\mathfrak{b}(s_1'))\ a)$，其中 P_ϕ 在式（6）中定义。

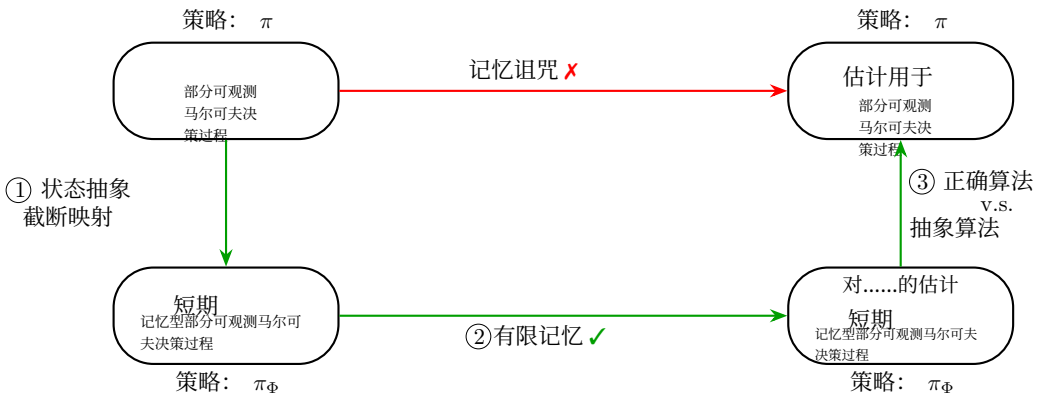

图 2：FDVF 分析流程

接下来，我们验证 POMDP $\mathcal{P}_T$ 的信念马尔可夫决策过程 $\mathcal{M}_T$ 确实与 $\mathcal{P}$的信念马尔可夫决策过程 $\mathcal{M}$ 的抽象同构。注意，对于 $\mathcal{P}_T$中的每个历史序列，只需取该序列最后的 T 个元素，即可唯一确定其状态。也就是说，$\mathcal{P}_T$ 中的信念状态采用独热编码。因此， $\mathcal{P}_T$ 的信念马尔可夫决策过程与 $\mathcal{P}_T$的隐藏底层马尔可夫决策过程同构。根据上述定义，该底层马尔可夫决策过程自然与 $\mathcal{M}_T$同构。

至此，构造和证明均已完成。 □

备注 8. 如上文及正文所述，由于这个短期记忆部分可观测马尔可夫决策过程是由一个抽象映射 ϕ诱导的，并且该抽象映射 ϕ 保证映射到同一代表状态的所有信念状态彼此接近（假设 *6*），我们可以直接应用定理 *2* 的结论来控制抽象误差。

请注意，策略截断在这里至关重要。这不仅是为了直接重用定理 *2* 的结论，也是因为存在“记忆诅咒”——策略的记忆可能会严重影响理论保证的质量。

D.1 实际算法与抽象算法

实际算法与 抽象算法之间的差异来自三个方面 s:

1. $\mu(a_h,\ \tau_h^+) = \pi_e(a_h|\tau_h^+)/\pi_b(a_h|\tau_h^+)$ 与截断版本 $\mu(a_h,\ \tau_{[h-T+1:\ h]}^+) = \pi_e(a_h|\tau_{[h-T+1:\ h]}^+)/\pi_b(a_h|\tau_{[h-T+1:\ h]}^+)$ 之间的差异。该差异可由以下引理控制：**引理 10**。 若假设

设 *7* 成立，则对于任意满足条件 *1* 的 $\varepsilon > 0$，当 $T \geq T_1(\varepsilon)$ 时，我们有 e,

$$|\mu(a_h, \tau_h^+) - \mu(a_h, \tau_{[h-T+1:h]}^+)| \leq \frac{2(C_\mu + 1)L_\pi \varepsilon}{\min_h \min_{a_h, \tau_h^+} \pi_b(a_h|\tau_h^+)} \tag{54}$$

证明。 根据条件 1，我们有

$$\begin{aligned}
|\mu(a_h, \tau_h^+) - \mu(a_h, \tau_{[h-T+1:h]}^+)| \leq & \frac{|\pi_e(a_h|\tau_h^+) - \pi_e(a_h|\tau_{[h-T+1:h]}^+)|}{\pi_b(a_h|\tau_h^+)} \\
& + \pi_e(a_h|\tau_{[h-T+1:h]}^+)\left|\frac{1}{\pi_b(a_h|\tau_h^+)} - \frac{1}{\pi_b(a_h|\tau_{[h-T+1:h]}^+)}\right| \qquad (55) \\
\leq & \frac{L_\pi \varepsilon}{\min_h \min_{a_h, \tau_h^+} \pi_b(a_h|\tau_h^+)}\left(1 + \frac{\pi_e(a_h|\tau_{[h-T+1:h]}^+)}{\pi_b(a_h|\tau_{[h-T+1:h]}^+)}\right) \qquad (56) \\
\leq & \frac{L_\pi \varepsilon}{\min_h \min_{a_h, \tau_h^+} \pi_b(a_h|\tau_h^+)}(2 + 2C_\mu) \qquad (57)
\end{aligned}$$

由此证明了该引理。 □

2. $V(f'_h, \tau'_h)$ 与 $V(f'_h, \tau'_{[h-T+1:h]})$ 之间的差异。这要求函数类能够快速遗忘历史信息，如下所述：

假设9 (快速遗忘函数类)。考虑用于估计的函数类 $\mathcal{V}$： $\mathcal{F} = (\mathcal{F}' \times \mathcal{H}) \to \mathbb{R}$。对于所有 $\varepsilon > 0$，存在 $T \in \mathbb{N}^+$，使得对于所有 $V \in \mathcal{V}$，

$$|V(f_h, \tau_h) - V(f_{[h:h+T]}, \tau_{[h-T+1:h]})| \leq \|\mathcal{V}\|_\infty \varepsilon \tag{58}$$

T 的合适取值构成 ε的一个函数，记为 $T_2(\varepsilon)$。

请注意，这里的关键假设是，扩展未来中的"历史"会快速遗忘。由于在 FDVF 的原始文献中， [Uehara 等人， 2023], 未来默认会被长度为 M_F 的范围截断。

3. 真实部分可观测马尔可夫决策过程与抽象短期记忆部分可观测马尔可夫决策过程之间的数据生成分布差异。这种差异源于两个方面：首先，行为策略被截断；其次，部分可观测马尔可夫决策过程的转移概率存在细微差异。

令 $w^{\phi}(f_1)$表示用于衡量该分布偏移的重要性权重。于是我们定义：

$$w^{\phi_T}(f_1) := \frac{\pi_b^{\phi_T}(a_1|\tau_1^+)}{\pi_b(a_1|\tau_1^+)} \cdot \frac{P^{\phi_T}(o_2|\tau_2)}{P(o_2|\tau_2)} \cdot \cdots \cdot \frac{\pi_b^{\phi_T}(a_H|\tau_H^+)}{\pi_b(a_H|\tau_H^+)} \tag{59}$$

在部分可观测马尔可夫决策过程与策略均为快速遗忘的假设下，我们有如下引理：

引理11. 若假设*6*和*7*成立，则对任意 $\varepsilon > 0$，在 $T \geq \max\{T_0(\varepsilon)\ T_1(\varepsilon)\}$的条件下，当 ε 足够小，即当条件*2*满足时，我们有

$$|w^{\phi_T}(f_1)| \leq 1 + \frac{3H\varepsilon}{\min\{\min_h \min_{o_h,\tau_h} P(o_h|\tau_h), \min_h \min_{a_h,\tau_h^+} \pi_b(a_h|\tau_h^+)/L_\pi\}} \tag{60}$$

证明. 我们首先证明，利用条件2，对任意 h， $\left|1 - \frac{\pi_b^{\phi_T}(a_h|\tau_h^+)}{\pi_b(a_h|\tau_h^+)}\right| \leq \frac{L_\pi \varepsilon}{\min_h \min^{a_h,\tau_h^+} \pi_b(a_h|\tau_h^+)}$。

然后利用如下事实：

$$P^{\phi_T}(o_h|\tau_h) = P^{\phi_T}(o_h|\mathbf{b}(\tau_{h-1}^+), a_{h-1}) = \mathbb{E}_{b\sim p_{\phi_T}(\mathbf{b}(\tau_{h-1}^+))}[P(o_h|b, a_{h-1})] \tag{61}$$

以及对所有 $b \in \phi_T^{-1}(\phi_T(\mathbf{b}(\tau_{h-1}^+)))$，均有 $\|b - \mathbf{b}(\tau_{h-1}^+)\|_1 \leq \varepsilon$，这在假设 6中给出。由此可得

$$\begin{aligned}
|P(o_h|\tau_h) - P^{\phi_T}(o_h|\tau_h)| &= |P(o_h|\mathbf{b}(\tau_{h-1}^+), a_{h-1}) - \mathbb{E}_{b\sim p_{\phi_T}(\mathbf{b}(\tau_{h-1}^+))}[P(o_h|b, a_{h-1})]| \\
&= |\mathbb{E}_{b\sim p_{\phi_T}(\mathbf{b}(\tau_{h-1}^+))}[P(o_h|\mathbf{b}(\tau_{h-1}^+), a_{h-1}) - P(o_h|b, a_{h-1})]| \\
&\leq \mathbb{E}_{b\sim p_{\phi_T}(\mathbf{b}(\tau_{h-1}^+))}[|P(o_h|\mathbf{b}(\tau_{h-1}^+), a_{h-1}) - P(o_h|b, a_{h-1})|] \qquad (62) \\
&\leq \varepsilon \qquad (63)
\end{aligned}$$

其中最后一步用到了引理 5。现在我们有 $\left|1 - \frac{P^{\phi_T}(o_h|\tau_h)}{P(o_h|\tau_h)}\right| \leq \frac{\varepsilon}{\text{最小值}_h \text{最小值}_{o_h,\ \tau_h} P(o_h|\tau_h)}$。

Therefore, $w^{\phi_T} \leq (1 + \frac{\varepsilon}{\min\{\min_h \min_{o_h,\tau_h} P(o_h|\tau_h), \min_h \min_{a_h,\tau_h^+} \pi_b(a_h|\tau_h^+)/L_\pi\}})^H$, combining Condition 2 we have $w^{\phi_T} \leq 1 + \frac{3H\varepsilon}{\min\{\min_h \min_{o_h,\tau_h} P(o_h|\tau_h), \min_h \min_{a_h,\tau_h^+} \pi_b(a_h|\tau_h^+)/L_\pi\}}$. □

综上所述，通过考虑所有误差来源，我们得到以下定理。

引理12。定义

$$\mathcal{E}_{\mathcal{V},\Theta}(V,\theta) := \sum_{h=1}^{H} \mathbb{E}_{\mathcal{D}}[\{\mu(a_h,\tau_h^+)(r_h + V(f_{h+1})) - V(f_h)\}\theta(\tau_h) - \frac{1}{2}\theta(\tau_h)^2] \tag{64}$$

$$\mathcal{E}^{\phi_T}_{\mathcal{V},\Theta}(V,\theta) := \sum_{h=1}^{H} \mathbb{E}_{\mathcal{D}}[w^{\phi_T}(\{\mu(a_h,\tau^+_{h-T+1:h})(r_h + V(\phi_T(f_{h+1}))) - V(\phi_T(f_h))\}\theta(\tau_h) - \frac{1}{2}\theta(\tau_h)^2)] \tag{65}$$

其中 $\phi_T(f_h) = \phi_T(f'_h, \tau_h) = (f'_h, \phi_T(\tau_h))$。若假设*6*、*7*和*9*均成立，则对于任意满足条件 *1* 和 *2* 的 $\varepsilon > 0$*s*，以及任意 $V \in \mathcal{V}$、$\theta \in \Theta$，有：

$$|\mathcal{E}_{\mathcal{V},\Theta}(V,\theta) - \mathcal{E}^{\phi_T}_{\mathcal{V},\Theta}(V,\theta)| \le \frac{1}{3} L_{\mathcal{E}} \varepsilon \tag{66}$$

和 $T =$ 最大值$\{T_0(\varepsilon)\ T_1(\varepsilon)\ T_2(\varepsilon)\}$。

证明。 该引理的证明使用三角不等式，将引理 10、假设9和引理11中描述的误差相加，即可得到结论。 □

定理16。定义

$$\mathcal{E}_{\mathcal{V}}(V) := \max_{\theta\in\Theta} \mathcal{E}_{\mathcal{V},\Theta}(V,\theta), \quad \mathcal{E}^{\phi_T}_{\mathcal{V}}(V) := \max_{\theta\in\Theta} \mathcal{E}^{\phi}_{\mathcal{V},\Theta}(V,\theta) \tag{67}$$

若假设*6*、*7*，并且*9*均成立，则对于任意满足条件 *1* 和 *2* 的 $\varepsilon > 0$，以及任意 $V \in \mathcal{V}$，我们有 *e:*

$$|\mathcal{E}_{\mathcal{V}}(V) - \mathcal{E}^{\phi_T}_{\mathcal{V}}(V)| \le \frac{2}{3} L_{\mathcal{E}} \varepsilon \tag{68}$$

且 $T = \max\{T_0(\varepsilon)\ T_1(\varepsilon)\ T_2(\varepsilon)\}$。

证明。 这里用到如下观察：若 $\forall\theta$，$|\mathcal{E}_{\mathcal{V},\Theta}(V,\theta) - \mathcal{E}^{\phi_T}_{\mathcal{V},\Theta}(V,\theta)| \le \frac{1}{3} L_{\mathcal{E}}\varepsilon$，则 $|\sup_\theta \mathcal{E}_{\mathcal{V},\Theta}(V,\theta) - \sup_\theta \mathcal{E}^{\phi_T}_{\mathcal{V},\Theta}(V,\theta)| \le 2 \cdot \frac{1}{3} L_{\mathcal{E}}\varepsilon$。 □

定理 8 的证明。

证明。 为证明该定理，我们需要证明 $L^{[2]}_{\phi} := \|\mathcal{V}\|_\infty$ 满足 $|\mathbb{E}_{\pi_b}[\hat{V}(f_1)] - \mathbb{E}_{\pi_b^\phi}[\hat{V}(f_1)]| \le L^{[2]}_{\phi}\varepsilon$。

回顾我们的构造，短期记忆部分可观测马尔可夫决策过程的动态过程保证了第一个 T 动作-观测与真实的部分可观测马尔可夫决策过程具有完全相同的动态过程。另请注意，$\forall f_1^{[1]}$，$f_1^{[2]} \in \mathcal{F}_1$，若 $f_1^{[1]}$，$f_1^{[2]}$共享前 T对动作和观测，则 $|\hat{V}(f_1^{[1]}) - \hat{V}(f_1^{[2]})| \le \|\mathcal{V}\|_\infty \varepsilon$，这由函数类 $\mathcal{V}$的性质（假设9）可知。接下来

$$\begin{aligned} |\mathbb{E}_{\pi_b}[\hat{V}(f_1)] - \mathbb{E}_{\pi_b^\phi}[\hat{V}(f_1)]| &\le \mathbb{E}_{f_{1:T}\sim\pi_b}[|\mathbb{E}_{f_{T+1}\sim\pi_b}[\hat{V}(f_1)|f_{1:T}] - \mathbb{E}_{f_{T+1}\sim\pi_b^\phi}[\hat{V}(f_1)|f_{1:T}]|] \\ &\le \|\mathcal{V}\|_\infty \varepsilon \end{aligned} \tag{69}$$

这证明了该定理。 □

D.2 FDVF 的理论保证

定理 9 的证明。

证明。 令 $\hat{V}:=\arg\min_{V\in\mathcal{V}}\mathcal{E}_{\mathcal{V}}(V)$，相应地令 $\hat{V}_\phi:=\arg\min_{V\in\mathcal{V}}\mathcal{E}^{\phi_T}_{\mathcal{V}}(V)$。我们的首要目标是证明其概率大于 $1-\delta$,

$$|J^\phi(\pi_e^\phi)-\mathbb{E}_{\pi_b^\phi}[\hat{V}(f_1)]|\le\sqrt{H}\cdot\max_{h\in[H]}\sup_{V\in\mathcal{V}}\sqrt{\frac{\mathbb{E}_{\pi_e^\phi}[(\mathcal{B}^{(\mathcal{S},\mathcal{H}_T)}V)(s_h,\tau_{[h-T+1:h]})^2]}{\mathbb{E}_{\pi_b^\phi}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2]}}\cdot\sqrt{\frac{cHC_{\mathcal{V}}^2C_\mu}{n}\log\frac{|\mathcal{V}||\Theta|}{\delta}+L_{\mathcal{E}}\varepsilon} \tag{70}$$

为此，我们遵循 [Zhang 和 Jiang 提供的证明， 2024],定义 $X^\phi_{V,h}:=\mu(a_h,\tau_h^+)(r_h+V(\phi_T(f_{h+1})))-V(\phi_T(f_h))$, $X_{V,h}:=\mu(a_h,\tau_h^+)(r_h+V(f_{h+1}))-V(f_h)$，然后 $\mathcal{E}^{\phi_T}_{\mathcal{V}}(V)=\frac{1}{2}\max_{\theta\in\Theta}\sum_{h=1}^H\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^\phi_{V,h})^2-(X^\phi_{V,h}-\theta(\tau_h))^2)]$,，并将取得最大值的此类 θ 记为 $\hat{\theta}_V$。类似地， $\mathcal{E}_{\mathcal{V}}(V)=\frac{1}{2}\max_{\theta\in\Theta}\sum_{h=1}^H\mathbb{E}_{\mathcal{D}}[X^2_{V,h}-(X_{V,h}-\theta(\tau_h))^2]$，取得最大值的 θ 记为 $\hat{\theta}^0_V$。根据 [Zhang 和 Jiang 的定理 2 证明第一部分（内层最大化分析）中所述、利用伯恩斯坦不等式的集中论证， 2024], 我们得到一个论证，表明以大于 $1-\delta/2$ 的概率，对任意$V\in\mathcal{V},\theta\in\Theta$,

$$\left|\sum_{h=1}^H\mathbb{E}_{\mathcal{D}}[w^{\phi_T}(\hat{\theta}_V(\tau_h)-X^\phi_{V,h})^2]-\sum_{h=1}^H\mathbb{E}_{\mathcal{D}}[w^{\phi_T}(X^\phi_{V,h}-(\mathcal{B}^{\mathcal{H}}V)(\tau_h))^2]\right|\le\frac{675HC_{\mathcal{V}}^2C_\mu\|w^{\phi_T}\|_\infty}{n}\cdot\log\frac{4|\mathcal{V}||\Theta|}{\delta}=:\eta \tag{71}$$

于是我们有

$$\begin{aligned}
&\sum_{h=1}^H\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^\phi_{\hat{V},h})^2-(X^\phi_{\hat{V},h}-(\mathcal{B}^{\mathcal{H}}\hat{V})(\tau_h))^2)]\\
\le&\sum_{h=1}^H\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^\phi_{\hat{V},h})^2-(\hat{\theta}_{\hat{V}}(\tau_h)-X^\phi_{\hat{V},h})^2)]+\eta &(72)\\
\le&\sum_{h=1}^H\mathbb{E}_{\mathcal{D}}[X^2_{\hat{V},h}-(\hat{\theta}_{\hat{V}}(\tau_h)-X_{\hat{V},h})^2]+\frac{1}{3}L_{\mathcal{E}}\varepsilon+\eta &(73)\\
\le&\sum_{h=1}^H\mathbb{E}_{\mathcal{D}}[X^2_{\hat{V},h}-(\hat{\theta}^0_{\hat{V}}(\tau_h)-X_{\hat{V},h})^2]+\frac{1}{3}L_{\mathcal{E}}\varepsilon+\eta &(74)\\
\le&\sum_{h=1}^H\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^\phi_{\hat{V}_\phi,h})^2-(\hat{\theta}_{\hat{V}_\phi}(\tau_h)-X^\phi_{\hat{V}_\phi,h})^2)]+L_{\mathcal{E}}\varepsilon+\eta &(75)
\end{aligned}$$

这里，(72) 使用了 (71)。(73) 使用了引理12，以及以下事实：$\mathcal{E}_{\mathcal{V},\Theta}(\hat{V},\hat{\theta}_{\hat{V}})=\sum_{h=1}^H\mathbb{E}_{\mathcal{D}}[X^2_{\hat{V},h}-(\hat{\theta}_{\hat{V}}(\tau_h)-X_{\hat{V},h})^2]$，并且 $\mathcal{E}^\phi_{\mathcal{V},\Theta}(\hat{V},\hat{\theta}_{\hat{V}})=\sum_{h=1}^H\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^{\phi\hat{}V,h})^2-(\hat{\theta}_{\hat{V}}(\tau_h)-X^{\phi\hat{}V,h})^2)]$。之后，(74) 使用了 $\hat{\theta}^0_{\hat{V}}=\arg\max_{\theta\in\Theta}\mathcal{E}_{\mathcal{V},\Theta}(\hat{V},\theta)$ 这一事实，因此 $\mathcal{E}_{\mathcal{V},\Theta}(\hat{V},\hat{\theta}^0_{\hat{V}})\ge\mathcal{E}_{\mathcal{V},\Theta}(\hat{V},\hat{\theta}_{\hat{V}})$。(75) 使用了定理16，以及以下事实：$\min_{V\in\mathcal{V}}\mathcal{E}_{\mathcal{V}}(V)=\sum_{h=1}^H\mathbb{E}_{\mathcal{D}}[X^2_{\hat{V},h}-(\hat{\theta}^0_{\hat{V}}(\tau_h)-X^{\hat{}V,h})^2]$，并且 $\min_{V\in\mathcal{V}}\mathcal{E}^\phi_{\mathcal{V}}(V)=\sum_{h=1}^H\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^{\phi\hat{}V_\phi,h})^2-(\hat{\theta}^{\hat{}V_\phi}(\tau_h)-X^{\phi\hat{}V_\phi,h})^2)]$。注意到 $\min_{V\in\mathcal{V}}\mathcal{E}_{\mathcal{V}}\le\min_{V\in\mathcal{V}}\mathcal{E}^\phi_{\mathcal{V}}(V)+\sup_{V\in\mathcal{V}}|\mathcal{E}_{\mathcal{V}}(V)-\mathcal{E}^\phi_{\mathcal{V}}(V)|$，即可完成 (75) 的推导。

在.....之后 注意到抽象可实现性假设 $\exists V^\phi_{\mathcal{F}}\in\mathcal{V}$，并且对于任意 $V^\phi_{\mathcal{F}}$，我们有 e

$$\begin{aligned}
&\sum_{h=1}^H\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^\phi_{\hat{V}_\phi,h})^2-(\hat{\theta}_{\hat{V}_\phi}(\tau_h)-X^\phi_{\hat{V}_\phi,h})^2)]\\
\le&\sum_{h=1}^H\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^\phi_{V^\phi_{\mathcal{F}},h})^2-(\hat{\theta}_{V^\phi_{\mathcal{F}}}(\tau_h)-X^\phi_{V^\phi_{\mathcal{F}},h})^2)] &(76)\\
\le&\sum_{h=1}^H\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^\phi_{V^\phi_{\mathcal{F}},h})^2-(X^\phi_{V^\phi_{\mathcal{F}},h}-(\mathcal{B}^{\mathcal{H}}V^\phi_{\mathcal{F}})(\tau_h))^2)]+\eta=\eta &(77)
\end{aligned}$$

其中，倒数第二个不等式利用了 $\hat{} V_\phi$ 的最小性，最后一个不等式利用了式 (71)。最后一个等式是根据 $V_{\mathcal{F}}^{\phi}$ 的定义得出的，因为它是贝尔曼残差算子 $\mathcal{B}^{\mathcal{H}}$ 的零点。

结合（77）和（75），我们得到

$$\sum_{h=1}^{H}\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X_{\hat{V},h}^{\phi})^2-(X_{\hat{V},h}^{\phi}-(\mathcal{B}^{\mathcal{H}}\hat{V})(\tau_h))^2)]\le L_{\mathcal{E}}\varepsilon+\frac{1350HC_{\mathcal{V}}^2C_\mu\|w^{\phi_T}\|_\infty}{n}\cdot\log\frac{4|\mathcal{V}||\Theta|}{\delta} \tag{78}$$

下一步与 [Zhang 和 Jiang, 2024], 中的公式（17）完全相同，借助伯恩斯坦不等式可得，对于任意 $V\in\mathcal{V}$，其成立的概率大于 $1-\delta/2$,

$$\begin{aligned}&\left|\sum_{h=1}^{H}\{\mathbb{E}_{\mathcal{D}}-\mathbb{E}_{\pi_b^{\phi}}\}[w^{\phi_T}((X_{V,h}^{\phi})^2-(X_{V,h}^{\phi}-(\mathcal{B}^{\mathcal{H}}V)(\tau_h))^2)]\right|\\&\le\sqrt{\frac{58HC_{\mathcal{V}}^2C_\mu\|w^{\phi_T}\|_\infty}{n}\cdot\log\frac{4|\mathcal{V}|}{\delta}\cdot\sum_{h=1}^{H}\mathbb{E}_{\pi_b^{\phi}}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2]}+\frac{27HC_{\mathcal{V}}^2C_\mu\|w^{\phi_T}\|_\infty}{n}\cdot\log\frac{4|\mathcal{V}|}{\delta}.\end{aligned} \tag{79}$$

因此，结合（78）和（79），我们得到

$$\begin{aligned}\sum_{h=1}^{H}\mathbb{E}_{\pi_b^{\phi}}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2]&\le L_{\mathcal{E}}\varepsilon+\frac{1377HC_{\mathcal{V}}^2C_\mu\|w^{\phi_T}\|_\infty}{n}\cdot\log\frac{4|\mathcal{V}||\Theta|}{\delta}\\&\quad+\sqrt{\frac{58HC_{\mathcal{V}}^2C_\mu\|w^{\phi_T}\|_\infty}{n}\cdot\log\frac{4|\mathcal{V}||\Theta|}{\delta}\cdot\sum_{h=1}^{H}\mathbb{E}_{\pi_b^{\phi}}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2]},\end{aligned} \tag{80}$$

求解该式后，得到最终结果:

$$\begin{aligned}\sum_{h=1}^{H}\mathbb{E}_{\pi_b^{\phi}}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2]&\le\frac{1406HC_{\mathcal{V}}^2C_\mu\|w^{\phi_T}\|_\infty}{n}\cdot\log\frac{4|\mathcal{V}||\Theta|}{\delta}+L_{\mathcal{E}}\varepsilon\\&+\sqrt{\frac{80707H^2C_{\mathcal{V}}^4C_\mu^2\|w^{\phi_T}\|_\infty^2}{n^2}\cdot\left(\log\frac{4|\mathcal{V}||\Theta|}{\delta}\right)^2+\frac{29HC_{\mathcal{V}}^2C_\mu\|w^{\phi_T}\|_\infty}{n}\cdot\log\frac{4|\mathcal{V}||\Theta|}{\delta}\cdot2L_{\mathcal{E}}\varepsilon}\qquad(81)\\&\le\frac{(1406+\sqrt{80707+29C})HC_{\mathcal{V}}^2C_\mu\|w^{\phi_T}\|_\infty}{n}\cdot\log\frac{4|\mathcal{V}||\Theta|}{\delta}+L_{\mathcal{E}}\varepsilon\qquad(82)\end{aligned}$$

其中，最后一步要求 $2L_{\mathcal{E}}\varepsilon\le\frac{CHC_{\mathcal{V}}^2C_\mu\|w^{\phi_T}\|_\infty}{n}\cdot\log\frac{4|\mathcal{V}||\Theta|}{\delta}$。注意，在条件2下，$\|w^{\phi_T}\|_\infty\le e$，因此该要求由条件3涵盖。

接下来，利用贝尔曼残差算子的望远镜性质以及前文提到的事实 $\|w^{\phi_T}\|_\infty\le e$ as，即可证明（70）。

既然我们已经得到（70），只需利用定理 8 的结果即可证明该定理，该结果表明:

$$|J(\pi_e)-J^{\phi}(\pi_e^{\phi})|+|\mathbb{E}_{\pi_b}[\hat{V}(f_1)]-\mathbb{E}_{\pi_b^{\phi}}[\hat{V}(f_1)]|\le L_\phi\varepsilon. \tag{83}$$

我们通过应用元定理3来证明该定理。请注意，应用元定理时，我们指定 $\mathfrak{est}^{\phi}(\hat{}V)=\mathbb{E}_{\pi_b^{\phi}}[\hat{}V(f_1)]$ 和 $\mathfrak{est}^{\phi}(V_\phi^{\pi})=J^{\phi}(\pi_e^{\phi})$，后者是抽象系统上的真实估计。

□

推论 3。 在满足定理 *9* 的条件下，以大于 $1-\delta$ 的概率，我们有:

$$\begin{aligned}|J(\pi_e)-\mathbb{E}_{\pi_b}[\hat{V}(f_1)]|\le\inf_{\substack{\varepsilon\ge0\\D(\varepsilon)}}\Bigg(&L_\phi\varepsilon+\sqrt{H}\cdot\max_{h\in[H]}\sup_{V\in\mathcal{V}}\sqrt{\frac{\mathbb{E}_{\pi_e^{\phi}}[(\mathcal{B}^{(\mathcal{S},\mathcal{H}_T)}V)(s_h,\tau_{[h-T+1:h]})^2]}{\mathbb{E}_{\pi_b^{\phi}}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2]}}\\&\cdot\sqrt{\frac{cHC_{\mathcal{V}}^2C_\mu}{n}\log\frac{|\mathcal{V}||\Theta|}{\delta}+L_{\mathcal{E}}\varepsilon}\Bigg)\end{aligned} \tag{84}$$

其中 $D(\varepsilon)$ 表示满足抽象可实现性、贝尔曼完备性以及条件 *1*、*2* 和 *3* 的此类 ε 。

证明。 结合引理3和定理 9，可得该结果。 □

推论 2 的证明

证明。 在定理 9 中选择 $L_{\mathcal{E}}\varepsilon = \frac{c' H C_{\mathcal{V}}^2 C_\mu}{n} \log \frac{|\mathcal{V}||\Theta|}{\delta}$，其中 c' 为某个常数，即可得到该结果。 □

D.3 更简单的流程：仅抽象策略

定理 10 的证明。

证明。 新的 L_ϕ 通过结合定理 12 与定理 8 得到，新的 $L_{\mathcal{E}}$ 与引理12的结果类似。证明的其余部分与定理 9 的证明完全相同。 □

E 为什么我们的覆盖更好

例1的详细说明 在本例中，我们考虑一个具有光滑性结构的信念空间（第 5.3 [节，Lee 等人， 2007]），其定义如下:

$\mathcal{B}$ 是 $|\mathcal{S}|$维向量空间中的有界子集，假设每个信念 $b \in \mathcal{B}$ 都可以表示为 m 个基向量的线性组合，且基向量和线性系数的大小均有常数 C作为上界。那么，我们的信念空间的覆盖数按$O((C|\mathcal{S}|L_{\mathcal{E}}m)^m \cdot (\frac{32R^4}{\text{最大值} n(1-\gamma)^4} \cdot \text{对数} \frac{2|\mathcal{F}|}{\delta}) \frac{m}{2})$的尺度增长。我们假设覆盖相对于其最坏情况（即覆盖数）呈次线性多项式增长，具体而言，其幂次为 $\frac{1}{2m}$。因此，我们得到有限样本保证$O(\frac{(C|\mathcal{S}|L_{\mathcal{E}}mR^2}{\text{最大值}) \frac{1}{4}(1-\gamma) \frac{3}{2}} \cdot (\frac{1}{n} \text{对数} \frac{|\mathcal{F}|}{\delta}) \frac{1}{8})$。请注意，我们只假设次线性多项式增长，而非对数增长，因为后者的条件过强，可能会直接消除指数增长。

定理 4 的证明

证明。我们通过构造 $d^D(\tau') = \sum_{\{\tau : \tilde{\phi}_T(\tau) = \tau'\}} d^{\pi_b}(\tau)$ 来证明该定理。然后注意到， $d_\phi^{\pi_e^\phi}(\tau') = \sum_{\{\tau : \tilde{\phi}_T(\tau) = \tau'\}} d^{[\pi_e^\phi]_{\text{true}}}(\tau)$ 根据抽象 ϕ_T 的定义会自动满足。此外，不难注意到，在独热信念状态场景中，$\frac{d^{[\pi_e^\phi]\text{true}}(s_h, \tau_h)}{d^{\pi_b}(s_h, \tau_h)} = \frac{d^{[\pi_e^\phi]_{\text{true}}}(\tau_h)}{d^{\pi_b}(\tau_h)}$，对于由我们所构造的 ϕ_T 诱导的短期记忆部分可观测马尔可夫决策过程，情况完全相同。这里， $\phi_T : \mathcal{H} \to \mathcal{H}_T \subset \mathcal{H}$。

接下来，考虑两个 σ-代数 $\mathcal{A} := \mathcal{P}(\mathcal{H})$ 和 $\mathcal{D} := \tilde{\phi}_T^{-1}(\mathcal{P}(\mathcal{H}_T))$，显然有 $\mathcal{D} \subset \mathcal{A}$。定义概率点测度 P^{π_e} 和 P^{π_b}，它们分别对应于 σ-代数$\mathcal{A}$上的权重函数 $d^{[\pi_e^\phi]_{\text{true}}}$ 和 d^{π_b}，那么该概率测度也可以限制到更小的 σ-代数 $\mathcal{D}$ 上。容易注意到，我们试图比较的两个项与 χ^2-散度一致，其中对于左侧，我们使用较粗的 σ-代数 $\mathcal{D}$，对于右侧则使用更细的 σ-代数 $\mathcal{A}$。

然后，我们利用 χ^2-散度的变分表示，并注意到

$$\chi^2_{\mathcal{D}}(P^{\pi_e} \| P^{\pi_b}) = \sup_{g \in \mathcal{M}(\mathcal{D})} \mathbb{E}_{P^{\pi_e}}[g(\tau_h)] - \mathbb{E}_{P^{\pi_b}}[g(\tau_h)^2/4 + g(\tau_h)] \tag{85}$$

$$\chi^2_{\mathcal{A}}(P^{\pi_e} \| P^{\pi_b}) = \sup_{g \in \mathcal{M}(\mathcal{A})} \mathbb{E}_{P^{\pi_e}}[g(\tau_h)] - \mathbb{E}_{P^{\pi_b}}[g(\tau_h)^2/4 + g(\tau_h)] \tag{86}$$

由于 $\mathcal{D} \subset \mathcal{A}$，任何 $\mathcal{D}$ 可测的 g 也是 $\mathcal{A}$ 可测的，因此

$$\chi^2_{\mathcal{D}}(P^{\pi_e} \| P^{\pi_b}) \leq \chi^2_{\mathcal{A}}(P^{\pi_e} \| P^{\pi_b}) \tag{87}$$

这证明了该定理。 □

定理 5 的证明

证明。按照定理 4 中的方式构造 d^D，然后令 $w^\star(\tau_h) = d^{[\pi_e^\phi]\text{true}}(\tau_h)\, d^D(\tau_h)$，并且 $\tau_h^\star$ 是在取得最大值时。类似地，令 $w^*(\tilde{\phi}_T(\tau)) = d^{\pi_e^\phi}(\tilde{\phi}_T(\tau))\, d^D(\tilde{\phi}_T(\tau))$，且 $\tilde{\phi}_T(\tau^*)$ 是在取得最大值时。显然，$\tilde{}$ 使得 $\forall \tau_h$ 满足 $\phi_T(\tau_h) = \phi_T(\tau^*)$、$w^\star(\tau_h) \le w^\star(\tau_h^\star)$。记 $\tau_h' := \arg\max_{\tilde{\phi}_T(\tau_h)=\tilde{\phi}_T(\tau^*)} w^\star(\tau_h)$，则 $w^*(\tilde{\phi}_T(\tau^*)) = \sum_{\tilde{\phi}_T(\tau_h)=\tilde{\phi}_T(\tau^*)} d^{[\pi_e^\phi]\text{true}}(\tau_h) \sum_{\tilde{\phi}_T(\tau_h)=\tilde{\phi}_T(\tau^*)} d^D(\tau_h)$。注意到

$$\frac{\sum_{\tilde{\phi}_T(\tau_h)=\tilde{\phi}_T(\tau^*)} d^{[\pi_e^\phi]_{\text{true}}}(\tau_h)}{\sum_{\tilde{\phi}_T(\tau_h)=\tilde{\phi}_T(\tau^*)} d^D(\tau_h)} \le \frac{\sum_{\tilde{\phi}_T(\tau_h)=\tilde{\phi}_T(\tau^*)} d^D(\tau_h)\cdot d^{[\pi_e^\phi]_{\text{true}}}(\tau_h')/d^D(\tau_h')}{\sum_{\tilde{\phi}_T(\tau_h)=\tilde{\phi}_T(\tau^*)} d^D(\tau_h)} = w^\star(\tau_h')$$

$\tilde{}$因此，$w^*(\phi_T(\tau^*)) \le w^\star(\tau_h') \le w^\star(\tau_h^\star)$，这证明了该定理。 □

F 与深度抽象的关系

本节讨论我们的方法与显式构建抽象的离线策略评估方法之间的关系。我们的方法纯粹将抽象用作分析工具：我们分析现有算法而不对其进行修改，只是揭示这些算法何时以及如何因信念空间平滑性而获得更优的保证。相反，其他一些方法会主动构建抽象，以简化离线策略评估。在这种情况下，在抽象系统（因此比原始系统更简单）上运行的算法可能获得更小的误差保证。本节将简要比较我们的思路与深度抽象 [Hao 等人， 2024]。

在 [Hao 等人， 2024], 中，他们设计了一种方法，利用传统抽象技术在马尔可夫决策过程中构建深度抽象，即递归地应用两种不同的抽象方法，以获得深度抽象。并且可以证明，随着抽象层级加深，抽象系统的方差单调递减。

与我们的设定比较。

1. **抽象类型与严格程度的差异：**本文中的抽象要求在每一步都满足前向模型无关条件或后向模型无关条件。众所周知，双模拟无论是精确版本还是近似版本，都是一个很难满足的强条件；在基数呈指数级且结构有限的信念空间中，这一条件尤其严格。此外，由于采用的是传统的抽象方法，它不要求状态空间具有度量结构。相比之下，我们的抽象基于信念空间上的 ε-网，利用了该空间的度量几何，并且无论部分可观测马尔可夫决策过程是否满足结构性假设，都能统一应用于广泛的此类问题。

2. **解决时域诅咒：**事实上， [Hao 等人 2024] 优雅地证明了，均方误差会随着抽象层级加深而单调下降。然而，要通过抽象应对记忆/时域诅咒，就必须分析抽象空间中的覆盖如何改善。值得注意的是，将他们的分析直接应用于部分可观测马尔可夫决策过程后可以发现，假设 2 和 4 在 $O(1)$ 中隐含了一个指数常数。该常数源于函数类 $\mathcal{W}$ 的有界性，其中包括边缘化重要性采样比 $\hat{}w^\pi$，并且根据假设 2，该比值被假定为有限。在不存在时域/记忆诅咒的马尔可夫决策过程中，这样的处理是可以接受的；但在部分可观测马尔可夫决策过程中，关键在于考察抽象如何影响这一指数项。

G 未来算法设计

虽然本文侧重于理论框架，但从稳定性的视角出发，我们的分析自然启发了未来工作的具体算法思路。

1. **稳定性正则化训练：** 在贝尔曼误差最小化或价值函数拟合中加入额外的惩罚项

$$\lambda \mathbb{E}_{\mathcal{D}^{\otimes 2}}[\mathbb{I}(\|\hat{b}_1 - \hat{b}_2\|_1 \le \epsilon)\cdot |V(\hat{b}_1) - V(\hat{b}_2)|].$$

2. **训练后稳定性选择：** 训练多个候选策略，然后选择在信念邻域上测得经验稳定性最高的策略。从理论上讲，这等同于上述惩罚方法，如 $\lambda \to 0$。

H 局限性

尽管我们的通用结果已被证明不会比原始覆盖假设更差，但在某些情况下，信念空间的度量性质也可能无法改善覆盖。最简单的情形是，每段历史都对应一个唯一的独热信念状态，而该部分可观测马尔可夫决策过程仅等价于一个具有指数级状态空间的马尔可夫决策过程。在这种情况下，信念度量是离散度量，对于$\forall b_1$， $b_2 \in \mathcal{B}$，$b_1 \neq b_2 \to \|b_1 - b_2\|_1 = 2$，覆盖数恰好等于该空间的基数，因而呈指数级。这揭示了我们的分析在信念空间稀疏或缺乏某些特定光滑性结构时的局限性。然而，从信息论角度来看，如 [Zhang 和 Jiang、 2025]，所示，部分可观测马尔可夫决策过程的离策略评估问题在最一般的情形下始终受到时域诅咒的影响，这意味着必须利用结构假设或系统的特定性质，才能取得有意义的进展。

另一项局限是，样本量相对于时域 H 过大。请注意，在我们的结果（例如推论 1）所给出的有限样本论证中，抽象覆盖取决于近似水平 ε，而该水平设为 $O(n^{-1})$。在这种情况下，如果 n 变得过大， $O(n^{-1})$-覆盖数将趋近于空间 $\mathcal{B}$ 本身的基数，而该基数相对于时域 H 呈指数增长。这也会使我们的分析失去意义。因此，在考虑有限时域部分可观测马尔可夫决策过程时，时域应相对于样本量足够大，才能使我们的结果成立。