arXiv:2302.01333v1 [cs.LG] 2 Feb 2023

揭示型部分可观测马尔可夫决策过程中的学习下界

Fan Chen* Huan Wang† Caiming Xiong† Song Mei‡ Yu Bai†

2023年2月3日

摘要

本文研究了具有挑战性的 *partiallyobservable* 设定下强化学习（RL）的基本极限。尽管人们早已认识到，在最坏情况下，部分可观测马尔可夫决策过程（POMDP）中的学习需要指数级数量的样本，但近期大量研究表明，在 *revealing condition*条件下可以实现多项式级样本复杂度——这是一种自然条件，要求可观测量能够揭示有关未观测潜在状态的一些信息。然而，对于揭示型部分可观测马尔可夫决策过程中的学习，其基本极限仍远未得到充分了解；现有下界相当初步，与当前最佳上界之间还存在显著差距。

我们为揭示型部分可观测马尔可夫决策过程中的学习建立了有力的 PAC 下界和遗憾下界。我们的下界以乘法形式对所有相关问题参数呈多项式级依赖，并显著缩小了与当前最佳上界之间的差距，为未来研究奠定了坚实基础。具体而言，对于 *multi-step* 揭示型部分可观测马尔可夫决策过程，我们证明：(1) PAC 样本复杂度对潜在状态空间的依赖至少为 $\Omega(S^{1.5})$ ，明显比完全可观测马尔可夫决策过程中的 $\widetilde{\Theta}(S)$ 标度更难；(2) 任何多项式级次线性遗憾至少为 $\Omega(T^{2/3})$, ，这表明它与 *single-step* 情形存在根本差异，在该情形下可以实现$\widetilde{\mathcal{O}}(\sqrt{T})$ 遗憾。从技术上讲，我们的困难实例构造改编自 *distribution testing*中的技术；这一构造在强化学习文献中尚属首次，也可能具有独立研究价值。

1 引言

部分可观测性——即智能体只能观察到有关系统真实底层状态的部分信息——在强化学习（RL）的现实应用中无处不在，也是强化学习面临的核心挑战之一（Kaelbling等人，1998；Sutton和Barto，2018）。已知，在标准的部分可观测马尔可夫决策过程（POMDP）模型中进行学习，比在其完全可观测对应模型中困难得多——在最坏情况下，要在长时域部分可观测马尔可夫决策过程中找到近最优策略，所需样本数至少随时域长度呈指数增长（Krishnamurthy等人，2016）。这种指数级难度源于这样一个事实：若不对部分可观测马尔可夫决策过程的结构施加进一步限制，智能体可能无法观测到有关系统真实底层状态的任何有用信息。这与完全可观测（表格型）马尔可夫决策过程中的学习形成鲜明对比：无需进一步假设，多项式数量的样本便已充分且必要（Kearns和Singh，2002；Jaksch等人，2010；Azar等人，2017；Jin 等人，2018；Zhang等人，2020 年；Domingues等人，2021）。

为克服这一困难性结果，近期研究致力于寻找能够实现样本高效学习的额外结构条件。一种自然的方案是揭示条件（Jin 等人，2020a；Liu 等人，2022a），从宏观上说，该条件要求可观测量（观测和动作）能够揭示有关底层潜在状态的某些信息，从而排除前述可观测量完全无法提供信息的最坏情况。具体而言，单步揭示条件（Jin 等人，2020a）要求潜在状态的（即时）发射概率具有良态性，也就是说，不同状态能够通过其发射结果在概率上加以区分。多步揭示条件（Liu

*北京大学。邮箱： chern@pku.edu.cn
†Salesforce Research。邮箱： {huan.wang,cxiong,yu.bai}@salesforce.com
‡加州大学伯克利分校。邮箱： songmei@berkeley.edu

表1：揭示型部分可观测马尔可夫决策过程学习的下界与当前最佳上界汇总，我们的贡献以灰色单元格突出显示。此处给出的速率仅关注对 S、O、A、α^{-1} 和 T（或 ε^{-1}）的依赖，省略了 $\mathrm{poly}(H)$ 以及所有多对数因子。为简化表述，我们还假设 $O \geqslant \Omega(SA)$（在我们的上界中）以及 $A^H \gg \mathrm{poly}(H、S、O、A^m、\alpha^{-1}、T)$。对于遗憾下界，我们还忽略了与 T 取最小值的情况（因为存在平凡的 $O(T)$ 遗憾上界）。 *通过先探索后利用转换获得。

问题	PAC 样本复杂度		遗憾	
	上界	下界	上界	下界
一步 α-揭示	$\tilde{\mathcal{O}}\left(\frac{S^2OA}{\alpha^2\varepsilon^2}\right)$ (Chen 等人，2022a)	$\Omega\left(\frac{SO^{1/2}A}{\alpha^2\varepsilon^2}\right)$ （定理 4）	$\tilde{\mathcal{O}}\left(\sqrt{\frac{S^2O^2A}{\alpha^2}\cdot T}\right)$ （定理 8）	$\Omega\left(\sqrt{\frac{SO^{1/2}A}{\alpha^2}\cdot T}\right)$ （推论 7）
m-步骤（$m \geqslant 2$） α-揭示	$\tilde{\mathcal{O}}\left(\frac{S^2OA^m}{\alpha^2\varepsilon^2}\right)$ (Chen 等人，2022a)	$\Omega\left(\frac{(S^{3/2}+SA)O^{1/2}A^{m-1}}{\alpha^2\varepsilon^2}\right)$ （定理5）	$\tilde{\mathcal{O}}\left(\left(\frac{S^2OA^m}{\alpha^2}\right)^{1/3}T^{2/3}\right)$ (Chen 等人，2022a)*	$\Omega\left(\left(\frac{SO^{1/2}A^m}{\alpha^2}\right)^{1/3}T^{2/3}\right)$ （定理6）

等人，2022a）将单步情形推广开来，要求多步发射—动作概率具有良态性——即在当前潜在状态下，给定采取特定动作序列后，接下来 $m \geqslant 2$ 步中观测到某个观测序列的概率。

Jin 等人（2020a）和 Liu 等人（2022a）最初设计了用于学习单步和多步揭示型部分可观测马尔可夫决策过程的样本高效算法，此后近期大量研究进一步发展了这些算法（Cai 等人，2022；Wang 等人，2022；Uehara 等人，2022b；Zhan 等人，2022；Chen 等人，2022a；Liu 等人，2022b；Zhong 等人，2022）。对于在 m 步揭示型部分可观测马尔可夫决策过程中寻找 ε 近最优策略，这些结果给出的 PAC 样本复杂度（所需的交互回合数）随状态数、观测数、动作序列数（长度为 m）、时域以及 $(1/\alpha)$ 呈多项式级增长，其中 $\alpha > 0$ 为揭示常数，且 $(1/\varepsilon)$, 当前最佳界由 Chen 等人（2022a）给出。

尽管取得了这些进展，但揭示型部分可观测马尔可夫决策过程中的学习基本极限仍然知之甚少。首先，目前关于揭示型部分可观测马尔可夫决策过程的下界研究十分有限：现有下界要么在其速率方面仍较为初步（Liu 等人，2022a），要么通过从完全可观测设置直接归约得到，因而无法体现部分可观测性的挑战（详见第 2.2 节的详细讨论）。这类下界留下了许多基本问题，例如最优 PAC 样本复杂度对 α 的依赖关系：当前最好的下界按 α^{-1} 进行缩放，而当前最好的上界则要求 α^{-2}。其次，目前关于揭示型部分可观测马尔可夫决策过程学习的最佳上界，大多是通过并非专门针对部分可观测马尔可夫决策过程设计的通用算法获得的（Chen 等人，2022a；Liu 等人，2022b；Zhong 等人，2022）。这些算法为包括揭示型部分可观测马尔可夫决策过程在内的大量强化学习问题提供了统一的分析框架，但尚不清楚这些分析（以及由此得到的上界）是否揭示了揭示型部分可观测马尔可夫决策过程的基本极限。

本文为学习揭示型部分可观测马尔可夫决策过程建立了强有力的样本复杂度下界。我们的贡献可概括如下。

- 我们为学习单步（第 3.1 节）和多步（第 3.2 节）揭示型部分可观测马尔可夫决策过程建立了 PAC 下界。我们的下界首次以乘法形式随所有相关问题参数共同变化，并解答了关于学习揭示型部分可观测马尔可夫决策过程基本极限的若干开放问题。值得注意的是，我们为多步情形建立的 PAC 下界按 $\Omega(S^{1.5})$ 缩放，其中 S 是~潜在状态空间的大小，这显著难于完全可观测的马尔可夫决策过程；后者的 $\Theta(S)$ 是极小极大最优标度。此外，我们的下界与当前最佳上界之间的差距相当小，可为进一步的细粒度研究提供起点。

- 我们为相同情形建立了遗憾下界。或许令人意外的是，我们证明了多步揭示型部分可观测马尔可夫决策过程存在 $\Omega(T^{2/3})$ 遗憾下界（第 4 节）。我们的构造揭示了关于多步情形的一些新见解，并表明它与单步情形存在根本差异；在单步情形中，$\tilde{\mathcal{O}}(\sqrt{T})$ 遗憾是可实现的。

- 严格来说，我们的 l 下界通过将均匀性检验问题嵌入揭示型 g

部分可观测马尔可夫决策过程，尤其是嵌入 m步揭示型组合锁来获得；这是我们困难实例构造的核心（第5节）。证明还使用了信息论技术，例如 Ingster 方法来界定某些散度；这些技术在强化学习文献中尚属新颖。

- 我们将讨论对一般强化学习理论，尤其是决策估计系数（DEC）框架（第6.2节）的一些其他有趣影响。

我们在表1中，将主要结果与当前最佳上界进行比较。

1.1 相关工作

一般部分可观测马尔可夫决策过程的学习难度 众所周知，在最坏情况下，学习部分可观测马尔可夫决策过程中的近最优策略在计算上是困难的（Papadimitriou 和 Tsitsiklis，1987；Mossel 和 Roch，2005）。关于学习，Krishnamurthy 等人（2016）和 Jin 等人（2020a）利用组合锁这一困难实例证明，学习情节式部分可观测马尔可夫决策过程所需的样本量至少随时域 H 呈指数级增长。Kearns 等人（1999）和 Even-Dar 等人（2005）提出了用于学习情节式部分可观测马尔可夫决策过程的算法，其样本复杂度随 A^H 缩放。通过界定贝尔曼秩（Jiang 等人，2017；Du 等人，2021；Jin 等人，2021）或可覆盖性（Xie 等人，2022），也可以得到类似的样本复杂度。

揭示型部分可观测马尔可夫决策过程 Jin 等人(2020a) 在欠完备部分可观测马尔可夫决策过程中提出了单步揭示条件，并通过设计一种谱型学习算法证明，该条件足以实现部分可观测马尔可夫决策过程的样本高效学习。 Liu 等人 (2022a,c) 将多步揭示条件引入过完备部分可观测马尔可夫决策过程，并开发了乐观最大似然估计（OMLE）算法，以实现高效学习。 Cai 等人 (2022); Wang 等人 (2022) 将这些结果扩展到在揭示条件的若干变体下高效学习线性部分可观测马尔可夫决策过程。 Golowich 等人 (2022b,a) 证明，在可观测条件（揭示条件的一种变体）下进行近似规划，可以采用准多项式时间算法。

目前，关于揭示型部分可观测马尔可夫决策过程学习的唯一已有下界由 Liu 等人 (2022a) 给出。他们改造了 Krishnamurthy等人 (2016) 的组合锁困难实例，构造出一个 m步 1步揭示部分可观测马尔可夫决策过程，并证明其具有 $\Omega(A^{m-1})$的样本复杂度下界，用于学习一个 $1/2$-最优策略。我们的下界通过一种复杂得多的困难实例构造，大幅改进了他们的结果；该构造将组合锁与用于学习马尔可夫决策过程的树形困难实例（Domingues等人，2021）以及均匀性检验的困难实例（Paninski，2008 年；Canonne，2020 年）结合起来。与均匀性检验的下界类似，我们下界的证明建立在 Ingster 方法（Ingster 和 Suslina，2012）之上。

其他结构条件 其他能够实现部分可观测马尔可夫决策过程样本高效学习的条件包括反应性（Jiang 等人，2017）、可解码性（Efroni 等人，2022）、结构化潜在马尔可夫决策过程（Kwon 等人，2021）、学习短记忆策略（Uehara 等人，2022b）、确定性转移（Uehara 等人，2022a）以及正则预测状态表示（PSR）（Zhan 等人，2022）。Chen 等人 (2022a)；Liu 等人 (2022b)；Zhong 等人 (2022) 为预测状态表示提出了统一结构条件，涵盖了大多数现有的可处理类别，包括揭示型部分可观测马尔可夫决策过程、可解码部分可观测马尔可夫决策过程和正则预测状态表示。

2 预备知识

部分可观测马尔可夫决策过程（POMDP） 一个分回合的部分可观测马尔可夫决策过程（POMDP）由元组$M = \{H, \mathcal{S}, \mathcal{O}, \mathcal{A}, \{\mathbb{T}_h\}_{h\in[H]}, \{\mathbb{O}_h\}_{h\in[H]}, \{r_h\}_{h\in[H]}, \mu_1\}$定义，其中 $H \in \mathbb{Z}_{\geqslant 1}$为时域长度； $(\mathcal{S}, \mathcal{O}, \mathcal{A})$ 分别为（潜在）状态、观测和动作空间，其基数分别为 (S, O, A)； $\mathbb{O}_h(\cdot|\cdot):\mathcal{S} \to \Delta(\mathcal{O})$ 为步骤 h 的发射动态（我们将其表示为发射矩阵 $\mathbb{O}_h \in \mathbb{R}^{\mathcal{O}\times\mathcal{S}}$）； $\mathbb{T}_h(\cdot|\cdot,\cdot)\colon \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ 为潜在状态上的转移动态（我们将其表示为转移矩阵 $\mathbb{T}_h \in \mathbb{R}^{\mathcal{S}\times(\mathcal{S}\times\mathcal{A})}$）； $r_h(\cdot,\cdot)\colon \mathcal{O}\times\mathcal{A} \to [0, 1]$为（可能是随机的）奖励函数； $\mu_1 = \mathbb{T}_0(\cdot) \in \Delta(\mathcal{S})$指定初始状态的分布。在每个步骤 $h \in [H]$，给定潜在状态 s_h（智能体无法观测该状态），系统会发出观测 $o_h \sim \mathbb{O}_h(\cdot|s_h)$，接收智能体给出的动作 $a_h \in \mathcal{A}$，发出奖励 $r_h(o_h, a_h)$，然后以马尔可夫方式转移到下一个潜在状态 $s_{h+1} \sim \mathbb{T}_h(\cdot|s_h, a_h)$。

我们用 $\tau = (o_1, a_1, \ldots, o_H, a_H) = (o_{1:H}, a_{1:H})$ 表示智能体观测到的完整观测与动作历史，并用 $\tau_h = (o_{1:h}, a_{1:h})$ 表示截至步骤 $h \in [H]$的部分历史。策略由一组动作分布 $\pi = \{\pi_h(\cdot|\tau_{h-1}, o_h) \in \Delta(\mathcal{A})\}_{h,\tau_{h-1},o_h}$给出，其中 $\pi_h(\cdot|\tau_{h-1}, o_h)$ 指定在历史 (τ_{h-1}, o_h)条件下 a_h 的分布。我们将 Π 记为所有策略的集合。任意策略 π的价值函数记为 $V_M(\pi) = \mathbb{E}_M^\pi[\sum_{h=1}^H r_h(o_h, a_h)]$，其中 $\mathbb{E}_M^\pi$指定模型 M和策略 π下 $(o_{1:H}, a_{1:H})$的分布律。模型 M的最优价值函数记为 $V_M^\star = \max_{\pi\in\Pi} V_M(\pi)$。不失一般性，我们假设总奖励不超过1，即对于任意 $(o_{1:H}, a_{1:H}) \in (\mathcal{O} \times \mathcal{A})^H$，均有 $\sum_{h\in[H]} r_h(o_h, a_h) \leqslant 1$。

学习目标 我们考虑在老虎机反馈（探索设定）下学习部分可观测马尔可夫决策过程：智能体在固定的（未知）部分可观测马尔可夫决策过程模型 M 下进行 $T \in \mathbb{N}_+$ 回合交互。在每个回合中，智能体执行某个策略 $\pi^{(t)}$，并观测轨迹 $\tau^{(t)}$和奖励 $r_1^{(t)}$。

我们考虑两种标准的学习目标：PAC 学习和无遗憾学习。在 PAC 学习中，目标是输出一个近最优策略 $\hat{\pi}$，使得 $V_M^\star - V_M(\hat{\pi}) \leqslant \varepsilon$，并尽可能减少交互回合数。在无遗憾学习中，目标是最小化遗憾

$$\mathbf{Regret}(T) := \sum_{t=1}^T \left(V_M^\star - V_M\left(\pi^{(t)}\right)\right),$$

，若某个算法满足以下条件，则称其为无遗憾算法：**遗憾**$(T) = o(T)$ 关于 T是次线性的。众所周知，无遗憾学习并不比 PAC 学习容易，因为任何无遗憾算法都可以通过标准的在线到批处理转换（例如，Jin 等人 (2018）转换为 PAC 学习算法，该算法输出平均策略 $\bar{\pi} := \frac{1}{T}\sum_{t=1}^T \pi^{(t)}$经过 T 个交互回合后。

2.1 揭示型部分可观测马尔可夫决策过程

我们考虑揭示型部分可观测马尔可夫决策过程（Jin 等人，2020a；Liu 等人，2022a），这是一类结构化的部分可观测马尔可夫决策过程子类，已知其可以样本高效地学习。对于任意 $m \geqslant 1$，定义部分可观测马尔可夫决策过程 M 在步骤 $h \in [H - m + 1]$ 的 m-步发射-动作矩阵$\mathbb{M}_{h,\,m} \in \mathbb{R}^{\mathcal{O}^m\mathcal{A}^{m-1}\times\mathcal{S}}$ 为

$$[\mathbb{M}_{h,m}]_{(\mathbf{o},\mathbf{a}),s} := \mathbb{P}_M(o_{h:h+m-1} = \mathbf{o}|s_h = s, a_{h:h+m-2} = \mathbf{a}). \tag{1}$$

在特殊情形下，若 $m = 1$ （单步情形），则有 $\mathbb{M}_{h,\,1} = \mathbb{O}_h \in \mathbb{R}^{\mathcal{O}\times\mathcal{S}}$，即发射-动作矩阵退化为发射矩阵。对于 $m \geqslant 2$， m-步发射-动作矩阵 $\mathbb{M}_{h,\,m}$ 通过编码发射-动作概率推广了发射矩阵，即从任意潜在状态 $s \in \mathcal{S}$ 出发，并在接下来的$m - 1$ 步中采取任意动作序列 $\mathbf{a} \in \mathcal{A}^{m-1}$ 时，观测到任意观测序列 $\mathbf{o} \in \mathcal{O}^m$的概率。

如果一个部分可观测马尔可夫决策过程的发射-动作矩阵 $\{\mathbb{M}_{h,m}\}_{h\in[H-m+1]}$ 存在算子范数有界的广义左逆，则称其为 m步揭示型。

定义 1 (m步 α揭示型部分可观测马尔可夫决策过程)。 对于 $m \geqslant 1$ 和 $\alpha > 0$，如果存在矩阵 $\mathbb{M}_{h\,,m}^+ \in \mathbb{R}^{\mathcal{S}\times\mathcal{O}^m\mathcal{A}^{m-1}}$，使其满足 $\mathbb{M}_{h\,,m}^+\mathbb{M}_{h,m}\mathbb{T}_{h-1} = \mathbb{T}_{h-1}$ （$\mathbb{M}_{h,m}$的广义左逆），且对任意 $h \in [H - m + 1]$均成立，则称部分可观测马尔可夫决策过程模型 M 为 m步揭示型。此外，如果每个 $\mathbb{M}_{h\,,m}^+$ 还存在范数不超过 α^{-1}的 $(* \to 1)$-算子范数，则称部分可观测马尔可夫决策过程模型 M 为 m步 α揭示型：

$$\|\mathbb{M}_{h,m}^+\|_{*\to 1} := \max_{\|\mathbf{x}\|_* \leqslant 1} \|\mathbb{M}_{h,m}^+\mathbf{x}\|_1 \leqslant \alpha^{-1}, \tag{2}$$

其中，对于任意向量 $\mathbf{x} = (\mathbf{x}(\mathbf{o},\mathbf{a}))_{\mathbf{o}\in\mathcal{O}^m,\mathbf{a}\in\mathcal{A}^{m-1}}$，我们用以下形式表示其星范数

$$\|\mathbf{x}\|_* := \left[\sum_{\mathbf{a}\in\mathcal{A}^{m-1}}\left(\sum_{\mathbf{o}\in\mathcal{O}^m}|\mathbf{x}(\mathbf{o},\mathbf{a})|\right)^2\right]^{1/2}.$$

令 $\alpha_m(M)$——模型 m步揭示常数， M表示满足 $\alpha > 0$ 且使(2)成立的最大可能值，因此 M 是 m步 α揭示，当且仅当 $\alpha_m(M) \geqslant \alpha$。

在定义1中，广义左逆的存在要求矩阵$\mathbb{M}_{h,\,m}$在$\mathbb{T}_{h-1}$的列空间中具有满秩，这确保了从上一步可达的不同状态在信息论意义上可通过下一步m观测和$m-1$动作加以区分。揭示条件作为这一满秩条件的定量版本，确保状态能够从可观测量中以概率方式“揭示”出来，并实现样本高效学习（Liu 等人，2022a）。

在仅考虑多项式可学习性时，由于范数之间的等价性，(2) 中具体范数的选择并不重要。我们选择$(*\to 1)$-范数，这与已有工作（Liu 等人，2022a，b；Chen 等人，2022a）不同；不过，这使我们的下界与已有上界之间的差距更小。

单步与多步 我们强调，当$m=1$时，发射-动作矩阵$\mathbb{M}_{h,\,1}=\mathbb{O}_h$不涉及动作的影响。这使其在性质上有别于$m\geqslant 2$的多步 情形，这一点将在我们的结果中体现出来。

此外，我们证明，任何m-步 α-揭示型部分可观测马尔可夫决策过程也是$(m+1)$-步 α-揭示型的，但反之则不成立（证明见附录 C.1；这一结果虽然直观，但我们未能在文献中找到相关记载）。因此，随着m增大，m-步揭示型部分可观测马尔可夫决策过程的类别会严格扩大，因而学习难度不会降低。

命题 2 (m步揭示型$\subsetneq$$(m+1)$步揭示型)。对于任意$m\geqslant 1$以及任意时域为$H\geqslant m+1$的部分可观测马尔可夫决策过程$M$，我们有$\alpha_{m+1}(M)\geqslant\alpha_m(M)$。因此，任何$m$步 α揭示型部分可观测马尔可夫决策过程也是一个$(m+1)$步 α揭示型部分可观测马尔可夫决策过程。反过来，存在一个$(m+1)$步揭示型部分可观测马尔可夫决策过程，它不是m步揭示型部分可观测马尔可夫决策过程。

2.2 已知上界和下界

上界 已知，学习揭示型部分可观测马尔可夫决策过程可以达到多项式样本复杂度上界（Liu 等人，2022a，b；Chen 等人，2022a）。当前学习揭示型部分可观测马尔可夫决策过程的最佳 PAC 样本复杂度见下述结果，该结果可通过将Chen 等人 (2022a，b) 的结果适配到我们对揭示条件的定义而直接得到（参见附录 C.2）。

定理 3（揭示型部分可观测马尔可夫决策过程的 PAC 上界（Chen 等人，2022a））。存在一些算法（OMLE，Exploratory E2D 和 MOPS），能够以高概率在ε步揭示型部分可观测马尔可夫决策过程中，找到一个m-最优策略，并且该过程有α步，所需回合数不超过

$$T\leqslant\widetilde{\mathcal{O}}\left(\frac{S^2OA^m(1+SA/O)H^3}{\alpha^2\varepsilon^2}\right) \tag{3}$$

交互回合。

下界 现有关于揭示型部分可观测马尔可夫决策过程学习的下界研究稀少且处于初步阶段。目前，唯一适用于m-步 α-揭示型部分可观测马尔可夫决策过程的 PAC 下界是

$$\Omega(\min\left\{\tfrac{1}{\alpha H},A^{H-1}\right\}+A^{m-1})$$

由 Liu 等人（2022a，定理6和9）给出，适用于学习一个$\varepsilon=\Theta(1)$-最优策略，但该下界不随模型参数S、O或$(1/\varepsilon)$变化（当ε较小时）。

此外，揭示型部分可观测马尔可夫决策过程将两种完全可观测模型作为特例：具有H步、最小$\{S$、$O\}$个状态和A个动作的（完全可观测）马尔可夫决策过程（其中$\alpha=1$）；以及具有O个上下文和A个动作的上下文老虎机。根据这两种情形下标准的 PAC 下界（Dann 和 Brunskill，2015；Lattimore 和 Szepesv´ari，2020 年；Domingues等人，2021），[1]，这意味着一个

$$\Omega\big((H\min\{S,O\}A+OA)/\varepsilon^2\big)$$

对于任意$m\geqslant 1$和$\alpha\leqslant 1$，m-步骤 α-揭示型部分可观测马尔可夫决策过程的 PAC 下界。

[1]将总奖励缩放至$[0,\ 1]$。

上述两个下界与上界之间都存在显著差距（3）。事实上，上界随 乘法地 变化，涉及 S、 A^m、 O、 α^{-1} 和 $1/\varepsilon^2$, ，而合并后的下界远小于这种乘法增长。

3 PAC 下界

我们为单步情形（第 3.1 节）和多步情形（第 3.2 节）的揭示型部分可观测马尔可夫决策过程建立 PAC 下界。我们首先陈述并讨论结果，然后在第 5 节概述多步情形的证明。

3.1 单步揭示型部分可观测马尔可夫决策过程

我们首先建立单步情形的 PAC 下界。证明见附录 E。

定理 4 （单步揭示型部分可观测马尔可夫决策过程的 PAC 下界）。 对于任意 $O \geqslant S \geqslant 5$, $A \geqslant 3$, $H \geqslant 4\log_2 S$, $\alpha \in (0, \frac{1}{5H}]$, $\varepsilon \in (0, 0.01]$, 存在一个单步揭示型部分可观测马尔可夫决策过程族 $\mathcal{M}$，满足 $|\mathcal{S}| \leqslant S$, $|\mathcal{O}| \leqslant O$, $|\mathcal{A}| = A$, 且对所有 $M \in \mathcal{M}$ 均有 $\alpha_1(M) \geqslant \alpha$，使得对于任何与环境交互 T 回合并返回一个 π^{out} 的算法 $\mathfrak{A}$，若其以至少 $3/4$ 的概率对所有 $M \in \mathcal{M}$ 满足 $V_M^\star - V_M(\pi^{\text{out}}) < \varepsilon$，则必有

$$T \geqslant c \cdot \min\left\{\frac{SO^{1/2}AH}{\alpha^2\varepsilon^2}, \frac{SA^{H/2}H}{\varepsilon^2}\right\}, \tag{4}$$

其中 $c > 0$ 是一个绝对常数。

定理 4 中的下界（以及后续的下界）涉及两个项的最小值，其中第二项通过指数级缩放[2] 在 H 中的形式“封顶”了下界，因此不那么重要。主项 $\Omega(S\sqrt{O}AH/(\alpha^2\varepsilon^2))$ 随 $1/\alpha^2, 1/\varepsilon^2,$ 和 (S, O, A) 以 乘法 方式呈多项式级增长。这是揭示型部分可观测马尔可夫决策过程的首个此类结果，相较于已有下界有显著改进（参见第 2.2节）。

推论 定理 4 表明，对 $(S, A, O, 1/\alpha, 1/\varepsilon)$ 的乘法依赖在当前最佳 PAC 上界 $\tilde{\mathcal{O}}(S^2OA(1+SA/O)/(\alpha^2\varepsilon^2))$（定理 3；忽略 H）中确有必要，并解决了关于揭示型部分可观测马尔可夫决策过程学习的若干开放问题：

- 这确定了对 α 的最优依赖关系为 $\Theta(\alpha^{-2})$（结合我们的下界与 $\mathcal{O}(\alpha^{-2})$ 上界），而此前关于 α 的最佳下界为 $\Omega(\alpha^{-1})$ (Liu 等人, 2022a)。
- 对于 $(\alpha$ 和 $\varepsilon)$ 的联合依赖关系，它表明至少需要 $1/(\alpha^2\varepsilon^2)$ 个样本。这排除了在一般情形下获得更优速率的可能性，例如单步揭示型部分可观测马尔可夫决策过程中针对 $\tilde{\mathcal{O}}(\max\{1/\alpha^2, 1/\varepsilon^2\})$ 的上界（转移为确定性的（Jin 等人, 2020a））。
- 这意味着样本复杂度中，其他参数（尤其是 $1/(\alpha^2\varepsilon^2)$）之外还必须乘上一个多项式因子 (O)，证实较大的观测空间确实会显著影响学习。

最后，与当前最佳 PAC 上界相比，下界 $\Omega(SO^{1/2}A/(\alpha^2\varepsilon^2))$ 涵盖了所有参数；在丰富观测区域 $O \geqslant \Omega(SA)$ 中，它与上界相差 $S\sqrt{O}$ 倍。这为未来的研究奠定了坚实基础。

3.2 多步揭示型部分可观测马尔可夫决策过程

通过类似的困难实例构造（更多细节见第5节），我们建立了多步情形下的 PAC 下界，其中 $m \geqslant 2$ （证明见附录G）。

[2]一个 $\tilde{\mathcal{O}}(\text{多项式}(S, O, H)A^H/\varepsilon^2)$ 对于 *any* 部分可观测马尔可夫决策过程（不一定是揭示型）（Even-Dar 等人, 2005）确实可以达到 PAC 上界；另请参阅 Uehara 等人 （2022b）。

定理 5（多步揭示型部分可观测马尔可夫决策过程的 PAC 下界）。对于任意 $m \geqslant 2$, $O \geqslant S \geqslant 10$, $A \geqslant 3$, $H \geqslant 8 \log_2 S + 2m$, $\alpha \in (0, 0.1]$, $\varepsilon \in (0, 0.01]$，存在一个由 m 步揭示型部分可观测马尔可夫决策过程组成的族 $\mathcal{M}$，满足$|\mathcal{S}| \leqslant S$, $|\mathcal{O}| \leqslant O$, $|\mathcal{A}| = A$，且对所有 $M \in \mathcal{M}$均有 $\alpha_m(M) \geqslant \alpha$。因此，任何与环境交互并返回一个 π^{out} 的算法 $\mathfrak{A}$，若其对所有 $M \in \mathcal{M}$以至少 $3/4$ 的概率满足 $V_M^\star - V_M(\pi^{\text{out}}) < \varepsilon$，则必须有

$$T \geqslant c_m \cdot \min\left\{\frac{(S^{1.5} \vee SA)O^{1/2}A^{m-1}H}{\alpha^2\varepsilon^2}, \frac{SA^{H/2}H}{\varepsilon^2}\right\},$$

其中，对于某个绝对常数 $c_0 > 0$，有 $c_m = c_0/m$。

多步情形（定理 5）的主要区别在于其更高的 A 依赖 $\Omega(A^{m-1})$，这表明上界（定理 3）中的 A^m 依赖在本质上已无法改进。此外，定理 $S^{1.5}$5 中的缩放高于定理 4，因此即使不考虑 A依赖，该结果在性质上也比单步情形更强。这是因为这里的困难实例实际上是对单步情形的强化，而非直接改编；这种强化利用了多步揭示的特性。相关讨论见第 5.3 节。

同样，与目前最佳的 PAC 上界 $S^2OA^m(1+SA/O)/(\alpha^2\varepsilon^2)$（定理 3）相比，定理 5中的下界与当前最佳上界之间存在 $\sqrt{SOA} \wedge S\sqrt{O}$ 差距。我们认为，在当前困难实例下，从下界方面无法改进这一差距中的 $\sqrt{SO}$ 因子；相关讨论见第 6.3节。

$\sqrt{O}$ **依赖性**单步和多步情形的下界均按 $\sqrt{O}$ 其 O-依赖性进行缩放。这种缩放源于规模为 $\mathcal{O}(O)$的均匀性检验任务的复杂度。该任务嵌入了揭示型部分可观测马尔可夫决策过程的困难实例中，其样本复杂度为 $\Theta(\sqrt{O}/\varepsilon^2)$（Paninski，2008 年；Diakonikolas 等人，2014 年；Canonne，2020 年）。困难实例的构造将在第5节中详细介绍。

4 多步情形的遗憾下界

现在我们转而建立遗憾下界。我们证明，令人惊讶的是，对于 m步揭示型部分可观测马尔可夫决策过程，只要 $m \geqslant 2$满足任意条件，非平凡的多项式遗憾（既不随 T 线性增长，也不随 H指数增长）至少为 $\Omega(T^{2/3})$。证明见附录F。

定理 6 ($\Omega(T^{2/3})$ 多步揭示型部分可观测马尔可夫决策过程的遗憾下界)。对于任意 $m \geqslant 2$、$O \geqslant S \geqslant 8$、$A \geqslant 3$、$H \geqslant 8\log_2 S + 2m$、$\alpha \in (0, 0.1]$、$T \geqslant 1$，存在一个由 m步揭示型部分可观测马尔可夫决策过程构成的族 $\mathcal{M}$，其满足$|\mathcal{S}| \leqslant S$、$|\mathcal{O}| \leqslant O$、$|\mathcal{A}| = A$，且对所有 $M \in \mathcal{M}$均有 $\alpha_m(M) \geqslant \alpha$。因此，对任意算法 $\mathfrak{A}$，都有

$$\max_{M \in \mathcal{M}} \mathbb{E}_M^{\mathfrak{A}}[\mathbf{Regret}] \geqslant c_m \cdot \min\left\{\left(\frac{SO^{1/2}A^mH}{\alpha^2}\right)^{1/3} T^{2/3}, \sqrt{SA^{H/2}HT}, T\right\},$$

其中 $c_m = c_0/m$ 对某个绝对常数 $c_0 > 0$成立。

目前，最好的次线性遗憾（关于其他问题参数为多项式）确实为 $T^{2/3}$，这是通过将 PAC 结果（Chen 等人，2022a）采用标准的先探索后利用式转换得到的。定理6排除了取得改进（例如达到 $\sqrt{T}$）的可能性，因为它表明 $T^{2/3}$ 其实是一个基本极限。

证明直觉定理6 中使用的困难实例与某个 PAC 困难实例相同（见第5节）。然而，定理6 依赖于一个关键的新观测，由此得出 $\Omega(T^{2/3})$ 遗憾下界。具体而言，对于多步揭示型部分可观测马尔可夫决策过程，我们可以设计一个困难实例，使得以下两类长度为 $m-1$的动作序列互不相交：

- 揭示型动作序列，其产生的观测能够揭示有关真实潜在状态的信息；
- 高奖励动作序列。

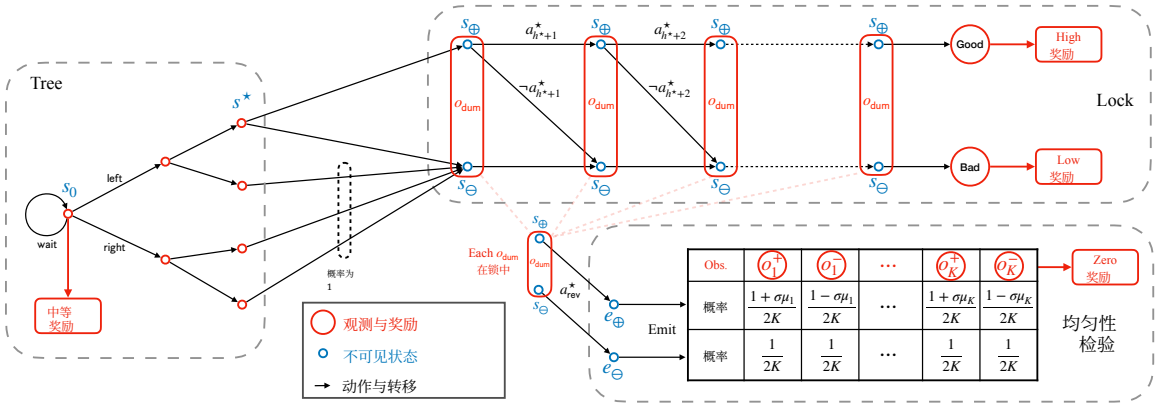

图1：两步揭示型POMDP困难实例简化版本的示意图。该实例由三个分量组成：树、锁和均匀性检验。在树中，所有转移都是确定且完全可观测的，智能体可以完全控制如何从 s_0 移动到叶节点。如果在树的任意叶节点采取任何动作，树都会随机转移到锁中；但存在一个唯一的（未知）状态 $s^\star$、步骤 $h^\star$和动作 $a^\star$，在此处智能体以正概率转移到 $s_\oplus$。在锁中，智能体无法观测潜在状态 $\{s_\oplus$、$s_\ominus\}$，因此必须输入正确的密码 $\mathbf{a}^\star$，才能留在 $s_\oplus$，最终获得高奖励。智能体还可以在任意 $a^\star_{\text{rev}}$ 采取揭示动作o_{dum}，转移到均匀性检验分量；在该分量中，智能体会收到一个观测，该观测会略微揭示先前的潜在状态是 $s_\oplus$ 还是 $s_\ominus$。更详细的说明请参见第5.1 节。

多步揭示条件（定义1）允许这样的构造。直观而言，这是因为它对 $\mathbb{M}_{h,m} \in \mathbb{R}^{\mathcal{O}^m \mathcal{A}^{m-1} \times \mathcal{S}}$ 存在广义左逆的要求相当宽松，而且可以通过精心设计某个子集动作序列上的发射-动作概率来实现。换言之，多步揭示条件允许只有部分动作序列具有揭示性，例如那些获得的奖励相当次优的动作序列。

这样的困难实例排除了高效实现探索-利用权衡的可能性，因为探索（采取揭示动作）和利用（采取高奖励动作）无法同时进行。因此，最好的做法就是采用先探索后利用类型的算法[3]，其遗憾通常为 $\Theta(T^{2/3})$ (Lattimoreand Szepesv´ari，2020 年)。

与单步情形的区别 定理 6 揭示了多步与单步设定之间的根本区别：已知单步揭示型部分可观测马尔可夫决策过程具有 $\tilde{\mathcal{O}}(\sqrt{T})$ 遗憾上界（Liu 等人，2022a）。直观而言，区别在于，在单步揭示型部分可观测马尔可夫决策过程中，智能体无需采取特定动作来获取潜在状态的信息，因此，信息获取（探索）与采取高奖励动作（利用）可以始终同时实现。

迈向 $\sqrt{T}$ 更强假设下的遗憾 自然会有人问，能否通过适当加强多步揭示条件（但仍弱于单步揭示条件）来规避 $\Omega(T^{2/3})$ 下界。基于上述直觉，一个可能的方向是进一步要求所有动作序列（长度为 $m-1$）都必须揭示有关潜在状态的信息。我们将此留作未来工作的一个问题。

5 证明概述

现在，我们概述困难实例构造和下界证明的技术细节。附录 F 中给出了一个简化版的多步揭示型困难实例，用于证明

[3]或者，也可以采用一种老虎机式算法，不执行揭示动作，而是通过暴力尝试直接识别最优策略，这对应于 $\sqrt{A^H T}$ 定理中的6。

PAC 和遗憾下界（定理 5 & 6）。为简单起见，我们以 2 步情形（$m = 2$）描述我们的构造；所得部分可观测马尔可夫决策过程的示意图见图 1。

5.1 困难实例的构造

与现有完全可观测设置下的下界相比，得到我们的下界的一项主要挑战，是刻画部分可观测性的难度，即其对 O 和 α^{-1}的依赖。

两步揭示组合锁 为体现这一难度，我们设计的基本分量是一个“两步揭示组合锁”（参见图 1中的“锁”部分），它是对Liu 等人 (2022a)；Jin 等人 (2020a)提出的非揭示组合锁的改造。该锁由两个隐藏状态 $s_\oplus, s_\ominus$ 以及一个未知的“正确”动作序列（即“密码”）$\mathbf{a}^\star_{h^\star+1:H}$组成。停留在 $s_\oplus$ 的唯一方法是在每个步骤 h都采取正确动作 $\mathbf{a}^\star_h$，而且只有步骤 H 的状态 $s_\oplus$ 会带来高奖励。因此，学习最优策略的任务等同于识别每个步骤的正确动作 $\mathbf{a}^\star_h$。我们使隐藏状态 $s_\oplus, s_\ominus$ 不可观测（发出虚拟观测 o_{dum}），因此智能体的一种朴素策略是从头猜测序列 $\mathbf{a}^\star$，这会导致 $\exp(\Omega(H))$ 的样本复杂度。

我们设计中的一个核心要素是：每个步骤都有一个唯一且已知的揭示动作 $a^\star_{\mathsf{rev}}$，并且它始终不同于正确动作。执行 $a^\star_{\mathsf{rev}}$ 会使系统从潜在状态 $s_\oplus$ 转移到 $e_\oplus$，随后从分布 $\mu_\oplus \in \Delta(\mathcal{O})$中发射一个观测；同样，系统也会从 $s_\ominus$ 转移到 $e_\ominus$，随后从分布$\mu_\ominus \in \Delta(\mathcal{O})$中发射一个观测。发射一次后，系统会确定性地转移到奖励为 0 的吸收终止状态。

均匀性检验 我们借鉴均匀性检验领域的技术（Canonne，2020，2022），选取 $\{\mu_\oplus、\ \mu_\ominus\}$，使它们尽可能难以区分，同时确保该部分可观测马尔可夫决策过程仍满足α-揭示条件。具体而言，令 $\mu_\ominus$ =Unif$(\mathcal{O})$ 为 $\mathcal{O}$[4]上的均匀分布。已知，要检验一个分布是否为均匀分布，与一个接近均匀分布且满足 $D_{\mathrm{TV}}(\mu_\oplus、\ \mu_\ominus) \asymp \sigma$ 的 $\mu_\oplus$ 相区分，需要 $\Theta(\sqrt{O}/\sigma^2)$ 个样本（Paninski，2008 年）。此外，针对 $\mu_\oplus$ 的最坏情况先验具有如下形式：$\mu_\oplus =$ Unif $(\mathcal{O}) + \sigma\mu/O,$，其中 $\mu \sim$ Unif$(\{(+1、\ -1)、\ (-1、\ +1)\}^{O/2})$。我们在困难实例中采用了 $\mu_\ominus$ 和 $\mu_\oplus$ 这样的选取（参见图1中的“均匀性检验”部分），这也能确保该部分可观测马尔可夫决策过程满足 $\Theta(\sigma^{-1})$-揭示条件。

树形马尔可夫决策过程；奖励 为了进一步体现下界中的 HSA 倍数因子，我们在组合锁之前进一步嵌入了一个完全可观测的 树形马尔可夫决策过程 （Domingues等人，2021）。该树是一棵具有 S 个叶节点的平衡二叉树，并具有确定性转移（因此，到达哪个叶节点完全由动作序列决定）和完全可观测性。树的所有叶节点都会转移到组合锁（即 $\{s_\oplus、\ s_\ominus\}$ 中的一个）。然而，存在唯一的 $(h^\star、\ s^\star、\ \mathbf{a}^\star)$，使得只有在 $a_{h^\star} = a^\star$ 处采取$s_{h^\star} = s^\star$ 并在第 $h^\star$ 步时，才有 ε 的概率转移到 $s_\oplus$；在叶节点处的其他所有选择都以概率 1 转移到 $s_\ominus$ （参见图 1中的“树”部分）。

我们进一步设计奖励函数，使得智能体必须正确识别底层参数 $(h^\star、\ s^\star、\ a^\star)$，才能学习到 $\Theta(\varepsilon)$ 近最优策略。

5.2 下界的计算

根据我们的构造，要在这个困难实例中学习一个 ε 近最优策略，智能体必须识别$(h^\star,\ s^\star,\ a^\star)$；而这只能通过尝试所有“入口” $(s,\ a,\ h)$ 并在两者之间进行测试来实现

$$H_0 : \mathbb{P}(s_{h+1} = s_\oplus | s_h = s, a_h = a) = 0,$$
$$H_1 : \mathbb{P}(s_{h+1} = s_\oplus | s_h = s, a_h = a) = \varepsilon.$$

对于每个入口。正如我们所展示的，要实现这一点，智能体必须（1）从头开始猜测密码$\mathbf{a}^\star$ （使用 $\Omega\left(A^{H-h}/\varepsilon^2\right)$ 个样本），或者（2）取 $a^\star_{\mathsf{rev}}$ 并使用这些

[4]具体来说，我们选取 $\mu_\oplus$，$\mu_\ominus$ 作为 $subset$ 上的均匀性检验困难实例，其中 $subset$ 是 $\mathcal{O}$ 的子集，规模为 $2K = \Theta(O)$。为便于表述，这里使用全集 $\mathcal{O}$。

观测。后一个任务最终等价于在以下两者之间进行检验

$$H_0' = \mu_\ominus, \quad H_1' = \varepsilon\mu_\oplus + (1-\varepsilon)\mu_\ominus,$$

其中 $\mu_\ominus$ 是定义在 $2K = \Theta(O)$ 个元素上的均匀分布，而 $\mu_\oplus$ 则从均匀性检验的最坏情况先验中抽取。在 H_0' 和 H_1' 之间进行区分，是一个参数为 $\sigma\varepsilon$的均匀性检验任务，需要 $n \geqslant \Omega(\sqrt{O}/(\varepsilon\sigma)^2)$ 个样本（Paninski，2008 年）。

通过仔细的信息论论证，上述论证将得到一个 PAC 下界

$$\Theta(SAH) \times \Omega\Big(\min\Big\{\frac{\sqrt{O}}{\sigma^2\varepsilon^2}, \frac{A^{\Theta(H)}}{\varepsilon^2}\Big\}\Big),$$

用于学习 2 步 $\Theta(\sigma^{-1})$-揭示型部分可观测马尔可夫决策过程。该速率与我们实际的 PAC 下界（定理 5）相近（尽管略差于后者）。同一个困难实例还进一步给出了一个 $\Omega(T^{2/3})$ 遗憾下界（尽管速率略差于定理 6）；详见附录 F.8中的计算。

我们指出，上述计算是启发式的；要使这些论证严格化，需要借助信息论论证——在我们的情形下，即 Ingster 方法（Ingster 和 Suslina，2012 年）（参见附录D &引理 E.5作为示例）——来界定任意算法在不同困难实例上诱导的分布之间的散度。

5.3 关于实际构造的说明

上述两步困难实例是对定理5& 6的证明中所用实际实例的简化，在若干方面有所不同。实际构造稍显复杂，还包含以下额外要素：

- 对于 m 步情形，为了得到随 A^m 缩放的下界，我们修改上述构造，使智能体每 $(m-1)$ 步只能执行一次 $a^{\star\text{揭示}}$，并将 $a^{\star\text{揭示}}$替换为一组 $|\mathcal{A}_{\text{揭示}}| = \Theta(A)$ 动作，它们共同带来一个 $A^{m-1} \times A = A^m$ 因子。

- 我们还通过将单个组合锁替换为 $\Theta(S)$个并行锁，获得它们共用同一密码，但发射概率各不相同。我们证明，此设置下的学习至少与对 $\Theta(SO)$ 个元素进行均匀性检验一样困难，因此多出一个 $\sqrt{S}$因子。 _

6 讨论

6.1 单步情形的遗憾

如前所述，单步揭示型部分可观测马尔可夫决策过程不可能存在像多步情形那样的 $\Omega(T^{2/3})$ 遗憾下界，因为可以达到 $\tilde{\mathcal{O}}(\sqrt{T})$ 上界。尽管如此，通过从 PAC 下界（定理 $(\sqrt{T})$4）直接归约，并使用马尔可夫不等式和标准的在线到批处理转换，我们得到了匹配的 Ω遗憾下界，具体如下。

推论 7 (单步揭示型部分可观测马尔可夫决策过程的遗憾下界**)。** 在与定理 *4* 相同的设定下，同一族 $\mathcal{M}$ 单步 α 揭示型部分可观测马尔可夫决策过程满足：对于任意算法 $\mathfrak{A}$，

$$\max_{M\in\mathcal{M}} \mathbb{E}_M^{\mathfrak{A}}[\mathbf{Regret}] \geqslant c_0 \cdot \min\Big\{\sqrt{\frac{SO^{1/2}AH}{\alpha^2}T}, \sqrt{SA^{H/2}HT}, T\Big\}, \tag{5}$$

其中 $c_0 > 0$ 是一个绝对常数。

作为对比，当前单步揭示型部分可观测马尔可夫决策过程的最佳遗憾上界为 $\tilde{\mathcal{O}}(\sqrt{S^3O^3A^2(1+SA/O)\alpha^{-4}} \cdot T \times$ 多项式$(H))$ (Liu 等人，2022b)[5]，，它至少比 $\sqrt{S^2O^{2.5}A\alpha^{-2}}$ 大一个 (5) 中的主项。这里我们给出一个精细得多的遗憾上界，将这一差距缩小到 $\sqrt{SO^{1.5}}$，并且重要的是，确定了对 α 的依赖关系。

[5]由他们的结果转换而来，其中揭示常数定义在 $(2 \to 2)$-范数下。

定理 8（单步揭示型部分可观测马尔可夫决策过程的遗憾上界）。 存在算法（OMLE、E2D-TA 和 MOPS），能够与任意单步 α-揭示型部分可观测马尔可夫决策过程 M 交互，并实现遗憾

$$\mathbf{Regret} \leqslant \widetilde{\mathcal{O}}\left(\sqrt{\frac{S^2O^2A(1+SA/O)H^3}{\alpha^2} \cdot T}\right) \tag{6}$$

以高概率。

我们在更广泛的一类序贯决策问题上证明了定理 8 ，这类问题称为强 B 稳定预测状态表示 (参见附录 H.1），其中单步揭示型部分可观测马尔可夫决策过程是其特例。证明与针对 B 稳定预测状态表示的 PAC 学习分析大体平行（Chen 等人, 2022a），详见附录 H。

6.2 DEC 方法的启示

决策估计系数（DEC）（Foster 等人，2021）为建立任意一般强化学习问题的样本复杂度下界提供了另一种潜在方法。然而，我们在此证明，对于揭示型部分可观测马尔可夫决策过程，DEC 给出的任何下界必然严格弱于我们的下界。

例如，对于 PAC 学习，已知 m 步揭示型部分可观测马尔可夫决策过程的探索型决策估计系数（EDEC）存在一个上界 $\mathrm{edec}_\gamma \leqslant \widetilde{\mathcal{O}}(SA^mH^2\alpha^{-2}/\gamma)$ (Chen 等人 (2022a）；另见命题 C.2），因此，通过对 EDEC 取下界所得到的任何 PAC 下界至多为 $\Omega(SA^mH^2\alpha^{-2}/\varepsilon^2)$ (Chen 等人, 2022b）。这样的下界必然至少比我们的定理 5 小$\sqrt{O}(1 \vee \sqrt{S/A})$, 倍，而且重要的是，它不会随 O 多项式级增长。

我们的下界还对决策估计系数理论有其他有趣的启示：尽管 E2D 等算法能够针对模型类(Foster 等人，2021；Chen 等人，2022b）给出以 DEC 和对数覆盖数表示的样本复杂度上界，但若不作进一步假设，就不能用价值类或策略类的对数覆盖数取而代之，从而对Foster 等人(2021)中留下的相关问题给出了否定答案（参见附录I.1，其中有详细讨论）。

6.3 力求弥合差距

最后，作为一个重要的开放问题，我们的下界与当前最佳上界之间仍存在较小差距，尤其体现在 $(S,\ O)$ 依赖关系上。例如，对于多步揭示型部分可观测马尔可夫决策过程，PAC 下界 $\Omega(S^{1.5}\sqrt{O}A^{m-1}/(\alpha^2\varepsilon^2))$ (定理 5)与上界（定理 $\sqrt{SOA}$ 3）之间仍有差距。尽管我们认为 A 这一因子是分析造成的，可予以消除，但剩余的$\sqrt{SO}$ 这一因子无法在沿用当前这类困难实例的情况下纳入下界——存在一种专门针对这类实例设计的算法，可通过对树进行暴力枚举，并在组合锁中进行均匀性检验，达到 $\widetilde{\mathcal{O}}(S^{1.5}\sqrt{O}A^m/(\alpha^2\varepsilon^2))$上界（附录I.2）。

弥合这$\sqrt{SO}$ 一差距，可能需要借助其他困难实例来证明更强的下界——例如，将其他问题嵌入分布检验中（Canonne，2020 年）——或者给出更精确的上界；我们将这些留作未来工作。

7 结论

本文为揭示型部分可观测马尔可夫决策过程这一重要且易于处理的类别中的部分可观测强化学习建立了样本复杂度下界。我们的下界首次以乘法形式呈现状态数、动作数、观测数和揭示常数的多项式级增长，并表明这些下界与当前最佳上界之间的差距相当温和。我们的工作为未来的细粒度研究奠定了坚实基础，并提出了许多有趣的问题，例如从任一方向弥合这些差距，或有意义地加强多步揭示假设，以允许 $\sqrt{T}$ 遗憾。

参考文献

Alekh Agarwal、Sham Kakade、Akshay Krishnamurthy 和 Wen Sun。Flambe：低秩马尔可夫决策过程的结构复杂度与表征学习。神经信息处理系统进展，33：20095–20107，2020 年。Mohammad Gheshlaghi Azar、Ian Osband 和 R´emi Munos。强化学习的极小极大遗憾界。载于国际机器学习会议，第 263–272 页。PMLR，2017 年。Qi Cai、Zhuoran Yang 和 Zhaoran Wang。部分观测下的强化学习：具有可证明样本效率的线性函数近似。载于国际机器学习会议，第 2485–2522 页。PMLR，2022。Cl´ement L Canonne。分布检验综述：你的数据量很大，但它是蓝色的吗？计算理论，第 1–100 页，2020 年。Cl´ement L Canonne。分布检验的主题与技术。2022。Fan Chen、Yu Bai 和 Song Mei。具有 B-稳定性的部分可观测强化学习：统一的结构条件与高效样本算法。*arXiv* 预印本 *arXiv:2209.14990*，2022a。Fan Chen、Song Mei 和 Yu Bai。基于决策估计系数的强化学习统一算法：无遗憾、PAC 与无奖励学习。*arXiv* 预印本 *arXiv:2209.11745*，2022b。Christoph Dann 和 Emma Brunskill。回合式固定时域强化学习的样本复杂度。神经信息处理系统进展，28，2015 年。Ilias Diakonikolas、Daniel M Kane 和 Vladimir Nikishkin。结构化分布的恒等性检验。载于第二十六届 *ACM-SIAM*离散算法研讨会论文集，第 1841–1854 页。SIAM，2014 年。Omar Darwiche Domingues、Pierre M´enard、Emilie Kaufmann 和 Michal Valko。有限马尔可夫决策过程中的回合式强化学习：重新审视极小极大下界。载于算法学习理论，第 578–598 页。PMLR，2021 年。Simon Du、Sham Kakade、Jason Lee、Shachar Lovett、Gaurav Mahajan、Wen Sun 和 Ruosong Wang。双线性类：强化学习中可证明泛化的结构框架。载于国际机器学习会议，第 2826–2836 页。PMLR，2021 年。Yonathan Efroni、Chi Jin、Akshay Krishnamurthy 和 Sobhan Miryoosefi。利用短期记忆实现可证明的强化学习。*arXiv* 预印本 *arXiv:2202.03983*，2022。Eyal Even-Dar、Sham M Kakade 和 Yishay Mansour。无需重置的部分可观测马尔可夫决策过程强化学习。2005 年。Dylan J Foster、Sham M Kakade、Jian Qian 和 Alexander Rakhlin。交互式决策的统计复杂度。*arXiv* 预印本 *arXiv:2112.13487*，2021 年。Dylan J Foster、Alexander Rakhlin、Ayush Sekhari 和 Karthik Sridharan。论对抗式决策的复杂度。*arXiv* 预印本 *arXiv:2206.13063*，2022。Noah Golowich、Ankur Moitra 和 Dhruv Rohatgi。无需计算上难以处理的预言机，在可观测部分可观测马尔可夫决策过程中进行学习。*arXiv* 预印本 *arXiv:2206.03446*，2022a。Noah Golowich、Ankur Moitra 和 Dhruv Rohatgi。以准多项式时间规划可观测部分可观测马尔可夫决策过程。*arXiv* 预印本 *arXiv:2201.04735*，2022b。Yuri Ingster 和 IA Suslina。高斯模型下的非参数拟合优度检验，第 169 卷。Springer Science & Business Media，2012 年。

Thomas Jaksch、Ronald Ortner 和 Peter Auer。《强化学习的近最优遗憾界》。Journal of Machine Learning Research，11(51):1563–1600，2010 年。网址：http://jmlr.org/papers/v11/ jaksch10a.html。Nan Jiang、Akshay Krishnamurthy、Alekh Agarwal、John Langford 和 Robert E Schapire。《具有低贝尔曼秩的上下文决策过程可进行 PAC 学习》。载于国际机器学习会议，页码 1704–1713。PMLR，2017 年。Chi Jin、Zeyuan Allen-Zhu、Sebastien Bubeck 和 Michael I Jordan。《Q-learning 是否可被证明是高效的？》。载于神经信息处理系统进展，第 31 卷，2018 年。Chi Jin、Sham Kakade、Akshay Krishnamurthy 和 Qinghua Liu。《欠完备部分可观测马尔可夫决策过程的样本高效强化学习》。载于神经信息处理系统进展，第 33 卷，页码 18530–18539，2020a。Chi Jin、Zhuoran Yang、Zhaoran Wang 和 Michael I Jordan。《基于线性函数近似的可证明高效强化学习》。载于学习理论会议，页码 2137–2143。PMLR，2020b。Chi Jin、Qinghua Liu 和 Sobhan Miryoosefi。《贝尔曼 eluder 维数：强化学习问题的新型丰富类别及样本高效算法》。载于神经信息处理系统进展，第 34 卷，页码 13406–13418，2021 年。Leslie Pack Kaelbling、Michael L Littman 和 Anthony R Cassandra。《部分可观测随机领域中的规划与行动》。Artificial intelligence，101(1-2):99–134，1998 年。Michael Kearns 和 Satinder Singh。《多项式时间内的近最优强化学习》。Machine learning，49(2):209–232，2002 年。Michael Kearns、Yishay Mansour 和 Andrew Ng。《利用可重用轨迹对大型部分可观测马尔可夫决策过程进行近似规划》。载于神经信息处理系统进展，第 12 卷，1999 年。Akshay Krishnamurthy、Alekh Agarwal 和 John Langford。《基于丰富观测的 PAC 强化学习》。载于神经信息处理系统进展，第 29 卷，2016 年。Jeongyeol Kwon、Yonathan Efroni、Constantine Caramanis 和 Shie Mannor。《潜在马尔可夫决策过程的强化学习：遗憾保证与下界》。载于神经信息处理系统进展，第 34 卷，页码 24523–24534，2021 年。Tor Lattimore 和 Csaba Szepesv´ari。《赌博机算法》。剑桥大学出版社，2020 年。Qinghua Liu、Alan Chung、Csaba Szepesv´ari 和 Chi Jin。《部分可观测强化学习何时不再令人望而生畏？》。arXiv 预印本 arXiv:2204.08967，2022a。Qinghua Liu、Praneeth Netrapalli、Csaba Szepesvari 和 Chi Jin。《乐观最大似然估计——一种用于部分可观测序贯决策的通用基于模型算法》。arXiv 预印本 arXiv:2209.14997，2022b。Qinghua Liu、Csaba Szepesv´ari 和 Chi Jin。《部分可观测马尔可夫博弈的样本高效强化学习》。arXiv 预印本 arXiv:2206.01315，2022c。Elchanan Mossel 和 S´ebastien Roch。《学习非奇异系统发育树和隐马尔可夫模型》。载于第三十七届 ACM 计算理论年度研讨会论文集，页码 366–375，2005 年。Liam Paninski。《基于巧合的检验：在离散数据采样极其稀疏时检验均匀性》。IEEE Transactions on Information Theory，54(10):4750–4755，2008 年。Christos H Papadimitriou 和 John N Tsitsiklis。《马尔可夫决策过程的复杂性》。Mathematics of operations research，12(3):441–450，1987 年。Igal Sason 和 Sergio Verd´u。《f-散度不等式》。IEEE Transactions on Information Theory，62(11): 5973–6006，2016。Richard S Sutton 和 Andrew G Barto。《强化学习：导论》。MIT 出版社，2018 年。

Masatoshi Uehara、Ayush Sekhari、Jason D Lee、Nathan Kallus 和 Wen Sun。《具有潜在确定性和条件嵌入的部分可观测马尔可夫决策过程中计算高效的 PAC 强化学习》。*arXiv* 预印本 *arXiv:2206.12081*，2022a。

Masatoshi Uehara、Ayush Sekhari、Jason D Lee、Nathan Kallus 和 Wen Sun。部分可观测动力系统中的可证明高效强化学习。*arXiv* 预印本 *arXiv:2206.12020*，2022b。

Sara A Van de Geer。*M*估计中的经验过程，第 6 卷。剑桥大学出版社，2000 年。

Lingxiao Wang、Qi Cai、Zhuoran Yang 和 Zhaoran Wang。嵌入以控制部分观测系统：具有可证明样本效率的表征学习。*arXiv* 预印本 *arXiv:2205.13476*，2022。

Tengyang Xie、Dylan J Foster、Yu Bai、Nan Jiang 和 Sham M Kakade。覆盖在在线强化学习中的作用。*arXiv* 预印本 *arXiv:2210.04157*，2022。

Wenhao Zhan、Masatoshi Uehara、Wen Sun 和 Jason D Lee。面向预测状态表征的 PAC 强化学习。*arXiv* 预印本 *arXiv:2207.05738*，2022。

Zihan Zhang、Yuan Zhou 和 Xiangyang Ji。基于参考优势分解的近最优无模型强化学习。神经信息处理系统进展，33：15198–15207，2020 年。

Han Zhong、Wei Xiong、Sirui Zheng、Liwei Wang、Zhaoran Wang、Zhuoran Yang 和 Tong Zhang。用于交互式决策的后验采样框架。*arXiv* 预印本 *arXiv:2211.01962*，2022。

A 技术工具

引理 A.1。 对于正实数 A、 B、 T、 $\varepsilon_0 > 0$，有

$$\sup_{\varepsilon\in(0,\varepsilon_0]}\left(\varepsilon T \wedge \frac{A}{\varepsilon^2} \wedge \frac{B}{\varepsilon}\right) \geqslant A^{1/3}T^{2/3} \wedge \sqrt{BT} \wedge \varepsilon_0 T.$$

引理 *A.1*的证明。 假设 $R > 0$ 满足 $R \geqslant \varepsilon T \wedge \frac{A}{\varepsilon^2} \wedge \frac{B}{\varepsilon}$ 对所有 $\varepsilon \in (0, \varepsilon_0]$。那么，对于每个$\varepsilon \in (0, \varepsilon_0]$，要么 $\varepsilon \leqslant \frac{R}{T}$，要么 $\varepsilon \geqslant \sqrt{\frac{A}{R}}$，要么 $\varepsilon \geqslant \frac{B}{R}$。因此，

$$(0, \varepsilon_0] \subseteq (0, \frac{R}{T}] \cup [\sqrt{\frac{A}{R}}, +\infty) \cup [\frac{B}{R}, +\infty).$$

因此， $\frac{R}{T} \geqslant \varepsilon_0$、 $\sqrt{\frac{A}{R}} \leqslant \frac{R}{T}$ 或 $\frac{B}{R} \leqslant \frac{R}{T}$。综合这三种情况，我们得到

$$R \geqslant \varepsilon_0 T \wedge A^{1/3}T^{2/3} \wedge \sqrt{BT}.$$

□

引理 A.2. 假设 $(R_t)_{t\geqslant 1}$ 是适应于滤过 $(\mathcal{F}_t)_{t\geqslant 1}$的一列正随机变量，且T 是一个停止时间（即对于 $t \geqslant 1$， R_t 是 $\mathcal{F}_t$-可测的，且事件 $\{\mathsf{T} \leqslant t\} \in \mathcal{F}_t$）。那么有

$$\mathbb{E}\left[\prod_{t=1}^{\mathsf{T}} R_t \times \prod_{t=1}^{\mathsf{T}} \mathbb{E}[R_t|\mathcal{F}_{t-1}]^{-1}\right] = 1.$$

等价地，

$$\mathbb{E}\left[\prod_{t=1}^{\mathsf{T}} R_t \times \exp\left(-\sum_{t=1}^{\mathsf{T}} \log \mathbb{E}[R_t|\mathcal{F}_{t-1}]\right)\right] = 1.$$

引理A.2 可通过反复应用塔性质立即得出 s.

引理 A.3. 假设随机变量 X 是 σ-次高斯的，即 $\mathbb{E}[\exp(tX)] \leqslant \exp\left(\frac{\sigma^2 t^2}{2}\right)$ 对任意$t \in \mathbb{R}$均成立。则对所有 $t \geqslant 0$，我们有

$$\mathbb{E}[\exp(t\,|X|)] \leqslant \exp\left(\max\left\{\sigma^2 t^2, \frac{4}{3}\sigma t\right\}\right).$$

引理 *A.3* 的证明。 对于任意 $x \geqslant 1$，我们有

$$\mathbb{E}[\exp(t\,|X|)] \leqslant \mathbb{E}[\exp(xt\,|X|)]^{\frac{1}{x}} \leqslant (\mathbb{E}[\exp(xtX)] + \mathbb{E}[\exp(-xtX)])^{\frac{1}{x}} \leqslant 2^{\frac{1}{x}} \exp\left(\frac{\sigma^2 t^2 x}{2}\right) = \exp\left(\frac{\sigma^2 t^2 x}{2} + \frac{\log 2}{x}\right).$$

我们分两种情况讨论：1. 若 $\sigma t \geqslant \sqrt{2\log 2}$，则在上述不等式中取 $x = 1$，可得$\mathbb{E}[\exp(t\,|X|)] \leqslant \exp(\sigma^2 t^2)$。2. 若 $\sigma t < \sqrt{2\log 2}$，则在上述不等式中取 $x = \sqrt{2\log_{\sigma t}{}^2} > 1$，可得 $\mathbb{E}[\exp(t\,|X|)] \leqslant \exp(\sqrt{2\log 2}\,\sigma t) \leqslant \exp(\frac{4}{3}\sigma t)$。综合这两种情况，即完成证明。 □

对于可测空间 $(\mathcal{X}, \mathcal{F})$ 上基准测度为 μ 的概率分布 $\mathbb{P}$ 和 $\mathbb{Q}$，我们定义它们之间的全变差距离和海灵格距离如下： $\mathbb{P}$, $\mathbb{Q}$

$$D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)| = \frac{1}{2}\int_{\mathcal{X}} \left|\frac{d\mathbb{P}}{d\mu}(x) - \frac{d\mathbb{Q}}{d\mu}(x)\right| d\mu(x),$$
$$D_{\mathrm{H}}^2(\mathbb{P}, \mathbb{Q}) = \int_{\mathcal{X}} \left(\sqrt{\frac{d\mathbb{P}}{d\mu}} - \sqrt{\frac{d\mathbb{Q}}{d\mu}}\right)^2 d\mu.$$

当 $\mathbb{P} \ll \mathbb{Q}$ 时，我们还可以定义 $\mathbb{P}$, $\mathbb{Q}$ 之间的 KL散度和 χ^2-散度如下：

$$\mathrm{KL}(\mathbb{P} \,\|\, \mathbb{Q}) = \mathbb{E}_{\mathbb{P}}\left[\log \frac{d\mathbb{P}}{d\mathbb{Q}}\right], \qquad \chi^2(\mathbb{P} \,\|\, \mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}\left[\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)^2\right] - 1.$$

引理 A.4。 设 $\mathbb{P}$、$\mathbb{Q}$、$\mathbb{P}'$、$\mathbb{Q}'$ 是定义在 $(\mathcal{X}$ 上的四个概率测度，$\mathcal{F})$，且 Ω 是一个满足$\mathbb{P}|_\Omega = \mathbb{P}'|_\Omega$、$\mathbb{Q}|_\Omega = \mathbb{Q}'|_\Omega$的事件。那么有

$$D_{\mathrm{TV}}(\mathbb{P}', \mathbb{Q}') \geqslant D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) - \mathbb{P}(\Omega^c).$$

引理 *A.4*的证明。 设 μ是定义在 $(\mathcal{X}$上的基准测度，$\mathcal{F})$ 满足 $\mathbb{P}$, $\mathbb{P}'$, $\mathbb{Q}$, $\mathbb{Q}'$ 相对于 μ 具有密度（例如，$\mu = (\mathbb{P} + \mathbb{P}' + \mathbb{Q} + \mathbb{Q}')/4$）。为简化记号，我们用 $\mathbb{P}(x)$ 表示 $d\mathbb{P}(x)/d\mu(x)$，并用 dx 表示 $\mu(dx)$。于是有

$$\begin{aligned}
2D_{\mathrm{TV}}(\mathbb{P}', \mathbb{Q}') &= \int_{\mathcal{X}} |\mathbb{P}'(x) - \mathbb{Q}'(x)|\, dx = \int_{\Omega} |\mathbb{P}'(x) - \mathbb{Q}'(x)|\, dx + \int_{\Omega^c} |\mathbb{P}'(x) - \mathbb{Q}'(x)|\, dx \\
&\geqslant \int_{\Omega} |\mathbb{P}'(x) - \mathbb{Q}'(x)|\, dx + |\mathbb{P}'(\Omega^c) - \mathbb{Q}'(\Omega^c)| \\
&= \int_{\Omega} |\mathbb{P}(x) - \mathbb{Q}(x)|\, dx + |\mathbb{P}(\Omega^c) - \mathbb{Q}(\Omega^c)| \\
&\geqslant \int_{\Omega} |\mathbb{P}(x) - \mathbb{Q}(x)|\, dx + \mathbb{P}(\Omega^c) + \mathbb{Q}(\Omega^c) - 2\mathbb{P}(\Omega^c) \\
&\geqslant \int_{\Omega} |\mathbb{P}(x) - \mathbb{Q}(x)|\, dx + \int_{\Omega^c} |\mathbb{P}(x) - \mathbb{Q}(x)|\, dx - 2\mathbb{P}(\Omega^c) \\
&= 2D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q}) - 2\mathbb{P}(\Omega^c).
\end{aligned}$$

证明至此完成。 □

引理 A.5（散度不等式，参见例如 Sason 和 Verd´u （2016））。 对于两个概率测度 $\mathbb{P}$， $\mathbb{Q}$在（$\mathcal{X}$， $\mathcal{F}$)上，有

$$2D_{\mathrm{TV}}\left(\mathbb{P},\mathbb{Q}\right)^2 \leqslant \mathrm{KL}(\mathbb{P}\parallel\mathbb{Q}) \leqslant \log\left(1+\chi^2(\mathbb{P}\parallel\mathbb{Q})\right).$$

引理 A.6 (海灵格条件化引理，参见例如 Chen 等人 (2022a，引理 A.1))。 对于任意一对随机变量 $(X,\ Y)$，有

$$\mathbb{E}_{X\sim\mathbb{P}_X}\left[D_{\mathrm{H}}^2\left(\mathbb{P}_{Y|X},\mathbb{Q}_{Y|X}\right)\right] \leqslant 2D_{\mathrm{H}}^2\left(\mathbb{P}_{X,Y},\mathbb{Q}_{X,Y}\right).$$

B 预测状态表征基础与 B-稳定性

以下关于预测状态表征（PSR）和 B-稳定性条件的记号摘自（Chen 等人，2022a）。

具有观测的序贯决策过程 一个分回合序贯决策过程由元组 $\{H,\ \mathcal{O},\ \mathcal{A},\ \mathbb{P},\ \{r_h\}_{h\in[H]}\}$,定义，其中 $H\in\mathbb{Z}_{\geqslant 1}$是时域长度； $\mathcal{O}$是观测空间； $\mathcal{A}$是动作空间； $\mathbb{P}$ 指定转移动态，使得初始观测服从 $o_1\sim\mathbb{P}_0(\cdot)\in\Delta(\mathcal{O})$，并且给定直至步骤 h的历史 $\tau_h:=(o_1,\ a_1,\ \cdots,\ o_h,\ a_h)$时，观测服从 $o_{h+1}\sim\mathbb{P}(\cdot|\tau_h)$； $r_h:\ \mathcal{O}\times\mathcal{A}\to[0,\ 1]$是第 h 步的奖励函数，我们假设它是 $(o_h,\ a_h)$的已知确定性函数。

在分回合序贯决策过程中，策略 $\pi=\{\pi_h:\ (\mathcal{O}\times\mathcal{A})^{h-1}\times\mathcal{O}\to\Delta(\mathcal{A})\}_{h\in[H]}$ 是一组 H 函数。在步骤 $h\in[H]$, 运行策略 π 的智能体会观察到观测 o_h ，并根据历史 $(\tau_{h-1}、\ o_h)=(o_1、\ a_1、\ \ldots\ldots、\ o_{h-1}、\ a_{h-1}、\ o_h)$采取动作$a_h\sim\pi_h(\cdot|\tau_{h-1},o_h)\in\Delta(\mathcal{A})$。随后，智能体会收到奖励 $r_h(o_h,a_h)$，环境则根据 $\tau_h=(o_1、\ a_1、\ \cdots、\ o_h、\ a_h)$（如果 $h<H$）生成下一个观测 $o_{h+1}\sim\mathbb{P}(\cdot|\tau_h)$ 。采取 a_H 后，回合立即终止。

对于任意 $\tau_h=(o_1,a_1,\cdots,o_h,a_h)$，我们记

$$\begin{aligned}\mathbb{P}(\tau_h)&:=\mathbb{P}(o_{1:h}|a_{1:h})=\prod_{h'\leqslant h}\mathbb{P}(o_{h'}|\tau_{h'-1}),\\ \pi(\tau_h)&:=\prod_{h'\leqslant h}\pi_{h'}(a_{h'}|\tau_{h'-1},o_{h'}),\\ \mathbb{P}^\pi(\tau_h)&:=\mathbb{P}(\tau_h)\times\pi(\tau_h).\end{aligned}$$

那么， $\mathbb{P}^\pi(\tau_h)$ 是在执行 π时观测到 τ_h （前 h 个步骤）的概率。

预测状态表示、核心测试集和预测状态A 测试 t 是未来观测和动作的序列（即 $t\in\mathfrak{T}:=\bigcup_{W\in\mathbb{Z}_{\geqslant 1}}\mathcal{O}^W\times\mathcal{A}^{W-1}$）。对于某个长度为 $W\geqslant 1$的测试 $t_h=(o_{h:h+W-1},\ a_{h:h+W-2})$，我们将给定（可达）历史 τ_{h-1}时测试 t_h成功的概率定义为 $\mathbb{P}(t_h|\tau_{h-1}):=\mathbb{P}(o_{h:h+W-1}|\tau_{h-1}$；其中$(a_{h:h+W-2}))$，即在给定历史 τ_{h-1}的条件下，如果智能体确定性地执行动作 $a_{h:h+W-2}$，则观测到 $o_{h:h+W-1}$的概率。我们遵循以下约定：如果对于任意 π都有 $\mathbb{P}^\pi(\tau_{h-1})=0$ ，则 $\mathbb{P}(t|\tau_{h-1})=0$。

定义 B.1（预测状态表示、核心测试集和预测状态）。 对于任意 $h\in[H]$，我们称集合 $\mathcal{U}_h\subset\mathfrak{T}$ 为步骤 h 处的核心测试集，如果满足以下条件：对于任意 $W\in\mathbb{Z}_{\geqslant 1}$、任意可能的未来（即测试） $t_h=(o_{h:h+W-1}$ $a_{h:h+W-2})\in\mathcal{O}^W\times\mathcal{A}^{W-1}$，都存在一个向量 $b_{t_h,\ h}\in\mathbb{R}^{\mathcal{U}_h}$，使得

$$\mathbb{P}(t_h|\tau_{h-1})=\langle b_{t_h,h},[\mathbb{P}(t|\tau_{h-1})]_{t\in\mathcal{U}_h}\rangle,\qquad\forall\tau_{h-1}\in\mathcal{T}^{h-1}:=(\mathcal{O}\times\mathcal{A})^{h-1}. \tag{7}$$

我们将向量 $\mathbf{q}(\tau_{h-1}):=[\mathbb{P}(t|\tau_{h-1})]_{t\in\mathcal{U}_h}$ 称为步骤 h 时的预测状态（约定：若 τ_{h-1} 不可达，则$\mathbf{q}(\tau_{h-1})=0$ ），并将$\mathbf{q}_0:=[\mathbb{P}(t)]_{t\in\mathcal{U}_1}$ 称为初始预测状态。线性 *PSR* 是一种配备核心测试集 $\{\mathcal{U}_h\}_{h\in[H]}$的序贯决策过程。

定义 $\mathcal{U}_{A,h} := \{$**动作**：$(\mathbf{o},\mathbf{a}) \in \mathcal{U}_h$，其中某个$\mathbf{o} \in \bigcup_{W\in\mathbb{N}^+} \mathcal{O}^W\}$属于 $\mathcal{U}_h$，作为其中的"核心动作"集合（可能包含空序列），且 $U_A := \max_{h\in[H]} |\mathcal{U}_{A,h}|$。另定义 $\mathcal{U}_{H+1} := \{o_{\mathrm{dum}}\}$ 以简化记号。假设核心测试集 $(\mathcal{U}_h)_{h\in[H]}$ 已知，且在同一 PSR 模型类中保持不变。

定义 B.2 (PSR 秩)。给定一个 *PSR*，其PSR 秩 定义为 $d_{\mathsf{PSR}} := \max^{h\in[H]}\mathrm{rank}(D_h)$，其中$D_h := [\mathbf{q}(\tau_h)]_{\tau_h\in\mathcal{T}^h} \in \mathbb{R}^{\mathcal{U}_{h+1}\times\mathcal{T}^h}$ 是由步骤 $h \in [H]$ 的预测状态构成的矩阵。

对于部分可观测马尔可夫决策过程，显然 $d_{\mathsf{PSR}} \leqslant S$，无论核心测试集如何。

B-表示 (Chen 等人，2022a) 引入了 PSR 的 *B*-表示 这一概念，它在其一般结构条件及其分析中发挥着基础性作用。

定义 B.3 (B-表示)。对于核心测试集为 $(\mathcal{U}_h)_{h\in[H]}$ 的 *PSR*，*B*-表示是一组矩阵 $\{(\mathbf{B}_h(o_h,a_h) \in \mathbb{R}^{\mathcal{U}_{h+1}\times\mathcal{U}_h})_{h,\ o_h,\ a_h}, \mathbf{q}_0 \in \mathbb{R}^{\mathcal{U}_1}\}$，使得对于任意 $0 \leqslant h \leqslant H$、策略 π、历史$\tau_h = (o_{1:h}, a_{1:h}) \in \mathcal{T}^h$ 和核心测试 $t_{h+1} = (o_{h+1:h+W}, a_{h+1:h+W-1}) \in \mathcal{U}_{h+1}$，量 $\mathbb{P}(\tau_h, t_{h+1})$，即采取动作 $a_{1:h+W-1}$ 后观察到 $o_{1:h+W}$ 的概率，可以分解为

$$\mathbb{P}(\tau_h, t_{h+1}) = \mathbb{P}(o_{1:h+W}|\mathrm{do}(a_{1:h+W-1})) = \mathbf{e}_{t_{h+1}}^\top \cdot \mathbf{B}_{h:1}(\tau_h)\cdot\mathbf{q}_0, \tag{8}$$

其中 $\mathbf{e}_{t_{h+1}} \in \mathbb{R}^{\mathcal{U}_{h+1}}$ 是 $t_{h+1} \in \mathcal{U}_{h+1}$ 的指示向量，并且

$$\mathbf{B}_{h:1}(\tau_h) := \mathbf{B}_h(o_h,a_h)\mathbf{B}_{h-1}(o_{h-1},a_{h-1})\cdots\mathbf{B}_1(o_1,a_1).$$

基于预测状态表示的 B-表示，Chen 等人 (2022a) 提出了以下结构条件，用于预测状态表示中的样本高效学习。

定义 B.4 (B-稳定性（Chen 等人，2022a））。如果一个预测状态表示具有参数 $\Lambda_{\mathsf{B}} \geqslant 1$ （以下也称为 Λ_{B}-稳定），则它存在一个 *B*-表示，使得对于所有步骤 $h \in [H]$、策略 π和 $x \in \mathbb{R}^{\mathcal{U}_h}$，均有

$$\sum_{\tau_{h:H}=(o_h,a_h,\cdots,o_H,a_H)} \pi(\tau_{h:H}) \times |\mathbf{B}_H(o_H,a_H)\cdots\mathbf{B}_h(o_h,a_h)x| \leqslant \Lambda_{\mathsf{B}} \max\left\{\|x\|_*, \|x\|_{\Pi'}\right\}, \tag{9}$$

对于任意向量 $x = (x(t))_{t\in\mathcal{U}_h}$，我们将其 $(1,2)$-范数记为

$$\|x\|_* := \left(\sum_{\mathbf{a}\in\mathcal{U}_{A,h}} \left(\sum_{\mathbf{o}:(\mathbf{o},\mathbf{a})\in\mathcal{U}_h} |x(\mathbf{o},\mathbf{a})|\right)^2\right)^{1/2},$$

并将其 Π'-范数记为

$$\|x\|_{\Pi'} := \max_{\bar\pi} \sum_{t\in\overline{\mathcal{U}}_h} \bar\pi(t)\,|x(t)|\,,$$

其中 $\overline{\mathcal{U}}_h := \{t \in \mathcal{U}_h : \nexists t' \in \mathcal{U}_h$，满足 t 是 $t'\}$ 的前缀。

等价地，(9) 可以写作 $\|\mathcal{B}x\|_\Pi \leqslant \Lambda_{\mathsf{B}}$最大值 $\{\|x\|_*, \|x\|_{\Pi'}\}$，其中对于每个步骤 h，向量 $x \in \mathbb{R}^{\mathcal{U}_h}$，我们记作

$$\|\mathcal{B}_{H:h}x\|_\Pi := \max_\pi \sum_{\tau_{h:H}} \pi(\tau_{h:H}) \times |\mathbf{B}_{H:h}(\tau_{h:H})x|\,. \tag{10}$$

Chen 等人 (2022a) 表明，B-稳定性能够实现 PAC 学习的样本效率，我们将结果总结为以下定理。

定理 B.5 (学习预测状态表示的 PAC 上界)。假设 Θ 是一个具有相同核心测试集 $\{\mathcal{U}_h\}_{h\in[H]}$的预测状态表示类，且每个 $\theta \in \Theta$ 都存在一个 Λ_{B}-稳定的 *B*-表示，其预测状态表示秩至多为 d。那么，存在算法（OMLE/Explorative E2D/MOPS），能够以至少 $1-\delta$的概率找到一个 ε-最优策略，所需时间不超过

$$T \leqslant \widetilde{\mathcal{O}}\left(\frac{\Lambda_{\mathsf{B}}^2 dAU_A H^2 \log(\mathcal{N}_\Theta(1/T)/\delta)}{\varepsilon^2}\right) \tag{11}$$

交互回合，其中 $\mathcal{N}_\Theta$ 是 Θ 的覆盖数（参见 *Chen 等人 (2022a，定义 A.4)*）。

当 Θ是部分可观测马尔可夫决策过程的一个子类时，我们有 $\log\mathcal{N}_\Theta(1/T) = \widetilde{\mathcal{O}}\big(H(S^2A+SO)\big)$ (Chen 等人，2022a)。因此，要从上述一般定理推导出定理 3，还需对 Λ_{B} 在 m-步骤 α-揭示型部分可观测马尔可夫决策过程上的上界进行约束，这将在附录 C.2 中完成。

C 第 2 节的证明

C.1 命题 2 的证明

任取一个部分可观测马尔可夫决策过程 M，我们首先证明 $\alpha_{m+1}(M) \geqslant \alpha_m(M)$。根据 $\alpha_{m+1}(M)$(定义1)，只需证明以下结果即可。

引理 C.1. 对于任意 $h \in [H-m]$，以及任意选取的广义左逆 $\mathbb{M}_{h,m}^{+}$（对应 $\mathbb{M}_{h,m}$），矩阵$\mathbb{M}_{h,m+1}$ 存在一个广义左逆$\mathbb{M}_{h,m+1}^{+}$ 使得

$$\left\|\mathbb{M}_{h,m+1}^{+}\right\|_{*\to 1} \leqslant \left\|\mathbb{M}_{h,m}^{+}\right\|_{*\to 1}.$$

命题 2 的逆命题部分可通过具体例子直接证明。特别是，附录 F 中的构造很容易给出这样一个例子（见注F.10）。

引理 *C.1* 的证明。 固定任意一个动作 $\tilde{a} \in \mathcal{A}$。考虑矩阵 $F_{\tilde{a}} \in \mathbb{R}^{\mathcal{O}^m\mathcal{A}^{m-1}\times\mathcal{O}^{m+1}\mathcal{A}^m}$，它定义为（与之对应的唯一矩阵）由以下线性算子给出：

$$[F_{\tilde{a}}\mathbf{x}](\mathbf{o}_{h:h+m-1}, \mathbf{a}_{h:h+m-2}) := \sum_{o\in\mathcal{O}} \mathbf{x}(\mathbf{o}_{h:h+m-1}o, \mathbf{a}_{h:h+m-2}\tilde{a}), \quad \text{for all } \mathbf{x} \in \mathbb{R}^{\mathcal{O}^{m+1}\mathcal{A}^m}.$$

我们首先证明 $F_{\tilde{a}}\mathbb{M}_{h,m+1} = \mathbb{M}_{h,m}$。事实上，

$$\begin{aligned}
&[F_{\tilde{a}}\mathbb{M}_{h,m+1}]_{\mathbf{o}_{h:h+m-1}\mathbf{a}_{h:h+m-2},s} = \sum_{o\in\mathcal{O}} [\mathbb{M}_{h,m+1}]_{(\mathbf{o}_{h:h+m-1}o)(\mathbf{a}_{h:h+m-2}\tilde{a}),s} \\
&= \sum_{o\in\mathcal{O}} \mathbb{P}(o_{h:h+m} = \mathbf{o}_{h:h+m-1}o | a_{h:h+m-1} = \mathbf{a}_{h:h+m-2}\tilde{a}, s_h = s) \\
&= \mathbb{P}(o_{h:h+m-1} = \mathbf{o}_{h:h+m-1} | a_{h:h+m-2} = \mathbf{a}_{h:h+m-2}, s_h = s) = [\mathbb{M}_{h,m}]_{\mathbf{o}_{h:h+m-1}\mathbf{a}_{h:h+m-2},s}
\end{aligned}$$

对于任意 $(\mathbf{o}_{h:h+m-1}\mathbf{a}_{h:h+m-2}, s)$，这验证了该结论。因此，对于任意广义左逆 $\mathbb{M}_{h}^{+}{}_{,m}$，我们可以取

$$\mathbb{M}_{h,m+1}^{+} := \mathbb{M}_{h,m}^{+}F_{\tilde{a}}.$$

该矩阵满足 $\mathbb{M}_{h}^{+}{}_{,m+1}\mathbb{M}_{h,m+1}\mathbb{T}_{h-1} = \mathbb{M}_{h}^{+}{}_{,m}F_{\tilde{a}}\mathbb{M}_{h,m+1}\mathbb{T}_{h-1} = \mathbb{M}_{h}^{+}{}_{,m}\mathbb{M}_{h,m}\mathbb{T}_{h-1} = \mathbb{T}_{h-1}$，因此确实是 $\mathbb{M}_{h,m+1}$ 的广义左逆。此外，

$$\left\|\mathbb{M}_{h,m+1}^{+}\right\|_{*\to 1} = \left\|\mathbb{M}_{h,m}^{+}F_{\tilde{a}}\right\|_{*\to 1} \leqslant \left\|\mathbb{M}_{h,m}^{+}\right\|_{*\to 1} \|F_{\tilde{a}}\|_{*\to *},$$

因此，只需证明 $\|F_{\tilde{a}}\|_{*\to *} \leqslant 1$。为此，注意对于任意满足 $\|\mathbf{x}\|_*^2 \leqslant 1$ 的 $\mathbf{x} \in \mathbb{R}^{\mathcal{O}^{m+1}\mathcal{A}^m}$，我们有

$$\begin{aligned}
\|F_{\tilde{a}}\mathbf{x}\|_*^2 &= \sum_{\mathbf{a}_{h:h+m-2}\in\mathcal{A}^{m-1}} \left(\sum_{\mathbf{o}_{h:h+m-1}\in\mathcal{O}^m} \left| \sum_{o\in\mathcal{O}} \mathbf{x}(\mathbf{o}_{h:h+m-1}o, \mathbf{a}_{h:h+m-2}\tilde{a}) \right| \right)^2 \\
&\leqslant \sum_{\mathbf{a}_{h:h+m-2}\in\mathcal{A}^{m-1}} \left(\sum_{\mathbf{o}_{h:h+m}\in\mathcal{O}^{m+1}} |\mathbf{x}(\mathbf{o}_{h:h+m}, \mathbf{a}_{h:h+m-2}\tilde{a})| \right)^2 \\
&\leqslant \sum_{\mathbf{a}_{h:h+m-1}\in\mathcal{A}^{m}} \left(\sum_{\mathbf{o}_{h:h+m}\in\mathcal{O}^{m+1}} |\mathbf{x}(\mathbf{o}_{h:h+m}, \mathbf{a}_{h:h+m-1})| \right)^2 = \|\mathbf{x}\|_*^2.
\end{aligned}$$

这证明了 $\|F_{\tilde{a}}\|_{*\to *} \leqslant 1$，从而得出所需结果。 □

定理 C.2 的证明3

我们将从学习预测状态表示的一般结果（定理 3 ，即定理 B.5）推导出定理 Chen 等人， 2022a。为应用定理 B.5，我们首先引用以下命题，该命题基本上指出，任意 m步α揭示型部分可观测马尔可夫决策过程都是 B 稳定的，且 $\Lambda_{\mathsf{B}} \leqslant \alpha^{-1}$。

命题 C.2。 任何 m-步骤的 α-揭示型部分可观测马尔可夫决策过程都是具有核心测试集 $\mathcal{U}_h = (\mathcal{O} \times_{\mathcal{A}})^{\min\{m-1,\, H-h\}} \times \mathcal{O}$的 α^{-1}-稳定预测状态表示，即它具有一个 $\Lambda_{\mathsf{B}} \leqslant \alpha^{-1}$-稳定的 B-表示。

因此，对于 $\mathcal{M}$ 一类 m-步骤 α-揭示型部分可观测马尔可夫决策过程， $\mathcal{M}$ 它也是一类具有共同核心测试集的预测状态表示，其中每个 $M \in \mathcal{M}$ 都是 α^{-1}-稳定的，其 PSR 秩至多为 S ，并且 $U_A = A^{m-1}$。于是，定理 B.5表明，可以使用 ε-最优策略来学习 $\mathcal{M}$ ，所需样本复杂度为 OMLE、 Explorative E2D或 MOPS。

$$\widetilde{\mathcal{O}}\left(\frac{SA^m H^2 \log(\mathcal{N}_{\mathcal{M}}(1/T)/\delta)}{\alpha^2 \varepsilon^2}\right),$$

并且我们还有 $\log \mathcal{N}_{\mathcal{M}}(1/T) = \widetilde{\mathcal{O}}\big(H(S^2A + SO)\big)$ (Chen 等人， 2022a）。结合这些事实即可完成定理 3的证明。

□

命题 *C.2* 的证明。 Chen 等人(2022a, 附录 B.3.3)表明，任何 m-步骤 α-揭示型部分可观测马尔可夫决策过程M 都是一个以 $\mathcal{U}_h = (\mathcal{O} \times \mathcal{A})^{\min\{m-1, H-h\}} \times \mathcal{O}$为核心测试集的 α^{-1}-稳定预测状态表示，并明确为其构造了如下 B-表示：当 $h \leqslant H - m$ 时，设

$$\mathbf{B}_h(o, a) = \mathbb{M}_{h+1} \mathbb{T}_{h,a} \operatorname{diag}\left(\mathbb{O}_h(o|\cdot)\right) \mathbb{M}_h^+, \qquad h \in [H-m], \tag{12}$$

当 $h > H - m$ 时，取

$$\mathbf{B}_h(o_h, a_h) = \left[\mathbb{1}(t_h = (o_h, a_h, t_{h+1}))\right]_{(t_{h+1}, t_h) \in \mathcal{U}_{h+1} \times \mathcal{U}_h} \in \mathbb{R}^{\mathcal{U}_{h+1} \times \mathcal{U}_h}, \tag{13}$$

其中 $\mathbb{1}(t_h = (o_h, a_h, t_{h+1}))$ 当 t_h 等于 (o_h, a_h, t_{h+1})时为 1，否则为 0。

接着，根据 Chen 等人 (2022a, 引理 B.13)，对于任意 $1 \leqslant h \leqslant H$, $x \in \mathbb{R}^{|\mathcal{U}_h|}$，有

$$\|\mathcal{B}_{H:h} x\|_{\Pi} = \text{最大}_{\pi} \sum_{\tau_{h:H}} \|\mathbf{B}_H(o_H, a_H) \cdots \mathbf{B}_h(o_h, a_h) x\|_1 \times \pi(\tau_{h:H}) \leqslant \text{最大}\left\{\left\|\mathbb{M}_h^+ x\right\|_1, \|x\|_{\Pi'}\right\} \leqslant \alpha^{-1} \text{最大}\left\{\|x\|_*, \|x\|_{\Pi'}\right\}.$$

因此，（12）和（13中给出的 α^{-1}-稳定 B-表示确实成立，从而完成了证明。

□

D Ingster 方法基础

本节首先介绍分析困难实例时常用的基本记号，然后陈述 Ingster 和 Suslina Ingster 和 Suslina (2012）提出的证明信息论下界的 Ingster 方法。回顾一下，我们已在附录 B中介绍了序贯决策过程的形式化定义。

序贯决策过程的算法序贯决策过程的算法 $\mathfrak{A}$ （回合数 T 固定）由一组 HT函数 $\mathfrak{A} = \{\pi_t^{\mathfrak{A}}, h\}_{h \in [H], t \in [T]}$构成，其中 $\pi_t^{\mathfrak{A}}, h$将所有过去的历史记录与当前观测组成的元组 $(\tau^{(1)}, \cdots, \tau^{(t-1)}, \tau_{h-1}^{(t)}, o_h^{(t)})$ 映射为一个动作分布 $\Delta(\mathcal{A})$，再从中采样得到下一个动作 $a_h^{(t)} \sim \pi_t^{\mathfrak{A}}, h(\cdot|\tau^{(1}: t-1), \tau_{h-1}^{(t)}, o_h^{(t)})$。交互结束时，算法通过取 $\pi^{\text{out}} = \pi_{\text{output}}^{\mathfrak{A}}(\tau^{1:T})$ 输出一个 $\pi^{\text{out}} \in \Pi$。

对于任意算法 $\mathfrak{A}$（回合数 T 固定），我们将 $\mathbb{P}_M^{\mathfrak{A}}$记为模型 M 和算法 $\mathfrak{A}$下 $(\tau^{(1)}, \tau^{(2)}, \cdots, \tau^{(T)})$的分布律。需要指出的是，尽管我们的表述似乎只允许确定性算法，其中每个 $\pi_t^{\mathfrak{A}}, h$ 都是到 $\Delta(\mathcal{A})$的确定性映射，但实际上也允许随机化算法：任何随机化算法都可以写成由 ω 参数化的确定性算法$\mathfrak{B}(\omega)$ 的混合，其中 ω 服从分布 $\omega \sim \zeta$；此外，对于任意 $\mathfrak{B}(\omega)$ 和 ζ，都存在一个确定性算法 $\mathfrak{A}$ ，使得由 $\mathfrak{B}$ 和 $\mathfrak{A}$ 诱导的 $\tau^{1:T}$ 的边缘分布相同，即 $\mathbb{E}_{\omega \sim \zeta}[\mathbb{P}_M^{\mathfrak{B}(\omega)}(\cdot)] = \mathbb{P}_M^{\mathfrak{A}}(\cdot)$。

具有随机停止时间的算法 我们的分析需要考虑具有随机停止时间的算法。具有随机停止时间的算法 $\mathfrak{A}$（交互次数至多为 T）由一组 HT 函数 $\{\pi_{t,h}^{\mathfrak{A}}\}_{h\in[H],t\in[T]}$ 以及一个退出条件 exit 指定，其中 $\pi_{t,h}^{\mathfrak{A}}$ 是第 t 个回合第 h步的策略，而 exit 是一个确定性函数，使得

$$\mathsf{exit}(\tau^{(1)},\cdots,\tau^{(t)})\in\{\mathsf{TRUE},\mathsf{FALSE}\}.$$

一旦退出$(\tau^{(1)},\ \cdots,\ \tau^{(\mathsf{T})})=$TRUE 或 $\mathsf{T}=T$，算法 $\mathfrak{A}$会在第 T 个回合结束时终止。随机变量 T（由退出条件 exit 确定）显然是一个停止时间。我们记 $\mathbb{P}_M^{\mathfrak{A}}$为模型 M和算法 $\mathfrak{A}$下 $(\tau^{(1)},\ \tau^{(2)},\ \cdots,\ \tau^{(\mathsf{T})})$的分布律。

以下　　引理及相关讨论适用于具有或不具有随机停止时间的算法　　.

引理 D.1 (Ingster 方法)。 对于一族序贯决策过程 $(\mathbb{P}_M)_{M\in\mathcal{M}}$、关于 $\mathcal{M}$的分布 ζ、一个参考模型 $0\in\mathcal{M}$, 以及一个与环境交互 T 个回合的算法 $\mathfrak{A}$ （其中 T 为停止时间），有

$$1+\chi^2(\mathbb{E}_{M\sim\zeta}\left[\mathbb{P}_M^{\mathfrak{A}}\right]\parallel\mathbb{P}_0^{\mathfrak{A}})=\mathbb{E}_{M,M'\sim_{\mathrm{iid}}\zeta}\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_M(\tau^{(t)})\mathbb{P}_{M'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right].$$

证明。 我们只需考虑 $\mathfrak{A}$ 具有随机停止时间 T 的情况。根据定义， $\mathbb{P}_M^{\mathfrak{A}}$ 支撑于以下集合:

$$\Omega_{0\circ}=\left\{\omega=\tau^{(1:\mathsf{T})}:\forall t<\mathsf{T},\mathsf{exit}(\tau^{(1:t)})=\mathsf{FALSE},\ \text{and either }\mathsf{T}=T\text{ or }\mathsf{exit}(\tau^{(1:\mathsf{T})})=\mathsf{TRUE}\right\}$$

对于任意 $(\tau^{(1)},\cdots,\tau^{(\mathsf{T})})\in\Omega_0$,，我们有

$$\begin{aligned}\mathbb{P}_M^{\mathfrak{A}}(\tau^{(1)},\cdots,\tau^{(\mathsf{T})})&=\prod_{t=1}^{\mathsf{T}}\mathbb{P}_M^{\mathfrak{A}}(\tau^{(t)}|\tau^{(1:t-1)})\\&=\prod_{t=1}^{\mathsf{T}}\prod_{h=1}^{H}\mathbb{P}_M(o_h^{(t)}|\tau_{1:h}^{(t)})\times\pi_{t,h}^{\mathfrak{A}}(a_h^{(t)}|\tau^{(1:t-1)},\tau_{1:h}^{(t)},o_h^{(t)})\\&=\prod_{t=1}^{\mathsf{T}}\mathbb{P}_M(\tau^{(t)})\times\prod_{t=1}^{\mathsf{T}}\prod_{h=1}^{H}\pi_{t,h}^{\mathfrak{A}}(a_h^{(t)}|\tau^{(1:t-1)},\tau_{1:h}^{(t)},o_h^{(t)}).\end{aligned}\tag{14}$$

因此，根据 χ^2 散度的定义，我们有

$$\begin{aligned}1+\chi^2(\mathbb{E}_{M\sim\zeta}\left[\mathbb{P}_M^{\mathfrak{A}}\right]\parallel\mathbb{P}_0^{\mathfrak{A}})&=\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\left(\frac{\mathbb{E}_{M\sim\zeta}\left[\mathbb{P}_M^{\mathfrak{A}}(\tau^{(1)},\cdots,\tau^{(\mathsf{T})})\right]}{\mathbb{P}_0^{\mathfrak{A}}(\tau^{(1)},\cdots,\tau^{(\mathsf{T})})}\right)^2\right]\\&=\mathbb{E}_{M,M'\sim\zeta}\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\frac{\mathbb{P}_M^{\mathfrak{A}}(\tau^{(1)},\cdots,\tau^{(\mathsf{T})})\mathbb{P}_{M'}^{\mathfrak{A}}(\tau^{(1)},\cdots,\tau^{(\mathsf{T})})}{\mathbb{P}_0^{\mathfrak{A}}(\tau^{(1)},\cdots,\tau^{(\mathsf{T})})^2}\right]\\&=\mathbb{E}_{M,M'\sim\zeta}\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_M(\tau^{(t)})\mathbb{P}_{M'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right],\end{aligned}$$

其中，最后一个等式由式（14）得出。这就证明了该引理。 □

因此，为了给 $\chi^2(\mathbb{E}_{M\sim\zeta}\left[\mathbb{P}_M^{\mathfrak{A}}\right]\parallel\mathbb{P}_0^{\mathfrak{A}})$ 找到上界，我们只需给该量找到上界

$$\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_M(\tau^{(t)})\mathbb{P}_{M'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right]=\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\prod_{h=1}^{H}\frac{\mathbb{P}_M(o_h^{(t)}|\tau_{h-1}^{(t)})\mathbb{P}_{M'}(o_h^{(t)}|\tau_{h-1}^{(t)})}{\mathbb{P}_0(o_h^{(t)}|\tau_{h-1}^{(t)})^2}\right].\tag{15}$$

为此，我们将利用以下事实（该事实源自引理 A.2 和（15））：

$$\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_M(\tau^{(t)})\mathbb{P}_{M'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\cdot\exp\left(-\sum_{t=1}^{\mathsf{T}}\sum_{h=1}^{H}\log I_{M,M'}(\tau_{h-1}^{(t)})\right)\right]=1, \tag{16}$$

其中 $I_{M,\ M'}(\tau_{h-1})$ 定义为

$$I_{M,M'}(\tau_{h-1}):=\mathbb{E}_0\left[\frac{\mathbb{P}_M(o_h|\tau_{h-1})\mathbb{P}_{M'}(o_h|\tau_{h-1})}{\mathbb{P}_0(o_h|\tau_{h-1})^2}\middle|\tau_{h-1}\right]. \tag{17}$$

提前停止算法 考虑算法 $\mathfrak{A}$ 与环境交互固定数量的回合 T，并考虑退出准则 exit。我们定义提前停止算法 $\mathfrak{A}$(exit)，它执行算法 $\mathfrak{A}$，直到满足 exit $=$ TRUE（或达到 T）。显然，$\mathfrak{A}$(exit) 是一个具有随机停止时间的算法。关于将算法 $\mathfrak{A}$ 更改为其停止版本 $\mathfrak{A}$(exit)后，全变差距离$D_{\mathrm{TV}}\left(\mathbb{E}_{M\sim\zeta}\left[\mathbb{P}_M^{\mathfrak{A}}\right],\ \mathbb{P}_0^{\mathfrak{A}}\right)$ 受到多大影响，我们有以下引理。

引理 D.2. 有

$$D_{\mathrm{TV}}\left(\mathbb{E}_{M\sim\zeta}\left[\mathbb{P}_M^{\mathfrak{A}(\mathsf{exit})}\right],\mathbb{P}_0^{\mathfrak{A}(\mathsf{exit})}\right)\geqslant D_{\mathrm{TV}}\left(\mathbb{E}_{M\sim\zeta}\left[\mathbb{P}_M^{\mathfrak{A}}\right],\mathbb{P}_0^{\mathfrak{A}}\right)-\mathbb{P}_0^{\mathfrak{A}}(\exists t<T,\mathsf{exit}(\tau^{(1:t)})=\mathsf{TRUE}).$$

证明。 我们考虑事件 $\Omega=\{\omega=\tau^{(1:\ T)}:\ \ \forall t<T,\ \text{退出}(\tau^{(1:\ t)})=\mathsf{FALSE}\}$。为证明该引理，只需验证 $\mathbb{P}_M^{\mathfrak{A}}|_\Omega=\mathbb{P}_M^{\mathfrak{A}(\mathsf{exit})}|_\Omega$，然后应用引理 A.4。

事实上，对于 $\omega=\tau^{(1:\ T)}\in\Omega$，我们有，对所有 $t<T$，退出$(\tau^{(1:\ t)})=\mathsf{FALSE}$。那么，由 (14) 可得

$$\mathbb{P}_M^{\mathfrak{A}(\mathsf{exit})}(\tau^{(1:T)})=\prod_{t=1}^{T}\mathbb{P}_M(\tau^{(t)})\times\prod_{t=1}^{T}\prod_{h=1}^{H}\pi_{t,h}^{\mathfrak{A}}(a_h^{(t)}|\tau^{(1:t-1)},\tau_{1:h}^{(t)},o_h^{(t)})=\mathbb{P}_M^{\mathfrak{A}}(\tau^{(1:T)}),$$

因此 $\mathbb{P}_M^{\mathfrak{A}(\mathsf{exit})}(\omega)=\mathbb{P}_M^{\mathfrak{A}}(\omega)$ 对于任意 $\omega\in\Omega$。应用引理A.4 即可证明该引理。 □

E 定理4的证明

我们首先在附录E.1中构造一族困难实例。我们在命题E.1中陈述这族困难实例的 PAC 下界。定理4 随后可直接由命题E.1 推论得出。

E.1 困难实例的构造及定理4的证明

我们考虑以下一族单步揭示型部分可观测马尔可夫决策过程 $\mathcal{M}$，其具有一组超参数(ε，σ，n，K，H)。$\mathcal{M}$ 中的所有部分可观测马尔可夫决策过程具有相同的时域长度 H、状态空间 $\mathcal{S}$、动作空间 $\mathcal{A}$和观测空间 $\mathcal{O}$，定义如下。

- 状态空间 $\mathcal{S}=\mathcal{S}_{\text{树}}\bigsqcup\{s_\oplus,\ s_\ominus\}$，其中 $\mathcal{S}_{\text{树}}$是一棵层级为 n 的二叉树（因此$|\mathcal{S}_{\text{树}}|=2^n-1$）。令 s_0 为 $\mathcal{S}_{\text{树}}$的根节点，并令 $\mathcal{S}_{\text{叶节点}}$为 $\mathcal{S}_{\text{树}}$的叶节点集合，其中$|\mathcal{S}_{\text{叶节点}}|=2^{n-1}$。
- 观测空间 $\mathcal{O}=\mathcal{S}_{\text{树}}\bigsqcup\{o_1^+,\ o_1^-,\ \cdots,\ o_K^+,\ o_K^-\}\bigsqcup\{\text{好},\ \text{坏}\}$。请注意，这里我们稍微滥用了记号，重复使用 $\mathcal{S}_{\text{树}}$ 来表示状态集合及其对应的观测集合，因为每个状态 $s\in\mathcal{S}_{\text{树}}\subset\mathcal{S}$ 都对应一个唯一的观测 $o_s\in\mathcal{S}_{\text{树}}\subset\mathcal{O}$；在上下文明确时，我们也将其记为 s。

- 动作空间 $\mathcal{A}=\{0,\ \mathbf{1},\ \cdots,\ A-1\}$。

模型参数 每个非零部分可观测马尔可夫决策过程模型 $M=M_{\theta,\ \mu}\in\mathcal{M}\backslash\{M_0\}$ 由两个参数(θ，μ)指定。这里 $\mu\in\{-1,\ +1\}^K$，以及 $\theta=(h^\star,\ s^\star,\ a^\star,$ **一个** $\star)$，其中

- $s^\star\in\mathcal{S}_{\mathsf{leaf}},\ a^\star\in\mathcal{A}_c:=\{1,\cdots,A-1\}$.

- $h^\star \in \{n+1, \cdots, H-1\}$。
- $\mathbf{a}^\star = (\mathbf{a}^\star_{h^\star+1}, \ldots, \mathbf{a}^\star_{H-1}) \in \mathcal{A}^{H-h^\star-1}$ 是一个由 $h^\star+1, \cdots, H-1$索引的动作序列 .

对于任意部分可观测马尔可夫决策过程 $M_{\theta,\mu}$，其发射动态和转移动态 $\mathbb{P}_{\theta,\mu} := \mathbb{P}_{M_{\theta,\mu}}$ 定义如下。

发射动态

- 在状态 $s \in \mathcal{S}_{\text{树}}$下，智能体总是接收其自身 s 作为观测（与之对应的唯一观测）。
- 在状态 $s_\oplus$ 和步骤 $h < H$下，发射动态定义为

$$\mathbb{O}_{h;\mu}(o_i^+|s_\oplus) = \frac{1+\sigma\mu_i}{2K}, \qquad \mathbb{O}_{h;\mu}(o_i^-|s_\oplus) = \frac{1-\sigma\mu_i}{2K}, \qquad \forall i \in [K].$$

- 在状态 $s_\ominus$ 以及步骤 $h < H$，观测从 $\mathcal{O}_o := \{o_1^+, o_1^-, \cdots, o_K^+, o_K^-$中均匀抽取 $\}$:

$$\mathbb{O}_h(o_i^+|s_\ominus) = \mathbb{O}_h(o_i^-|s_\ominus) = \frac{1}{2K}, \qquad \forall i \in [K].$$

  此处省略下标 μ，以强调该动力学不依赖于 μ .
- 在步骤 H，$\{s_\oplus, s_\ominus\}$ 处的发射动态为

$$\mathbb{O}_H(\mathsf{good}|s_\oplus) = \frac{3}{4}, \qquad \mathbb{O}_H(\mathsf{bad}|s_\oplus) = \frac{1}{4},$$
$$\mathbb{O}_H(\mathsf{good}|s_\ominus) = \frac{1}{4}, \qquad \mathbb{O}_H(\mathsf{bad}|s_\ominus) = \frac{3}{4}.$$

转移动态 在每个回合中，智能体总是从 s_0 开始。

- 在树的任意节点 $s \in \mathcal{S}_{\text{树}} \backslash \mathcal{S}_{\text{叶子}}$处，都有三种可用动作：等待 $=0$、左 $=1$ 和右 $=2$。智能体可以选择等待以留在 s，选择左以转移到 s 的左子节点，或选择右以转移到 s的右子节点。[6]
- 在任意 $s \in \mathcal{S}_{\text{叶子}}$处，智能体可以选择等待动作 $=0$ 以留在 s（即 $\mathbb{P}(s|s, \text{等待}) = 1$）；否则，对于$s \in \mathcal{S}_{\text{叶子}}$，$h \in [H-1]$，$a \neq$ 等待（即 $a \in \mathcal{A}_c$），

$$\mathbb{P}_{h;\theta}(s_\oplus|s,a) = \varepsilon \cdot \mathbb{1}(h = h^\star, s = s^\star, a = a^\star),$$
$$\mathbb{P}_{h;\theta}(s_\ominus|s,a) = 1 - \varepsilon \cdot \mathbb{1}(h = h^\star, s = s^\star, a = a^\star),$$

  其中，我们使用下标 θ 强调转移概率 $\mathbb{P}_{h;\theta}$ 对 θ的依赖。换言之，在步骤 h，处于状态 $s \in \mathcal{S}_{\mathsf{leaf}}$时，执行 $a \in \mathcal{A}_c$后，任何叶节点都会转移到 $\{s_\oplus, s_\ominus\}$之一；只有在状态 $s^\star$ 的步骤 $h^\star$ 执行 $a^\star$，才会以较小的概率 ε转移到状态 $s_\oplus$；在其他情况下，系统都会以概率 $\mathbf{1}$ 转移到状态 $s_\ominus$。
- 在状态 $s_\oplus$下，我们设定

$$\mathbb{P}_{h;\theta}(s_\oplus|s_\oplus,a) = \begin{cases} 1, & a = \mathbf{a}^\star_h, \\ 0, & a \neq \mathbf{a}^\star_h, \end{cases}, \qquad \mathbb{P}_{h;\theta}(s_\ominus|s_\oplus,a) = \begin{cases} 0, & a = \mathbf{a}^\star_h, \\ 1, & a \neq \mathbf{a}^\star_h. \end{cases}$$

- 状态 $s_\ominus$ 是一个吸收状态，即 $\mathbb{P}_h(s_\ominus|s_\ominus,a) = 1$ 对所有 $a \in \mathcal{A}$均成立。

奖励奖励函数已知（且仅取决于观测）：前 $H-1$ 个步骤不提供奖励；在步骤 H，我们设定 $r_H(\text{好}) = 1$、$r_H(\text{坏}) = 0$、$r_H(s_0) = (1+\varepsilon)/4,$，而对于任何其他 $o \in \mathcal{O}$，则设定 $r_H(o) = 0$。

[6]对于动作 $a \in \{3, \cdots, A-1\}$，a的效果与等待相同。

参考模型我们用 M_0 （或简称 0）指代零模型（参考模型）。零模型M_0 的转移和发射与任何非零模型相同，但智能体无论采取什么动作，总会在 $s \in \mathcal{S}_{\mathsf{leaf}}$ 等待时到达 $s_{\ominus}$，并且 $h \in [H-1]$（即 $\mathbb{P}_{h;M_0}(s_{\ominus}|s, a) = 1$ 对于任何 $s \in \mathcal{S}_{\mathsf{leaf}}$、 $a \in \mathcal{A}_c$、$h \in [H-1]$）。在此模型中， $s_{\oplus}$ 不可达，因此我们无需指定 $s_{\oplus}$处的发射动态。

我们在以下命题中给出上述部分可观测马尔可夫决策过程模型类的 PAC 学习样本复杂度下界 $\mathcal{M}$ ，并在附录 E.2中证明该命题。

命题 E.1。对于给定的 $\varepsilon \in (0, 0.1], \sigma \in (0, \frac{1}{2H}]$, $n \geqslant 1$, $K \geqslant 1$, $H \geqslant 4n$，我们在上文构造的模型类 $\mathcal{M}$ 满足以下性质：

1. $|\mathcal{S}| = 2^n + 1$, $|\mathcal{O}| = 2^n + 2K + 1$, $|\mathcal{A}| = A$。
2. 对于每个 $M \in \mathcal{M}$ （包括零模型 M_0）， M 均为单步可揭示，且 $\alpha_1(M)^{-1} \leqslant 1 + \frac{2}{\sigma}$ 。
3. $\log|\mathcal{M}| \leqslant K \log 2 + H \log A + \log(SAH)$.
4. 假设算法 $\mathfrak{A}$ 与环境交互 T 个回合，并返回 π^{out} ，使得

$$\mathbb{P}_M^{\mathfrak{A}}\left(V_M^\star - V_M(\pi^{\mathrm{out}}) < \frac{\varepsilon}{8}\right) \geqslant \frac{3}{4}$$

对任意 $M \in \mathcal{M}$ 均成立。那么必有

$$T \geqslant \frac{1}{20000} \min\left\{\frac{|\mathcal{S}_{\mathsf{leaf}}|\, K^{1/2} AH}{\sigma^2 \varepsilon^2}, \frac{|\mathcal{S}_{\mathsf{leaf}}|\, A^{H/2} H}{\varepsilon^2}\right\},$$

这里回顾一下， $|\mathcal{S}_{\mathsf{leaf}}| = 2^{n-1}$。

定理 4 的证明 在命题 E.1 中，适当选择 σ、 n、 K，并选择重新缩放后的 ε，即可得到定理 4。更具体地说，我们可以将 $n \geqslant 1$ 取为满足 $2^n \leqslant \min\{S-1, (O-1)/2\}$, 的最大整数，并取 $K = \left\lfloor \frac{O-2^n-1}{2} \right\rfloor \geqslant \frac{O-1}{4}$、$\varepsilon' = \varepsilon/8$, 和 $\sigma = \frac{2}{\alpha^{-1}-1} \leqslant \frac{1}{2H}$。将参数 $(\varepsilon'、\ \sigma、\ n、\ K、\ H)$ 代入命题 E.1，即完成定理 4 的证明。 □

E.2 命题 E.1 的证明

本节中陈述的所有命题和引理均在附录 E.3-E.6中给出证明。

断言 1 可直接通过计算构造中状态、观测和动作的数量得出$\mathcal{M}$。断言 3 则由我们有 $|\mathcal{M}| = |\{(h^\star,\ s^\star,\ a^\star,$ 一个$^\star)\}| \times |\{\pm 1\}^K| + 1 \leqslant HSA \times A^H \times 2^K$ 得出。取对数即可得到该断言。

断言 2 可直接由以下命题得出，其证明见附录 E.3。

命题 E.2。对于任意 $M \in \mathcal{M}$, M 都是具有 $\alpha_1(M)^{-1} \leqslant \frac{2}{\sigma} + 1$ 的单步可揭示模型。

现在我们证明断言 4（样本复杂度下界）。首先利用以下引理，根据 $\mathcal{M}$ 的结构，将 PAC 学习问题与检验问题联系起来。直观而言，该引理指出，任何 $M \neq 0$ 的近最优策略都无法“停留”在 s_0，而模型$M = 0$ 的近最优策略则必须“停留”在 s_0。该引理的证明见附录 E.4。

引理 E.3 （将策略次优性与停留概率联系起来）。 对于任意 $M \in \mathcal{M}$ 满足 $M \neq 0$以及任意策略 π，均有

$$V_M^\star - V_M(\pi) \geqslant \frac{\varepsilon}{4} \mathbb{P}_M^\pi(o_H = s_0). \tag{18}$$

另一方面，对于参考模型 0 和任意策略 π，我们有

$$V_0^\star - V_0(\pi) \geqslant \frac{\varepsilon}{4} \mathbb{P}_0^\pi(o_H \neq s_0). \tag{19}$$

注意　概率 $\mathbb{P}_M^{\pi}(o_H = s_0)$ 实际上并不依赖于模型 $M \in \mathcal{M}$，即 .

$$\mathbb{P}_M^{\pi}(o_H = s_0) = \mathbb{P}_0^{\pi}(o_H = s_0).$$

这是因为智能体离开 s_0 后，对于任何模型 $M \in \mathcal{M}$，都不会再回来。接下来，我们定义 $w(\pi) := \mathbb{P}_0^{\pi}(o_H = s_0)$。注意， π^{out} 是依赖于观测历史$\tau^{1:T}$的输出策略，因此 $w(\pi^{\mathrm{out}})$是观测历史 $\tau^{1:T}$的确定性函数。

根据引理 E.3 以及我们的假设 $\mathbb{P}_M^{\mathfrak{A}}\left(V_M^{\star} - V_M(\pi^{\mathrm{out}}) < \frac{\varepsilon}{8}\right) \geqslant \frac{3}{4}$对于任意 $M \in \mathcal{M}$均成立，我们有

$$\mathbb{P}_0^{\mathfrak{A}}\left(1 - w(\pi^{\mathrm{out}}) < \frac{1}{2}\right) \geqslant \frac{3}{4}, \qquad \text{while} \qquad \mathbb{P}_M^{\mathfrak{A}}\left(w(\pi^{\mathrm{out}}) < \frac{1}{2}\right) \geqslant \frac{3}{4}, \qquad \forall M \neq 0.$$

现在，我们将 $\mu \sim \mathsf{Unif}(\{\pm 1\}^K)$ 视为参数 μ上的均匀先验。对于任意固定的 θ，当 $\mu \sim \mathsf{Unif}(\{\pm 1\}^K)$时，我们考虑对非零模型 $M = (\theta, \mu)$ 上的上述量取平均，

$$\mathbb{E}_{\mu \sim \mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\left(w(\pi^{\mathrm{out}}) < \frac{1}{2}\right) = \mathbb{E}_{\mu \sim \mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\left(w(\pi^{\mathrm{out}}) < \frac{1}{2}\right)\right] \geqslant \frac{3}{4}.$$

然而，我们还有

$$\mathbb{P}_0^{\mathfrak{A}}\left(w(\pi^{\mathrm{out}}) < \frac{1}{2}\right) = 1 - \mathbb{P}_0^{\mathfrak{A}}\left(w(\pi^{\mathrm{out}}) \geqslant \frac{1}{2}\right) \leqslant 1 - \mathbb{P}_0^{\mathfrak{A}}\left(w(\pi^{\mathrm{out}}) > \frac{1}{2}\right) \leqslant \frac{1}{4}.$$

因此，根据全变差距离的定义，我们必有

$$D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}, \mathbb{E}_{\mu \sim \mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\right) \geqslant \left|\mathbb{P}_0^{\mathfrak{A}}\left(w(\pi^{\mathrm{out}}) < \frac{1}{2}\right) - \mathbb{E}_{\mu \sim \mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\left(w(\pi^{\mathrm{out}}) < \frac{1}{2}\right)\right| \geqslant \frac{1}{2}. \tag{20}$$

作为证明的核心，我们现在使用（20）来推导关于 T 的下界。回顾一下， $\mathbb{P}_M^{\mathfrak{A}}$ 是通过让 $\mathfrak{A}$ 与模型 M 交互而诱导出的 $(\tau^{(1)}, \tau^{(2)}, \cdots, \tau^{(\mathsf{T})})$ 的分布律。对于任意事件 $E \subseteq (\mathcal{O} \times \mathcal{A})^H$，我们将 E 的访问次数记为

$$N(E) := \sum_{t=1}^{T} \mathbb{1}(\tau^{(t)} \in E).$$

由于 $N(E)$是 $\tau^{(1:\ T)}$的函数，因此对于任意 $M \in \mathcal{M}$，我们可以讨论其在分布 $\mathbb{P}_M^{\mathfrak{A}}$下的期望。我们在此给出以下引理，讨论某些良好事件的期望访问次数下界；其证明见附录 E.5。

引理 E.4。 固定一个 $\theta = (h^{\star}$、 $s^{\star}$、 $a^{\star}$， 一个$^{\star})$。我们考虑以下事件

$$E_{\mathsf{rev},h}^{\theta} := \left\{o_{h^{\star}} = s^{\star}, a_{h^{\star}:h} = (a^{\star}, \mathbf{a}_{h^{\star}+1:h}^{\star})\right\}, \ \forall h \in \{h^{\star}+1, \ldots, H-2\},$$
$$E_{\mathsf{correct}}^{\theta} := \{o_{h^{\star}} = s^{\star}, a_{h^{\star}:H-1} = (a^{\star}, \mathbf{a}^{\star})\}.$$

那么，对于任意满足 $\delta :=$ $D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}、\ \mathbb{E}_{\mu \sim \mathsf{unif}}\left[\mathbb{P}_{\theta,\ \mu}^{\mathfrak{A}}\right]\right) > 0$ 的算法 $\mathfrak{A}$，我们有

$$\textit{either} \sum_{h=h^{\star}}^{H-2} \mathbb{E}_0^{\mathfrak{A}}\left[N(E_{\mathsf{rev},h}^{\theta})\right] \geqslant \frac{\delta^3\sqrt{K}}{54\varepsilon^2\sigma^2} - \frac{H\delta}{6}, \qquad \textit{or} \qquad \mathbb{E}_0^{\mathfrak{A}}\left[N(E_{\mathsf{correct}}^{\theta})\right] \geqslant \frac{\delta^3}{54\varepsilon^2} - \frac{\delta}{6}.$$

应用引理 E.4 对于任意参数元组 $\theta = (h^{\star},\ s^{\star},\ a^{\star},\ \mathbf{a}^{\star})$ 满足 $\delta = \frac{1}{2}$，我们得到

$$\text{either} \sum_{h=h^{\star}}^{H-2} \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{rev},h}^{(h^{\star},s^{\star},a^{\star},\mathbf{a}^{\star})}\right)\right] \geqslant \frac{\sqrt{K}}{1000\varepsilon^2\sigma^2}, \quad \text{or} \quad \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{correct}}^{(h^{\star},s^{\star},a^{\star},\mathbf{a}^{\star})}\right)\right] \geqslant \frac{1}{1000\varepsilon^2}, \tag{21}$$

根据我们的选择，$\varepsilon \in (0,\ 0.1]$ 且 $\sigma \in (0,\ \frac{1}{2H}]$。

固定一个元组 $(h^\star,\ s^\star,\ a^\star)$ 满足 $s^\star \in \mathcal{S}_{\mathsf{leaf}},\ a^\star \in \mathcal{A}_c,\ h^\star \in [n+1,\ \frac{H}{2}]$。根据（21），我们知道，对所有 $\mathbf{a} \in \mathcal{A}^{H-h^\star-1}$，都有

$$\sum_{h=h^\star}^{H-2} \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{rev},h}^{(h^\star,s^\star,a^\star,\mathbf{a})}\right)\right] + A^{H-h^\star-1}\cdot \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,\mathbf{a})}\right)\right] \geqslant \frac{1}{1000}\min\left\{\frac{\sqrt{K}}{\varepsilon^2\sigma^2}, \frac{A^{H/2-1}}{\varepsilon^2}\right\} =: \omega. \quad (22)$$

注意，根据定义，

$$\begin{aligned}\sum_{\mathbf{a}\in\mathcal{A}^{H-h^\star-1}} \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,\mathbf{a})}\right)\right] &= \sum_{\mathbf{a}\in\mathcal{A}^{H-h^\star-1}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star:H-1}=(a^\star,\mathbf{a}))]\\ &= \mathbb{E}_0^{\mathfrak{A}}\left[\sum_{\mathbf{a}\in\mathcal{A}^{H-h^\star-1}} N(o_{h^\star}=s^\star, a_{h^\star:H-1}=(a^\star,\mathbf{a}))\right]\\ &= \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)],\end{aligned}$$

类似地，对于每个 $h \in [h^\star,\ H-2]$，都有

$$\begin{aligned}\sum_{\mathbf{a}\in\mathcal{A}^{H-h^\star-1}} \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{rev},h}^{(h^\star,s^\star,a^\star,\mathbf{a})}\right)\right] &= \sum_{\mathbf{a}\in\mathcal{A}^{H-h^\star-1}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star:h}=(a^\star,\mathbf{a}_{h^\star+1:h}))]\\ &= \sum_{\mathbf{a}_{h^\star+1:h}\in\mathcal{A}^{h-h^\star}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star:h}=(a^\star,\mathbf{a}_{h^\star+1:h}))]\cdot \sum_{\mathbf{a}_{h+1:H-1}\in\mathcal{A}^{H-h-1}} 1\\ &= \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)]\cdot A^{H-h-1}.\end{aligned}$$

因此，将界 (22) 对所有 $\mathbf{a} \in \mathcal{A}^{H-h^\star-1}$求和，可得

$$\begin{aligned}A^{H-h^\star-1}\omega = \sum_{\mathbf{a}\in\mathcal{A}^{H-h^\star-1}} \omega &\leqslant \sum_{\mathbf{a}\in\mathcal{A}^{H-h^\star-1}}\left[\sum_{h=h^\star}^{H-2} \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{rev},h}^{(h^\star,s^\star,a^\star,\mathbf{a})}\right)\right] + A^{H-h^\star-1}\cdot \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,\mathbf{a})}\right)\right]\right]\\ &= \left(\sum_{h=h^\star}^{H-2} A^{H-h-1} + A^{H-h^\star-1}\right)\mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)]\\ &\leqslant 3A^{H-h^\star-1}\mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)],\end{aligned}$$

其中，最后一个不等式是因为 $\sum_{h=h^\star}^{H-2} A^{H-h-1} = \frac{A^{H-h}-A}{A-1} \leqslant 2A^{H-h-1}$ for $A \geqslant 3$.

因此，我们证明了对于每个 $s^\star \in \mathcal{S}_{\mathsf{leaf}},\ a^\star \in \mathcal{A}_c,\ h^\star \in [n+1,\ \frac{H}{2}]$，都有 $\mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)] \geqslant \frac{\omega}{3}$。对所有此类 $(h^\star,\ s^\star,\ a^\star)$求和，可得

$$|\mathcal{S}_{\mathsf{leaf}}|\,|\mathcal{A}_c|\left(\left\lfloor\frac{H}{2}\right\rfloor - n\right)\cdot\frac{\omega}{3} \leqslant \sum_{s^\star\in\mathcal{S}_{\mathsf{leaf}}}\sum_{a^\star\in\mathcal{A}_c}\sum_{h^\star=n+1}^{\lfloor H/2\rfloor-1} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)] \leqslant T,$$

其中第二个不等式成立，是因为事件 $\{o_{h^\star}=s^\star,\ a_{h^\star}=a^\star\}$ 互不相交。代入 $|\mathcal{A}_c| = A-1$、$H \geqslant 4n$ 以及（22）中 ω 的定义，即完成命题 E.1 的证明。 □

E.3 命题 E.2的证明

我们首先考虑 $M = M_{\theta,\ \mu}$ 的情况。在步骤 $h < H$，发射矩阵 $\mathbb{O}_{h;\mu}$ 可以写成（对行和列进行某种置换后）

$$\mathbb{O}_{h;\mu} = \begin{bmatrix} \frac{\mathbb{1}_{2K}+\sigma\tilde{\mu}}{2K} & \frac{\mathbb{1}_{2K}}{2K} & 0_{2K\times\mathcal{S}_{\mathsf{tree}}} \\ 0_{\mathcal{S}_{\mathsf{tree}}\times 1} & 0_{\mathcal{S}_{\mathsf{tree}}\times 1} & I_{\mathcal{S}_{\mathsf{tree}}\times\mathcal{S}_{\mathsf{tree}}} \\ 0_{2\times 1} & 0_{2\times 1} & 0_{2\times\mathcal{S}_{\mathsf{tree}}} \end{bmatrix} \in \mathbb{R}^{\mathcal{O}\times\mathcal{S}},$$

其中 $\tilde{\mu} = [\mu;\ -\mu] \in \{-1,\ 1\}^{2K}$，且 $\mathbb{1} = \mathbb{1}_{2K}$ 是 $\mathbb{R}^{2K}$ 中的列向量，其所有元素均为 1。简单计算可得

$$\left[\frac{\mathbb{1}+\sigma\tilde{\mu}}{2K}, \frac{\mathbb{1}}{2K}\right]^{\dagger\top} = \left[\frac{1}{\sigma}\tilde{\mu}, \mathbb{1} - \frac{1}{\sigma}\tilde{\mu}\right],$$

其 1-范数有界于 $\frac{2}{\sigma} + 1$。因此 $\left\|\mathbb{O}_h^{\dagger};\ \mu\right\|_1 \leqslant \frac{2}{\sigma} + 1$。

类似地，对于 $h = H$，$\mathbb{O}_H$ 具有如下形式（对行和列进行某种置换后）

$$\mathbb{O}_H = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} & 0_{1\times\mathcal{S}_{\text{tree}}} \\ \frac{1}{4} & \frac{3}{4} & 0_{1\times\mathcal{S}_{\text{tree}}} \\ 0_{\mathcal{S}_{\text{tree}}\times 1} & 0_{\mathcal{S}_{\text{tree}}\times 1} & I_{\mathcal{S}_{\text{tree}}\times\mathcal{S}_{\text{tree}}} \\ 0_{2K\times 1} & 0_{2K\times 1} & 0_{2K\times\mathcal{S}_{\text{tree}}} \end{bmatrix} \in \mathbb{R}^{\mathcal{O}\times\mathcal{S}}.$$

注意到 $\begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{3}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{3}{2} \end{bmatrix}$，因此 $\left\|\mathbb{O}_H^{\dagger}\right\|_1 \leqslant 2$。

最后，根据定义 1，并注意到 $\mathbb{M}_{h,\ 1} = \mathbb{O}_h$，对所有 $h \in [H]$取伪逆作为广义左逆 $\mathbb{M}_{h\ ,\ 1}^{\dagger} = \mathbb{O}_h^{\dagger}$，可得 $(\alpha_1(M))^{-1} \leqslant \max\left\{\frac{1}{\sigma} + 2,\ 2\right\} = \frac{2}{\sigma} + 1$。

接下来考虑情形 $M = 0$。在此情形下，$s_{\oplus}$ 不可达，因此对于每个步骤 h，我们可以考虑 $\mathbb{O}_h$ 的广义左逆，其定义为

$$\mathbb{O}_h^{+} := [\mathbb{1}(\mathbb{O}_h(o|s) > 0)]_{(s,o)} \in \mathbb{R}^{\mathcal{S}\times\mathcal{O}},$$

按照约定，对于所有 $o \in \mathcal{O}$，$\mathbb{1}(\mathbb{O}_h(o|s_{\oplus}) > 0) = 0$ as $\mathbb{O}_h(\cdot|s_{\oplus})$ 未定义。随后可直接验证，对所有状态 $s \neq s_{\oplus}$，$\mathbb{O}_h^{+}\mathbb{O}_h\mathbf{e}_s = \mathbf{e}_s$（因为根据我们的构造，支撑集 $\text{supp}(\mathbb{O}_h(\cdot|s))$ 彼此不相交）。显然，$\|\mathbb{O}_h^{+}\|_{1\to 1} \leqslant 1$，因此 $(\alpha_1(M))^{-1} \leqslant 1$，证明完毕。 □

E.4 引理 E.3 的证明

根据定义，对于任意模型 $M \in \mathcal{M}$ 和策略 π，

$$\begin{aligned} V_M(\pi) = \mathbb{E}_M^{\pi}[r_H(o_H)] &= \frac{1+\varepsilon}{4}\mathbb{P}_M^{\pi}(o_H = s_0) + \mathbb{P}_M^{\pi}(o_H = \text{good}) \\ &= \frac{1+\varepsilon}{4}\mathbb{P}_M^{\pi}(o_H = s_0) + \frac{3}{4}\mathbb{P}_M^{\pi}(s_H = s_{\oplus}) + \frac{1}{4}\mathbb{P}_M^{\pi}(s_H = s_{\ominus}), \end{aligned}$$

接下来我们证明以下等式，这一等式由我们的构造可得：

$$\begin{aligned} \mathbb{P}_M^{\pi}(o_H = \text{good}) =& \mathbb{P}_M(o_H = \text{good}|s_H = s_{\oplus}) \cdot \mathbb{P}_M^{\pi}(s_H = s_{\oplus}) + \mathbb{P}_M(o_H = \text{good}|s_H = s_{\ominus}) \cdot \mathbb{P}_M^{\pi}(s_H = s_{\ominus}) \\ =& \frac{3}{4}\mathbb{P}_M^{\pi}(s_H = s_{\oplus}) + \frac{1}{4}\mathbb{P}_M^{\pi}(s_H = s_{\ominus}). \end{aligned}$$

接下来，我们分别证明 $M = 0$ 和 $M \neq 0$ 的情形。

情况 1：$M = 0$。在这种情况下，$s_{\oplus}$不可达，因此我们有 $V_0^{\star} = \max_{\pi} V_0(\pi) = \max\left\{\frac{1+\varepsilon}{4},\ \frac{1}{4}\right\} = \frac{1+\varepsilon}{4}$，当停留在 s_0 时可达到。因此，对于任意策略 π，

$$\begin{aligned} V_0^{\star} - V_0(\pi) =& \frac{1+\varepsilon}{4} - \frac{1+\varepsilon}{4}\mathbb{P}_0^{\pi}(o_H = s_0) - \frac{1}{4}\mathbb{P}_0^{\pi}(s_H = s_{\ominus}) \\ =& \frac{1+\varepsilon}{4}\mathbb{P}_0^{\pi}(o_H \neq s_0) - \frac{1}{4}\mathbb{P}_0^{\pi}(s_H = s_{\ominus}) \\ =& \frac{1}{4}(\mathbb{P}_0^{\pi}(o_H \neq s_0) - \mathbb{P}_0^{\pi}(s_H = s_{\ominus})) + \frac{\varepsilon}{4}\mathbb{P}_0^{\pi}(o_H \neq s_0) \\ \geqslant& \frac{\varepsilon}{4}\mathbb{P}_0^{\pi}(o_H \neq s_0). \end{aligned}$$

情况2： $M=(\theta,\ \mu)$ 对于某个 $\theta=(h^\star,\ s^\star,\ a^\star,$ 一个$^\star)$。在这种情况下，只有当 $o_{h^\star}=s^\star$ 和 $a_{h^\star}=a^\star$ 时，$s_\oplus$ 才可达，并且

$$\mathbb{P}_M^\pi(s_H=s_\oplus)=\mathbb{P}_M^\pi(s_H=s_\oplus|o_{h^\star}=s^\star,a_{h^\star}=a^\star)\mathbb{P}_M^\pi(o_{h^\star}=s^\star,a_{h^\star}=a^\star)\leqslant\varepsilon\mathbb{P}_M^\pi(o_{h^\star}=s^\star,a_{h^\star}=a^\star)\leqslant\varepsilon,$$

其中，当 π 是满足 $o_{h^\star}=s^\star,\ a_{h^\star}=a^\star,\ a_{h^\star+1:\ H-1}=\mathbf{a}^\star$ 的任意确定性策略时，等号可以成立。因此，在这种情况下 $V_M^\star=\max_\pi V_M(\pi)=\max\left\{\frac{1+\varepsilon}{4},\ \frac{3\varepsilon}{4}+\frac{1-\varepsilon}{4}\right\}=\frac{1+2\varepsilon}{4}$，且

$$\begin{aligned}V_M^\star-V_M(\pi)=&\frac{1+2\varepsilon}{4}-\frac{1+\varepsilon}{4}\mathbb{P}_M^\pi(o_H=s_0)-\frac{3}{4}\mathbb{P}_M^\pi(s_H=s_\oplus)-\frac{1}{4}\mathbb{P}_M^\pi(s_H=s_\ominus)\\=&\frac{\varepsilon}{4}\mathbb{P}_M^\pi(o_H=s_0)+\frac{1+2\varepsilon}{4}\mathbb{P}_M^\pi(o_H\neq s_0)-\frac{3}{4}\mathbb{P}_M^\pi(s_H=s_\oplus)-\frac{1}{4}\mathbb{P}_M^\pi(s_H=s_\ominus)\\\geqslant&\frac{\varepsilon}{4}\mathbb{P}_M^\pi(o_H=s_0)+\frac{\varepsilon}{2}\mathbb{P}_M^\pi(o_H\neq s_0)-\frac{1}{2}\mathbb{P}_M^\pi(s_H=s_\oplus)\\\geqslant&\frac{\varepsilon}{4}\mathbb{P}_M^\pi(o_H=s_0),\end{aligned}$$

其中，第一个不等式是因为 $\mathbb{P}_M^\pi(s_H=s_\oplus)+\mathbb{P}_M^\pi(s_H=s_\ominus)\leqslant\mathbb{P}_M^\pi(o_H\neq s_0)$ 事件的包含关系。

□

E.5 引理 E.4 的证明

我们先证明引理 E.4 的以下版本：在 $\mathbb{P}_0^{\mathfrak{A}}$ 下，访问次数几乎必然有界；然后通过使用截断论证将问题化归到这一情形，从而证明引理 E.4 。

引理 E.5。假设算法 $\mathfrak{A}$（可能具有随机停止时间 T）满足 $\sum_h N(E^\theta_{\text{rev},\ h})\leqslant N_o$，并且在 $\mathbb{P}_0^{\mathfrak{A}}$ 下几乎必然满足 $N(E^\theta_{\text{correct}})\leqslant N_r$，其中 N_o、N_r 为某些固定值。那么

$$either\ \overline{N}_o\geqslant\frac{\delta^2\sqrt{K}}{4\varepsilon^2\sigma^2},\quad or\ \overline{N}_r\geqslant\frac{\delta^2}{4\varepsilon^2},$$

其中 $\delta=D_{\text{TV}}\left(\mathbb{P}_0^{\mathfrak{A}},\mathbb{E}_{\mu\sim\text{unif}}\left[\mathbb{P}_{\theta,\ \mu}^{\mathfrak{A}}\right]\right)$。

引理 *E.5* 的证明。 根据引理 D.1，我们有

$$1+\chi^2(\mathbb{E}_{\mu\sim\text{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\parallel\mathbb{P}_0^{\mathfrak{A}})=\mathbb{E}_{\mu,\mu'\sim\text{unif}}\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_{\theta,\mu}(\tau^{(t)})\mathbb{P}_{\theta,\mu'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right].$$

为了给上述量一个上界，我们调用以下引理，它是为上述χ^2内积（Canonne，2022，第 3.1 节）在 $\mathbb{P}_{\theta,\ \mu}/\mathbb{P}_0$ 和 $\mathbb{P}_{\theta,\ \mu'}/\mathbb{P}_0$ （证明见附录 E.6）的上界提供关键步骤。

引理 E.6 （χ^2内积的上界）。 在引理 *E.5* 的条件下（固定 θ），对于任意 $\mu,\ \mu'\in\{-1,\ 1\}^K$，

$$\mathbb{E}_0^{\mathfrak{A}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_{\theta,\mu}(\tau^{(t)})\mathbb{P}_{\theta,\mu'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right]\leqslant\exp\left(\overline{N}_o\cdot\frac{C\sigma^2\varepsilon^2}{K}\left|\langle\mu,\mu'\rangle\right|+\frac{4}{3}C\varepsilon^2\overline{N}_r\right).\tag{23}$$

其中 $C:=(1+\sigma)^{2H}\leqslant e$ 作为 $\sigma\leqslant\frac{1}{2H}$。

现在假设引理E.6 成立，并继续证明引理 E.5。对 μ、$\mu'\sim$均匀分布$(\{-1,\ +1\}^K)$取期望，得到 (23)

$$1+\chi^2(\mathbb{E}_{\mu\sim\text{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\parallel\mathbb{P}_0^{\mathfrak{A}})=\mathbb{E}_{\mu,\mu'\sim\text{unif}}\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_{\theta,\mu}(\tau^{(t)})\mathbb{P}_{\theta,\mu'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right]$$

$$\leqslant \mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\left[\exp\left(\overline{N}_o\cdot\frac{C\sigma^2\varepsilon^2}{K}\left|\langle\mu,\mu'\rangle\right|+\frac{4}{3}C\varepsilon^2\overline{N}_r\right)\right].$$

注意到 μ_i、μ_i'是独立同分布的均匀分布($\{\pm1\}$)，因此 $\mu_1\mu_1'$、$\cdots$、$\mu_K\mu_K'$是独立同分布的均匀分布($\{\pm1\}$)。根据 Hoeffding 引理，有 $\mathbb{E}_{\mu、\mu'\sim均匀分布}\left[\text{指数}\left(x\sum_{i=1}^K\mu_i\mu_i'\right)\right]\leqslant\text{指数}\left(Kx^2/2\right)$，对所有 $x\in\mathbb{R}$均成立，因此根据引理 A.3，可得

$$\mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\left[\exp\left(\frac{C\overline{N}_o\sigma^2\varepsilon^2}{K}\left|\langle\mu,\mu'\rangle\right|\right)\right]\leqslant\exp\left(\max\left\{\frac{C^2\sigma^4\varepsilon^4\overline{N}_o^2}{K},\frac{4}{3}\frac{C\sigma^2\varepsilon^2\overline{N}_o}{\sqrt{K}}\right\}\right).$$

因此，结合上述不等式与引理 A.5，我们得到

$$2\delta^2=2D_{\mathrm{TV}}\left(\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right],\mathbb{P}_0^{\mathfrak{A}}\right)^2\leqslant\log\left(1+\chi^2(\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\|\,\mathbb{P}_0^{\mathfrak{A}})\right)\leqslant\max\left\{\frac{4}{3}\frac{\overline{N}_oC\sigma^2\varepsilon^2}{\sqrt{K}},\frac{\overline{N}_o^2C^2\sigma^4\varepsilon^4}{K}\right\}+\frac{4}{3}C\varepsilon^2\overline{N}_r.$$

那么，要么有 $N_r\geqslant\frac{3\delta^2}{4C\varepsilon^2}$，要么成立

$$\max\left\{\frac{4}{3}\frac{\overline{N}_oC\sigma^2\varepsilon^2}{\sqrt{K}},\frac{\overline{N}_o^2C^2\sigma^4\varepsilon^4}{K}\right\}\geqslant\delta^2,$$

这意味着 $\frac{N_oC\sigma^2\varepsilon^2}{\sqrt{K}}\geqslant$最小$\left\{\frac{4}{3},\ \frac{3}{4}\delta^2\right\}=\frac{3}{4}\delta^2$(因为 $\delta\leqslant1$)。利用 $C\leqslant e$即可完成引理 E.5的证明。

□

引理 *E.4* 的证明。 我们采用截断型论证，将引理 E.4 归约为引理 E.5。取 $N_o=\left\lceil6\delta^{-1}\mathbb{E}_0^{\mathfrak{A}}\left[\sum_{h=h^\star}^{H-2}N(E_{\mathsf{rev},h}^\theta)\right]\right\rceil$ 和 $N_r=\left\lceil6\delta^{-1}\mathbb{E}_0^{\mathfrak{A}}N(E_{\mathsf{correct}}^\theta)\right\rceil$。根据马尔可夫不等式，我们有

$$\mathbb{P}_0^{\mathfrak{A}}\left(\sum_{h=h^\star}^{H-2}N(E_{\mathsf{rev},h}^\theta)\geqslant\overline{N}_o\right)\leqslant\frac{\delta}{6},\qquad\mathbb{P}_0^{\mathfrak{A}}\left(N(E_{\mathsf{correct}}^\theta)\geqslant\overline{N}_r\right)\leqslant\frac{\delta}{6}.$$

因此，我们可以为算法 $\mathfrak{A}$考虑以下退出准则：

$$\mathsf{exit}(\tau^{(1:T')})=\mathsf{TRUE}\quad\text{iff}\quad\sum_{t=1}^{T'}\sum_{h=h^\star}^{H-2}\mathbb{I}\left(\tau^{(t)}\in E_{\mathsf{rev},h}^\theta\right)\geqslant\overline{N}_o\ \text{or}\ \sum_{t=1}^{T'}\mathbb{I}\left(\tau^{(t)}\in E_{\mathsf{correct}}^\theta\right)\geqslant\overline{N}_r.$$

退出条件诱导出一个停止时间 T_{exit}，并且我们有

$$\mathbb{P}_0^{\mathfrak{A}}(\exists t<T,\mathsf{exit}(\tau^{(1:t)})=\mathsf{TRUE})\leqslant\mathbb{P}_0^{\mathfrak{A}}\left(\sum_{h=h^\star}^{H-2}N(E_{\mathsf{rev},h}^\theta)\geqslant\overline{N}_o\ \text{or}\ N(E_{\mathsf{correct}}^\theta)\geqslant\overline{N}_r\right)\leqslant\frac{\delta}{6}+\frac{\delta}{6}\leqslant\frac{\delta}{3}.$$

因此，我们可以考虑带有退出准则 $\mathfrak{A}$(退出) 的提前停止算法 退出 (参见附录 D)，根据引理D.2 我们有

$$D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}(\mathsf{exit})},\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}(\mathsf{exit})}\right]\right)\geqslant D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}},\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\right)-\mathbb{P}_0^{\mathfrak{A}}(\exists t<T,\mathsf{exit}(\tau^{(1:t)})=\mathsf{TRUE})\geqslant\frac{2\delta}{3}.$$

注意，根据　我们对退出和停止时间 T_{exit} 的定义，在 $\mathfrak{A}$(退出) 的执行过程中，我们也有

$$\sum_{t=1}^{T_{\mathsf{exit}}-1}\sum_{h=h^\star}^{H-2}\mathbb{I}\left(\tau^{(t)}\in E_{\mathsf{rev},h}^\theta\right)<\overline{N}_o,\qquad\sum_{t=1}^{T_{\mathsf{exit}}-1}\mathbb{I}\left(\tau^{(t)}\in E_{\mathsf{correct}}^\theta\right)<\overline{N}_r.$$

因此，算法 $\mathfrak{A}$(exit) 确保

$$\sum_{h=h^\star}^{H-2}N(E_{\mathsf{rev},h}^\theta)=\sum_{t=1}^{T_{\mathsf{exit}}}\sum_{h=h^\star}^{H-2}\mathbb{I}\left(\tau^{(t)}\in E_{\mathsf{rev},h}^\theta\right)\leqslant\overline{N}_o+H-1,\qquad N(E_{\mathsf{correct}}^\theta)=\sum_{t=1}^{T_{\mathsf{exit}}}\mathbb{I}\left(\tau^{(t)}\in E_{\mathsf{correct}}^\theta\right)\leqslant\overline{N}_r.$$

将引理 E.5 应用于算法 $\mathfrak{A}(\mathsf{exit})$（以及 $\delta' = \frac{2}{3}\delta$），我们可以得到

$$\text{either} \quad \frac{\delta^2\sqrt{K}}{9\varepsilon^2\sigma^2} \leqslant \overline{N}_o + H - 1 \leqslant 6\delta^{-1}\mathbb{E}_0^{\mathfrak{A}}\left[\sum_{h=h^\star}^{H-2} N(E^\theta_{\mathsf{rev},h})\right] + H, \qquad \text{or} \quad \frac{\delta^2}{9\varepsilon^2} \leqslant \overline{N}_r \leqslant 6\delta^{-1}\mathbb{E}_0^{\mathfrak{A}}\big[N(E^\theta_{\mathsf{correct}})\big] + 1,$$

重新排列后即可得到所需结果。 □

引理 E.6的证明

在整个证明过程中，参数 θ, μ, μ' 均保持不变。

根据附录 D中的讨论，利用 (16)，我们有

$$\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}} \frac{\mathbb{P}_M(\tau^{(t)})\mathbb{P}_{M'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2} \cdot \exp\left(-\sum_{t=1}^{\mathsf{T}}\sum_{h=1}^{H} \log I(\tau^{(t)}_{h-1})\right)\right] = 1, \tag{24}$$

其中，对于截至步骤 $l \in [H]$ 的任意部分轨迹 τ_l， $I(\tau_l)$ 定义为

$$I(\tau_l) := \mathbb{E}_0\left[\frac{\mathbb{P}_{\theta,\mu}(o_{l+1}|\tau_l)\mathbb{P}_{\theta,\mu'}(o_{l+1}|\tau_l)}{\mathbb{P}_0(o_{l+1}|\tau_l)^2}\middle|\tau_l\right].$$

注意，模型 $\mathbb{P}_{\theta,\mu}$ 和 $\mathbb{P}_0$ 仅在从 $s_{h^\star} = s^\star, a_{h^\star} = a^\star$ 到 $s_\oplus$ 的转移以及状态 $s_\oplus$处的状态转移动态上有所不同。因此，对于任意（可达的）轨迹 $\tau_l = (o_1, a_1, \cdots, o_l, a_l)$， $\mathbb{P}_{\theta,\mu}(o_{l+1} =|\tau_l) \neq \mathbb{P}_0(o_{l+1} = \cdot|\tau_l)$ 当且仅当 $o_{h^\star} = s^\star, a_{h^\star} = a^\star$。换言之，$I(\tau_l) = 1$ 当 $\tau_l \notin \{o_{h^\star} = s^\star, a_{h^\star} = a^\star\}$时。

接下来我们计算 $I(\tau_l)$ 对于 $\tau_l \in \{o_{h^\star} = s^\star,\ a_{h^\star} = a^\star\}$。根据我们的构造，我们有

$$\begin{aligned}\mathbb{P}_{\theta,\mu}(o_{l+1} = o|\tau_l) &= \mathbb{P}_{\theta,\mu}(o_{l+1} = o|s_{l+1} = s_\oplus)\cdot\mathbb{P}_{\theta,\mu}(s_{l+1} = s_\oplus|\tau_l)\\ &\quad + \mathbb{P}_{\theta,\mu}(o_{l+1} = o|s_{l+1} = s_\ominus)\cdot\mathbb{P}_{\theta,\mu}(s_{l+1} = s_\ominus|\tau_l)\\ &= (\mathbb{O}_{l;\mu}(o|s_\oplus) - \mathbb{O}_l(o|s_\ominus))\cdot\mathbb{P}_{\theta,\mu}(s_{l+1} = s_\oplus|\tau_l) + \mathbb{O}_l(o|s_\ominus).\end{aligned} \tag{25}$$

注意，如果 $\tau_l \notin E_{\mathsf{rev},\,l}$，那么 s_{l+1} 必定是 $s_\ominus$，因此 $\mathbb{P}_{\theta,\,\mu}(o_{l+1} = \cdot|\tau_l) = \mathbb{O}_h(\cdot|s_\ominus) = \mathbb{P}_0(o_{l+1} = \cdot|\tau_l)$这意味着 $I(\tau_l) = 1$。

接下来考虑 $\tau_l \in E_{\mathsf{rev},\,l}$ 的情况，即 $a_{h^\star+1:\,l} = \mathbf{a}^\star_{h^\star+1:\,l}$:

$$\begin{aligned}\mathbb{P}_{\theta,\mu}(s_{l+1} = s_\oplus|\tau_l) &= \mathbb{P}_{\theta,\mu}(s_{l+1} = s_\oplus|o_{h^\star} = s^\star, a_{h^\star} = a^\star, o_{h^\star+1:l}, a_{h^\star+1:l})\\ &= \frac{\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}, s_{l+1} = s_\oplus|o_{h^\star} = s^\star, a_{h^\star} = a^\star, a_{h^\star+1:l})}{\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}|o_{h^\star} = s^\star, a_{h^\star} = a^\star, a_{h^\star+1:l})}\\ &= \frac{\varepsilon\cdot\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}|s_{h^\star+1} = s_\oplus, a_{h^\star+1:l})}{\varepsilon\cdot\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}|s_{h^\star+1} = s_\oplus, a_{h^\star+1:l}) + (1-\varepsilon)\cdot\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}|s_{h^\star+1} = s_\ominus, a_{h^\star+1:l})}\\ &= \frac{\varepsilon}{\varepsilon + (1-\varepsilon)\cdot\frac{\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}|s_{h^\star+1}=s_\ominus, a_{h^\star+1:l})}{\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}|s_{h^\star+1}=s_\oplus, a_{h^\star+1:l})}},\end{aligned}$$

其中第三个等式是因为 $\mathbb{P}_{\theta,\,\mu}(s_{h^\star+1} = s_\oplus|o_{h^\star} = s^\star,\ a_{h^\star} = a^\star) = \varepsilon$。注意，

$$\beta_{\tau_l} := \frac{\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}|s_{h^\star+1} = s_\oplus, a_{h^\star+1:l})}{\mathbb{P}_{\theta,\mu}(o_{h^\star+1:l}|s_{h^\star+1} = s_\ominus, a_{h^\star+1:l})} = \prod_{h=h^\star+1}^{l}\frac{\mathbb{O}_{h;\mu}(o_h|s_\oplus)}{\mathbb{O}_h(o_h|s_\ominus)} \leqslant (1+\sigma)^{l-h^\star},$$

其中，只要 τ_l 可达（即 $o_{h^\star+1:l} \in \mathcal{O}^{l-h^\star}$），该不等式就由我们对 $\mathbb{O}$ 的构造成立。因此，对于

$$c_{\tau_l} := \mathbb{P}_{\theta,\mu}(s_{l+1} = s_\oplus|\tau_l) = \frac{\beta_{\tau_l}}{\varepsilon\beta_{\tau_l} + 1 - \varepsilon},$$

我们有 $c_{\tau_l} \leqslant (1+\sigma)^H = \sqrt{C}$。注意，由 (25) 以及上面的等式可得

$$\text{when } l < H-1, \qquad \mathbb{P}_{\theta,\mu}(o_{l+1} = o_i^+ | \tau_l) = \frac{1 + c_{\tau_l}\varepsilon\sigma\mu_i}{2K}, \qquad \mathbb{P}_{\theta,\mu}(o_{l+1} = o_i^- | \tau_l) = \frac{1 - c_{\tau_l}\varepsilon\sigma\mu_i}{2K} \qquad \forall i \in [K],$$

$$\text{when } l = H-1, \qquad \mathbb{P}_{\theta,\mu}(o_H = \mathsf{good} | \tau_{H-1}) = \frac{1 + 2c_{\tau_{H-1}}\varepsilon}{4}, \qquad \mathbb{P}_{\theta,\mu}(o_H = \mathsf{bad} | \tau_{H-1}) = \frac{3 - 2c_{\tau_{H-1}}\varepsilon}{4}.$$

另一方面，当 $l < H-1$, $\mathbb{P}_{\theta,\ \mu}(o_{l+1} = \cdot | \tau_l) =$ 均匀分布$(\{o_1^+,\ o_1^-,\ \cdots,\ o_K^+,\ o_K^-\})$。因此，

$$\begin{aligned}
I(\tau_l) =& \mathbb{E}_0\left[\frac{\mathbb{P}_{\theta,\mu}(o_{l+1}|\tau_l)\mathbb{P}_{\theta,\mu'}(o_{l+1}|\tau_l)}{\mathbb{P}_0(o_{l+1}|\tau_l)^2}\middle| \tau_l\right] \\
=& \frac{1}{2K}\sum_{o\in\mathcal{O}_o}\frac{\mathbb{P}_{\theta,\mu}(o_{l+1}=o|\tau_l)\mathbb{P}_{\theta,\mu'}(o_{l+1}=o|\tau_l)}{\mathbb{P}_0(o_{l+1}=o|\tau_l)^2} \\
=& \frac{1}{2K}\sum_{i=1}^{K}(1+c_{\tau_l}\varepsilon\sigma\mu_i)(1+c_{\tau_l}\varepsilon\sigma\mu_i') + (1-c_{\tau_l}\varepsilon\sigma\mu_i)(1-c_{\tau_l}\varepsilon\sigma\mu_i') \\
=& 1 + \frac{c_{\tau_l}^2\varepsilon^2\sigma^2}{K}\sum_{i=1}^{K}\mu_i\mu_i' \leqslant 1 + \frac{C\varepsilon^2\sigma^2}{K}\left|\langle\mu,\mu'\rangle\right|.
\end{aligned}$$

类似地，当 $l = H-1$ 时，我们可以计算得出

$$I(\tau_{H-1}) = \mathbb{E}_0\left[\frac{\mathbb{P}_{\theta,\mu}(o_H|\tau_{H-1})\mathbb{P}_{\theta,\mu'}(o_H|\tau_{H-1})}{\mathbb{P}_0(o_H|\tau_{H-1})^2}\middle|\tau_{H-1}\right] = 1 + \frac{4}{3}c_{\tau_{H-1}}^2\varepsilon^2 \leqslant 1 + \frac{4}{3}C\varepsilon^2.$$

因此，结合上述所有事实，我们可以得出结论：

$$\begin{cases}
I(\tau_l) = 1, & l \leqslant h^\star, \\
I(\tau_l) \leqslant 1 + \mathbb{1}\left(\tau_l \in E_{\mathsf{rev},l}^\theta\right)\cdot\frac{C\varepsilon^2\sigma^2}{K}\left|\langle\mu,\mu'\rangle\right|, & h^\star < l < H-1, \\
I(\tau_{H-1}) \leqslant 1 + \mathbb{1}\left(\tau_{H-1} \in E_{\mathsf{correct}}^\theta\right)\cdot\frac{4}{3}C\varepsilon^2, & l = H-1,
\end{cases}$$

其中我们根据定义使用了 $E_{\mathsf{correct}}^\theta = E_{\mathsf{rev},\ H-1}^\theta$ 这一事实。因此，利用 $\log(1+x) \leqslant x$ 这一事实，我们有

$$\begin{aligned}
\sum_{t=1}^{\mathsf{T}}\sum_{l=0}^{H-1}\log I(\tau_l^{(t)}) =& \sum_{t=1}^{\mathsf{T}}\sum_{l=h^\star+1}^{H-1}\log I(\tau_l^{(t)}) \\
\leqslant& \sum_{t=1}^{\mathsf{T}}\sum_{l=h^\star+1}^{H-2}\mathbb{1}\left(\tau_l^{(t)} \in E_{\mathsf{rev},l}^\theta\right)\cdot\frac{C\varepsilon^2\sigma^2}{K}\left|\langle\mu,\mu'\rangle\right| + \mathbb{1}\left(\tau_{H-1}^{(t)} \in E_{\mathsf{correct}}^\theta\right)\cdot\frac{4}{3}C\varepsilon^2 \\
=& \sum_{l=h^\star+1}^{H-2}N\left(E_{\mathsf{rev},l}^\theta\right)\cdot\frac{C\varepsilon^2\sigma^2}{K}\left|\langle\mu,\mu'\rangle\right| + N\left(E_{\mathsf{correct}}^\theta\right)\cdot\frac{4}{3}C\varepsilon^2 \\
\leqslant& \overline{N}_o\cdot\frac{C\varepsilon^2\sigma^2}{K}\left|\langle\mu,\mu'\rangle\right| + \overline{N}_r\cdot\frac{4}{3}C\varepsilon^2.
\end{aligned}$$

将上述不等式代入（24）即可完成引理 E.6 的证明。 □

F 定理 6 的证明

我们首先在附录 F.1 中构造一族困难实例。我们在命题F.1 中给出这族困难实例的遗憾下界。定理 6 随即由命题 F.1 直接推得。命题 F.1 还蕴含了定理 5 中所述 PAC 下界的一部分。

F.1 困难实例的构造与定理 6 的证明

我们考虑以下一族 m步揭示型部分可观测马尔可夫决策过程 $\mathcal{M}$，它们具有一组超参数$(\varepsilon,\ \sigma,\ n,\ m,\ K,\ H)$。$\mathcal{M}$ 中的所有部分可观测马尔可夫决策过程共享状态空间 $\mathcal{S}$、动作空间 $\mathcal{A}$、观测空间 $\mathcal{O}$ 和时域长度 H，其定义如下。

- 状态空间为 $\mathcal{S}=\mathcal{S}_{\text{树}}\bigsqcup\{s_{\oplus},\ s_{\ominus},\ e_{\oplus},\ e_{\ominus},$ 终止状态$\}$，其中 $\mathcal{S}_{\text{树}}$ 是一棵层数为 n 的二叉树（因此$|\mathcal{S}_{\text{树}}|=2^n-1$）。令 s_0 为 $\mathcal{S}_{\text{树}}$的根节点，$\mathcal{S}_{\text{叶节点}}$ 为 $\mathcal{S}_{\text{树}}$的叶节点集合，且$|\mathcal{S}_{\text{叶节点}}|=2^{n-1}$。
- 观测空间为 $\mathcal{O}=\mathcal{S}_{\text{树}}\bigsqcup\{o_1^+,\ o_1^-,\ \cdots,\ o_K^+,\ o_K^-\}\bigsqcup\{$锁，好，坏，终止状态$\}$。[7]
- 动作空间为 $\mathcal{A}=\{0,\ \mathbf{1},\ \cdots,\ A-1\}$。

我们进一步定义 $\mathcal{A}_{\text{rev}}=\{0,\ \mathbf{1},\ \cdots,\ A_1-1\}$，$\mathcal{A}_{\text{tr}}=\{A_1,\ \cdots,\ A-1\}$，其中 $A_1=1+\lfloor A/6\rfloor$。

模型参数 每个非零部分可观测马尔可夫决策过程模型 $M=M_{\theta,\ \mu}\in\mathcal{M}\backslash\{M_0\}$ 由两个参数$(\theta,\ \mu)$指定。这里 $\mu\in\{-1,\ +1\}^K$，且 $\theta=(h^\star,\ s^\star,\ a^\star,\ a^\star_{\text{rev}},\ \mathbf{a}^\star)$，其中

- $s^\star\in\mathcal{S}_{\text{叶}},\ a^\star\in\mathcal{A}_c:=\{1,\ \cdots,\ A-1\},\ a^\star{}_{\text{rev}}\in\mathcal{A}_{\text{rev}}$。
- $h^\star\in\mathcal{H}:=\{h=n+lm:\ h<H,\ l\in\mathbb{Z}_{\geqslant 0}\}$。
- $\mathbf{a}^\star=(\mathbf{a}^\star_{h^\star+1},\ \ldots\ldots,\ \mathbf{a}^\star_{H-1})\in\mathcal{A}^{H-h^\star-1}$是一个由 $h^\star+1,\ \cdots,\ H-1$索引的动作序列，使得当$h\in\mathcal{H}$时，我们有 $\mathbf{a}^\star_h\in\mathcal{A}_{\text{tr}}$。我们用 $\mathcal{A}_{\text{code }h^\star}$ 表示所有此类 $\mathbf{a}^\star$的集合。

我们的构造确保，只有在步骤 $h\in\mathcal{H}$ 和状态 $s_h\in\{s_{\oplus},\ s_{\ominus}\}$时，智能体才能采取$\mathcal{A}_{\text{rev}}$ 中的动作，并转移到 $\{e_{\oplus},\ e_{\ominus}\}$。

对于任意部分可观测马尔可夫决策过程 $M_{\theta,\ \mu}$，其系统动力学 $\mathbb{P}_{\theta,\ \mu}:=\mathbb{P}_{M_{\theta,\ \mu}}$ 定义如下。

发射动态 在状态 $s\in\mathcal{S}_{\text{tree}}\cup\{\text{terminal}\}$，智能体总是收到 s 其自身作为观测（与之对应的唯一观测）。

- 在状态 $e_{\oplus}$，发射动态定义为

$$\mathbb{O}_\mu(o_i^+|e_{\oplus})=\frac{1+\sigma\mu_i}{2K},\qquad \mathbb{O}_\mu(o_i^-|e_{\oplus})=\frac{1-\sigma\mu_i}{2K},\qquad \forall i\in[K],$$

  其中省略下标 h，因为发射分布不依赖于 h .
- 在状态 $e_{\ominus}$，观测从 $\mathcal{O}_o:=\{o_1^+,\ o_1^-,\ \cdots,\ o_K^+,\ o_K^-\}$,中均匀抽取，即 $\mathbb{O}(\cdot|e_{\ominus})$ =均匀分布$(\mathcal{O}_o)$。
- 在状态 $s\in\{s_{\oplus},s_{\ominus}\}$ 和步骤 $h\in[H-1]$，智能体始终接收到锁作为观测；在步骤H，$\{s_{\oplus},s_{\ominus}\}$ 处的发射动态由以下内容给出

$$\mathbb{O}_H(\mathsf{good}|s_{\oplus})=\frac{3}{4},\qquad \mathbb{O}_H(\mathsf{bad}|s_{\oplus})=\frac{1}{4},$$
$$\mathbb{O}_H(\mathsf{good}|s_{\ominus})=\frac{1}{4},\qquad \mathbb{O}_H(\mathsf{bad}|s_{\ominus})=\frac{3}{4}.$$

转移动态 在每个回合中，智能体始终从状态 s_0 开始。

- 在 $s\in\mathcal{S}_{\text{树}}\backslash\mathcal{S}_{\text{叶节点}}$的任意节点上，有三种类型的可用动作：等待 $=0$、向左 $=1$ 和向右 $=2$，智能体可以选择等待以停留在 s，选择向左转移到 s 的左子节点，或选择向右转移到 s的右子节点。

[7]与附录 E类似，此处我们稍微滥用记号，将 $\mathcal{S}_{\text{tree}}$ 同时用于表示状态集合及其对应的观测集合，即每个状态 $s\in\mathcal{S}_{\text{tree}}\subset\mathcal{S}$ 都对应一个唯一的观测 $o_s\in\mathcal{S}_{\text{tree}}\subset\mathcal{O}$，当上下文明确时，我们也将其记为 s 。

- 在任意 $s \in \mathcal{S}_{\text{叶节点}}$上，智能体可以执行等待动作 $= 0$ 以停留在 s（即 $\mathbb{P}(s|s, \text{等待}) = 1$）；否则，对于 $s \in \mathcal{S}_{\text{叶节点}}$， $h \in [H-1]$， $a \neq$ 等待，

$$\mathbb{P}_{h;\theta}(s_{\oplus}|s,a) = \varepsilon \cdot \mathbb{1}(h = h^\star, s = s^\star, a = a^\star),$$
$$\mathbb{P}_{h;\theta}(s_{\ominus}|s,a) = 1 - \varepsilon \cdot \mathbb{1}(h = h^\star, s = s^\star, a = a^\star).$$

  其中我们使用下标 θ 来强调对 θ的依赖。换言之，在步骤 $h^\star$，在任意叶节点采取任意动作后，智能体都会转移到 $\{s_{\oplus}, s_{\ominus}\}$之一；只有在 $s^\star$采取$a^\star$ 时，智能体才会以较小的概率 ε转移到状态 $s_{\oplus}$；在其他任何情况下，智能体都会以概率 **1** 转移到状态 $s_{\ominus}$。

- 无论采取什么动作，状态 $s \in \{e_{\oplus}, e_{\ominus}\}$ 都会转移到终止状态。
- 终止状态是一个吸收状态。
- 在状态 $s_{\ominus}$:
  – 对于步骤 $h \in \mathcal{H}$ 和 $a \in \mathcal{A}_{\mathsf{rev}}$，我们设定 $\mathbb{P}_{h;\theta}(e_{\ominus}|s_{\ominus}, a) = 1$，即采取 $a \in \mathcal{A}_{\mathsf{rev}}$ 总会转移到 $e_{\ominus}$。– 对于步骤 $h \notin \mathcal{H}$ 或 $a \in \mathcal{A}_{\mathsf{tr}}$，我们设定 $\mathbb{P}_{h;\theta}(s_{\ominus}|s_{\ominus}, a) = 1$，即采取此类动作总会停留在 $s_{\ominus}$。
- 在状态 $s_{\oplus}$，我们只需指定步骤 $h \geqslant h^\star + 1$的转移动态:
  – 对于步骤 $h \in \mathcal{H}_{>h^\star} = \mathcal{H} \cap \{h > h^\star\}$ 和 $a \in \mathcal{A}_{\mathsf{rev}}$，我们设定

$$\mathbb{P}_{h;\theta}(e_{\oplus}|s_{\oplus}, a) = \mathbb{1}(a = a^\star_{\mathsf{rev}}), \qquad \mathbb{P}_{h;\theta}(e_{\ominus}|s_{\oplus}, a) = \mathbb{1}(a \neq a^\star_{\mathsf{rev}}).$$

  换言之，在步骤 $h \in \mathcal{H}_{>h^\star}$ 和状态 $s_h \in \{s_{\oplus}, s_{\ominus}\}$（对应于 $o_h =$ 锁）下，智能体可以采取动作 $a_h \in \mathcal{A}_{\mathsf{rev}}$ 转移到 $\{e_{\oplus}, e_{\ominus}\}$；但只有在 $s_{\oplus}$ 采取 $a_h = a^\star_{\mathsf{rev}}$ “正确地”行动时，智能体才能转移到 $e_{\oplus}$；在其他任何情况下，智能体都将以概率 **1** 转移到状态 $e_{\ominus}$。注意， $\mathcal{H} = \{h = n + lm : \ h < H, \ l \in \mathbb{Z}_{\geqslant 0}\}$，因此我们只允许智能体每隔 m 步骤采取一次揭示动作$a^\star_{\mathsf{rev}}$，这确保了我们的构造是 $(m+1)$-步骤揭示的。

  – 对于步骤 $h \notin \mathcal{H}$ 或 $a \in \mathcal{A}_{\mathsf{tr}}$，我们设定

$$\mathbb{P}_{h;\theta}(s_{\oplus}|s_{\oplus}, a) = \mathbb{1}(a = \mathbf{a}^\star_h), \qquad \mathbb{P}_{h;\theta}(s_{\ominus}|s_{\oplus}, a) = \mathbb{1}(a \neq \mathbf{a}^\star_h).$$

奖励 奖励函数是已知的（且仅取决于观测）：在最初的 $H-1$ 个步骤中，不提供奖励；在步骤 H，我们设定 $r_H(\text{好}) = 1$、 $r_H(\text{坏}) = 0$，以及对任何其他 $o \in \mathcal{O}$设定 $r_H(s_0) = (1+\varepsilon)/4$, $r_H(o) = 0$。

参考模型 我们用 M_0 （或简写为 **0**）表示零模型（参考模型）。零模型M_0 的转移和发射与任何非零模型相同，但智能体在 $s \in \mathcal{S}_{\mathsf{leaf}}$处采取任意动作 $a \neq$ 等待时，总会到达 $s_{\ominus}$ $h \in [H-1]$（即 $\mathbb{P}_{h;M_0}(s_{\ominus}|s, \ a) = 1$ 对于任意 $s \in \mathcal{S}_{\mathsf{leaf}}$、 $a \in \mathcal{A}_c$、$h \in [H-1]$）。在此模型中， $s_{\oplus}$ 不可达（因此 $e_{\oplus}$也不可达），所以我们无需指定 $s_{\oplus}$， $e_{\oplus}$处的转移和发射动态。

我们在以下命题中给出上述部分可观测马尔可夫决策过程模型类的期望遗憾下界和 PAC 学习样本复杂度下界 $\mathcal{M}$，并在附录F.2中证明该命题。

命题 F.1。 对于给定的 $\varepsilon \in (0、\ 0.1]$、 $\sigma \in (0、\ 1]$、 m、 $n \geqslant 1$、 $K \geqslant 2$、 $H \geqslant 8n + m + 1$，上述模型类$\mathcal{M}$ 满足以下性质。

1. $|\mathcal{S}| = 2^n + 4$, $|\mathcal{O}| = 2^n + 2K + 3$, $|\mathcal{A}| = A$。
2. 对于每个 $M \in \mathcal{M}$， M 都是具有 $\alpha_{m+1}(M)^{-1} \leqslant 1 + \frac{2}{\sigma}$ 的 $(m+1)$ 步揭示过程。
3. $\log|\mathcal{M}| \leqslant K \log 2 + H \log A + \log(SAH)$。

4. 假设算法 $\mathfrak{A}$ 与环境交互 T 回合，那么

$$\max_{M\in\mathcal{M}} \mathbb{E}_M^{\mathfrak{A}}[\mathbf{Regret}] \geqslant \frac{1}{120000}\min\left\{\frac{|\mathcal{S}_{\mathsf{leaf}}|\,K^{1/2}A^{m+1}H}{m\sigma^2\varepsilon^2}, \frac{|\mathcal{S}_{\mathsf{leaf}}|\,A^{H/2}H}{m\varepsilon}, \varepsilon T\right\},$$

其中我们回顾一下 $|\mathcal{S}_{\mathsf{leaf}}| = 2^{n-1}$。

5. 假设 算法 $\mathfrak{A}$ 与环境交互 T 个回合，并返回 π^{out}，使得 at

$$\mathbb{P}_M^{\mathfrak{A}}\left(V_M^\star - V_M(\pi^{\mathrm{out}}) < \frac{\varepsilon}{8}\right) \geqslant \frac{3}{4}.$$

对于任意 $M\in\mathcal{M}$，都必须满足

$$T \geqslant \frac{1}{60000}\min\left\{\frac{|\mathcal{S}_{\mathsf{leaf}}|\,K^{1/2}A^{m+1}H}{\sigma^2\varepsilon^2}, \frac{|\mathcal{S}_{\mathsf{leaf}}|\,A^{H/2}H}{\varepsilon^2}\right\}.$$

定理 6 的证明 我们只需在应用命题 F.1 时适当地选择参数。更具体地，给定 $(S、O、A、H、\alpha、m)$，我们可以令 $m' = m-1$，并将 $n\geqslant 1$ 取为满足 $2^n \leqslant$ 的最大整数，其中取 $\{S-4、(O-5)/2\}$, 中的最小值；再取 $K = \left\lfloor \frac{O-2^n-3}{2}\right\rfloor \geqslant \frac{O-5}{4}$、$\varepsilon' = \varepsilon/8$, 和 $\overline{\sigma = \frac{2}{\alpha^{-1}-1}} \leqslant \overline{1。}$ 对于任意固定的 $\varepsilon\in\overline{(0、\ 0.1]}$，将参数 $(\varepsilon、\sigma、n、m'、K、H)$ 应用于命题 F.1，可得到一个模型类 $\mathcal{M}_\varepsilon$，使得对于任意算法 $\mathfrak{A}$,

$$\max_{M\in\mathcal{M}_\varepsilon} \mathbb{E}_M^{\mathfrak{A}}[\mathbf{Regret}] \geqslant c_0\min\left\{\frac{SO^{1/2}A^mH}{m\alpha^2\varepsilon^2}, \frac{SA^{H/2}H}{m\varepsilon}, \varepsilon T\right\},$$

其中 c_0 是一个普适常数。接下来，我们可以取使上述不等式右侧达到最大值的 $\varepsilon\in(0,\ 0.1]$ 并应用引理 A.1 即可完成定理 6的证明。 □

注 F.2. 定理 6 （以及定理 5）中的要求 $S\leqslant O$ 实际上可以放宽为 $S\leqslant O^m$。当前构造中之所以要求 $S\leqslant O$，是因为我们直接将 $\mathcal{S}_{\mathsf{tree}}$ 嵌入观测空间 $\mathcal{O}$，也就是说，对于每个状态 $s\in\mathcal{S}_{\mathsf{tree}}$，它都会发出相应的 $o_s\in\mathcal{O}$。然而，当$O^m \geqslant |\mathcal{S}_{\mathsf{tree}}| \gg O$时，我们也可以采用嵌入 $\mathcal{S}_{\mathsf{tree}}\to\mathcal{O}^m$；也就是说，对于每个满足 $s\mapsto(o_s^{(1)},\ \cdots,\ o_s^{(m)})$的状态 $s\in\mathcal{S}_{\mathsf{tree}}$，它会在步骤 h发出 $o^{(h\ 模\ s\qquad m)}\in\mathcal{O}$。

F.2 命题 F.1 的证明

本节陈述的所有命题和引理均在附录 F.3-F.7中证明。

断言 1 可通过计算 $\mathcal{M}$ 中模型的状态、观测和动作数量直接得出。由于 $|\mathcal{M}| = |\{(h^\star,\ s^\star,\ a^\star,\ a_{\mathsf{rev}}^\star,\ \mathbf{a}^\star)\}| \times \left|\{\pm1\}^K\right| + 1 \leqslant HSA^H\times 2^K$，因此可得断言 3。取对数即可得到该断言。

断言 2 可由该引理推出，其证明见附录 F.3。

引理 F.3。 对于每个 $M\in\mathcal{M}$，都有 $\alpha_{m+1}(M)^{-1} \leqslant \frac{2}{\sigma}+1$。 –

现在我们证明断言4和5。与命题 E.1的证明类似，我们先将学习问题与检验问题联系起来。回顾一下，$\mathbb{P}_M^{\mathfrak{A}}$是由算法 $(\tau^{(1)}、\tau^{(2)}、\cdots、\tau^{(\mathsf{T})})$和模型 $\mathfrak{A}$诱导的分布律 M。对于任意事件 $E\subseteq(\mathcal{O}\times\mathcal{A})^H$，我们将 E的访问次数记为

$$N(E) := \sum_{t=1}^{T}\mathbb{1}(\tau^{(t)}\in E).$$

由于 $N(E)$是 $\tau^{(1_{的函数:}\ T)}$，我们可以讨论它在分布 $\mathbb{P}_M^{\mathfrak{A}}$下对任意$M\in\mathcal{M}$的期望。我们首先将期望遗憾与某些“坏”事件的期望访问次数联系起来，得到以下引理，其证明见附录 F.4。

引理 F. 4（将遗憾与访问次数联系起来）。对于任意 $M \in \mathcal{M}$ 使得 $M \neq 0$，有 *at*

$$\mathbb{E}_M^{\mathfrak{A}}[\mathbf{Regret}] \geqslant \frac{\varepsilon}{4}\mathbb{E}_M^{\mathfrak{A}}[N(o_H = s_0)]. \tag{26}$$

另一方面，对于参考模型 0，我们有

$$\mathbb{E}_0^{\mathfrak{A}}[\mathbf{Regret}] \geqslant \frac{\varepsilon}{4}\mathbb{E}_0^{\mathfrak{A}}[N(o_H \neq s_0)] + \frac{1}{4}\mathbb{E}_0^{\mathfrak{A}}[N(E_{\mathsf{rev}})]. \tag{27}$$

其中我们定义 $E_{\mathsf{rev}} := \{\tau : \text{对某些 } h \in \mathcal{H}, o_h = \text{锁 } a_h \in \mathcal{A}_{\mathsf{rev}}\}$。

另一方面，对于任意策略 π，我们有

$$V_M^\star - V_M(\pi) \geqslant \frac{\varepsilon}{4}\mathbb{P}_M^\pi(o_H = s_0) \ \ \forall M \neq 0, \qquad \textit{and} \qquad V_0^\star - V_0(\pi) \geqslant \frac{\varepsilon}{4}\mathbb{P}_0^\pi(o_H \neq s_0). \tag{28}$$

因此，我们可以通过类似于附录 $\mu \sim$均匀分布$(\{-1, +1\}^K)$ 参数 μ的先验分布的论证，将遗憾（或输出策略的次优性）与全变差距离（在 E.2 中）联系起来，从而得到以下引理，其证明见附录 F.5。

引理 F.5. 假设以下陈述之一对算法 $\mathfrak{A}$ 成立：

(*a*) 对于任意模型 $M \in \mathcal{M}$，$\mathbb{E}_M^{\mathfrak{A}}[\textbf{遗憾}] \leqslant T\varepsilon/32$。
(*b*) 对于任意模型 $M \in \mathcal{M}$，算法 $\mathfrak{A}$ 输出一个策略 π^{out}，使得 $\mathbb{P}_M^{\mathfrak{A}}\left(V_M^\star - V_M(\pi^{\mathrm{out}}) < \frac{\varepsilon}{8}\right) \geqslant \frac{3}{4}$。

于是我们有

$$D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\right) \geqslant \frac{1}{2}, \qquad \forall\theta. \tag{29}$$

根据断言 4（或 5）中的假设，以下我们只需考虑（29）对所有 θ 都成立的情况。我们将利用（29）推导 $\mathbb{E}_0^{\mathfrak{A}}[N(o_H \neq s_0)]$ 和 $\mathbb{E}_0^{\mathfrak{A}}[N(E_{\mathsf{rev}})]$的下界，从而得到以下引理，其证明见附录 F.6。

引理 F.6。固定一个 $\theta = (h^\star, s^\star, a^\star, a_{\mathsf{rev}}^\star, \mathbf{a}^\star)$。我们考虑事件

$$E_{\mathsf{rev}}^\theta := \left\{o_{h^\star} = s^\star, a_{h^\star:h} = (a^\star, \mathbf{a}_{h^\star+1:h-1}^\star, a_{\mathsf{rev}}^\star) \textit{ for some } h \in \mathcal{H}_{>h^\star}\right\},$$
$$E_{\mathsf{correct}}^\theta := \{o_{h^\star} = s^\star, a_{h^\star:H-1} = (a^\star, \mathbf{a}^\star)\}.$$

那么对于任意满足 δ 的算法 $\mathfrak{A}$：$= D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\right) > 0$，我们有

$$\textit{either } \mathbb{E}_0^{\mathfrak{A}}\left[N(E_{\mathsf{rev}}^\theta)\right] \geqslant \frac{\delta^3\sqrt{K}}{18\varepsilon^2\sigma^2} - \frac{\delta}{6}, \ \textit{or } \mathbb{E}_0^{\mathfrak{A}}\left[N(E_{\mathsf{correct}}^\theta)\right] \geqslant \frac{\delta^3}{18\varepsilon^2} - \frac{\delta}{6}.$$

应用 引理 F.6 对于任意参数元组 $\theta = (h^\star, s^\star, a^\star, a_{\mathsf{rev}}^\star, \text{一个}^\star)$ 满足 $\delta = \frac{1}{2}$，我们得到

$$\text{either } \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{rev}}^{(h^\star,s^\star,a^\star,a_{\mathsf{rev}}^\star,\mathbf{a}^\star)}\right)\right] \geqslant \frac{\sqrt{K}}{300\varepsilon^2\sigma^2}, \quad \text{or} \quad \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,a_{\mathsf{rev}}^\star,\mathbf{a}^\star)}\right)\right] \geqslant \frac{1}{300\varepsilon^2}, \tag{30}$$

当我们选择 $\varepsilon \in (0, 0.1]$ 时。

固定一个满足 $h^\star \in \mathcal{H}$ 和 $h^\star \leqslant n + m\lfloor H/10m\rfloor$, $s^\star \in \mathcal{S}_{\mathsf{leaf}}$, $a^\star \in \mathcal{A}_c$ 的元组 $(h^\star, s^\star, a^\star)$。根据（30），我们知道，对于所有 $\mathbf{a}^\star \in \mathcal{A}_{\mathsf{code},\ h^\star}$、$a_{\mathsf{rev}}^\star \in \mathcal{A}_{\mathsf{rev}}$、$\theta = (h^\star, s^\star, a^\star, a_{\mathsf{rev}}^\star, \mathbf{a}^\star)$ 以及实常数 $r \geqslant 0$，都有

$$\begin{aligned}
&|\mathcal{A}_{\mathsf{rev}}|\, A^{m-1} \cdot \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{rev}}^{(h^\star,s^\star,a^\star,a_{\mathsf{rev}}^\star,\mathbf{a}^\star)}\right)\right] + r\,|\mathcal{A}_{\mathsf{code},h^\star}| \cdot \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,a_{\mathsf{rev}}^\star,\mathbf{a}^\star)}\right)\right] \\
&\geqslant \frac{1}{300}\min\left\{\frac{|\mathcal{A}_{\mathsf{rev}}|\, A^{m-1}\sqrt{K}}{\varepsilon^2\sigma^2}, \frac{r\,|\mathcal{A}_{\mathsf{code},h^\star}|}{\varepsilon^2}\right\} \geqslant \frac{1}{300}\min\left\{\frac{|\mathcal{A}_{\mathsf{rev}}|\, A^{m-1}\sqrt{K}}{\varepsilon^2\sigma^2}, \frac{rA^{H/2-1}}{\varepsilon^2}\right\} =: \omega_r,
\end{aligned}$$

其中，最后一个不等式可通过直接计算得出（参见引理 F.7）。注意 th at

$$
\begin{aligned}
&\sum_{a^\star_{\mathsf{rev}}\in\mathcal{A}_{\mathsf{rev}},\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}\Big[N\Big(E_{\mathsf{rev}}^{(h^\star,s^\star,a^\star,a^\star_{\mathsf{rev}},\mathbf{a}^\star)}\Big)\Big] \\
&= \sum_{a^\star_{\mathsf{rev}}\in\mathcal{A}_{\mathsf{rev}},\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}\big[N\big(o_{h^\star}=s^\star, a_{h^\star:h}=(a^\star,\mathbf{a}^\star_{h^\star+1:h-1},a^\star_{\mathsf{rev}}) \text{ for some } h\in\mathcal{H}_{>h^\star}\big)\big] \\
&\leqslant \sum_{a^\star_{\mathsf{rev}}\in\mathcal{A}_{\mathsf{rev}},\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}\big[N\big(o_{h^\star}=s^\star, a_{h^\star:h^\star+m-1}=(a^\star,\mathbf{a}^\star_{h^\star+1:h^\star+m-1}), a_h=a^\star_{\mathsf{rev}} \text{ for some } h\in\mathcal{H}_{>h^\star}\big)\big] \\
&= \sum_{a^\star_{\mathsf{rev}}\in\mathcal{A}_{\mathsf{rev}},\mathbf{a}\in\mathcal{A}^{m-1}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star:h^\star+m-1}=(a^\star,\mathbf{a}), a_h=a^\star_{\mathsf{rev}} \text{ for some } h\in\mathcal{H}_{>h^\star})]\cdot \sum_{\substack{\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}\\ \mathbf{a}^\star \text{ begins with } \mathbf{a}}} 1 \\
&= \sum_{a^\star_{\mathsf{rev}}\in\mathcal{A}_{\mathsf{rev}},\mathbf{a}\in\mathcal{A}^{m-1}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star:h^\star+m-1}=(a^\star,\mathbf{a}), a_h=a^\star_{\mathsf{rev}} \text{ for some } h\in\mathcal{H}_{>h^\star})]\cdot \frac{|\mathcal{A}_{\mathsf{code},h^\star}|}{A^{m-1}} \\
&= \frac{|\mathcal{A}_{\mathsf{code},h^\star}|}{A^{m-1}}\cdot \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star, a_h\in\mathcal{A}_{\mathsf{rev}} \text{ for some } h\in\mathcal{H}_{>h^\star})],
\end{aligned}
$$

其中，第二行是由于事件的包含关系，第四行根据我们对 $\mathcal{A}_{\mathsf{code},h^\star}$ 的定义得出，最后一行则是因为事件 $\{o_{h^\star}=s^\star, a_{h^\star:h^\star+m-1}=(a^\star,\mathbf{a}), a_h=a^\star_{\mathsf{rev}}$（对某个 $h\in\mathcal{H}_{>h^\star}\}$）两两不相交，且它们的并集恰为 $\{o_{h^\star}=s^\star, a_{h^\star}=a^\star, a_h\in\mathcal{A}_{\mathsf{rev}}$（对某个 $h\in\mathcal{H}_{>h^\star}\}$）。类似地，我们有

$$
\begin{aligned}
\sum_{a^\star_{\mathsf{rev}}\in\mathcal{A}_{\mathsf{rev}},\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}\Big[N\Big(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,a^\star_{\mathsf{rev}},\mathbf{a}^\star)}\Big)\Big] &= \sum_{a^\star_{\mathsf{rev}}\in\mathcal{A}_{\mathsf{rev}},\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star:H-1}=(a^\star,\mathbf{a}^\star))] \\
&= |\mathcal{A}_{\mathsf{rev}}|\cdot \sum_{\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star:H-1}=(a^\star,\mathbf{a}^\star))] \\
&= |\mathcal{A}_{\mathsf{rev}}|\cdot \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star, a_{h^\star+1:H-1}\in\mathcal{A}_{\mathsf{code},h^\star})].
\end{aligned}
$$

综合以上所有事实，我们得到

$$
\begin{aligned}
\omega_r &\leqslant \frac{1}{|\mathcal{A}_{\mathsf{rev}}|\,|\mathcal{A}_{\mathsf{code},h^\star}|}\sum_{a^\star_{\mathsf{rev}}\in\mathcal{A}_{\mathsf{rev}}}\sum_{\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star}}\Big(|\mathcal{A}_{\mathsf{rev}}|\,A^{m-1}\cdot\mathbb{E}_0^{\mathfrak{A}}\Big[N\Big(E_{\mathsf{rev}}^{(h^\star,s^\star,a^\star,a^\star_{\mathsf{rev}},\mathbf{a}^\star)}\Big)\Big]+r\,|\mathcal{A}_{\mathsf{code},h^\star}|\cdot\mathbb{E}_0^{\mathfrak{A}}\Big[N\Big(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,a^\star_{\mathsf{rev}},\mathbf{a}^\star)}\Big)\Big]\Big) \\
&\leqslant \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star, a_h\in\mathcal{A}_{\mathsf{rev}} \text{ for some } h\in\mathcal{H}_{>h^\star})]+r\mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)]
\end{aligned}
$$

注意，上述不等式对任意给定的 $s^\star\in\mathcal{S}_{\mathsf{leaf}}, a^\star\in\mathcal{A}_c, h^\star\in\mathcal{H}$（满足 $h^\star\leqslant n+m\lfloor H/10m\rfloor$,）以及任意 $r\geqslant 0$ 都成立。因此，我们可以对所有满足 $0\leqslant l\leqslant\lfloor H/10m\rfloor$, 的 $s^\star\in\mathcal{S}_{\mathsf{leaf}}, a^\star\in\mathcal{A}_c, h^\star=n+lm\in\mathcal{H}$ 求和，得到

$$
\begin{aligned}
&|\mathcal{S}_{\mathsf{leaf}}|\,|\mathcal{A}_c|\,(\lfloor H/10m\rfloor+1)\cdot\min\left\{\frac{|\mathcal{A}_{\mathsf{rev}}|\,A^{m-1}\sqrt{K}}{300\varepsilon^2\sigma^2},\frac{rA^{H/2-1}}{300\varepsilon^2}\right\}=\sum_{s^\star\in\mathcal{S}_{\mathsf{leaf}}}\sum_{a^\star\in\mathcal{A}_c}\sum_{\substack{h^\star=n+lm:\\ 0\leqslant l\leqslant\lfloor H/10m\rfloor}}\omega \\
&\leqslant \sum_{s^\star\in\mathcal{S}_{\mathsf{leaf}}}\sum_{a^\star\in\mathcal{A}_c}\sum_{\substack{h^\star=n+lm:\\ 0\leqslant l\leqslant\lfloor H/10m\rfloor}}\mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star, a_h\in\mathcal{A}_{\mathsf{rev}} \text{ for some } h\in\mathcal{H}_{>h^\star})]+r\mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)] \\
&\leqslant \mathbb{E}_0^{\mathfrak{A}}[N(E_{\mathsf{rev}})]+r\mathbb{E}_0^{\mathfrak{A}}[N(o_H\neq s_0)],
\end{aligned}
$$

其中，最后一个不等式是因为

$$
\bigsqcup_{s^\star\in\mathcal{S}_{\mathsf{leaf}},a^\star\in\mathcal{A}_c,h^\star\in\mathcal{H}}\{o_{h^\star}=s^\star,\ a_{h^\star}=a^\star,\ a_h\in\mathcal{A}_{\mathsf{rev}}\text{对于某个}h\in\mathcal{H}_{>h^\star}\}\subseteq\{\text{对于某个 } h\in\mathcal{H},\ o_h=\text{锁 } a_h\in\mathcal{A}_{\mathsf{rev}}\}=E_{\mathsf{rev}},
$$

以及 $\bigsqcup_{s^\star\in\mathcal{S}_{\mathsf{leaf}},\ a^\star\in\mathcal{A}_c,\ h^\star\in\mathcal{H}}\{o_{h^\star}=s^\star,\ a_{h^\star}=a^\star\}\subseteq\{o_H\neq s_0\}$。代入我们选择的 $|\mathcal{A}_c|=A-1\geqslant\frac{2}{3}A$, $|\mathcal{A}_{\mathsf{rev}}|=1+\lfloor A/6\rfloor\geqslant A/6$ $\lfloor H/10m\rfloor+1\geqslant H/10m$, ，由此得出以下结论的证明： –

结论： 只要满足（29），就有

$$\mathbb{E}_0^{\mathfrak{A}}[N(E_{\mathsf{rev}})] + r\mathbb{E}_0^{\mathfrak{A}}[N(o_H \neq s_0)] \geqslant \frac{|\mathcal{S}_{\mathsf{leaf}}|\, H}{30000m} \cdot \min\left\{\frac{A^{m+1}\sqrt{K}}{\varepsilon^2\sigma^2}, \frac{rA^{H/2}}{\varepsilon^2}\right\}, \qquad \forall r \geqslant 0. \tag{31}$$

根据上述事实推导断言 4 时，我们注意到，要么 (1) $\mathbb{E}_M^{\mathfrak{A}}[$**遗憾**$] > T\varepsilon/32$ 对于某个 $M \in \mathcal{M}$成立，要么 (2) $\mathbb{E}_M^{\mathfrak{A}}[$**遗憾**$] \leqslant T\varepsilon/32$ 对于任意 $M \in \mathcal{M}$成立；接着根据引理 F.5，式 (29) 成立，因此我们有

$$\mathbb{E}_0^{\mathfrak{A}}[\mathbf{Regret}] \geqslant \frac{1}{4}\mathbb{E}_0^{\mathfrak{A}}[N(E_{\mathsf{rev}})] + \frac{\varepsilon}{4}\mathbb{E}_0^{\mathfrak{A}}[N(o_H \neq s_0)] \geqslant \frac{|\mathcal{S}_{\mathsf{leaf}}|\, H}{120000m} \cdot \min\left\{\frac{A^{m+1}\sqrt{K}}{\varepsilon^2\sigma^2}, \frac{A^{H/2}}{\varepsilon}\right\}$$

在式 ($r = \varepsilon$ 中取值 31)。结合这两种情况，我们完成了命题 F.1中断言 4 的证明。

类似地，假设断言 5 中的条件成立，由此可知式 (29) 成立（根据引理 F.5）。然后我们可以在式 ($r = 1$ 31) 中取值，以得到

$$2T \geqslant \mathbb{E}_0^{\mathfrak{A}}[N(E_{\mathsf{rev}})] + \mathbb{E}_0^{\mathfrak{A}}[N(o_H \neq s_0)] \geqslant \frac{|\mathcal{S}_{\mathsf{leaf}}|\, H}{30000m} \cdot \min\left\{\frac{A^{m+1}\sqrt{K}}{\varepsilon^2\sigma^2}, \frac{A^{H/2}}{\varepsilon^2}\right\},$$

由此完成断言 4 的证明。附录 F.2 的证明至此结束。 □

引理 F.7。 只要 $|\mathcal{A}_{\mathsf{rev}}| = A_1 \leqslant 1 + \lfloor A/6 \rfloor$, 我们就有 $|\mathcal{A}_{\mathsf{code},\ h^\star}| \geqslant A^{H/2-1}$ 对于 $h^\star \in \mathcal{H}$ 满足 $h^\star \leqslant n + m\lfloor H/10m \rfloor$ 的情况。

证明。 我们记 $H_0 = \lfloor (H-n)/m \rfloor,$ ，并假设 $h^\star = n + ml$。回顾可知

$$\mathcal{A}_{\mathsf{code},h^\star} := \left\{\mathbf{a}^\star = (\mathbf{a}^\star_{h^\star+1}, \ldots, \mathbf{a}^\star_{H-1}) \in \mathcal{A}^{H-h^\star-1} : \mathbf{a}^\star_h \in \mathcal{A}_{\mathsf{tr}}, \forall h \in \mathcal{H}_{>h^\star}\right\}.$$

因此，注意到 $|\mathcal{H}_{>h^\star}| = H_0 - l$、 $|\mathcal{A}| = A$、 $|\mathcal{A}_{\mathsf{tr}}| = A - A_1$，我们有

$$|\mathcal{A}_{\mathsf{code},h^\star}| = A^{H-h^\star-1-(H_0-l)} \times (A - A_1)^{H_0-l}.$$

因此，我们只需证明

$$H - h^\star - 1 - (H_0 - l) + \frac{\log(A - A_1)}{\log A}(H_0 - l) \geqslant \frac{H}{2} - 1. \tag{32}$$

注意，只要 $A_1 \leqslant 1 + \lfloor A/6 \rfloor,$成立，就有 $\frac{\log(A-A_1)}{\log A} \geqslant \frac{\log 2}{\log 3} =: w$。利用这一事实并重新整理，我们可以看到，(32)在以下条件下成立

$$l \leqslant \frac{\frac{H}{2} - n - (1-w)H_0}{m - 1 + w} =: l_0.$$

现在，利用我们的假设 $H \geqslant 10n$，我们有

$$l_0 \geqslant \frac{\frac{H}{2} - n - (1-w)(H-n)}{mw} = \frac{(w-0.5)H - wn}{mw} \geqslant \frac{(w-0.5)H - 0.1wH}{mw} \geqslant \frac{H}{10m},$$

其中，最后一个不等式利用了 $w > \frac{5}{8}$。因此，只要 $l \leqslant \lfloor H/10m \rfloor$（即 $h^\star \leqslant n + m\lfloor H/10m \rfloor$），我们就有 $l \leqslant l_0$，这意味着(32)，从而完成证明。 □

F.3 引理F.3

这里的思路与命题 E.2 的证明类似，但由于我们的构造更为复杂，直接描述 $\mathbb{M}_{h,m+1}$ 可能会非常繁琐（尽管实际上其中只有少数几项非零）。因此，为了给 $\mathbb{M}_{h,m+1}$ 求上界，我们将借助以下引理，使讨论更加简洁。

引理 F.8. 对于 $m \geqslant 1$, $h \in [H-m]$, 我们考虑一个 $\in \mathcal{A}^m$

$$\mathbb{M}_{h,\mathbf{a}} := [\mathbb{P}(o_{h:h+m} = \mathbf{o} | s_h = s, a_{h:h+m-1} = \mathbf{a})]_{\mathbf{o} \in \mathcal{O}^{m+1}, s \in \mathcal{S}} \in \mathbb{R}^{\mathcal{O}^{m+1} \times \mathcal{S}}.$$

那么有

$$\min_{\mathbb{M}^+_{h,m+1}} \left\| \mathbb{M}^+_{h,m+1} \right\|_{*\to 1} \leqslant \min_{\mathbb{M}^+_{h,\mathbf{a}}} \left\| \mathbb{M}^+_{h,\mathbf{a}} \right\|_{1\to 1},$$

其中，$\min_{\mathbb{M}^+_h, \mathbf{a}}$ 取遍所有满足以下条件的 $\mathbb{M}^+_h, \mathbf{a}$ 使得 $\mathbb{M}^+_h, \mathbf{a}\mathbb{M}_h, \mathbf{a}\mathbb{T}_{h-1} = \mathbb{T}_{h-1}$ （参见定义 1）。

引理 F.8 的证明。 注意，给定一个 $\mathbf{a} \in \mathcal{A}^m$, $\mathbb{M}^+_h, \mathbf{a}$ 满足 $\mathbb{M}^+_h, \mathbf{a}\mathbb{M}_h, \mathbf{a}\mathbb{T}_{h-1} = \mathbb{T}_{h-1}$, 我们可以如下构造 $\mathbb{M}_{h,m}$ 的广义左逆:

$$\mathbb{M}^+_{h,m} = \begin{bmatrix} \vdots \\ \mathbb{1}(\mathbf{a}' = \mathbf{a}) \mathbb{M}^+_{h,\mathbf{a}} \\ \vdots \end{bmatrix}_{\mathbf{a}' \in \mathcal{A}^m},$$

并且显然 $\left\| \mathbb{M}^+_h, m \right\|_{*\to 1} \leqslant \left\| \mathbb{M}^+_h, \mathbf{a} \right\|_{1\to 1}$。 □

在下文中，对于任意矩阵 M，我们记

$$\gamma(M) := \min_{M^+ : M^+ M = I} \left\| M^+ \right\|_{1\to 1}.$$

引理 F.9。 固定一个步骤 h 和一个状态集合 $\mathcal{S}_h$。假设 $\mathcal{S}_h$ 包含所有满足 $\exists (s', a) \in \mathcal{S} \times \mathcal{A}, \mathbb{T}_{h-1}(s|s', a) > 0$ 的 $s \in \mathcal{S}$。此外，假设 $\mathcal{S}_h$ 可以划分为 $\mathcal{S}_h = \bigsqcup_{i=1}^n \mathcal{S}^i_h$，使得对于每个 $i \neq j, s \in \mathcal{S}^i_h, s' \in \mathcal{S}^j_h$,

$$\operatorname{supp}(\mathbb{M}_{h,\mathbf{a}}(\cdot|s)) \bigcap \operatorname{supp}(\mathbb{M}_{h,\mathbf{a}}(\cdot|s')) = \varnothing,$$

也就是说，不同的 $\mathcal{S}^i_h$ 所发出的观测各不相同。[8] 于是有

$$\min_{\mathbb{M}^+_{h,\mathbf{a}}} \left\| \mathbb{M}^+_{h,\mathbf{a}} \right\|_{1\to 1} \leqslant \max \left\{ \gamma\big(\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}^1_h)\big), \cdots, \gamma(\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}^n_h)) \right\},$$

其中

$$\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}') := [\mathbb{P}(o_{h:h+m} = \mathbf{o} | s_h = s, a_{h:h+m-1} = \mathbf{a})]_{\mathbf{o} \in \mathcal{O}^{m+1}, s \in \mathcal{S}'} \in \mathbb{R}^{\mathcal{O}^{m+1} \times \mathcal{S}'}, \qquad \textit{for } \mathcal{S}' \subset \mathcal{S}_h.$$

引理的证明 F.9。 我们首先注意到 $\min_{\mathbb{M}^+_h, \mathbf{a}} \left\| \mathbb{M}^+_h, \mathbf{a} \right\|_{1\to 1} \leqslant \gamma(\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h))$，因为矩阵 $\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h)$ 直接给出了 $\mathbb{M}_{h,\mathbf{a}}$ 的广义左逆（因为 $\mathcal{S}_h$ 包含所有满足 $\exists (s', a) \in \mathcal{S} \times \mathcal{A}, \mathbb{T}_{h-1}(s|s', a) > 0$ 的 $s \in \mathcal{S}$）。

接下来，由于每个 $\mathcal{S}^i_h$ 的可能观测集合彼此不相交，矩阵 $\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h)$ 可以写成（至多相差行和列的置换，且所有空项均为零）

$$\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h) = \begin{bmatrix} \mathbb{M}_{h,\mathbf{a}}(\mathcal{S}^1_h) & & & \\ & \mathbb{M}_{h,\mathbf{a}}(\mathcal{S}^2_h) & & \\ & & \ddots & \\ & & & \mathbb{M}_{h,\mathbf{a}}(\mathcal{S}^n_h) \end{bmatrix}.$$

[8] 特别是，如果对每个 $i \neq j$、 $s \in \mathcal{S}^i_h$、 $s' \in \mathcal{S}^j_h$，都有 $\operatorname{supp}(\mathbb{O}_h(\cdot|s)) \bigcap \operatorname{supp}(\mathbb{O}_h(\cdot|s')) = \varnothing$，那么该条件就满足。

因此，假设对每个 i ，我们都有一个$\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h^i)$的左逆 $\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h^i)^+$ ，那么我们可以构造$\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h)$ 的一个左逆，如下所示：

$$\mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h)^+ = \begin{bmatrix} \mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h^1)^+ & & & \\ & \mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h^2)^+ & & \\ & & \ddots & \\ & & & \mathbb{M}_{h,\mathbf{a}}(\mathcal{S}_h^n)^+ \end{bmatrix},$$

因此我们推导出 $\gamma(\mathbb{M}_{h,\ \mathbf{a}}(\mathcal{S}_h)) \leqslant \max\left\{\gamma\left(\mathbb{M}_{h,\ \mathbf{a}}(\mathcal{S}_h^1)\right),\ \cdots,\ \gamma(\mathbb{M}_{h,\ \mathbf{a}}(\mathcal{S}_h^n))\right\}$。 □

一个重要的观察是，对于矩阵 $M \in \mathbb{R}^{m\times 1}$，我们有 $\gamma(M) \leqslant \frac{1}{\|M\|_1}$。因此，当$M$ 的元素之和等于 1 时，$\gamma(M) \leqslant 1$。利用上述引理，我们现在给出引理 F.3 的证明。

引理 *F.3* 的证明。 我们首先证明零模型 0 是 1 步 1-可揭示的。在该模型中，状态 $s_\oplus$和 $e_\oplus$不可达，因此对于每一步 h，我们考虑集合 $\mathcal{S}' = \mathcal{S}_{\text{tree}} \bigsqcup \{s_\ominus,\ e_\ominus,\ \text{终止}\}$。对于不同的状态 $s,\ s' \in \mathcal{S}'$，根据我们的构造，$\mathbb{O}_h(\cdot|s)$ 和 $\mathbb{O}_h(\cdot|s')$ 的支撑集互不相交，因此应用引理 F.9 可得

$$\min_{\mathbb{O}_h^+} \left\|\mathbb{O}_h^+\right\|_{1\to 1} \leqslant \max_{s\in\mathcal{S}'} \gamma(\mathbb{O}_h(s)) \leqslant 1.$$

应用命题 2 即可完成对零模型 0 的证明。

接下来考虑非零模型 $M = M_{\theta,\ \mu} \in \mathcal{M}\backslash\{M_0\}$。根据我们的构造，对于 $h \leqslant h^\star$，状态$s_\oplus$ 和 $e_\oplus$ 不可达，因此，沿用零模型中的相同论证，我们得到 $\min_{\mathbb{M}_h^+,\ m+1} \left\|\mathbb{M}_h^+,\ m+1\right\|_{*\to 1} \leqslant \min_{\mathbb{O}_h^+} \left\|\mathbb{O}_h^+\right\|_{1\to 1} \leqslant 1$。

因此，我们只需对固定步骤 $h > h^\star$，界定数量 $\min_{\mathbb{M}_h^+,\ m+1} \left\|\mathbb{M}_h^+,\ m+1\right\|_{*\to 1}$。在这种情况下，存在一个 $l \in \mathcal{H}$ 使得 $h \leqslant l \leqslant h+m-1$，我们记作 $r = l-h+1$。根据引理 C.1，我们只需界定 $\min_{\mathbb{M}_h^+,\ r+1} \left\|\mathbb{M}_h^+,\ r+1\right\|_{*\to 1}$。考虑动作序列 $\mathbf{a} = (\mathbf{a}_{h:l-1}^\star,\ a_{\text{rev}}^\star) \in \mathcal{A}^r$，并将 $\mathcal{S}$ 划分为

$$\mathcal{S} = \bigsqcup_{s\in\mathcal{S}_{\text{tree}}} \{s\} \sqcup \{s_\oplus, s_\ominus\} \sqcup \{e_\oplus, e_\ominus\} \sqcup \{\text{terminal}\}.$$

不难直接验证，在$M_{\theta\text{中},\ \mu}$对于状态 $s,\ s'$ 来自上述划分中的不同子集时，$\mathbb{M}_{h,\ \mathbf{a}}(\cdot|s)$ 和$\mathbb{M}_{h\text{的}\mathbf{a}}(\cdot|s')$ 支撑集互不相交。于是，我们可以应用引理F.8 和引理F.9，从而得到

$$\min_{\mathbb{M}_{h,r+1}^+} \left\|\mathbb{M}_{h,r+1}^+\right\|_{*\to 1} \leqslant \min_{\mathbb{M}_{h,\mathbf{a}}^+} \left\|\mathbb{M}_{h,\mathbf{a}}^+\right\|_{1\to 1} \leqslant \max\left\{1, \gamma(\mathbb{M}_{h,\mathbf{a}}(\{s_\oplus, s_\ominus\})), \gamma(\mathbb{M}_{h,\mathbf{a}}(\{e_\oplus, e_\ominus\}))\right\}.$$

因此，在接下来 我们只需考虑矩阵$\mathbb{M}_{h,\ \mathbf{a}}(\{s_\oplus,\ s_\ominus\})$ 和$\mathbb{M}_{h,\ \mathbf{a}}(\{e_\oplus,\ e_\ominus\}$的左逆).

(1) 矩阵 $\mathbb{M}_{h,\ \mathbf{a}}(\{s_\oplus,\ s_\ominus\})$。根据我们的构造，在 $s_h = s_\oplus$ 处取一个将导致 $o_{h:l}$ = 锁和$o_{l+1} \sim \mathbb{O}_\mu(\cdot|e_\oplus)$；在 $s_h = s_\ominus$ 处取一个将导致 $o_{h:l}$ = 锁和 $o_{l+1} \sim \mathbb{O}_\mu(\cdot|e_\ominus)$。因此，$\mathbb{M}_{h,\ \mathbf{a}}(\{s_\oplus,\ s_\ominus\})$可写为（至多进行行置换）

$$\mathbb{M}_{h,\mathbf{a}}(\{s_\oplus, s_\ominus\}) = \begin{bmatrix} \frac{\mathbb{1}_{2K}+\sigma\tilde{\mu}}{2K} & \frac{\mathbb{1}_{2K}}{2K} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{\mathcal{O}^{r+1}\times 2},$$

其中，$\tilde{\mu} = [\mu;\ -\mu] \in \{-1,\ 1\}^{2K}$，$\mathbb{1} = \mathbb{1}_{2K}$ 是 $\mathbb{R}^{2K}$ 中的向量，其所有元素均为一。类似于命题 E.2，我们可以直接验证$\gamma(\mathbb{M}_{h,\ \mathbf{a}}(\{s_\oplus,\ s_\ominus\})) \leqslant \frac{2}{\sigma} + 1$。

(2) 矩阵 $\mathbb{M}_{h,\ \mathbf{a}}(\{e_\oplus,\ e_\ominus\})$。根据我们的构造，在 $s_h = e_\oplus$ 处，我们有 $o_h \sim \mathbb{O}_\mu(\cdot|e_\oplus)$，并且 $o_{h+1:l+1}$ = 是终止状态；在 $s_h = e_\ominus$ 处，我们有 $o_h \sim \mathbb{O}_\mu(\cdot|e_\ominus)$，并且 $o_{h+1:l+1}$ = 是终止状态。因此，$\mathbb{M}_{h,\ \mathbf{a}}(\{e_\oplus,\ e_\ominus\})$ 也可以写成（至多相差行置换）

$$\mathbb{M}_{h,\mathbf{a}}(\{e_\oplus, e_\ominus\}) = \begin{bmatrix} \frac{\mathbb{1}_{2K}+\sigma\tilde{\mu}}{2K} & \frac{\mathbb{1}_{2K}}{2K} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{\mathcal{O}^{r+1}\times 2},$$

因此我们也有 $\gamma(\mathbb{M}_{h,\ \mathbf{a}}(\{e_\oplus,\ e_\ominus\})) \leqslant \frac{1}{\sigma} + 2$。 –

结合上述两种情形可得

$$\min_{\mathbb{M}^+_{h,m+1}} \left\|\mathbb{M}^+_{h,m+1}\right\|_{*\to 1} \leqslant \min_{\mathbb{M}^+_{h,r+1}} \left\|\mathbb{M}^+_{h,r+1}\right\|_{*\to 1} \leqslant \min_{\mathbb{M}^+_{h,\mathbf{a}}} \left\|\mathbb{M}^+_{h,\mathbf{a}}\right\|_{1\to 1} \leqslant \frac{2}{\sigma} + 1,$$

从而完成了引理 F.3的证明。 □

注 F.10。 从上述证明中不容易看出，部分可观测马尔可夫决策过程 $M = M_{\theta,\mu}$ 对于任意参数 (θ, μ)都不是 m-步揭示的。事实上，对于 $\theta = (h^\star, s^\star, a^\star, a^\star_{\mathsf{rev}}, \mathbf{a}^\star)$，我们可以证明矩阵 $\mathbb{M}_{h^\star+1,m}$ 不 存在广义左逆。这是因为对于任意一个 $\in \mathcal{A}^{m-1}$，我们有

$$\mathbb{P}_{\theta,\mu}(o_{h^\star+1:\ h^\star+m} = \cdot\,|s_{h^\star+1} = s_\oplus, a_{h^\star+1:h^\star+m-1} = \mathbf{a}) = \mathbb{P}_{\theta,\mu}(o_{h^\star+1:h^\star+m} = \cdot\,|s_{h^\star+1} = s_\ominus, a_{h^\star+1:h^\star+m-1} = \mathbf{a}),$$

因为这两个分布都支撑于虚拟观测锁$^{\otimes m}$。然而，显然$\mathbf{e}^{s_\oplus}, \mathbf{e}^{s_\ominus} \in$ 列跨$(\mathbb{T}_{h^\star})$，因此若 $\mathbb{M}_{h^\star+1,m}$ 存在广义左逆 $\mathbb{M}^+_{h^\star+1,m}$，则$\mathbf{e}^{s_\oplus} = \mathbb{M}^+_{h^\star+1,m}\mathbb{M}_{h^\star+1,m}\mathbf{e}^{s_\oplus} = \mathbb{M}^+_{h^\star+1,m}\mathbb{M}_{h^\star+1,m}\mathbf{e}^{s_\ominus} = \mathbf{e}^{s_\ominus}$，矛盾！因此，我们可以得出结论：$\mathbb{M}_{h^\star+1,m}$ 不存在广义左逆，因此 M 不是 m-步揭示的。

F.4 引理 F.4 的证明

下文将证明(26) 和 (27)。此证明与引理 E.3的证明非常相似。 (28) 的证明也非常相似，因此为简洁起见省略。

注意，根据**遗憾** 的定义以及我们对奖励函数的构造，可得

$$\begin{aligned}\mathbb{E}^{\mathfrak{A}}_M[\mathbf{Regret}] &= T\cdot V^\star_M - \mathbb{E}^{\mathfrak{A}}_M\left[\sum_{t=1}^T r_H(o_H^{(t)})\right] = T\cdot V^\star_M - \mathbb{E}^{\mathfrak{A}}_M\left[\frac{1+\varepsilon}{4}\cdot N(o_H = s_0) + N(o_H = \mathsf{good})\right] \\ &= \left(V^\star_M - \frac{1+\varepsilon}{4}\right)\mathbb{E}^{\mathfrak{A}}_M[N(o_H = s_0)] + V^\star_M\mathbb{E}^{\mathfrak{A}}_M[N(o_H \neq s_0)] - \mathbb{E}^{\mathfrak{A}}_M[N(o_H = \mathsf{good})]\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}^{\mathfrak{A}}_M[N(o_H = \mathsf{good})] &= \mathbb{E}^{\mathfrak{A}}_M\left[\sum_{t=1}^T \mathbb{E}_M\left[\mathbb{1}(o_H^{(t)} = \mathsf{good})\,\middle|\,\tau_{H-1}^{(t)}\right]\right] \\ &= \mathbb{E}^{\mathfrak{A}}_M\left[\sum_{t=1}^T\sum_{\tau_{H-1}}\mathbb{P}_M(o_H = \mathsf{good}|\tau_{H-1})\cdot\mathbb{1}(\tau_{H-1}^{(t)} = \tau_{H-1})\right] \\ &= \sum_{\tau_{H-1}}\mathbb{E}^{\mathfrak{A}}_M[N(\tau_{H-1})]\cdot\mathbb{P}_M(o_H = \mathsf{good}|\tau_{H-1}).\end{aligned}$$

我们分别针对 $M \neq 0$ 和情况 $M = 0$ 证明该结果。

情况 1: $M = (\theta,\ \mu) \neq 0$。在这种情况下，我们有

$$\mathbb{P}_M(o_H = \mathsf{good}|\tau_{H-1}) = \frac{3}{4}\mathbb{P}_M(s_H = s_\oplus|\tau_{H-1}) + \frac{1}{4}\mathbb{P}_M(s_H = s_\ominus|\tau_{H-1}) \leqslant \frac{1}{4} + \frac{1}{2}\mathbb{P}_M(s_H = s_\oplus|\tau_{H-1}) \leqslant \frac{1}{4} + \frac{1}{2}\varepsilon,$$

因为根据我们的构造， $\mathbb{P}_M(s_H = s_\oplus|\tau_{H-1}) \leqslant \varepsilon$ 。因此，我们已证明

$$\mathbb{E}^{\mathfrak{A}}_M[N(o_H = \mathsf{good})] \leqslant \left(\frac{1}{4} + \frac{1}{2}\varepsilon\right)\mathbb{E}^{\mathfrak{A}}_M[N(o_H \neq s_0)].$$

注意，通过这种方式我们还可以证明 $V^\star_M = \frac{1+2\varepsilon}{4}$ 。因此，结合上述方程，我们得出

$$\mathbb{E}^{\mathfrak{A}}_M[\mathbf{Regret}] \geqslant \frac{\varepsilon}{4}\mathbb{E}^{\mathfrak{A}}_M[N(o_H = s_0)].$$

情况2： $M=0$。在这种情况下，$s_{\oplus}$ 不可达，因此我们有

$$\mathbb{P}_0(o_H=\mathsf{good}|\tau_{H-1})=\frac{1}{4}\mathbb{P}_0(s_H=s_{\ominus}|\tau_{H-1})\leqslant\frac{1}{4}.$$

另请注意，对于 任意轨迹 $\tau\in E_{\mathsf{rev}}$，我们都有 $\mathbb{P}_0(o_H=\text{好}|\tau_{H-1})=0$。因此，我们已经证明

$$\mathbb{E}_0^{\mathfrak{A}}[N(o_H=\mathsf{good})]\leqslant\frac{1}{4}\mathbb{E}_0^{\mathfrak{A}}[N(\{o_H\neq s_0\}-E_{\mathsf{rev}})]=\frac{1}{4}\mathbb{E}_0^{\mathfrak{A}}[N(o_H\neq s_0)]-\frac{1}{4}\mathbb{E}_0^{\mathfrak{A}}[N(E_{\mathsf{rev}})].$$

通过这种方式，我们还可以证明 $V_0^\star=\frac{1+\varepsilon}{4}\cdot o$ 因此，我们可以得出结论

$$\begin{aligned}\mathbb{E}_0^{\mathfrak{A}}[\mathbf{Regret}]&=\frac{1+\varepsilon}{4}\mathbb{E}_0^{\mathfrak{A}}[N(o_H\neq s_0)]-\mathbb{E}_0^{\mathfrak{A}}[N(o_H=\mathsf{good})]\\&\geqslant\frac{\varepsilon}{4}\mathbb{E}_0^{\mathfrak{A}}[N(o_H\neq s_0)]+\frac{1}{4}\mathbb{E}_0^{\mathfrak{A}}[N(E_{\mathsf{rev}})].\end{aligned}$$

这就完成了引理 F.4的证明。 □

F.5 引理 F.5的证明

我们首先考虑情形 (a)，即假设 $\mathbb{E}_M^{\mathfrak{A}}[\textbf{遗憾}]\leqslant T\varepsilon/32$ 对所有 $M\in\mathcal{M}$。根据马尔可夫不等式和 (26) 以及(27)，有

$$\begin{aligned}\mathbb{P}_0^{\mathfrak{A}}(N(o_H\neq s_0)\geqslant T/2)&\leqslant\frac{1}{4},\\\mathbb{P}_M^{\mathfrak{A}}(N(o_H=s_0)\geqslant T/2)&\leqslant\frac{1}{4},\qquad\forall M\neq 0.\end{aligned}$$

特别地，对于任意固定的 θ，我们考虑 $M=(\theta,\ \mu)$ 的先验分布，其中 $\mu\sim$均匀分布$(\{-1,\ 1\}^K)$，则

$$\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right](N(o_H=s_0)\geqslant T/2)\leqslant\frac{1}{4}.$$

然而，我们还有

$$\mathbb{P}_0^{\mathfrak{A}}(N(o_H=s_0)\geqslant T/2)=\mathbb{P}_0^{\mathfrak{A}}(N(o_H\neq s_0)\leqslant T/2)=1-\mathbb{P}_0^{\mathfrak{A}}(N(o_H\neq s_0)>T/2)\geqslant\frac{3}{4},$$

然后根据全变差距离的定义，有

$$D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}},\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\right)\geqslant\frac{1}{2}.$$

情形 (b) 的证明 可由与 (20) 的证明相同的论证得出，因此省略。 □

F.6 引理 F.6 的证明

我们首先证明引理 F.6 的以下版本：增加一个条件，即在 $\mathbb{P}_0^{\mathfrak{A}}$ 下，访问计数几乎必然有界。然后，我们通过截断论证将问题化归到此情形，从而证明引理 F.6 。

为了给上述量提供上界，我们将用到以下引理，它是界定 χ^2-内积的关键步骤（Canonne，2022，第 3.1 节），该内积取自 $\mathbb{P}_{\theta,\ \mu}/\mathbb{P}_0$ 和 $\mathbb{P}_{\theta,\ \mu'}/\mathbb{P}_0$ （证明见附录 F.7）。

引理 F.11 （χ^2-内积的界）。假设算法 $\mathfrak{A}$ （其停止时间 T 可以是随机的）几乎必然满足 $N(E_{\mathsf{rev}}^\theta)\leqslant N_o$ 和 $N(E_{\mathsf{correct}}^\theta)\leqslant N_r$，其中 N_o、N_r 为某些固定值。那么

$$either\ \overline{N}_o\geqslant\frac{3}{4}\frac{\delta^2\sqrt{K}}{\varepsilon^2\sigma^2},\ or\ \overline{N}_r\geqslant\frac{3}{4}\frac{\delta^2}{\varepsilon^2},$$

其中 $\delta=D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}},\ \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\ \mu}^{\mathfrak{A}}\right]\right)$。

引理 *F.11*。 由引理 D.1，可得

$$1+\chi^2\left(\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\,\|\,\mathbb{P}_0^{\mathfrak{A}}\right)=\mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_{\theta,\mu}(\tau^{(t)})\mathbb{P}_{\theta,\mu'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right].$$

在下面的引理（证明见附录 F.7）中，我们对上述等式的左端进行界定。

引理 F.12。 在引理 *F.11*的条件下，对于任意 $\mu,\ \mu'\in\{-1,\ 1\}^K$，

$$\mathbb{E}_0^{\mathfrak{A}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_{\theta,\mu}(\tau^{(t)})\mathbb{P}_{\theta,\mu'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right]\leqslant\exp\left(\overline{N}_o\cdot\frac{\sigma^2\varepsilon^2}{K}\left|\left\langle\mu,\mu'\right\rangle\right|+\frac{4}{3}\varepsilon^2\overline{N}_r\right).\tag{33}$$

利用引理 F.12，我们可以对 (33 关于 $\mu,\ \mu'\sim$均匀分布$(\{-1,\ +1\}^K)$取期望，然后

$$\begin{aligned}1+\chi^2\left(\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\,\|\,\mathbb{P}_0^{\mathfrak{A}}\right)=&\mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\mathbb{P}_{\theta,\mu}(\tau^{(t)})\mathbb{P}_{\theta,\mu'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2}\right]\\\leqslant&\mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\left[\exp\left(\overline{N}_o\cdot\frac{\sigma^2\varepsilon^2}{K}\left|\left\langle\mu,\mu'\right\rangle\right|+\frac{4}{3}\varepsilon^2\overline{N}_r\right)\right].\end{aligned}$$

注意到 $\mu_i,\ \mu_i'$服从 $(\{\pm1\})$上的独立同分布均匀分布，因此 $\mu_1\mu_1',\ \cdots,\ \mu_K\mu_K'$服从 $(\{\pm1\})$上的独立同分布均匀分布。根据 Hoeffding 引理，可得 $\mathbb{E}_{\mu,\ \mu'\sim\text{均匀分布}}\left[\text{指数}\ \left(x\sum_{i=1}^K\mu_i\mu_i'\right)\right]\leqslant\text{指数}\ \left(Kx^2/2\right)$对所有 $x\in\mathbb{R}$均成立，因此根据引理 A.3，可得

$$\mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\left[\exp\left(\frac{\overline{N}_o\sigma^2\varepsilon^2}{K}\left|\left\langle\mu,\mu'\right\rangle\right|\right)\right]\leqslant\exp\left(\max\left\{\frac{\sigma^4\varepsilon^4\overline{N}_o^2}{K},\frac{4}{3}\frac{\sigma^2\varepsilon^2\overline{N}_o}{\sqrt{K}}\right\}\right).$$

因此，结合上述不等式与引理 A.5，我们得到

$$2\delta^2=2D_{\mathrm{TV}}\left(\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right],\mathbb{P}_0^{\mathfrak{A}}\right)^2\leqslant\log\left(1+\chi^2\left(\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\,\|\,\mathbb{P}_0^{\mathfrak{A}}\right)\right)\leqslant\max\left\{\frac{4}{3}\frac{\overline{N}_o\sigma^2\varepsilon^2}{\sqrt{K}},\frac{\overline{N}_o^2\sigma^4\varepsilon^4}{K}\right\}+\frac{4}{3}\varepsilon^2\overline{N}_r.$$

那么，我们要么有 $\overline{N_r}\geqslant\frac{3\delta^2}{4\varepsilon^2}$，要么有

$$\max\left\{\frac{4}{3}\frac{\overline{N}_o\sigma^2\varepsilon^2}{\sqrt{K}},\frac{\overline{N}_o^2\sigma^4\varepsilon^4}{K}\right\}\geqslant\delta^2,$$

这意味着 $\frac{N_o\sigma^2\varepsilon^2}{\sqrt{K}}\geqslant$最小值 $\left\{\frac{4}{3},\ \frac{3}{4}\delta^2\right\}=\frac{3}{4}\delta^2$（因为 $\delta\leqslant1$）。通过重新整理即可完成引理 F.11 的证明。

□

引理 *F.6* 的证明。 我们采用一种截断型论证，将引理 F.6 化归为引理 F.11，这与引理 E.4的证明类似。

令我们取 $N_o=\left\lceil6\delta^{-1}\mathbb{E}_0^{\mathfrak{A}}\left[N(E_{\mathsf{rev}}^\theta)\right]\right\rceil$和 $N_r=\left\lceil6\delta^{-1}\mathbb{E}_0^{\mathfrak{A}}\left[N(E_{\mathsf{correct}}^\theta)\right]\right\rceil$。根据马尔可夫不等式，我们有

$$\mathbb{P}_0^{\mathfrak{A}}\left(N(E_{\mathsf{rev}}^\theta)\geqslant\overline{N}_o\right)\leqslant\frac{\delta}{6},\qquad\mathbb{P}_0^{\mathfrak{A}}\left(N(E_{\mathsf{correct}}^\theta)\geqslant\overline{N}_r\right)\leqslant\frac{\delta}{6}.$$

因此，我们可以为算法 $\mathfrak{A}$考虑以下退出条件：

$$\mathsf{exit}(\tau^{(1:T')})=\mathsf{TRUE}\quad\text{iff}\quad\sum_{t=1}^{T'}\mathbb{I}\left(\tau^{(t)}\in E_{\mathsf{rev}}^\theta\right)\geqslant\overline{N}_o\ \text{or}\ \sum_{t=1}^{T'}\mathbb{I}\left(\tau^{(t)}\in E_{\mathsf{correct}}^\theta\right)\geqslant\overline{N}_r.$$

退出条件会产生一个停止时间 T_{exit}，并且我们有

$$\mathbb{P}_0^{\mathfrak{A}}(\exists t<T, \mathsf{exit}(\tau^{(1:t)})=\mathsf{TRUE}) \leqslant \mathbb{P}_0^{\mathfrak{A}}\left(N(E_{\mathsf{rev}}^{\theta}) \geqslant \overline{N}_o \text{ or } N(E_{\mathsf{correct}}^{\theta}) \geqslant \overline{N}_r\right) \leqslant \frac{\delta}{6}+\frac{\delta}{6} \leqslant \frac{\delta}{3}.$$

因此，我们可以考虑提前停止算法 $\mathfrak{A}$(退出)，其退出准则为 退出 (参见附录 D)，根据引理D.2 可得

$$D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}(\mathsf{exit})}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}(\mathsf{exit})}\right]\right) \geqslant D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\right) - \mathbb{P}_0^{\mathfrak{A}}(\exists t<T, \mathsf{exit}(\tau^{(1:t)})=\mathsf{TRUE}) \geqslant \frac{2\delta}{3}.$$

注意，根据　我们对退出和停止时间 T_{exit} 的定义，在 $\mathfrak{A}$(退出) 的执行过程中，也有

$$\sum_{t=1}^{T_{\mathsf{exit}}-1} \mathbb{1}\left(\tau^{(t)} \in E_{\mathsf{rev}}^{\theta}\right) < \overline{N}_o, \qquad \sum_{t=1}^{T_{\mathsf{exit}}-1} \mathbb{1}\left(\tau^{(t)} \in E_{\mathsf{correct}}^{\theta}\right) < \overline{N}_r.$$

因此，算法 $\mathfrak{A}$(退出) 确保

$$N(E_{\mathsf{rev},h}^{\theta}) = \sum_{t=1}^{T_{\mathsf{exit}}} \mathbb{1}\left(\tau^{(t)} \in E_{\mathsf{rev}}^{\theta}\right) \leqslant \overline{N}_o, \qquad N(E_{\mathsf{correct}}^{\theta}) = \sum_{t=1}^{T_{\mathsf{exit}}} \mathbb{1}\left(\tau^{(t)} \in E_{\mathsf{correct}}^{\theta}\right) \leqslant \overline{N}_r.$$

将引理 F.11 应用于算法 $\mathfrak{A}$(退出) (以及 $\delta' = \frac{2}{3}\delta$)，可得 –

$$\text{either } \frac{\delta^2\sqrt{K}}{3\varepsilon^2\sigma^2} \leqslant \overline{N}_o \leqslant 6\delta^{-1}\mathbb{E}_0^{\mathfrak{A}}\left[N(E_{\mathsf{rev}}^{\theta})\right]+1, \qquad \text{or } \frac{\delta^2}{3\varepsilon^2} \leqslant \overline{N}_r \leqslant 6\delta^{-1}\mathbb{E}_0^{\mathfrak{A}}\left[N(E_{\mathsf{correct}}^{\theta})\right]+1,$$

重新整理后，即可得到引理 F.6的所需结论。 □

引理 F.7 的证明 F.12

在整个证明过程中，参数 θ，μ，μ' 均固定。

根据附录D中的讨论，利用 (16)，我们得到

$$\mathbb{E}_{\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\mathbb{P}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}} \frac{\mathbb{P}_M(\tau^{(t)})\mathbb{P}_{M'}(\tau^{(t)})}{\mathbb{P}_0(\tau^{(t)})^2} \cdot \exp\left(-\sum_{t=1}^{\mathsf{T}}\sum_{h=1}^{H} \log I(\tau_{h-1}^{(t)})\right)\right] = 1, \tag{34}$$

其中，我们将任意截至步骤 $l\in[H]$ 的部分轨迹 τ_l 定义为 $I(\tau_l)$

$$I(\tau_l) := \mathbb{E}_0\left[\frac{\mathbb{P}_{\theta,\mu}(o_{l+1}|\tau_l)\mathbb{P}_{\theta,\mu'}(o_{l+1}|\tau_l)}{\mathbb{P}_0(o_{l+1}|\tau_l)^2}\middle|\tau_l\right].$$

注意，模型 $\mathbb{P}_{\theta,\ \mu}$ 和 $\mathbb{P}_0$ 仅在从 $s_{h^\star}=s^\star$，$a_{h^\star}=a^\star$ 到 $s_\oplus$ 的转移以及分量 $\{s_\oplus,\ e_\oplus\}$ 的动力学上有所不同。因此，对于任意（可达）轨迹 $\tau_l=(o_1,\ a_1,\ \cdots,\ o_l,\ a_l)$，我们可以考察 $\mathbb{P}_{\theta,\ \mu}(o_{l+1}=\cdot|\tau_l)\neq\mathbb{P}_0(o_{l+1}=\cdot|\tau_l)$ 的推论：

1. 显然，$o_{h^\star}=s^\star, a_{h^\star}=a^\star$ （即 $l\geqslant h^\star+1$，并且在 s_0 执行动作 $a_{1:h^\star-1}$ 将在步骤 $h^\star$ 得到 $s^\star$）。2. 要么 $a_{h^\star+1:\ l}=(\mathbf{a}_{h^\star+1:\ l-1}^\star,\ a_{\mathsf{rev}}^\star)$ 对某个 $l\in\mathcal{H}_{>h^\star}$ 成立，要么 $l=H-1$ 且 $a_{h^\star+1:\ H-1}=\mathbf{a}^\star$。

因此，对于 $l\in\mathcal{H}_{>h^\star}$，我们定义

$$E_{\mathsf{rev},l}^{\theta} := \left\{o_{h^\star}=s^\star, a_{h^\star:l}=(a^\star, \mathbf{a}_{h^\star+1:l-1}^\star, a_{\mathsf{rev}}^\star)\right\}.$$

另外，回顾一下我们定义了 $E_{\mathsf{correct}}^{\theta}:=\{o_{h^\star}=s^\star,\ a_{h^\star:\ H-1}=(a^\star,\ \mathbf{a}^\star)\}$。那么，如果 $\mathbb{P}_{\theta,\ \mu}(\cdot|\tau_l)\neq\mathbb{P}_0(\cdot|\tau_l)$，则要么 $l\in\mathcal{H}_{>h^\star}$，$\tau_l\in E_{\mathsf{rev},\ l}^{\theta}$，要么 $l=H-1$，$\tau_{H-1}\in E_{\mathsf{correct}}^{\theta}$。换句话说，对于任意 τ_l（即在 $\mathbb{P}_0$ 下可达的），除了这两种情况外，我们都有 $I(\tau_l)=1$，而接下来只需计算这两种情况下的 $I(\tau_l)$。

情况 1： $l\in\mathcal{H}_{>h^\star}$，$\tau_l\in E_{\mathsf{rev},\ l}^{\theta}$。在这种情况下，我们有

$$\mathbb{P}_{\theta,\ \mu}(o_{l+1}=o|\tau_l) = \mathbb{P}_{\theta,\mu}(o_{l+1}=o|s_{l+1}=e_\oplus)\mathbb{P}_{\theta,\mu}(s_{l+1}=e_\oplus|\tau_l) + \mathbb{P}_{\theta,\mu}(o_{l+1}=o|s_{l+1}=e_\ominus)\mathbb{P}_{\theta,\mu}(s_{l+1}=e_\ominus|\tau_l$$

$$= (\mathbb{O}_\mu(o|e_\oplus) - \mathbb{O}(o|e_\ominus)) \cdot \mathbb{P}_{\theta,\mu}(s_{l+1} = e_\oplus|\tau_l) + \mathbb{O}(o|e_\ominus),$$

其中 第二个等式成立，是因为在给定 τ_l 的条件下，我们有 $s_{l+1} \in \{e_\oplus,\ e_\ominus\}$。现在，我们 h ave

$$\begin{aligned}\mathbb{P}_{\theta,\mu}(s_{l+1} = e_\oplus|\tau_l) &= \mathbb{P}_{\theta,\mu}(s_{l+1} = e_\oplus|o_{h^\star} = s^\star, a_{h^\star:l} = (a^\star, \mathbf{a}^\star_{h^\star+1:l-1}, a^\star_{\mathsf{rev}})) \\ &= \mathbb{P}_{\theta,\mu}(s_{h^\star+1} = s_\oplus|o_{h^\star} = s^\star, a_{h^\star} = a^\star) = \varepsilon.\end{aligned}$$

因此，根据 $\mathbb{O}_\mu(\cdot|e_\oplus)$ 和 $\mathbb{O}(\cdot|e_\ominus)$的定义，我们可以得出

$$\mathbb{P}_{\theta,\mu}(o_{l+1} = o_i^+|\tau_l) = \frac{1+\varepsilon\sigma\mu_i}{2K}, \qquad \mathbb{P}_{\theta,\mu}(o_{l+1} = o_i^-|\tau_l) = \frac{1-\varepsilon\sigma\mu_i}{2K}, \qquad \forall i \in [K].$$

另一方面，显然 $\mathbb{P}_0(o_{l+1} = \cdot|\tau_l) =$均匀分布($\{o_1^+,\ o_1^-,\ \cdots,\ o_K^+,\ o_K^-\}$)。因此，有

$$\begin{aligned}I(\tau_l) =& \mathbb{E}_0\left[\frac{\mathbb{P}_{\theta,\mu}(o_{l+1}|\tau_l)\mathbb{P}_{\theta,\mu'}(o_{l+1}|\tau_l)}{\mathbb{P}_0(o_{l+1}|\tau_l)^2}\middle|\tau_l\right] \\ =& \frac{1}{2K}\sum_{o\in\mathcal{O}_o}\frac{\mathbb{P}_{\theta,\mu}(o_{l+1} = o|\tau_l)\mathbb{P}_{\theta,\mu'}(o_{l+1} = o|\tau_l)}{\mathbb{P}_0(o_{l+1} = o|\tau_l)^2} \\ =& \frac{1}{2K}\sum_{i=1}^K (1+\varepsilon\sigma\mu_i)(1+\varepsilon\sigma\mu_i') + (1-\varepsilon\sigma\mu_i)(1-\varepsilon\sigma\mu_i') \\ =& 1 + \frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K \mu_i\mu_i' = 1 + \frac{\varepsilon^2\sigma^2}{K}\langle\mu,\mu'\rangle.\end{aligned}$$

情况2： $l = H-1.\tau_{H-1} \in E^\theta_{\mathsf{correct}}$。在这种情况下，分布 $\mathbb{P}(o_H = \cdot|\tau_{H-1})$ 支撑于 {好、坏}。与情况 1 类似，我们有

$$\begin{aligned}\mathbb{P}_{\theta,\mu}(o_H = \cdot|\tau_l) &= \mathbb{P}_{\theta,\mu}(o_H = \cdot|s_H = s_\oplus)\mathbb{P}_{\theta,\mu}(s_H = s_\oplus|\tau_{H-1}) + \mathbb{P}_{\theta,\mu}(o_H = \cdot|s_H = s_\ominus)\mathbb{P}_{\theta,\mu}(s_H = s_\ominus|\tau_{H-1}) \\ &= (\mathbb{O}_H(\cdot|s_\oplus) - \mathbb{O}_H(\cdot|s_\ominus)) \cdot \mathbb{P}_{\theta,\mu}(s_H = s_\oplus|\tau_{H-1}) + \mathbb{O}_H(\cdot|s_\ominus),\end{aligned}$$

其中第二个等式是因为在给定 $\tau_{H-1} \in E^\theta_{\mathsf{correct}}$的条件下，有 $s_H \in \{s_\oplus,\ s_\ominus\}$。现在，我们有

$$\mathbb{P}_{\theta,\ \mu}(s_H = s_\oplus|\tau_{H-1}) = \mathbb{P}_{\theta,\ \mu}(s_H = s_\oplus|o_{h^\star} = s^\star、\ a_{h^\star:\ H-1} = (a^\star,\ \text{一个}^\star)) = \mathbb{P}_{\theta,\ \mu}(s_{h^\star+1} = s_\oplus|o_{h^\star} = s^\star、\ a_{h^\star} = a^\star) = \varepsilon.$$

因此，根据 $\mathbb{O}_H(\cdot|s_\oplus)$ 和 $\mathbb{O}_H(\cdot|s_\ominus)$的定义，我们有

$$\mathbb{P}_{\theta,\mu}(o_H = o|\tau_{H-1}) = \begin{cases}\frac{1+2\varepsilon}{4}, & o = \mathsf{good}, \\ \frac{3-2\varepsilon}{4}, & o = \mathsf{bad}.\end{cases}$$

另一方面，显然 $\mathbb{P}_0(o_H =$好$|\tau_{H-1}) = \frac{1}{4}$，$\mathbb{P}_0(o_H =$坏$|\tau_{H-1}) = \frac{3}{4}$。因此，在这种情况下，我们有

$$I(\tau_{H-1}) = \mathbb{E}_0\left[\frac{\mathbb{P}_{\theta,\mu}(o_H|\tau_{H-1})\mathbb{P}_{\theta,\mu'}(o_H|\tau_{H-1})}{\mathbb{P}_0(o_H|\tau_{H-1})^2}\middle|\tau_{H-1}\right] = \frac{1}{4}\times\left(\frac{(1+2\varepsilon)/4}{1/4}\right)^2 + \frac{3}{4}\times\left(\frac{(3-2\varepsilon)/4}{3/4}\right)^2 = 1 + \frac{4}{3}\varepsilon^2.$$

结合上述两种情况，我们得到

$$I(\tau_l) = \begin{cases}1 + \frac{\varepsilon^2\sigma^2}{K}\langle\mu,\mu'\rangle, & l \in \mathcal{H}_{>h^\star}, \tau_l \in E^\theta_{\mathsf{rev},l}, \\ 1 + \frac{4}{3}\varepsilon^2, & l = H-1, \tau_{H-1} \in E^\theta_{\mathsf{correct}}, \\ 1, & \text{otherwise.}\end{cases}$$

因此，对于每个 $t \in$ [T],

$$\sum_{l=0}^{H-1}\log I(\tau_l^{(t)}) = \sum_{l=0}^{H-1}\mathbb{1}\left(l \in \mathcal{H}_{>h^\star}, \tau_l^{(t)} \in E^\theta_{\mathsf{rev},l}\right)\cdot\log\left(1 + \frac{\varepsilon^2\sigma^2}{K}\langle\mu,\mu'\rangle\right) + \mathbb{1}\left(\tau_{H-1}^{(t)} \in E^\theta_{\mathsf{correct}}\right)\cdot\log\left(1+\frac{4}{3}\varepsilon^2\right)$$

$$\leqslant \sum_{l\in\mathcal{H}_{>h^\star}} \mathbb{1}\left(\tau_l^{(t)} \in E^\theta_{\mathsf{rev},l}\right)\cdot\frac{\varepsilon^2\sigma^2}{K}\left\langle\mu,\mu'\right\rangle + \mathbb{1}\left(\tau_{H-1}^{(t)}\in E^\theta_{\mathsf{correct}}\right)\cdot\frac{4}{3}\varepsilon^2$$

$$= \mathbb{1}\left(\tau_H^{(t)} \in E^\theta_{\mathsf{rev}}\right)\cdot\frac{\varepsilon^2\sigma^2}{K}\left\langle\mu,\mu'\right\rangle + \mathbb{1}\left(\tau_H^{(t)}\in E^\theta_{\mathsf{correct}}\right)\cdot\frac{4}{3}\varepsilon^2,$$

最后一个等式成立是因为 $E^\theta_{\mathsf{rev}} = \bigsqcup_{l:\ l\in\mathcal{H}_{>h^\star}} E^\theta_{\mathsf{rev},\ l}$。对 $t\in[\mathrm{T}]$ 求和，可得

$$\sum_{t=1}^{\mathsf{T}}\sum_{l=0}^{H-1}\log I(\tau_l^{(t)}) \leqslant \sum_{t=1}^{\mathsf{T}}\mathbb{1}\left(\tau_H^{(t)}\in E^\theta_{\mathsf{correct}}\right)\cdot\frac{4}{3}\varepsilon^2 + \mathbb{1}\left(\tau_H^{(t)}\in E^\theta_{\mathsf{rev}}\right)\cdot\frac{\varepsilon^2\sigma^2}{K}\left\langle\mu,\mu'\right\rangle$$

$$= N(E^\theta_{\mathsf{correct}})\cdot\frac{4}{3}\varepsilon^2 + N(E^\theta_{\mathsf{rev}})\cdot\frac{\varepsilon^2\sigma^2}{K}\left\langle\mu,\mu'\right\rangle$$

$$\leqslant \overline{N}_r\cdot\frac{4}{3}\varepsilon^2 + \overline{N}_o\cdot\frac{\varepsilon^2\sigma^2}{K}\left|\left\langle\mu,\mu'\right\rangle\right|.$$

将上述不等式代入（34），即可完成引理 F.12 的证明。 □

F.8 第 5 节中困难实例的遗憾计算

对于第 5 节中给出的困难实例，我们注意到，任何算法要么会遭受 $\Omega(\varepsilon T)$ 遗憾，要么必须在 $(h^\star、\ s^\star、\ a^\star)$ 的交互回合内成功识别它们，这就需要至少 $\Omega(SAH\times\sqrt{O/(\sigma^2\varepsilon^2)})$ 个回合采取揭示动作，且每次都存在 $\Theta(1)$的次优性；或者至少 $\Omega(SAH\times A^{\Theta(H)}/\varepsilon^2)$个回合尝试所有可能的动作序列，且每次都存在 $\Theta(\varepsilon)$的次优性。由此得到遗憾下界

$$\Omega\left(SAH\times\min\left\{\tfrac{\sqrt{O}}{\sigma^2\varepsilon^2}, \tfrac{A^{\Theta(H)}}{\varepsilon^2}\cdot\varepsilon\right\}\right)\wedge\Omega(\varepsilon T).$$

对 $\varepsilon>0$进行优化后，我们得到一个 $\Omega(T^{2/3})$类型的遗憾下界（对于 $T\ll A^{\mathcal{O}(H)}$），其形式与定理 6 类似（但速率略差）。

G 定理 5 的证明

我们首先在附录 G.1中构造一族困难实例。随后在命题 G.1中给出这族困难实例的 PAC 下界。定理 5 由命题 G.1和命题 F.1结合即可得出。

G.1 困难实例的构造及定理 5的证明

我们考虑以下一族 m步揭示型部分可观测马尔可夫决策过程 $\mathcal{M}$，该族由一组超参数$(\varepsilon,\ \sigma,\ n,\ m,\ K,\ L,\ H)$确定。$\mathcal{M}$ 中的所有部分可观测马尔可夫决策过程共享状态空间 $\mathcal{S}$、动作空间 $\mathcal{A}$、观测空间 $\mathcal{O}$和时域长度 H，其定义如下。

- 状态空间 $\mathcal{S}=\mathcal{S}_{\text{树}}\bigsqcup_{j=1}^{L}\left\{s^j_\oplus,\ s^j_\ominus,\ e^j_\oplus,\ e^j_\ominus,\ \text{终端}^j\right\}$，其中 $\mathcal{S}_{\text{树}}$ 是一棵层级为 n 的二叉树（因此$|\mathcal{S}_{\text{树}}|=2^n-1$）。令 s_0为 $\mathcal{S}_{\text{树}}$的根节点，并令 $\mathcal{S}_{\text{叶节点}}$ 为 $\mathcal{S}_{\text{树}}$的叶节点集合，其中$|\mathcal{S}_{\text{叶节点}}|=2^{n-1}$。
- 观测空间 $\mathcal{O}=\mathcal{S}_{\text{树}}\bigsqcup\left\{o_1^+,\ o_1^-,\ \cdots,\ o_K^+,\ o_K^-\right\}\bigsqcup\{\text{锁},\ \text{好},\ \text{坏}\}\bigsqcup_{j=1}^{L}\{\text{锁}^j,\ \text{终止}^j\}$。
- 动作空间 $\mathcal{A}=\{0,\ \mathbf{1},\ \cdots,\ A-1\}$。

我们进一步定义揭示 $=0\in\mathcal{A}$，$\mathcal{A}_c=\{1,\ \cdots,\ A-1\}$.

模型参数 每个非零部分可观测马尔可夫决策过程模型 $M=M_{\theta,\ \mu}\in\mathcal{M}\backslash\{M_0\}$ 由参数 $(\theta,\ \mu)$指定，其中 $\mu\in\{-1,\ +1\}^{L\times K}$，且 $\theta=(h^\star,\ s^\star,\ a^\star,\ \mathbf{a}^\star)$，其中

- $s^\star\in\mathcal{S}_{\mathsf{leaf}}$, $a^\star\in\mathcal{A}_c:=\{1,\cdots,A-1\}$.

- $h^\star \in \mathcal{H} := \{h = n + lm\colon\ l \in \mathbb{Z}_{\geqslant 0},\ \ h < H\}$。

- **动作**$^\star = (\textbf{动作}^\star_{h^\star+1},\ \ldots\ldots,\ \textbf{动作}^\star_{H-1}) \in \mathcal{A}^{H-h^\star-1}$ 是一个由 $h^\star+1$、$\cdots$、$H-1$索引的动作序列，使得当$h \in \mathcal{H}$时，我们有**动作**$^\star_h \neq$ 揭示。我们用 $\mathcal{A}_{\mathsf{code}\ h^\star}$ 表示所有此类**动作***的集合。

我们的构造将确保，只有在步骤 $h \in \mathcal{H}$ 和状态 $s_h \in \{s_\oplus、\ s_\ominus\}$下，智能体才能观测到锁^j，并采取揭示动作，转移至 $\left\{e^j_\oplus、\ e^j_\ominus\right\}$。

对于任意部分可观测马尔可夫决策过程 $M_{\theta,\ \mu}$，其系统动力学 $\mathbb{P}_{\theta,\ \mu} := \mathbb{P}_{M_{\theta,\ \mu}}$ 定义如下。

发射动态在终止状态 $s \in \mathcal{S}_{\mathsf{tree}} \cup \{\mathsf{terminal}\}$，智能体总是将状态自身（对应的唯一观测）作为观测接收。

- 在状态 $e^j_\oplus$下，发射动态定义为

$$\mathbb{O}_\mu(o_i^+|e^j_\oplus) = \frac{1+\sigma\mu_{j,i}}{2K}, \qquad \mathbb{O}_\mu(o_i^-|e^j_\oplus) = \frac{1-\sigma\mu_{j,i}}{2K}, \qquad \forall i \in [K],$$

  其中我们省略下标 h，因为发射分布不依赖于 h。

- 在状态 $e^j_\ominus$，观测从以下集合中均匀抽取：$\mathcal{O}_o := \left\{o_1^+,\ o_1^-,\ \cdots,\ o_K^+,\ o_K^-\right\}$,即 $\mathbb{O}(\cdot|e^j_\ominus)$ =均匀分布$(\mathcal{O}_o)$。

- 在状态 $s \in \{s^j_\oplus,\ s^j_\ominus\}$:

  – 在步骤 $h \in \mathcal{H}$，智能体始终接收到锁^j 作为观测。– 对于不属于 $\mathcal{H}$ 的步骤 $h \leqslant H-1$，智能体始终接收到锁作为观测。– 在步骤 H，$\{s^j_\oplus,\ s^j_\ominus\}$ 处的发射动态为

$$\mathbb{O}_H(\mathsf{good}|s^j_\oplus) = \frac{3}{4}, \qquad \mathbb{O}_H(\mathsf{bad}|s^j_\oplus) = \frac{1}{4},$$
$$\mathbb{O}_H(\mathsf{good}|s^j_\ominus) = \frac{1}{4}, \qquad \mathbb{O}_H(\mathsf{bad}|s^j_\ominus) = \frac{3}{4}.$$

转移动态 在每个回合中，智能体始终从 s_0 开始。

- 在任意节点 $s \in \mathcal{S}_{\text{树}} \backslash \mathcal{S}_{\text{叶}}$，都有三种可用动作：等待 $=0$、左 $=1$ 和右 $=2$，智能体可以执行等待以停留在 s，执行左以转移到 s 的左子节点，执行右以转移到 s的右子节点。

- 在任意 $s \in \mathcal{S}_{\text{叶节点}}$，智能体都可以执行等待动作 $=0$ 以停留在 s（即 $\mathbb{P}(s|s,\ \text{等待}) = 1$）；否则，对于$s \in \mathcal{S}_{\text{叶节点}},\ \ h \in [H-1], a \neq$ 等待，

$$\mathbb{P}_{h;\theta}(s^j_\oplus|s,a) = \frac{\varepsilon}{L} \cdot \mathbb{1}(h = h^\star, s = s^\star, a = a^\star),$$
$$\mathbb{P}_{h;\theta}(s^j_\ominus|s,a) = \frac{1}{L} - \frac{\varepsilon}{L} \cdot \mathbb{1}(h = h^\star, s = s^\star, a = a^\star).$$

- 无论采取什么动作，状态 $s \in \{e^j_\oplus,\ e^j_\ominus\}$ 都会转移到终止状态^j。

- 状态终止状态^1，$\cdots$，终止状态^L 是吸收状态。

- 在状态 $s \in \{s^j_\oplus,\ s^j_\ominus\}$:

  – 对于步骤 $h \in \mathcal{H}$，我们设定

$$\mathbb{P}_{h;\theta}(e^j_\oplus|s^j_\oplus, \mathsf{reveal}) = 1, \qquad \mathbb{P}_{h;\theta}(e^j_\ominus|s^j_\ominus, \mathsf{reveal}) = 1.$$

  换言之，在步骤 $h \in \mathcal{H}$ 和状态 $s \in \{s^j_\oplus,\ s^j_\ominus\}$（对应于 $o =$ 锁^j）下，智能体可以执行揭示动作 reveal，分别转移至 $\{e^j_\oplus,\ e^j_\ominus\}$。注意，$\mathcal{H} = \{h = n + lm\colon\ h < H,\ l \in \mathbb{Z}_{\geqslant 0}\}$，因此我们只允许智能体每隔 m 步执行一次揭示动作 reveal，从而确保我们的构造具有 $(m+1)$步揭示性。

– 对于 $h \notin \mathcal{H}$ 或 $a \neq$ 揭示，我们设定

$$\mathbb{P}_{h;\theta}(s^j_{\oplus}|s^j_{\oplus},a) = \mathbb{1}(a = \mathbf{a}^\star_h), \qquad \mathbb{P}_{h;\theta}(s^j_{\ominus}|s^j_{\oplus},a) = \mathbb{1}(a \neq \mathbf{a}^\star_h),$$
$$\mathbb{P}_{h;\theta}(s^j_{\ominus}|s^j_{\ominus},a) = 1.$$

奖励 奖励函数是已知的（且仅依赖于观测）：在前 $H-1$ 个步骤中，不给予奖励；在第 H 步，我们设定 $r_H(\text{好}) = 1$、$r_H(\text{坏}) = 0$，以及对于任何其他 $o \in \mathcal{O}$，设定 $r_H(s_0) = (1+\varepsilon)/4$, 和 $r_H(o) = 0$。

参考模型 我们使用 M_0（或简称为 0）来指代零模型（参考模型）。零模型 M_0 的转移和发射与任何非零模型相同，不同之处在于，智能体在 $s \in \mathcal{S}_{\mathsf{leaf}}$ 处采取任何动作 $a \neq$ 等待后，总会到达 $s^j_{\ominus}$（服从 $j \sim$ 均匀分布$([L])$）和 $h \in [H-1]$（即，对于任何 $s \in \mathcal{S}_{\mathsf{leaf}}$、$a \in \mathcal{A}_c$、$h \in [H-1]$，$\mathbb{P}_{h;M_0}(s^j_{\ominus}|s,\ a) = \frac{1}{L}$）。在该模型中，$\{s^1_{\oplus},\ e^1_{\oplus},\ \cdots,\ s^L_{\oplus},\ e^L_{\oplus}\}$ 中的状态均不可达，因此我们无需指定这些状态下的转移和发射动态。

我们在以下命题中总结所构造困难实例的结果，并在附录 G.2 中证明该命题。

命题 G.1。对于给定的 $\varepsilon \in (0, 0.1], \sigma \in (0,1],\ m, n \geqslant 1,\ K, L \geqslant 1,\ H \geqslant 8n+m+1$，上述模型类 $\mathcal{M}$ 满足以下性质。

1. $|\mathcal{S}| = 2^n + 5L,\ |\mathcal{O}| = 2^n + 2K + 2L + 3,\ |\mathcal{A}| = A$。
2. 对每个 $M \in \mathcal{M}$，M 都是具有 $\alpha_{m+1}(M)^{-1} \leqslant 1 + \frac{2}{\sigma}$ 的 $(m+1)$-步骤揭示模型。
3. $\log|\mathcal{M}| \leqslant LK \log\ 2 + H \log\ A + \log(SAH)$.
4. 假设 该算法 $\mathfrak{A}$ 与环境交互 T 回合，并返回满足以下条件的 π^{out} hat

$$\mathbb{P}^{\mathfrak{A}}_M\left(V^\star_M - V_M(\pi^{\text{out}}) < \frac{\varepsilon}{8}\right) \geqslant \frac{3}{4}.$$

对任意 $M \in \mathcal{M}$。那么必有

$$T \geqslant \frac{1}{10000m} \min\left\{\frac{|\mathcal{S}_{\mathsf{leaf}}|\sqrt{LK}A^m H}{\sigma^2\varepsilon^2}, \frac{|\mathcal{S}_{\mathsf{leaf}}|\ A^{H/2}H}{\varepsilon^2}\right\}.$$

定理 5 的证明 应用命题 G.1 时，我们需要适当地选择参数。具体而言，给定 (S, O, A, H, α, m)，我们可以令 $m' = m-1$，并取满足 $2^n \leqslant S/4$, 的最大整数作为 $n \geqslant 1$；再取 $L = \lfloor (S-2^n)/5 \rfloor$, $K = \left\lfloor \frac{O - 2^n - 2L - 3}{2} \right\rfloor \gtrsim O$（因为 $O \geqslant S \geqslant 10$）、$\varepsilon' = \varepsilon/8$, 和 $\sigma = \frac{2}{\alpha^{-1}-1} \leqslant 1$。将参数 $(\varepsilon, \sigma, n, m', K, L, H)$ 代入命题 G.1，可得到一个模型类 $\mathcal{M}$，其中包含 m 步、α 揭示型部分可观测马尔可夫决策过程。若存在算法 $\mathfrak{A}$，它与环境交互 T 回合，并返回一个 π^{out}，使得对于所有 $M \in \mathcal{M}$，以至少 $3/4$ 的概率满足 $V^\star_M - V_M(\pi^{\text{out}}) < \varepsilon$，那么

$$T \geqslant \frac{c_0}{m} \min\left\{\frac{S^{3/2}O^{1/2}A^{m-1}H}{\alpha^2\varepsilon^2}, \frac{SA^{H/2}H}{\varepsilon^2}\right\},$$

其中，c_0 是一个普适常数。

此外，我们也可以改用命题F.1 （主张 5），并类似地得到一个包含 $\mathcal{M}'$ 个m步的 α-揭示型部分可观测马尔可夫决策过程的模型类，使得若存在一种算法 $\mathfrak{A}$ 与环境交互 T个回合并返回一个 π^{out}满足 $V^\star_M - V_M(\pi^{\text{out}}) < \varepsilon$的结果，且其概率至少为 $3/4$对所有 $M \in \mathcal{M}'$均成立，则

$$T \geqslant \frac{c_0'}{m} \min\left\{\frac{SO^{1/2}A^m H}{\alpha^2\varepsilon^2}, \frac{SA^{H/2}H}{\varepsilon^2}\right\},$$

其中， c_0' 是一个普适常数。

结合这两种情况，定理5的证明即告完成。 □

G.2 命题G.1的证明

本节中陈述的所有命题和引理均在附录G.3-G.4中证明。

断言 1 可通过计算 $\mathcal{M}$ 中模型的状态、观测和动作数量直接得出。断言 3 成立，因为我们有 $|\mathcal{M}| = |\{(h^\star, s^\star, a^\star, \mathbf{a}^\star)\}| \times \left|\{\pm 1\}^{L\times K}\right| + 1 \leqslant HSA^H \times 2^{LK}$。取对数即可得到该断言。

断言 2 可由该引理推出，其证明见附录G.3。

引理 G.2。 对于每个 $M \in \mathcal{M}$，均有 $\alpha_{m+1}(M)^{-1} \leqslant \frac{2}{\sigma} + 1$。 –

根据我们的构造，我们可以将输出策略的次优性与模型之间的全变差距离联系起来（在参数 $\mu \sim$ 均匀分布$(\{-1, +1\}^{L\times K})$下），其论证类似于附录E.2中的论证。我们在以下引理中总结结果，为简洁起见，略去其证明。

引理 G.3（将学习与检验联系起来）。 有

$$V_M^\star - V_M(\pi) \geqslant \frac{\varepsilon}{4}\mathbb{P}_M^\pi(o_H = s_0) \quad \forall M \neq 0, \qquad \text{and} \qquad V_0^\star - V_0(\pi) \geqslant \frac{\varepsilon}{4}\mathbb{P}_0^\pi(o_H \neq s_0).$$

因此，假设算法 $\mathfrak{A}$ 输出一个策略 π^{out}，使得 $\mathbb{P}_M^{\mathfrak{A}}\big(V_M^\star - V_M(\pi^{\text{out}}) < \frac{\varepsilon}{8}\big) \geqslant \frac{3}{4}$对任意模型 $M \in \mathcal{M}$均成立，则有

$$D_{\text{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}, \mathbb{E}_{\mu\sim\text{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\right) \geqslant \frac{1}{2}, \qquad \forall \theta. \tag{35}$$

接下来，我们利用（35）推导某些良好事件的期望访问次数的下界，进而推出 T的下界，从而得到以下引理，其证明见附录 G.4。

引理 G.4。 固定一个 $\theta = (h^\star、\ s^\star、\ a^\star、\mathbf{a}^\star)$。我们考虑事件

$$E_{\text{reach}}^\theta := \left\{o_{h^\star} = s^\star, a_{h^\star:h^\star+m-1} = (a^\star, \mathbf{a}^\star_{h^\star+1:h^\star+m-1})\right\},$$
$$E_{\text{correct}}^\theta := \left\{o_{h^\star} = s^\star, a_{h^\star:H-1} = (a^\star, \mathbf{a}^\star)\right\}.$$

那么，对于任何满足 $\mathfrak{A}\ \delta := D_{\text{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}、\ \mathbb{E}_{\mu\sim\text{unif}}\left[\mathbb{P}_{\theta,\ \mu}^{\mathfrak{A}}\right]\right) > 0$ 的算法，我们有

$$either\ \mathbb{E}_0^{\mathfrak{A}}\left[N(E_{\text{reach}}^\theta)\right] \geqslant \frac{\delta^3\sqrt{LK}}{18\varepsilon^2\sigma^2} - \frac{\delta}{6},\ or\ \mathbb{E}_0^{\mathfrak{A}}\left[N(E_{\text{correct}}^\theta)\right] \geqslant \frac{\delta^3}{18\varepsilon^2} - \frac{\delta}{6}.$$

应用引理 G.4 对于任意参数元组 $\theta = (h^\star,\ s^\star,\ a^\star,$ **取**$^\star)$ 满足 $\delta = \frac{1}{2}$，可得–

$$\text{either } \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\text{reach}}^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}\right)\right] \geqslant \frac{\sqrt{LK}}{300\varepsilon^2\sigma^2}, \quad \text{or} \quad \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\text{correct}}^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}\right)\right] \geqslant \frac{1}{300\varepsilon^2}, \tag{36}$$

根据我们的选择 $\varepsilon \in (0,\ 0.1]$。

固定一个元组 $(h^\star,\ s^\star,\ a^\star)$，使得 $h^\star \in \mathcal{H}$ 且 $h^\star \leqslant n + m\lfloor H/10m\rfloor$, $s^\star \in \mathcal{S}_{\text{leaf}}$, $a^\star \in \mathcal{A}_{\text{co}}$。根据（36），我们知道，对于所有 $\mathbf{a}^\star \in \mathcal{A}_{\text{code},\ h^\star}$，都有

$$\begin{aligned} &A^{m-1}\cdot\mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\text{reach}}^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}\right)\right] + |\mathcal{A}_{\text{code},h^\star}|\cdot\mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\text{correct}}^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}\right)\right] \\ &\geqslant \frac{1}{300}\min\left\{\frac{A^{m-1}\sqrt{LK}}{\varepsilon^2\sigma^2}, \frac{|\mathcal{A}_{\text{code},h^\star}|}{\varepsilon^2}\right\} \geqslant \frac{1}{300}\min\left\{\frac{A^{m-1}\sqrt{LK}}{\varepsilon^2\sigma^2}, \frac{A^{H/2-1}}{\varepsilon^2}\right\} =: \omega, \end{aligned} \tag{37}$$

其中，最后一个不等式利用了这一事实：$|\mathcal{A}_{\mathsf{code},\ h^\star}| \geqslant A^{H/2-1}$ 对于 $h^\star \leqslant n+m\lfloor H/10m \rfloor,$，该结论可由直接计算得出（引理 F.7）。注意，根据我们对 E_{reach} 的定义，

$$\begin{aligned}
\sum_{\mathbf{a}^\star \in \mathcal{A}^{H-h^\star-1}} \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{reach}}^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}\right)\right] &= \sum_{\mathbf{a}^\star \in \mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}\left[N\left(o_{h^\star}=s^\star, a_{h^\star:h^\star+m-1}=(a^\star,\mathbf{a}^\star_{h^\star+1:h^\star+m-1})\right)\right] \\
&= \sum_{\mathbf{a}\in\mathcal{A}^{m-1}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star:h^\star+m-1}=(a^\star,\mathbf{a}))] \cdot \sum_{\substack{\mathbf{a}^\star\in\mathcal{A}_{\mathsf{code},h^\star} \\ \mathbf{a}^\star \text{ begins with } \mathbf{a}}} 1 \\
&= \sum_{\mathbf{a}\in\mathcal{A}^{m-1}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star:h^\star+m-1}=(a^\star,\mathbf{a}))] \cdot \frac{|\mathcal{A}_{\mathsf{code},h^\star}|}{A^{m-1}} \\
&= \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)] \cdot \frac{|\mathcal{A}_{\mathsf{code},h^\star}|}{A^{m-1}}.
\end{aligned}$$

同样，根据我们对 E_{correct} 的定义，有

$$\begin{aligned}
\sum_{\mathbf{a}^\star \in \mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}\right)\right] &= \sum_{\mathbf{a}^\star \in \mathcal{A}_{\mathsf{code},h^\star}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star:H-1}=(a^\star,\mathbf{a}^\star))] \\
&= \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star, a_{h^\star+1:H-1}\in\mathcal{A}_{\mathsf{code},h^\star})] \leqslant \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)].
\end{aligned}$$

因此，对所有 $\mathbf{a}\in\mathcal{A}_{\mathsf{code},h^\star}$，将 (37) 取平均，并利用上述等式，可得

$$\begin{aligned}
\omega \leqslant& \frac{1}{|\mathcal{A}_{\mathsf{code},h^\star}|} \sum_{\mathbf{a}\in\mathcal{A}_{\mathsf{code},h^\star}} \left[A^{m-1}\cdot \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{reach}}^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}\right)\right] + |\mathcal{A}_{\mathsf{code},h^\star}|\cdot \mathbb{E}_0^{\mathfrak{A}}\left[N\left(E_{\mathsf{correct}}^{(h^\star,s^\star,a^\star,\mathbf{a}^\star)}\right)\right]\right] \\
\leqslant& 2\ \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star, a_{h^\star}=a^\star)].
\end{aligned}$$

现在，我们已经证明，对于每个满足条件 $h^\star \leqslant n+m\lfloor H/10m \rfloor$ 的 $s^\star\in\mathcal{S}_{\mathsf{leaf}}, a^\star\in\mathcal{A}_c, h^\star\in\mathcal{H}$，都有 $\mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star,a_{h^\star}=a^\star)] \geqslant \frac{\omega}{2}$。对所有这样的 $(h^\star,s^\star,a^\star)$ 求和，可得

$$\frac{|\mathcal{S}_{\mathsf{leaf}}|\,|\mathcal{A}_c|\,(\lfloor H/10m\rfloor+1)}{600}\min\left\{\frac{A^{m-1}\sqrt{LK}}{\varepsilon^2\sigma^2}, \frac{A^{H/2-1}}{\varepsilon^2}\right\} \leqslant \sum_{s^\star\in\mathcal{S}_{\mathsf{leaf}}}\sum_{a^\star\in\mathcal{A}_c}\sum_{\substack{h^\star=n+lm: \\ 0\leqslant l\leqslant\lfloor H/10m\rfloor}} \mathbb{E}_0^{\mathfrak{A}}[N(o_{h^\star}=s^\star,a_{h^\star}=a^\star)] \leqslant T,$$

其中，第二个不等式成立是因为事件 $(\{o_{h^\star}=s^\star, a_{h^\star}=a^\star\})_{h^\star,s^\star,a^\star}$ 两两互斥。代入 $|\mathcal{A}_c| = A-1 \geqslant \frac{2}{3}A$，$\lfloor H/10m\rfloor+1 \geqslant H/10m$ 即可完成命题 G.1的证明。 □

引理G.2的证明

该证明与引理 F.3的证明非常相似，只需稍作修改。

情况 1： 我们首先证明零模型 0 是 1 步 1-可揭示的。在此模型中，$\{s_\oplus^1、\ e_\oplus^1、\ \cdots、\ s_\oplus^L、\ e_\oplus^L\}$中的状态均不可达，因此对于每个步骤 h，我们考虑集合 $\mathcal{S}' = \mathcal{S}_{\mathsf{tree}} \sqcup \bigsqcup_{j=1}^{L}\left\{s_\ominus^j、\ e_\ominus^j、\ 终止^j\right\}$。对于不同的状态 $s、\ s'\in\mathcal{S}'$，根据我们的构造，$\mathbb{O}_h(\cdot|s)$ 和 $\mathbb{O}_h(\cdot|s')$ 的支撑集互不相交，因此应用引理 F.9 可得

$$\min_{\mathbb{O}_h^+}\left\|\mathbb{O}_h^+\right\|_{1\to1} \leqslant \max_{s\in\mathcal{S}'}\gamma(\mathbb{O}_h(s)) \leqslant 1.$$

应用命题 2 即可完成零模型 0 的证明。

情况2： 接下来我们考虑模型$M = M_{\theta,\ \mu}\in\mathcal{M}$。根据我们的构造，对于 $h\leqslant h^\star$，$\{s_\oplus^1$中的状态、$e_\oplus^1、\ \cdots、\ s_\oplus^L、\ e_\oplus^L\}$ 中的状态均不可达，因此通过与零模型中相同的论证，我们得到

$$\min_{\mathbb{M}_{h,m+1}^+}\left\|\mathbb{M}_{h,m+1}^+\right\|_{*\to1} \leqslant \min_{\mathbb{O}_h^+}\left\|\mathbb{O}_h^+\right\|_{1\to1} \leqslant 1.$$

因此，我们只需界定数量 $\min_{\mathbb{M}_h^+,\, m+1} \left\|\mathbb{M}_h^+,\, m+1\right\|_{*\to 1}$，对于固定的步骤 $h > h^\star$。在此情况下，存在一个 $l \in \mathcal{H}$ 使得 $h \leqslant l \leqslant h+m-1$，我们记 $r = l-h+1$。根据引理 C.1，我们只需界定 $\min_{\mathbb{M}_h^+,\, r+1} \left\|\mathbb{M}_h^+,\, r+1\right\|_{*\to 1}$。考虑动作序列 $\mathbf{a} = (\mathbf{a}_{h:l-1}^\star,$ 揭示$) \in \mathcal{A}^r$，并将$\mathcal{S}$ 划分为

$$\mathcal{S} = \bigsqcup_{s \in \mathcal{S}_{\mathsf{tree}}} \{s\} \sqcup \bigsqcup_{j=1}^{L} \{s_\oplus^j, s_\ominus^j\} \sqcup \{e_\oplus^j, e_\ominus^j\} \sqcup \left\{\mathsf{terminal}^j\right\}.$$

不难直接验证，在 M_θ 中，μ对于状态 s，s' 来自上述划分中的不同子集，$\mathbb{M}_{h,\,\mathbf{a}}(\cdot|s)$ 和 $\mathbb{M}_{h\,\text{的}\mathbf{a}}(\cdot|s')$ 支撑集互不相交。于是，我们可以应用引理 F.8 和引理 F.9，从而得到

$$\min_{\mathbb{M}_{h,r+1}^+} \left\|\mathbb{M}_{h,r+1}^+\right\|_{*\to 1} \leqslant \min_{\mathbb{M}_{h,\mathbf{a}}^+} \left\|\mathbb{M}_{h,\mathbf{a}}^+\right\|_{1\to 1} \leqslant \max_j \left\{1, \gamma\Big(\mathbb{M}_{h,\mathbf{a}}(\{s_\oplus^j, s_\ominus^j\})\Big), \gamma\Big(\mathbb{M}_{h,\mathbf{a}}(\{e_\oplus^j, e_\ominus^j\})\Big)\right\}.$$

因此，接下来我们只需考虑矩阵 $\mathbb{M}_{h,\,\mathbf{a}}(\{s_\oplus^j,\ s_\ominus^j\})$ 和 $\mathbb{M}_{h,\,\mathbf{a}}(\{e_\oplus^j,\ e_\ominus^j\})$对于每个 $j \in [L]$ 的左逆。

（1）矩阵 $\mathbb{M}_{h,\,\mathbf{a}}(\{s_\oplus^j,\ s_\ominus^j\})$。根据我们的构造，在 $s_h = s_\oplus^j$ 处采取**动作** 将导致 $o_{h:l-1} =$ 锁、$o_l =$ 锁^{j}和 $o_{l+1} \sim \mathbb{O}_\mu(\cdot|e_\oplus^j)$；在 $s_h = s_\ominus^j$ 处采取**动作** 将导致 $o_{h:l-1} =$ 锁、$o_l =$ 锁^j 和 $o_{l+1} \sim \mathbb{O}_\mu(\cdot|e_\ominus^j)$。因此，$\mathbb{M}_{h,\,\mathbf{a}}(\{s_\oplus^j,\ s_\ominus^j\})$ 可以写成（对行进行置换后）

$$\mathbb{M}_{h,\mathbf{a}}(\{s_\oplus^j, s_\ominus^j\}) = \begin{bmatrix} \frac{\mathbb{1}_{2K} + \sigma\tilde{\mu}_j}{2K} & \frac{\mathbb{1}_{2K}}{2K} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{\mathcal{O}^{r+1} \times 2},$$

其中 $\tilde{\mu}_j = [\mu_j; -\mu_j] \in \{-1,\ 1\}^{2K}$，$\mathbb{1} = \mathbb{1}_{2K}$ 是 $\mathbb{R}^{2K}$ 中的一个向量，其所有分量均为 1。与命题 E.2 类似，我们可以直接验证 $\gamma(\mathbb{M}_{h,\,\mathbf{a}}(\{s_\oplus,\ s_\ominus\})) \leqslant \frac{2}{\sigma} + 1$。

(2) 矩阵 $\mathbb{M}_{h,\,\mathbf{a}}(\{e_\oplus^j,\ e_\ominus^j\})$。根据我们的构造，在 $s_h = e_\oplus^j$ 处，我们有 $o_h \sim \mathbb{O}_\mu(\cdot|e_\oplus^j)$ 和 $o_{h+1:l+1} =$终止^j；在 $s_h = e_\ominus$ 处，我们有 $o_h \sim \mathbb{O}_\mu(\cdot|e_\ominus^j)$ 和 $o_{h+1:l+1} =$ 终止^j。因此，$\mathbb{M}_{h,\,\mathbf{a}}(\{s_\oplus^j,\ s_\ominus^j\})$ 也可以写成

$$\mathbb{M}_{h,\mathbf{a}}(\{e_\oplus^j, e_\ominus^j\}) = \begin{bmatrix} \frac{\mathbb{1}_{2K} + \sigma\tilde{\mu}_j}{2K} & \frac{\mathbb{1}_{2K}}{2K} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{\mathcal{O}^{r+1} \times 2},$$

因此，我们也有 $\gamma\Big(\mathbb{M}_{h,\,\mathbf{a}}(\{e_\oplus^j,\ e_\ominus^j\})\Big) \leqslant \frac{2}{\sigma} + 1$。

合并上述两种情形可得

$$\min_{\mathbb{M}_{h,m+1}^+} \left\|\mathbb{M}_{h,m+1}^+\right\|_{*\to 1} \leqslant \min_{\mathbb{M}_{h,r+1}^+} \left\|\mathbb{M}_{h,r+1}^+\right\|_{*\to 1} \leqslant \min_{\mathbb{M}_{h,\mathbf{a}}^+} \left\|\mathbb{M}_{h,\mathbf{a}}^+\right\|_{1\to 1} \leqslant \frac{2}{\sigma} + 1,$$

由此完成引理 G.2的证明。 □

G.4 引理的证明G.4

与引理 F.6的证明类似，我们只需证明以下引理，而引理 G.4可通过归约论证得证（参见附录 F.6）。

引理 G.5。 假设算法 $\mathfrak{A}$（其随机停止时间可能为 T）几乎必然满足 $N(E_{\mathsf{reach}}^\theta) \leqslant N_o$和 $N(E_{\mathsf{correct}}^\theta) \leqslant N_r$，其中 N_o、N_r 为某些固定值。那么

$$either\ \overline{N}_o \geqslant \frac{3}{4}\frac{\delta^2\sqrt{LK}}{\varepsilon^2\sigma^2},\ or\ \overline{N}_r \geqslant \frac{3}{4}\frac{\delta^2}{\varepsilon^2},$$

其中 $\delta = D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}}, \mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_\theta^{\mathfrak{A},\mu}\right]\right)$。

证明。固定一个 $\theta=(h^\star,s^\star,a^\star,\mathbf{a}^\star)$。回忆一下，我们定义了 $E^\theta_{\mathsf{reach}}:=\{o_{h^\star}=s^\star,a_{h^\star:h^\star+m-1}=(a^\star,\mathbf{a}^\star_{h^\star+1:h^\star+m-1})\}$，并进一步定义

$$E^\theta_{\mathsf{rev}}:=\{o_{h^\star}=s^\star,a_{h^\star:h}=(a^\star,\mathbf{a}^\star_{h^\star+1:h-1},\mathsf{reveal})\text{ for some }h\in\mathcal{H}_{>h^\star}\}.$$

对于任意模型 $M\in\mathcal{M}_\theta:=\{M_{\theta,\ \mu}:\ \mu\in\{-1,\ +1\}^{L\times K}\}\cup\{0\},$，我们考虑以下"增广"系统动力学 $\mathbb{P}_M$：

1. 对于每个回合，在交互 $\tau_H\sim\mathbb{P}_M$ 结束后，环境会生成一个额外的观测$o_{H+1}=o^{\mathsf{aug}}$。

2. 如果 $\tau_H\notin E^\theta_{\mathsf{reach}}$ 或 $\tau_H\in E^\theta_{\mathsf{rev}}$，则 $o^{\mathsf{aug}}=$ 为虚设值。
3. 如果 $\tau_H\in E^\theta_{\mathsf{reach}}-E^\theta_{\mathsf{rev}}$，那么在 $\tau_H=(o_1,\ a_1,\ \cdots,\ o_H,\ a_H)$中，我们有 $o_{h^\star+m}=$锁^j，其中某个 $j\in[L]$满足条件，随后环境生成 o^{aug} 作为

$$M=M_{\theta,\mu}:\qquad \overline{\mathbb{P}}_{\theta,\mu}(o^{\mathsf{aug}}=o_i^+|\tau_H)=\frac{1+\varepsilon\sigma\mu_{j,i}}{2K},\qquad \overline{\mathbb{P}}_{\theta,\mu}(o^{\mathsf{aug}}=o_i^+|\tau_H)=\frac{1-\varepsilon\sigma\mu_{j,i}}{2K},\qquad \forall i\in[K],$$

并且对于 $M=0$，$\overline{\mathbb{P}_0}(o^{\mathsf{aug}}=\cdot|\tau_H)=$服从均匀分布$(\{o_1^+,\ o_1^-,\ \cdots,\ o_K^+,\ o_K^-\})$。

显然，对于每个 $M\in\mathcal{M}_\theta$，$\overline{\mathbb{P}}_M$ 仍是一个序贯决策过程。在这样的构造下，每个策略 π 都会诱导出一个关于 $\tau=(\tau_H,o^{\mathsf{aug}})\sim\mathbb{P}_M^\pi$ 的分布，而算法 $\mathfrak{A}$ 则会诱导出一个关于 $\tau^{(1)},\cdots,\tau^{(\mathsf{T})}\sim\overline{\mathbb{P}}_0^{\mathfrak{A}}$ 的分布。根据数据处理不等式，我们有

$$D_{\mathrm{TV}}\left(\mathbb{P}_0^{\mathfrak{A}},\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\mathbb{P}_{\theta,\mu}^{\mathfrak{A}}\right]\right)\leqslant D_{\mathrm{TV}}\left(\overline{\mathbb{P}}_0^{\mathfrak{A}},\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\overline{\mathbb{P}}_{\theta,\mu}^{\mathfrak{A}}\right]\right).$$

因此，根据引理D.1，我们只需对

$$1+\chi^2(\mathbb{E}_{\mu\sim\mathsf{unif}}\left[\overline{\mathbb{P}}_{\theta,\mu}^{\mathfrak{A}}\right]\|\,\overline{\mathbb{P}}_0^{\mathfrak{A}})=\mathbb{E}_{\mu,\mu'\sim\mathsf{unif}}\mathbb{E}_{\overline{\tau}^{(1)},\cdots,\overline{\tau}^{(\mathsf{T})}\sim\overline{\mathbb{P}}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau}^{(t)})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau}^{(t)})}{\overline{\mathbb{P}}_0(\overline{\tau}^{(t)})^2}\right].$$

为了进一步界定上述量，我们调用以下引理（证明见附录G.5）.

引理 G.6 (关于 χ^2内积的界)。在引理 *G.5* 的条件下（固定 θ），对任意 $\mu,\ \mu'\in\{-1,\ 1\}^K$,

$$\overline{\mathbb{E}}_0^{\mathfrak{A}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau}^{(t)})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau}^{(t)})}{\overline{\mathbb{P}}_0(\overline{\tau}^{(t)})^2}\right]\leqslant\exp\left(\overline{N}_o\cdot\frac{\sigma^2\varepsilon^2}{LK}\left|\langle\mu,\mu'\rangle\right|+\frac{4}{3}\varepsilon^2\overline{N}_r\right).\tag{38}$$

由引理 G.6，通过标准论证即可得到所需结果（例如，参见引理 F.11的证明）。

□

G.5 引理的证明G.6

我们首先证明以下引理，它是引理 G.6的单回合版本。

引理 G.7. 对于任意策略 π和参数 θ,μ,μ'，均有

$$\mathbb{E}_{\overline{\tau}\sim\overline{\mathbb{P}}_0^\pi}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau})}{\overline{\mathbb{P}}_0(\overline{\tau})^2}\exp\left(-\mathbb{1}(\overline{\tau}\in E^\theta_{\mathsf{reach}})\cdot\frac{\varepsilon^2\sigma^2}{LK}\langle\mu,\mu'\rangle-\mathbb{1}(\overline{\tau}\in E^\theta_{\mathsf{correct}})\cdot\frac{4}{3}\varepsilon^2\right)\right]\leqslant 1.\tag{39}$$

引理 *G.7* 的证明。下文中的所有期望和条件期望均相对于$\tau=(\tau_H,\ o^{\mathsf{aug}})\sim\mathbb{P}_0^\pi$。

与引理 F.12（见附录 F.7）的证明类似，我们分析的核心仍是计算以下量$I(\tau_l)$，其定义为

$$I(\tau_l) := \mathbb{E}_{o_{l+1}\sim\overline{\mathbb{P}}_0(\tau_l)}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(o_{l+1}|\tau_l)\overline{\mathbb{P}}_{\theta,\mu'}(o_{l+1}|\tau_l)}{\overline{\mathbb{P}}_0(o_{l+1}|\tau_l)^2}\right] = \begin{cases} \mathbb{E}\left[\dfrac{\mathbb{P}_{\theta,\mu}(o_{l+1}|\tau_l)\mathbb{P}_{\theta,\mu'}(o_{l+1}|\tau_l)}{\mathbb{P}_0(o_{l+1}|\tau_l)^2}\middle|\tau_l\right], & l < H, \\ \mathbb{E}_{o^{\mathsf{aug}}\sim\mathbb{P}_0(\tau_H)}\left[\dfrac{\overline{\mathbb{P}}_{\theta,\mu}(o^{\mathsf{aug}}|\tau_H)\overline{\mathbb{P}}_{\theta,\mu'}(o^{\mathsf{aug}}|\tau_H)}{\overline{\mathbb{P}}_0(o^{\mathsf{aug}}|\tau_H)^2}\right], & l = H. \end{cases} \tag{40}$$

基本上，根据引理 A.2，我们有

$$1 = \mathbb{E}_{\overline{\tau}\sim\overline{\mathbb{P}}_0^{\pi}}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau}|\tau_{h^\star+m})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau}|\tau_{h^\star+m})}{\overline{\mathbb{P}}_0(\overline{\tau}|\tau_{h^\star+m})^2}\cdot\exp\left(-\sum_{l=h^\star+m}^{H}\log I(\tau_l)\right)\middle|\tau_{h^\star+m}\right]. \tag{41}$$

下文中，我们首先计算每个（可达的） τ_l对应的 $I(\tau_l)$ 。

一个重要的观测是，对于轨迹 $\tau_l = (o_1,\ a_1,\ \cdots,\ o_l,\ a_l)$，其中 $l < H$，如果 $\mathbb{P}_{\theta,\ \mu}(o_{l+1} = \cdot|\tau_l) \neq \mathbb{P}_0(o_{l+1} = \cdot|\tau_l)$，那么

1. 显然， $o_{h^\star} = s^\star$、 $a_{h^\star} = a^\star$ （即， $l \geqslant h^\star + 1$ ，并从 s_0 采取动作 $a_{1:h^\star-1}$ 将在步骤 $h^\star$导致 $s^\star$ ）。2. 以下两种情况之一成立： $a_{h^\star+1:\ l} = (\mathbf{a}^\star_{h^\star+1:\ l-1}$，对于某个 $l \in \mathcal{H}_{>h^\star}$，揭示）；或者 $l = H-1$ 且 $a_{h^\star+1:\ H-1} = \mathbf{a}^\star$。

因此，对于每个 $l \in \mathcal{H}_{>h^\star}$ ，我们定义

$$E^\theta_{\mathsf{rev},l} := \left\{o_{h^\star} = s^\star, a_{h^\star:l} = (a^\star, \mathbf{a}^\star_{h^\star+1:l-1}, \mathsf{reveal})\right\}.$$

那么，如果 $\mathbb{P}_{\theta,\ \mu}(\cdot|\tau_l) \neq \mathbb{P}_0(\cdot|\tau_l)$，则要么是（情况 1）$l \in \mathcal{H}_{>h^\star}$， $\tau_{H-1} \in E^\theta_{\mathsf{rev},\ l}$，要么是（情况2） $l = H-1$， $\tau_l \in E^\theta_{\mathsf{correct}}$，要么是（情况 3） $l = H$， $\tau_H \in E^\theta_{\mathsf{reach}} - E^\theta_{\mathsf{rev}^\circ}$

接下来，我们分别计算这三种情况下的 $I(\tau_l)$。为简化讨论，我们考虑事件 $\mathcal{L}_j := \left\{o_{h^\star+m} = \text{锁}^j\right\}(j \in [L])$ 。

情况 1： $l \in \mathcal{H}_{>h^\star}$， $\tau_l \in E^\theta_{\mathsf{rev},\ l}$。在这种情况下，存在一个 $j \in [L]$ ，使得 $o_{h^\star+m} = \text{锁}^j$，即 $\tau_l \in \mathcal{L}_j$。换句话说，观察到 τ_l 意味着，对于 $h < h' \leqslant l$， $s_{h'} \in \{s^j_\oplus,\ s^j_\ominus\}$ ，并且 $s_{l+1} \in \{e^j_\oplus,\ e^j_\ominus\}$ （因为 $a_l =$ 揭示）。因此，

$$\begin{aligned}\mathbb{P}_{\theta,\mu}(o_{l+1} = o|\tau_l) &= \mathbb{P}_{\theta,\mu}(o_{l+1} = o|s_{l+1} = e^j_\oplus)\mathbb{P}_{\theta,\mu}(s_{l+1} = e^j_\oplus|\tau_l) + \mathbb{P}_{\theta,\mu}(o_{l+1} = o|s_{l+1} = e^j_\ominus)\mathbb{P}_{\theta,\mu}(s_{l+1} = e^j_\ominus|\tau_l) \\ &= \left(\mathbb{O}_\mu(o|e^j_\oplus) - \mathbb{O}(o|e^j_\ominus)\right)\cdot\mathbb{P}_{\theta,\mu}(s_{l+1} = e^j_\oplus|\tau_l) + \mathbb{O}(o|e^j_\ominus),\end{aligned}$$

请注意，根据我们的构造，

$$\begin{aligned}\mathbb{P}_{\theta,\mu}(s_{l+1} = e^j_\oplus|\tau_l) &= \mathbb{P}_{\theta,\mu}(s_l = s^j_\oplus|\tau_{l-1}, o_l) \\ &= \mathbb{P}_{\theta,\mu}(s_l = s^j_\oplus|o_{h^\star} = s^\star, a_{h^\star:l-1} = (a^\star, \mathbf{a}^\star_{h^\star+1:l-1}), \mathcal{L}_j) \\ &= \mathbb{P}_{\theta,\mu}(s_{h^\star+1} = s^j_\oplus|o_{h^\star} = s^\star, a_{h^\star:l-1} = (a^\star, \mathbf{a}^\star_{h^\star+1:l-1}), \mathcal{L}_j) \\ &= \mathbb{P}_{\theta,\mu}(s_{h^\star+1} = s^j_\oplus|o_{h^\star} = s^\star, a_{h^\star:l-1} = a^\star, s_{h^\star+1} \in \{s^j_\oplus, s^j_\ominus\}) \\ &= \varepsilon,\end{aligned}$$

其中，第一个等号成立是因为 $s_{l+1} = e^j_\oplus$ 当且仅当 $s_l = s^j_\oplus$， $a_l =$ 揭示；第二个不等式成立是因为在 $o_{h^\star+1:\ l}$ 中只有锁和锁^j ；第三个等号成立是因为 $s_l = s^j_\oplus$ 当且仅当$s_{h^\star+1} = s^j_\oplus$， $a_{h^\star+1:\ l-1} =$ 一个$^\star_{h^\star+1:\ l-1}$。将上述等式与我们的 $\mathbb{O}_\mu(\cdot|e^j_\oplus)$ 和$\mathbb{O}(\cdot|e^j_\ominus)$ 定义结合起来可得

$$\mathbb{P}_{\theta,\mu}(o_{l+1} = o_i^+|\tau_l) = \frac{1+\varepsilon\sigma\mu_{j,i}}{2K}, \qquad \mathbb{P}_{\theta,\mu}(o_{l+1} = o_i^-|\tau_l) = \frac{1-\varepsilon\sigma\mu_{j,i}}{2K}, \qquad \forall i \in [K].$$

另一方面，显然 $\mathbb{P}_0(o_{l+1}=\cdot|\tau_l)=$均匀分布$(\{o_1^+,\ o_1^-,\ \cdots,\ o_K^+,\ o_K^-\})$。因此，有

$$I(\tau_l)=\frac{1}{2K}\sum_{o\in\mathcal{O}_o}\frac{\mathbb{P}_{\theta,\mu}(o_{l+1}=o|\tau_l)\mathbb{P}_{\theta,\mu'}(o_{l+1}=o|\tau_l)}{\mathbb{P}_0(o_{l+1}=o|\tau_l)^2}=1+\frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K\mu_{j,i}\mu'_{j,i},\quad \text{for any } \tau_l\in E^\theta_{\mathsf{rev},l}\cap\mathcal{L}_j. \tag{42}$$

情况2： $l=H-1,\tau_{H-1}\in E^\theta_{\mathsf{correct}}$。在这种情况下，通过与引理 F.12（附录 F.7，情况2）的证明完全相同的计算，我们可以得到

$$I(\tau_{H-1})=1+\frac{4}{3}\varepsilon^2,\quad \text{for any } \tau_{H-1}\in E^\theta_{\mathsf{correct}}. \tag{43}$$

情况3： $l=H,\tau_H\in E^\theta_{\mathsf{reach}}-E^\theta_{\mathsf{rev}}$。假设对于某些 $j\in[L]$, $\tau_H\in\left(E^\theta_{\mathsf{reach}}-E^\theta_{\mathsf{rev}}\right)\cap\mathcal{L}_j$，那么根据我们对 $\mathbb{P}$ 的构造，有

$$I(\tau_H)=1+\frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K\mu_{j,i}\mu'_{j,i},\quad \text{for any } \tau_H\in\left(E^\theta_{\mathsf{reach}}-E^\theta_{\mathsf{rev}}\right)\cap\mathcal{L}_j. \tag{44}$$

结合（42）（43）（44）可知，对于任意以 τ_H 开头的 $\tau_{h^\star+m}\in E^\theta_{\mathsf{reach}}\cap\mathcal{L}_j$，

$$\begin{aligned}
\sum_{l=h^\star+m}^H\log I(\tau_l)=&\sum_{l\in\mathcal{H}_{>h^\star}}\mathbb{1}\left(\tau_l\in E^\theta_{\mathsf{rev},l}\right)\cdot\log\left(1+\frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K\mu_{j,i}\mu'_{j,i}\right)+\mathbb{1}\left(\tau_{H-1}\in E^\theta_{\mathsf{correct}}\right)\cdot\log\left(1+\frac{4}{3}\varepsilon^2\right)\\
&+\mathbb{1}\left(\tau_H\in E^\theta_{\mathsf{reach}}-E^\theta_{\mathsf{rev}}\right)\cdot\log\left(1+\frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K\mu_{j,i}\mu'_{j,i}\right)\\
=&\,\mathbb{1}\left(\tau_H\in E^\theta_{\mathsf{rev}}\right)\cdot\log\left(1+\frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K\mu_{j,i}\mu'_{j,i}\right)+\mathbb{1}\left(\tau_H\in E^\theta_{\mathsf{correct}}\right)\cdot\log\left(1+\frac{4}{3}\varepsilon^2\right)\\
&+\mathbb{1}\left(\tau_H\in E^\theta_{\mathsf{reach}}-E^\theta_{\mathsf{rev}}\right)\cdot\log\left(1+\frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K\mu_{j,i}\mu'_{j,i}\right)\\
=&\,\mathbb{1}\left(\tau_H\in E^\theta_{\mathsf{reach}}\right)\cdot\log\left(1+\frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K\mu_{j,i}\mu'_{j,i}\right)+\mathbb{1}\left(\tau_H\in E^\theta_{\mathsf{correct}}\right)\cdot\log\left(1+\frac{4}{3}\varepsilon^2\right),
\end{aligned}$$

其中，第二个等号成立是因为 $E^\theta_{\mathsf{rev}}=\bigsqcup_l E^\theta_{\mathsf{rev},l}$。我们还有，对于 $\tau_{h^\star+m}\notin E^\theta_{\mathsf{reach}}$, $\sum_{l=h^\star+m}^H\log I(\tau_l)=0$。将 $\sum_{l=h^\star+m}^H\log I(\tau_l)$ 的值代入（41），并利用根据我们的构造有 $\mathbb{P}_{\theta,\mu}(\tau_{h^\star+m})=\mathbb{P}_0(\tau_{h^\star+m})$ 这一事实，可得

$$\mathbb{E}_{\overline{\tau}\sim\overline{\mathbb{P}}_0^\pi}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau})}{\overline{\mathbb{P}}_0(\overline{\tau})^2}\cdot\left(1+\mathbb{1}(\tau_H\in E^\theta_{\mathsf{correct}})\cdot\frac{4}{3}\varepsilon^2\right)^{-1}\middle|\tau_{h^\star+m}\right]=\begin{cases}1+\frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K\mu_{j,i}\mu'_{j,i}, & \text{for } \tau_{h^\star+m}\in E^\theta_{\mathsf{reach}}\cap\mathcal{L}_j,\\ 1, & \text{if } \tau_{h^\star+m}\notin E^\theta_{\mathsf{reach}}.\end{cases}$$

注意，在 $\mathbb{P}_0$ 中，给定 $\tau_{h^\star+m-1}\in E^\theta_{\mathsf{reach}}$, $o_{h^\star+m}$ 在 $\left\{\text{锁}^1、\ \cdots、\ \text{锁}^L\right\}$ 上服从均匀分布，因此

$$\begin{aligned}
&\mathbb{E}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau})}{\overline{\mathbb{P}}_0(\overline{\tau})^2}\cdot\left(1+\mathbb{1}(\tau_H\in E^\theta_{\mathsf{correct}})\cdot\frac{4}{3}\varepsilon^2\right)^{-1}\middle|\tau_{h^\star+m-1}\right]\\
=&\,\mathbb{E}\left[\mathbb{E}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau})}{\overline{\mathbb{P}}_0(\overline{\tau})^2}\cdot\left(1+\mathbb{1}(\tau_H\in E^\theta_{\mathsf{correct}})\cdot\frac{4}{3}\varepsilon^2\right)^{-1}\middle|\tau_{h^\star+m}\right]\middle|\tau_{h^\star+m-1}\right]\\
=&\,\frac{1}{L}\sum_{j=1}^L\left(1+\frac{\varepsilon^2\sigma^2}{K}\sum_{i=1}^K\mu_{j,i}\mu'_{j,i}\right)=1+\frac{\varepsilon^2\sigma^2}{LK}\left\langle\mu,\mu'\right\rangle.
\end{aligned}$$

另一方面，对于 $\tau_{h^\star+m-1} \notin E^\theta_{\mathsf{reach}}$，

$$\mathbb{E}_{\overline{\tau}\sim\overline{\mathbb{P}}_0^\pi}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau})}{\overline{\mathbb{P}}_0(\overline{\tau})^2}\cdot\left(1+\mathbb{1}(\overline{\tau}\in E^\theta_{\mathsf{correct}})\cdot\frac{4}{3}\varepsilon^2\right)^{-1}\middle|\,\tau_{h^\star+m-1}\right]=1.$$

因此，对 $\tau_{h^\star+m-1}$ 取期望可得

$$\mathbb{E}_{\overline{\tau}\sim\overline{\mathbb{P}}_0^\pi}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau})}{\overline{\mathbb{P}}_0(\overline{\tau})^2}\cdot\left(1+\mathbb{1}(\overline{\tau}\in E^\theta_{\mathsf{correct}})\cdot\frac{4}{3}\varepsilon^2\right)^{-1}\cdot\left(1+\mathbb{1}(\overline{\tau}\in E^\theta_{\mathsf{reach}})\cdot\frac{\varepsilon^2\sigma^2}{LK}\left\langle\mu,\mu'\right\rangle\right)^{-1}\right]=1.$$

利用 $(1+x)^{-1}\geqslant\exp(-x)$ 这一事实即可完成证明。 □

引理 G.7 证明完毕后，我们继续证明引理 G.6。将引理 G.7 应用于算法 $\mathfrak{A}$，可得对每个 $t\in[\mathsf{T}]$,

$$\mathbb{E}_{\overline{\tau}^{(1)},\cdots,\overline{\tau}^{(\mathsf{T})}\sim\overline{\mathbb{P}}_0^{\mathfrak{A}}}\left[\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau}^{(t)})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau}^{(t)})}{\overline{\mathbb{P}}_0(\overline{\tau}^{(t)})^2}\cdot\exp\left(-\mathbb{1}(\overline{\tau}^{(t)}\in E^\theta_{\mathsf{reach}})\cdot\frac{\varepsilon^2\sigma^2}{LK}\left\langle\mu,\mu'\right\rangle-\mathbb{1}(\overline{\tau}^{(t)}\in E^\theta_{\mathsf{correct}})\cdot\frac{4}{3}\varepsilon^2\right)\middle|\,\overline{\tau}^{(1:t-1)}\right]\leqslant 1.$$

因此，根据鞅性质，有

$$\begin{aligned}
1&\geqslant\mathbb{E}_{\overline{\tau}^{(1)},\cdots,\overline{\tau}^{(\mathsf{T})}\sim\overline{\mathbb{P}}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau}^{(t)})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau}^{(t)})}{\overline{\mathbb{P}}_0(\overline{\tau}^{(t)})^2}\cdot\exp\left(-\mathbb{1}(\overline{\tau}^{(t)}\in E^\theta_{\mathsf{reach}})\cdot\frac{\varepsilon^2\sigma^2}{LK}\left\langle\mu,\mu'\right\rangle-\mathbb{1}(\overline{\tau}^{(t)}\in E^\theta_{\mathsf{correct}})\cdot\frac{4}{3}\varepsilon^2\right)\right]\\
&=\mathbb{E}_{\overline{\tau}^{(1)},\cdots,\overline{\tau}^{(\mathsf{T})}\sim\overline{\mathbb{P}}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau}^{(t)})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau}^{(t)})}{\overline{\mathbb{P}}_0(\overline{\tau}^{(t)})^2}\times\exp\left(-\sum_{t=1}^{T}\mathbb{1}(\overline{\tau}^{(t)}\in E^\theta_{\mathsf{reach}})\cdot\frac{\varepsilon^2\sigma^2}{LK}\left\langle\mu,\mu'\right\rangle-\sum_{t=1}^{T}\mathbb{1}(\overline{\tau}^{(t)}\in E^\theta_{\mathsf{correct}})\cdot\frac{4}{3}\varepsilon^2\right)\right]\\
&=\mathbb{E}_{\overline{\tau}^{(1)},\cdots,\overline{\tau}^{(\mathsf{T})}\sim\overline{\mathbb{P}}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau}^{(t)})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau}^{(t)})}{\overline{\mathbb{P}}_0(\overline{\tau}^{(t)})^2}\times\exp\left(-N(E^\theta_{\mathsf{reach}})\cdot\frac{\varepsilon^2\sigma^2}{LK}\left\langle\mu,\mu'\right\rangle-N(E^\theta_{\mathsf{correct}})\cdot\frac{4}{3}\varepsilon^2\right)\right]\\
&\geqslant\mathbb{E}_{\overline{\tau}^{(1)},\cdots,\overline{\tau}^{(\mathsf{T})}\sim\overline{\mathbb{P}}_0^{\mathfrak{A}}}\left[\prod_{t=1}^{\mathsf{T}}\frac{\overline{\mathbb{P}}_{\theta,\mu}(\overline{\tau}^{(t)})\overline{\mathbb{P}}_{\theta,\mu'}(\overline{\tau}^{(t)})}{\overline{\mathbb{P}}_0(\overline{\tau}^{(t)})^2}\times\exp\left(-\overline{N}_o\cdot\frac{\varepsilon^2\sigma^2}{LK}\left|\left\langle\mu,\mu'\right\rangle\right|-\overline{N}_r\cdot\frac{4}{3}\varepsilon^2\right)\right].
\end{aligned}$$

两边相乘 乘以 $\exp\left(N_o\cdot\frac{\sigma^2\varepsilon^2}{LK}\left|\langle\mu,\ \mu'\rangle\right|+\frac{4}{3}\varepsilon^2N_r\right)$ 即可完成引理 G.6 的证明 . □

H 单步揭示型部分可观测马尔可夫决策过程的遗憾

在本节中，我们将定理 8 推广到一类更广泛的序贯决策问题，称为强 B 稳定预测状态表示，并据此推导出单步揭示型部分可观测马尔可夫决策过程这一特殊情形下的保证。证明与 B 稳定预测状态表示的 PAC 学习分析（Chen 等人，2022a）基本平行，我们沿用其中的记号：下文用 θ 表示预测状态表示模型，用 Θ 表示预测状态表示模型类。

H.1 强 B 稳定预测状态表示

我们回顾附录 B 中预测状态表示和 B 稳定性的定义。为了建立学习预测状态表示的 $\sqrt{T}$遗憾上界，我们引入以下结构条件。

定义 H.1（强 B-稳定性）。若一个预测状态表示具有某种 B-表示，使得对所有步骤 $h\in[H]$和策略 π, $x\in\mathbb{R}^{\mathcal{U}_h}$ 均成立，则称其具有参数 $\Lambda_{\mathsf{B}}\geqslant 1$ 的强 B-稳定性（以下简称为Λ_{B}-强稳定）。

$$\sum_{\tau_{h:H}}\pi(\tau_{h:H})\times\left|\mathbf{B}_{H:h}(\tau_{h:H})x\right|\leqslant\Lambda_{\mathsf{B}}\sum_{t_h\in\mathcal{U}_h}\pi(t_h)\times\left|x(t_h)\right|. \tag{45}$$

为简化记号，从现在起我们假设，对于每个步骤 h, $\mathcal{U}_h=(\mathcal{O}\times\mathcal{A})^{m_h-1}\times\mathcal{O}$，对于某个$m_h\in\mathbb{Z}_{\geqslant 1}$，并定义 $\Omega_h:=(\mathcal{O}\times\mathcal{A})^{m_h-1}$；对于任意一般的 $\mathcal{U}_h$，我们的结果也成立，只需使用稍微复杂一些的记号。

算法 1 Optimistic Maximum Likelihood Estimation (OMLE) (Liu et al., 2022a; Chenet al., 2022a)

1: **输入：** 模型类 Θ，参数 $\beta > 0$。2: **初始化：** $\Theta^1 = \Theta$, $\mathcal{D} = \{\}$。
3: **对于**迭代 $k = 1, \ldots, T$ **执行**4: 设置 $(\theta^k, \pi^k) = \arg\max_{\theta \in \Theta^k, \pi} V_\theta(\pi)$。
5: 执行 π^k 以收集一条轨迹 τ^k，并将 (π^k, τ^k) 加入 $\mathcal{D}$。6: 更新置信集

$$\Theta^{k+1} = \left\{\widehat{\theta} \in \Theta : \sum_{(\pi,\tau)\in\mathcal{D}} \log \mathbb{P}^\pi_{\widehat{\theta}}(\tau) \geqslant \max_{\theta\in\Theta} \sum_{(\pi,\tau)\in\mathcal{D}} \log \mathbb{P}^\pi_\theta(\tau) - \beta\right\}.$$

命题 H.2 (强 B 稳定预测状态表示的误差分解)。 假设两个 *PSR* 模型 $\theta, \bar{\theta}$ 分别具有 $\{\{\mathbf{B}^\theta_h(o_h, a_h)\}_{h,o_h,a_h}, \mathbf{q}^\theta_0\}$ 和 $\{\{\mathbf{B}^{\bar{\theta}}_h(o_h, a_h)\}_{h,o_h,a_h}, \mathbf{q}^{\bar{\theta}}_0\}$ 作为 *B*-表示。定义

$$\mathcal{E}^{\bar{\theta}}_{\theta,h}(\pi, \tau_{h-1}) := \frac{1}{2} \max_{\pi' \in \Pi_h(\pi)} \sum_{\tau_{h:H}} \pi'(\tau_{h:H}|\tau_{h-1}) \times \left|\mathbf{B}^\theta_{H:h+1}(\tau_{h+1:H}) \left(\mathbf{B}^\theta_h(o_h, a_h) - \mathbf{B}^{\bar{\theta}}_h(o_h, a_h)\right) \mathbf{q}^{\bar{\theta}}(\tau_{h-1})\right|,$$

$$\mathcal{E}^{\bar{\theta}}_{\theta,0}(\pi) := \frac{1}{2} \max_{\pi' \in \Pi_0(\pi)} \sum_{\tau_{1:H}} \pi'(\tau_{1:H}) \times \left|\mathbf{B}^\theta_{H:1}(\tau_{1:H}) \left(\mathbf{q}^\theta_0 - \mathbf{q}^{\bar{\theta}}_0\right)\right|,$$

其中我们定义

$$\Pi_h(\pi) := \left\{\pi' : \pi'|_{\mathcal{O}\times\mathcal{A}\times\Omega_{h+1}} = \pi|_{\mathcal{O}\times\mathcal{A}\times\Omega_{h+1}}\right\}, \qquad \Pi_0(\pi) := \left\{\pi' : \pi'|_{\Omega_1} = \pi|_{\Omega_1}\right\},$$

即， $\Pi_h(\pi)$ 是所有满足以下条件的策略 π' 的集合：对于所有 $(o_h, a_h, \omega_{h+1}) \in \mathcal{O} \times \mathcal{A} \times \Omega_{h+1}$, $\pi'(o_h, a_h, \omega_{h+1}|\tau_{h-1}) = \pi(o_h, a_h, \omega_{h+1}|\tau_{h-1})$。
则以下结论成立。

1.（性能分解）有

$$D_{\mathrm{TV}}\left(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_{\bar{\theta}}\right) \leqslant \mathcal{E}^{\bar{\theta}}_{\theta,0}(\pi) + \sum_{h=1}^{H} \mathbb{E}^\pi_{\bar{\theta}}\left[\mathcal{E}^{\bar{\theta}}_{\theta,h}(\pi, \tau_{h-1})\right],$$

其中，对于 $h \in [H]$，期望 $\mathbb{E}^\pi_{\bar{\theta}}$ 是在模型 θ 和策略 π下对 τ_{h-1} 取的。

2.（用海灵格距离界定误差）假设 θ 是 Λ_{B}-强稳定的，且 $\{\{\mathbf{B}^\theta_h(o_h, a_h)\}_{h,o_h,a_h}, \mathbf{q}^\theta_0\}$满足稳定性条件 (45)。对于任意步骤 h，策略 π，有

$$\mathbb{E}^\pi_{\bar{\theta}}\left[\mathcal{E}^{\bar{\theta}}_{\theta,h}(\pi, \tau_{h-1})^2\right] \leqslant 2\Lambda^2_{\mathsf{B}} D^2_{\mathrm{H}}\left(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_{\bar{\theta}}\right).$$

以及 $(\mathcal{E}^{\bar{\theta}}_{\theta,0}(\pi))^2 \leqslant \Lambda^2_{\mathsf{B}} D^2_{\mathrm{H}}\left(\mathbb{P}^\pi_\theta, \mathbb{P}^\pi_{\bar{\theta}}\right)$。

H.2 算法与保证

在本节中，我们陈述 $\sqrt{T}$算法的遗憾保证 OMLE （算法1，（Liu 等人，2022a；Chen 等人，2022a））。其证明见附录H.4，该证明改编自对（探索型） OMLE 算法的分析，见Chen 等人（2022a）。我们还指出，定理OMLE 中的遗憾上界8也可以直接针对单步揭示型部分可观测马尔可夫决策过程得到，只需利用Liu 等人（2022a）的思想加强Chen 等人（2022a）中的分析。

定理 H.3。 假设每个 $\theta \in \Theta$ 都具有 Λ_{B}-强稳定性（定义 *H.1*），且秩为 $\theta^\star \in \Theta$ 的真实模型 $d_{\mathsf{PSR}} \leqslant d$。那么，对于某个绝对常数 $\beta = C$，取 $(\mathcal{N}_\Theta(1/T))/\delta)$ *log* $C > 0$，其概率为

至少为 $1-\delta$时，算法 *1* 可达到

$$\sum_{t=1}^{T} V^{\star} - V_{\theta^{\star}}(\pi^t) \leqslant \mathcal{O}\left(\sqrt{\Lambda_{\mathsf{B}}^2 OAU_{\mathcal{T}} dH^2 \iota \beta T}\right) \tag{46}$$

其中 $U_{\mathcal{T}} := \max_h |\Omega_h|$， $\iota := \log\left(1 + TdOAU_{\mathcal{T}}\Lambda_{\mathsf{B}} R_{\mathsf{B}}\right)$ 其中 $R_{\mathsf{B}} := 1 + \max_{h,\ o,\ a} \|\mathbf{B}_h(o,\ a)\|_1$。

采用与 Chen 等人 （2022a，附录 G）完全平行的分析，我们可以证明 E2D-TA （Chen 等人，2022b）和 MOPS （Chen 等人，2022a，算法 4）均可达到与定理 8相同的遗憾保证。

定理 H.4。 设 Θ 是一个具有相同核心测试集 $\{\mathcal{U}_h\}_{h\in[H]}$的 *PSR* 类，且每个 $\theta \in \Theta$ 都具有一个 Λ_{B}-强稳定的 *B*-表示（参见定义 *H.1*），并且 *PSR* 秩为 $d_{\mathsf{PSR}} \leqslant d$。那么，对于 dec 和 psc 这两个由 *Chen* 等人（*2022b*）引入的系数，有

$$\overline{\mathrm{dec}}_{\gamma}(\Theta) \leqslant \mathcal{O}\left(\frac{\Lambda_{\mathsf{B}}^2 OAU_{\mathcal{T}} dH^2}{\gamma}\right), \qquad \mathrm{psc}_{\gamma}(\Theta) \leqslant \mathcal{O}\left(\frac{\Lambda_{\mathsf{B}}^2 OAU_{\mathcal{T}} dH^2}{\gamma}\right).$$

因此，我们可以应用 *Chen* 等人 （*2022b*，定理 *D.1*）（对于 MOPS）以及 *Chen* 等人（*2022b*，定理 *C.7*）（对于 E2D-TA）来证明，在适当选择参数后， MOPS 和 E2D-TA 均可达到遗憾值

$$\mathbf{Regret} \leqslant \mathcal{O}\left(\sqrt{\Lambda_{\mathsf{B}}^2 OAU_{\mathcal{T}} dH^2 \log(\mathcal{N}_{\Theta}(1/T))/\delta)T}\right), \tag{47}$$

概率至少为 $1-\delta$。

定理 *8* 的证明。为应用定理 H.3，我们首先注意到，命题 C.2 直接推出，任何单步 α-揭示型模型都是强 B-稳定 PSR，且 $\Lambda_{\mathsf{B}} \leqslant \alpha^{-1}$ 对于所有 $\mathcal{U}_h = \mathcal{O}$ 和核心测试集 h，均有 H.3 因此，应用定理 $\mathcal{M}$ 可知，对于由单步 α-揭示型部分可观测马尔可夫决策过程构成的模型类，OMLE 可实现的遗憾为

$$\mathbf{Regret} \leqslant \widetilde{\mathcal{O}}\left(\sqrt{\alpha^{-2} SOAH^2 \log \mathcal{N}_{\mathcal{M}}(1/T) \cdot T}\right),$$

当 $U_{\mathcal{T}} = 1$， $d \leqslant S$， $\Lambda_{\mathsf{B}} \leqslant \alpha^{-1}$， $R_{\mathsf{B}} \leqslant \alpha^{-1}$ 以及 $\iota = \widetilde{\mathcal{O}}(1)$ 时，同理， E2D-TA 和 MOPS 也能达到相同的遗憾上界。注意到 $\log \mathcal{N}_{\mathcal{M}}(1/T) = \widetilde{\mathcal{O}}\big(H(S^2A + SO)\big)$ (Chen 等人， 2022a) 即完成证明。

□

H.3 命题 H.2 的证明

断言 1 可直接由 Chen 等人的(2022a引理 D.1 的证明得出。接下来，我们证明断言 2。

固定一个步骤 $h \in [H]$。一个重要的观测是，根据 Λ_{B} 的强 θ (定义 H.1)，对于任意 $\pi' \in \Pi_h(\pi)$，我们有 $\forall x \in \mathbb{R}^{\mathcal{U}_h}$

$$\sum_{\tau_{h:H}} \pi'(\tau_{h:H}|\tau_{h-1}) \times \left|\mathbf{B}^{\theta}_{H:h}(\tau_{h:H})x\right| \leqslant \Lambda_{\mathsf{B}} \sum_{t_h \in \mathcal{U}_h} \pi'(t_h|\tau_{h-1}) \times |x(t_h)| = \Lambda_{\mathsf{B}} \sum_{t_h \in \mathcal{U}_h} \pi(t_h|\tau_{h-1}) \times |x(t_h)|, \tag{48}$$

类似地，对于 $\forall x \in \mathbb{R}^{\mathcal{U}_{h+1}}$,

$$\sum_{\tau_{h+1:H}} \pi'(\tau_{h+1:H}|\tau_h) \times \left|\mathbf{B}^{\theta}_{H:h+1}(\tau_{h+1:H})x\right| \leqslant \Lambda_{\mathsf{B}} \sum_{t_{h+1} \in \mathcal{U}_{h+1}} \pi(t_{h+1}|\tau_h) \times |x(t_{h+1})|. \tag{49}$$

因此，使用以下公式:
$\left(\mathbf{B}^{\theta}_h(o_h,\ a_h) - \mathbf{B}^{\bar{\theta}}_h(o_h,\ a_h)\right)\mathbf{q}^{\bar{\theta}}(\tau_{h-1}) = \mathbf{B}^{\theta}_h(o_h,a_h)\left(\mathbf{q}^{\bar{\theta}}(\tau_{h-1}) - \mathbf{q}^{\theta}(\tau_{h-1})\right) + \left(\mathbf{B}^{\theta}_h(o_h,a_h)\mathbf{q}^{\theta}(\tau_{h-1}) - \mathbf{B}^{\bar{\theta}}_h(o_h,a_h)\mathbf{q}^{\bar{\theta}}(\tau_{h-1})\right),$

我们有

$$2\mathcal{E}^{\bar{\theta}}_{\theta,h}(\pi, \tau_{h-1}) = \max_{\pi' \in \Pi_h(\pi)} \sum_{\tau_{h:H}} \pi'(\tau_{h:H}|\tau_{h-1}) \times \left|\mathbf{B}^{\theta}_{H:h+1}(\tau_{h+1:H}) \left(\mathbf{B}^{\theta}_h(o_h,a_h) - \mathbf{B}^{\bar{\theta}}_h(o_h,a_h)\right) \mathbf{q}^{\bar{\theta}}(\tau_{h-1})\right|$$

$$
\begin{aligned}
&\leqslant \max_{\pi' \in \Pi_h(\pi)} \sum_{\tau_{h:H}} \pi'(\tau_{h:H}|\tau_{h-1}) \times \left| \mathbf{B}^{\theta}_{H:h}(\tau_{h:H}) \left(\mathbf{q}^{\theta}(\tau_{h-1}) - \mathbf{q}^{\bar{\theta}}(\tau_{h-1}) \right) \right| \\
&\quad + \max_{\pi' \in \Pi_h(\pi)} \sum_{\tau_{h:H}} \pi'(\tau_{h:H}|\tau_{h-1}) \times \left| \mathbf{B}^{\theta}_{H:h+1}(\tau_{h+1:H}) \left(\mathbf{B}^{\theta}_{h}(o_h, a_h)\mathbf{q}^{\theta}(\tau_{h-1}) - \mathbf{B}^{\bar{\theta}}_{h}(o_h, a_h)\mathbf{q}^{\bar{\theta}}(\tau_{h-1}) \right) \right| \\
&\leqslant \Lambda_{\mathsf{B}} \sum_{t_h \in \mathcal{U}_h} \pi(t_h|\tau_{h-1}) \times \left| \mathbf{e}_{t_h}^{\top} \left(\mathbf{q}^{\theta}(\tau_{h-1}) - \mathbf{q}^{\bar{\theta}}(\tau_{h-1}) \right) \right| \\
&\quad + \Lambda_{\mathsf{B}} \sum_{o_h, a_h} \sum_{t_{h+1} \in \mathcal{U}_{h+1}} \pi(o_h, a_h, t_{h+1}|\tau_{h-1}) \times \left| \mathbf{e}_{t_{h+1}}^{\top} \left(\mathbf{B}^{\theta}_{h}(o_h, a_h)\mathbf{q}^{\theta}(\tau_{h-1}) - \mathbf{B}^{\bar{\theta}}_{h}(o_h, a_h)\mathbf{q}^{\bar{\theta}}(\tau_{h-1}) \right) \right|,
\end{aligned}
$$

其中 l　　这一不等式使用了（48）和（49）。注意，$\mathbf{q}^{\theta}(\tau_{h-1}) = [\mathbb{P}_{\theta}(t_h|\tau_{h-1})]_{t_h \in \mathcal{U}_h}$，因此

$$
\begin{aligned}
&\sum_{t_h \in \mathcal{U}_h} \pi(t_h|\tau_{h-1}) \times \left| \mathbf{e}_{t_h}^{\top} \left(\mathbf{q}^{\theta}(\tau_{h-1}) - \mathbf{q}^{\bar{\theta}}(\tau_{h-1}) \right) \right| \\
&= \sum_{t_h \in \mathcal{U}_h} \pi(t_h|\tau_{h-1}) \times |\mathbb{P}_{\theta}(t_h|\tau_{h-1}) - \mathbb{P}_{\bar{\theta}}(t_h|\tau_{h-1})| \\
&\leqslant D_{\mathrm{TV}} \left(\mathbb{P}^{\pi}_{\theta}(\tau_{h:H} = \cdot|\tau_{h-1}), \mathbb{P}^{\pi}_{\bar{\theta}}(\tau_{h:H} = \cdot|\tau_{h-1}) \right).
\end{aligned}
$$

此外，根据 B-表示的定义（参见定义 B.3），我们有

$$
\left[\mathbf{B}^{\theta}_{h}(o, a)\mathbf{q}^{\theta}(\tau_{h-1}) \right](t_{h+1}) = \mathbb{P}_{\theta}(t_{h+1}|\tau_{h-1}, o, a) \times \mathbb{P}_{\theta}(o|\tau_{h-1}) = \mathbb{P}_{\theta}(o, a, t_{h+1}|\tau_{h-1}),
$$

因此

$$
\begin{aligned}
&\sum_{o_h, a_h} \sum_{t_{h+1} \in \mathcal{U}_{h+1}} \pi(o_h, a_h, t_{h+1}|\tau_{h-1}) \times \left| \mathbf{e}_{t_{h+1}}^{\top} \left(\mathbf{B}^{\theta}_{h}(o_h, a_h)\mathbf{q}^{\theta}(\tau_{h-1}) - \mathbf{B}^{\bar{\theta}}_{h}(o_h, a_h)\mathbf{q}^{\bar{\theta}}(\tau_{h-1}) \right) \right| \\
&= \sum_{o_h, a_h} \sum_{t_{h+1} \in \mathcal{U}_{h+1}} \pi(o_h, a_h, t_{h+1}|\tau_{h-1}) \times |\mathbb{P}_{\theta}(o_h, a_h, t_{h+1}|\tau_{h-1}) - \mathbb{P}_{\bar{\theta}}(o_h, a_h, t_{h+1}|\tau_{h-1})| \\
&= \sum_{o_h, a_h} \sum_{t_{h+1} \in \mathcal{U}_{h+1}} \left| \mathbb{P}^{\pi}_{\theta}(o_h, a_h, t_{h+1}|\tau_{h-1}) - \mathbb{P}^{\pi}_{\bar{\theta}}(o_h, a_h, t_{h+1}|\tau_{h-1}) \right| \\
&\leqslant D_{\mathrm{TV}} \left(\mathbb{P}^{\pi}_{\theta}(\tau_{h:H} = \cdot|\tau_{h-1}), \mathbb{P}^{\pi}_{\bar{\theta}}(\tau_{h:H} = \cdot|\tau_{h-1}) \right).
\end{aligned}
$$

结合上述不等式，我们已经证明

$$
\mathcal{E}^{\bar{\theta}}_{\theta,h}(\pi, \tau_{h-1}) \leqslant \Lambda_{\mathsf{B}} D_{\mathrm{TV}} \left(\mathbb{P}^{\pi}_{\theta}(\tau_{h:H} = \cdot|\tau_{h-1}), \mathbb{P}^{\pi}_{\bar{\theta}}(\tau_{h:H} = \cdot|\tau_{h-1}) \right)
$$

对于任意步骤 $h \in [H]$。因此，我们可以利用这一事实 $D_{\mathrm{TV}} \leqslant D_{\mathrm{H}}$ 并应用引理 A.6 以获得　　ain

$$
\begin{aligned}
\mathbb{E}^{\pi}_{\bar{\theta}} \left[\mathcal{E}^{\bar{\theta}}_{\theta,h}(\pi, \tau_{h-1})^2 \right] &\leqslant \Lambda^2_{\mathsf{B}} \mathbb{E}^{\pi}_{\bar{\theta}} \left[D_{\mathrm{TV}} \left(\mathbb{P}^{\pi}_{\theta}(\tau_{h:H} = \cdot|\tau_{h-1}), \mathbb{P}^{\pi}_{\bar{\theta}}(\tau_{h:H} = \cdot|\tau_{h-1}) \right)^2 \right] \\
&\leqslant \Lambda^2_{\mathsf{B}} \mathbb{E}^{\pi}_{\bar{\theta}} \left[D^2_{\mathrm{H}} \left(\mathbb{P}^{\pi}_{\theta}(\tau_{h:H} = \cdot|\tau_{h-1}), \mathbb{P}^{\pi}_{\bar{\theta}}(\tau_{h:H} = \cdot|\tau_{h-1}) \right) \right] \leqslant 2\Lambda^2_{\mathsf{B}} D^2_{\mathrm{H}} \left(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\bar{\theta}} \right).
\end{aligned}
$$

类似的论证也可以证明 $(\mathcal{E}^{\bar{\theta}}_{\theta,\ 0}(\pi))^2 \leqslant \Lambda^2_{\mathsf{B}} D_{\mathrm{TV}} \left(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\bar{\theta}} \right)^2 \leqslant \Lambda^2_{\mathsf{B}} D^2_{\mathrm{H}} \left(\mathbb{P}^{\pi}_{\theta}, \mathbb{P}^{\pi}_{\bar{\theta}} \right.$　　$\left.\right)$.　　□

H.4 定理H.3 的证明

定理H.3 的证明使用了 OMLE 算法的以下快速率保证，这是标准结果（例如 Van de Geer (2000)；Agarwal 等人 (2020 年)，一个简单的证明可见（Chen 等人，2022a，附录 E））。

命题 H.5（最大似然估计的保证）。 假设我们在算法 $\beta \geqslant 2\log \mathcal{N}_{\Theta}(1/T) + 2\log(1/\delta) + 2$ 中选择 *1*。那么，以至少 $1 - \delta$ 的概率，以下结论成立：

(a) For all $k \in [K]$, $\theta^{\star} \in \Theta^k$;

(b) 对所有 $k \in [K]$ 以及任意 $\theta \in \Theta^k$，均有

$$\sum_{t=1}^{k-1} D_{\mathrm{H}}^2\left(\mathbb{P}_\theta^{\pi^t}, \mathbb{P}_{\theta^\star}^{\pi^t}\right) \leqslant 2\beta.$$

接下来我们证明定理 H.3。我们采用命题 $\mathcal{E}_\theta^{\bar{\theta}}{}_{、h}(\pi_{、}\ \tau_{h-1})$ 中的定义，并将其简记为 H.2 以及$\mathcal{E}_{k、h}^{\star}=\mathcal{E}_{\theta^k}^{\theta^\star}{}_{、h}$。我们还以命题 H.5中的事件成功为条件。

步骤 1。 根据命题 H.5，有 $\theta^\star \in \Theta$。因此，$V_{\theta^k}(\pi^k) \geqslant V_\star$，且根据命题 H.2，我们有

$$\begin{aligned}
\sum_{t=1}^{k}\left(V_\star - V_{\theta^\star}(\pi^t)\right) &\leqslant \sum_{t=1}^{k}\left(V_{\theta^t}(\pi^t) - V_{\theta^\star}(\pi^t)\right) \leqslant \sum_{t=1}^{k} D_{\mathrm{TV}}\left(\mathbb{P}_{\theta^t}^{\pi^t}, \mathbb{P}_{\theta^\star}^{\pi^t}\right) \\
&\leqslant \sum_{t=1}^{k} 1 \wedge \left(\mathcal{E}_{t,0}^\star(\pi^t) + \sum_{h=1}^{H} \mathbb{E}_{\pi^t}\left[\mathcal{E}_{t,h}^\star(\pi^t, \tau_{h-1})\right]\right) \\
&\leqslant \sum_{t=1}^{k}\left(1 \wedge \mathcal{E}_{t,0}^\star(\pi^t) + \sum_{h=1}^{H} 1 \wedge \mathbb{E}_{\pi^t}\left[\mathcal{E}_{t,h}^\star(\pi^t, \tau_{h-1})\right]\right),
\end{aligned} \tag{50}$$

其中期望 取值于 $\mathbb{E}_{\pi^t}$ 之上 $\tau_{h-1} \sim \mathbb{P}_{\theta^\star}^{\pi^t}$。另一方面，根据命题 H.2，我们有 e

$$\mathbb{E}_{\pi^t}\left[\mathcal{E}_{k,h}^\star(\pi^t, \tau_{h-1})^2\right] \leqslant 2\Lambda_{\mathsf{B}}^2 D_{\mathrm{H}}^2\left(\mathbb{P}_{\theta^k}^{\pi^t}, \mathbb{P}_{\theta^\star}^{\pi^t}\right), \qquad \mathcal{E}_{k,0}^\star(\pi^t)^2 \leqslant \Lambda_{\mathsf{B}}^2 D_{\mathrm{H}}^2\left(\mathbb{P}_{\theta^k}^{\pi^t}, \mathbb{P}_{\theta^\star}^{\pi^t}\right).$$

此外，根据命题 H.5，我们有 $\sum_{t=1}^{k-1} D_{\mathrm{H}}^2\left(\mathbb{P}_{\theta^k}^{\pi^t},\ \mathbb{P}_{\theta^\star}^{\pi^t}\right) \leqslant 2\beta$. 因此，结合上面两个等式可得

$$\sum_{t<k} \mathbb{E}_{\pi^t}[\mathcal{E}_{k,h}^\star(\pi^t, \tau_{h-1})^2] \leqslant 4\Lambda_{\mathsf{B}}^2\beta, \qquad \forall k \in [K], 0 \leqslant h \leqslant H. \tag{51}$$

步骤 2。 我们希望利用广义 50-51-Eluder论证，将性能分解（ℓ_2）与平方 B 误差界（$h \in [H]$）联系起来，$h = 0$。

情况 1： $h \in [H]$。我们记作 $m = m_{h+1}$，使得 $\mathcal{U}_{h+1} = (\mathcal{O} \times \mathcal{A})^{m-1} \times \mathcal{O}$，$\Omega_{h+1} = (\mathcal{O} \times \mathcal{A})^{m-1}$。根据定义，

$$\begin{aligned}
\mathcal{E}_{k,h}^\star(\pi^t, \tau_{h-1}) :=& \frac{1}{2} \max_{\pi' \in \Pi_h(\pi^t)} \sum_{\tau_{h:H}} \pi'(\tau_{h:H}|\tau_{h-1}) \times \left|\mathbf{B}_{H:h+1}^k(\tau_{h+1:H})\left(\mathbf{B}_h^k(o_h, a_h) - \mathbf{B}_h^\star(o_h, a_h)\right)\mathbf{q}^\star(\tau_{h-1})\right|, \\
=& \frac{1}{2} \max_{\pi'} \sum_{\tau_{h:H}} \pi'(\tau_{h+m:H}|\tau_{h+m-1}) \times \pi^t(\tau_{h:h+m-1}|\tau_{h-1}) \times \left|\mathbf{B}_{H:h+1}^k(\tau_{h+1:H})\left(\mathbf{B}_h^k(o_h, a_h) - \mathbf{B}_h^\star(o_h, a_h)\right)\mathbf{q}^\star(\tau_{h-1})\right| \\
=& \frac{1}{2} \sum_{o_h, a_h} \sum_{\omega_{h+1} \in \Omega_{h+1}} \pi^t(o_h, a_h, \omega_{h+1}|\tau_{h-1}) \left\|\mathcal{B}_{H:h+m}^k \cdot \mathbf{B}_{h+m-1:h+1}^k(\omega_{h+1})\left(\mathbf{B}_h^k(o_h, a_h) - \mathbf{B}_h^\star(o_h, a_h)\right)\mathbf{q}^\star(\tau_{h-1})\right\|_\Pi,
\end{aligned}$$

其中，在最后一个等式中，我们采用了（10）中引入的记号。

为衔接（50）和（51），我们使用以下广义 ℓ_2-Eluder 引理，该引理可直接通过推广 Chen 等人 (2022a, 命题 C.1 和推论 C.2）（对应于以下结果在 $N = 1$ 时的特殊情形）得到。

引理 H.6（广义 ℓ_2-Eluder 论证）。假设我们有一列函数 $\{f_{k,l} : \mathbb{R}^n \to \mathbb{R}\}_{(k,l)\in[K]\times[N]}$：

$$f_{k,l}(x) := \max_{r \in \mathcal{R}} \sum_{j=1}^{J} |\langle x, y_{k,l,j,r}\rangle|,$$

它由向量族 $\{y_{k,l,j,r}\}_{(k,j,r)\in[K]\times[J]\times\mathcal{R}} \subset \mathbb{R}^n$ 给出。进一步假设存在 $L>0$，使得 $f_{k,l}(x) \leqslant L\|x\|_1$。

进一步考虑向量序列 $(x_{t,\ l,\ i})_{(t,\ l,\ i)\in[K]\times[N]\times\mathcal{I}}$，满足以下条件

$$\sum_{t=1}^{k-1}\mathbb{E}_{i\sim q_t}\left[\left(\sum_{l=1}^{N} f_{k,l}(x_{t,l,i})\right)^2\right] \leqslant \beta_k, \qquad \forall k\in[K],$$

由 $(x_{t,\ l,\ i})$ 张成的子空间维数至多为 d。那么有

$$\sum_{t=1}^{k} 1\wedge\mathbb{E}_{i\sim q_t}\left[\sum_{l=1}^{N} f_{t,l}(x_{t,l,i})\right] \leqslant \sqrt{4Nd\Big(k+\sum_{t=1}^{k}\beta_t\Big)\log\Big(1+kdL\max_i\|x_i\|_1\Big)}, \qquad \forall k\in[K].$$

我们将采用以下三个准备步骤来应用引理 H.6。

1. 我们定义

$$x_{t,l,i} := \pi^t(o_h^l, a_h^l, \omega_{h+1}^l|\tau_{h-1}^i)\times\mathbf{q}^\star(\tau_{h-1}^i)\in\mathbb{R}^{\mathcal{U}_h},$$
$$y_{k,l,j,\pi} := \frac{1}{2}\pi(\tau_{h+m:H}^j)\times\left[\mathbf{B}_{H:h+m}^k(\tau_{h+1:H}^j)\mathbf{B}_{h+m-1:h+1}^k(\omega_{h+1}^l)\left(\mathbf{B}_h^k(o_h^l,a_h^l)-\mathbf{B}_h^\star(o_h^l,a_h^l)\right)\right]^\top\in\mathbb{R}^{\mathcal{U}_h},$$

其中 $\{\tau_{h-1}^i\}_i$是所有可能的 $\tau_{h-1}\in(\mathcal{O}\times\mathcal{A})^{h-1}$，$\{\tau_{h+m:\ H}^j=(o_{h+m},\ a_{h+m},\ \cdots,\ o_H,\ a_H)\}_{j=1}^n$的排序（因此也是 $\tau_{h+m:H}$ 的排序），$\{(o_h^l,\ a_h^l,\ \omega_{h+1}^l)\}_{l=1}^N$ 是$\mathcal{O}\times\mathcal{A}\times\Omega_{h+1}$ 的排序（因此也是 $N=OA|\Omega_{h+1}|\leqslant OAU_{\mathcal{T}}$的排序），$\pi$ 是任意一个从步骤 h开始的策略。然后我们定义

$$f_{k,l}(x)=\max_\pi\sum_j|\langle y_{k,l,j,\pi},x\rangle|, \qquad x\in\mathbb{R}^{\mathcal{U}_h}.$$

根据定义可得

$$\mathcal{E}_{k,h}^\star(\pi^t,\tau_{h-1}^i)=\sum_{l=1}^{N}\pi^t(o_h^l,a_h^l,\omega_{h+1}^l|\tau_{h-1}^i)\times f_{k,l}(\mathbf{q}^\star(\tau_{h-1}^i))=\sum_{l=1}^{N}f_{k,l}(x_{t,l,i}).$$

2. 根据 $\theta^\star$ 的 PSR 秩小于或等于 d这一假设，我们有 $\mathrm{dimspan}(x_{t,l,i})\leqslant d$。此外，根据定义，我们有$\|x_{t,l,\ i}\|_1\leqslant U_A\leqslant U_{\mathcal{T}}$。

3. 接下来只需验证 f_k 关于 1-范数是利普希茨连续的。显然,

$$\begin{aligned} f_{k,l}(\mathbf{q}) &\leqslant \frac{1}{2}\left[\left\|\mathcal{B}_{H:h}^k\mathbf{q}\right\|_\Pi+\max_{o,a}\left\|\mathcal{B}_{H:h+1}^k\mathbf{B}_h^\star(o,a)\mathbf{q}\right\|_\Pi\right] \\ &\leqslant \frac{1}{2}\left[\Lambda_{\mathsf{B}}\|\mathbf{q}\|_1+\Lambda_{\mathsf{B}}\max_{o,a}\|\mathbf{B}_h^\star(o,a)\mathbf{q}\|_1\right]\leqslant\frac{1}{2}\Lambda_{\mathsf{B}}R_{\mathsf{B}}\|\mathbf{q}\|_1. \end{aligned}$$

因此，我们可以取 $L=\frac{1}{2}\Lambda_{\mathsf{B}}R_{\mathsf{B}}$，以确保 $f_{k,l}(x)\leqslant L\|x\|_1$。

因此，应用引理 H.6 可得

$$\sum_{t=1}^{k}1\wedge\mathbb{E}_{\pi^t}\left[\mathcal{E}_{t,h}^\star(\pi^t,\tau_{h-1})\right]\leqslant\mathcal{O}\left(\sqrt{\Lambda_{\mathsf{B}}^2Nd\iota\beta k}\right)\leqslant\mathcal{O}\left(\sqrt{\Lambda_{\mathsf{B}}^2OAU_{\mathcal{T}}d\iota\beta k}\right). \tag{52}$$

情况 1 证毕。

情况2： $h=0$。本情况的推导类似于

$$\sum_{t=1}^{k}1\wedge\mathcal{E}_{t,0}^\star(\pi^t)\leqslant\mathcal{O}\left(\sqrt{\Lambda_{\mathsf{B}}^2OAU_{\mathcal{T}}\iota\beta k}\right). \tag{53}$$

结合这两种情况，我们得到

$$\sum_{t=1}^{k}\left(V_\star - V_{\theta^\star}(\pi^t)\right) \overset{(i)}{\leqslant} \sum_{t=1}^{k} 1 \wedge \mathcal{E}^\star_{t,0}(\pi^t) + \sum_{h=1}^{H}\sum_{t=1}^{k} 1 \wedge \mathbb{E}_{\pi^t}\left[\mathcal{E}^\star_{t,h}(\pi^t, \tau_{h-1})\right]$$
$$\overset{(ii)}{\leqslant} \mathcal{O}\left(\sqrt{\Lambda_{\mathsf{B}}^2 OAU_{\mathcal{T}}\iota\beta k}\right) + H\cdot\mathcal{O}\left(\sqrt{\Lambda_{\mathsf{B}}^2 OAU_{\mathcal{T}}d\iota\beta k}\right) \leqslant \mathcal{O}\left(\sqrt{H^2\Lambda_{\mathsf{B}}^2 Nd\iota\beta k}\right),$$

其中 (i) 使用 (50); (ii）使用上述两个情况 （53）和 （52）。至此，定理 H.3的证明完成 □

I 其他讨论

I.1 DEC 中不存在通用的样本复杂度 + 价值函数类／策略类的对数覆盖数

DEC 理论的典型保证Foster 等人(2021); Chen 等人 (2022b)断言，对于任意模型类$\mathcal{M}$ 和策略类 Π，E2D 算法可达到

$$\mathbb{E}[\mathbf{Regret}] \leqslant \mathcal{O}(1)\cdot\min_{\gamma>0}\left(T\cdot\mathrm{dec}^{\mathrm{H}}_{\gamma}(\mathcal{M}) + \gamma\log|\mathcal{M}|\right). \tag{54}$$

Foster 等人 (2021) 还指出，令 $\mathrm{co}(\mathcal{M})$ 表示 $\mathcal{M}$ (的凸包（即 $M\in\mathcal{M}$的所有混合模型构成的集合），存在一种 E2D 变体可达到

$$\mathbb{E}[\mathbf{Regret}] \leqslant \mathcal{O}(1)\cdot\min_{\gamma>0}\left(T\cdot\mathrm{dec}^{\mathrm{H}}_{\gamma}(\mathrm{co}(\mathcal{M})) + \gamma\log|\Pi|\right).$$

然而，$\mathrm{dec}^{\mathrm{H}}_{\gamma}(\mathrm{co}(\mathcal{M}))$ 通常规模很大，难以计算——例如，当 $\mathcal{M}$ 是所有表格型马尔可夫决策过程构成的类时，$\mathrm{dec}^{\mathrm{H}}_{\gamma}(\mathrm{co}(\mathcal{M}))$ 随 S呈指数增长，H (Foster 等人，2022)。因此，自然会提出以下问题

> **问题：** 能否得到一个遗憾上界，将（$|\mathcal{M}|$ （54）中的 log 项替换为 log $|\Pi|$ 或 log $|\mathcal{F}_{\mathcal{M}}|$（其中 $\mathcal{F}_{\mathcal{M}}$ 是由 $\mathcal{M}$诱导的某个价值函数类）？

当模型类 $\mathcal{M}$ 本身远大于价值类（例如 Q 函数类）时，上述问题尤其值得关注；例如， $\mathcal{M}$ 是一个具有已知特征 Jin 等人，2020b 的线性马尔可夫决策过程类， $\phi(s,a)$ 但 $\mu(s')$未知。此外，对于某些特定问题类，例如表格型马尔可夫决策过程，将（$|\mathcal{M}|$ （54）中的 log 项替换为 log $|\Pi|$ 可能会带来不错的改进。此时我们可以令 Π 为确定性马尔可夫策略类，其 $\log|\Pi| = \widetilde{\mathcal{O}}(SH)$ 小于 log $|\mathcal{M}| = \widetilde{\mathcal{O}}(\log\mathcal{N}_{\mathcal{M}}) = \widetilde{\mathcal{O}}\left(S^2AH\right)$，差距达 SA 倍。

然而，我们关于揭示型部分可观测马尔可夫决策过程的下界——具体而言，附录F中的困难实例构造——为这个问题提供了一个（部分）否定的答案。为简单起见，考虑 $m = 2$情形，并假设 $A^H \gg$多项式$(S,\ O,\ A,\ \alpha^{-1},\ T)$）关于我们的模型类，我们有以下基本事实 $\mathcal{M}$。

1. $\mathcal{M}$ 的结构确保，任何可能的最优策略都是一个确定性动作序列（不依赖于历史），因此我们可以考虑 Π = {确定性动作序列}，其中~对数 $\Pi = \mathcal{O}(H)$。2. Chen 等人 (2022a的一般结果表明，只要 $\mathcal{M}$ 是 2 步 α揭示型部分可观测马尔可夫决策过程的一个子类，就有 $\mathrm{edec}_\gamma(\mathcal{M}) \leqslant \widetilde{\mathcal{O}}(SA^2H^2\alpha^{-2}/\gamma)$——其中 edec 是 dec 的 PAC 学习类比——这意味着 $\mathrm{dec}_\gamma(\mathcal{M}) \leqslant \widetilde{\mathcal{O}}(\sqrt{SA^2H^2\alpha^{-2}/\gamma})$ (Chen 等人，2022b）。3. 命题F.1指出，对于任意算法，族 $\mathcal{M}$ 中的最坏情况遗憾下界为 $\Omega\big((S\sqrt{O}A^2H\alpha^{-2})^{1/3}T^{2/3}\big)$。

—

请注意，遗憾下界中包含一个 poly(O) 因子，而该因子并未出现在 DEC 的上界中。这引出了以下

事实：在不对问题作出更多结构性假设的情况下，形如以下形式的遗憾上界

$$\mathbb{E}[\mathbf{Regret}] \leqslant \mathcal{O}(1) \cdot \min_{\gamma>0} \left(T \cdot \mathrm{dec}_{\gamma}^{\mathrm{H}}(\mathcal{M}) + \gamma \log |\Pi| \right) \tag{55}$$

无法实现。

上述事实是因为，如果 (55) 可以实现，那么结合前面提到的 DEC 上界，就会得到一个不随 poly(O)变化的遗憾上界，这与下界矛盾。

类似地，如果将每个 POMDP $M \in \mathcal{M}$ 视为一个马尔可夫决策过程，并将每段历史 τ_h 视为一个“超级状态”，那么自然有， M 的 Q函数为

$$Q_M^\star(\tau_h) = \mathbb{E}_M^{\pi_M^\star}\left[\sum_{h'=1}^{H} r_{h'} \middle| \tau_h \right], \qquad \tau_h \in (\mathcal{O} \times \mathcal{A})^h, 0 \leqslant h \leqslant H,$$

~其中， $\pi_M^\star$是 M的最优策略。对于我们的族 $\mathcal{M}$，不难验证 log $\mathcal{Q}_\mathcal{M} = \mathcal{O}(H)$，其中 $\mathcal{Q}_\mathcal{M} = \{Q_M^\star : M \in \mathcal{M}\}$。因此，如果将价值类取为这样的 Q 函数类，对上面问题的回答也是否定的。

I.2 定理 5困难实例的算法

我们提出一种暴力算法 $\mathfrak{A}$，用于学习附录 G 中给出的困难实例类（用于证明定理 5），其 PAC 样本复杂度为 $\widetilde{\mathcal{O}}\big(S^{3/2}O^{1/2}A^m H/(\alpha^2\varepsilon^2)\big)$。算法 $\mathfrak{A}$包含两个阶段：

1. 阶段1：对于每个 $h \in \mathcal{H}$， $s \in \mathcal{S}_{\mathsf{leaf}}$， $a \in \mathcal{A}_c$，**一个** $\in \mathcal{A}^{m-1}$，算法会花费 N_1 回合访问 $o_h = s$，执行动作 $a_{h:h+m} = (a,$ **一个**，揭示)，并观察 $(o_{h+m},\ o_{h+m+1})$。观测到的 $(o_{h+m},\ o_{h+m+1})$随后应满足联合分布

$$\mathbb{P}(\mathsf{lock}^j, o_i^+) = \frac{1+\sigma\varepsilon\mu_{j,i}}{2KL}, \qquad \mathbb{P}(\mathsf{lock}^j, o_i^-) = \frac{1-\sigma\varepsilon\mu_{j,i}}{2KL}, \qquad \forall (j,i) \in [L] \times [K]$$

若 $(h,\ s,\ a,$ **一个**$) = (h^\star,\ s^\star,\ a^\star,$ **一个**$^\star_{h^\star+1:\ h^\star+m-1})$，并满足分布 均匀分布$(\{$锁$^1,\ \cdots,$ 锁$^L\} \times \mathcal{O}_o)$ 否则。使用标准的均匀性检验算法（Canonne，2020 年），我们可以区分

$$H_0 : (h,s,a,\mathbf{a}) = (h^\star, s^\star, a^\star, \mathbf{a}^\star_{h^\star+1:h^\star+m-1}),$$
$$H_1 : (h,s,a,\mathbf{a}) \neq (h^\star, s^\star, a^\star, \mathbf{a}^\star_{h^\star+1:h^\star+m-1})$$

以高概率使用 $N_1 = \widetilde{\mathcal{O}}\big(\sqrt{KL/(\sigma^2\varepsilon^2)}\big)$ 个样本，对每个固定的 $(h,\ s,\ a,$ **一个**$)$。因此，阶段1所需的总样本量为 $|\mathcal{S}_{\mathsf{leaf}}|\, H\, |\mathcal{A}_c|\, A^{m-1} \times \widetilde{\mathcal{O}}\big(\sqrt{KL/(\sigma^2\varepsilon^2)}\big)$。 ——

2. 阶段 2：阶段1完成后，算法可以正确识别参数 $(h^\star,\ s^\star,\ a^\star,\ \mathbf{a}^\star_{h^\star+1:\ h^\star+m-1})$（如果 $M \neq 0$）或找出 $M = 0$。后一种情况下，算法可以直接终止，并输出 $M = 0$ 的最优策略。前一种情况下，算法需要继续学习密码$\mathbf{a}^\star_{h^\star+m:\ H-1}$：

   • 对于每个 $h = h^\star + m,\ h^\star + 2m,\ \cdots$：– 对于每个$\mathbf{a} \in \mathcal{A}^m$，通过花费 N_1 个回合访问 $o_{h^\star} = s^\star$，执行动作 $a_{h^\star:\ h+m} = (a,\ \mathbf{a}^\star_{h^\star+1:\ h-1},\ \mathbf{a},$ 揭示)并观察 $(o_{h+m},\ o_{h+m+1})$，检验$\mathbf{a} = \mathbf{a}^\star_{h:\ h+m-1}$ 是否成立。– 基于与阶段1相同的理由，并根据我们的选择 $N_1 = \widetilde{\mathcal{O}}\big(\sqrt{KL/(\sigma^2\varepsilon^2)}\big)$, ，我们可以使用标准均匀性检验算法，以较高概率学习$\mathbf{a}^\star_{h:\ h+m-1}$。 ——

   算法一旦学会 $M = (h^\star,\ s^\star,\ a^\star,$ **一个**$^\star)$，便会终止并输出 M 的最优策略。阶段 2 所需的总样本量至多为 $A^m H \times \widetilde{\mathcal{O}}\big(\sqrt{KL/(\sigma^2\varepsilon^2)}\big)$ 个样本。

其中 总结一下，我们在上文构造的暴力算法 $\mathfrak{A}$ 可以用样本量为 $\mathcal{M}$ 学会 e

$$|\mathcal{S}_{\mathsf{leaf}}|\, H\, |\mathcal{A}_c|\, A^{m-1} \times \widetilde{\mathcal{O}}\left(\frac{\sqrt{KL}}{\sigma^2\varepsilon^2}\right) + A^m H \times \widetilde{\mathcal{O}}\left(\frac{\sqrt{KL}}{\sigma^2\varepsilon^2}\right) \leqslant \widetilde{\mathcal{O}}\left(\frac{S^{3/2}O^{1/2}A^m H}{\alpha^2\varepsilon^2}\right),$$

其中的界由我们对 σ， $\mathcal{S}_{\mathsf{leaf}}$， $\mathcal{A}_c$， K， L 的选择决定。