arXiv:2602.10938v1 [cs.LO] 11 Feb 2026

具有先验语义和全称语义的多环境马尔可夫决策过程

Benjamin Bordais，Jean-Fran.cois Raskin

布鲁塞尔自由大学

摘要

多环境马尔可夫决策过程（MEMDP）为一个马尔可夫决策过程配备多个概率转移函数（每种可能的环境对应一个），因此状态可观测，而环境不可观测。以往研究探讨了两种语义：（i）全称语义，由对手选择环境；（ii）先验语义，在执行前根据固定分布抽取一次环境。我们阐明了这两种语义之间的关系。对于奇偶目标，我们证明定性问题（即值为一）在两种语义下等价，并为具有先验语义的多环境马尔可夫决策过程的一般值提出了一种新算法。具体而言，我们证明，对于具有奇偶目标的多环境马尔可夫决策过程，其先验值可由一种空间高效算法以任意精度近似；等价地，当概率采用一元编码时，相关间隙问题可在多项式空间内判定（否则可在指数空间内判定）。随后，我们证明，全称值等于所有信念上的先验值下确界。由此，我们为全称间隙问题得到了一个复杂度相同的新算法（一元编码概率时为多项式空间，一般情况下为指数空间），改进了此前需要双指数空间的过程。最后，我们指出，具有先验语义的多环境马尔可夫决策过程构成了部分可观测马尔可夫决策过程的一个重要且易于处理的子类：我们的算法利用了信念熵永不增加这一事实，并证明任何具有此性质的部分可观测马尔可夫决策过程都能有效地归约为先验多环境马尔可夫决策过程，这表明先验多环境马尔可夫决策过程涵盖了一个广泛且具有实际意义的部分可观测马尔可夫决策过程子类。

1 引言

多环境马尔可夫决策过程（MEMDP）最初于 [15]，中提出，是对经典马尔可夫决策过程的推广。它们用于建模这样一类决策问题：系统状态完全可观测，而随机 环境 未知但固定，并在执行前从有限个候选环境中选定或抽取。形式上，一个多环境马尔可夫决策过程由一组共用的状态空间和动作空间，以及每个可能环境各自对应的概率转移函数组成。控制器在行动时并不会被明确告知当前运行的是哪个环境。直观而言，这一模型介于完全可观测的马尔可夫决策过程与部分可观测马尔可夫决策过程之间：它刻画了一类自然的部分信息问题，同时避免了一般部分可观测马尔可夫决策过程中许多不可判定性障碍。

早期研究引入了多环境马尔可夫决策过程，并证明：对于部分可观测马尔可夫决策过程而言不可判定的一些关键综合问题，在多环境马尔可夫决策过程中变得可判定（有时甚至能高效求解），从而说明该模型是一种兼具可处理性与表达能力的部分可观测马尔可夫决策过程变体 [15]。后续研究进一步发展了针对 ω-正则目标的理论，并厘清了算法方面的整体格局 [6, 19]。除基础理论价值外，多环境马尔可夫决策过程也被视为一种实用的建模工具：其结构使信念更新比一般部分可观测马尔可夫决策过程更廉价，并支持情境推荐以及概率模型中的参数不确定性等应用 [3]。

通用语义与先验语义。 以往的研究从两种不同语义出发研究多环境马尔可夫决策过程：[15, 19, 6] 在通用语义下研究该模型，而 [3] 则在先验语义下研究该模型。在通用（对抗）语义下，对手选择固定控制器策略所要面对的环境；策略的价值是其在最坏环境中

满足目标的概率。在先验语义下，环境最初从一个固定分布（先验）中随机抽取；其价值是相应的期望。在这两种情况下，所选环境都不会透露给控制器。我们聚焦于奇偶目标，它是表达 ω-正则性质的一种典型方式，并建立这两种语义之间的结构性和算法性联系。

在介绍我们的主要贡献之前，我们先指出两种语义之间存在一种定性等价关系。在几乎必然和极限必然意义下，值为 1 的概念在两种语义中一致：在全称语义下，存在达到值 1（分别地，任意接近 1）的策略，当且仅当存在先验值为 1（分别地，任意接近 1）的策略。这种对应关系使我们能够在两种语义之间迁移值为 1 的结果（见推论2）。

贡献。我们的主要贡献可概括如下：

1. **先验值的近似**。我们给出一种算法，可将多环境马尔可夫决策过程的先验值近似到任意精度。这是我们贡献中技术难度最高的部分。更具体地说，我们展示了如何通过算法求解先验语义下的 ε-间隙问题。该问题要求给定一个多环境马尔可夫决策过程 M、关于运行环境的先验信念 b 、一个奇偶目标 W、一个阈值 $0 < \alpha < 1$和一个精度$\varepsilon > 0$，然后作出如下回答：是，如果先验值至少为 α；否，如果该值至多为 $\alpha - \epsilon$；除此之外不作要求。若概率采用一元编码，我们的算法运行时间为 PSPACE；否则为 EXPSPACE（见定理6）。

2. **从先验语义到通用语义**。 我们还证明，通用 值等于所有先验信念对应的先验 值的下确界（见定理 7）。结合这一结果以及先验值关于先验信念的 1-Lipschitz 连续性，我们得到了一个新的通用 ε-间隙问题算法，其复杂度与先验值问题相同（PSPACE对于一元编码概率，EXPSPACE 一般情况下），见定理10，改进了此前 [6] 中提出的双指数空间过程。

3. **具有先验语义的多环境马尔可夫决策过程是一个易处理的部分可观测马尔可夫决策过程子类**。 多环境马尔可夫决策过程易于处理的一个关键原因是隐藏环境固定不变，这限制了信念动态：关于运行环境的信念熵（信息论中的概念）不会增加。我们证明，实际上，任何信念熵不会增加的部分可观测马尔可夫决策过程都可以归约为一个先验多环境马尔可夫决策过程（代价为指数级）。因此，我们的算法适用范围不局限于“纯粹的”多环境马尔可夫决策过程，还涵盖了部分可观测马尔可夫决策过程的一个重要子类，同时避免了 ω-正则目标所涉及的经典不可判定现象（见定理15）。

相关工作。 如上所述，Raskin 和 Sankur 在 [15] 中首次在全称语义下引入了多环境马尔可夫决策过程，作为一种形式化模型，用于处理环境固定但未知、且从有限个候选模型中选取的决策场景。在这一框架下，若干定性 ω-正则综合问题变得可判定，这与部分可观测马尔可夫决策过程（POMDP）的相应情形形成对比。后续工作进一步细化了定性复杂度格局。对于几乎必然满足 Rabin 目标，Suilen 等人建立了一个 PSPACE 完全的判定过程 [19]。对于奇偶目标，Chatterjee 等人证明，值为 1 问题在一般情况下是 PSPACE 完全的（而当环境数量固定时，可在 PTIME 内求解）；他们还证明纯策略足以实现值 1，并提出一种双指数空间近似方案来计算该值 [6]。在先验语义下，多环境马尔可夫决策过程还被提出作为具有折扣收益的部分可观测马尔可夫决策过程的一个计算上易处理的子类，其优势在于能够更高效地进行信念状态更新，且信念熵具有非递增性质 [3]。

作为背景，经典（完全可观测）马尔可夫决策过程的定性问题可在多项式时间内求解 [2], ，而类似的部分可观测马尔可夫决策过程问题则要困难得多，因为其中大多数都是不可判定的 [14,12], ，目前已知只有有界时域问题是可判定的 [14]。因此，在这一背景下，多环境马尔可夫决策过程占据着特殊的位置：它们保留了部分信息

表达能力，同时避免了部分可观测马尔可夫决策过程中出现的关键不可判定性障碍。本文取得的正面结果补充并细化了 [15, 3, 19, 6]。

2 定义

考虑一个非空集合 Q。函数 $d: Q \to [0,1]$ 的支撑集Supp(d)是集合 Supp(d) := $\{q \in Q \mid d(q) > 0\}$。若函数 $d: Q \to [0,1]$ 具有可数支撑集，且 $\sum_{q\in Q} d(q) = 1$，则称其为 Q 上的分布。我们用 $\mathcal{D}(Q)$ 表示集合 Q 上所有概率分布的集合。若概率分布 $d \in \mathcal{D}(Q)$ 满足 $|\text{Supp}(b)| = 1$，则称其为狄拉克分布。对于所有集合 A 和$\rho = q_0 \cdot (a_1, q_1) \cdots (a_n, q_n) \in Q \cdot (A \cdot Q)^*$，我们令 last($\rho$) := $q_n \in Q$。

马尔可夫决策过程与策略。马尔可夫决策过程（简称 MDP）G 是一个元组 G $= (Q, A, \delta)$，其中 Q 是非空有限状态集，A 是非空有限动作集，而 δ: $Q \times A \to \mathcal{D}(Q)$ 是转移函数，它将每个状态-动作对映射到后继状态上的概率分布。$Q \cdot (A \cdot Q)^*$ 中的元素（分别地，$(Q \cdot A)^\omega$ 中的元素）称为有限运行（分别地，无限运行）；Q^*（分别地，Q^ω）中的元素称为有限路径（分别地，无限路径）。(Q, A) 上的策略是函数 σ: $Q \cdot (A \cdot Q)^* \to \mathcal{D}(A)$，它将有限运行映射到动作上的概率分布。我们用 Strat(Q, A) 表示 (Q, A) 上所有策略的集合。

目标。考虑一个非空有限集 Q。一个目标 $W \subseteq Q^\omega$ 是一个博雷尔集: $W \in$博雷尔(Q)。我们关注奇偶目标，即满足以下条件的目标 $W \in$ 博雷尔(Q)：存在一个标记函数 $f: Q \to \mathbb{N}$，使得 W 是所有无限路径的集合，其中无限次出现的最高标签为偶数: $W = \{\rho \in Q^\omega \mid \max\{n \in \mathbb{N} \mid \forall i \in \mathbb{N}, \exists j \geq i: f(\rho_j) = n\}$ 为偶数$\}$。

考虑一个马尔可夫决策过程 G $= (Q, A, \delta)$。策略 σ: $Q \cdot (A \cdot Q)^* \to \mathcal{D}(A)$ 在有限运行集合上诱导出一个概率测度，该测度可以自然地扩展为无限运行上的相应博雷尔σ-代数。我们首先定义策略在给定有限运行 $\rho \in Q \cdot (A \cdot Q)^*$之后所诱导的概率测度。对于 $\pi \in Q \cdot (A \cdot Q)^*$，我们按照 π的长度通过归纳法定义值 $\mathbb{P}^\sigma_\rho$(G π) $\in [0,1]$ 如下。若 $|\pi| \leq |\rho|$，则令 $\mathbb{P}^\sigma_\rho$(G π) := 1，当 π是ρ的前缀时；否则令 $\mathbb{P}^\sigma_\rho$(G π) := 0。对于所有满足 ρ是 π的前缀的 $\pi \in Q \cdot (A \cdot Q)^*$，以及所有 $(a,t) \in A \times Q$，定义 $\mathbb{P}^\sigma_\rho$(G $\pi \cdot (a,t)$) := $\mathbb{P}^\sigma_\rho$(G π) $\cdot\, \sigma(\pi)(a) \cdot \delta(\text{last}(\pi)\ a)(t)$。

那么，对于所有 $\pi \in Q \cdot (A \cdot Q)^*$，我们将柱集 cyl($\pi$) := $\{\pi \cdot \rho' \mid \rho' \in (A \cdot Q)^\omega\} \in$Borel($Q \cdot A$) 的概率定义为 $\mathbb{P}^\sigma_\rho[\text{G}, \text{cyl}(\pi)]$: $\mathbb{P}^\sigma_\rho[\text{G}, \text{cyl}(\pi)] := \mathbb{P}^\sigma_\rho$(G π) $\in [0,1]$。我们将其自然地扩展为博雷尔集上的概率测度: $\mathbb{P}^\sigma_\rho[\text{G} \cdot]$: Borel($Q \cdot A$) $\to [0,1]$。通过投影，我们得到 Borel(Q) 中博雷尔集的概率测度。

多环境马尔可夫决策过程。多环境马尔可夫决策过程（简称 MEMDP） Γ 是一个元组$\Gamma = (Q, A, E, (\delta_e)_{e\in E})$，其中 E是一个非空的有限环境集，并且对于所有 $e \in E$，(Q, A, δ_e) 都是一个马尔可夫决策过程，记作 $\Gamma[e]$。除非另有说明，MEMDP Γ 指元组 $\Gamma = (Q, A, E, (\delta_e)_{e\in E})$。对于所有 $e \neq e' \in E$，若状态-动作对 $(q, a) \in Q \times A$ 满足 (e, e')-区分性条件 $\delta_e(q,a) \neq \delta_{e'}(q,a)$，则称其为 (e, e')-区分性状态-动作对。我们用 Dstg(Γ, e, e') 表示所有 (e, e')-区分性状态-动作对组成的集合。我们还令 Dstg(Γ) := $\cup_{e\neq e'\in E}$Dstg(Γ, e, e')。

多环境马尔可夫决策过程中的值。考虑一个多环境马尔可夫决策过程，以及一个我们希望最大化其概率的奇偶目标。在多环境马尔可夫决策过程中合成策略时，我们并不知道它在哪个环境中执行。因此，在定义所合成策略的值时，我们可以要求它在所有 环境中都表现良好，或者在给定关于运行环境的先验信念时，要求它在平均意义上表现良好。本文的目标是研究后者的“先验”值，并将其与前者的“唯一”值联系起来。具体而言，给定多环境马尔可夫决策过程 Γ、状态 $q \in Q$ 以及奇偶目标 $W \in$ 博雷尔 (Q)，我们定义 uni 值

$$\text{val}^{\text{uni}}_q(\Gamma, W) := \sup_{\sigma\in\text{Strat}(Q,A)} \text{val}^{\text{uni}}_q(\Gamma, \sigma, W) \text{ with } \text{val}^{\text{uni}}_q(\Gamma, \sigma, W) := \min_{e\in E} \mathbb{P}^\sigma_q[\Gamma[e], W]$$

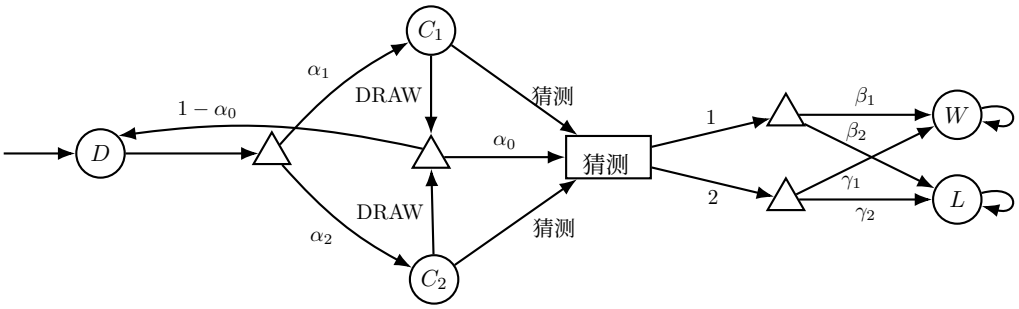

图1：该图描绘了一个多环境马尔可夫决策过程（灵感源自 [7]），用于建模一个简单的纸牌游戏。游戏使用一副包含两种牌（1 和 2）的牌堆。牌堆的组成由 α_1和 α_2参数化，其中 $\alpha_1+\alpha_2=1$；例如，如果牌 1 出现的次数是牌 2 的两倍，则 $\alpha_1=2\alpha_2$。每一回合，玩家可以选择 猜测 哪种牌的数量占多数，或要求从牌堆中额外 抽牌；提出抽牌请求后，游戏以 $1-\alpha_0$ 的概率继续进行，而以 α_0 的概率，玩家会被迫立即猜测。玩家的目标是到达获胜状态 W，即正确猜出牌堆中数量较多的牌的类型；通过适当选择参数即可实现这一点（例如，β 和 γ），这些参数决定了从猜测动作到 W 或 L的转移。

以及先验值，给定先验信念 $b\in\mathcal{D}(E)$

$$\mathsf{val}_q^{\mathsf{pr}}(\Gamma,b,W):=\sup_{\sigma\in\mathsf{Strat}(Q,A)}\mathsf{val}_q^{\mathsf{pr}}(\Gamma,b,\sigma,W)\text{ with }\mathsf{val}_q^{\mathsf{pr}}(\Gamma,b,\sigma,W):=\sum_{e\in E}b(e)\cdot\mathbb{P}_q^{\sigma}[\Gamma[e],W]$$

示例。 考虑图 1，它表示一个多环境马尔可夫决策过程。状态、动作和概率后继状态的定义与经典马尔可夫决策过程中的定义相同，关键区别在于，这里我们考虑对标记概率转移的参数进行多种赋值（即三角形节点出边上的参数）。

首先，假设我们要对一副牌进行建模，其中牌 1 有两张，牌 2 有一张。在这种情况下，抽到牌 1 的概率为 $\alpha_1=\frac{2}{3}$，而抽到牌 2 的概率为 $\alpha_2=\frac{1}{3}$。此外，假设玩家在猜牌之前始终可以选择再抽一次；这可通过设置 $\alpha_0=0$ 来表示。

为了对重复牌是牌 1 这一事实进行建模，我们将参数设为 $\beta_1=1$ and$\beta_2=0$，这表示猜牌 1 必然到达获胜状态 W，而猜牌 2 则会到达失败状态 L。同理，我们设置 $\gamma_1=0$ and $\gamma_2=1$。我们将这一参数赋值称为环境 E_1。我们还可以定义另一个环境，记为E_2，用于表示牌 2 重复的情况。此时，参数设为 $\alpha_1=\frac{1}{3}$ and $\alpha_2=\frac{2}{3}$，以及 $\beta_1=0$、 $\beta_2=1$、 $\gamma_1=1$ 和 $\gamma_2=0$。

最后，为了编码可达性目标（即到达状态 W），我们使用如下奇偶标记：除获胜状态 W 的标记为 2 外，所有状态的标记均为 1。在这种标记下，当且仅当对局最终到达 W，即玩家正确猜出重复牌时，奇偶目标才得以满足。

3 定性与定量问题

本节重点讨论如何判定是否存在具有足够好先验值的策略。我们首先研究定性问题（值为 1），然后转向定量问题。

3.1 定性问题

当取值（任意接近）1 时，通用设定与先验设定一致。

命题 *1*（证明 *A.1*）。考虑一个多环境马尔可夫决策过程 Γ、一个状态 $q \in Q$ 和一个奇偶目标 W。令 $b \in \mathcal{D}(E)$，使得 $\mathit{Supp}(b) = E$。那么 $\mathit{val}_q^{\mathsf{uni}}(\Gamma, W) = 1$ 当且仅当 $\mathit{val}_q^{\mathsf{pr}}(\Gamma, b, W) = 1$，并且对于所有 $\sigma \in \mathit{Strat}(Q, A)$：$\mathit{val}_q^{\mathsf{uni}}(\Gamma, \sigma, W) = 1$ 当且仅当 $\mathit{val}_q^{\mathsf{pr}}(\Gamma, b, \sigma, W) = 1$。

证明概要。如果 $\mathrm{val}_q^{\mathsf{uni}}(\Gamma, W) = 1$，那么存在一些策略，其唯一值可以任意接近 1，因此它们的先验值也可以任意接近 1。因此，$\mathrm{val}_q^{\mathsf{pr}}(\Gamma, W) = 1$。此外，如果某个策略相对于 b 的先验值至少为 $1 - d \cdot \varepsilon$，其中某个 $\varepsilon > 0$ 满足 $d := \min_{e \in E} b(e) > 0$，那么它在任何环境中的最小值（因为 $\mathrm{Supp}(b) = E$）至少为 $1 - \varepsilon$。因此，如果 $\mathrm{val}_q^{\mathsf{pr}}(\Gamma, W) = 1$，那么 $\mathrm{val}_q^{\mathsf{uni}}(\Gamma, W) = 1$。另一个等价关系显而易见。 □

因此，在具有奇偶目标的多环境马尔可夫决策过程中，无论采用通用设定还是先验设定，判定是否存在几乎必胜策略（值为 1）以及是否存在极限必胜策略（值可以任意接近 1）的复杂度都相同。由此可得下面的推论。

推论 2 (根据 [7, 定理 4, 13])。对于给定先验信念和奇偶目标的多环境马尔可夫决策过程，判断是否存在具有 pr值（可任意接近）*1* 的策略，是多项式空间完全问题。当环境数量固定时，可在多项式时间内求解。

示例。再考虑示例 2中的多环境马尔可夫决策过程。不难看出，没有任何策略能够保证以概率 1 获胜。事实上，经过有限次抽牌后，经验频率仍可能与典型情况相差甚远，因而错误地指示哪张卡片是重复牌。

尽管如此，对于每个 $\varepsilon > 0$，玩家都可以请求足够多次抽牌，使观察到这种误导性样本的概率降至 ε 以下。根据大数定律，这意味着，尽管几乎必然获胜是不可能的，玩家仍可通过选择适当的策略族（以 ε为索引），使获胜概率任意接近 1。

无论采用全称语义还是先验语义，只要先验分布对两种环境都赋予正概率，上述推理便适用。在这两种情况下，都不存在确定必胜策略，但存在一个能够证明极限必然获胜的策略族；特别是，该构造适用于这两种环境上的任意固定先验分布。

3.2 定量问题

现在我们转向定量问题。我们设计了一种算法，对于任意给定的精度水平 $\varepsilon > 0$，计算 pr 值 $\mathrm{val}_q^{\mathsf{pr}}(\Gamma, b, W)$ 的 ε-近似。随后，我们将采用这一近似过程来解决相关的间隙问题，其定义如下：给定参数 $\varepsilon > 0$，若 $\mathrm{val}_q^{\mathsf{pr}}(\Gamma, b, W) \geq \alpha$，则必须输出 YES；若 $\mathrm{val}_q^{\mathsf{pr}}(\Gamma, b, W) \leq \alpha - \varepsilon$，则必须输出 NO；其他情况下可任意作答。由于有些输入不要求给出特定输出，因此该间隙问题并非经典的判定问题，而是一个承诺问题 [10]。

我们的目标是根据先验环境信念 $b \in \mathcal{D}(E)$ 计算 pr 值的近似。在多环境马尔可夫决策过程中，每当遇到区分性状态-动作对时，环境信念都会更新：当我们访问转移 $(q, a, q') \in Q \times A \times Q$ 使得$\delta_e(q, a)(q') > \delta_{e'}(q, a)(q')$ 对某个 $e \neq e' \in E$ 成立时，信念会更新，以反映 e 的似然增加以及 e'中的似然降低。下面我们正式定义这一信念更新。

定义*3*（信念更新）。考虑一个多环境马尔可夫决策过程 $\Gamma = (Q, A, E, (\delta_e)_{e \in E})$。令 $b \in \mathcal{D}(E)$。对于所有 $(q, a) \in Q \times A$，我们令 $p[b, q, a] \in \mathcal{D}(Q)$ 定义如下：对于所有 $q' \in Q$：$p[b, q, a](q') := \sum_{e \in E} b(e) \cdot \delta_e(q, a)(q')$。然后，对于所有 $q' \in Q$，如果 $p[b, q, a](q') = 0$，我们令 $\lambda[b, q, a, q'] \in \mathcal{D}(E)$ 为任意分布；否则，对于所有 $e \in E$：$\lambda[b, q, a, q'](e) := \frac{b(e) \cdot \delta_e(q, a)(q')}{p[b, q, a](q')}$。

给定任意初始状态 $q \in Q$ 和先验信念 $b \in \mathcal{D}(E)$，我们定义一个（“似然”）函数 $lk_b : Q \cdot (A \cdot Q)^* \to \mathcal{D}(E)$，它会迭代更新环境信念：对于所有 $q \in Q$，$lk_b(q) := b \in E$；并且，对于所有 $\rho \cdot (a, q') \in Q \cdot (A \cdot Q)^+$，我们令 $lk_b(\rho \cdot (a, q')) := \lambda[lk_b(\rho), last(\rho), a, q']$。

示例。让我们通过正在讨论的示例来说明信念更新的运作方式。假设两个环境上的先验由分布 $b = (\frac{1}{4}, \frac{3}{4})$ 给出，因此初始信念偏向环境 E_2（即偏向牌 2 被重复的情况）。现在假设第一次抽牌后，我们观测到牌 2，也就是说，我们到达状态 2。这一观测应当增强我们对运行环境确实为 E_2 的信心。应用上述更新规则，后验信念变为 $\lambda[b, D, a, C_2](E_2) = \frac{3}{4} \cdot \frac{\frac{2}{3}}{\frac{1}{3} \cdot \frac{1}{4} + \frac{2}{3} \cdot \frac{3}{4}} = \frac{3}{4} \cdot \frac{24}{21} = \frac{6}{7}$，且 $\lambda[b, D, a, C_2](E_1) = \frac{1}{7}$。反之，如果第一次抽到牌 1，那么后验信念将变为 $\lambda[b, D, a, C_1](E_2) = \frac{3}{4} \cdot \frac{\frac{1}{3}}{\frac{2}{3} \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{3}{4}} = \frac{3}{4} \cdot \frac{12}{15} = \frac{3}{5}$，且 $\lambda[b, D, a, C_1](E_1) = \frac{2}{5}$。

我们用于计算pr值近似的算法，会根据信念支撑集中环境的数量递归运行。当支撑集中只有一个环境时，实际上我们面对的是一个马尔可夫决策过程，可以在多项式时间内计算所有状态 [2]的值。现在假设先验信念的支撑集大小至少为 2。如果某个环境e中的信念降为 0，这是因为我们访问了一个转移 (q, a, q')，满足 $\delta_e(q, a)(q') = 0$，那么我们可以用支撑集更小的信念递归调用该算法。或者，某个环境中的信念也可能接近 0。在这种情况下，根据pr值的 1-Lipschitz 连续性，将该信念截断为支撑集更小的信念几乎不会造成损害（就值的变化而言）。

引理 *4*（证明 *A.2*）。考虑一个多环境马尔可夫决策过程 Γ 和一个奇偶目标 W。对于所有信念 $b, b' \in \mathcal{D}(E)$，有 $|val_q^{pr}(\Gamma, b, W) - val_q^{pr}(\Gamma, b', W)| \leq$ 差异(b, b')，其中差异$(b, b') := \frac{1}{2} \cdot \sum_{e \in E} |b(e) - b'(e)|$。此外，对于所有满足 $|Supp(b)| \geq 2$ 的信念 $b \in \mathcal{D}(E)$，存在 $Truncate(b) \in \mathcal{D}(E)$，使得 $Supp(Truncate(b)) \subsetneq Supp(b)$ 且 $Diff(b, Truncate(b)) \leq min_{e \in Supp(b)} b(e)$。

每次访问区分性状态-动作对时，都会更新环境信念。然而，即使访问了许多区分性状态-动作对，也可能不存在任何一个环境的信念降至 0。尽管如此，如果访问了足够多的区分性状态-动作对，那么以高概率会有某个环境的信念降至接近 0。

定理 *5*（证明*A.3*）。考虑一个 $MEMDP\ \Gamma = (Q, A, E, (\delta_e)_{e \in E})$ 和一个先验环境信念 $b \in \mathcal{D}(E)$。令 $k := |E|$，并令 l 表示写出 Γ 中出现的任意概率所需的最大比特数。对于所有 $\varepsilon > 0$，令：$NeverSmallBelief(b, \varepsilon) := \{\rho \in Q \cdot (A \cdot Q)^* \mid \forall \pi \sqsubseteq \rho, \forall e \in E : lk_b(\pi)(e) > \varepsilon\}$，其中 $\pi \sqsubseteq \rho$ 表示 π 是 ρ 的前缀。对于所有 $\rho \in (Q \cdot A)^\omega$，令 $nb^{dstg}(\rho) \in \mathbb{N} \cup \{\infty\}$ 表示 ρ 访问区分性状态 - 动作对的次数。

那么，对于所有 $\varepsilon > 0$，令 $O\ poly\ k\ 2^l\ x$ $\sigma \in Q\ A$ $e \in E$ $:= \lceil \log(\frac{1}{\varepsilon})\rceil$，存在 $(\Gamma,) \in \mathbb{N}$，使得 $(\Gamma,)((, ,))$ 成立，从而对所有策略 $Strat(,)$ 和环境 $:$ 均成立：

$$\mathbb{P}_q^\sigma(\Gamma[e], \{\mathsf{nb}_{\mathsf{dstg}} \geq m(\Gamma, \varepsilon)\} \cap \mathsf{NeverSmallBelief}(b, \varepsilon)) \leq \varepsilon$$

该定理是本文的一项关键结果，其证明相当复杂。下文仅介绍该证明的两个主要组成部分；所有细节请见附录 A.3.2，其中量 $m(\Gamma, \varepsilon)$ 的定义见附录 A.3.1。

证明概要。假设某个 $e \in \mathsf{Supp}(b)$ 是运行环境。考虑另一个环境 $e' \neq e \in \mathsf{Supp}(b)$，并为简单起见，假设对于所有 $(q, a) \in Q \times A$，$\mathsf{Supp}(\delta_e(q, a)) = \mathsf{Supp}(\delta_{e'}(q, a))$。我们的目标是证明，无论策略 σ 如何，以高概率，访问了足够多 (e, e')-区分状态-动作对的运行 ρ 都满足 $\mathsf{lk}_b(\rho)(e')$ 可以任意小。我们关注环境 e 和 e' 中信念的比值，该比值初始时等于 $b(e')\ b(e)$。经过某次运行 $\rho \in Q \cdot (A \cdot Q)^*$ 后，它等于 $\mathsf{lk}_b(\rho)(e')\ \mathsf{lk}_b(\rho)(e) = b(e')\ b(e) \cdot \prod_{\pi \cdot (a,q) \sqsubseteq \rho} \frac{\delta_{e'}}{\delta_e}(\mathsf{last}(\pi)\ a)(q)\ (\mathsf{last}(\pi)\ a)(q)$。由于随机变量之和比它们的乘积更容易分析，我们旨在证明 $(*)$：“无论策略 σ 如何，以高概率，如果 ρ 访问了足够多 (e, e')-区分状态-动作对，那么总和 $\sum_{\pi \cdot (a,q) \sqsubseteq \rho} \log(\delta_{e'}(\mathsf{last}(\pi)\ a)(q)\ \delta_e(\mathsf{last}(\pi)\ a)(q))$ 足够小。”证明这一事实的一个关键结果是，存在某个 $\eta < 0$，使得对于所有

$(q,\ a) \in \mathsf{Dstg}(\Gamma,\ e,\ e')$: $\sum_{q' \in Q} \delta_e(q,\ a)(q') \cdot \log(\frac{\delta_{e'}(q,\ a)(q')}{\delta_e(q,\ a)(q')}) \leq \eta$。这是算术平均数与几何平均数之差所导致的。

为了证明（$*$），我们想使用霍夫丁不等式 [11]（这也是 [7] 中使用的核心工具）。非正式地说，这个不等式意味着：如果我们考虑 $n \in \mathbb{N}$ 个相互独立的实值随机变量，那么它们的和大于其期望值的概率会随着 n 指数级下降。在我们的例子中，我们可以考虑随机变量 X_k，其中 $k \in \mathbb{N}$，它们将访问到的第 k 次转移（q、a、q'）$\in Q \times A \times Q$（满足（q、a）$\in \mathsf{Dstg}$（Γ、e、e'））映射为实数值 $\log(\frac{\delta_{e'}(q,a)(q')}{\delta_e(q,a)(q')}) \in \mathbb{R}$。然而，这些随机变量显然并不相互独立，因此无法应用霍夫丁不等式。尽管如此，由于我们可以证明这些随机变量的期望值都不超过 $\eta < 0$，因此我们可以改写霍夫丁不等式的证明，从而证明：无论策略如何 σ，以关于 n 指数级小的概率，n 个随机变量 X_1、.....、X_n 的和都不会显著高于 $n \cdot \eta < 0$。

总的来说，如果我们恰当地选择 n，就可以得到：在环境 e中，对于所有策略 σ，在对环境的信念 e' 始终不低于 ε 的情况下，访问至少 n (e, e')个区分性状态-动作对的概率至多为 $\varepsilon > 0$。数量 $m(\Gamma, \varepsilon)$的选择方式保证了：如果至少访问了 $m(\Gamma, \varepsilon)$个区分性状态-动作对，那么对于所有环境$e \in E$，都存在一个环境 $e' \in E$，使得至少访问 n (e, e')个区分性状态-动作对。定理得证。□

示例。让我们考虑前面所述的、环境为 E_1 和 E_2 的运行示例，只是现在考虑一个较小的 $\alpha_0 > 0$，这样策略就能以较高概率在必须做出猜测之前收集许多观测。正如我们在上面关于信念更新的示例中所看到的，访问状态 C_2 会增加对环境 E_2的信念，而访问状态C_1 会增加对环境 E_1的信念。上述定理的一个推论是，以较高概率，访问揭示卡片的状态的次数会极度偏向于 C_1（并且对 E_2 的信念会降至接近0）或 C_2（并且对 E_1 的信念会降至接近0）。一旦出现这种情况，我们就可以几乎无害地截断信念（回顾引理4），从而得到一个常规马尔可夫决策过程，并可在多项式时间内计算其值。

解决间隙问题的算法。我们设计了算法 1，给定一个多环境马尔可夫决策过程Γ、一个奇偶目标 W、一个先验环境信念 $b \in \mathcal{D}(E)$ 以及一些 $\gamma > 0$，该算法会为每个状态 $q \in Q$计算一个值 v_q，使得 $|\mathsf{val}^{\mathsf{pr}}_q(\Gamma, b, W) - v_q| \leq \gamma$。算法还接受一个整数 $n \in \mathbb{N}$ 作为额外参数，该整数限定了环境信念被截断之前可访问的区分性状态-动作对的最大数量。调用该算法时应使用 $n := m(\Gamma, \frac{\gamma}{3|E|})$[1]。该算法递归地处理信念支撑集中环境的数量；基本情形在第 1 行处理：对于具有奇偶目标的马尔可夫决策过程，我们可以在多项式时间内计算所有状态的值。接着，如果我们已访问足够多的区分性状态-动作对（$n = 0$），或者最小的正信念值足够接近 0，我们就会递归调用算法，处理支撑集更小的截断环境信念（第 2–4 行）。否则，对于每个区分性状态-动作对（q, a）以及满足$p[b, q, a](q') > 0$的状态 $q' \in Q$（回顾定义3），我们通过递归调用算法来计算该状态 q' 在更新后的信念 $\lambda[b, q, a, q'] \in \mathcal{D}(E)$ 下的值 $v_{q,a,q'} \in [0, 1]$；递归调用时使用的界限 n'，若信念的支撑集已缩小，则重置为 $m(\Gamma, \frac{\gamma}{3|E|})$，否则保持为 $n-1$（第 5–8 行）。随后，我们通过添加两个新的汇状态 q_{win} 和 q_{lose}，将多环境马尔可夫决策过程转化为马尔可夫决策过程：把所有满足 (q, a) 是区分状态-动作对的转移 $(q, a, q') \in Q \times A \times Q$，以概率 $v_{q,a,q'}$ 重定向至状态 q_{win}，并以概率 $1 - v_{q,a,q'}$重定向至状态 q_{lose}；我们还将奇偶目标 W 改为奇偶目标 W'，使得在 q_{win} 上循环为获胜，而在 q_{lose} 上循环为失败（第 9 行）。（这一变换的形式化描述见

[1]我们使用 $\frac{\gamma}{3|E|}$ 而不是 γ，因为该算法通过截断环境信念进行了多次近似，这些近似的总和应至多为 γ。

算法 1 MEMDP-Prior-Parity(Γ, W, b, n, γ)

输入：多环境马尔可夫决策过程 $\Gamma = (Q, A, E, (\delta_e)_{e\in E})$，奇偶目标 W，信念 $b \in \mathcal{D}(E)$，$n \in \mathbb{N}$，$\gamma > 0$

1: **if** $|\text{支持集}(b)| = 1$ **then return** MDP-Parity(Γ, W) : $Q \to [0,1]$
2: **if** $n = 0$ 或 $\text{最小值}_{e\in\text{Supp}(b)}\, b(e) \le \frac{\gamma}{3|E|}$ **then**
3: $b' \leftarrow$截断(b) ▷ 如引理 4：支持集(b') $\subsetneq$支持集(b)
4: **返回** MEMDP-Prior-Parity(Γ, W, b', $m(\Gamma, \frac{\gamma}{3|E|})$ γ)
5: **对于** $(q, a) \in \text{Dstg}(\Gamma)$ 和 $q' \in$支持集($p[b,q,a]$) **执行**
6: $b' \leftarrow \lambda[b,q,a,q']$ ▷ 如定义3
7: **若** 支持集(b') $\subsetneq$支持集(b) **则** $n' \leftarrow \frac{\gamma}{3|E|}$ **否则** $n' \leftarrow n-1$
8: $v_{q,a\,q'} \leftarrow$ MEMDP-Prior-Parity(Γ, W, b', n', m)(q')
9: (G W') $\leftarrow$从MEMDP构造MDP(Γ, W, b, v)
10: **返回**MDP-Parity(G W') : $Q \to [0,1]$

附录 A.6.1。）最后，我们可以计算该马尔可夫决策过程在目标W'下各状态的值（第 10 行）。这些值对应于多环境马尔可夫决策过程 Γ的值。

现在考虑算法 1 的复杂度。首先请注意，每当它递归调用自身时，要么信念的支撑集缩小，要么（重置为 $m(\Gamma, \frac{\gamma}{3|E|})$ 的）已访问区分状态-动作对数量上界减小。因此，递归深度上界为 $|E| \times m(\Gamma, \frac{\gamma}{3|E|})$。此外，用于描述环境信念的比特数会随着更新和截断的次数线性增长。事实上，该算法占用的空间为 $O(poly(|Q|, |A|, |E|, |\log(\gamma)|, x_b, m(\Gamma, \frac{\gamma}{3|E|})))$，其中 x_b 是表示先验信念所需的比特数 b。算法 1 的所有细节见附录 A.6。

间隙问题可以通过调用算法 1 并使用 $\gamma := \varepsilon/3$，然后将结果与 α进行比较来求解。根据 $m(\Gamma, \frac{\gamma}{3|E|})$ 的界限，由定理 5 可得下述定理。

定理 6。对于一个多环境马尔可夫决策过程 Γ，给定先验信念 $b \in \mathcal{D}(E)$（以二进制编码给出）、奇偶目标、阈值 $0 < \alpha < 1$（以二进制编码给出）以及精度 $\varepsilon > 0$（以二进制编码给出），带有pr-值的间隙问题可以在指数空间中判定。如果多环境马尔可夫决策过程中涉及的概率以一元编码给出，则该间隙问题可以在多项式空间中判定。

4 关联pr-值与uni-值

中所述的过程 [7] 可以求解带有 uni-值的间隙问题，其所需空间对于 $|Q|$（状态数）呈双指数增长。另一方面，当概率以一元编码给出时，我们已经给出了一个在多项式空间内求解带有pr-值的间隙问题的算法；在其他情况下，该算法需要指数空间。现在，我们的目标是联系 pr-值与 uni-值，从而利用我们的算法更高效地求解带有 uni-值的间隙问题。

显然，对于任何先验信念，唯一值都小于或等于 pr 值。事实上，唯一值等于所有可能先验信念下的 pr 值的下确界。

定理 7. 考虑一个多环境马尔可夫决策过程 $\Gamma = (Q, A, E, (\delta_e)_{e\in E})$、一个状态 $q \in Q$ 和一个奇偶目标$W \subseteq$博雷尔(Q)。我们有：$\mathsf{val}^{\mathsf{uni}}_q(\Gamma, W) =$下确界$_{b\in\mathcal{D}(E)} \mathsf{val}^{\mathsf{pr}}_q(\Gamma, b, W)$。

示例。我们重新审视这个贯穿全文的示例，并按如下方式修改环境 E_1 和 E_2。在这两个环境中，我们设定 $\alpha_0 = 1$，因此玩家必须在抽一次牌后立即猜出重复牌。若环境的其他方面均保持不变，则唯一值为 $\frac{2}{3}$；采用根据唯一观察到的牌来猜测对应环境的策略即可达到该值。在两个环境上采用均匀先验时，先验值与这个 $\frac{2}{3}$ 值相同。现在我们引入非对称环境。在 E_1中，牌 1 重复两次，牌 2 重复一次（因此 $\alpha_1 = \frac{3}{5}$）；而在 E_2中，牌 2 重复两次（因此

$\alpha_1 = \frac{1}{4}$）。在这种情况下，唯一值仍等于 2_3；然而，先验值的下确界现在是在先验 b 下取得的，该先验略微偏向较不利的环境，即 $b(E_1) := \frac{5}{9}$，而不是在环境上的均匀先验下取得。

为了证明定理 7，我们引入混合策略的概念。给定状态集 Q和动作集 A，在 (Q, A) 上的混合策略是一个概率分布 $\tau \in \mathcal{D}(\text{Strat}(Q, A))$，定义在策略之上（尽管集合 $\text{Strat}(Q, A)$ 不可数，但我们考虑的概率分布具有可数支撑）。在马尔可夫决策过程 $\mathsf{G} = (Q, A, \delta)$ 中，我们自然地定义由混合策略 $\tau \in \mathcal{D}(\text{Strat}(Q, A))$ 诱导的概率测度 $\mathbb{P}^{\tau}_{\rho}[\mathsf{G}\ \cdot] : \text{Borel}(Q) \to [0, 1] : \mathbb{P}^{\tau}_{\rho}[\mathsf{G}\ \cdot] := \sum_{\sigma \in \text{Strat}(Q,A)} \tau(\sigma) \cdot \mathbb{P}^{\sigma}_{\rho}[\mathsf{G}\ \cdot]$。对于所有博雷尔目标 $W \in \text{Borel}(Q)$，给定先验信念 $b \in \mathcal{D}(E)$，我们自然地将唯一值 $\text{val}^{\text{uni}}_q(\Gamma, \tau, W) \in [0, 1]$ 和先验值 $\text{val}^{\text{uni}}_q(\Gamma, b, \tau, W) \in [0, 1]$ 扩展到混合策略 $\tau \in \mathcal{D}(\text{Strat}(Q, A))$。然后，我们可以将混合策略所能达到的唯一值上确界与所有先验信念下先验值的下确界联系起来。

引理 *8*（证明 *B.1*）。考虑一个多环境马尔可夫决策过程 Γ、一个状态 $q \in Q$ 和一个奇偶目标 W。我们有：$\sup_{\tau \in \mathcal{D}(\text{Strat}(Q,A))} \inf_{b \in \mathcal{D}(E)} \text{val}^{\text{pr}}_q(\Gamma, b, \tau, W) = \inf_{b \in \mathcal{D}(E)} \sup_{\tau \in \mathcal{D}(\text{Strat}(Q,A))} \text{val}^{\text{pr}}_q(\Gamma, b, \tau, W)$。因此：$\sup_{\tau \in \mathcal{D}(\text{Strat}(Q,A))} \text{val}^{\text{uni}}_q(\Gamma, \tau, W) = \inf_{b \in \mathcal{D}(E)} \text{val}^{\text{pr}}_q(\Gamma, b, W)$。

证明思路。第一个等式直接源于冯·诺伊曼极小极大定理的一个标准推广 [20]，，该推广成立是因为环境集合是有限的 [18]。此外，对于所有非空集合 X 和 $f : X \to [0, 1]$，我们有 $\sup_{d \in \mathcal{D}(X)} \sum_{x \in X} d(x) \cdot f(x) = \sup_{x \in X} f(x)$，且 $\inf_{d \in \mathcal{D}(X)} \sum_{x \in X} d(x) \cdot f(x) = \inf_{x \in X} f(x)$。因此，对于所有 $\tau \in \mathcal{D}(\text{Strat}(Q, A))$，都有 $\inf_{b \in \mathcal{D}(E)} \text{val}^{\text{pr}}_q(\Gamma, b, \tau, W) = \text{val}^{\text{uni}}_q(\Gamma, \tau, W)$；并且对于所有 $b \in \mathcal{D}(E)$，都有 $\sup_{\tau \in \mathcal{D}(\text{Strat}(Q,A))} \text{val}^{\text{pr}}_q(\Gamma, b, \tau, W) = \text{val}^{\text{pr}}_q(\Gamma, b, W)$。 □

为了证明定理 7，现在只需证明，在 $\mathcal{D}(\text{Strat}(Q, A))$ 中，混合策略所达到的 uni 值不会高于 $\text{Strat}(Q, A)$ 中的策略，正如下方引理所述。

引理 9 (证明 B.2)。考虑一个多环境马尔可夫决策过程 Γ、一个状态 $q \in Q$以及一个奇偶目标 $W \in$ 博雷尔 (Q)。我们有：$\sup_{\tau \in \mathcal{D}(\text{Strat}(Q,A))} \text{val}^{\text{uni}}_q(\Gamma, \tau, W) = \text{val}^{\text{uni}}_q(\Gamma, W)$。

证明概要。给定任意 $\tau \in \mathcal{D}(\text{Strat}(Q, A))$，我们定义 $\sigma \in \text{Strat}(Q, A)$，使得对所有 $e \in E$ 以及博雷尔目标 $W \in \text{Borel}(Q)$：$\mathbb{P}^{\tau}_q[\Gamma[e], W] = \mathbb{P}^{\sigma}_q[\Gamma[e], W]$。因此，$\text{val}^{\text{uni}}_q(\Gamma, \tau, W) = \text{val}^{\text{uni}}_q(\Gamma, \sigma, W) \leq \text{val}^{\text{uni}}_q(\Gamma, W)$。引理得证。 □

如何使用定理 7。 引理 8和 9 结合起来可推出定理 7。该定理表明，计算 uni-值相当于计算所有先验信念下 pr-值的下确界。此外，引理 4 表明，由两个相近信念诱导的 pr-值相差不大。因此，我们可以通过计算足够多的先验信念对应的 ε-近似 uni-值来得到对pr-值的近似；这些先验信念可以紧密覆盖所有信念构成的集合。由此我们推导出下面的定理，它显著改进了 [7]中确立的双指数空间复杂度。

定理 10 (证明 B.3)。对于给定奇偶目标、阈值 $0 < \alpha < 1$ （二进制编码）和精度 $\varepsilon > 0$ （二进制编码）的多环境马尔可夫决策过程，uni-值的间隙问题可在指数空间内判定。如果概率采用一元编码给出，则该间隙问题可在 多项式空间内判定。

证明概要。令 $N _= \lceil \frac{1}{\varepsilon} \rceil$ $k = |E|$ $S = \{b \in \mathcal{D}\ E \mid \forall e \in E\ \ b\ e \in \{\frac{x}{k \cdot 2^{N+1}} __: 0 \leq x \leq k \cdot 2^{N+1}\}\}$ $|S| = O\ 2^{k^2 \cdot\ N+1}$ $b \in \mathcal{D}\ E$ $b' \in S$ log( ) , : , and : ( ) ,( ) 。我们有 (()，并且对于所有 ( )，都存在使得 差异$(b, b') \leq \frac{\varepsilon}{2}$。枚举 S中的所有信念，可以在关于 k和 N的多项式空间内完成。此外，S 中的所有信念都可以用关于 k 和 N的多项式位数来描述。因此，对 S 中的任意信念执行算法 1，其中 $\gamma := \varepsilon/2$，需要关于输入的指数空间（如果概率以一元编码给出，则需要关于输入的多项式空间）。 □

现在，我们的目标是利用定理 7 推导近似 pr 值的复杂度下界。对于以一元编码表示的概率，我们已经确立了多项式空间上界。文献 [16, 定理 26] 指出，在概率以一元编码表示的双环境 MEMDP 中，具有 uni 值和可达性目标的间隙问题是 NP难的。因此，下面我们着重讨论如何近似双环境马尔可夫决策过程中的 uni 值。如下文所述，只需调用多项式次近似 pr 值的预言机，即可完成这项工作，从而迁移上述 NP难性结果。

命题 11 (证明B.4)。 在一个具有两个环境的多环境马尔可夫决策过程 Γ 中，给定奇偶目标W和精度 $\varepsilon > 0$，令 $N := \lceil \log(\frac{1}{\varepsilon}) \rceil$，我们可以在关于 ε的多项式时间内计算唯一值的 N-近似，并且只需 $O(N)$ 次调用预言机，该预言机计算信念上的 γ-近似的 pr值，且信念的描述所需比特数关于 γ 为多项式。 N

证明概要。令 $E = \{e_1, e_2\}$。我们考虑 $f\colon [0,1] \to [0,1]$ 使得，对于所有 $x \in [0,1]$, $f(x) := \mathrm{val}_q^{\mathrm{pr}}(\Gamma, b_x, W)$，其中 $b_x \in \mathcal{D}(E)$ 满足 $b_x(e_1) := e$。根据定理 7，下确界 f值就是唯一值。函数 f是拟凸的：对于所有 $x < y < z \in [0,1]$: $f(y) \leq \max(f(x)\ f(z))$。这意味着，对于所有 $x < y < z < t$，若 $f(z) < f(y)$，则 $\inf_{u\in[x,t]} f(u) = \inf_{u\in[y,t]} f(u)$；若 $f(y) < f(z)$，则 $\inf_{u\in[x,t]} f(u) = \inf_{u\in[x,z]} f(u)$。基于这一观察，我们设计了一种类似二分查找的过程：通过在长度不断缩小的区间$[x,t]$ 中搜索，找到最小 f值的一个近似，直到区间长度足够小（这就足够了，因为 f是 1-Lipschitz 连续的）。

□

推论 12 (证明B.5)。 对于概率采用一元编码的双环境 *MEMDP*，除非 P-值间隙问题且目标为奇偶目标，否则无法在多项式时间内判定 P =NP。

5 多环境马尔可夫决策过程的熵刻画

在多环境马尔可夫决策过程中进行对局时，我们只能部分了解自己所处的位置：我们知道当前状态，却不知道运行环境。实际上，多环境马尔可夫决策过程是部分可观测马尔可夫决策过程（简称 POMDP）的一种特殊形式；也就是说，在这类马尔可夫决策过程中，我们在一组底层状态上进行对局，但只能通过观测间接获取状态信息，而且不同状态可能对应相同的观测。形式化地说，POMDP 是一个元组 $\Lambda = (Q, A, \delta, \Omega\ O)$，其中 (Q, A, δ) 是一个马尔可夫决策过程，Ω 是一个非空的观测集合，且 $O\colon Q \to \Omega$。一个多环境马尔可夫决策过程 Γ 自然会导出 POMDP $\Lambda(\Gamma) = (Q', A, \delta, \Omega\ O)$，满足 $Q' := Q \times E$；对于所有 $(q, e, q', a) \in Q \times E \times Q \times A$,$\delta((q,e)\ a)((q',e)) := \delta_e(q,a)(q')$；$\Omega := Q$；并且对于所有 $(q,e) \in Q \times E$, $O(q,e) := q$。POMDP 中的策略是 Strat$(\Omega\ A)$ 中的函数。对于任意策略 $\sigma \in$ Strat$(\Omega\ A)$，以及任意状态 $q \in Q$，我们自然地定义概率测度 $\mathbb{P}_q^\sigma[\Lambda, \cdot]\colon$ Borel$(Q) \to [0,1]$。给定 $b \in \mathcal{D}(Q)$ 和$W \in$ Borel(Q)，令 $\mathrm{val}(\Lambda, b, W) := \sup_{\sigma\in\mathrm{Strat}(O,A)} \sum_{q\in Q} b(q) \cdot \mathbb{P}_q^\sigma[\Lambda, W]$。

本节旨在刻画部分可观测马尔可夫决策过程中的多环境马尔可夫决策过程。为此，我们研究 PODMPs 中的信念。在部分可观测马尔可夫决策过程中进行对局时，我们从对当前状态的初始信念[2]开始；该信念会根据所采取的动作和收集到的观测进行更新。

定义 *13*（部分可观测马尔可夫决策过程中的信念）。考虑一个部分可观测马尔可夫决策过程 $\Lambda = (Q, A, \delta, \Omega\ O)$。信念是一个概率分布 $b \in \mathcal{D}(Q)$。给定 $b \in \mathcal{D}(Q)$、 $a \in A$ 和 $o \in \Omega$，我们令 $p(b,a,o) \in [0,1]$ 表示在给定 (b,a) 时 o 的似然： $p(b,a,o) := \sum_{q\in O^{-1}(o)} \sum_{q'\in Q} b(q') \cdot \delta(q',a)(q)$。

我们令 $\mathit{Comp}(b,a) := \{o \in \Omega \mid p(b,a,o) > 0\} \neq \emptyset$ 表示与 (b,a) 兼容的观测集合。对于所有 $o \in \mathit{Comp}(b,a)$，我们令 $\lambda[b,a,o] \in \mathcal{D}(Q)$ 表示更新后的信念，其定义如下，对于 $q \in Q$: $\lambda[b,a,o](q) := 0$ 当 $O(q) \neq o$； $\lambda[b,a,o](q) := \frac{\sum_{q'\in Q} b(q')\cdot\delta(q',a)(q)}{p(b,a,o)}$ 否则。

在多环境马尔可夫决策过程中对局时，如果我们确切知道当前环境（即环境信念是狄拉克分布），那么这种状态永远不会改变。由多环境马尔可夫决策过程导出的部分可观测马尔可夫决策过程称为

[2]在部分可观测 马尔可夫决策过程中，我们始终对当前状态有一个初始信念。这就是为什么，尽管在 he“先验”语义下的 多环境马尔可夫决策过程是部分可观测马尔可夫决策过程的特殊类型，但在“通用”语义下的多环境马尔可夫决策过程并非如此。

狄拉克保持的，也就是说，对于所有 $(q,\ a) \in Q \times A$（其中 q 视为狄拉克信念），都有：$\forall o \in \mathsf{Comp}(q,\ a)$：$|\mathsf{Supp}(\lambda[q,a,o])| = 1$。

正如 [4], 所指出的，可以通过（信息论中的）熵这一概念，采用另一种方式来描述由多环境马尔可夫决策过程诱导的部分可观测马尔可夫决策过程的行为 [17]。信念的熵 $H(b)$ 定义为信念 $b \in \mathcal{D}(Q)$ 的 $H(b)\colon = -\sum_{q\in Q} b(q)\log(b(q)) \geq 0$。信念的熵衡量该信念所承载的信息量：熵越高，承载的信息就越少。当且仅当信念是狄拉克分布时，其熵为零；熵的最大值为 $\log(|Q|)$，在均匀概率分布下达到。

在多环境马尔可夫决策过程中，我们对当前环境的信息量从不会减少：访问的区分状态-动作对越多，我们就越清楚自己正在什么环境中对局。正如 [4], 所确立的，这对应于由多环境马尔可夫决策过程诱导的部分可观测马尔可夫决策过程具有非增熵 这一性质；也就是说，对于所有 $b \in \mathcal{D}(Q)$ 和 $a \in A$：$H(b) \geq \sum_{o\in\mathsf{Comp}(b,\ a)} p(b,\ a,\ o) \cdot H(\lambda[q,a,o])$。

显然，熵不增的部分可观测马尔可夫决策过程是狄拉克保持型的。事实上，反过来也成立：所有狄拉克保持型部分可观测马尔可夫决策过程都具有非增熵。（这点较难证明。）

命题 14（证明C.2）。一个部分可观测马尔可夫决策过程具有非增熵，当且仅当它是狄拉克保持型的 g.

上述命题有两方面的益处。首先，熵不增的部分可观测马尔可夫决策过程这一概念很自然，因为它对应于关于当前状态的知识单调变化的部分可观测马尔可夫决策过程。然而，从熵不增的部分可观测马尔可夫决策过程的定义中，并不容易看出如何有效判定一个部分可观测马尔可夫决策过程是否满足这一性质。现在可以看出，这可以在多项式时间内完成，因为只需枚举所有状态、动作和观测的三元组，就能检查一个部分可观测马尔可夫决策过程是否为狄拉克保持型。其次，给定任意一个具有观测相容奇偶目标的狄拉克保持型部分可观测马尔可夫决策过程[3]，（即由标记函数 $f\colon Q \to \mathbb{N}$ 诱导的奇偶目标），使得对于任意 $O(q_1) = O(q_2)$，$f(q_1) = f(q_2)$，以及一个初始信念，我们可以构造一个规模大一个指数级的多环境马尔可夫决策过程，并为其配备一个奇偶目标和一个先验环境信念，使得该部分可观测马尔可夫决策过程中的值与该多环境马尔可夫决策过程中的 pr 值相同。下文将对此作出形式化陈述。

定理 15（证明 C.3）。考虑一个狄拉克保持型部分可观测马尔可夫决策过程 $\Lambda = (Q, A, \delta, \Omega\ O)$、一个观测相容奇偶目标 $W \in$博雷尔(Q)，以及一个初始信念 $b \in \mathcal{D}(Q)$。那么，我们可以在指数时间内计算出一个多环境马尔可夫决策过程 $\Gamma = (Q', A, E, (\delta_e)_{e\in E})$、一个奇偶目标 $W' \in$ 博雷尔(Q')，以及一个环境信念 $b' \in \mathcal{D}(E)$，使得存在一个特定状态 $q' \in Q'$，对于该状态，$val(\Lambda, b, W) = val^{\mathsf{pr}}_{q'}(\Gamma, b', W')$。

我们得到，除指数级膨胀外，多环境马尔可夫决策过程恰好等价于熵非增的部分可观测马尔可夫决策过程。此外，注意到，任意具有观测兼容奇偶目标的部分可观测马尔可夫决策过程上的极限必然 [9] 和间隙 [13] 判定问题都是不可判定的。因此，狄拉克保持型部分可观测马尔可夫决策过程构成了部分可观测马尔可夫决策过程的一个性质尤为良好的子类。

[3]观测相容奇偶目标，或称可见目标，是部分可观测马尔可夫决策过程中常见的一项假设。它们定义了玩家能够观测到的目标：在对局中行动并观测所经历的观测序列之后，玩家可以判定自己是否获胜；例如，参见 [5]。

参考文献

[1] JM Aldaz. 算术平均数与几何平均数不等式的自我改进。arXiv预印本 arXiv:0807.1788，2008年。[2] Christel Baier 和 Joost-Pieter Katoen。《模型检测原理》。麻省理工学院出版社，2008年。[3] Krishnendu Chatterjee、Martin Chmel´ík、Deep Karkhanis、Petr Novotn´y 和 Am´elie Royer。多环境马尔可夫决策过程：高效分析与应用。载于 J. Christopher Beck、Olivier Buffet、J¨org Hoffmann、Erez Karpas 和 Shirin Sohrabi（编），《第三十届国际自动规划与调度会议论文集》，法国南锡，2020年10月26日至30日，第48–56页。AAAI出版社，2020年。[4] Krishnendu Chatterjee、Martin Chmel´ık、Deep Karkhanis、Petr Novotn´y 和 Am´elie Royer。多环境马尔可夫决策过程：高效分析与应用。载于 J. Christopher Beck、Olivier Buffet、J¨org Hoffmann、Erez Karpas 和 Shirin Sohrabi（编），《第三十届国际自动规划与调度会议论文集》，法国南锡，2020年10月26日至30日，第48–56页。AAAI出版社，2020年。[5] Krishnendu Chatterjee、Laurent Doyen 和 Thomas A. Henzinger。部分可观测马尔可夫决策过程的定性分析。载于 Petr Hlinen´y 和 Anton´ın Kucera（编），《2010年计算机科学数学基础》，第35届国际研讨会，MFCS 2010，捷克共和国布尔诺，2010年8月23日至27日。论文集，《计算机科学讲义》第6281卷，第258–269页。施普林格，2010年。[6] Krishnendu Chatterjee、Laurent Doyen、Jean-Fran¸cois Raskin 和 Ocan Sankur。具有奇偶目标的多环境马尔可夫决策过程的值问题。载于 Keren Censor-Hillel、Fab-rizio Grandoni、Jo¨el Ouaknine 和 Gabriele Puppis（编），《第52届自动机、语言与编程国际研讨会》，ICALP 2025，丹麦奥胡斯，2025年7月8日至11日，《LIPIcs》第334卷，第150:1–150:17页。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik，2025年。[7] Krishnendu Chatterjee、Laurent Doyen、Jean-Fran¸cois Raskin 和 Ocan Sankur。具有奇偶目标的多环境马尔可夫决策过程的值问题。载于 Keren Censor-Hillel、Fab-rizio Grandoni、Jo¨el Ouaknine 和 Gabriele Puppis（编），《第52届自动机、语言与编程国际研讨会》，ICALP 2025，2025年7月8日至11日，丹麦奥胡斯，《LIPIcs》第334卷，第150:1–150:17页。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik，2025年。[8] Krishnendu Chatterjee、Laurent Doyen、Jean-Fran¸cois Raskin 和 Ocan Sankur。具有奇偶目标的多环境马尔可夫决策过程的值问题。CoRR，abs/2504.15960，2025年。[9] Hugo Gimbert 和 Youssouf Oualhadj。有限词上的概率自动机：可判定与不可判定问题。载于 Samson Abramsky、Cyril Gavoille、Claude Kirchner、Friedhelm Meyer auf der Heide 和 Paul G. Spirakis（编），《自动机、语言与编程》，第37届国际研讨会，ICALP 2010，法国波尔多，2010年7月6日至10日。论文集，第二部分，《计算机科学讲义》第6199卷，第527–538页。施普林格，2010年。[10] Oded Goldreich。关于承诺问题：综述。载于 Oded Goldreich、Arnold L. Rosenberg 和 Alan L. Selman（编），《理论计算机科学：纪念 Shimon Even 文集》，《计算机科学讲义》第3895卷，第254–290页。施普林格，2006年。[11] Wassily Hoeffding。有界随机变量之和的概率不等式。《美国统计学会杂志》58(301):13–30，1963年。

[12] Omid Madani、Steve Hanks 和 Anne Condon。概率规划及相关随机优化问题的不可判定性。人工智能，147(1-2):5– 34，2003。[13] Omid Madani、Steve Hanks 和 Anne Condon。概率规划及相关随机优化问题的不可判定性。人工智能，147(1-2):5–34，2003。[14] Christos H. Papadimitriou 和 John N. Tsitsiklis。马尔可夫决策过程的复杂度。运筹学数学，12(3):441–450，1987。[15] Jean-Fran¸cois Raskin 和 Ocan Sankur。多环境马尔可夫决策过程。载于 Venkatesh Raman 和 S. P. Suresh 编，第*34*届软件技术与理论计算机科学基础国际会议（*FSTTCS 2014*），印度新德里，*2014*年*12*月*15–17*日，LIPIcs 第29卷，第531–543页。Schloss Dagstuhl - Leibniz 信息学中心，2014年。[16] Jean-Fran¸cois Raskin 和 Ocan Sankur。多环境马尔可夫决策过程。*CoRR*，abs/1405.4733，2014年。[17] Claude E. Shannon。通信的数学理论。*ACM SIGMOBILE Mob.Comput. Commun. Rev.*，5(1):3–55，2001年。[18] Maurice Sion。关于一般极小极大定理。1958年。[19] Marnix Suilen、Marck van der Vegt 和 Sebastian Junges。多环境马尔可夫决策过程中几乎必然满足 Rabin 目标的 PSPACE算法。载于 Rupak Majumdar 和 Alexandra Silva 编，第*35*届并发理论国际会议（*CONCUR 2024*），加拿大卡尔加里，*2024*年*9*月*9–13*日，LIPIcs 第311卷，第40:1–40:17页。Schloss Dagstuhl - Leibniz 信息学中心，2024年。[20] John Von Neumann 和 Oskar Morgenstern。《博弈论与经济行为》，第2次修订版。1947年。

A 节补充内容3

A.1 命题1

证明。如果 $\mathsf{val}_q^{\mathsf{uni}}(\Gamma, W) = 1$，那么对所有 $\varepsilon > 0$，存在 $\sigma \in \mathsf{Strat}(Q, A)$，使得对所有 $e \in E$，$\mathbb{P}_q^\sigma[\Gamma[e], W] \geq 1 - \varepsilon$。我们得到：

$$\mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W) \geq \sum_{e \in E} b(e) \cdot \mathbb{P}_q^\sigma[\Gamma[e], W] \geq \sum_{e \in E} b(e) \cdot (1 - \varepsilon) = 1 - \varepsilon$$

因此，$\mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W) = 1$。现在假设 $\mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W) = 1$。令 $\varepsilon > 0$。令 $d := \min_{e \in E} b(e) > 0$。存在某个 $\sigma \in \mathsf{Strat}(Q, A)$，使得 $\sum_{e \in E} b(e) \cdot \mathbb{P}_q^\sigma[\Gamma[e], W] \geq 1 - d \cdot \varepsilon$。反设存在某个 $e_m \in E$ 使得 $\mathbb{P}_q^\sigma[\Gamma[e_m], W] < 1 - \varepsilon$。在这种情况下：

$$\sum_{e \in E} b(e) \cdot \mathbb{P}_q^\sigma[\Gamma[e], W] < \sum_{e \in E \setminus \{e_m\}} b(e) + b(e_m) \cdot (1 - \varepsilon) = 1 - b(e_m) \cdot \varepsilon \leq 1 - d \cdot \varepsilon$$

因此产生矛盾。事实上，对于所有 $e \in E$，都有 $\mathbb{P}_q^\sigma[\Gamma[e_m], W] \geq 1 - \varepsilon$。对于所有 $\varepsilon > 0$，这样的策略 $\sigma \in \mathsf{Strat}(Q, A)$ 的存在均有保证，因此我们有 $\mathsf{val}_q^{\mathsf{uni}}(\Gamma, W) = 1$。

现在考虑策略 $\sigma \in \mathsf{Strat}(Q, A)$。我们有：

$$\begin{aligned}\mathsf{val}_q^{\mathsf{uni}}(\Gamma, \sigma, W) = 1 &\Leftrightarrow \forall e \in E,\ \mathbb{P}_q^\sigma[\Gamma[e], W] = 1 \Leftrightarrow \forall e \in E,\ b(e) \cdot \mathbb{P}_q^\sigma[\Gamma[e], W] = b(e) \\ &\Leftrightarrow \sum_{e \in E} b(e) \cdot \mathbb{P}_q^\sigma[\Gamma[e], W] = 1 \Leftrightarrow \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \sigma, W) = 1\end{aligned}$$

□

A.2 引理的证明4

证明。首先，请注意：

$$\sum_{\substack{e \in E \\ b(e) \geq b'(e)}} (b(e) - b'(e)) = \sum_{e \in E} (b(e) - b'(e)) - \sum_{\substack{e \in E \\ b'(e) \geq b(e)}} (b(e) - b'(e)) = \sum_{\substack{e \in E \\ b'(e) \geq b(e)}} (b'(e) - b(e))$$

因此：

$$\begin{aligned}\mathsf{Diff}(b, b') = \frac{1}{2} \sum_{e \in E} |b(e) - b'(e)| &= \frac{1}{2} \left(\sum_{\substack{e \in E \\ b(e) \geq b'(e)}} (b(e) - b'(e)) + \sum_{\substack{e \in E \\ b'(e) \geq b(e)}} (b'(e) - b(e)) \right) \\ &= \sum_{\substack{e \in E \\ b(e) \geq b'(e)}} (b(e) - b'(e)) = \sum_{\substack{e \in E \\ b'(e) \geq b(e)}} (b'(e) - b(e))\end{aligned}$$

现在，考虑策略 $\sigma \in \mathsf{Strat}(Q, A)$。我们有：

$$\begin{aligned}\sum_{e \in E} b(e) \cdot \mathbb{P}_q^\sigma[\Gamma[e], W] &= \sum_{e \in E} b'(e) \cdot \mathbb{P}_q^\sigma[\Gamma[e], W] + \sum_{e \in E} (b(e) - b'(e)) \cdot \mathbb{P}_q^\sigma[\Gamma[e], W] \\ &\leq \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b', W) + \mathsf{Diff}(b, b')\end{aligned}$$

由于这对所有策略 $\sigma \in \mathsf{Strat}(Q, A)$ 都成立，我们得到：$\mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W) \leq \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b', W) + \mathsf{Diff}(b, b')$。由对称性可得：$\mathsf{val}_q^{\mathsf{pr}}(\Gamma, b', W) \leq \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W) + \mathsf{Diff}(b, b')$。

令 $b \in \mathcal{D}(E)$，使得 $|\mathsf{Supp}(b)| \geq 2$。令 $e \in E$，使得 $\min_{e' \in E} b(e') = b(e)$。存在某个 $e' \in \mathsf{Supp}(b) \setminus \{e\}$。我们如下定义 $\mathsf{Truncate}(b) \in \mathcal{D}(E)$：$\mathsf{Truncate}(b)(e) := 0$；$\mathsf{Truncate}(b)(e') := b(e) + b(e')$，且对所有 $e'' \in E \setminus \{e, e'\}$，$\mathsf{Truncate}(b)(e'') := b(e'')$。于是：

$$\begin{aligned}\mathsf{Diff}(b, \mathsf{Truncate}(b)) &= \frac{1}{2} \cdot \sum_{e'' \in E} |b(e'') - \mathsf{Truncate}(b)(e'')| \\ &= \frac{1}{2} \cdot \big(|b(e) - \mathsf{Truncate}(b)(e)| + |b(e') - \mathsf{Truncate}(b)(e')|\big) \\ &= b(e)\end{aligned}$$

□

A.3 定理的证明 5

A.3.1 $m(\Gamma, \varepsilon)$ 的定义

下面介绍一些记号，用于表示量 $m(\Gamma, \varepsilon)$。

定义 16. 考虑一个多环境马尔可夫决策过程 $\Gamma = (Q, A, E, (\delta_e)_{e \in E})$。对于 $m \in \{\text{最大}, \text{最小}\}$，我们定义：

$$\mathsf{ratio}_m(\Gamma) := m\left\{ \frac{\delta_e(q,a)(q')}{\delta_{e'}(q,a)(q')} \mid e, e' \in E,\ (q,a,q') \in Q \times A \times Q,\ \delta_{e'}(q,a)(q') > 0 \right\}$$

如果上述集合为空，则令 $\text{比率}_{\min}(\Gamma) = 1 = \text{比率}_{\max}(\Gamma)$。如果情况并非如此，则有 $\text{比率}_{\min}(\Gamma) < 1 < \text{比率}_{\max}(\Gamma)$。我们还定义：

$$\mathsf{ratio}^{>1}_{\min}(\Gamma) := \min\left\{ \frac{\delta_e(q,a)(q')}{\delta_{e'}(q,a)(q')} \mid e, e' \in E,\ (q,a,q') \in Q \times A \times Q,\ \delta_e(q,a)(q') > \delta_{e'}(q,a)(q') > 0 \right\}$$

如果上述集合为空，则令 $\text{比率}^{>1}_{\min}(\Gamma) := 1$。否则，有 $\text{比率}^{>1}_{\min}(\Gamma) > 1$。

我们还令

$$p_{\mathsf{min}}(\Gamma) := \min\{\delta_e(q,a)(q') \mid (q,a,q') \in Q \times A \times Q,\ \delta_e(t,a)(t') > 0\} > 0$$

我们令：

$$\iota(\Gamma) := \min\left(\left(\sqrt{\mathsf{ratio}^{>1}_{\min}(\Gamma)} - 1 \right)^2, 1 \right) \ \textit{and}\ \eta(\Gamma) := \log(1 - p_{\mathsf{min}(\Gamma)} \cdot \iota(\Gamma))$$

注意，如果 $\text{比值}^{>1}_{\min}(\Gamma) = 1$，则 $\iota(\Gamma) = 0 = \eta(\Gamma)$；否则 $\iota(\Gamma) > 0$ 以及 $\eta(\Gamma) < 0$。最后，我们令：

- $n_1(\Gamma, \varepsilon) := \left\lceil \frac{\text{对数函数}(\varepsilon/(2 \cdot |E|))}{\text{对数函数}(1 - p_{\text{最小值}(\Gamma)})} \right\rceil$
- $n_2(\Gamma, \varepsilon) := \lceil 2|\text{对数函数}(\varepsilon)| + n_1(\Gamma, \varepsilon) \cdot \text{对数函数}(\text{比值}_{\text{最大值}(\Gamma)}) \rceil$
- 如果 $\eta(\Gamma) = 0$，则我们令 $n_3(\Gamma, \varepsilon) := 0$，否则：

$$n_3(\Gamma, \varepsilon) := \left\lceil \frac{2 \cdot |\log(\varepsilon/(2 \cdot |E|))| \cdot (\log(\frac{\mathsf{ratio}_{\max}(\Gamma)}{\mathsf{ratio}_{\min}(\Gamma)}))^2}{\eta(\Gamma)^2} + \frac{2 \cdot n_2(\Gamma, \varepsilon)}{|\eta(\Gamma)|} \right\rceil$$

- $m(\Gamma, \varepsilon) := |E| \cdot (n_1(\Gamma, \varepsilon) + n_3(\Gamma, \varepsilon))$

下面证明 $m(\Gamma, \varepsilon)$ 满足定理中陈述的等式。

引理*17.* 考虑一个多环境马尔可夫决策过程 $\Gamma = (Q, A, E, (\delta_e)_{e\in E})$。令 $k := |E|$，l 表示书写 Γ 中任意一个概率所需的最大比特数，并令 $x := \lceil \text{对数函数}(\frac{1}{\varepsilon}) \rceil$。那么，上文定义的 $m(\Gamma, \varepsilon)$ 满足 $m(\Gamma, \varepsilon) = O(poly(k, 2^l, x))$。

证明。首先，我们有 $p_{\min}(\Gamma) \geq \frac{1}{2^l}$。因此，$|\log(1 - p_{\min}(\Gamma))| \geq p_{\min}(\Gamma) \geq \frac{1}{2^l}$。于是，$n_1(\Gamma, \varepsilon) = O(2^l \cdot (x + k))$。此外，$\log(\text{ratio}_{\max}(\Gamma)) \leq l$，因此 $n_2(\Gamma, \varepsilon) = O(2^l \cdot l \cdot (x + k))$。

此外，假设 $\eta(\Gamma) > 0$，从而 $\text{ratio}^{>1}_{\min}(\Gamma) > 1$。若 $\text{ratio}^{>1}_{\min}(\Gamma) \geq 4$，则 $\iota(\Gamma) = 1$；否则，我们有 $\sqrt{\text{ratio}^{>1}_{\min}(\Gamma)} - 1 = \text{ratio}^{>1}\ \min(\quad \Gamma\ -1) \text{ min}(\ ) + 1 \geq \frac{1}{3}(\text{ratio}^{>1}_{\min}(\Gamma) - 1)$，因此 $\left(\sqrt{\text{ratio}^{>1}_{\min}(\Gamma)} - 1\right)^2 \geq \frac{1}{9}(\text{ratio}^{>1}_{\min}(\Gamma) - 1)^2$。由于 $\text{ratio}^{>1}_{\min}(\Gamma)$ 等于两个概率的比值，而这两个概率可用 l 比特编码，因此 $\text{ratio}^{>1}_{\min}(\Gamma) - 1 \geq \frac{1}{2^{2l}}$。所以，$|\eta(\Gamma)| = |\log(1 - p_{\min(\Gamma)\cdot\iota(\Gamma)})| \geq p_{\min(\Gamma)\cdot\iota(\Gamma)} \geq \frac{1}{9} \cdot \frac{1}{2^{5l}}$。于是，我们有 $n_3(\Gamma, \varepsilon) = O(2^{10l} \cdot (4l)^2 \cdot (x+k) + 2^{6l} \cdot l \cdot (x+k))$。引理得证。□

A.3.2 定理的证明

我们在马尔可夫决策过程中定义运行赋值的概念（即将运行映射为实数值的函数）。具体而言，我们分两步进行。首先，我们证明，若某个历史估值属于特定类型，则已访问历史值之和与其期望值相差甚远的概率很小；为此，我们将霍夫丁不等式调整到我们的设定中。其次，我们证明，定理5 的结果可以理解为关于这类历史估值的陈述。我们先正式引入历史估值的概念。

定义18。考虑一个马尔可夫决策过程 $\mathsf{G} = (Q, A, \delta)$。令 $a, b \in \mathbb{R}$ 满足 $a < 0 < b$。一个 $[a, b]$-赋值是一个函数 $v\colon Q \cdot (A \cdot Q)^* \to [a, b]$。给定 $a \leq \eta \leq 0$，若一个 $[a, b]$-赋值 $v\colon Q \cdot (A \cdot Q)^* \to [a, b]$ 满足以下条件，则称其为 η-有界：对所有 $\rho \in q \cdot (A \times Q)^*$ 和 $\alpha \in A$：

- 要么对所有 $q' \in Q$，都有 $v(\rho \cdot (\alpha, q')) = 0$；
- 要么 $\sum_{q'\in Q} \delta(q, \alpha)(q') \cdot v(\rho \cdot (\alpha, q')) \leq \eta$；我们用 $\mathsf{NZ}_v \subseteq Q \times A$（“非零”）表示此类状态-动作对 (q, α) 的集合。

我们还引入其他几个有用的记号。首先，定义：

$$\mathsf{NZ}^+_v := \{\pi = (q_0, a_0) \cdots (q_n, a_n) \in (Q \cdot A)^+ \mid (q_n, a_n) \in \mathsf{NZ}_v \ \textit{and}\ \forall i \leq n-1,\ (q_i, a_i) \notin \mathsf{NZ}_v\}$$

此外，对所有 $\rho = (q_0, a_0) \cdots \in (Q \cdot A)^\omega$，定义：

$$\mathsf{IndNZ}_v(\rho) := \{i \in \mathbb{N} \mid (q_i, a_i) \in \mathsf{NZ}_v\}$$

对所有 $n \in \mathbb{N}$，令 $\mathsf{IndNZ}^n_v(\rho) \subseteq \mathsf{IndNZ}_v(\rho)$ 表示 $\min(n, |\mathsf{IndNZ}_v(\rho)|)$ 个最小索引组成的集合，这些索引属于 $\mathsf{IndNZ}_v(\rho)$。我们还令 $\mathsf{AZ}^\omega_v := \{\rho \mid \mathsf{IndNZ}_v(\rho) = \emptyset\}$（“恒为零”）。

最后，对于所有 $n \in \mathbb{N}$，我们分别定义函数 $\mathsf{s}^n_v\colon (Q \cdot A)^\omega \to \mathbb{R}$ 和 $\mathsf{c}^n_v\colon (Q \cdot A)^\omega \to \mathbb{N}$，前者对 NZ_v-动作之后出现的 n 个首个值求和，后者统计访问过的 NZ_v-动作数量。具体而言，对于所有 $\rho \in (Q \cdot A)^\omega$，定义如下：

$$\mathsf{s}^n_v(\rho) := \sum_{i\in \mathsf{IndNZ}^n_v(\rho)} v(\rho_{\leq i}) \in [n \cdot a, n \cdot b]$$
$$\mathsf{c}^n_v(\rho) := |\mathsf{IndNZ}^n_v(\rho)| \leq n$$

A.4 适用于霍夫丁引理

既然我们已经引入了运行赋值，首要目标就是证明下面的引理。

引理 19. 考虑一个马尔可夫决策过程 $\mathsf{G} = (Q, A, \delta)$ 以及一个状态 $q \in Q$。令 $a < 0 < b \in \mathbb{R}$，并令 $a \le \eta \le 0$。考虑一个 $[a,b]$-赋值 $v\colon Q \cdot (A \cdot Q)^* \to [a,b]$，它是 η-有界的。对于所有 $n \in \mathbb{N}$、所有 $t > 0$ 以及所有策略 $\sigma \in \mathsf{Strat}(Q, A)$，均有：

$$\mathbb{P}_q^\sigma[\mathsf{G}, \{\mathsf{s}_v^n \ge n \cdot \eta + t\} \cap \{\mathsf{c}_v^n = n\}] \le \exp\left(-\frac{2t^2}{n(b-a)^2}\right)$$

由于 η 是在执行任意 NZ_v-动作后赋值 v 的期望值上界，因此事件 $\{\mathsf{s}_v^n \ge n \cdot \eta + t\}$ 对应于这样一些无限路径：由 s_v^n 计算得出的、在 NZ_v-动作后出现的前 n 个 v-值之和与期望值至少相差 t。

该引理可视为将霍夫丁不等式推广到我们的 η-有界运行赋值情形。不过，原则上不能直接应用霍夫丁不等式，因为描述 s_v^n 随机演化的随机变量并不独立。尽管如此，该引理的证明与推导霍夫丁不等式所采用的步骤非常相似；特别是，它依赖于下文回顾的霍夫丁引理。

引理 20 (霍夫丁引理)。 考虑一个有限集 Q、某个 $a \le b \in \mathbb{R}$、一个赋值 $v\colon Q \to [a,b]$ 以及一个分布 $d \in \mathcal{D}(Q)$。令 $\mu := \sum_{q \in Q} d(q) \cdot v(q)$。对于所有 $s > 0$，有：

$$\sum_{q \in Q} d(q) \cdot \exp(s(v(q) - \mu)) \le \exp\left(\frac{s^2(b-a)^2}{8}\right)$$

我们现在通过归纳法建立下面的引理，并在每个归纳步骤中调用上述引理。

引理 21. 考虑一个马尔可夫决策过程 $\mathsf{G} = (Q, A, \delta)$ 和一个状态 $q \in Q$。设 $a < 0 < b \in \mathbb{R}$ 并设 $a \le \eta \le 0$。考虑一个 $[a,b]$-赋值 $v\colon Q \cdot (A \times Q)^* \to [a,b]$，它是 η-有界的。对于所有 $n \in \mathbb{N}$和 $M \ge 0$，我们定义函数 $f^{n,M}_v\colon (Q \cdot A)^\omega \to \mathbb{R}$ 如下，对于所有 $\rho \in (Q \cdot A)^\omega$：

$$f_v^{n,M}(\rho) := \begin{cases} \mathsf{s}_v^n(\rho) & \text{if } \mathsf{c}_v^n(\rho) = n \\ -M & \text{otherwise} \end{cases}$$

设 $s > 0$。对于所有 $n \in \mathbb{N}$，对于所有 $M \ge 0$，对于所有策略 $\sigma \in \mathsf{Strat}(Q, A)$，我们有：

$$\mathbb{E}_q^\sigma[\mathsf{G}, \exp(s(f_v^{n,M} - n \cdot \eta))] \le \exp\left(n\frac{s^2(b-a)^2}{8}\right) + n \cdot \exp\left(s\left(-M + n \cdot \max(|a|,|b|)\right) + n\frac{s^2(b-a)^2}{8}\right)$$

证明。我们通过对 n 使用归纳法来证明该结果。情形 $n = 0$ 是直接的，因为对于所有 $M \ge 0$，$f^{0,M}_v = 0$。现在假设该性质对某个 $n \in \mathbb{N}$ 成立。设 $M \ge 0$。注意，对于所有 $\pi \in \mathsf{NZ}_v^+$，对于所有 $q' \in Q$，以及对于所有 $\rho \in q' \cdot A \cdot (Q \cdot A)^\omega$，我们有：

$$f_v^{n+1,M}(\pi \cdot \rho) = v(\pi \cdot q') + f_v^{n, M + v(\pi \cdot q')}(\rho)$$

对于所有 $\rho = (q_0,\ a_0) \cdots (q_n,\ a_n) \in$ 自然数集$_v^+$，令 $\delta(\rho) := \delta(q_n,\ a_n)$。然后，根据霍夫丁引理，对于所有 $\rho \in$ 自然数集$_v^+$，由于 $\mu := \sum_{q' \in Q} \delta(\rho)(q') \cdot v(\rho \cdot q') \le \eta \le 0$，我们有：

$$\sum_{q' \in Q} \delta(\rho)(q') \cdot \exp(s(v(\rho \cdot q')) - \eta) \le \sum_{q' \in Q} \delta(\rho)(q') \cdot \exp(s(v(\rho \cdot q')) - \mu) \le \exp\left(\frac{s^2(b-a)^2}{8}\right)$$

令 $c := \max(|a|, |b|)$。注意，由于 $a \le \eta \le b$，我们有 $-\eta \le c$。对于所有 $\rho \in \mathsf{NZ}_v^{\mathsf{run}}$，令 $\mathsf{last}(\rho) \in Q$ 表示 ρ 访问的最后一个状态；我们还令 $\sigma_\rho \in \mathsf{Strat}(Q, A)$ 定义如下：对于所有$\pi \in Q \cdot (A \cdot Q)^*$, $\sigma_\rho(\pi) := \sigma(\rho \cdot \pi)$。首先，我们有：

$$\begin{aligned}\mathbb{E}_q^\sigma[\mathsf{G}, \exp(s(f_v^{n+1,M} - (n+1) \cdot \eta))] = &\sum_{\rho \in \mathsf{NZ}_v^+} \mathbb{P}_q^\sigma[\mathsf{G}, \rho] \cdot \mathbb{E}_{\mathsf{last}(\rho)}^{\sigma_\rho}[\mathsf{G}, \exp(s(f_v^{n+1,M} - (n+1) \cdot \eta))] \\ &+ \mathbb{P}_q^\sigma[\mathsf{G}, \mathsf{AZ}_v^\omega] \cdot \exp(s(-M - (n+1) \cdot \eta))\end{aligned}$$

with

$$\mathbb{P}_q^{\sigma}[\mathsf{G}, \mathsf{AZ}_v^{\omega}] \cdot \exp(s(-M-(n+1)\cdot\eta)) \leq \exp\left(s(-M+(n+1)\cdot c)+(n+1)\cdot\frac{s^2(b-a)^2}{8}\right)$$

令 $\rho \in \mathsf{NZ}_v^{+}$。我们有：

$$\mathbb{E}^{\sigma_\rho}_{\mathsf{last}(\rho)}[\mathsf{G}, \exp(s(f_v^{n+1,M}-(n+1)\cdot\eta))] = \sum_{q'\in Q}\delta(\rho)(q')\cdot\mathbb{E}^{\sigma_\rho}_{q'}[\mathsf{G}, \exp(s(f_v^{n,M+v(\rho\cdot q')}-n\cdot\eta+v(\rho\cdot q')-\eta))]$$

现在，令 $q' \in Q$ 和 $x := \mathbb{E}^{\sigma_\rho}_{q'}[\mathsf{G}, \exp(s(f^{n,M+v(\rho\cdot q')}_v - n\cdot\eta + v(\rho\cdot q') - \eta))]$。我们有：

$$\begin{aligned}
x &= \exp(s(v(\rho\cdot q')-\eta))\cdot\mathbb{E}^{\sigma}_{q'}[\mathsf{G}, \exp(s(f^{n,M+v(\rho\cdot q')}_v - n\cdot\eta))] \\
&\leq \exp(s(v(\rho\cdot q')-\eta))\cdot\left[\exp\left(n\frac{s^2(b-a)^2}{8}\right)+n\exp\left(s\left(-M-v(\rho\cdot q')+n\cdot c+n\frac{s^2(b-a)^2}{8}\right)\right)\right] \\
&\leq \exp\left(\frac{s^2(b-a)^2}{8}\right)\cdot\left[\exp\left(n\frac{s^2(b-a)^2}{8}\right)+n\exp(s(-M-v(\rho\cdot q')+n\cdot c+n\frac{s^2(b-a)^2}{8}))\right] \\
&\leq \exp\left((n+1)\frac{s^2(b-a)^2}{8}\right)+n\exp\left(s(-M+(n+1)\cdot c)+(n+1)\frac{s^2(b-a)^2}{8}\right)
\end{aligned}$$

这对所有 $q' \in Q$均成立，因此，我们有：

$$\begin{aligned}
\mathbb{E}^{\sigma_\rho}_{\mathsf{last}(\rho)}[\mathsf{G}, \exp(s(f_v^{n+1,M}-(n+1)\cdot\eta))] &\leq \exp\left((n+1)\frac{s^2(b-a)^2}{8}\right) \\
&\quad + n\exp\left(s(-M+(n+1)\cdot c)+(n+1)\frac{s^2(b-a)^2}{8}\right)
\end{aligned}$$

这对所有 $\rho \in \mathsf{NZ}_v^{+}$均成立，因此：

$$\begin{aligned}
\mathbb{E}^{\sigma}_{q}[\mathsf{G}, \exp(s(f_v^{n+1,M}-(n+1)\cdot\eta))] &\leq \exp\left((n+1)\frac{s^2(b-a)^2}{8}\right) \\
&\quad + n\exp\left(s(-M+(n+1)\cdot c)+(n+1)\frac{s^2(b-a)^2}{8}\right) \\
&\quad + \exp\left(s(-M+(n+1)\cdot c)+(n+1)\frac{s^2(b-a)^2}{8}\right) \\
&\leq \exp\left((n+1)\frac{s^2(b-a)^2}{8}\right) \\
&\quad + (n+1)\exp\left(s(-M+(n+1)\cdot c)+(n+1)\frac{s^2(b-a)^2}{8}\right)
\end{aligned}$$

归纳法至此完成。 □

我们可以开始证明引理19。

证明。 令 $t > 0$。对于所有 $M > -n\cdot\eta - t$，我们有 $\{\mathsf{s}_v^n \geq n\cdot\eta + t\} \cap \{\mathsf{c}_v^n = n\} = \{f_v^{n,M} \geq n\cdot\eta + t\}$。此外，对于所有 $s > 0$，由马尔可夫不等式可得：

$$\begin{aligned}
\mathbb{P}_q^{\sigma}[\mathsf{G}, \{f_v^{n,M} \geq n\cdot\eta + t\}] &= \mathbb{P}_q^{\sigma}[\mathsf{G}, \{s(f_v^{n,M} - n\cdot\eta) \geq s\cdot t\}] \\
&= \mathbb{P}_q^{\sigma}[\mathsf{G}, \{\exp(s(f_v^{n,M} - n\cdot\eta)) \geq \exp(s\cdot t)\}] \\
&\leq \exp(-s\cdot t)\cdot\mathbb{E}_q^{\sigma}[\mathsf{G}, \exp(s(f_v^{n,M} - n\cdot\eta))]
\end{aligned}$$

由于 $n \cdot \exp(-s(M + n \cdot \max(|a|, |b|)) + n\frac{s^2(b-a)^2}{8}) \to_{M\to\infty} 0$，根据引理 21可得：

$$\mathbb{P}_q^\sigma[\mathsf{G}\ \{\mathsf{s}_v^n \geq n \cdot \eta + t\} \cap \{\mathsf{c}_v^n \geq n\}] \leq \exp(-s \cdot t) \cdot \exp\left(n\frac{s^2(b-a)^2}{8}\right) = \exp\left(-st + n\frac{s^2(b-a)^2}{8}\right)$$

这对所有 $s > 0$ 都成立，特别是对 $s = \frac{4t}{n(b-a)^2}$，我们得到：

$$\mathbb{P}_q^\sigma[\mathsf{G}, \{\mathsf{s}_v^n \geq n \cdot \eta + t\} \cap \{\mathsf{c}_v^n \geq n\}] \leq \exp\left(-\frac{4t^2}{n(b-a)^2} + \frac{16t^2}{8n(b-a)^2}\right) = \exp\left(-\frac{2t^2}{n(b-a)^2}\right)$$

□

A.5 基于信念更新的赋值

作为一项初步结果，我们来证明，在某个环境中进行对局时，对该环境的信念不会降为 0。

引理 *22*。设 $\Gamma = (Q, A, E, (\delta_e)_{e\in E})$ 是一个多环境马尔可夫决策过程。考虑一个状态 $q \in Q$ 和一个策略$\sigma \in Strat(Q, A)$。对于所有环境 $e \in E$ 和信念 $b \in \mathcal{D}(E)$，若 $b(e) > 0$，则对于所有$\rho \in Q \cdot (A \cdot Q)^*$，若 $\mathbb{P}_q^\sigma(\Gamma[e], \rho) > 0$，有：$lk_b(\rho)(e) > 0$。

证明。我们通过对 $\rho \in Q \cdot (A \cdot Q)^*$ 作归纳法来证明该结果。当 $\rho \in Q$ 时，这显然成立。现在，假设对于某个 $\rho \in Q \cdot (A \cdot Q)^*$，该结论成立，并令 $(a, t) \in A_{\mathsf{last}(\rho)} \times Q$。若$\mathbb{P}_q^\sigma(\Gamma[e], \rho \cdot (a, t)) > 0$，则 $\mathbb{P}_q^\sigma(\Gamma[e], \rho) \cdot \sigma(\rho)(a) \cdot \delta_e(\mathsf{last}(\rho)\ a)(t) > 0$，因此 $\mathsf{lk}_b(\rho)(e) > 0$。于是：$\mathsf{lk}_b(\rho \cdot (a, t))(e) = \frac{1}{\mathbb{P}[\Gamma_b, \rho]} \mathsf{lk}(\ )\ \mathsf{last}(\rho)\ a](t) \cdot \mathsf{lk}_b(\rho)(e) \cdot \delta_e(\mathsf{last}(\rho)\ a)(t) > 0$。 □

我们的目标现在是应用引理 19 来证明定理 5。为此，我们描述马尔可夫决策过程中有界的赋值，这些赋值可以通过多环境马尔可夫决策过程中的信念更新获得。在定义这种赋值之前，我们先引入揭示状态-动作对这一概念。

定义 23. 考虑一个多环境马尔可夫决策过程 $\Gamma = (Q, A, E, (\delta_e)_{e\in E})$。设 $e \neq e' \in E$。如果存在某个状态 $q' \in Q$，使得$\delta_{e'}(q, a)(q') = 0$且 $\delta_e(q, a)(q') > 0$，则称状态-动作对 $(q, a) \in Q \times A$ 是 e'/e-揭示的。

对于所有 $\rho \in Q \cdot (A \cdot Q)^*$，我们用 $\mathsf{nb}_{\mathsf{rev}}(e', e, \rho) \in \mathbb{N}$ 表示 ρ 中 e'/e-揭示性状态-动作对出现的次数。

现在我们可以定义所考虑的历史估值。

定义 24. 考虑一个多环境马尔可夫决策过程 $\Gamma = (Q, A, E, (\delta_e)_{e\in E})$。设 $q \in Q$和 $e \neq e' \in E$。我们将历史估值 $v[e', e]$： $Q \cdot (A \cdot Q)^* \to \mathbb{R}$ 定义如下： $v[e', e](q) := 0$并且对于所有$\rho = q_0 \cdot (a_1, q_1) \cdots (a_n, q_n) \in Q \cdot (A \cdot Q)^*$:

$$v[e', e](\rho) := \begin{cases} 0 & \text{if } a_n \text{ is } e'/e\text{-revealing} \\ 0 & \text{if } \delta_e(q_{n-1}, a_n)(q_n) = 0 \\ \log\left(\frac{\delta_{e'}(q_{n-1}, a_n)(q_n)}{\delta_e(q_{n-1}, a_n)(q_n)}\right) & \text{otherwise} \end{cases}$$

根据定义，在马尔可夫决策过程 $\Gamma[e]$ 中，路径 ρ 上访问到的 $v[e', e]$ 值之和的指数，与环境 e' 和 e 在 ρ 处的信念之比密切相关；这里假设该路径在任意策略下都以正概率出现。下文将对此作出形式化表述。

引理 25. 设 $\Gamma = (Q, A, E, (\delta_e)_{e\in E})$ 是一个多环境马尔可夫决策过程。考虑两个环境 $e \neq e' \in E$，以及一个满足 $b(e) > 0$ 的先验信念 $b \in \mathcal{D}(E)$ 。考虑任意策略 $\sigma \in \mathsf{Strat}(Q, A)$ 和某个$\rho \in Q \cdot (A \cdot Q)^*$，使得 $\mathbb{P}_q^\sigma[\Gamma[e], \rho] > 0$。那么：

$$\frac{b(e')}{b(e)} \cdot 2^{(\sum_{\pi \sqsubseteq \rho} v[e', e](\pi))} \cdot \mathsf{ratio}_{\max}(\Gamma)^{\mathsf{nb}_{\mathsf{rev}}(e', e, \rho)} \geq \frac{\mathsf{lk}_b(\rho)(e')}{\mathsf{lk}_b(\rho)(e)}$$

其中，我们将 $\pi \sqsubseteq \rho$ 写作 $\pi \in Q \cdot (A \cdot Q)^*$，表示 ρ 的前缀。

证明。首先，对于所有 $\rho \in Q \cdot (A \cdot Q)^*$ 满足 $\mathbb{P}_q^\sigma[\Gamma[e], \rho] > 0$的情况，我们有 $\mathsf{lk}_b(\rho)(e) > 0$，根据引理 22。

接着，我们对满足 $\mathbb{P}_q^\sigma[\Gamma[e], \rho] > 0$ 的 $\rho \in Q \cdot (A \cdot Q)^*$ 使用归纳法证明该结果。对于 $\rho = q$，结论显然成立，因为 $v[e', e](q) = 0$、$\mathsf{nb}_{\mathsf{rev}}(e', e, q) = 0$，且对于所有 $f \in \{e, e'\}$，$\mathsf{lk}_b(q)(f) = b(f)$。

现在假设该性质对某个 $\rho \in Q \cdot (A \cdot Q)^*$ 成立，并令 $(a, q) \in A \times Q$ 满足 $\rho \cdot (a, q) \in Q \cdot (A \cdot Q)^*$ 且 $\mathbb{P}_q^\sigma[\Gamma[e], \rho \cdot (a, q)] > 0$，因此 $\delta_e(\mathsf{last}(\rho))(a)(q) > 0$。我们有：

$$\frac{\mathsf{lk}_b(\rho \cdot (a,q))(e')}{\mathsf{lk}_b(\rho \cdot (a,q))(e)} = \frac{\frac{1}{\lambda[\mathsf{lk}_b(\rho), \mathsf{last}(\rho), a, q]} \cdot \mathsf{lk}_b(\rho)(e') \cdot \delta_{e'}(\mathsf{last}(\rho), a)(q)}{\frac{1}{\lambda[\mathsf{lk}_b(\rho), \mathsf{last}(\rho), a, q]} \cdot \mathsf{lk}_b(\rho)(e) \cdot \delta_e(\mathsf{last}(\rho), a)(q)} = \frac{\mathsf{lk}_b(\rho)(e')}{\mathsf{lk}_b(\rho)(e)} \cdot \frac{\delta_{e'}(\mathsf{last}(\rho), a)(q)}{\delta_e(\mathsf{last}(\rho), a)(q)}$$

接着，分两种情况：

- 如果 $(\mathsf{last}(\rho)\ a)$ 是 e'/e-揭示的，那么 $\mathsf{nbrev}(e', e, \rho \cdot (a, q)) = \mathsf{nbrev}(e', e, \rho) + 1$ 且 $v[e', e](\rho \cdot (a, q)) = 0$。此外，$\delta_{e'}(\mathsf{last}(\rho)\ a)(q) \le \delta_e(\mathsf{last}(\rho)\ a)(q) \le \mathsf{ratio}_{\max}(\Gamma)$。因此：

$$\begin{aligned}\frac{\mathsf{lk}_b(\rho \cdot (a,q))(e')}{\mathsf{lk}_b(\rho \cdot (a,q))(e)} &\le \frac{\mathsf{lk}_b(\rho)(e')}{\mathsf{lk}_b(\rho)(e)} \cdot \mathsf{ratio}_{\max}(\Gamma) \le \frac{b(e')}{b(e)} \cdot 2^{\left(\sum_{\pi \sqsubseteq \rho} v[e', e](\pi)\right)} \cdot \mathsf{ratio}_{\max}(\Gamma)^{\mathsf{nb}_{\mathsf{rev}}(e', e, \rho) + 1} \\ &= \frac{b(e')}{b(e)} \cdot 2^{\left(\sum_{\pi \sqsubseteq \rho \cdot (a,q)} v[e', e](\pi)\right)} \cdot \mathsf{ratio}_{\max}(\Gamma)^{\mathsf{nb}_{\mathsf{rev}}(e', e, \rho \cdot (a,q))}\end{aligned}$$

- 如果 a 不是 e'/e-揭示的，那么 $\mathsf{nbrev}(e', e, \rho \cdot (a, q)) = \mathsf{nbrev}(e', e, \rho)$，并且 $v[e', e](\rho \cdot (a, q)) = \log(\delta_{e'}(\mathsf{last}(\rho)\ a)(q)\ \delta_e(\mathsf{last}(\rho)\ a)(q))$，因此 $2^{v[e', e](\rho \cdot (a,q))} = \delta_{e'}(\mathsf{last}(\rho)\ a)(q)\ \delta_e(\mathsf{last}(\rho)\ a)(q)$。因此：

$$\frac{\mathsf{lk}_b(\rho \cdot (a,q))(e')}{\mathsf{lk}_b(\rho \cdot (a,q))(e)} = \frac{\mathsf{lk}_b(\rho)(e')}{\mathsf{lk}_b(\rho)(e)} \cdot 2^{v[e', e](\rho \cdot (a,q))} \le \frac{b(e')}{b(e)} \cdot 2^{\left(\sum_{\pi \sqsubseteq \rho \cdot (a,q)} v[e', e](\pi)\right)} \cdot \mathsf{ratio}_{\max}(\Gamma)^{\mathsf{nb}_{\mathsf{rev}}(e', e, \rho \cdot (a,q))}$$

归纳法至此完成。 □

此外，该赋值 $v[e', e]$ 是 $\eta(\Gamma)$-有界的。

引理 26。 考虑一个 *MEMDP* $\Gamma = (Q, A, E, (\delta_e)_{e \in E})$，以及 $e \neq e' \in E$。在 *MDP* $\Gamma[e]_v[e', e]$：$Q \cdot (A \cdot Q)^* \to \mathbb{R}$ 中，这是一个 $[\log(\mathsf{ratio}_{\min}(\Gamma))、\log(\mathsf{ratio}_{\max}(\Gamma))]$-赋值，且它是 $\eta(\Gamma)$-有界的。

该引理的证明依赖于下文回顾的算术平均数与几何平均数之差的下界。

引理 27 ([1]中的定理 1。设 $1 \le k \in \mathbb{N}$。对于所有 $1 \le i \le k$，考虑某些 $0 \le x_i \in \mathbb{R}$ 和 $0 < \alpha_i \in \mathbb{R}$，并假设 $\sum_{i=1}^k \alpha_i = 1$。那么，我们有：

$$\prod_{i=1}^k x_i^{\alpha_i} \le \sum_{i=1}^k \alpha_i \cdot x_i - \sum_{i=1}^k \alpha_i \left(\sqrt{x_i} - \sum_{i=1}^k \alpha_k \sqrt{x_k}\right)^2$$

现在我们继续证明引理 26。

证明。对于所有 $(q, a, q') \in Q \times A \times Q$，如果 $\delta_e(q, a)(q') > 0$，则有：

$$\mathsf{ratio}_{\min}(\Gamma) \le \frac{\delta_{e'}(q, a)(q')}{\delta_e(q, a)(q')} \le \mathsf{ratio}_{\max}(\Gamma)$$

由此可得，$v[e', e]$ 是一个 $[\log(\mathsf{ratio}_{\min}(\Gamma)), \log(\mathsf{ratio}_{\max}(\Gamma))]$-赋值。

现在考虑一些 $\rho \in Q \cdot (A \cdot Q)^*$ 和 $a \in A$，使得 $(\mathsf{last}(\rho)\ a) \in \mathsf{NZ}^{v[e', e]}$。根据 $v[e', e]$的定义，二元组 $(\mathsf{last}(\rho)\ a)$ 不是 e'/e-揭示的，并且我们有 $\delta_e(q, a) \neq \delta_{e'}(q, a)$。令 $S := \mathsf{Supp}(\delta_e(q, a)) \subseteq Q$。对于所有 $s \in S$，令 $p_s := \delta_e(\mathsf{last}(\rho)\ a)(s) > 0$，以及 $r_s := \delta_{e'}(\mathsf{last}(\rho)\ a)(s) > 0$

（因为 a 不是 e'/e-揭示的）。根据定义，我们有 $\sum_{q'\in Q}\delta_e(\mathsf{last}(\rho)\ a)(q')\cdot v(\rho\cdot(a,\ q'))=\sum_{s\in S}p_s\cdot v(\rho\cdot(a,\ s))$。此外：

$$\sum_{s\in S}p_s\cdot v(\rho\cdot(a,s))=\sum_{s\in S}p_s\cdot\log\left(\frac{r_s}{p_s}\right)=\log\left(\prod_{s\in S}\left(\frac{r_s}{p_s}\right)^{p_s}\right)$$

此外，根据引理27，我们有：

$$\prod_{s\in S}\left(\frac{r_s}{p_s}\right)^{p_s}\leq\sum_{s\in S}p_s\cdot\frac{r_s}{p_s}-\sum_{s\in S}p_s\left(\sqrt{\frac{r_s}{p_s}}-\sum_{s'\in S}p_{s'}\sqrt{\frac{r_{s'}}{p_{s'}}}\right)^2$$

如果 $\sum_{s\in S}r_s<1$，那么 $\sum_{s\in S}r_s\leq 1-p_{\min}(\Gamma)\leq 1-p_{\min}(\Gamma)\cdot\iota(\Gamma)$，因此 $\prod_{s\in S}\left(\frac{r_s}{p_s}\right)^{p_s}\leq 1-p_{\min}(\Gamma)\cdot\iota(\Gamma)$。

现在假设 $\sum_{s\in S}r_s=1$。那么，由于 $\delta_e(q,\ a)\neq\delta_{e'}(q,\ a)$，可知存在某个 $t\in S$ 使得 $\delta_{e'}(q,\ a)(t)=r_t>p_t=\delta_e(q,\ a)(t)$，因此根据比率$^{>1}_{\min}(\Gamma)$的定义，$\frac{r_t}{p_t}\geq$比率$^{>1}_{\min}(\Gamma)>1$。此外，根据平方根函数的凹性以及詹森不等式，我们有：

$$\sum_{s'\in S}p_{s'}\sqrt{\frac{r_{s'}}{p_{s'}}}\leq\sqrt{\sum_{s'\in S}p_{s'}\frac{r_{s'}}{p_{s'}}}=\sqrt{\sum_{s'\in S}r_{s'}}=1$$

因此：

$$\sum_{s\in S}p_s\left(\sqrt{\frac{r_s}{p_s}}-\sum_{s'\in S}p_{s'}\sqrt{\frac{r_{s'}}{p_{s'}}}\right)^2\geq p_t\left(\sqrt{\frac{r_t}{p_t}}-\sum_{s'\in S}p_{s'}\sqrt{\frac{r_{s'}}{p_{s'}}}\right)^2\geq p_{\min}(\Gamma)\left(\sqrt{\mathsf{ratio}^{>1}_{\min}(\Gamma)}-1\right)^2$$

因此：

$$\prod_{s\in S}\left(\frac{r_s}{p_s}\right)^{p_s}\leq 1-p_{\min}(\Gamma)\left(\sqrt{\mathsf{ratio}^{>1}_{\min}(\Gamma)}-1\right)^2\leq 1-p_{\min}(\Gamma)\cdot\iota(\Gamma)$$

总之，我们得到：

$$\sum_{q'\in Q}\delta_e(\mathsf{last}(\rho),a)(q')\cdot v(\rho\cdot(a,q'))=\sum_{s\in S}p_s\cdot v(\rho\cdot(a,s))=\log\left(\prod_{s\in S}\left(\frac{r_s}{p_s}\right)^{p_s}\right)\leq\log(1-p_{\min}(\Gamma)\cdot\iota(\Gamma))=\eta(\Gamma)$$

因此，$v[e',e]$ 是 $\eta(\Gamma)$-有界的。 □

借助上述引理，我们可以将引理19的结果应用于赋值$v[e',e]$；这样一来，我们将证明，如果访问了足够多的可区分且非 e'/e-揭示动作，那么两个环境中的信念之比以高概率足够小（这将在下一页处理）。剩下的就是处理 e'/e-揭示动作。下文的引理将对此进行证明。

引理 28。 考虑一个多环境马尔可夫决策过程 $\Gamma=(Q,A,E,(\delta_e)_{e\in E})$，以及 $e\neq e'\in E$。对于所有 $k\in\mathbb{N}$ 和先验信念 $b\in\mathcal{D}(E)$，令：

$$\mathsf{Reveal}^k(\Gamma,b,e',e):=\{\rho\in Q\cdot(A\cdot Q)^*\mid \mathsf{nb}_{\mathsf{rev}}(e',e,\rho)\geq k\ \textit{and}\ \mathsf{lk}_b(\rho)(e')\neq 0\}$$

那么，对于所有 $n\in\mathbb{N}$、所有 $q\in Q$ 以及所有策略 $\sigma\in\mathsf{Strat}(Q,A)$，有：

$$\mathbb{P}^\sigma_q(\Gamma[e],\mathsf{Reveal}^n(\Gamma,b,e',e))\leq(1-p_{\min}(\Gamma))^n$$

证明。 我们对 $n \in \mathbb{N}$ 使用归纳法来证明该结果。对于 $n = 0$，这显然成立。现在假设该性质对某个 $n \in \mathbb{N}$ 成立。令 $q \in Q$。我们令：

$$\mathsf{FirstRev} := \{\pi = (q_0,\ a_0) \cdots (q_n,\ a_n) \in (Q \cdot A)^* \mid (q_n,\ a_n) \text{ 是揭示的，而 } \forall i \leq n-1,\ (q_i,\ a_i) \text{ 不是揭示的}\}$$

对于所有 $\rho \in \mathsf{FirstRev}$，我们令 $\sigma_\rho \in \mathsf{Strat}(Q, A)$ 定义如下：对于所有 $\pi \in Q \cdot (A \cdot Q)^*$、$\sigma_\rho(\pi) := \sigma(\rho \cdot \pi)$（类似于引理 21 的证明）。对于所有 $\rho \in \mathsf{FirstRev}$，我们考虑某个 $q_\rho \in Q$ 使得 $\delta_{e'}(\rho)(q_\rho) = 0$ 且 $\delta_e(\rho)(q_\rho) > 0$。注意，$\sum_{q' \in Q \setminus \{q_\rho\}} \delta_e(\rho)(q) = 1 - \delta_e(\rho)(q_\rho) \leq 1 - p_{\min}(\Gamma)$。根据 lk_b 的定义，对于某次运行 $\rho \in Q \cdot (A \cdot Q)^*$ 的所有前缀 π，如果 $\mathsf{lk}_b(\pi)(e') = 0$，那么 $\mathsf{lk}_b(\rho)(e') = 0$。因此，令 $x := \mathbb{P}_q^\sigma(\Gamma[e], \mathsf{Reveal}^{n+1}(\Gamma, b, e', e))$，我们有：

$$\begin{aligned}
x &= \sum_{\rho \in \mathsf{FirstRev}} \mathbb{P}_q^\sigma(\Gamma[e], \rho) \cdot \sum_{q' \in Q} \delta_e(\rho)(q') \cdot \mathbb{P}_{q'}^{\sigma_\rho}(\Gamma[e], \mathsf{Reveal}^n(\Gamma, \mathsf{lk}_b(\rho \cdot q')), e', e) \\
&= \sum_{\rho \in \mathsf{FirstRev}} \mathbb{P}_q^\sigma(\Gamma[e], \rho) \cdot \sum_{q' \in Q \setminus \{q_\rho\}} \delta_e(\rho)(q') \cdot \mathbb{P}_{q'}^{\sigma_\rho}(\Gamma[e], \mathsf{Reveal}^n(\Gamma, \mathsf{lk}_b(\rho \cdot q')), e', e) \\
&\leq \sum_{\rho \in \mathsf{FirstRev}} \mathbb{P}_q^\sigma(\Gamma[e], \rho) \cdot \sum_{q' \in Q \setminus \{q_\rho\}} \delta_e(\rho)(q') \cdot (1 - p_{\min}(\Gamma))^n \\
&\leq \sum_{\rho \in \mathsf{FirstRev}} \mathbb{P}_q^\sigma(\Gamma[e], \rho) \cdot (1 - p_{\min}(\Gamma))^{n+1} \leq (1 - p_{\min}(\Gamma))^{n+1}
\end{aligned}$$

归纳法至此完成。 □

定理的证明 5。 现在，我们已经具备了证明定理5所需的所有要素。

证明。 考虑环境 $e \in E$、先验信念 $b \in \mathcal{D}(E)$ 和策略 $\sigma \in \mathsf{Strat}(Q,\ A)$。如果存在某个环境 $e \in E$，使得 $b(e) \leq \varepsilon$，那么 $\mathsf{NeverSmallBelief}(b,\ \varepsilon) = \emptyset$。因此，假设情况并非如此。考虑某个 $e' \neq e \in E$。

首先注意到 $\{\mathsf{nb}_{\mathsf{rev}}(e', e) \geq n_1(\Gamma, \varepsilon)\} \cap \mathsf{NeverSmallBelief}(b, \varepsilon) \subseteq \mathsf{Reveal}^{n_1(\Gamma,\varepsilon)}(\Gamma, b, e', e)$，因此根据引理 28，我们有：

$$\mathbb{P}_q^\sigma(\Gamma[e], \{\mathsf{nb}_{\mathsf{rev}}(e', e) \geq n_1(\Gamma, \varepsilon)\} \cap \mathsf{NeverSmallBelief}(b, \varepsilon)) \leq (1 - p_{\min}(\Gamma))^{n_1(\Gamma,\varepsilon)} \leq \frac{\varepsilon}{3|E|}$$

现在注意到有两种情况。首先假设 $\eta(\Gamma) = 0$。这等价于 $\iota(\Gamma) = 0$且 $n_3(\Gamma,\ \varepsilon) = 0$。这也意味着不存在不具有揭示性的区分状态-动作对。因此：

$$\left\{\mathsf{nb}_{\mathsf{dstg}}(e', e) \geq \frac{m(\Gamma, \varepsilon)}{|E|}\right\} \subseteq \{\mathsf{nb}_{\mathsf{rev}}(e', e) \geq n_1(\Gamma, \varepsilon)\}$$

因此，我们有：

$$\mathbb{P}_q^\sigma\left(\Gamma[e], \left\{\mathsf{nb}_{\mathsf{dstg}}(e', e) \geq \frac{m(\Gamma, \varepsilon)}{|E|}\right\} \cap \mathsf{NeverSmallBelief}(b, \varepsilon)\right) \leq \frac{\varepsilon}{|E|}$$

现在假设 $\eta(\Gamma) < 0$、$\iota(\Gamma) > 0$ 以及 $n_3(\Gamma,\ \varepsilon) \geq 1$。那么，根据定义，我们有：

$$\begin{aligned}
\left\{\mathsf{nb}_{\mathsf{dstg}}(e', e) \geq \frac{m(\Gamma, \varepsilon)}{|E|}\right\} \subseteq (\{\mathsf{nb}_{\mathsf{dstg}}(e', e) - \mathsf{nb}_{\mathsf{rev}}(e', e) \geq n_3(\Gamma, \varepsilon)\} \cap \{\mathsf{nb}_{\mathsf{rev}}(e', e) < n_1(\Gamma, \varepsilon)\}) \\
\bigcup \{\mathsf{nb}_{\mathsf{rev}}(e', e) \geq n_1(\Gamma, \varepsilon)\}
\end{aligned}$$

考虑任意满足 $\mathsf{nb}_{\mathsf{rev}}(e', e, \rho) \leq n_1(\Gamma, \varepsilon)$、$\mathbb{P}_q^\sigma(\Gamma[e], \rho) > 0$ 且 $\sum_{\pi \sqsubseteq \rho} v[e', e](\pi) \leq -n_2(\Gamma, \varepsilon)$ 的 $\rho \in Q \cdot (A \cdot Q)^*$。根据引理 25，我们有：

$$\begin{aligned}
\frac{\mathsf{lk}_b(\rho)(e')}{\mathsf{lk}_b(\rho)(e)} &\leq \frac{b(e')}{b(e)} \cdot 2^{\left(\sum_{\pi \sqsubseteq \rho} v[e', e](\pi)\right)} \cdot \mathsf{ratio}_{\max}(\Gamma)^{\mathsf{nb}_{\mathsf{rev}}(e', e, \rho)} \\
&\leq \frac{1}{\varepsilon} \cdot 2^{-n_2(\Gamma, \varepsilon)} \cdot \mathsf{ratio}_{\max}(\Gamma)^{\mathsf{nb}_{\mathsf{rev}}(e', e, \rho)} \leq \frac{1}{\varepsilon} \cdot \varepsilon^2 = \varepsilon
\end{aligned}$$

由此可得，对于所有 $k \in \mathbb{N}$:

$$\mathbb{P}_q^\sigma(\Gamma[e], \{\mathsf{s}^k_{v[e',e]} \le -n_2(\Gamma,\varepsilon)\} \cap \{\mathsf{nb}_{\mathsf{rev}}(e',e) < n_1(\Gamma,\varepsilon)\} \cap \mathsf{NeverSmallBelief}(b,\varepsilon)) = 0$$

此外，根据引理 26，在马尔可夫决策过程 $\Gamma[e], v[e',e]$ 是一个 [对数函数(比值$_{\min}(\Gamma)$),对数函数(比值$_{\max}(\Gamma)$)]-赋值，且其 $\eta(\Gamma)$-有界。因此，根据引理 19，对所有 $n \in \mathbb{N}$ 和 $t > 0$，我们有:

$$\mathbb{P}_q^\sigma[\Gamma[E], \{\mathsf{s}^n_{v[e',e]} \ge n \cdot \eta(\Gamma) + t\} \cap \{\mathsf{c}^n_v \ge n\}] \le \exp\left(-\frac{2t^2}{n \cdot (\log(\frac{\mathsf{ratio}_{\max}(\Gamma))}{\mathsf{ratio}_{\min}(\Gamma))})^2}\right)$$

令 $n := n_3(\Gamma,\varepsilon) \in \mathbb{N}$ 且 $t := -\frac{n\cdot\eta(\Gamma)}{2} > 0$。由于 $n \ge 2\cdot n_2(\Gamma,\varepsilon)\cdot|\eta(\Gamma)|$，我们有:

$$n \cdot \eta(\Gamma) + t = \frac{n \cdot \eta(\Gamma)}{2} \le -n_2(\Gamma,\varepsilon)$$

因此，$\{\mathsf{s}^n_{v[e',e]} \ge -n_2(\Gamma,\varepsilon)\} \subseteq \{\mathsf{s}^n_{v[e',e]} \ge n\cdot\eta(\Gamma) + t\}$。此外，由于 $n \ge 2\cdot|$对数函数$(\varepsilon)| \cdot$ (对数函数$(\frac{\mathsf{ratio}_{\max}(\Gamma)}{\text{比值}_{\min}(\Gamma)})$)$^2 \cdot \eta(\Gamma)^2$，我们有:

$$-\frac{2t^2}{n \cdot (\log(\frac{\mathsf{ratio}_{\max}(\Gamma))}{\mathsf{ratio}_{\min}(\Gamma))})^2} = -\frac{n\cdot\eta(\Gamma)^2}{2\cdot(\log(\frac{\mathsf{ratio}_{\max}(\Gamma))}{\mathsf{ratio}_{\min}(\Gamma))})^2} \le \log(\varepsilon) \le \ln(\varepsilon)$$

综上，我们得到:

$$\mathbb{P}_q^\sigma[\Gamma[E], \{\mathsf{s}^{n_3(\Gamma,\varepsilon)}_{v[e',e]} \ge -n_2(\Gamma,\varepsilon)\} \cap \{\mathsf{c}^{n_3(\Gamma,\varepsilon)}_v \ge n_3(\Gamma,\varepsilon)\}] \le \frac{\varepsilon}{2\cdot|E|}$$

因此，令 $X_1 := \{\mathsf{nb}_{\mathsf{rev}}(e',\ e) < n_1(\Gamma,\ \varepsilon)\}$:

$$\begin{aligned}
&\mathbb{P}_q^\sigma[\Gamma[E], \{\mathsf{c}^{n_3(\Gamma,\varepsilon)_v} \ge n_3(\Gamma,\varepsilon)\} \cap X_1 \cap \mathsf{NeverSmallBelief}(b,\varepsilon)] \\
&\le \mathbb{P}_q^\sigma[\Gamma[E], \{\mathsf{s}^{n_3(\Gamma,\varepsilon)}_{v[e',e]} \ge -n_2(\Gamma,\varepsilon)\} \cap \{\mathsf{c}^{n_3(\Gamma,\varepsilon)}_v \ge n_3(\Gamma,\varepsilon)\} \cap X_1 \cap \mathsf{NeverSmallBelief}(b,\varepsilon)] \\
&+ \mathbb{P}_q^\sigma[\Gamma[E], \{\mathsf{s}^{n_3(\Gamma,\varepsilon)}_{v[e',e]} \le -n_2(\Gamma,\varepsilon)\} \cap \{\mathsf{c}^{n_3(\Gamma,\varepsilon)}_v \ge n_3(\Gamma,\varepsilon)\} \cap X_1 \cap \mathsf{NeverSmallBelief}(b,\varepsilon)] \\
&\le \mathbb{P}_q^\sigma[\Gamma[E], \{\mathsf{s}^{n_3(\Gamma,\varepsilon)}_{v[e',e]} \ge -n_2(\Gamma,\varepsilon)\} \cap \{\mathsf{c}^{n_3(\Gamma,\varepsilon)}_v \ge n_3(\Gamma,\varepsilon)\}] \\
&+ \mathbb{P}_q^\sigma[\Gamma[E], \{\mathsf{s}^{n_3(\Gamma,\varepsilon)}_{v[e',e]} \le -n_2(\Gamma,\varepsilon)\} \cap X_1 \cap \mathsf{NeverSmallBelief}(b,\varepsilon)] \\
&\le \frac{\varepsilon}{2\cdot|E|}
\end{aligned}$$

最后，根据 $v[e',e]$ 的定义，状态-动作对 $(q,\ a) \in Q \times A$ 属于 $\mathsf{NZ}^{v[e',e]}$，当且仅当它是 $(e',\ e)$-区分的，且不是 e'/e-揭示的。因此，对于所有 $\rho \in Q\cdot(A\cdot Q)^*$，我们有:

$$\{\mathsf{nb}_{\mathsf{dstg}}(e',e,\rho) - \mathsf{nb}_{\mathsf{rev}}(e',e,\rho) \ge n_3(\Gamma,\varepsilon)\} = \{\mathsf{c}^{n_3(\Gamma,\varepsilon)}_v(\rho) \ge n_3(\Gamma,\varepsilon)\}$$

因此，我们有:

$$\mathbb{P}_q^\sigma(\Gamma[e], \{\mathsf{nb}_{\mathsf{dstg}}(e',e) - \mathsf{nb}_{\mathsf{rev}}(e',e) \ge n_3(\Gamma,\varepsilon)\} \cap X_1 \cap \mathsf{NeverSmallBelief}(b,\varepsilon)) \le \frac{\varepsilon}{2\cdot|E|}$$

我们得到:

$$\begin{aligned}
&\mathbb{P}_q^\sigma\left(\Gamma[e], \left\{\mathsf{nb}_{\mathsf{dstg}}(e',e) \ge \frac{m(\Gamma,\varepsilon)}{|E|}\right\} \cap X_1 \cap \mathsf{NeverSmallBelief}(b,\varepsilon)\right) \\
&\le \mathbb{P}_q^\sigma(\Gamma[e], \{\mathsf{nb}_{\mathsf{rev}}(e',e) \ge n_1(\Gamma,\varepsilon)\} \cap \mathsf{NeverSmallBelief}(b,\varepsilon)) \\
&+ \mathbb{P}_q^\sigma(\Gamma[e], \{\mathsf{nb}_{\mathsf{dstg}}(e',e) - \mathsf{nb}_{\mathsf{rev}}(e',e) \ge n_3(\Gamma,\varepsilon)\} \cap X_1 \cap \mathsf{NeverSmallBelief}(b,\varepsilon)) \\
&\le \frac{\varepsilon}{2\cdot|E|} + \frac{\varepsilon}{2\cdot|E|} = \frac{\varepsilon}{|E|}
\end{aligned}$$

这对所有 $e' \neq e \in E$ 都成立。此外，我们有：

$$\{\mathsf{nb}_{\mathsf{dstg}} \geq m(\Gamma,\varepsilon)\} \subseteq \bigcup_{e' \in E \setminus \{e\}} \left\{ \mathsf{nb}_{\mathsf{dstg}}(e,e') \geq \frac{m(\Gamma,\varepsilon)}{|E|} \right\}$$

因此：

$$\begin{aligned}
&\mathbb{P}_q^\sigma(\Gamma[e], \{\mathsf{nb}_{\mathsf{dstg}} \geq m(\Gamma,\varepsilon)\} \cap \mathsf{NeverSmallBelief}(b,\varepsilon)) \\
&\quad \leq \sum_{e' \in E \setminus \{e\}} \mathbb{P}_q^\sigma \left(\Gamma[e], \left\{ \mathsf{nb}_{\mathsf{dstg}}(e',e) \geq \frac{m(\Gamma,\varepsilon)}{|E|} \right\} \cap \mathsf{NeverSmallBelief}(b,\varepsilon) \right) \\
&\quad \leq \sum_{e' \in E \setminus \{e\}} \frac{\varepsilon}{|E|} \leq \varepsilon
\end{aligned}$$

□

A.6 算法的补充说明1

A.6.1 从多环境马尔可夫决策过程定义马尔可夫决策过程

下面正式定义算法1第 9 行构造的马尔可夫决策过程。

定义 *29.* 考虑一个多环境马尔可夫决策过程 Γ、一个奇偶目标 W 和一个环境信念 $b \in \mathcal{D}(E)$。令 $S(\Gamma,b) := \{(q,a,q') \in Q \times A \times Q \mid (q,a) \in \mathit{Dstg}(\Gamma)\ q' \in \mathit{Supp}(p[b,q,a])\}$。考虑一个赋值 $v\colon S(\Gamma,b) \to [0,1]$。令 $w_{v,b}\colon \mathit{Dstg}(\Gamma) \to [0,1]$ 通过如下方式定义，对所有 $(q,a) \in Q \times A$：

$$w_{v,b}(q,a) := \sum_{q' \in \mathsf{Supp}(p[b,q,a])} \sum_{e \in E} b(e) \cdot \delta_e(q,a)(q') \cdot v(q,a,q')$$

接着，我们定义马尔可夫决策过程 $\mathsf{G}(\Gamma,W,b,v)$ 和奇偶目标 W'（两者合称为马尔可夫决策过程-来自-多环境马尔可夫决策过程(Γ,W,b,v)，即算法 *1*中的定义）如下：

- $\mathsf{G}(\Gamma,W,b,x) := (Q',A,\delta)$，其中 $Q' := Q \cup \{q_{\text{获胜}}, q_{\text{失败}}\}$，其中 $q_{\text{获胜}}$ 和 $q_{\text{失败}}$ 是两个不属于 Q 的新状态；且 $\delta\colon Q' \times A \to \mathcal{D}(A)$，满足：
  - 对所有 $a \in A$, $\delta(q_{\mathsf{win}},a)(q_{\mathsf{win}}) := 1$，以及 $\delta(q_{\mathsf{lose}},a)(q_{\mathsf{lose}}) := 1$；- 对所有 $(q,a) \in Q \times A \setminus \mathit{Dstg}(\Gamma)$，令 $\delta(q,a) := \delta_e(q,a) \in \mathcal{D}(Q)$，对于任意 $e \in E$（它们都相同，因为 (q,a) 是不可区分状态-动作对）。- 对所有 $(q,a) \in Q \times \mathit{Dstg}(\Gamma)$，令：

$$\begin{aligned}
\delta(q,a)(q_{\mathsf{win}}) &:= w_{v,b}(q,a) \\
\delta(q,a)(q_{\mathsf{lose}}) &:= 1 - w_{v,b}(q,a)
\end{aligned}$$

- 考虑定义奇偶目标 W 的函数 $f\colon Q \to \mathbb{N}$。奇偶目标 W'由函数 $g\colon Q' \to \mathbb{N}$ 定义，该函数在 Q 上与 f 相同，并且满足 $g(q_{\text{获胜}}) := 0$，以及 $g(q_{\text{失败}}) := 1$；因此，我们有 $(q_{\text{获胜}})^\omega \in W'$ 和 $(q_{\text{失败}})^\omega \notin W'$。

在下面确立这一构造所保证的性质之前，我们先引入一些记号。

定义*30.* 考虑一个多环境马尔可夫决策过程 Γ、一个奇偶目标 W、一个环境信念 $b \in \mathcal{D}(E)$，以及一个赋值 $v\colon S(\Gamma,b) \to [0,1]$。令 $\mathit{FirstD} := \{\pi = (q_0,a_0) \cdots (q_n,a_n) \in (Q \times A)^+ \mid (q_n,a_n) \in \mathit{Dstg}(\Gamma)\ \forall i \leq n, (q_i,a_i) \notin \mathit{Dstg}(\Gamma)\} \subseteq (Q \times A)^+$。对于所有 $\pi \in \mathit{FirstD}$，令 $(q,a) \in Q \times A$ 表示 π 最近访问的状态 - 动作对，则对于所有 $e \in E$，令：$\delta_e(\pi) := \delta_e(q,a) \in \mathcal{D}(Q)$;$\delta(\pi) := \delta(q,a) \in \mathcal{D}(Q')$, $w_{v,b}(\pi) := w_{v,b}(q,a) \in [0,1]$, $p[b,\pi] := p[b,q,a] \in \mathcal{D}(e)$，并且对于所有$q' \in Q$：$v(\pi,q') := v(q,a,q')$ 和 $\lambda(b,\pi,q') := \lambda(b,q,a,q')$。

对于所有策略 $\sigma \in \mathsf{Strat}(Q,A)$ 以及 $q \in Q$，我们将 $x(\Gamma,q,\sigma,b,v) \in [0,1]$ 定义为 y:

$$x(\Gamma,q,\sigma,b,v) := \sum_{\pi \in \mathsf{FirstD}} \sum_{q' \in \mathsf{Supp}(p[b,\pi])} \sum_{e \in E} \mathbb{P}_q^\sigma[\Gamma[e],\pi] \cdot b(e) \cdot \delta_e(\pi)(q') \cdot |v(\pi,q') - \mathsf{val}_{q'}^{\mathsf{pr}}(\Gamma,\lambda(b,\pi,q'),W)|$$

我们将在下文证明这一构造所确保的性质。

引理 *31*。考虑一个多环境马尔可夫决策过程 Γ、一个奇偶目标 W、一个环境信念 $b \in \mathcal{D}(E)$，以及一个赋值 $v\colon S(\Gamma, b) \to [0,1]$。对于 $G := G(\Gamma, W, b, x)$，以及 W' 定义 *29* 中的马尔可夫决策过程和奇偶目标，对于所有 $q \in Q$，我们有：

$$|\mathsf{val}^{\mathsf{pr}}_q(\Gamma, b, W) - \mathsf{val}_q(\mathsf{G}, W')| \leq \sup_{\sigma \in \mathsf{Strat}(Q,A)} x(\Gamma, q, \sigma, b)$$

其中 $\mathsf{val}_q(\mathsf{G}\ W') := \sup_{\sigma \in \mathsf{Strat}(Q',A)} \mathbb{P}^{\sigma}_q[\mathsf{G}\ W']$。

Proof. 我们令 $\mathsf{NoD} := \{\rho = (q_0, a_0) \cdots (Q \times A)^{\omega} \mid \forall i \in \mathbb{N}, (q_i, a_i) \notin \mathsf{Dstg}(\Gamma)\} \subseteq (Q \times A)^{\omega}$.

令 $q \in Q$。考虑一个策略 $\sigma_Q \in \mathsf{Strat}(Q, A)$。对于所有 $\pi \in \mathsf{FirstD}$，我们令 $\sigma^{\pi}_Q \in \mathsf{Strat}(Q, A)$，使得对于所有 $\pi' \in Q \cdot (A \cdot Q)^*$, $\sigma^{\pi}_Q(\pi') := \sigma_Q(\pi \cdot \pi')$。我们有：

$$\forall e \in E,\ \mathbb{P}^{\sigma_Q}_q[\Gamma[e], W] = \mathbb{P}^{\sigma_Q}_q[\Gamma[e], W \cap \mathsf{NoD}] + \sum_{\pi \in \mathsf{FirstD}} \sum_{q' \in \mathsf{Supp}(p[b,\pi])} \mathbb{P}^{\sigma_Q}_q[\Gamma[e], \pi] \cdot \delta_e(\pi)(q') \cdot \mathbb{P}^{\sigma^{\pi}_Q}_{q'}[\Gamma[e], W]$$

同样，考虑策略 $\sigma_{Q'} \in \mathsf{Strat}(Q', A)$。我们有：

$$\begin{aligned}\mathbb{P}^{\sigma_{Q'}}_q[\mathsf{G}, W'] &= \mathbb{P}^{\sigma_{Q'}}_q[\mathsf{G}, W' \cap \mathsf{NoD}] + \sum_{\pi \in \mathsf{FirstD}} \mathbb{P}^{\sigma_Q}_q[\mathsf{G}, \pi] \cdot w_{v,b}(\pi) \\ &= \mathbb{P}^{\sigma_{Q'}}_q[\mathsf{G}, W' \cap \mathsf{NoD}] + \sum_{\pi \in \mathsf{FirstD}} \mathbb{P}^{\sigma_Q}_q[\mathsf{G}, \pi] \cdot \sum_{q' \in \mathsf{Supp}(p[b,\pi])} \sum_{e \in E} b(e) \cdot \delta_e(\pi)(q') \cdot v(\pi, q')\end{aligned}$$

现在假设策略 σ_Q 和 $\sigma_{Q'}$ 满足：对于 FirstD 中运行的所有前缀 $\pi \in Q \cdot (A \cdot Q)^*$，都有 $\sigma_Q(\pi) = \sigma_{Q'}(\pi)$。在这种情况下，我们有：

$$\forall e \in E,\ \mathbb{P}^{\sigma_Q}_q[\Gamma[e], W \cap \mathsf{NoD}] = \mathbb{P}^{\sigma_{Q'}}_q[\mathsf{G}, W' \cap \mathsf{NoD}]$$

并且，对于所有 $\pi \in$FirstD：

$$\forall e \in E,\ \mathbb{P}^{\sigma_Q}_q[\Gamma[e], \pi] = \mathbb{P}^{\sigma_{Q'}}_q[\mathsf{G}, \pi]$$

因此，我们有：

$$\mathsf{val}^{\mathsf{pr}}_q(\Gamma, b, \sigma_Q, W) - \mathbb{P}^{\sigma_{Q'}}_q[\mathsf{G}, W'] = \sum_{\pi \in \mathsf{FirstD}} \sum_{q' \in \mathsf{Supp}(p[b,\pi])} \sum_{e \in E} \mathbb{P}^{\sigma_Q}_q[\mathsf{G}, \pi] \cdot \Big(b(e) \cdot \delta_e(\pi)(q') \cdot (\mathbb{P}^{\sigma^{\pi}_Q}_{q'}[\Gamma[e], W] - v(\pi, q')) \Big)$$

现在，考虑任意策略 $\sigma_Q \in \mathsf{Strat}(Q, A)$，并考虑一个策略 $\sigma_{Q'} \in \mathsf{Strat}(Q', A)$，使得对于 FirstD 中运行的所有前缀 $\pi \in Q \cdot (A \cdot Q)^*$，都有 $\sigma_Q(\pi) = \sigma_{Q'}(\pi)$。注意，对于所有$\pi \in$ FirstD 和 $q' \in \mathsf{Supp}(p[b, \pi])$，我们有：

$$\sum_{e \in E} \lambda(b, \pi, q')(e) \cdot \mathbb{P}^{\sigma^{\pi}_Q}_{q'}[\Gamma[e], W] \leq \mathsf{val}^{\mathsf{pr}}_{q'}(\Gamma, \lambda(b, \pi, q'), W)$$

其中，对于所有 $e \in E$：

$$\lambda(b, t, a, q')(e) = \frac{b(e) \cdot \delta_e(\pi)(q')}{\sum_{e' \in E} b(e') \cdot \delta_{e'}(\pi)(q')}$$

因此：

$$\sum_{e \in E} b(e) \cdot \delta_e(\pi)(q') \cdot \mathbb{P}^{\sigma^{\pi}_Q}_{q'}[\Gamma[e], W] \leq \sum_{e \in E} b(e) \cdot \delta_e(\pi)(q') \cdot \mathsf{val}^{\mathsf{pr}}_{q'}(\Gamma, \lambda(b, \pi, q'), W)$$

因此，我们有：

$$\begin{aligned}\mathsf{val}_q^{\mathsf{pr}}(\Gamma,b,\sigma_Q,W) &= \sum_{\pi\in\mathsf{FirstD}}\sum_{q'\in\mathsf{Supp}(p[b,\pi])}\sum_{e\in E}\mathbb{P}_q^{\sigma_Q}[\mathsf{G},\pi]\cdot\Big(b(e)\cdot\delta_e(\pi)(q')\cdot(\mathbb{P}_{q'}^{\sigma_Q^\pi}[\Gamma[e],W]-v(\pi,q'))\Big)\\ &\qquad + \mathbb{P}_q^{\sigma_{Q'}}[\mathsf{G},W']\\ &\le \sum_{\pi\in\mathsf{FirstD}}\sum_{q'\in\mathsf{Supp}(p[b,\pi])}\sum_{e\in E}\mathbb{P}_q^{\sigma_Q}[\mathsf{G},\pi]\cdot\Big(b(e)\cdot\delta_e(\pi)(q')\cdot(\mathsf{val}_{q'}^{\mathsf{pr}}(\Gamma,\lambda(b,\pi,q'),W)-v(\pi,q'))\Big)\\ &\qquad + \mathbb{P}_q^{\sigma_{Q'}}[\mathsf{G},W']\\ &\le x(\Gamma,q,\sigma_Q,b)+\mathbb{P}_q^{\sigma_{Q'}}[\mathsf{G},W'] \le \sup_{\sigma\in\mathsf{Strat}(Q,A)} x(\Gamma,q,\sigma,b)+\mathsf{val}_q(\mathsf{G},W')\end{aligned}$$

由于这对所有 $\sigma_{Q'}\in\mathsf{Strat}(Q,A)$，我们有：$(\text{值}_q^{\mathsf{pr}}(\Gamma,b,W)-\mathsf{val}_q(\mathsf{G},W'))\le\sup_{\sigma\in\mathsf{Strat}(Q,A)}x(\Gamma,q,\sigma,b,v)$。

现在考虑任意策略 $\sigma_{Q'}\in\mathsf{Strat}(Q',A)$。令 $\varepsilon>0$。令 $\sigma_Q\in\mathsf{Strat}(Q,A)$ 为 Γ 中的一个策略，该策略：与策略$\sigma_{Q'}$ 在 FirstD 中运行的 $Q\cdot(A\cdot Q)^*$ 的所有前缀上相同；此外，对于所有 $\pi\in\mathsf{FirstD}$，该策略 σ_Q 满足：策略 $\sigma_Q^\pi\in\mathsf{Strat}(Q,A)$ 对所有 $q'\in\mathsf{Supp}(p[b,\pi])$ 均保证：

$$\sum_{e\in E}\lambda(b,\pi,q')(e)\cdot\mathbb{P}_{q'}^{\sigma_Q^\pi}[\Gamma[e],W]\ge\mathsf{val}_{q'}^{\mathsf{pr}}(\Gamma,\lambda(b,\pi,q'),W)-\varepsilon$$

因此

$$\sum_{e\in E}b(e)\cdot\delta_e(\pi)(q')\cdot\mathbb{P}_{q'}^{\sigma_Q^\pi}[\Gamma[e],W]\ge\sum_{e\in E}b(e)\cdot\delta_e(\pi)(q')\cdot\mathsf{val}_{q'}^{\mathsf{pr}}(\Gamma,\lambda(b,\pi,q'),W)-\varepsilon$$

在这种情况下，我们有：

$$\begin{aligned}\mathbb{P}_q^{\sigma_{Q'}}[\mathsf{G},W'] &= \sum_{\pi\in\mathsf{FirstD}}\sum_{q'\in\mathsf{Supp}(p[b,\pi])}\sum_{e\in E}\mathbb{P}_q^{\sigma_Q}[\mathsf{G},\pi]\cdot\Big(b(e)\cdot\delta_e(\pi)(q')\cdot(v(\pi,q')-\mathbb{P}_{q'}^{\sigma_Q^\pi}[\Gamma[e],W])\Big)\\ &\qquad + \mathsf{val}_q^{\mathsf{pr}}(\Gamma,b,\sigma_Q,W)\\ &\le \sum_{\pi\in\mathsf{FirstD}}\sum_{q'\in\mathsf{Supp}(p[b,\pi])}\sum_{e\in E}\mathbb{P}_q^{\sigma_Q}[\mathsf{G},\pi]\cdot\Big(b(e)\cdot\delta_e(\pi)(q')\cdot(v(\pi,q')-\mathsf{val}_{q'}^{\mathsf{pr}}(\Gamma,\lambda(b,\pi,q'),W))\Big)\\ &\qquad + \mathsf{val}_q^{\mathsf{pr}}(\Gamma,b,\sigma_Q,W)+\varepsilon\\ &\le x(\Gamma,q,\sigma_Q,b)+\mathsf{val}_q^{\mathsf{pr}}(\Gamma,b,\sigma_Q,W)+\varepsilon \le \sup_{\sigma\in\mathsf{Strat}(Q,A)} x(\Gamma,q,\sigma,b,v)+\mathsf{val}_q^{\mathsf{pr}}(\Gamma,b,W)+\varepsilon\end{aligned}$$

由于这对所有 $\sigma_{Q'}\in\mathsf{Strat}(Q',A)$ 和 $\varepsilon>0$ 都成立，因此我们有：$(\mathsf{val}_q(\mathsf{G},W')-\mathsf{val}_q^{\mathsf{pr}}(\Gamma,b,W))\le\sup^{\sigma\in\mathsf{Strat}(Q,A)}x(\Gamma,q,\sigma,b)$。证明至此结束。 □

基于上述引理，我们现在确立一个结果，该结果将在算法 1 的正确性证明中发挥关键作用（归纳步骤）。首先介绍用于表述所关注性质的核心概念。

定义32。考虑一个多环境马尔可夫决策过程 Γ、一个奇偶目标 W 以及一个环境信念 $b\in\mathcal{D}(E)$。对于所有 $n\ge1$ 和 $\gamma\in[0,1]$，我们定义 $\mathsf{N}(\Gamma,b,n,\gamma):=\{\mathsf{nb}_{\mathsf{dstg}}\ge n\}\cap\mathsf{NeverSmallBelief}(b,\gamma)$。那么，对于所有状态 $q\in Q$ 和 $\eta\ge0$，若实数值 $x_q\in[0,1]$ 满足以下条件，则称其为(n,γ,η,b,q)-近似：

$$|\mathsf{val}_q^{\mathsf{pr}}(\Gamma,b,W)-x_q|\le\eta+\mathsf{val}_q^{\mathsf{pr}}(\Gamma,b,\mathsf{N}(\Gamma,b,n,\gamma))$$

我们可以陈述并证明所需的引理。

引理 33。考虑一个多环境马尔可夫决策过程 Γ、一个奇偶目标 W 和一个环境信念 $b \in \mathcal{D}(E)$。令 $n \in \mathbb{N}$ 和 $\gamma \in [0,1]$ 满足 $\min_{e \in \mathsf{Supp}(b)} b(e) > \gamma$。令 $\eta \geq 0$。考虑一个赋值 $v\colon S(\Gamma, b) \to [0,1]$，并假设对于所有 $(q,a,q') \in S(\Gamma,b)$，值 $v(q,a,q')$ 都是 $(n,\gamma,\eta,\lambda(b,q,a,q')\ q')$-近似。那么，对于 $G := G(\Gamma, W, b, x)$，以及定义 29 中的马尔可夫决策过程和奇偶目标，对于所有 $q \in Q$，值 $val_q(G\ W')$ 都是 $(n+1,\gamma,\eta,b,q)$-近似。

证明。 根据引理 31，我们有：

$$|\mathsf{val}^{\mathsf{pr}}_q(\Gamma, b, W) - \mathsf{val}_q(\mathsf{G}, W')| \leq \sup_{\sigma \in \mathsf{Strat}(Q,A)} x(\Gamma, q, \sigma, b, v)$$

令 $\sigma \in \mathsf{Strat}(Q,\ A)$。对于所有 $\pi \in \mathsf{FirstD}$ 和 $q' \in \mathsf{Supp}(p[b,\pi])$，$v(\pi,\ q')$ 是一个 $(n,\ \gamma,\ \eta,\ \lambda(b,\ \pi,\ q')\ q')$-近似，因此：

$$\begin{aligned} x(\Gamma,q,\sigma,b,v) &= \sum_{\pi \in \mathsf{FirstD}} \sum_{q' \in \mathsf{Supp}(p[b,\pi])} \sum_{e \in E} \mathbb{P}^{\sigma}_q[\Gamma[e],\pi] \cdot b(e) \cdot \delta_e(\pi)(q') \cdot \underbrace{|v(\pi,q') - \mathsf{val}^{\mathsf{pr}}_{q'}(\Gamma, \lambda(b,\pi,q'), W)|}_{\leq \eta + \mathsf{val}^{\mathsf{pr}}_q(\Gamma, \lambda(b,\pi,q'), \mathsf{N}(\Gamma, \lambda(b,\pi,q'), n, \gamma))} \\ &\leq \eta + \sum_{\pi \in \mathsf{FirstD}} \sum_{q' \in \mathsf{Supp}(p[b,\pi])} \sum_{e \in E} \mathbb{P}^{\sigma}_q[\Gamma[e],\pi] \cdot b(e) \cdot \delta_e(\pi)(q') \cdot \mathsf{val}^{\mathsf{pr}}_q(\Gamma, \lambda(b,\pi,q'), \mathsf{N}(\Gamma, \lambda(b,\pi,q'), n, \gamma)) \end{aligned}$$

令 $\varepsilon > 0$。我们考虑一个策略 $\sigma_a \in \mathsf{Strat}(Q,\ A)$，使得 σ_a 和 σ 在 FirstD 中所有运行的前缀上相同；并且对于所有 $\pi \in \mathsf{FirstD}$，策略 σ_a 满足：对于所有 $q' \in \mathsf{Supp}(p[b,\pi])$，$\sigma^{\pi}_s$ 确保：

$$\mathsf{val}^{\mathsf{pr}}_q(\Gamma, \lambda(b,\pi,q'), \sigma^{\pi}_a, \mathsf{N}(\Gamma, \lambda(b,\pi,q'), n, \gamma)) \geq \mathsf{val}^{\mathsf{pr}}_q(\Gamma, \lambda(b,\pi,q'), \mathsf{N}(\Gamma, \lambda(b,\pi,q'), n, \gamma)) - \varepsilon$$

也就是说：

$$\sum_{e \in E} \lambda(b,\pi,q')(e) \cdot \mathbb{P}^{\sigma^{\pi}_a}_{q'}[\Gamma[e], \mathsf{N}(\Gamma, \lambda(b,\pi,q'), n, \gamma)] \geq \mathsf{val}^{\mathsf{pr}}_q(\Gamma, \lambda(b,\pi,q'), \mathsf{N}(\Gamma, \lambda(b,\pi,q'), n, \gamma)) - \varepsilon$$

根据 $\lambda(b,\ \pi,\ q')$ 的定义，这意味着：

$$\begin{aligned} &\sum_{e \in E} b(e) \cdot \delta_e(\pi,q') \cdot \mathbb{P}^{\sigma^{\pi}_a}_{q'}[\Gamma[e], \mathsf{N}(\Gamma, \lambda(b,\pi,q'), n, \gamma)] + \varepsilon \geq \\ &\qquad \sum_{e \in E} b(e) \cdot \delta_e(\pi,q') \cdot \mathsf{val}^{\mathsf{pr}}_q(\Gamma, \lambda(b,\pi,q'), \mathsf{N}(\Gamma, \lambda(b,\pi,q'), n, \gamma)) \end{aligned}$$

此外，根据 FirstD 的定义，对于所有 $\pi \in \mathsf{FirstD}$，概率 $\mathbb{P}^{\sigma}_q[\Gamma[e],\pi] \in [0,1]$ 不依赖于环境 $e \in E$。因此，我们有：

$$\begin{aligned} x(\Gamma,q,\sigma,b,v) &\leq \eta + \sum_{\pi \in \mathsf{FirstD}} \sum_{q' \in \mathsf{Supp}(p[b,\pi])} \sum_{e \in E} \mathbb{P}^{\sigma}_q[\Gamma[e],\pi] \cdot b(e) \cdot \delta_e(\pi)(q') \cdot \mathsf{val}^{\mathsf{pr}}_q(\Gamma, \lambda(b,\pi,q'), \mathsf{N}(\Gamma, \lambda(b,\pi,q'), n, \gamma)) \\ &\leq \eta + \sum_{\pi \in \mathsf{FirstD}} \sum_{q' \in \mathsf{Supp}(p[b,\pi])} \sum_{e \in E} \mathbb{P}^{\sigma}_q[\Gamma[e],\pi] \cdot b(e) \cdot \delta_e(\pi)(q') \cdot \mathbb{P}^{\sigma^{\pi}_a}_{q'}[\Gamma[e], \mathsf{N}(\Gamma, \lambda(b,\pi,q'), n, \gamma)] + \varepsilon \\ &= \eta + \sum_{\pi \in \mathsf{FirstD}} \sum_{q' \in \mathsf{Supp}(p[b,\pi])} \sum_{e \in E} \mathbb{P}^{\sigma_a}_q[\Gamma[e],\pi] \cdot b(e) \cdot \delta_e(\pi)(q') \cdot \mathbb{P}^{\sigma^{\pi}_a}_{q'}[\Gamma[e], \mathsf{N}(\Gamma, \lambda(b,\pi,q'), n, \gamma)] + \varepsilon \\ &= \eta + \sum_{e \in E} b(e) \cdot \mathbb{P}^{\sigma_a}_q[\Gamma[e], \mathsf{N}(\Gamma, b, n+1, \gamma)] + \varepsilon \\ &\leq \eta + \mathsf{val}^{\mathsf{pr}}_q(\Gamma, b, \mathsf{N}(\Gamma, b, n+1, \gamma)) + \varepsilon \end{aligned}$$

第二个等式源于以下事实：$\min_{e \in \mathsf{Supp}(b)} b(e) > \gamma$；并且根据 FirstD 的定义，对于所有 $\pi \in \mathsf{FirstD}$ 和 $q' \in \mathsf{Supp}(p[b,\pi])$，我们有 $\mathsf{lk}_b(\pi \cdot q') = \lambda(b,\ \pi,\ q')$。

由于该不等式对所有策略 $\sigma \in \mathsf{Strat}(Q,\ A)$ 和 $\varepsilon > 0$ 都成立，我们可推得：

$$\sup_{\sigma \in \mathsf{Strat}(Q,A)} x(\Gamma,q,\sigma,b,v) \leq \eta + \mathsf{val}^{\mathsf{pr}}_q(\Gamma, b, \mathsf{N}(\Gamma, b, n+1, \gamma))$$

该引理随后由引理31推出。 □

A.6.2 算法1的正确性

我们考虑算法1的一个略作修改的版本，其中在第 8 行，我们所考虑的值$v_{q,\ a,\ q'}$被截断，并采用$(\frac{\varepsilon}{3|E|}\cdot m^{1}(\frac{\varepsilon}{3|E|}))$-近似，从而它可以写在$\lceil$对数函数$(\frac{\varepsilon}{3|E|}\cdot m^{1}(\frac{\varepsilon}{3|E|}))\rceil$比特上。

下面给出这个略作修改的算法1版本的正确性证明。

定理 *34.* 考虑一个多环境马尔可夫决策过程 Γ、一个奇偶目标 W，以及一个环境信念 $b\in\mathcal{D}(E)$。对于所有 $\varepsilon>0$，在 $n:=m(\Gamma,\frac{\varepsilon}{3|E|})$ 和 $\gamma:=\varepsilon$ 上执行算法 *1*（在第 *8* 行进行截断）会输出 $f\colon Q\to[0,1]$，使得对于所有 $q\in Q$，都有 $|val_q^{pr}(\Gamma,b,W)-f(q)|\le\varepsilon$。

证明。我们对 $1\le|\mathsf{Supp}(b)|\le E$ 进行归纳，证明输出 $f\colon Q\to[0,1]$ 算法1的结果满足：对于所有 $q\in Q$，都有 $|\mathsf{val}_q^{pr}(\Gamma,b,W)-f(q)|\le|\mathsf{Supp}(b)|\cdot\frac{\varepsilon}{|E|}$。

当 $|\mathsf{Supp}(b)|=1$ 时，这显然成立，因为在这种情况下，算法 1 不会进行任何近似。现在假设对于某个 $j\le|E|-1$，所有 $1\le i\le j$ 都满足该结果。我们通过对 $t\in\mathbb{N}$ 作归纳法来证明：对于所有满足 $|\mathsf{Supp}(b)|=j+1$ 的信念 $b\in\mathcal{D}(E)$，算法 1 在 b 上执行，并取 $n:=t$ 和 $\gamma:=\varepsilon$ 时，其输出 $f\colon Q\to[0,1]$ 满足：对于所有 $q\in Q$，$|\mathsf{val}_q^{pr}(\Gamma,b,W)-f(q)|\le|\mathsf{Supp}(b)|\cdot\frac{\varepsilon}{|E|}-\frac{2\varepsilon}{3|E|}+t\cdot\frac{\varepsilon}{3|E|}\cdot m^{1}(\frac{\varepsilon}{3|E|})+\mathsf{val}_q^{pr}(\Gamma,b,\mathsf{N}(\Gamma,b,t,\frac{\varepsilon}{3|E|}))$。

我们先考虑 $t=0$ 的情形，并令 $b\in\mathcal{D}(E)$，使得 $|\mathsf{Supp}(b)|=j+1$。令$q\in Q$。如果 $\min_{e\in\mathsf{Supp}(b)}b(e)>\frac{\varepsilon}{3|E|}$，那么 $\mathsf{val}_q^{pr}(\Gamma,b,\mathsf{N}(\Gamma,b,t,\frac{\varepsilon}{3|E|}))=1$，因此不等式成立。否则，对于 $b':=\mathsf{Truncate}(b)$，有 $\mathsf{Diff}(b,b')\le\frac{\varepsilon}{3|E|}$，因此由引理 4 得：$|\mathsf{val}_q^{pr}(\Gamma,b,W)-\mathsf{val}_q^{pr}(\Gamma,b',W)|\le\frac{\varepsilon}{3|E|}$。此外，$|\mathsf{Supp}(b')|<|\mathsf{Supp}(b)|$，因此根据归纳假设，对于状态 q（第 4 行），算法在 b' 上递归运行，并采用 $n:=m(\Gamma,\frac{\varepsilon}{3|E|})$ 和 $\gamma:=\varepsilon$，计算得到的值 $x_q\in[0,1]$ 满足 $|\mathsf{val}_q^{pr}(\Gamma,b',W)-x_q|\le|\mathsf{Supp}(b')|\cdot\frac{\varepsilon}{|E|}$。因此，我们有：$|\mathsf{val}_q^{pr}(\Gamma,b,W)-x_q|\le|\mathsf{Supp}(b')|\cdot\frac{\varepsilon}{|E|}+\frac{\varepsilon}{3|E|}\le|\mathsf{Supp}(b)|\cdot\frac{\varepsilon}{|E|}-\frac{2\varepsilon}{3|E|}$。

现在假设对于某个 $t\in\mathbb{N}$，归纳假设成立，并令 $b\in\mathcal{D}(E)$ 满足 $|\mathsf{Supp}(b)|=j+1$。设 $q\in Q$。我们考虑在 b 上执行算法且满足 $n:=t+1$ 和 $\gamma=\varepsilon$ 的情况。如果 $\min_{e\in\mathsf{Supp}(b)}b(e)\le\frac{\varepsilon}{3|E|}$，则与上述情况相同：算法在 b' 上递归调用，并使用 $=\mathsf{Truncate}(b)$ 以及 $n:=m(\Gamma,\frac{\varepsilon}{3|E|})$ 和 $\gamma:=\varepsilon$，针对状态 q 计算出的值 $x_q\in[0,1]$（第 4 行）满足 $|\mathsf{val}_q^{pr}(\Gamma,b,W)-x_q|\le|\mathsf{Supp}(b)|\cdot\frac{\varepsilon}{|E|}-\frac{2\varepsilon}{3|E|}$。

现在假设 $\min_{e\in\mathsf{Supp}(b)}b(e)>\frac{\varepsilon}{3|E|}$，因此进入第 5 行的 for 循环。对于所有 $(q,\ a,\ q')\in S(\Gamma,\ b)$，通过递归调用算法 1（第 8 行）计算得到一个值 $v_{q,\ a,\ q'}\in[0,1]$。注意，值 $v_{q,\ a,\ q'}\in[0,1]$ 是在调用算法 1。考虑 $b':=\lambda[b,q,a,q']\in\mathcal{D}(E)$，有两种情况：

- 如果 $\mathsf{Supp}(b')\subsetneq\mathsf{Supp}(b)$，那么 $v_{q,a,q'}$ 来自于在 b' 上执行算法调用的结果，其中$n:=m(\Gamma,\frac{\varepsilon}{3|E|})$ 以及$\gamma:=\varepsilon$。根据我们的归纳假设（关于支撑集的大小），有 $|\mathsf{valpr}_{q'}(\Gamma,b',W)-v_{q,a,q'}|\le|\mathsf{Supp}(b')|\cdot\frac{\varepsilon}{|E|}+\frac{\varepsilon}{3|E|}\cdot m^{1}(\frac{\varepsilon}{3|E|})\le|\mathsf{Supp}(b)|\cdot\frac{\varepsilon}{|E|}-\frac{2\varepsilon}{3|E|}+\frac{\varepsilon}{3|E|}\cdot m^{1}(\frac{\varepsilon}{3|E|})$。

- 否则，$v_{q,a,q'}$ 来自于在 b' 上执行算法调用的结果，其中 $n:=t$ 以及$\gamma:=\varepsilon$。根据我们的归纳假设（关于 n），有 $|\mathsf{valpr}_{q'}(\Gamma,b',W)-v_{q,a,q'}|\le|\mathsf{Supp}(b)|\cdot\frac{\varepsilon}{|E|}-\frac{2\varepsilon}{3|E|}+t\cdot\frac{\varepsilon}{3|E|}\cdot m^{1}(\frac{\varepsilon}{3|E|})+\mathsf{valpr}_{q'}(\Gamma,b',\mathsf{N}(\Gamma,b',t,\frac{\varepsilon}{3|E|}))+\frac{\varepsilon}{3|E|}\cdot m^{1}(\frac{\varepsilon}{3|E|})=|\mathsf{Supp}(b)|\cdot\frac{\varepsilon}{|E|}-\frac{2\varepsilon}{3|E|}+(t+1)\cdot\frac{\varepsilon}{3|E|}\cdot m^{1}(\frac{\varepsilon}{3|E|})+\mathsf{valpr}_{q'}(\Gamma,b',\mathsf{N}(\Gamma,b',t,\frac{\varepsilon}{3|E|}))$.

这对所有 $(q,a,q')\in S(\Gamma,b)$ 均成立。因此，根据引理 33（对于 $\gamma=\frac{\varepsilon}{3|E|}$ 和 $\eta=|\mathsf{Supp}(b)|\cdot\frac{\varepsilon}{|E|}-\frac{2\varepsilon}{3|E|}+(t+1)\cdot\frac{\varepsilon}{3|E|}\cdot m^{1}(\frac{\varepsilon}{3|E|})$），由在 b 上使用 $n:=t+1$ 和 $\gamma=\varepsilon$ 运行的算法在第 9 行计算得到的马尔可夫决策过程 G 和奇偶目标 W' 满足 $|\mathsf{val}_q^{pr}(\Gamma,b,W)-\mathsf{val}_q(\mathsf{G}\ W')|\le$

$|\mathrm{Supp}(b)| \cdot \frac{\varepsilon}{|E|} - \frac{2\varepsilon}{3|E|} + (t+1) \cdot \frac{\varepsilon}{3|E|} \cdot m\left(\Gamma, \frac{\varepsilon}{3|E|}\right) + \mathrm{val}_q^{\mathrm{pr}}(\Gamma, b, \mathsf{N}(\Gamma, b, t+1, \frac{\varepsilon}{3|E|}))$. 关于 $n := t+1$ 的归纳假设成立，因为 MDP-Parity 算法计算出的值恰好是 $\mathrm{val}_q(\mathsf{G}\ W')$。

事实上，归纳假设（关于 n）对所有 $t \in \mathbb{N}$都成立。特别是，在 b上执行算法 1，且 $n := m(\Gamma, \frac{\varepsilon}{3|E|})$和 $\gamma := \varepsilon$时，其输出$f\colon Q \to [0,1]$满足：对所有 $q \in Q$，$|\mathrm{val}_q^{\mathrm{pr}}(\Gamma, b, W) - f(q)| \leq |\mathrm{Supp}(b)| \cdot \frac{\varepsilon}{|E|} - \frac{\varepsilon}{3|E|} + \mathrm{val}_q^{\mathrm{pr}}(\Gamma, b, \mathsf{N}(\Gamma, b, m(\Gamma, \frac{\varepsilon}{3|E|}) \frac{\varepsilon}{3|E|}))$。此外，由定理 5可知，对所有 $q \in Q$：$\mathrm{val}_q^{\mathrm{pr}}(\Gamma, b, \mathsf{N}(\Gamma, b, m(\Gamma, \frac{\varepsilon}{3|E|}) \frac{\varepsilon}{3|E|})) \leq \frac{\varepsilon}{3|E|}$。因此，对所有 $q \in Q$，都有 $|\mathrm{val}_q^{\mathrm{pr}}(\Gamma, b, W) - f(q)| \leq |\mathrm{Supp}(b)| \cdot \frac{\varepsilon}{|E|}$。

因此，当信念支持集的大小等于 $j+1$时，归纳假设（关于信念的支持集）成立。事实上，对于所有 $1 \leq j \leq |E|$，它都成立。因此，对于所有信念 b，对于所有 $\varepsilon > 0$，在 $n := m(\Gamma, \frac{\varepsilon}{3|E|})$ 和 $\gamma := \varepsilon$ 上执行算法 1，会输出一个函数 $f\colon Q \to [0,1]$，使得对于所有 $q \in Q$，我们都有 $|\mathrm{val}_q^{\mathrm{pr}}(\Gamma, b, W) - f(q)| \leq |\mathrm{Supp}(b)| \cdot \frac{\varepsilon}{|E|} \leq \varepsilon$。 □

A.6.3 算法1 的复杂度

我们先陈述一个关于表示更新后的信念所需比特数的引理。

引理 *35*。考虑一个多环境马尔可夫决策过程 Γ 和一个信念 $b \in \mathcal{D}(E)$。假设存在 $N \in \mathbb{N}$，使得对于所有 $e \in E$，$b(e) = \frac{x_e}{N}$，其中某个 $0 \leq x_e \leq N$。还假设存在 $T \in \mathbb{N}$，使得对于所有 $e \in E$，以及 $(q, a, q') \in Q \times A \times Q$，我们有 $\delta_e(q,a)(q') = \frac{y_{q,a,q',e}}{T}$，其中某些 $0 \leq y_{q,a,q',e} \leq T$。那么，对于所有满足 $p[b,q,a,q'] > 0$ 的 $(q,a,q') \in Q \times A \times Q$，存在 $N' \in \mathbb{N}$，使得 $\log(N') \leq \log(N) + \log(T) + \log(|E|)$；并且对于所有 $e \in E$，存在 $0 \leq z \leq N'$，使得$\lambda[b,q,a,q'](e) = \frac{z}{N'}$。

证明。令 $(q,\ a,\ q') \in Q \times A \times Q$ 满足 $p[b,q,a,q'] > 0$，并令 $e \in E$。我们有：

$$\lambda[b,q,a,q'](e) = \frac{b(e) \cdot \delta_e(q,a)(q')}{\sum_{e' \in E} b(e') \cdot \delta_{e'}(q,a)(q')} = \frac{\frac{x_e}{N} \cdot \frac{y_{q,a,q',e}}{T}}{\sum_{e' \in E} \frac{x_{e'}}{N} \cdot \frac{y_{q,a,q',e'}}{T}} = \frac{x_e \cdot y_{q,a,q',e}}{\sum_{e' \in E} x_{e'} \cdot y_{q,a,q',e'}}$$

分母 $\sum_{e' \in E} x_{e'} \cdot y_{q,\ a,\ q',\ e'}$ 与所考虑的环境 e 无关。此外：

$$\log\left(\sum_{e' \in E} x_{e'} \cdot y_{q,a,q',e'}\right) \leq \log\left(\sum_{e' \in E} N \cdot T\right) = \log(|E| \cdot N \cdot T) = \log(|E|) + \log(N) + \log(T)$$

□

现在我们可以陈述并证明算法 1 的复杂度。请注意，我们已经在一个略作修改的版本上证明了该算法的正确性（定理34），其中第 8 行中我们考虑的值 $v_{q,\ a,\ q'}$经过截断，并采用（$\frac{\varepsilon}{3|E|} \cdot m\left(\frac{\varepsilon}{3|E|}\right)$）近似，因此可以用 $\lceil \rceil$对数函数($\frac{\varepsilon}{3|E|} \cdot m\left(\frac{\varepsilon}{3|E|}\right)$)$\rceil$比特表示。我们在对第 8 行作出相同修改的情况下，证明算法 1的复杂度。

定理 *36*. 考虑一个多环境马尔可夫决策过程 Γ、一个奇偶目标 W、一个环境信念 $b \in \mathcal{D}(E)$，以及 $\varepsilon > 0$。令 $n := |Q|$，$k := |E|$，$r := |A|$，$x := \lceil \frac{1}{\varepsilon} \rceil$和 $l \in \mathbb{N}$（分别为 $h \in \mathbb{N}$）表示在 Γ 中写出概率（分别为 b 中的概率）所需的最大比特数。那么，算法 *1*（在第 *8* 行进行截断），对于 $n := m(\Gamma, \frac{\varepsilon}{3|E|})$ 和 $\gamma := \varepsilon$，在空间$O(poly(n, r, k, h, x, 2^l))$内执行。

证明。正如论文正文部分所述，算法的每次递归调用要么针对支撑集更小的信念，要么针对截断前待访问的可区分状态-动作对数量上界更小的情况。因此，根据定理 $k \cdot m(\Gamma, \frac{\varepsilon}{3k})$，算法的递归深度上界为 $m(\Gamma, \frac{\varepsilon}{3k}) = O(poly(k, 2^l, x))$。5。

此外，先验信念 b 可以表示为所有概率具有相同分母的形式（如引理 35 所示），且表示该分母至多需要 $k \cdot h$ 比特。多环境马尔可夫决策过程中的所有概率也可以采用同样的方式表示，其分母至多需要 $l' := k \cdot n^2 \cdot r \cdot l$ 比特（因为我们需要处理所有四元组 $(q, a, q', e) \in Q \times A \times Q \times E$）。因此，每次递归调用中处理的信念，在所有环境下都用相同的分母表示。信念被截断时，分母保持不变（参见引理 4 的证明）。此外，信念最多更新 $k \cdot m(\Gamma, \frac{\varepsilon}{3k})$ 次；根据引理 35，这只会使表示分母所需的比特数增加 $(l' + \text{对数函数}(k)) \cdot k \cdot m(\Gamma, \frac{\varepsilon}{3k})$。因此，所有递归调用中用于表示信念的比特数为 $O(poly(n, r, k, h, x, 2^l))$。

此外，第 8 行中的值会被截断，因此可以用 $O(poly(k, x, 2^l))$ 个比特表示。因此，在对算法 1 的所有（递归）调用中，第 9 行构建的马尔可夫决策过程都具有 $n+2$ 个状态，其概率可以用 $O(poly(n, r, k, h, x, 2^l))$ 个比特表示（回顾定义 29）。因此，对 MDP-Parity 的调用都能在关于 $O(poly(n, r, k, h, x, 2^l))$ 的多项式时间内执行（[2]）。总之，我们可以得出，算法 1 在 $O(poly(n, r, k, h, x, 2^l))$ 空间内执行。□

B 章节补充内容4

B.1 引理 8 的证明

在开始引理的证明之前，我们先形式化定义混合策略的 uni 值和 pr 值，其定义方式与策略 $\sigma \in \mathsf{Strat}(Q, A)$ 类似。

定义37。 现在考虑一个多环境马尔可夫决策过程 $\Gamma = (Q, A, E, (\delta_e)_{e \in E})$。对于所有状态 $q \in Q$、信念$b \in \mathcal{D}(E)$以及混合策略 $\tau \in \mathcal{D}(\mathsf{Strat}(Q, A))$，令：

$$\mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \tau, W) := \sum_{e \in E} b(e) \cdot \mathbb{P}_q^{\tau}[\Gamma[e], W] = \sum_{e \in E} \sum_{\tau \in \mathsf{Strat}(Q,A)} b(e) \cdot \tau(\sigma) \cdot \mathbb{P}_q^{\sigma}[\Gamma[e], W]$$

$$\mathsf{val}_q^{\mathsf{uni}}(\Gamma, \tau, W) := \min_{e \in E} \mathbb{P}_q^{\tau}[\Gamma[e], W] = \min_{e \in E} \sum_{\sigma \in \mathsf{Strat}(Q,A)} \tau(\sigma) \cdot \mathbb{P}_q^{\sigma}[\Gamma[e], W]$$

现在我们继续证明引理 8。

证明。 考虑任意两个非空集合 A, B和 $f\colon A \times B \to [0,1]$。设 $g\colon \mathcal{D}(A) \times \mathcal{D}(B) \to [0,1]$（其概率分布具有可数支撑）满足：对于所有 $(\tau_A, \tau_B) \in \mathcal{D}(A) \times \mathcal{D}(B)$，均有 $g(\tau_A, \tau_B) := \sum_{a \in A} \sum_{b \in B} \tau_A(a) \cdot \tau_B(b) \cdot f(a, b)$。那么，如果 A 或 B 是有限的：

$$\sup_{\tau_a \in \mathcal{D}(A)} \inf_{\tau_b \in \mathcal{D}(B)} g(\tau_a, \tau_b) = \inf_{\tau_b \in \mathcal{D}(B)} \sup_{\tau_a \in \mathcal{D}(A)} g(\tau_a, \tau_b)$$

这一结果对应于冯·诺伊曼极小极大定理 [20]，当 A 和 B 均为有限集时；其中一个集合无限（可能不可数）的情形是标准推广，参见例如[18]。应用于我们的情形，可得：

$$\sup_{\tau \in \mathcal{D}(\mathsf{Strat}(Q,A))} \inf_{b \in \mathcal{D}(E)} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \tau, W) = \inf_{b \in \mathcal{D}(E)} \sup_{\tau \in \mathcal{D}(\mathsf{Strat}(Q,A))} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \tau, W)$$

考虑任意 $\tau \in \mathcal{D}(\mathsf{Strat}(Q, A))$。令 $e \in E$满足 $\mathsf{val}_q^{\mathsf{pr}}(\Gamma, e, \tau, W) = \min_{e' \in E} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, e', \tau, W)$。那么，对于所有 $b \in \mathcal{D}(E)$，我们有：

$$\mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \tau, W) = \sum_{e' \in E} b(e') \cdot \mathsf{val}_q^{\mathsf{pr}}(\Gamma, e', \tau, W) \geq \sum_{e' \in E} b(e') \cdot \mathsf{val}_q^{\mathsf{pr}}(\Gamma, e, \tau, W) = \mathsf{val}_q^{\mathsf{pr}}(\Gamma, e, \tau, W)$$

因此：$\inf_{b \in \mathcal{D}(E)} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \tau, W) = \min_{e' \in E} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, e', \tau, W) = \mathsf{val}_q^{\mathsf{uni}}(\Gamma, \tau, W)$。

类似地，考虑任意 $b \in \mathcal{D}(E)$ 和 $\varepsilon > 0$。令 $\sigma \in \mathsf{Strat}(Q, A)$，使得 $\mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \sigma, W) \geq \sup_{\sigma' \in \mathsf{Strat}(Q, A)} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \sigma', W) - \varepsilon$。那么，对于所有 $\tau \in \mathcal{D}(\mathsf{Strat}(Q, A))$，我们有：

$$\begin{aligned}\mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \tau, W) &= \sum_{\sigma' \in \mathsf{Strat}(Q,A)} \tau(\sigma') \cdot \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \sigma', W) \leq \sum_{\sigma' \in \mathsf{Strat}(Q,A)} \tau(\sigma') \cdot \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \sigma, W) + \varepsilon \\ &= \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \sigma, W) + \varepsilon \leq \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W) + \varepsilon\end{aligned}$$

由于这对所有 $\varepsilon > 0$ 都成立，我们得到 $\sup^{\tau \in \mathcal{D}(\mathsf{Strat}(Q,A))} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, \tau, W) = \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W)$。总之，我们有：

$$\sup_{\tau \in \mathcal{D}(\mathsf{Strat}(Q,A))} \mathsf{val}_q^{\mathsf{uni}}(\Gamma, \tau, W) = \inf_{b \in \mathcal{D}(E)} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W)$$

□

B.2 引理9的证明

B在继续证明引理9之前，我们先来看看它如何推出定理7

.

定理7的证明。由引理8和9，我们有：

$$\mathsf{val}_q^{\mathsf{uni}}(\Gamma, W) = \sup_{\tau \in \mathcal{D}(\mathsf{Strat}(Q,A))} \mathsf{val}_q^{\mathsf{uni}}(\Gamma, \tau, W) = \inf_{b \in \mathcal{D}(E)} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W)$$

定理得证。 □

现在我们继续证明引理9。请注意，这实际上与[8,引理 19]的证明类似。

证明。考虑一个多环境马尔可夫决策过程 Γ、一个状态 $q \in Q$ 和一个混合策略 $\tau \in \mathcal{D}(\mathsf{Strat}(Q, A))$。对于所有 $\rho \in Q \cdot (A \cdot Q)^*$、$e \in E$ 以及 $\sigma \in \mathsf{Strat}(Q, A)$，我们令 $p[\sigma, e, \rho] := \tau(\sigma) \cdot \mathbb{P}_q^\sigma[\Gamma[e], \rho]$ 和 $p[\tau, e, \rho] := \sum_{\sigma \in \mathsf{Strat}(Q,A)} p[\sigma, e, \rho] = \mathbb{P}_q^\tau[\Gamma[e], \rho]$。现在考虑任意一个 $\rho \in Q \cdot (A \cdot Q)^*$。我们来证明，对于所有满足 $p[\tau, e, \rho], p[\tau, e', \rho] \neq 0$ 的 $e, e' \in E$，都有：对于所有 $\sigma \in \mathsf{Strat}(Q, A)$：

$$\frac{p[\sigma, e, \rho]}{p[\tau, e, \rho]} = \frac{p[\sigma, e', \rho]}{p[\tau, e', \rho]}$$

事实上，考虑任意 $e \in E$ 和 $\sigma \in \mathsf{Strat}(Q, A)$。将 ρ 记作 $\rho = q_0 \cdot (a_0, q_1) \cdots (a_{n-1}, q_n)$，其中 $q_0 := q_\circ$。对于所有 $0 \leq i \leq n$，令 $\rho_{\leq i} := q_0 \cdots (a_{i-1}, q_i)$。（特别地，$\rho_{\leq 0} = q_\circ$。）我们有：

$$\mathbb{P}_q^\sigma[\Gamma[e], \rho] = \Pi_{i=0}^{n-1} \sigma(\rho_{\leq i})(a_i) \cdot \delta_e(q_i, a_i)(q_{i+1})$$

因此：

$$\frac{p[\sigma, e, \rho]}{p[\tau, e, \rho]} = \frac{\tau(\sigma) \cdot \prod_{i=0}^{n-1} \sigma(\rho_{\leq i})(a_i) \cdot \delta_e(q_i, a_i)(q_{i+1})}{\sum_{\sigma' \in \mathsf{Strat}(Q,A)} \tau(\sigma') \cdot \prod_{i=0}^{n-1} \sigma'(\rho_{\leq i})(a_i) \cdot \delta_e(q_i, a_i)(q_{i+1})} = \frac{\tau(\sigma) \cdot \prod_{i=0}^{n-1} \sigma(\rho_{\leq i})(a_i)}{\sum_{\sigma' \in \mathsf{Strat}(Q,A)} \tau(\sigma') \cdot \prod_{i=0}^{n-1} \sigma'(\rho_{\leq i})(a_i)}$$

这不依赖于环境 e，因此 $\frac{p[\sigma,e,\rho]}{p[\tau,e,\rho]} = \frac{p[\sigma,e',\rho]}{p[\tau,e',\rho]}$。

现在，定义策略 $\sigma \in \mathsf{Strat}(Q, A)$。对于所有 $\rho \in Q \cdot (A \cdot Q)^*$，令 $\sigma(\rho) \in \mathcal{D}(A)$ 满足，对于所有 $a \in A$：

$$\sigma(\rho)(a) := \begin{cases} \sum_{\sigma' \in \mathsf{Strat}(Q,A)} \sigma'(\rho)(a) \cdot \frac{p[\sigma', e, \rho]}{p[\tau, e, \rho]} \\ \text{is arbitrary} \end{cases} \quad \begin{array}{l} \text{对于某个 } e \in E, \text{ 使得 } p[\tau, e, \rho] \neq 0 \text{（如果存在），} \\ \text{若 } p[\tau, e, \rho] = 0, \text{ 则对于所有 } e \in E \end{array}$$

考虑一个环境 $e \in E$。我们通过对 $\rho \in Q \cdot (A \cdot Q)^*$ 使用归纳法来证明 $\mathbb{P}_q^\sigma[\Gamma[e], \rho] = p[\tau, e, \rho]$。显然，这对所有 $\rho \in Q$ 都成立。现在假设对于某个 $\rho \in Q \cdot (A \cdot Q)^*$，该结论成立，并令 $(a, q') \in Q$。若 $p[\tau, e, \rho] = \mathbb{P}_q^\sigma[\Gamma[e], \rho] = 0$，则 $p[\tau, e, \rho \cdot (a, q')]$，$\mathbb{P}_q^\sigma[\Gamma[e], \rho \cdot (a, q')] = 0$。因此，我们假设 $p[\tau, e, \rho] = \mathbb{P}_q^\sigma[\Gamma[e], \rho] > 0$。我们有：

$$\begin{aligned}
\mathbb{P}_q^\sigma[\Gamma[e], \rho \cdot (a, q')] &= \mathbb{P}_q^\sigma[\Gamma[e], \rho] \cdot \sigma(\rho)(a) \cdot \delta_e(\mathsf{last}(\rho), a)(q') \\
&= p[\tau, e, \rho] \cdot \left(\sum_{\sigma' \in \mathsf{Strat}(Q,A)} \sigma'(\rho)(a) \cdot \frac{p[\sigma', e, \rho]}{p[\tau, e, \rho]} \right) \cdot \delta_e(\mathsf{last}(\rho), a)(q') \\
&= \sum_{\sigma' \in \mathsf{Strat}(Q,A)} \sigma'(\rho)(a) \cdot p[\sigma', e, \rho] \cdot \delta_e(\mathsf{last}(\rho), a)(q') \\
&= \sum_{\sigma' \in \mathsf{Strat}(Q,A)} p[\sigma', e, \rho \cdot (a, q')] \\
&= p[\tau, e, \rho \cdot (a, q')]
\end{aligned}$$

因此，性质对 $\rho \cdot (a, q') \in Q \cdot (A \cdot Q)^*$ 也成立。事实上，它对所有 $\rho \in Q \cdot (A \cdot Q)^*$ 都成立。由此可得，对所有 $\rho \in Q \cdot (A \cdot Q)^*$：

$$\mathbb{P}_q^\sigma[\Gamma[e], \mathsf{cyl}(\rho)] = \sum_{\sigma' \in \mathsf{Strat}(Q,A)} \tau(\sigma') \cdot \mathbb{P}_q^{\sigma'}[\Gamma[e], \mathsf{cyl}(\rho)] = \mathbb{P}_q^\tau[\Gamma[e], \mathsf{cyl}(\rho)]$$

根据 Carath´eodory 扩张定理，概率测度完全由柱集的测度确定。由于 $\mathbb{P}_q^\sigma[\Gamma[e], \cdot]$: 博雷尔$(Q)$ 和 $\sum_{\sigma' \in \mathsf{Strat}(Q, A)} \tau(\sigma') \cdot \mathbb{P}_q^{\sigma'}[\Gamma[e], \cdot]$: 博雷尔$(Q)$ 是概率测度，因此，对于所有博雷尔目标 $W \in$ 博雷尔(Q)：

$$\mathbb{P}_q^\sigma[\Gamma[e], W] = \mathbb{P}_q^\tau[\Gamma[e], W]$$

这对所有环境 $e \in E$ 都成立，因此：

$$\mathsf{val}_q^{\mathsf{uni}}(\Gamma, \tau, W) = \mathsf{val}_q^{\mathsf{uni}}(\Gamma, \sigma, W) \leq \mathsf{val}_q^{\mathsf{uni}}(\Gamma, W)$$

这对所有 $\sigma \in \mathcal{D}(\mathsf{Strat}(Q, A))$ 都成立。此外，我们有 $\mathsf{val}_q^{\mathsf{uni}}(\Gamma, W) \leq \sup_{\tau \in \mathcal{D}(\mathsf{Strat}(Q, A))} \mathsf{val}_q^{\mathsf{uni}}(\Gamma, \tau, W)$。

该引理得证。 □

B.3 定理 10 的证明

证明。与证明概要中一样，令 $N := \lceil \text{对数函数}(\frac{1}{\varepsilon}) \rceil$, $k := |E|$, and $S := \{b \in \mathcal{D}(E) \mid \forall e \in E, \mathcal{B}(\in \{\frac{x}{k \cdot 2^{N+1}} \mid 0 \leq x \leq k \cdot 2^{N+1}\}\}$ $|S| = O\left(2^{k^2 \cdot N}\right)$ $\varepsilon \geq \frac{1}{2^N}$)。我们确实有 ( )，同时也有 。

考虑任意 $b \in \mathcal{D}(E)$。定义 $b' \in S$ 使得 $\mathsf{Diff}(b, b') \leq \frac{\varepsilon}{2}$。考虑任意 $e \in E$。对于所有 $e' \in E \setminus \{e\}$，令 $b'(e') := \frac{x_{e'}}{k \cdot 2^{N+1}}$ 为满足 $b'(e) \leq b(e)$ 的最大 $0 \leq x \leq k \cdot 2^{N+1}$。因此，我们有 $0 \leq b(e') - b'(e') \leq \frac{1}{k \cdot 2^{N+1}}$。此外，我们有 $\sum_{e' \in E \setminus \{e\}} x_{e'} = k \cdot 2^{N+1} \cdot \left(\sum_{e' \in E \setminus \{e\}} b'(e') \right) \leq k \cdot 2^{N+1} \left(\sum_{e' \in E \setminus \{e\}} b(e') \right) \leq k \cdot 2^{N+1}$。于是，我们可以定义 $b'(e) := \frac{\sum_{e' \in E \setminus \{e\}} x_{e'}}{k \cdot 2^{N+1}}$，并且有 $b(e) \leq b'(e) \in [0, 1]$。我们定义了 $b' \in \mathcal{D}(E)$。如引理 4 的证明中所述，我们有：

$$\mathsf{Diff}(b, b') = \sum_{\substack{\tilde{e} \in E \\ b(\tilde{e}) \geq b'(\tilde{e})}} (b(\tilde{e}) - b'(\tilde{e}))$$

因此：

$$\mathsf{Diff}(b, b') = \sum_{\substack{\tilde{e} \in E \\ b(\tilde{e}) \geq b'(\tilde{e})}} (b(\tilde{e}) - b'(\tilde{e})) = \sum_{e' \in E \setminus \{e\}} (b(e') - b'(e')) \leq \sum_{e' \in E \setminus \{e\}} \frac{1}{k \cdot 2^{N+1}} \leq \frac{1}{2^{N+1}} \leq \frac{\varepsilon}{2}$$

因此，根据定理 7 和引理 4：

$$\mathsf{val}_q^{\mathsf{uni}}(\Gamma, W) = \inf_{b\in\mathcal{D}(E)} \text{值}_q^{\mathsf{pr}}(\Gamma, b, W) \leq \text{最小值}_{b\in S} \text{值}_q^{\mathsf{pr}}(\Gamma, b, W) \leq \inf_{b\in\mathcal{D}(E)} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W) + \frac{\varepsilon}{2} = \text{值}_q^{\mathsf{uni}}(\Gamma, W) + \frac{\varepsilon}{2}$$

于是：

$$|\min_{b\in S} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W) - \mathsf{val}_q^{\mathsf{uni}}(\Gamma, W)| \leq \frac{\varepsilon}{2}$$

此外，对任意 $b \in S$ 进行编码所需的比特数在 N 和 k 中均为多项式。因此，在任意具有 $\gamma := \varepsilon/2$ 和 $n := m(\Gamma, \frac{\varepsilon}{4k})$ 的信念 $b \in S$ 上执行算法 1，可在指数空间中（如果概率采用一元编码，则可在多项式空间中）计算出一个值 $v_b \in [0,1]$，使得 $|\mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W) - v_b| \leq \frac{\varepsilon}{2}$。令 $b_m \in S$ 满足 $\min_{b\in S} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W) = \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b_m, W)$，并令 $b_s \in S$ 满足 $\min_{b\in S} v_b = v_{b_s}$，则有：

$$-\frac{\varepsilon}{2} \leq \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b_m, W) - v_{b_m} \leq \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b_m, W) - v_{b_s} \leq \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b_s, W) - v_{b_s} \leq \frac{\varepsilon}{2}$$

因此，我们有：

$$|\min_{b\in S} \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b, W) - \min_{b\in S} v_b| \leq \frac{\varepsilon}{2}$$

于是：

$$|\min_{b\in S} v_b - \mathsf{val}_q^{\mathsf{uni}}(\Gamma, W)| \leq \varepsilon$$

S 中的环境信念可以在多项式空间内枚举，因此可以在指数空间中（如果概率采用一元编码，则可在多项式空间中）计算值 $\min_{b\in S} v_b$，定理得证。

□

B.4 命题的证明 11

我们首先介绍本小节将使用的对象。我们固定一个多环境马尔可夫决策过程Γ，其中 $E := \{e_1, e_2\}$，并考虑某个状态 $q \in Q$以及一个奇偶目标 $W \in$ 博雷尔(Q)。

为简化记号，我们令 $S := \mathsf{Strat}(Q, A)$。此外，由于只有两个环境 e_1 和 e_2，分布 $b \in \mathcal{D}(E)$完全由其赋予环境 e_1的值所确定。因此，对所有 $x \in [0,1]$，我们令 $b_x \in \mathcal{D}(E)$满足 $b_x(e_1) := x$ 且$b_x(e_2) := 1 - x$。然后我们令 $g\colon S \times [0,1] \to [0,1]$ 满足：对所有 $\sigma \in S$和 $x \in [0,1]$，$g(\sigma, x) := \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b_x, \sigma, W) = x \cdot \mathbb{P}_q^{\sigma}[\Gamma[e_1], W] + (1-x) \cdot \mathbb{P}_q^{\sigma}[\Gamma[e_2], W] \in [0,1]$。我们还令$f\colon [0,1] \to [0,1]$ 满足：对所有 $x \in [0,1]$，$f(x) := \mathsf{val}_q^{\mathsf{pr}}(\Gamma, b_x, W) = \sup_{\sigma\in S} g(\sigma, x)$。

现在，我们来证明函数 f 满足以下引理。

引理 38. 对于所有 $x < y < z \in [0,1]$，我们有：

$$f(y) \leq \min(f(x), f(z))$$

此外，考虑任意 $x < z \in [0,1]$。令 $y := \frac{z+x}{2}$，$\delta := z - x$，以及 $\eta := \max(|f(y) - f(x)|, |f(z) - f(y)|)$，则：

$$f(y) - \frac{2\eta}{\delta} \leq \inf_{t\in[0,1]} f(t) \leq f(y)$$

证明。首先考虑某个 $x < y < z \in [0,1]$。令 $\varepsilon > 0$和 $\sigma \in S$满足 $g(\sigma, y) \geq f(y) - \varepsilon$。令 $a_1 := g(\sigma, 1)$ 和 $a_2 := g(\sigma, 0)$。分两种情况：

- 若 $a_1 \geq a_2 \in [0,1]$，则 $g(\sigma, z) - g(\sigma, y) = (za_1 + (1-z)a_2) - (ya_1 + (1-y)a_2) = (z-y)(a_1 - a_2) \geq 0$。因此：$f(y) - \varepsilon \leq g(\sigma, y) \leq g(\sigma, z) \leq f(z) \leq \min(f(x), f(z))$；
- 若 $a_1 < a_2 \in [0,1]$，则 $g(\sigma, x) - g(\sigma, y) = (xa_1 + (1-x)a_2) - (ya_1 + (1-y)a_2) = (x-y)(a_1 - a_2) > 0$。因此：$f(y) - \varepsilon \leq g(\sigma, y) \leq g(\sigma, x) \leq f(x) \leq \min(f(x), f(z))$。

算法 2 Inf-f(x,t,ε)

输入： $x<t\in[0,1]$， $\varepsilon>0$

if $t-x\le\frac{\varepsilon}{2}$ **then return** Compute-f$(x,\frac{\varepsilon}{2})$
$y\leftarrow\frac{2x+t}{3}$, $a_y\leftarrow$ Compute-f$(y,\frac{\varepsilon^2}{48})$
$z\leftarrow\frac{x+2t}{3}$, $a_z\leftarrow$ Compute-f$(z,\frac{\varepsilon^2}{48})$
$u\leftarrow\frac{y+z}{2}$, $a_u\leftarrow$ Compute-f$(u,\frac{\varepsilon^2}{48})$
if $a_u>a_z+\frac{\varepsilon^2}{24}$ or $a_y>a_u+\frac{\varepsilon^2}{24}$ **then return** Inf-f(y,t,ε)
if $a_u>a_y+\frac{\varepsilon^2}{24}$ or $a_z>a_u+\frac{\varepsilon^2}{24}$ **then return** Inf-f(x,z,ε)
return a_u

由于这对所有 $\varepsilon>0$ 都成立，我们有 $f(y)\le\min(f(x),f(z))$。

现在考虑某个 $x<z\in[0,1]$，并令 $y:=\frac{z+x}{2}$， $\delta:=z-x$，以及 $\eta:=\max(|f(y)-f(x)|,|f(z)-f(x)|)$。令 $\varepsilon>0$并考虑某个 $\sigma\in S$使得 $g(\sigma,y)\ge f(y)-\varepsilon$。令$a_1:=g(\sigma,1)$ 以及 $a_2:=g(\sigma,0)$。我们有：$g(\sigma,y)=g(\sigma,z)+(z-y)(a_2-a_1)$。因此

$$f(y)-\varepsilon\le g(\sigma,y)=g(\sigma,z)+(z-y)(a_2-a_1)\le f(z)+\frac{\delta(a_2-a_1)}{2}$$

因此：

$$a_1-a_2\le\frac{2}{\delta}(f(z)-f(y)+\varepsilon)\le\frac{2}{\delta}(\eta+\varepsilon)$$

对称地，我们有： $g(\sigma,y)=g(\sigma,x)+(y-x)(a_1-a_2)$。因此

$$f(y)-\varepsilon\le g(\sigma,y)=g(\sigma,x)+(y-x)(a_1-a_2)\le f(x)+\frac{\delta(a_1-a_2)}{2}$$

因此：

$$a_2-a_1\le\frac{2}{\delta}(f(x)-f(y)+\varepsilon)\le\frac{2}{\delta}(\eta+\varepsilon)$$

由于这对所有 $\varepsilon>0$ 都成立，我们得到：

$$|a_1-a_2|\le\frac{2(\eta+\varepsilon)}{\delta}$$

现在，令 $a:=\min(a_1,a_2)$。我们有 $f(y)-\varepsilon\le g(\sigma,y)\le a+\frac{2}{(\eta+\varepsilon)\delta}$。此外，对于所有 $t\in[0,1]$，都有 $f(t)\ge g(\sigma,t)\ge a$。综上，我们得到：

$$f(y)-\varepsilon-\frac{2(\eta+\varepsilon)}{\delta}\le a\le\inf_{t\in[0,1]}f(t)\le f(y)$$

由于这对 $\varepsilon>0$ 成立，引理得证。 □

现在我们可以开始证明命题 11。

证明。考虑名为 Inf-f 的算法 2。算法 2 调用的算法 Compute-f 接收输入 $u\in[0,1]$ 以及某个 $\varepsilon>0$，并输出一个值 $v\in[0,1]$，使得 $|v-f(u)|\le\varepsilon$。实际上，调用 Compute-f(u,ε) 可能只需调用算法 1，其中 $b:=b_u$, $\gamma:=\varepsilon$ 和$n:=m(\frac{\varepsilon}{3|E|})$（见定理 34）。在此证明中，我们将对 Compute-f 的调用视为对一个预言机的调用，该预言机近似于 f (因此也近似于 $\mathrm{val}_q^{\mathrm{pr}}$)。

设 $\varepsilon>0$。我们对 $k\in\mathbb{N}$ 作归纳法证明：对于所有满足 $t-x\le\frac{\varepsilon}{2}\cdot(\frac{3}{2})^k$ 的 $x<t\in[0,1]$，在 (x,t,ε) 上调用算法 Inf-f，至多进行 $k-1$ 次递归调用（调用 Inf-f），并输出一个值 $v\in[0,1]$，使得 $|v-\sup_{u\in[x,t]}f(u)|\le\varepsilon$。

首先考虑满足 $t-x\le\frac{\varepsilon}{2}$ 的情形 $k=0$和 $x<t\in[0,1]$。此时，Inf-f(x,t,ε) 输出值 $v\in[0,1]$，该值由 Compute-f$(x,\frac{\varepsilon}{2})$ 返回，且没有递归调用（对 Inf-f），并满足

$|v - f(x)| \le \frac{\varepsilon}{2}$。此外，对于所有 $u \in [x,t]$，都有 $\mathrm{Diff}(b_x, b_u) = u - x \le \frac{\varepsilon}{2}$。因此，根据引理 4，有 $|f(u) - f(x)| \le \frac{\varepsilon}{2}$。于是，$|f(x) - \inf_{u\in[x,t]} f(u)| \le \frac{\varepsilon}{2}$，从而 $|v - \sup_{u\in[x,t]} f(u)| \le \varepsilon$。

现在假设对于某个 $k \in \mathbb{N}$，归纳性质成立，并考虑满足 $t - x \le \frac{\varepsilon}{2} \cdot (\frac{3}{2})^{k+1}$ 的某个 $x < t \in [0,1]$。如果 $t - x \le \frac{\varepsilon}{2}$，则与上面相同。否则，我们有：$y = \frac{2x+t}{3}$，$z = \frac{x+2t}{3}$，$u = \frac{y+z}{2}$，以及 $|a_y - f(y)| \le \frac{\varepsilon^2}{48}$，$|a_z - f(z)| \le \frac{\varepsilon^2}{48}$，$|a_u - f(u)| \le \frac{\varepsilon^2}{48}$。分三种情况。

- 如果 $a_u > a_z + \frac{\varepsilon^2}{24}$，那么 $f(u) > f(z)$。我们有 $y < u < z$，因此根据引理 38，有 $f(u) \le \max(f(y), f(z))$，进而 $f(y) \ge f(u) > f(z)$。此外，如果 $a_y > a_u + \frac{\varepsilon^2}{24}$，那么 $f(y) > f(u)$。因此：$f(y) > \min(f(u), f(z))$。

  于是，对于任意 $x \le \alpha < y$，由于 $\alpha < y < u < z$，我们有 $f(y) \le \max(f(\alpha), f(z))$，并且 $f(y) \le \max(f(\alpha), f(u))$，因此 $f(\alpha) \ge f(y) > \min(f(u), f(z))$。所以：$\inf_{\alpha\in[x,t]} f(\alpha) = \inf_{\alpha\in[y,t]} f(\alpha)$。此外，我们有：

$$t - y = t - \frac{2x+t}{3} = \frac{2}{3}(t-x) \le \frac{\varepsilon}{2} \cdot (\frac{3}{2})^k$$

  因此，根据归纳假设，Inf-f$(y,\ t,\ \varepsilon)$ 输出一个值 $v \in [0,1]$，递归调用（调用 Inf-f）的次数至多为 $k-1$，且满足 $|v - \inf_{\alpha\in[x,t]} f(\alpha)| = |v - \inf_{\alpha\in[y,t]} f(\alpha)| \le \varepsilon$。值 v 由 Inf-f$(x,\ t,\ \varepsilon)$ 输出，递归调用次数至多为 k。

- 若 $a_u > a_y + \frac{\varepsilon^2}{24}$，则 $f(u) > f(y)$。我们有 $y < u < z$，因此根据引理 38，可得 $f(u) \le \max(f(y), f(z))$，从而 $f(z) \ge f(u) > f(y)$。此外，若 $a_z > a_u + \frac{\varepsilon^2}{24}$，则 $f(z) > f(u)$。因此：$f(z) > \min(f(u), f(y))$。

  那么，对于任意 $z < \alpha \le t$，由于 $y < u < z < \alpha$，我们有 $f(z) \le \max(f(y), f(\alpha))$ 且 $f(z) \le \max(f(u), f(\alpha))$，从而 $f(\alpha) \ge f(z) > \min(f(y), f(u))$。因此：$\inf_{\alpha\in[x,t]} f(\alpha) = \inf_{\alpha\in[x,z]} f(\alpha)$。此外，我们有：

$$z - x = \frac{x+2t}{3} - x = \frac{2}{3}(t-x) \le \frac{\varepsilon}{2} \cdot (\frac{3}{2})^k$$

  因此，根据归纳假设，Inf-f(x,z,ε) 输出值 $v \in [0,1]$，至多进行 $k-1$ 次递归调用（调用 Inf-f），并满足 $|v - \inf_{\alpha\in[x,t]} f(\alpha)| = |v - \inf_{\alpha\in[x,z]} f(\alpha)| \le \varepsilon$。值 v 由 Inf-f(x,t,ε) 输出，至多进行 k 次递归调用。

- 如果以上条件均不适用，那么 $|a_u - a_y| \le \frac{\varepsilon^2}{24}$ 且 $|a_u - a_z| \le \frac{\varepsilon^2}{24}$。因此：$|f(u) - f(y)| \le |f(u) - a_u| + |f(y) - a_y| + |a_u - a_y| \le \frac{\varepsilon^2}{48} + \frac{\varepsilon^2}{48} + \frac{\varepsilon^2}{24} = \frac{\varepsilon^2}{12}$，类似地 $|f(u) - f(z)| \le \frac{\varepsilon^2}{12}$。因此，根据引理38，我们有：

$$|f(u) - \inf_{\alpha\in[x,t]} f(t)| \le \frac{2 \cdot \frac{\varepsilon^2}{12}}{z-y} = \frac{\varepsilon^2}{2(t-x)} \le \frac{\varepsilon^2}{\varepsilon} = \varepsilon$$

  最后这两个不等式源于以下事实：$z - y = \frac{t-x}{3}$ 和 $t - x \ge \frac{\varepsilon}{2}$。

总的来说，这个归纳假设对所有 $k \in \mathbb{N}$ 都成立。

现在，令 $k := \lceil \log(\frac{3}{2}) \cdot \log(\frac{2}{\varepsilon}) \rceil \in \mathbb{N}$。注意，$k = O(N)$ 且 $\frac{\varepsilon}{2} \cdot (\frac{3}{2})^k \ge 1$。于是，Inf-f$(0, 1, \varepsilon)$ 输出一个值 $v \in [0,1]$，最多进行 $k-1$ 次递归调用，使得 $|v - \mathrm{val}_q^{\mathrm{uni}}(\Gamma, W)| = |v - \sup_{u\in[0,1]} f(u)| \le \varepsilon$（根据定理 7，因为对于所有 $u \in [0,1]$，都有 $f(u) = \mathrm{val}_q^{\mathrm{pr}}(\Gamma, b_u, W)$）。此外，每次递归调用至多涉及三次对 Compute-f 的调用，其精度为 $\frac{\varepsilon}{2}$ 或 $\frac{\varepsilon^2}{48}$，二者均关于 ε 为多项式级别。另外，调用 Compute-f 时所用的 $[0,1]$ 中的值属于集合 $X := \{\frac{x}{3^k} \mid 0 \le x \le 3^k\}$，或是该集合 X 中两个值的平均值；因此，相应的信念可用多项式数量的比特表示（关于 N）。命题得证。 □

B.5 推论的证明 12

证明。 文献 [16, 定理 26] 表明，对于双环境 MEMDP 中概率以一元编码（精度以二进制编码）给出的、具有 一元编码值和可达性目标的[4] 间隙问题 (P1) 是NP难问题。现在，考虑双环境 MEMDP 中具有 pr编码值和奇偶目标的间隙问题 (P2)，其中概率以一元编码给出，信念和精度以二进制编码给出。若 (P2) 可在多项式时间内求解，则可通过二分搜索计算 γ近似的pr编码值（γ 以二进制编码给出）；只需考虑小于 γ的 (P2) 精度）。然而，根据命题 11，若能在多项式时间内计算 γ近似的pr编码值，则问题 (P1) 可在多项式时间内求解。因此，若 (P2) 可在多项式时间内求解，则 (P1) 也可；并且 $\mathsf{P} = \mathsf{NP}$。

□

C 第 5节补充内容

C.1 由多环境马尔可夫决策过程诱导的部分可观测马尔可夫决策过程是狄拉克保持的

命题 39. 对于所有多环境马尔可夫决策过程 Γ，部分可观测马尔可夫决策过程 $\Lambda(\Gamma)$ 都是狄拉克保持的。

证明。考虑任意 $(q, e) \in Q \times E$ 和动作 $a \in A$。对于所有 $(q', e') \in Q \times E$，我们有 $\delta((q, e)\ a)((q', e')) > 0$ 意味着 $e' = e$；并且 $O((q', e')) = q'$。因此，对于所有兼容观测 $q' \in \mathsf{Comp}((q, e)\ a) \subseteq Q$，我们有 $\lambda((q, e)\ a, q')((q', e)) = 1$，从而$H(\lambda((q, e)\ a, q')) = 0$。

□

C.2 命题14的证明

我们先快速介绍部分可观测马尔可夫决策过程的一种记号。

定义 *40.* 考虑一个狄拉克保持型部分可观测马尔可夫决策过程 $\Lambda = (Q, A, \delta, \Omega\ O)$. 对于所有 $q \in Q$、所有 $a \in A$以及所有 $o \in \mathit{Comp}(q, a)$，我们令 $\mathit{sc}(q, a, o) \in Q$ 满足 $\lambda(q, a, o)(\mathit{sc}(q, a, o)) = 1$。我们还令 $\mathit{Prec}(q, a, o) := \{q' \in Q \mid q = \mathit{sc}(q', a, o)\}$。

现在我们继续证明命题14。

证明。 首先假设 Λ 具有非增熵。考虑任意状态 $q \in Q$ 和动作 $a \in A$。我们有:

$$0 = H(q) \geq \sum_{o \in \mathsf{Comp}(q,a)} p(q, a, o) \cdot H(\lambda(q, a, o)) \geq 0$$

那么，对于所有 $o \in \mathsf{Comp}(q, a)$，由于 $p(q, a, o) > 0$，我们必然有 $H(\lambda(q, a, o)) = 0$。因此，Λ是狄拉克保持的。

现在假设 Λ是狄拉克保持的。

对于所有 $(q,\ a) \in Q \times A$, $o \in \mathsf{Comp}(q,\ a)$ 和 $q' \in O^{-1}(o)$，我们有:

$$\delta(q, a)(q') = \begin{cases} p(q, a, o) > 0 & \text{if } q' = \mathsf{sc}(q, a, o) \\ 0 & \text{otherwise} \end{cases}$$

因此，我们有:

$$\sum_{o \in \mathsf{Comp}(q,a)} p(q, a, o) = \sum_{o \in \mathsf{Comp}(q,a)} \delta(q, a)(\mathsf{sc}(q, a, o)) = \sum_{o \in \mathsf{Comp}(q,a)} \sum_{q' \in O^{-1}(o)} \delta(q, a)(q') = \sum_{q' \in Q} \delta(q, a)(q') = 1$$

[4]它们可以直接编码为奇偶目标。

I此外，对于所有信念 $b \in \mathcal{D}(Q)$、动作 $a \in A$ 和观测 $o \in \mathsf{Comp}(b,a)$:

$$\begin{aligned} p(b,a,o) &= \sum_{q \in O^{-1}(o)} \sum_{q' \in Q} b(q') \cdot \delta(q',a)(q) = \sum_{q \in O^{-1}(o)} \sum_{q' \in \mathsf{Prec}(q,a,o)} b(q') \cdot \delta(q',a)(q) \\ &= \sum_{q \in O^{-1}(o)} b(q) \cdot \delta(q,a)(\mathsf{sc}(q,a,o)) = \sum_{q \in O^{-1}(o)} b(q) \cdot p(q,a,o) \end{aligned}$$

现在，我们如下定义信念 define $\lambda'(b,a,o) \in \mathcal{D}(Q)$，对所有 $q \in Q$:

$$\lambda'(b,a,o)(q) := \begin{cases} 0 & \text{if } O(q) \neq o \\ \frac{1}{p(b,a,o)} \cdot b(q) \cdot p(q,a,o) & \text{otherwise} \end{cases}$$

我们来证明以下两个事实，对所有信念 $b \in \mathcal{D}(Q)$ 和动作 $a \in A$:

$$\forall o \in \mathsf{Comp}(b,a): \; H(\lambda'(b,a,o)) \geq H(\lambda(b,a,o)) \tag{1}$$

$$H(b) \geq \sum_{o \in \mathsf{Comp}(b,a)} p(b,a,o) \cdot H(\lambda'(b,a,o)) \tag{2}$$

我们先证明公式（1）。对于所有信念 $b \in \mathcal{D}(Q)$、动作 $a \in A$、兼容观测 $o \in \mathsf{Comp}(b,a)$ 以及 $q \in O^{-1}(o)$，我们有:

$$\lambda(b,a,o)(q) = \frac{1}{p(b,a,o)} \sum_{q' \in Q} b(q') \cdot \delta(q',a)(q) = \sum_{q' \in \mathsf{Prec}(q,a,o)} \frac{b(q') \cdot p(q',a,o)}{p(b,a,o)} = \sum_{q' \in \mathsf{Prec}(q,a,o)} \lambda'(b,a,o)(q')$$

因此，我们有:

$$\begin{aligned} H(\lambda(b,a,o)) &= -\sum_{q \in O^{-1}(o)} \lambda(b,a,o)(q) \cdot \log(\lambda(b,a,o)(q)) \\ &= -\sum_{q \in O^{-1}(o)} \left(\sum_{q' \in \mathsf{Prec}(q,a,o)} \lambda'(b,a,o)(q') \right) \cdot \log \left(\sum_{q' \in \mathsf{Prec}(q,a,o)} \lambda'(b,a,o)(q') \right) \\ &= -\sum_{q \in O^{-1}(o)} \sum_{q' \in \mathsf{Prec}(q,a,o)} \left(\lambda'(b,a,o)(q') \cdot \log \left(\sum_{q'' \in \mathsf{Prec}(q,a,o)} \lambda'(b,a,o)(q'') \right) \right) \\ &\leq -\sum_{q \in O^{-1}(o)} \sum_{q' \in \mathsf{Prec}(q,a,o)} \lambda'(b,a,o)(q') \cdot \log \left(\lambda'(b,a,o)(q') \right) \\ &= -\sum_{\substack{q' \in Q \\ o \in \mathsf{Comp}(q',a)}} \lambda'(b,a,o)(q') \cdot \log \left(\lambda'(b,a,o)(q') \right) \\ &= -\sum_{q' \in Q} \lambda'(b,a,o)(q') \cdot \log \left(\lambda'(b,a,o)(q') \right) = H(\lambda'(b,a,o)) \end{aligned}$$

倒数第二个等式源于以下事实：$\{q' \in \mathsf{Prec}(q,a,o) \mid q \in O^{-1}(o)\} = \{q' \in Q \mid \mathsf{sc}(q',a,o) = q,\ q \in O^{-1}(o)\} = \{q' \in Q \mid \mathsf{sc}(q',a,o) \in O^{-1}(o)\} = \{q' \in Q \mid o \in \mathsf{Comp}(q',a)\}$。最后一个等式源于以下事实：对于所有满足 $o \notin \mathsf{Comp}(q',a)$ 的 $q' \in Q$，我们有 $p(q',a,o) = 0$，因此 $\lambda'(b,a,o)(q') = 0$。这证明了公式（1）。

现在让我们转向公式（2）。令 $x := \sum_{o \in \mathsf{Comp}(b,a)} p(b,a,o) \cdot H(\lambda'(b,a,o))$，并根据

詹森不等式，由于对数函数是凹函数，我们有：

$$
\begin{aligned}
x &= \sum_{o\in\mathsf{Comp}(b,a)} p(b,a,o)\cdot\left(-\sum_{q\in Q}\lambda'(b,a,o)(q)\cdot\log(\lambda'(b,a,o)(q))\right)\\
&= -\sum_{o\in\mathsf{Comp}(b,a)}\sum_{q\in Q}\left[b(q)\cdot p(q,a,o)\cdot\log\left(\frac{b(q)\cdot p(q,a,o)}{p(b,a,o)}\right)\right]\\
&= -\sum_{o\in\mathsf{Comp}(b,a)}\sum_{q\in Q} b(q)\cdot p(q,a,o)\cdot\log\left(b(q)\right) - \sum_{o\in\mathsf{Comp}(b,a)}\sum_{q\in Q}\left[b(q)\cdot p(q,a,o)\cdot\log\left(\frac{p(q,a,o)}{p(b,a,o)}\right)\right]\\
&= -\sum_{q\in Q} b(q)\cdot\log\left(b(q)\right) + \sum_{o\in\mathsf{Comp}(b,a)} p(b,a,o)\cdot\left[\sum_{q\in Q}\frac{b(q)\cdot p(q,a,o)}{p(b,a,o)}\cdot\log\left(\frac{p(b,a,o)}{p(q,a,o)}\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&\le H(b) + \sum_{o\in\mathsf{Comp}(b,a)} p(b,a,o)\cdot\log\left(\sum_{q\in Q}\frac{b(q)\cdot p(q,a,o)}{p(b,a,o)}\cdot\frac{p(b,a,o)}{p(q,a,o)}\right)\\
&= H(b) + \sum_{o\in\mathsf{Comp}(b,a)} p(b,a,o)\cdot\log\left(\sum_{q\in Q} b(q)\right) = H(b)
\end{aligned}
$$

因此，等式（2）成立。由等式（1）和（2）可直接推出 $H(b)\ge\sum_{o\in\mathsf{Comp}(b,a)} p(b,a,o)\cdot H(\lambda(b,a,o))$. 因此，部分可观测马尔可夫决策过程 Λ 具有非增熵。 □

C.3 定理15的证明

下面我们介绍一些定义，以便从狄拉克保持型部分可观测马尔可夫决策过程定义一个多环境马尔可夫决策过程。请注意，我们将使用定义 40.<style>

定义 *41*。考虑一个狄拉克保持型部分可观测马尔可夫决策过程 $\Lambda=(Q,A,\delta,\Omega\ O)$。对于所有 $S\subseteq Q$，令 $T_S := \{(t_s)_{s\in S}\in(Q\cup\{\bot\})^S \mid \exists o\in O,\forall s\in S: t_s=\bot \text{ 或 } O(t_s)=o\}$。令 $\bot_S := (\bot_s)_{s\in S}\in T_S$。对于所有 $t\in T_S$ 和 $s\in S$，令 $t[s]\in Q\cup\{\bot\}$ 表示t中由 s索引的元素。我们还令 $O(t)\in\Omega$ 表示 t 中所有元素共有的观测（如果 $t=\bot_S$, $O(t)\in\Omega$，则任意定义）。

对于所有 $t\in T_s$、$a\in A$ 和 $o\in O$，我们令 $\mathsf{sc}(t,a,o)\in T_S$ 使得对于所有 $s\in S$ ：

$$
\mathsf{sc}(t,a,o)[s] := \begin{cases}\mathsf{sc}(t[s],a,o) & \text{if } o\in\mathsf{Comp}(t[s],a)\\ \bot & \text{otherwise}\end{cases}
$$

最后，对于所有 $\rho=\rho_0\cdots\rho_n\in(T_S)^+$ 和 $s\in S$，我们令 $\rho[s]:=\rho_0[s]\cdots\rho_n[s]\in(Q\cup\{\bot\})^+$。对 $\rho\in(T_S)^\omega$、$T_S\cdot(A\cdot T_S)^*$、$(T_S\cdot A)^\omega$ 分别与 $\rho[s]\in(Q\cup\{\bot\})^\omega$、$(Q\cup\{\bot\})\cdot(A\cdot(Q\cup\{\bot\}))^*$、$((Q\cup\{\bot\})\cdot A)^\omega$ 作类似处理。

设 $s\in S$。对于所有 $W\in$ 博雷尔(Q)，我们令 $W_s\in$ 博雷尔(T_S) 由 $\{\rho\in(T_S)^\omega\mid\rho[s]\in W\}$ 定义。我们还令 $\mathsf{Incomp}_s := \{\rho\in(T_S)^\omega\mid\exists i\in\mathbb{N},\ \rho_i[s]=\bot\}$。

现在，我们来定义一个由狄拉克保持型部分可观测马尔可夫决策过程导出的多环境马尔可夫决策过程。

定义 42. 考虑一个狄拉克保持型部分可观测马尔可夫决策过程 $\Lambda=(Q,A,\delta,\Omega\ O)$ 以及一个与观测兼容的 $S\subseteq Q$，即满足 $O(S)=\{o_S\}$，其中 $o_S\in\Omega$。我们令 $\Gamma(\Lambda)_S=(Q',A,E,(\delta_e)_{e\in E})$为如下定义的多环境马尔可夫决策过程：

- $Q' := T_S$;
- $E := S$;

• 对于所有 $t \in T_S$、 $a \in A$ 和 $s \in E = S$，我们令 $\delta_s(t,a) \in \mathcal{D}(Q') = \mathcal{D}(T_S)$ 满足：若 $t[s] = \bot$，则 $\delta_s(t,a)(\bot_S) := 1$；否则，对于所有 $o \in O$:

$$\delta_s(t,a)(\mathsf{sc}(t,a,o)) := \delta(t[s],a)(\mathsf{sc}(t[s],a,o)) = p[t[s],a,o]$$

我们令 $t_S^{\mathsf{init}} := (s)_{s\in S} \in T_S$ （因为 S 是观测相容的）。

现在考虑由函数 $f: Q \to \mathbb{N}$ 定义的观测兼容奇偶目标 $W \in$ 博雷尔(Q)，以及 $S \subseteq Q$。我们令 $f_S: T_S \to \mathbb{N}$，使得对于所有 $t \in T_S \setminus \{\bot_S\}$， $f_S(t) := f(q)$，对于任意满足 $O(q) = O(t)$ 的 $q \in Q$ （回忆一下， f 是观测相容的）； $f_S(\bot_S) := 1$。我们用 $W_S \in$ 博雷尔(T_S) 表示由 f_S 诱导的奇偶目标。

此定义满足以下性质。

引理 43. 考虑一个狄拉克保持型部分可观测马尔可夫决策过程 $\Lambda = (Q, A, \delta, \Omega\ O)$ 以及一个观测相容子集 $S \subseteq Q$。

1. 对所有观测兼容奇偶目标 $W \in$ 博雷尔(Q) 和 $s \in E = S$:

$$W_s = W_S \setminus \mathsf{Incomp}_s$$

2. 对所有 $\sigma \in \mathsf{Strat}(T_S, A)$，以及所有 $s \in S = E$，均有:

$$\mathbb{P}^{\sigma}_{t_S^{\mathsf{init}}}[\Gamma(\Lambda)_S[s], \mathsf{Incomp}_s] = 0$$

3. 对所有 $\kappa \in \mathsf{Strat}(\Omega\ A)$，存在 $\sigma(\kappa) \in \mathsf{Strat}(T_S, A)$，并且对所有 $\sigma \in \mathsf{Strat}(T_S, A)$，存在$\kappa(\sigma) \in \mathsf{Strat}(\Omega\ A)$，使得对所有博雷尔集 $W \in$ 博雷尔(Q)，以及 $s \in S = E$:

$$\mathbb{P}^{\kappa}_{s}[\Lambda, W] = \mathbb{P}^{\sigma(\kappa)}_{t_S^{\mathsf{init}}}[\Gamma(\Lambda)_S[s], W_s] \ \textit{and}\ \mathbb{P}^{\kappa(\sigma)}_{q}[\Lambda, W] = \mathbb{P}^{\sigma}_{t_S^{\mathsf{init}}}[\Gamma(\Lambda)_S[s], W_s]$$

证明。 1. 对所有 $\rho \in W_s$，均有 $\rho[s] \in W$。这意味着对所有 $i \in \mathbb{N}$，均有 $\rho_i[s] \neq \bot$，因此 $\rho \notin \mathsf{Incomp}_s$。事实上， $W_s \cap \mathsf{Incomp}_s = \emptyset$。

考虑任意 $\rho \in (T_S)^\omega \setminus \mathsf{Incomp}_s$。对所有 $i \in \mathbb{N}$，均有 $\rho_i[s] \neq \bot$；因此 $O(\rho_i) = O(\rho_i[s])$ 且 $f_S(\rho_i) = f(\rho_i[s])$。因此， $\rho \in W_S$ 当且仅当 $\rho[s] \in W$ 当且仅当 $\rho \in W_s$。于是: $W_s = W_S \setminus \mathsf{Incomp}_s$。

2. 令 $\sigma \in \mathsf{Strat}(T_S, A)$ 且 $s \in S = E$。我们对 $\rho \in T_S \cdot (A \cdot T_S)^*$ 进行归纳法，证明若 $\mathbb{P}^{\sigma}_{t_S^{\mathsf{init}}}[\Gamma(\Lambda)_S[s], \rho] > 0$，则 $\mathsf{last}(\rho)[s] \neq \bot$。对于 $\rho \in T_S$，这显然成立。现在假设该结论对某个 $\rho \in T_S \cdot (A \cdot T_S)^*$ 成立，并令 $(a,t) \in A \times T_S$ 满足 $\mathbb{P}^{\sigma}_{t_S^{\mathsf{init}}}[\Gamma(\Lambda)_S[s], \rho \cdot (a,t)] > 0$，其中:

$$\mathbb{P}^{\sigma}_{t_S^{\mathsf{init}}}[\Gamma(\Lambda)_S[s], \rho \cdot (a,t)] = \mathbb{P}^{\sigma}_{t_S^{\mathsf{init}}}[\Gamma(\Lambda)_S[s], \rho] \cdot \sigma(\rho)(a) \cdot \delta_s(\mathsf{last}(\rho), a)(t)$$

我们有 $\mathbb{P}^{\sigma}_{t_S^{\mathsf{init}}}[\Gamma(\Lambda)_S[s], \rho] > 0$，因此 $\mathsf{last}(\rho)[s] \neq \bot$。此外， $\delta_s(\mathsf{last}(\rho)\ a)(t) > 0$ 意味着存在 $o \in O$ 使得 $t = \mathsf{sc}(\mathsf{last}(\rho)\ a, o)$。在这种情况下，我们有 $\delta_s(\mathsf{last}(\rho)\ a)(t) = p[\mathsf{last}(\rho)[s], a, o]$，如果 $o \notin \mathsf{Comp}(\mathsf{last}(\rho)[s], a)$，其值为 0。因此， $o \in \mathsf{Comp}(\mathsf{last}(\rho)[s], a)$，且 $t[s] = \mathsf{sc}(\mathsf{last}(\rho)\ a, o)[s] = \mathsf{sc}(\mathsf{last}(\rho)[s], a, o) \in Q$。因此，归纳性质对 $\rho \cdot (a,t)$ 也成立。事实上，它对所有 $\rho \in T_S \cdot (A \cdot T_S)^*$ 都成立。因此，我们有:

$$\mathbb{P}^{\sigma}_{t_S^{\mathsf{init}}}[\Gamma(\Lambda)_S[s], \mathsf{Incomp}_s] = \sum_{\substack{\rho \in T_S \cdot (A \cdot T_S)^* \\ \mathsf{last}(\rho) = \bot}} \mathbb{P}^{\sigma}_{t_S^{\mathsf{init}}}[\Gamma(\Lambda)_S[s], \rho] = 0$$

3. 令 $o_S \in O$ 满足：对所有 $s \in S$， $O(s) = o_S$。

对于所有 $s \in S = E$，我们通过归纳法定义函数 $\mathsf{sc}^*_S\colon \Omega \cdot (A \times \Omega)^* \to T_S \cdot (A \cdot T_S)^*$ 如下：对于所有 $\rho \in \Omega \cdot (A \times \Omega)^* \setminus o_S \cdot (A \times \Omega)^*$，$\mathsf{sc}^*_S(\rho) := \bot_S$；此外，$\mathsf{sc}^*_S(o_S) := t^{\mathsf{init}}_S$，并且，对于所有 $\rho \in o_S \cdot (A \times \Omega)^*$ 和 $(a, o) \in A \times \Omega$：

$$\mathsf{sc}^*_S(\rho \cdot (a, o)) := \mathsf{sc}^*_S(\rho) \cdot (a, \mathsf{sc}(\mathsf{last}(\mathsf{sc}^*_S(\rho)), a, o)) \in T_S \cdot (A \cdot T_S)^*$$

对于所有 $s \in S$，我们令 $\mathsf{Compatible}_s(A, \Omega) := \{\rho \in o_s \cdot (A \times \Omega)^* \mid \mathsf{last}(\mathsf{sc}^*_S(\rho)[s]) \neq \bot\}$。我们令 $\mathsf{Compatible}_S(A, \Omega) := \cup_{s \in S} \mathsf{Compatible}_s(A, \Omega)$。对于所有 $\pi \in Q \cdot (A \cdot Q)^*$，我们令 $O(\pi) \in \Omega \cdot (A \cdot \Omega)^*$ 为与 π 关联的动作和观测序列。特别地，对于所有 $s \in S$ 和 $\rho \in \mathsf{Compatible}_s(A, \Omega)$，我们有 $O(\mathsf{sc}^*_S(\rho)[s]) = \rho$。

我们给出若干中间结果。

(a) 设 $s \in S$ 和 $\kappa \in \mathsf{Strat}(\Omega\ A)$。首先，我们通过对 $\pi \in s \cdot (A \cdot Q)^*$ 作归纳法，证明若 $\mathbb{P}^\kappa_s[\Lambda, \pi] > 0$，则 $\mathsf{sc}^*_S(O(\pi))[s] = \pi$。这对 $\pi = s$ 成立。现在假设它对某个 $\pi \in s \cdot (A \cdot Q)^*$ 成立，并令 $(a, q) \in A \times Q$。假设 $\mathbb{P}^\kappa_s[\Lambda, \pi \cdot (a, q)] > 0$。我们有：

$$\mathbb{P}^\kappa_s[\Lambda, \pi \cdot (a, q)] = \mathbb{P}^\kappa_s[\Lambda, \pi] \cdot \kappa(O(\pi))(a) \cdot \delta(\mathsf{last}(\pi), a)(q) > 0$$

因此，$\delta(\mathsf{last}(\pi)\ a)(q) > 0$，从而，对于 $o := O(q) \in \Omega$，$o \in \mathsf{Comp}(\mathsf{last}(\pi)\ a)$ 且 $\mathsf{sc}(\mathsf{last}(\pi)\ a, o) = q$。因此：

$$\begin{aligned}\mathsf{sc}^*_S(O(\pi \cdot (a, q)))[s] &= \mathsf{sc}^*_S(O(\pi) \cdot (a, o))[s] = (\mathsf{sc}^*_S(O(\pi)) \cdot (a, \mathsf{sc}(\mathsf{last}(\mathsf{sc}^*_S(O(\pi))), a, o)))[s] \\ &= \pi \cdot (a, \mathsf{sc}(\mathsf{last}(\mathsf{sc}^*_S(O(\pi))), a, o)[s]) = \pi \cdot (a, \mathsf{sc}(\mathsf{last}(\mathsf{sc}^*_S(O(\pi))[s]), a, o)) \\ &= \pi \cdot (a, \mathsf{sc}(\mathsf{last}(\pi), a, o)) = \pi \cdot (a, q)\end{aligned}$$

因此，对所有满足 $\mathbb{P}^\kappa_s[\Lambda, \pi] > 0$ 的 $\pi \in s \cdot (A \cdot Q)^*$，等式均成立。

(b) 令 $s \in S$ 和 $\kappa \in \mathsf{Strat}(\Omega\ A)$。我们证明，对于所有满足 $O(\pi) \notin \mathsf{Compatible}_s(A, \Omega)$ 的 $\pi \in s \cdot (A \cdot Q)^*$，都有 $\mathbb{P}^\kappa_s[\Lambda, \pi] = 0$。由于 $O(s) = o_S \in \mathsf{Compatible}_s(A, \Omega)$，这对 $\pi = s$ 成立。现在考虑某个 $\pi \in s \cdot (A \cdot Q)^*$ 和 $(a, q) \in A \times Q$。假设 $O(\pi \cdot (a, q)) \notin \mathsf{Compatible}_s(A, \Omega)$。我们有：

$$\mathbb{P}^\kappa_s[\Lambda, \pi \cdot (a, q)] = \mathbb{P}^\kappa_s[\Lambda, \pi] \cdot \kappa(O(\pi))(a) \cdot \delta(\mathsf{last}(\pi), a)(q)$$

若 $\mathbb{P}^\kappa_s[\Lambda, \pi] = 0$，则 $\mathbb{P}^\kappa_s[\Lambda, \pi \cdot (a, q)] = 0$。因此，假设 $\mathbb{P}^\kappa_s[\Lambda, \pi] > 0$。根据 (a) 项，这意味着 $\mathsf{last}(\mathsf{sc}^*_S(O(\pi))[s]) = \mathsf{last}(\pi) \in Q$。此外，由于 $O(\pi \cdot (a, q)) \notin \mathsf{Compatible}_s(A, \Omega)$，我们有 $\mathsf{sc}(\mathsf{last}(\mathsf{sc}^*_S(O(\pi))[s])\ a, o) = \mathsf{last}(\mathsf{sc}^*_S(O(\pi \cdot (a, q)))[s]) = \bot$。也就是说，$o \notin \mathsf{Comp}(\mathsf{last}(\mathsf{sc}^*_S(O(\pi))[s])\ a) = \mathsf{Comp}(\mathsf{last}(\pi)\ a)$。由于 $O(q) = o$，这意味着 $\delta(\mathsf{last}(\pi)\ a)(q) = 0$，因此 $\mathbb{P}^\kappa_s[\Lambda, \pi \cdot (a, q)] = 0$。

(c) 现在考虑任意策略 $\sigma \in \mathsf{Strat}(T_S, A)$。令 $s \in S$ 和 $\pi \in s \cdot (A \cdot Q)^*$。根据定义，我们有：

$$\mathsf{cyl}(\pi)_s = \{\rho \in (T_S \cdot A)^\omega \mid \rho[s] \in \mathsf{cyl}(\pi)\} = \{\rho \in (T_S \cdot A)^\omega \mid \rho[s] \in \pi \cdot (A \cdot Q)^\omega\}$$

我们将 $\mathsf{cyl}(\pi)^\perp_s \in$ 博雷尔$(T_S \cdot A)$ 定义如下：

$$\mathsf{cyl}(\pi)^\perp_s = \{\rho \in (T_S \cdot A)^\omega \mid \rho[s] \in \pi \cdot (A \cdot (Q \cup \{\bot\}))^\omega\}$$

我们有：

$$\mathsf{cyl}(\pi)_s \subseteq \mathsf{cyl}(\pi)^\perp_s \subseteq \mathsf{cyl}(\pi)_s \cup \mathsf{Incomp}_s$$

因此，根据条目 (b)，我们有：

$$\mathbb{P}^\sigma_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \mathsf{cyl}(\pi)_s] = \mathbb{P}^\sigma_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \mathsf{cyl}(\pi)^\perp_s]$$

此外，我们有：

$$\mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \mathsf{cyl}(\pi)^{\perp}_s] = \sum_{\substack{\rho \in T_S \cdot (A \cdot T_S)^* \\ \rho[s] = \pi}} \mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \rho]$$

我们通过对 $\pi \in s \cdot (A \cdot Q)^*$ 使用归纳法来证明：对于所有满足 $\rho[s] = \pi$ 的 $\rho \in T_S \cdot (A \cdot T_S)^*$，若 $\mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \rho] > 0$，则 $\rho = \mathsf{sc}^*_S(O(\pi))$。若$\pi = s$，这一点显然成立。现在，假设它对某个 $\pi \in s \cdot (A \cdot Q)^*$ 成立，并令 $(a, q) \in A \times Q$。考虑任意满足 $\rho \cdot (a, t)[s] = \pi \cdot (a, q)$ 的 $\rho \cdot (a, t) \in T_S \cdot (A \cdot T_S)^+$，并假设 $\mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \rho \cdot (a, t)] > 0$。我们有：

$$\mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \rho \cdot (a, t)] = \mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \rho] \cdot \sigma(\rho)(a) \cdot \delta_s(\mathsf{last}(\rho), a)(t) > 0$$

由于 $\mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \rho] > 0$，根据我们的归纳假设，我们有 $\rho = \mathsf{sc}^*_S(O(\pi))$。此外，由于 $\delta_s(\mathsf{last}(\rho)\ a)(t) > 0$，存在 $o \in \Omega$ 使得 $t = \mathsf{sc}(\mathsf{last}(\rho)\ a, o)$，其中 $q = t[s] = \mathsf{sc}(\mathsf{last}(\rho)[s], a, o)$，从而 $O(q) = O(t[s]) = o$。因此，我们有：

$$\begin{aligned}\mathsf{sc}^*_S(O(\pi \cdot (a, q))) &= \mathsf{sc}^*_S(O(\pi) \cdot (a, o)) = \mathsf{sc}^*_S(O(\pi)) \cdot (a, \mathsf{sc}(\mathsf{last}(\mathsf{sc}^*_S(O(\pi))), a, o)) \\ &= \rho \cdot (a, \mathsf{sc}(\mathsf{last}(\rho), a, o)) = \rho \cdot (a, t)\end{aligned}$$

事实上，我们的归纳假设对所有 $\pi \in s \cdot (A \cdot Q)^*$ 都成立。由此可得，对所有$\pi \in s \cdot (A \cdot Q)^*$:

$$\mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \mathsf{cyl}(\pi)_s] = \begin{cases} 0 & \text{if } \mathsf{sc}^*_S(O(\pi))[s] \neq \pi \\ \mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \mathsf{sc}^*_S(O(\pi))] & \text{otherwise} \end{cases}$$

(d) 现在，考虑两个策略 $\kappa \in \mathsf{Strat}(\Omega\ A)$ 和 $\sigma \in \mathsf{Strat}(T_S, A)$，并假设它们是 相似 的，即对于所有 $\rho \in \mathsf{Compatible}_S(A,\Omega)$，都有 $\kappa(\rho) = \sigma(\mathsf{sc}^*_S(\rho)) \in \mathcal{D}(A)$。我们对 $\pi \in s \cdot (A \cdot Q)^*$ 进行归纳，证明：

$$\mathbb{P}^{\kappa}_s[\Lambda, \pi] = \mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \mathsf{cyl}(\pi)_s]$$

这显然对 $\pi = s$ 成立。现在假设它对某个 $\pi \in s \cdot (A \cdot Q)^*$ 成立，并令 $(a, q) \in A \times Q$。令 $o := O(q) \in \Omega$。我们有：

$$\mathsf{sc}^*_S(O(\pi \cdot (a, q)) = \mathsf{sc}^*_S(O(\pi) \cdot (a, o)) = \mathsf{sc}^*_S(O(\pi)) \cdot (a, \mathsf{sc}(\mathsf{last}(\mathsf{sc}^*_S(O(\pi))), a, o))$$

如果 $\mathbb{P}^{\kappa}_s[\Lambda, \pi] = 0$，则 $\mathbb{P}^{\kappa}_s[\Lambda, \pi \cdot (a, q)] = 0$，$\mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \mathsf{sc}^*_S(O(\pi))] = 0$从而 $\mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \mathsf{sc}^*_S(O(\pi \cdot (a, q)))] = 0$。现假设 $\mathbb{P}^{\kappa}_s[\Lambda, \pi] > 0$。由第(b)项，这蕴含 $O(\pi) \in$ 相容$_s(A,\Omega)$；由第(a)项，令 $\rho := \mathsf{sc}^*_S(O(\pi))$，这蕴含 $\rho[s] = \pi$。现在，如果 $\delta(\mathsf{last}(\pi)\ a)(q) = \delta(\mathsf{last}(\rho)[s], a)(q) = 0$，则 $\mathbb{P}^{\kappa}_s[\Lambda, \pi \cdot (a, q)] = 0$并且要么 $o = O(q) \notin \mathsf{Comp}(\mathsf{last}(\rho)[s], a)$，要么 $o \in \mathsf{Comp}(\mathsf{last}(\rho)[s], a)$ 并且 $\mathsf{sc}(\mathsf{last}(\rho)[s], a, o) \neq q$。在这两种情况下，我们有 $\mathsf{last}(\mathsf{sc}^*_S(O(\pi \cdot (a, q))) \neq q$，从而 $\mathsf{sc}^*_S(O(\pi \cdot (a, q))[s] \neq \pi \cdot (a, q)$，这蕴含 $\mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \mathsf{cyl}(\pi \cdot (a, q))_s] = 0$。现假设 $\delta(\mathsf{last}(\pi)\ a)(q) > 0$，因此 $o \in \mathsf{Comp}(\mathsf{last}(\pi)\ a)$ 并且$q = \mathsf{sc}(\mathsf{last}(\pi)\ a, o) = \mathsf{sc}(\mathsf{last}(\rho)[s], a, o)$，这蕴含 $\mathsf{sc}^*_S(O(\pi \cdot (a, q))[s] = \pi \cdot (a, q)$。于是，我们有：

$$\begin{aligned}\mathbb{P}^{\kappa}_s[\Lambda, \pi \cdot (a, q)] &= \mathbb{P}^{\kappa}_s[\Lambda, \pi] \cdot \kappa(O(\pi))(a) \cdot \delta(\mathsf{last}(\pi), a)(q) \\ &= \mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \rho] \cdot \sigma(\rho)(a) \cdot \delta(\mathsf{last}(\rho)[s], a)(q) \\ &= \mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \rho] \cdot \sigma(\rho)(a) \cdot \delta_s(\mathsf{last}(\rho), a)(\mathsf{sc}(\mathsf{last}(\rho), a, o)) \\ &= \mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \rho \cdot (a, \mathsf{sc}(\mathsf{last}(\rho), a, o))] \\ &= \mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \mathsf{sc}^*_S(O(\pi \cdot (a, q)))] \\ &= \mathbb{P}^{\sigma}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s], \mathsf{cyl}(\pi \cdot (a, q))_s]\end{aligned}$$

因此该性质对 $\pi\cdot(a,q)\in s\cdot(A\cdot Q)^+$ 成立。事实上，它对所有 $\pi\in s\cdot(A\cdot Q)$ 都成立 *.

(e) 设 $s\in S$。考虑两个相似策略 $\kappa\in\mathsf{Strat}(\Omega\ A)$ 和 $\sigma\in\mathsf{Strat}(T_S,A)$。令 $\mathbb{P}_\kappa\colon\mathsf{Borel}(Q)\to[0,1]$，使得对于所有 $W\in\mathsf{Borel}(Q)$，都有 $\mathbb{P}_\kappa(W):=\mathbb{P}^\kappa_s[\Lambda,W]\in[0,1]$。我们还令 $\mathbb{P}_\sigma\colon\mathsf{Borel}(Q)\to[0,1]$，使得对于所有 $W\in\mathsf{Borel}(Q)$，都有 $\mathbb{P}_\sigma(W):=\mathbb{P}^\sigma_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s],W_s]$。显然，$\mathbb{P}_\kappa$ 是一个概率测度。对于 $\mathbb{P}_\sigma$ 也同样如此，因为 $\mathbb{P}_\sigma[\emptyset]=0$；$\mathbb{P}_\sigma[Q^\omega]=\mathbb{P}^\sigma_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s],(T_S\cdot A)^\omega\setminus W_s]=1$（由条目 (2)）；并且对于所有两两不相交的 $(W^n)_{n\in\mathbb{N}}\in\mathsf{Borel}(Q)^{\mathbb{N}}$，都有 $(\cup_{n\in\mathbb{N}}W^n)_s=\cup_{n\in\mathbb{N}}W^n_s$，因此 $\mathbb{P}_\sigma[\cup_{n\in\mathbb{N}}W_n]=\sum_{n\in\mathbb{N}}\mathbb{P}_\sigma[W_n]$。对于所有 $\rho=s\cdot q_1\cdots q_n\in s\cdot Q^*$，令 $\mathsf{Proj}(\rho):=\{s\cdot(a_1,q_1)\cdots(a_n,q_n)\in Q\cdot(A\cdot Q)^*\mid a_1,\ldots,a_n\in A\}$。于是，对于所有$\rho\in s\cdot Q^*$，由 (d) 可得：

$$\begin{aligned}\mathbb{P}_\kappa[\mathsf{cyl}(\rho)]=\mathbb{P}^\kappa_s[\Lambda,\mathsf{cyl}(\rho)]&=\sum_{\pi\in\mathsf{Proj}(\rho)}\mathbb{P}^\kappa_s[\Lambda,\mathsf{cyl}(\pi)]=\sum_{\pi\in\mathsf{Proj}(\rho)}\mathbb{P}^\kappa_s[\Lambda,\pi]\\&=\sum_{\pi\in\mathsf{Proj}(\rho)}\mathbb{P}^\sigma_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s],\mathsf{cyl}(\pi)_s]=\mathbb{P}^\sigma_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s],\mathsf{cyl}(\rho)_s]=\mathbb{P}_\sigma[\mathsf{cyl}(\rho)]\end{aligned}$$

此外，对于所有 $\rho\in Q^+\setminus s\cdot Q^*$，我们有 $\mathbb{P}_\kappa[\mathsf{cyl}(\rho)]=0=\mathbb{P}_\sigma[\mathsf{cyl}(\rho)]$。因此，两个概率测度 $\mathbb{P}_\kappa$ 和 $\mathbb{P}_\sigma$ 在所有柱集上相同，故由卡拉泰奥多里扩张定理可知，它们相等：$\mathbb{P}_\kappa=\mathbb{P}_\sigma$，也就是说，对于所有 $W\in$博雷尔(Q)：

$$\mathbb{P}^\kappa_s[\Lambda,W]=\mathbb{P}^\sigma_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s],W_s]$$

(f) 对于所有 $\sigma\in\mathsf{Strat}(T_S,\ A)$，令 $\kappa(\sigma)\in\mathsf{Strat}(\Omega\ A)$ 满足：对于所有 $\rho\in\Omega\cdot(A\cdot\Omega)^*$：$\kappa(\sigma)(\rho):=\sigma(\mathsf{sc}^*_S(\rho))$。那么，策略 $\kappa(\sigma)$ 和 σ 相似，因此由前一项可知，对于所有 $W\in$博雷尔(Q)，我们有：

$$\mathbb{P}^{\kappa(\sigma)}_s[\Lambda,W]=\mathbb{P}^\sigma_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s],W_s]$$

反过来，考虑任意 $\kappa\in\mathsf{Strat}(\Omega\ A)$。我们的目标是定义 $\sigma(\kappa)\in\mathsf{Strat}(T_S,A)$，使得策略 $\sigma(\kappa)$ 和 κ 相似。我们来说明函数 $\mathsf{sc}^*_S\colon\Omega\cdot(A\times\Omega)^*\to T_S\cdot(A\cdot T_S)^*$ 在 $\mathsf{Compatibles}(A,\Omega)$ 上是单射。事实上，考虑任意满足 $\mathsf{sc}^*_S(\rho)=\mathsf{sc}^*_S(\rho')$ 的 $\rho,\rho'\in\Omega\cdot(A\times\Omega)^*\cap\mathsf{Compatibles}(A,\Omega)$。存在某个$s\in S$，使得 $\rho,\rho'\in\Omega\cdot(A\times\Omega)^*\cap\mathsf{Compatible}_s(A,\Omega)$。因此，我们有 $\rho=O(\mathsf{sc}^*_S(\rho)[s])=O(\mathsf{sc}^*_S(\rho')[s])=\rho'$。于是，对于所有 $\rho\in T_S\cdot(A\cdot T_S)^*$，至多存在一个满足 $\rho=\mathsf{sc}^*_S(\pi_\rho)$ 的 $\pi_\rho\in\mathsf{Compatibles}(A,\Omega)$；若存在，我们令 $\sigma(\kappa)(\rho):=\kappa(\pi_\rho)$；否则，$\sigma(\kappa)(\rho)$ 任意定义。这样，策略 κ和 $\sigma(\kappa)$ 相似，因此由前一项可知，对于所有 $W\in$博雷尔(Q)，我们有：

$$\mathbb{P}^\kappa_s[\Lambda,W]=\mathbb{P}^{\sigma(\kappa)}_{t^{\mathsf{init}}_S}[\Gamma(\Lambda)_S[s],W_s]$$

这就证明了该引理。

□

现在我们可以继续证明定理15。

证明。 显然，给定狄拉克保持型部分可观测马尔可夫决策过程 $\Lambda=(Q,\ A,\ \delta,\ \Omega\ O)$、一个信念 $b\in\mathcal{D}(Q)$ 和一个与观测兼容的奇偶目标 $W\in$博雷尔(Q)，我们可以在指数时间内构造出多环境马尔可夫决策过程 $\Gamma(\Lambda)_S=(Q',\ A,\ E,\ (\delta_e)_{e\in E})$，其中 $S:=$支持集(b)[5]，奇偶目标为 W_S，且$|Q'|=|T_S|\le(n+1)^n$，对 $n:=|Q|$成立。我们将 b 本身用作 $\Gamma(\Lambda)_S$ 中的初始环境信念。

[5]不失一般性，我们假设 $\mathsf{Supp}(b)$ 是观测相容的，因为在部分可观测马尔可夫决策过程（POMDP）中首次进行对局时，我们会获得初始状态的观测。

此外，根据引理 43，对于所有 $\sigma \in \mathsf{Strat}(T_S, A)$，对于所有 $s \in S$，均有：

$$\mathbb{P}^{\sigma}_{t_S^{\mathsf{init}}}[\Gamma(\Lambda)_S[s], W_s] = \mathbb{P}^{\sigma}_{t_S^{\mathsf{init}}}[\Gamma(\Lambda)_S[s], W_S \setminus \mathsf{Incomp}_s] = \mathbb{P}^{\sigma}_{t_S^{\mathsf{init}}}[\Gamma(\Lambda)_S[s], W_S]$$

此外，根据43引理，对于所有 $\kappa \in \mathsf{Strat}(\Omega\ A)$，存在 $\sigma(\kappa) \in \mathsf{Strat}(T_S, A)$，使得：

$$\sum_{s \in S} b(s) \cdot \mathbb{P}^{\kappa}_{s}[\Lambda, W] = \sum_{s \in S} b(s) \cdot \mathbb{P}^{\sigma(\kappa)}_{t_S^{\mathsf{init}}}[\Gamma(\Lambda)_S[s], W_S] \leq \mathsf{val}^{\mathsf{pr}}_{t_S^{\mathsf{init}}}(\Gamma(\Lambda)_S, b, W_S)$$

此外，对于所有 $\sigma \in \mathsf{Strat}(T_S, A)$，存在 $\kappa(\sigma) \in \mathsf{Strat}(\Omega\ A)$，使得：

$$\sum_{s \in S} b(s) \cdot \mathbb{P}^{\sigma}_{t_S^{\mathsf{init}}}[\Gamma(\Lambda)_S[s], W_S] = \sum_{s \in S} b(s) \cdot \mathbb{P}^{\kappa(\sigma)}_{s}[\Lambda, W] \leq \mathsf{val}(\Lambda, b, W)$$

事实上，我们有：

$$\mathsf{val}(\Lambda, b, W) = \mathsf{val}^{\mathsf{pr}}_{t_S^{\mathsf{init}}}(\Gamma(\Lambda)_S, b, W_S)$$

定理得证。 □