arXiv:2301.11296v2 [cs.LO] 6 Dec 2024

多环境马尔可夫决策过程中的鲁棒几乎必然可达性

Marck van der Vegt、Nils Jansen和 Sebastian Junges

荷兰奈梅亨拉德堡德大学{marck.vandervegt, nils.jansen, sebastian.junges}@ru.nl

摘要。多环境马尔可夫决策过程（MEMDP）刻画了有限个共享状态但转移动态不同的马尔可夫决策过程集合。这些模型构成部分可观测马尔可夫决策过程（POMDP）的真子类。我们研究在多环境马尔可夫决策过程中综合能够鲁棒地满足几乎必然可达性质的策略，也就是说，*one* 能够在 *for all* 环境中满足某一性质的策略。对于部分可观测马尔可夫决策过程，判定鲁棒策略是否存在是一个 EXPTIME 完全问题。我们证明，对于多环境马尔可夫决策过程，该问题是 PSPACE 完全的，而一般而言，这些策略需要指数级记忆。我们利用理论结果开发并实现了一种算法，在为各种基准测试综合鲁棒策略方面展现出良好的效果。

1 引言

马尔可夫决策过程（MDP）是在不确定性下对序贯决策进行建模的标准形式化方法。一个典型目标是找到满足时序逻辑规范 [5] 的策略。诸如Storm [22] 和 Prism [30] 之类的概率模型检测器能够高效地计算此类策略。然而，一个问题是其对环境中潜在扰动的鲁棒性。MDP 无法刻画关于环境形态的此类不确定性。

多环境马尔可夫决策过程（MEMDP） [36,14] 包含一组定义在同一状态空间上的马尔可夫决策过程，称为环境。MEMDP 中的目标是找到一个在所有环境中均满足给定规范的单一策略。例如，对于系统动态未知的马尔可夫决策过程，MEMDP 是一种自然模型，其中多位领域专家提供了他们对动态转移的解释 [11]。这些不同的马尔可夫决策过程共同构成一个 MEMDP。MEMDP 也出现在其他领域：在安全领域，猜测一个（静态）密码便是一个自然的例子。在机器人学中，MEMDP 可刻画某些静态障碍物位置未知的情况。可以将 MEMDP 解释为马尔可夫决策过程的（不交）并集，其中智能体仅具有部分观测；即，每个 MEMDP 都可以转化为一个规模线性增大的部分可观测马尔可夫决策过程（POMDP） [27]。事实上，部分可观测马尔可夫决策过程的一些著名例实际上是 MEMDP，例如 RockSample [39] 和 Hallway [31]。求解部分可观测马尔可夫决策过程众所周知地困难 [32], ，因此，研究其自然子类是值得的。

我们考虑几乎必然成立的规约，即到达一组目标状态的概率必须为 1。在马尔可夫决策过程中，只需考虑无记忆

策略。此类策略可以通过图搜索 [5]高效地构造。对于多环境马尔可夫决策过程，我们考虑以下问题：

计算一个能在所有环境中几乎必然到达目标的策略。

这样的策略能针对一组马尔可夫决策过程稳健地满足几乎必然规约。我们的方法受到部分可观测马尔可夫决策过程相关工作的启发，构造了一个信念观测马尔可夫决策过程（BOMDP）[16]，用于跟踪马尔可夫决策过程的状态以及潜在环境上的信念（支撑集）。我们证明，在 BOMDP 中满足几乎必然性质的策略，在多环境马尔可夫决策过程中也满足该性质。

尽管 BOMDP 的规模呈指数级增长，远大于多环境马尔可夫决策过程，但我们利用其特有结构，设计了一种 PSPACE 算法来判定是否存在这样的稳健策略。该算法的核心是递归构造 BOMDP 的一个片段，并将其限制在信念支持集固定的情形下。这种方法之所以可行，是因为多环境马尔可夫决策过程中的信念具有单调性：一旦我们确定自己不处于某个特定环境中，就不会失去这一认知。这与部分可观测马尔可夫决策过程形成对比，在后者中，信念支持集并不具有单调性。这一区别至关重要：判定部分可观测马尔可夫决策过程中的几乎必然可达性是 EXPTIME 完全的 [37,19]。相比之下，判定多环境马尔可夫决策过程中是否存在满足几乎必然可达性的策略，确实是 PSPACE 完全的。我们通过从真量化布尔公式问题进行归约来证明其困难性。最后，我们无法期望通过此类算法提取出策略，因为多环境马尔可夫决策过程中的最小策略可能需要关于环境数量的指数级记忆。

PSPACE 算法本身会重复计算许多结果。出于实际考虑，我们设计了一种算法，能够迭代地探索 BOMDP 的部分内容。该算法还利用 MEMDP 结构，对存在获胜策略的状态集合进行泛化，并推导出高效的启发式方法来引导探索。结合这些要素，我们在模型检查器 Storm的基础上实现了一个高效且具有竞争力的原型。

相关工作。我们将相关工作分为三个方面。

MEMDP。已有研究探讨了环境数恰为两个时 MEMDP 的几乎必然可达性 [36]。我们将这些结果推广到任意多个环境。这并非易事：对于两个环境，该判定问题可以通过一个多项式时间例程解决 [36], ，而我们证明，对于任意数量的环境，该问题是 PSPACE 完全的。MEMDP 以及与之密切相关的模型，例如隐藏模型马尔可夫决策过程、隐藏参数马尔可夫决策过程、多模型马尔可夫决策过程和并发马尔可夫决策过程 [11,2,40,10] ，已被用于研究定量性质[1]。典型做法是针对部分可观测马尔可夫决策过程中的不可判定问题，考虑近似算法 [14] ，或调整强化学习算法 [3,28]。这些近似方法不适用于几乎必然性质。

部分可观测马尔可夫决策过程。可以构建一个底层的、可能是无限的信念马尔可夫决策过程 [27] ，它与该部分可观测马尔可夫决策过程相对应——使用模型检查器 [35,7,8] 来验证这个马尔可夫决策过程

[1] 隐藏参数马尔可夫决策过程与多环境马尔可夫决策过程不同，前者假设存在一个定义在马尔可夫决策过程上的先验分布。不过，对于几乎必然性质而言，这一区别无关紧要。

可以回答关于多环境马尔可夫决策过程的问题。对于部分可观测马尔可夫决策过程，几乎必然可达性可在指数时间 [37,19] 内判定，其方法是采用一种与我们的方法类似的构造。除了几乎必然可达性之外，大多数定性性质都不可判定 [4,15]。有两种专用算法将搜索限制在内存需求较小的策略范围内，并采用基于 SAT 的方法 [12,26] 处理这一 NP 难问题 [19]，这些算法已在 STORM中实现。我们将它们用作基线方法。

鲁棒模型。MEMDP 的高层表示在结构上与特征马尔可夫决策过程相似，后者表示一组马尔可夫决策过程。所提出的技术称为基于族的模型检测，可为该族中的每个马尔可夫决策过程计算策略，而我们的目标是为所有马尔可夫决策过程找到一个策略。区间马尔可夫决策过程和随机博弈不允许状态之间存在依赖关系，因此无法对诸如不同障碍物位置之类的特征进行建模。参数化马尔可夫决策过程假设不确定性可控，并且不考虑策略的鲁棒性。

贡献。我们确立了判定多环境马尔可夫决策过程中的几乎必然可达性问题具有 PSPACE 完全性，并表明策略的规模可能呈指数型增长。我们的迭代算法是首个专门针对多环境马尔可夫决策过程中几乎必然可达性的算法，它构建 BOMDP 的片段。一项实证评估表明，该迭代算法优于专用于部分可观测马尔可夫决策过程的方法。

2 问题陈述

本节将介绍一些背景知识，并对问题陈述进行形式化。

对于集合 X，Dist(X) 表示定义在 X 上的概率分布集合。对于给定的分布 $d \in \text{Dist}(X)$，其支撑集记为 Supp(d)。对于有限集合 X，令 unif(X) 表示均匀分布。dirac(x) 表示定义在 $x \in X$ 上的狄拉克分布。对于函数和分布，我们采用简写记法：$f = [x \mapsto a, y \mapsto b]$ 表示 $f(x) = a$ 以及 $f(y) = b$。我们用 $\mathcal{P}(X)$ 表示 X 的幂集。对于 $n \in \mathbb{N}$，我们记作 $[n] = \{i \in \mathbb{N} \mid 1 \leq i \leq n\}$。

定义 1（马尔可夫决策过程）。马尔可夫决策过程是一个元组 $\mathcal{M} = \langle S, A, \iota_{init}, p \rangle$，其中 S 是有限状态集合，A 是有限动作集合，$\iota_{init} \in \text{Dist}(S)$ 是初始状态分布，p: $S \times A \to \text{Dist}(S)$ 是转移函数。

转移函数是全定义的，也就是说，为方便记号，马尔可夫决策过程均为输入全定义。这一要求不会影响我们结果的普遍性。马尔可夫决策过程的一条路径是一个序列 $\pi = s_0 a_0 s_1 a_1 \dots s_n$，满足 $\iota_{\text{init}}(s_0) > 0$ 且 $p(s_i, a_i)(s_{i+1}) > 0$，对所有 $0 \leq i < n$ 均成立。π 的最后一个状态是 last (π) $= s_n$。所有有限路径的集合记为 PATH，而 PATH(S') 表示从 $S' \subseteq S$ 中的某个状态开始的路径。来自 S' 的可达状态集合为 Reachable(S')。若 S' =Supp(ι_{init})，则我们简称它们为可达状态。从分布 $d \in$ Dist(S) 出发、限制在可达状态上的马尔可夫决策过程为 ReachFragment($\mathcal{M}, d$)，其中 d 是新的初始分布。若状态 $s \in S$ 满足 Reachable($\{s\}$) $= \{s\}$，则该状态为吸收态。若每个状态都是吸收态，或无法从其后继状态到达，则该马尔可夫决策过程是无环的。

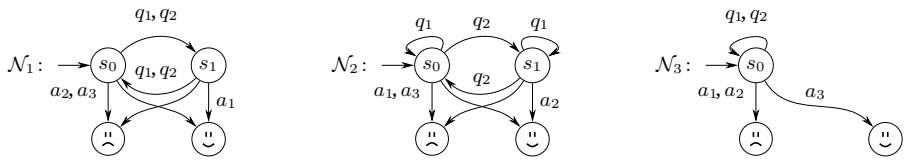

图 1：多环境马尔可夫决策过程示例

动作选择由策略 σ: $\text{PATH} \to \text{Dist}(A)$ 决定，该策略将路径映射为动作上的分布。形式为 σ: $S \to \text{Dist}(A)$ 的策略称为无记忆策略；若满足 σ: $\text{PATH} \to A$，则称为确定性策略；当 σ: $S \to A$ 时，则称为无记忆确定性策略。对于马尔可夫决策过程 $\mathcal{M}$，我们将从状态 s 出发，在策略 σ 下到达某个目标集合 $T \subseteq S$ 的概率记为 $\Pr_{\mathcal{M}}(s \to T \mid \sigma)$。更准确地说，$\Pr_{\mathcal{M}}(s \to T \mid \sigma)$ 表示在策略 σ 下，从 s 出发并到达 T 的所有路径的概率。若 s 按照 ι_{init} 分布，则使用 $\Pr_{\mathcal{M}}(T \mid \sigma)$。

定义 2（多环境马尔可夫决策过程）。多环境马尔可夫决策过程是一个元组 $\mathcal{N} = \langle S, A, \iota_{init}, \{p_i\}_{i \in I} \rangle$，其中 S, A, ι_{init} 与马尔可夫决策过程中的定义相同，而 $\{p_i\}_{i \in I}$ 是转移函数的集合，其中 I 是有限的环境索引集。

直观而言，多环境马尔可夫决策过程由一组马尔可夫决策过程（环境）构成，它们共享状态和动作，但转移概率各不相同。对于索引集为 I 且集合为 $I' \subseteq I$ 的多环境马尔可夫决策过程 $\mathcal{N}$，我们将环境限制定义为多环境马尔可夫决策过程 $\mathcal{N}_{\downarrow I'} = \langle S, A, \iota_{\text{init}}, \{p_i\}_{i \in I'} \rangle$。给定一个环境 $i \in I$，我们将其对应的马尔可夫决策过程记为 $\mathcal{N}_i = \langle S, A, \iota_{\text{init}}, p_i \rangle$。仅包含一个环境的多环境马尔可夫决策过程就是一个马尔可夫决策过程。路径和策略定义在多环境马尔可夫决策过程的状态和动作上，与马尔可夫决策过程中的策略并无不同。如果每个马尔可夫决策过程都是无环的，则该多环境马尔可夫决策过程也是无环的。

例1。 图1展示了一个包含三个环境的多环境马尔可夫决策过程 $\mathcal{N}_i$。智能体可以提出两个问题， q_1 和 q_2。响应可以是“切换”（$s_1 \leftrightarrow s_2$），也可以是“保持”（循环）。在 $\mathcal{N}_1$中，对 q_1 和 q_2 的响应是切换。在 $\mathcal{N}_2$中，对 q_1 的响应是保持，对 q_2 的响应是切换。智能体可以通过a_1、 a_2、 a_3来猜测环境。猜测 a_i 只有在环境$\{\smile\}$ 中才能到达目标 i。因此，智能体必须通过 q_1、 q_2 推断出环境，才能确保到达目标。■

定义3（几乎必然可达性）。几乎必然可达性质由一组 $T \subseteq S$ 目标状态定义。策略 σ 满足多环境马尔可夫决策过程 $\mathcal{N} = \langle S, A, \iota_{init}, \{p_i\}_{i \in I} \rangle$ 的性质T，当且仅当 $\forall i \in I$: $\Pr_{\mathcal{N}_i}(T \mid \sigma) = 1$。

换言之，当且仅当在每个马尔可夫决策过程中，到达 T 的概率均为 1 时，策略 σ 才满足几乎必然可达性质 T，此时称其为获胜策略。进一步地，如果从状态 s 出发时存在获胜策略，则称状态 $s \in S$ 为必胜状态。未获胜的策略和状态称为失败。现在我们来定义判定问题和策略问题：

给定一个多环境马尔可夫决策过程 $\mathcal{N}$ 和一个几乎必然可达性质 T。判定问题要求判断是否存在满足 T 的策略。策略问题要求在存在此类策略时计算出该策略。

在第4节中，我们讨论判定问题的计算复杂性。接下来，在第5节中，我们介绍用于解决策略问题的算法。其实现和评估的详情将在第6节中介绍。

3 化简为信念-观测马尔可夫决策过程

本节中，我们将策略问题（以及判定问题）化简为：在一个规模呈指数型增长的信念-观测马尔可夫决策过程中寻找策略。这一化简是构造 PSPACE 算法和实际实现的基础模块。本文各处陈述的证明等补充信息见技术报告 [42]。

3.1 将多环境马尔可夫决策过程解释为部分可观测马尔可夫决策过程

定义 4（部分可观测马尔可夫决策过程）。部分可观测马尔可夫决策过程（POMDP）是一个元组$\langle \mathcal{M}, Z, O\rangle$，包含一个马尔可夫决策过程 $\mathcal{M} = \langle S, A, \iota_{init}, p\rangle$、一个观测集合 Z 以及一个观测函数 $O\colon S \to Z$。

部分可观测马尔可夫决策过程是指状态带有观测标记的马尔可夫决策过程。我们将 O 提升到路径，并使用 $O(\pi) = O(s_1)a_1O(s_2)\dots O(s_n)$。我们使用基于观测的策略$\sigma$，即满足以下条件的策略：对于 $\pi, \pi' \in \textsc{Path}$, $O(\pi) = O(\pi')$，有 $\sigma(\pi) = \sigma(\pi')$。多环境马尔可夫决策过程可以转换为一个由以下不交并构成的部分可观测马尔可夫决策过程：

定义 5（并集部分可观测马尔可夫决策过程）。给定一个多环境马尔可夫决策过程 $\mathcal{N} = \langle S, A, \iota_{init}, \{p_i\}_{i\in I}\rangle$，我们定义其并集部分可观测马尔可夫决策过程为 $\mathcal{N}_\sqcup = \langle\langle S', A, \iota'_{init}, p'\rangle, Z, O\rangle$，其状态为 $S' = S \times I$，初始分布为 $\iota'_{init}(\langle s, i\rangle) = \iota_{init}(s) \cdot |I|^{-1}$，转移为 $p'(\langle s, i\rangle, a)(\langle s', i\rangle) = p_i(s, a)(s')$，观测为 $Z = S$，观测函数为 $O(\langle s, i\rangle) = s$。

策略可以观测状态 s，但无法知道当前所处的是哪个马尔可夫决策过程。这迫使任何基于观测的策略在所有环境中都采取相同的选择。

引理1. 给定多环境马尔可夫决策过程 $\mathcal{N}$，存在获胜策略，当且仅当存在基于观测的策略 σ，使得 $\Pr_{\mathcal{N}_\sqcup}(T \mid \sigma) = 1$。

该结论成立的原因如下：首先，部分可观测马尔可夫决策过程的任何基于观测的策略都可以应用于多环境马尔可夫决策过程；其次，反过来，多环境马尔可夫决策过程的任何策略都是基于观测的；第三，这些策略所诱导的马尔可夫链彼此同构。

3.2 信念-观测马尔可夫决策过程

对于部分可观测马尔可夫决策过程，无记忆策略并不足够，这使得策略的计算变得复杂。因此，我们加入历史所包含的信息——也就是截至某一时刻的路径。在多环境马尔可夫决策过程中，这一信息是

（环境）信念（支撑集） $J \subseteq I$，即与多环境马尔可夫决策过程中的某条路径相一致的环境集合。给定一个信念 $J \subseteq I$ 和一次状态-动作-状态转移 $s \xrightarrow{a} s'$，我们定义 $\mathsf{Up}(J, s, a, s') = \{i \in J \mid p_i(s, a, s') > 0\}$，即该转移存在的环境子集。对于一条路径 $\pi \cdot s \in \textsc{Path}$，我们递归地将其对应的信念定义为 $\mathcal{B}(\pi) \subseteq I$：

$$\mathcal{B}(s_0) = I \quad \text{and} \quad \mathcal{B}(\pi \cdot sas') = \mathsf{Up}(\mathcal{B}(\pi \cdot s), s, a, s')$$

多环境马尔可夫决策过程中的信念沿路径单调递减，也就是说，如果我们知道自己不处于某个特定环境中，这一点就会始终成立。

我们的目标是使用一种足以采用无记忆策略的模型。为此，我们将多环境马尔可夫决策过程转换为规模大得多的信念-观测马尔可夫决策过程 [16][2]。

定义 6（BOMDP）。对于一个多环境马尔可夫决策过程 $\mathcal{N} = \langle S, A, \iota_{init}, \{p_i\}_{i \in I} \rangle$，我们将其信念观测马尔可夫决策过程（BOMDP）定义为一个部分可观测马尔可夫决策过程 $\mathcal{G}_{\mathcal{N}} = \langle \langle S', A, \iota'_{init}, p' \rangle, Z, O \rangle$，其状态为 $S' = S \times I \times \mathcal{P}(I)$，初始分布为 $\iota'_{init}(\langle s, j, I \rangle) = \iota_{init}(s) \cdot |I|^{-1}$，转移关系为 $p'(\langle s, j, J \rangle, a)(\langle s', j, J' \rangle) = p_j(s, a, s')$，其中 $J' = \mathsf{Up}(J, s, a, s')$；观测为 $Z = S \times \mathcal{P}(I)$，观测函数为 $O(\langle s, j, J \rangle) = \langle s, J \rangle$。

与并集部分可观测马尔可夫决策过程相比，BOMDP 还会通过相应更新来跟踪信念。我们阐明 BOMDP 的路径与多环境马尔可夫决策过程之间的对应关系。对于多环境马尔可夫决策过程中的路径 π，我们可以在马尔可夫决策过程 $\mathcal{N}_j$ 中精确模拟这条路径，其中 $j \in \mathcal{B}(\pi)$。由于我们在状态中跟踪 $\mathcal{B}(\pi)$，因此可以根据 BOMDP 状态推断我们可能处于哪些环境中。

引理 2。对于多环境马尔可夫决策过程 $\mathcal{N}$及 BOMDP $\mathcal{G}_{\mathcal{N}}$的路径 $\langle s_1, j, J_1 \rangle a_1 \langle s_2, j, J_2 \rangle \ldots \langle s_n, j, J_n \rangle$，令 $j \in J_1$。则：$J_n \neq \emptyset$，且路径 $s_1 a_1 \ldots s_n$ 存在于马尔可夫决策过程 $\mathcal{N}_i$ 中，当且仅当 $i \in J_1 \cap J_n$。

因此，一条路径的信念可以由最后到达状态的观测唯一确定，这也就是“信念观测马尔可夫决策过程”这一名称的由来。

引理 3. 对于 BOMDP 中的任意一对路径 π, π'，有：

$$O(last(\pi)) = O(last(\pi')) \quad implies \quad \mathcal{B}(\pi) = \mathcal{B}(\pi').$$

为便于记号表示，我们定义 $S_J = \{\langle s, j, J \rangle \mid j \in J, s \in S\}$，并类似地记为 $Z_J = \{\langle s, J \rangle \mid s \in S\}$。我们将目标状态 T 提升为 BOMDP 中的状态：$T_{\mathcal{G}_{\mathcal{N}}} = \{\langle s, j, J \rangle \mid s \in T, J \subseteq I, j \in J\}$，并定义目标观测 $T_Z = O(T_{\mathcal{G}_{\mathcal{N}}})$。

定义 7（BOMDP 中的必胜）。设 $\mathcal{G}_{\mathcal{N}}$ 是一个目标观测为 T_Z 的 BOMDP。若对所有 $s \in O^{-1}(z)$，均有 $\Pr_{\mathcal{G}_{\mathcal{N}}}(s \to O^{-1}(T_Z) \mid \sigma) = 1$，则称基于观测的策略 σ 从某个观测 $z \in Z$ 出发是必胜的。

此外，若策略 σ 对初始分布 ι_{init} 是必胜的，则称其为必胜策略。若存在一个对观测 z 必胜的策略，则称该观测是获胜观测。获胜区域 $\mathsf{Win}^{T}_{\mathcal{G}_{\mathcal{N}}}$ 是所有获胜观测的集合。

BOMDP 中的几乎必然获胜对应于 MEMDP 中的必胜。

[2] 这种转换在符号表示上比经由并集部分可观测马尔可夫决策过程更简洁。

定理 1. 对于具有目标状态T 的 MEMDP $\mathcal{N}$，存在一个获胜策略，当且仅当对于具有目标状态 $T_{\mathcal{G}_\mathcal{N}}$的 BOMDP $\mathcal{G}_\mathcal{N}$，存在一个获胜策略。

直观而言，关键在于，对于几乎必然可达性，基于观测的无记忆策略已足够 [13]。对于任意此类策略，在并集部分可观测马尔可夫决策过程和 BOMDP 上诱导的马尔可夫链是双模拟的 [16]。

BOMDP 使策略搜索在概念上更简单。首先，由于无记忆策略足以实现几乎必然可达性，获胜区域不受固定策略的影响：对于分别在观测 z 和 z'中获胜的策略 σ 和 σ'，必定存在一个策略 $\hat{}\sigma$，同时在 z 和 z'中获胜。其次，可以在关于 BOMDP 大小的多项式时间内确定获胜区域 [16]。

3.3 BOMDP 的片段

为避免存储规模呈指数级的 BOMDP，我们只构建片段：可以选择任意一组观测作为前沿观测，并将具有这些观测的状态设为吸收态。我们稍后将讨论前沿的选择。

定义 8（切片信念观测马尔可夫决策过程）。对于一个 BOMDP $\mathcal{G}_\mathcal{N} = \langle\langle S, A, \iota_{init}, p\rangle, Z, O\rangle$以及一组前沿观测 $F \subseteq Z$，我们定义一个 BOMDP $\mathcal{G}_\mathcal{N}|F = \langle\langle S, A, \iota_{init}, p'\rangle, Z, O\rangle$，其满足：

$$\forall s \in S, a \in A\colon p'(s,a) = \begin{cases} \mathsf{dirac}(s) & \textit{if } O(s) \in F, \\ p(s,a) & \textit{otherwise.} \end{cases}$$

我们利用这个切片信念观测马尔可夫决策过程，推导出关于必胜状态集合的约束。

引理 4. 对于任意具有状态 S 和目标状态 T 的 BOMDP $\mathcal{G}_\mathcal{N}$，以及任意前沿观测 $F \subseteq Z$，均有：$\mathsf{Win}^T_{\mathcal{G}_\mathcal{N}|F} \subseteq \mathsf{Win}^T_{\mathcal{G}_\mathcal{N}} \subseteq \mathsf{Win}^{T\cup F}_{\mathcal{G}_\mathcal{N}|F}$。

将（非目标）观测设为吸收态会扩大失败观测集合，而添加目标状态会扩大获胜观测集合。

4 计算复杂性

上述 BOMDP $\mathcal{G}_\mathcal{N}$ 通过定理 1 得出一个指数时间和空间算法。我们可以避免指数级内存需求。本节证明，判定是否存在获胜策略是 PSPACE 完全问题。

定理 2。几乎必然可达性判定问题是 PSPACE 完全的。

该结果由下文的引理 11 和引理 10 得出。在第 4.3 节中，我们将证明，表示获胜策略本身可能需要指数空间。

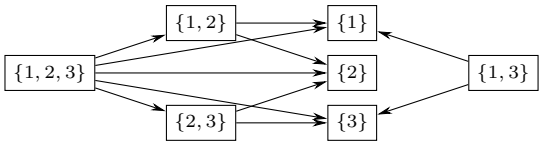

图 2：我们贯穿全文的示例中的环境图。

4.1 在 PSPACE 中判定多环境马尔可夫决策过程的几乎必然必胜性

我们设计了一种内存占用为多项式的算法。该算法利用了 BOMDP 中循环行为的局部性，这种局部性通过无环环境图以及与环境图中的节点相对应的局部 BOMDP 得以形式化。算法在环境图上递归，同时记忆多项局部BOMDP 的结果。

BOMDP 的图结构。首先，沿着多环境马尔可夫决策过程的一条路径，我们只能获得更多信息，因此能够排除某些环境 [14]。由于更新算子的单调性，对于任意 BOMDP，$\langle s, j, J\rangle \in \text{Reachable}(\langle s', j, J'\rangle)$ 蕴含 $J \subseteq J'$。我们定义一个由环境集合构成的图，用于描述信念支持集如何在一次运行过程中更新。

定义 9（环境图）。设 $\mathcal{N}$ 是一个多环境马尔可夫决策过程，p 是 $\mathcal{G}_{\mathcal{N}}$ 的转移函数。$\mathcal{N}$ 的环境图 $\text{GE}_{\mathcal{N}} = (\text{V}_{\mathcal{N}}, \text{E}_{\mathcal{N}})$ 是一个有向图，其顶点为 $\text{V}_{\mathcal{N}} = \mathcal{P}(I)$，边为

$$E_{\mathcal{N}} = \{\langle J, J'\rangle \mid \exists s, s' \in S, a \in A, j \in I. p(\langle s, j, J\rangle, a, \langle s', j, J'\rangle) > 0 \ \textit{and}\ J \neq J'\}.$$

例 2。图 2 展示了例 1 中多环境马尔可夫决策过程的环境图。它由不同的信念支持集组成。例如，从 $\{1, 2, 3\}$ 到 $\{2, 3\}$ 再到 $\{1\}$ 的转移，是由状态 s_0 下的动作 q_1 引起的，如图 1 所示。■

环境图中的路径抽象了 BOMDP 中的路径。由于我们不会创建形如 $\langle J, J\rangle$ 的边，因此将信念支持集保持不变的路径片段概括为一步。我们将这一想法形式化如下：设 $\pi = \langle s_1, j, J_1\rangle a_1 \langle s_2, j, J_2\rangle \ldots \langle s_n, j, J_n\rangle$ 是 BOMDP 中的一条路径。对于任意 $J \subseteq I$，若 $J_i = J$ 对所有 $i \in [n]$ 都成立，则称 π 是一条 J-局部路径。

引理 5. 对于具有环境图 $\text{GE}_{\mathcal{N}}$ 的多环境马尔可夫决策过程 $\mathcal{N}$，若 $\mathcal{G}_{\mathcal{N}}$ 中存在一条路径 $\pi = \pi_1 \ldots \pi_n$，使得每个 π_i 都是 J_i-局部的，则存在一条路径$J_1 \ldots J_n$。

环境图的形状对于我们开发的算法至关重要。

引理 6. 设 $\text{GE}_{\mathcal{N}} = (\text{V}_{\mathcal{N}}, \text{E}_{\mathcal{N}})$ 是多环境马尔可夫决策过程 $\mathcal{N}$ 的环境图。首先，$\text{E}_{\mathcal{N}}(J, J')$ 蕴含 $J' \subsetneq J$。因此，G 是无环的，且其最大路径长度为$|I|$。该图的最大出度为 $|S|^2|A|$。

关于 J, J' 的单调性可由信念更新的定义得出。出度的上界则是下面引理 9 的推论。

局部信念支持BOMDP。在继续之前，我们指出，BOMDP 中（未来的）动态转移只取决于当前状态和环境集合。更正式地说，我们将这一直觉表述如下。

引理 7. 设 $\mathcal{G}_{\mathcal{N}}$ 是一个状态为 S'的BOMDP。对于任意状态 $\langle s,\ j,\ J\rangle \in S'$，令$\mathcal{N}' =$ ReachFragment($\mathcal{N}_{\downarrow J}$, dirac($s$))且 $Y = \{\langle s, i, J\rangle \mid i \in J\}$。那么：

$$\mathsf{ReachFragment}(\mathcal{G}_{\mathcal{N}}, \mathsf{unif}(Y)) = \mathcal{G}_{\mathcal{N}'}.$$

关键之处在于，限制多环境马尔可夫决策过程并不会改变环境 $j \in J$的转移函数。此外，利用更新的单调性，我们只会到达其行为由环境 J决定的BOMDP状态。

这一思路使我们能够在局部分析BOMDP，并将结果推广到完整的BOMDP。我们将局部BOMDP定义为BOMDP中从 S_J内任意状态开始的部分。所有不在 $Z_J = S \times \{J\}$ 中的观测都设为吸收态。

定义10（局部BOMDP）。给定一个多环境马尔可夫决策过程 $\mathcal{N}$，其BOMDP为 $\mathcal{G}_{\mathcal{N}}$，以及一个环境集合 J。环境 J 对应的局部BOMDP是以下片段

$$\textsc{Loc}\mathcal{G}(J) = \mathsf{ReachFragment}(\mathcal{G}_{\mathcal{N}_{\downarrow J}} | F, \mathsf{unif}(S_J)) \quad \textit{where} \quad F = Z \setminus Z_J\ .$$

局部BOMDP的这一定义与完整BOMDP的一个片段相吻合。随后，我们将恰好在环境集合 $J' \subsetneq J$ 内获胜的观测标记为局部BOMDP中的获胜观测，并计算局部BOMDP中的所有获胜观测。这些观测在完整BOMDP中也是获胜的。引理 4 的以下具体化对此进行了形式化说明。

引理 8. 考虑一个 MEMDP $\mathcal{N}$ 以及一个环境子集 J。

$$O\left(\mathsf{Win}^{T'_{\mathcal{G}_{\mathcal{N}}}}_{\textsc{Loc}\mathcal{G}(J)}\right) \cap Z_J \ = \ O\left(\mathsf{Win}^{T_{\mathcal{G}_{\mathcal{N}}}}_{\mathcal{G}_{\mathcal{N}}}\right) \cap Z_J \quad \textit{with} \quad T'_{\mathcal{G}_{\mathcal{N}}} = T_{\mathcal{G}_{\mathcal{N}}} \cup (\mathsf{Win}^{T_{\mathcal{G}_{\mathcal{N}}}}_{\mathcal{G}_{\mathcal{N}}} \setminus S_J).$$

此外，局部BOMDP的规模在 MEMDP 的规模上是多项式有界的。

引理 9. 设 $\mathcal{N}$ 是一个具有状态 S 和动作 A 的 MEMDP。 $\textsc{Loc}\mathcal{G}$（J）至多有 $\mathcal{O}$（$|S|^2 \cdot |A| \cdot |J|$）个状态和 $\mathcal{O}$（$|S|^2 \cdot |A| \cdot |J|^2$）个转移 [3]。

一个 PSPACE 算法。我们给出用于 MEMDP 判定问题的算法1，该算法在环境图 [4] 中沿路径进行深度优先递归。我们首先说明该算法的正确性和空间复杂度。

引理10。 ASWINNING在算法1中以 PSPACE 求解该判定问题。

[3] 转移的数量就是 [4] 中非零项的数量。与深度优先搜索不同，我们不会记住之前访问过的节点。

p

算法1 搜索算法

1: 函数 SEARCH(多环境马尔可夫决策过程 $\mathcal{N} = \langle S, A, \{p_i\}_{i \in I} \iota_{\text{init}} \rangle$, $J \subseteq I$, $T \subseteq S$)
2: $T' \leftarrow \{\langle s, j, J \rangle \mid j \in J, s \in T\}$
3: 对于 J'，使得 $E_{\mathcal{N}}(J, J')$ 执行 ▷ 考虑环境图中的边（定义9）
4: $W_{J'} \leftarrow$ SEARCH($\mathcal{N}, J', T$) ▷ 递归!
5: $T' \leftarrow T' \cup \{\langle s, j, J' \rangle \mid j \in J, \langle s, J' \rangle \in W_{J'}\}$
6: 返回 $\text{Win}^{T'}_{\text{Loc}_{\mathcal{G}}(J)} \cap Z_J$ ▷ 如定义 10 所述构造 BOMDP，然后进行模型检查
7:
8: 函数 ASWINNING(多环境马尔可夫决策过程 $\mathcal{N} = \langle S, A, \{p_i\}_{i \in I} \iota_{\text{init}} \rangle$, $T \subseteq S$)
9: 返回 $O(Supp(\iota_{\text{init}})) \subseteq$ SEARCH($\mathcal{N}, I, T$)

为证明正确性，我们首先注意到 SEARCH($\mathcal{N}, J, T$) 计算的是 $\text{Win}^{T_{\mathcal{G}_\mathcal{N}}}_{\mathcal{G}_\mathcal{N}} \cap Z_J$。我们通过对环境图的结构进行归纳来证明这一点。对于所有没有出边的J，局部BOMDP与仅针对环境 J 的BOMDP相同（引理 7）。否则，注意第 5 行的 T' 与引理 8 中的定义相同，因此根据同一引理，我们返回 $\text{Win}^{T_{\mathcal{G}_\mathcal{N}}}_{\mathcal{G}_\mathcal{N}} \cap Z_J$。最后，根据定理 1，如果BOMDP初始状态的观测是必胜的，那么该多环境马尔可夫决策过程中就存在获胜策略。该算法会终止，因为它会递归遍历有限有向无环图的所有路径，参见引理 6。根据引理 9，前沿状态的数量因而有上界 $|S|^2 \cdot |A|$。因此，算法主体需要多项式空间，而最大递归深度（栈高度）为 $|I|$（引理 6）。综上，空间复杂度为 $\mathcal{O}(|S|^2 \cdot |A| \cdot |I|^2)$。

4.2 判定多环境马尔可夫决策过程中的几乎必然获胜是 PSPACE 困难的

该算法无法改进到超越 PSPACE 的复杂度。

引理 11。MEMDP 判定问题是 PSPACE 困难的。

即使对于无环多环境部分可观测马尔可夫决策过程，该困难性依然成立，并且用到了以下事实。

引理 12。如果无环多环境马尔可夫决策过程中存在获胜策略，那么也存在一个确定性获胜策略。

特别地，几乎必然可达性等价于避开汇状态。这是一种安全性质。对于安全性质，确定性策略就足够了，因为随机化只会访问额外的状态，这对安全性没有益处。

关于引理 11，我们概述从 PSPACE 完全的全称量化布尔公式问题 [20] 问题到 MEMDP 判定问题的多项式时间归约。设 Ψ是一个量化布尔公式，$\Psi = \exists x_1 \forall y_1 \exists x_2 \forall y_2 \ldots \exists x_n \forall y_n \big[\Phi\big]$ 其中 Φ 是一个合取范式布尔公式。问题是判定 Ψ 是否为真。

例 3。考虑量化布尔公式 $\Psi = \forall x \exists y \big[(x \vee y) \wedge (\neg x \vee \neg y)\big]$。我们为每个子句构造一个环境，见图 3[5]。

[5] 为简洁起见，我们展示一个略微简化的多环境马尔可夫决策过程。

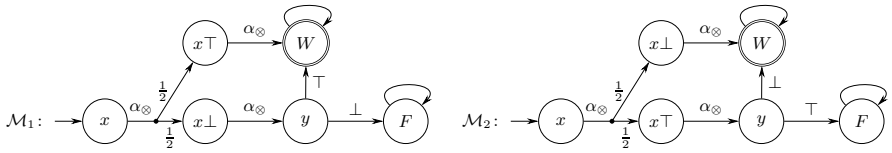

图 3：为量化布尔公式 $\forall x \exists y \big[(x \vee y) \wedge (\neg x \vee \neg y)\big]$ 构造的多环境马尔可夫决策过程。

状态空间为每个变量 $v \in V$设置的三个状态：状态 v 以及用于编码其赋值的状态 $v\top$ 和 $v\bot$ 。此外，我们还设置了专用目标 W和汇状态 F。我们考虑三种动作：动作 true（$\top$）和 false（$\bot$）在语义上描述了对存在量化变量的赋值。动作 any $\alpha_\otimes$ 用于所有其他状态。当且仅当子句中的某个文字被赋值为 true 时，每个环境都会到达目标状态。

直观而言，在这个例子中，策略应将 x 的否定赋给 y。形式上，由 $\sigma(\pi \cdot y) = \top$ 当且仅当 $x_\bot \in \pi$ 刻画的策略 σ 是获胜策略。 ■

由这一构造，我们还可以推导出以下定理。

定理 3. 判定是否存在无记忆获胜策略是 NP 完全问题。

NP 难性的证明对量化布尔公式的命题 SAT 片段采用了类似的构造，该片段不含全称量词。此外，无记忆策略问题属于 NP，因为可以非确定性地猜测一个（大小为多项式的）无记忆策略，并在每个环境中独立验证。

4.3 策略问题

策略是将历史映射到动作的函数，通常是无限对象。然而，我们可以从 BOMDP 中提取获胜策略，而 BOMDP 相对于多环境马尔可夫决策过程仅呈指数级增长。有限状态控制器 [34] 是策略的一种合适且广泛使用的表示形式，只需有限的记忆量。直观而言，记忆状态的数量反映了策略能够区分的历史等价类的数量。一般来说，我们无法指望找到比通过 BOMDP 得到的策略更小的策略。

定理 4. 存在一个多环境马尔可夫决策过程族 $\{\mathcal{N}^n\}_{n\geq 1}$，对于每个 n， $\mathcal{N}^n$ 都有 $2n$ 个环境和 $\mathcal{O}(n)$ 个状态，并且其中每个获胜策略对于 $\mathcal{N}^n$ 都至少需要 2^n 个记忆状态。

我们来说明这个见证。考虑一个多环境马尔可夫决策过程 $\{\mathcal{N}^n\}_n$ 的族，其中 $\mathcal{N}^n$ 包含 $2n$ 个马尔可夫决策过程， $4n$ 个状态被划分为两部分，并且每个状态至多有 $2n$ 个出边动作。我们在图4中概述了该多环境马尔可夫决策过程族。第一部分中，每个状态只有一个动作。记号如下：在状态 s_0 和马尔可夫决策过程 $\mathcal{N}_1^n$ 中，我们以概率1转移到状态 a_0；而在 $\mathcal{N}_2^n$ 中，我们以概率1转移到状态 b_0。在其他每个马尔可夫决策过程中，我们以二分之一的概率转移到任一状态。在状态 s_1 中，我们对环境3、4及所有其他环境进行类似的构造。一条路径 $s_0 b_1 \dots$ 因而与每个马尔可夫决策过程都一致

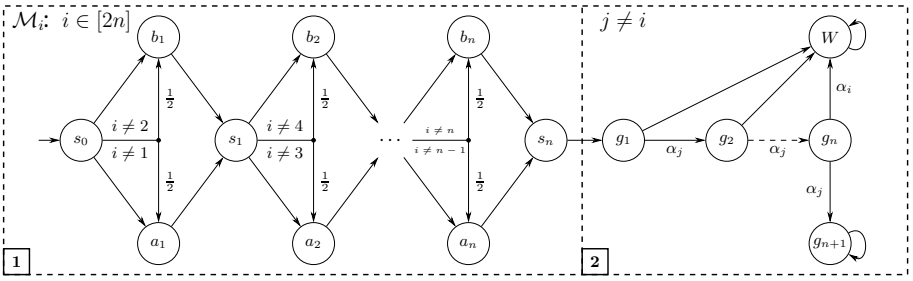

图 4：获胜策略所需指数级内存的见证。

除了 $\mathcal{N}_1^n$。第一部分结束于状态 s_n。根据构造，共有 2^n 条路径终止于 s_n。其中每条路径都与一组唯一的 n 个环境（不）一致。在第二部分中，策略可以通过为每个 $i \in [2n]$ 选择一个动作 α_i，来猜测 n 次环境。只有在马尔可夫决策过程 $\mathcal{N}_i^n$ 中，动作 α_i 才会通向目标状态。在所有其他马尔可夫决策过程中，该转移会从状态 g_j 通向 g_{j+1}。状态 g_{n+1} 在所有马尔可夫决策过程中都是吸收态。重要的是，在采取动作 α_i 并到达 g_{j+1} 后，至多还有一个马尔可夫决策过程与该路径不一致。

此系列中的每个多环境马尔可夫决策过程 $\mathcal{N}^n$ 都有一个获胜策略：当 $a_i \in \pi$ 时执行 $\sigma(\pi \cdot g_i) = \alpha_{2i-1}$，否则执行 $\sigma(\pi \cdot g_i) = \alpha_{2i}$。此外，到达状态 s_n时，有限记忆控制器的记忆状态必须反映与该历史相一致的精确环境集合。这样的集合共有 2^n 个。证明表明，如果存储的信息更少，就会有两条路径到达同一记忆状态，但与这两条路径相一致的环境集合不同。由于在多环境马尔可夫决策过程的第二部分中，我们只能通过 n 个动作排除 n 个环境，因此无法确保在每种环境中都获胜。

5 部分博弈探索算法

本节中，我们提出一种用于求解策略问题的算法。我们调整算法，使其侧重于运行时间而非内存复杂度，同时力求避免内存耗尽。我们采用若干关键要素，构造算法 1 的一种实用变体，并支持提取获胜策略。

首先，我们采用一种从 BOMDP 到信念随机博弈（BSG）的抽象，类似于 [46] that reduces the number of states and simplifies the iterative construction[6]。其次，我们调整并推广有界模型检测中的思想 [6] to build and model check only a fragment of the BSG, using explicit partial exploration approaches as in, e.g., [33,9,43,29]。第三，我们的探索不会持续扩展该片段，还可以利用迄今获得的模型检测结果对其进行剪枝。BSG 的结构由环境图刻画，这使该方法很有前景，并带来一些自然的启发式方法。第四，获胜区域的结构使我们能够

[6] 在撰写本文时，我们尚未发现适用于 BOMDP 的多项式时间算法。

将结果推广到未见过的状态。由此，我们在部分探索的背景下将 [26] 中的一个思想付诸实践。最后，我们将单个马尔可夫决策过程作为一种高效且重要的预处理步骤进行分析。下文将讨论这些要素。

抽象为信念支持博弈。我们简要回顾随机博弈（SG）。更多详情请参见 [38,17] 。

定义 11（随机博弈）。随机博弈是一个元组 $\mathcal{B} = \langle \mathcal{M}, S_1, S_2 \rangle$，其中$\mathcal{M} = \langle S, A, \iota_{init}, p \rangle$ 是一个马尔可夫决策过程，且(S_1, S_2)是 S的一个划分。

S_1 是玩家1的状态， S_2 是玩家2的状态。通常，我们也将（无记忆确定性）策略“划分”为两个函数 $\sigma_1\colon S_1 \to A$ 和$\sigma_1\colon S_2 \to A$。若对于所有σ_2，都有 $\Pr(T \mid \sigma_1, \sigma_2)$，则玩家1的策略 σ_1 对状态 s 是获胜的。我们继续使用 $\mathrm{Win}^T_{\mathcal{B}_\mathcal{N}}$表示具有获胜策略的状态集合。

我们应用基于博弈的抽象，将具有相同观测的状态归为一组。玩家1的状态刻画了 BOMDP 中的观测，即由多环境马尔可夫决策过程状态 s 与环境子集 J 构成的元组$\langle s, J \rangle$ 。玩家1选择动作 a，结果是玩家2的状态 $\langle \langle s, J \rangle, a \rangle$。随后，玩家2选择一个环境 $j \in J$，博弈模拟从 $\langle s, j, J \rangle$出发的转移，也就是说，它模拟多环境马尔可夫决策过程中从 s 出发的转移。形式化地说：

定义12（BSG）。设 $\mathcal{G}_\mathcal{N}$ 是一个 BOMDP，具有 $\mathcal{G}_\mathcal{N} = \langle \langle S, A, \iota_{init}, p \rangle, Z, O \rangle$。对于 $\mathcal{G}_\mathcal{N}$，信念支持博弈 $\mathcal{B}_\mathcal{N}$ 是一个随机博弈 $\mathcal{B}_\mathcal{N} = \langle \langle S', A', \iota'_{init}, p \rangle, S_1, S_2 \rangle$，其中 $S' = S_1 \cup S_2$ 与通常情况相同，玩家1的状态为 $S_1 = Z$，玩家2的状态为 $S_2 = Z \times A$，动作集为 $A' = A \cup I$，初始分布为 $\iota'_{init}(\langle s, I \rangle) = \sum_{i \in I} \iota_{init}(\langle s, i, I \rangle)$，而（部分）转移函数 p 分别针对玩家1和玩家2定义如下：

$$p'(z, a) = \mathsf{dirac}(\langle z, a \rangle) \qquad \textit{(Player 1)}$$

$$p'(\langle z, a \rangle, j, z') = p(\langle s, j, J \rangle, a, \langle s', j, J' \rangle) \textit{ with } z = \langle s, J \rangle, z' = \langle s', J' \rangle \qquad \textit{(Player 2)}$$

引理13。具有目标状态 T 的（无环）多环境马尔可夫决策过程 $\mathcal{N}$ 是必胜的，当且仅当信念支持博弈 $\mathcal{B}_\mathcal{N}$ 中存在一个以 T_Z 为目标状态的获胜策略。

因此，对于无环多环境马尔可夫决策过程，基于信念支持博弈的算法是可靠且完备的；然而，对于有环马尔可夫决策过程，它可能无法找到获胜策略。算法的其余部分基于信念支持博弈来表述；我们使用切片信念支持博弈作为切片 BOMDP 的信念支持博弈，或者等价地，将某些状态设为吸收态后的信念支持博弈。

主算法。我们概述用于策略问题的算法 2。我们跟踪必然观测集和失败观测集（BSG 中的状态）。初始时，目标状态为必胜状态。此外，通过简单的预处理，我们确定各个马尔可夫决策过程中的一些必胜状态和失败状态。

我们不断迭代，直到初始状态成为必胜状态或失败状态。我们的算法构造一个切片 BSG，并即时判断某个状态是否应为前沿状态，最终返回切片 BSG 和所使用的前沿状态。下面将讨论具体实现。对于切片 BSG，我们计算两次获胜区域：一次假设前沿状态为必胜状态，另一次假设前沿状态为失败状态。由此得到必胜状态和失败状态的近似，见引理 4。我们可以从必胜状态中提取一个随机化获胜策略 [13]。

算法 2 策略查找算法

1: 函数 FindPolicy(多环境马尔可夫决策过程 $\mathcal{N} = \langle S, A, \{p_i\}_{i \in I} \iota_{\text{init}} \rangle$, 目标状态 $T \subseteq S$)
2: $W \leftarrow \{\langle s, J\rangle \mid s \in T, J \subseteq I\}$; $L \leftarrow \emptyset$; $i \leftarrow 1$; $S_{\text{init}} \leftarrow \mathit{Supp}(\iota_{\text{init}}) \times \{I\}$
3: 当 $S_{\text{init}} \cap W \neq W$ 且 $S_{\text{init}} \cap L = \emptyset$ 时
4: $\langle \mathcal{B}, F\rangle \leftarrow$ GenerateGameSlice($\mathcal{N}, W, L, i$)
5: $W \leftarrow W \cup$贏$^{W}_{\mathcal{B}}$
6: $L \leftarrow L \cup S \setminus$贏$^{W \cup F}_{\mathcal{B}}$
7: $i \leftarrow i + 1$
8: 如果 $S_{\text{init}} \subseteq W$ 则返回 ExtractPolicy(W)，否则返回 $\perp$

算法 3 博弈生成算法

1: 函数 GenerateGameSlice(多环境马尔可夫决策过程 $\mathcal{N}$, W, L, i)
2: $Q \leftarrow \{s_\iota\}$; $E = \{s_\iota\}$
3: 当 $s \in Q$ 和 $|E| \leq$ Bound$[i]$ 存在时执行
4: $E \leftarrow E \cup \{s\}$ ▷将 s 标记为已探索
5: $\mathcal{B} \leftarrow \mathcal{B}_{\mathcal{N}}|\ (S \setminus E)$ ▷ 扩展博弈，剪除所有未探索的部分
6: $Q \leftarrow$ 可达($\mathcal{B}$) $\setminus (E \cup W \cup L)$ ▷ 添加新到达的状态
7: 返回 $\mathcal{B}, Q$

可靠性。假设 $\mathcal{B}_{\mathcal{N}}$ 确实是一个前沿为 F 的切片 BSG。则以下不变量成立：$W \subseteq \text{Win}^{T}_{\mathcal{B}_{\mathcal{N}}}$ 和 $L \cap \text{Win}^{T}_{\mathcal{B}_{\mathcal{N}}} = \emptyset$.。该不变量利用了这样一个事实：我们可以从切片 BSG（隐式地）切分出完整的 BSG，同时保留每个状态的必胜状态属性，具体形式化如下。在后续迭代中，我们只探索隐式切分出的 BSG。

引理 14. 给定 $W \subseteq \text{Win}^{T_{\mathcal{B}_{\mathcal{N}}}}_{\mathcal{B}_{\mathcal{N}}}$ 和 $L \subseteq S \setminus \text{Win}^{T_{\mathcal{B}_{\mathcal{N}}}}_{\mathcal{B}_{\mathcal{N}}}$: $\text{Win}^{T_{\mathcal{B}_{\mathcal{N}}}}_{\mathcal{B}_{\mathcal{N}}} = \text{Win}^{T_{\mathcal{B}_{\mathcal{N}}} \cup W}_{\mathcal{B}_{\mathcal{N}}|W \cup L}$

终止性取决于切片博弈生成。只需确保从长远来看，W 或 L 会增长，因为状态的数量有限。如果 W和 L 在若干次迭代中保持不变，$W \cup L$ 将被用作前沿。然后，新博弈就足以一次性确定 $s \in W$。

生成切片 BSG。算法 3 概述了切片 BSG 的生成过程。具体而言，我们从初始状态开始探索隐式 BSG，但将所有未显式探索的状态设为吸收态。在每次迭代中，我们首先检查 Q 中是否还有待探索的状态，以及 E 中已探索状态的数量是否低于阈值 Bound$[i]$。然后，我们从优先队列中取出一个状态并将其添加到 E。我们找到新的可达状态[7] 并将它们添加到队列 Q。

推广获胜状态和失败状态。我们的目标是确定游戏中的某个状态 $\mathcal{B}_{\mathcal{N}}$ 是必胜的，而无需对其进行探索。首先，观察：

引理 15。在多环境马尔可夫决策过程 $\mathcal{N}$ 中，获胜策略在 $\mathcal{N}_{\downarrow J}$ 中对任意 J都是必胜的。

[7] 在第 5 行中，我们并不是从头开始重新构建游戏 $\mathcal{B}$，而是逐步构建数据结构。同样，可达状态也是这一构建过程的直接副产品。

由此可直接得出关于两个环境 $J_1 \subseteq J_2$的以下结论：

$$\langle s, J_2\rangle \in \mathsf{Win}^T_{\mathcal{B}_\mathcal{N}} \quad \text{implies} \quad \langle s, J_1\rangle \in \mathsf{Win}^T_{\mathcal{B}_\mathcal{N}}.$$

因此，我们可以如下存储 W （对称地，也可以存储 L）。对于每个多环境马尔可夫决策过程状态 $s \in S$， $W_s = \{J \mid \langle s, J\rangle \in W\}$ 在偏序 $P = (I, \subset)$ 上是向下封闭的。这使得存储更加高效：我们只需存储两两比较下的极大元素集合，即反链，

$$W_s^{\max} = \{J \in W_s \mid \forall J' \in W_s \text{ with } J \not\subseteq J'\}.$$

要确定 $\langle s, J\rangle$ 是否为必胜，我们检查是否对某个$J' \in W_s^{\max}$， $J \subseteq J'$ 。将 J 加入 $W_s^{\max}$需要先移除所有 $J' \subseteq J$ ，再加入 J。不过请注意，在最坏情况下， $|W_s^{\max}|$ 对于 $|I|$ 仍然是指数型的。

启发式方法的选择。该算法提供了一定的自由度。我们通过实证评估以下几个方面。(1) 第 i 轮切片 BSG 的最大规模上限$[i]$ 至关重要。如果上限太小，集合 W 和 L 在每一轮中的增长都会很缓慢。权衡之所以更加复杂，是因为集合 W 和 L 可能推广到未见过的状态。(2) 对于固定的上限$[i]$，如何确定状态探索的优先级尚不明确。PSPACE 算法表明，深入探索是有益的；而广泛探索则最有利于推广到未见过的状态。(3) 最后，计算 W 和 L都会带来开销。如果存在获胜策略，我们只需计算 W。然而，计算 L 可能有助于剪除状态空间的部分区域。对于不可满足实例，计算 W 也有类似的观察。

注 1. 算法 2 经过少量调整后即可达到算法1所需的 PSPACE 算法要求。优先队列必须确保始终包含完整的（可达）局部 BSG，并优先探索 J 较小的状态 $\langle s, J\rangle$ 。此外， W 和 L 需要定期剪枝；如果将 W 剪枝至多项式大小的上限，我们就无法提取策略。实际上，我们可以将 W 中经剪枝的部分写入磁盘。

6 实验

我们重点介绍两个方面：（1）将我们的原型与现有的部分可观测马尔可夫决策过程基线方法进行比较；（2）考察探索启发式方法。技术报告 [42] 提供了实现、基准测试及更多结果的详细信息。

实现。我们基于算法 2，在概率模型检测器 Storm [22]之上提供了一种新颖的部分博弈探索（PAGE）原型。我们使用带整数常量的 Prism 语言表示多环境马尔可夫决策过程。对这些常量的每种赋值都会产生一个显式的马尔可夫决策过程。随机博弈使用现有的数据结构和图算法构建并求解。

设置。我们借鉴 POMDP 与 MEMDP 文献 [26,12,21]，创建了一组基准测试。我们考虑了可满足和不可满足

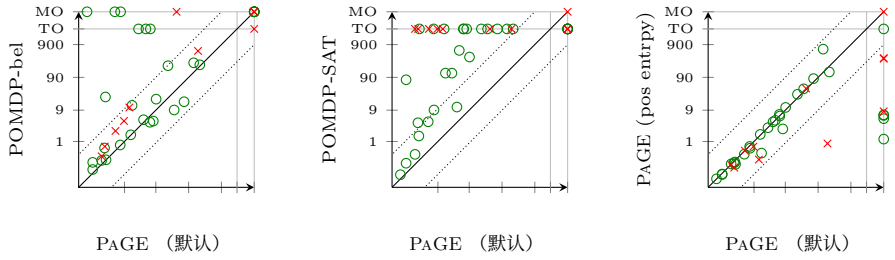

图 5：基线方法与新型 PAGE 算法的性能

的基准测试。在后一种情况下，不存在获胜策略。我们按照定义 5，从多环境马尔可夫决策过程构造部分可观测马尔可夫决策过程。作为基线方法，我们使用以下两种现有的部分可观测马尔可夫决策过程算法。对于几乎必然性质，信念马尔可夫决策过程构造 [7] 的效果类似于经过高效工程优化的游戏构造变体，但其适用范围针对更一般的定量性质。基于可满足性的 [26] 方法旨在寻找规模逐渐增大的策略。我们在配备 3GHz 英特尔酷睿 i9-10980XE 处理器的系统上评估所有基准测试。我们将时间限制设为 30 分钟，内存限制设为 32 GB。

结果。图 5 展示了不同配置[8]之间的性能对比（对数刻度）。绿色圆圈表示可满足的基准测试，红色叉号表示不可满足的基准测试。横轴为默认配置下的 PAGE 。第一幅图将其与信念马尔可夫决策过程构造进行比较。量身定制的启发式方法和信念支持集表示在几乎所有情况下都带来了显著优势。少数位于对角线下方的点，是构建状态空间时探索率较高所致。第二幅图将其与基于 SAT 的方法进行比较；该方法仅适用于寻找策略，无法证明策略不存在。这种方法隐式地搜索某一特定类别的策略，但其结构并不适用于某些多环境马尔可夫决策过程。第三幅图比较了默认配置下的 PAGE （以负熵作为优先级函数）与使用正熵的 PAGE。正如预期，不同的优先级对性能有显著影响。

表1展示了可满足和不可满足基准测试的概况。每张表都列出了多环境马尔可夫决策过程中的环境数量、状态数量以及每个状态的动作数。对于 PAGE，我们同时纳入默认配置（负熵）和变体（正熵）。对于这两种配置，我们都列出了运行时间以及构建出的 BSG 的最大规模。我们还列出了两种基线方法的运行时间。不出所料，待探索的状态数量是预测性能的良好指标，相对性能如图 5 所示。

[8] 图中的每个点 $\langle x, y \rangle$ 都表示一个基准测试：横轴上的配置用时 x 求解，而纵轴上的配置用时 y 求解。因此，对角线以上的点表示横轴上的配置更快。

表1：可满足与不可满足基准测试结果

				PAGE(正熵) t	PAGE(正熵) n	PAGE(负熵) t	PAGE(负熵) n	信念 t	可满足性 t
Grid	19	132	4	0.2	3002	0.2	3002	0.6	3.7
	39	152	4	0.4	9007	1.6	41029	12.6	121.3
	199	474	4	6.4	337177	MO		MO	TO
Catch	256	625	4	6.6	93614	5.9	41094	3.8	TO
	256	6561	4	40.1	749295	32.6	337899	9.1	TO
	256	14641	4	82.5	1826922	65.3	338079	16.2	TO
Exp	8	19	9	0.1	349	0.1	349	0.1	75.9
	20	43	21	131.4	192163	197.6	448443	217.6	TO
	24	51	25	TO		MO		MO	TO
Frogger	10	1200	4	0.2	1200	0.2	1200	22.7	1.4
	20	1200	4	0.4	1200	0.5	1200	MO	3.9
	80	4000	4	4.4	4000	4.4	4000	TO	597.3
	99	4000	4	5.9	8001	6.1	8001	TO	TO

		PAGE(正熵) t	PAGE(正熵) n	PAGE(负熵) t	PAGE(负熵) n	信念 t
MMind	16 21 16	0.1	1003	0.2	1445	0.3
	27 17 27	0.5	5167	0.5	7579	2.0
	32 25 32	0.6	7799	0.9	11809	4.2
	81 21 81	41.1	170291	38.6	296407	MO
Exp	20 42 21	0.8	9005	173.8	388127	576.1
	24 50 25	8.3	41022	MO		MO
	32 66 33	347.7	337177	MO		MO

7 结论

本文研究具有任意数量环境且以几乎必然可达性为目标的多环境马尔可夫决策过程。我们给出了新颖且紧致的复杂度界，并利用这些发现推导出一种新算法。该算法在一系列广泛的基准测试中优于针对部分可观测马尔可夫决策过程的方法。未来，我们将直接在 BOMDP [16] 上应用一种算法。

数据可用性声明

与本文相关的补充材料已在 Zenodo 上公开获取：
https://doi.org/10.5281/zenodo.7560675

参考文献

1. Roman Andriushchenko、Milan Ceska、Sebastian Junges、Joost-Pieter Katoen 和 Simon Stupinsk´y。PAYNT：用于概率程序归纳综合的工具。载于 *CAV*，*LNCS* 第 12759 卷，第 856–869 页。施普林格，2021年。2. Sebastian Arming、Ezio Bartocci、Krishnendu Chatterjee、Joost-Pieter Katoen 和 Ana Sokolova。通过部分可观测马尔可夫决策过程实现部分可观测马尔可夫决策过程的参数无关策略。载于 *QEST*，*LNCS* 第 11024 卷，第 53–70 页。施普林格，2018 年。3. Mohammad Gheshlaghi Azar 、Alessandro Lazaric 和 Emma Brunskill。有限模型集上的多臂老虎机序贯迁移。载于 *NIPS* 2220 2228, ，第 – 页。2013
4. Christel Baier、Marcus Gr¨oßer 和 Nathalie Bertrand。概率 ω-自动机。*J. ACM*，59(1):1:1–1:52，2012年。5. Christel Baier 和 Joost-Pieter Katoen。*Principles of model checking*。麻省理工学院出版社，2008年。6. Armin Biere、Alessandro Cimatti、Edmund M. Clarke、Ofer Strichman 和 Yun- shan Zhu。有界模型检测。*Adv. Comput.*，58:117–148，2003年。7. Alexander Bork、Sebastian Junges、Joost-Pieter Katoen 和 Tim Quatmann。无限时域部分可观测马尔可夫决策过程的验证。载于 *ATVA*，*LNCS* 第 12302 卷，第 288–304 页。施普林格，2020年。8. Alexander Bork、Joost-Pieter Katoen 和 Tim Quatmann。部分可观测马尔可夫决策过程中期望总回报的下近似。载于 *TACAS (2)*，*LNCS* 第 13244 卷，第 22–40 页。施普林格，2022年。9. Tom´as Br´azdil、Krishnendu Chatterjee、Martin Chmelik、Vojtech Forejt、Jan Kret´ınsk´y、Marta Z. Kwiatkowska、David Parker 和 Mateusz Ujma。利用学习算法验证马尔可夫决策过程。载于 *ATVA*，*LNCS* 第 8837 卷，第 98–114 页。施普林格，2014年。10. Peter Buchholz 和 Dimitri Scheftelowitsch。并发马尔可夫决策过程的加权奖励和计算。*Math. Methods Oper. Res.*，89(1):1–42，2019年。11. Iadine Chades、Josie Carwardine、Tara G. Martin、Samuel Nicol、R´egis Sabbadin 和 Olivier Buffet。混合可观测性马尔可夫决策过程：自适应管理问题建模的解决方案。载于 *AAAI*。美国人工智能协会出版社，2012年。12. Krishnendu Chatterjee、Martin Chmelik 和 Jessica Davies。一种基于符号可满足性的算法，用于在部分可观测马尔可夫决策过程中实现具有小型策略的几乎必然可达性。载于 *AAAI*，第 3225–3232 页。美国人工智能协会出版社，2016年。13. Krishnendu Chatterjee、Martin Chmelik、Raghav Gupta 和 Ayush Kanodia。部分可观测马尔可夫决策过程中的最优成本几乎必然可达性。*Artif. Intell.*，234:26–48，2016年。14. 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 *ICAPS*，第 48–56 页。美国人工智能协会出版社，2020年。15. Krishnendu Chatterjee、Martin Chmelik 和 Mathieu Tracol。关于带有 ω-正则目标的部分可观测马尔可夫决策过程，哪些问题是可判定的？载于 *CSL*，*LIPIcs* 第 23 卷，第 165–180 页。施普林格·莱布尼茨信息学中心，2013年。

16. Krishnendu Chatterjee、Martin Chmelik 和 Mathieu Tracol。关于带有 ω *J.* -正则目 *Comput. Syst. Sci.* 82 5 878 911, 2016
标的部分可观测马尔可夫决策过程，哪些问题是可判定的？, ( ): – 。17. Krishnendu Chatterjee、Marcin Jurdzinski 和 Thomas A. Henzinger。简单随机奇偶博弈。载于 *CSL*, *LNCS* 第 2803 卷，第 100–113 页。施普林格，2003年。18. Philipp Chrszon、Clemens Dubslaff、Sascha Kl¨uppelholz 和 Christel Baier。Pro-Feat：面向特征的工程，用于基于产品族的概率模型检测。*Formal Aspects Comput.*，30(1):45–75，2018年。19. Luca de Alfaro。无记忆部分信息策略下概率系统的验证是困难的。技术报告，加州大学伯克利分校，1999年。曾发表于 ProbMiV。20. M. R. Garey 和 David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness* 1979 。W. H. Freeman，。21. Arnd Hartmanns、Michaela Klauck、David Parker、Tim Quatmann 和 Enno Rui-jters。定量验证基准集。载于 *TACAS (1)*, *LNCS* 第 11427 卷，第 344–350 页。施普林格，2019年。22. Christian Hensel、Sebastian Junges、Joost-Pieter Katoen、Tim Quatmann 和 Matthias Volk。概率模型检测器 Storm。 *Int. J. Softw. Tools Technol.*, ( ): – 。 *Transf.* 24 4 589 610, 2022
23. Manfred Jaeger、Giorgio Bacci、Giovanni Bacci、Kim Guldstrand Larsen 和 Peter Gjøl Jensen。用非精确马尔可夫决策过程逼近欧几里得。载于 *ISoLA (1)*, *LNCS* 第 12476 卷，第 275–289 页。施普林格，2020年。24. Nils Jansen、Sebastian Junges 和 Joost-Pieter Katoen。马尔可夫模型中的参数综合：简明综述。 *CoRR*, abs/2207.06801, 2022年。25. Bengt Jonsson 和 Kim Guldstrand Larsen。概率过程的规约与细化。载于 *LICS*，第 266–277 页。IEEE计算机学会，1991年。26. Sebastian Junges、Nils Jansen 和 Sanjit A. Seshia。在部分可观测马尔可夫决策过程中确保几乎必然可达性。载于 *CAV (2)*, *LNCS* 第 12760 卷，第 602–625 页。施普林格，2021年。27. Leslie Pack Kaelbling、Michael L. Littman 和 Anthony R. Cassandra。在部分可观测随机领域中进行规划与行动。 *Artif. Intell.*，101(1-2):99–134，1998年。28. Robert Kirk、Amy Zhang、Edward Grefenstette 和 Tim Rockt¨aschel。深度强化学习中泛化问题综述。 *CoRR*, abs/2111.09794，2021年。29. Jan Kret´ınsk´y 和 Tobias Meggendorfer。关于核心：马尔可夫决策过程的部分探索框架。 *Log. Methods Comput. Sci.*，16(4)，2020年。30. Marta Kwiatkowska、Gethin Norman 和 Dave Parker。PRISM 4.0：概率实时系统验证。载于 *CAV*, *LNCS* 第 6806 卷，第 585–591 页。施普林格，2011年。31. Michael L. Littman、Anthony R. Cassandra 和 Leslie Pack Kaelbling。部分可观测环境中的策略学习：扩展规模。载于 *ICML*，第 362–370 页。Morgan Kaufmann，1995年。32. 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。 *Artif. Intell.*, 147(1-2):5–34，2003年。33. H. Brendan McMahan、Maxim Likhachev 和 Geoffrey J. Gordon。有界实时动态规划：具有单调上界和性能保证的实时动态规划。在 *ICML* 119 *ACM International Conference Proceeding* 中，卷 *Series* 569 576 2005
号为，页码 –。美国计算机协会，。

34. Nicolas Meuleau、Leonid Peshkin、Kee-Eung Kim 和 Leslie Pack Kaelbling。面向部分可观测环境的有限状态控制器学习。载于 *UAI*，第427–436页。Morgan Kaufmann，1999年。35. Gethin Norman、David Parker 和 Xueyi Zou。部分可观测概率系统的验证与控制。 *Real Time Syst.*，53(3):354–402，2017年。36. Jean-Fran¸cois Raskin 和 Ocan Sankur。多环境马尔可夫决策过程。载于 *FSTTCS*， *LIPIcs*第29卷，第531–543页。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik，2014年。37. John H. Reif。不完全信息双人博弈的复杂性。

J.，（）：–。

Comput. Syst. Sci. 29 2 274 301, 1984

38. L. S. Shapley。随机博弈*。 *Proceedings of the National Academy of Sciences*，39 10 1095 1100，1953年（）：–。39. Trey Smith 和 Reid G. Simmons。基于点的部分可观测马尔可夫决策过程算法：改进的分析与实现。载于 *UAI*，第542–547页。AUAI Press，2005年。40. Lauren N. Steimle、David L. Kaufman 和 Brian T. Denton。多模型马尔可夫决策过程。 *IISE Trans.*，53(10):1124–1139，2021年。41. 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 Rabin 目标的 PSPACE 算法。载于 *CONCUR*， *LIPIcs*第311卷，第40:1–40:17页。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik，2024年。42. Marck van der Vegt、Nils Jansen 和 Sebastian Junges。多环境马尔可夫决策过程中的鲁棒几乎必然可达性。 *CoRR*，abs/2301.11296，2023年。43. Matthias Volk、Sebastian Junges 和 Joost-Pieter Katoen。通过模型检测技术进行快速动态故障树分析。 *IEEE Trans. Ind. Informatics*，14(1):370–379，2018年。44. Wolfram Wiesemann、Daniel Kuhn 和 Ber¸c Rustem。鲁棒马尔可夫决策过程。 *Math. Oper. Res.*，38(1):153–183，2013年。45. Tobias Winkler、Sebastian Junges、Guillermo A. P´erez 和 Joost-Pieter Katoen。参数化马尔可夫决策过程中可达性的复杂性。载于 *CONCUR*， *LIPIcs*第140卷，第14:1–14:17页。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik，2019年。46. Leonore Winterer、Sebastian Junges、Ralf Wimmer、Nils Jansen、Ufuk Topcu、Joost-Pieter Katoen 和 Bernd Becker。通过基于博弈的抽象进行面向机器人规划的部分可观测马尔可夫决策过程策略综合。 *IEEE Trans. Autom. Control.*，66(3):1040–1054，2021年。

多环境马尔可夫决策过程中有限状态控制器的定义

作为策略的一种更具体表示形式，我们将使用有限状态控制器（有限状态控制器）。鉴于引理 12，以及定理4中的构造使用了无环多环境马尔可夫决策过程，我们可以简化问题，仅考虑确定性有限状态控制器。

定义 13（有限状态控制器）。有限状态控制器（FSC）是一个元组 $C = (Q, A, \delta, a, q_\iota)$，其中：

– Q 是有限状态集，即策略的记忆；– A 是有限动作集；– δ: $Q \times S \to Q$ 是转移函数；– a: $Q \times S \to A$ 是动作函数；– $q_\iota \in Q$ 是初始状态。

有限状态控制器是一种可用作马尔可夫决策过程（以及推广而言的多环境马尔可夫决策过程/BSG）策略的自动机。有限状态控制器和马尔可夫决策过程以交替方式彼此提供输入。首先，有限状态控制器提供初始动作 $a(q_\iota, s_\iota)$，随后马尔可夫决策过程执行该动作并转移到某个状态 s'。接着，有限状态控制器从记忆状态 q 转移到 $q' = \delta(q, s')$。此过程无限重复。我们将转移函数扩展到马尔可夫决策过程路径：

$$\delta^+(s_\iota) = q_\iota$$
$$\delta^+(s_\iota \dots s\ i\ s') = \delta(\delta^+(s_\iota \dots s), s')$$

直观地说，$\delta^+(\pi)$ 表示有限状态控制器处理 π 后到达的记忆状态。然后，我们可以使用该有限状态控制器定义策略 σ：

$$\sigma(\pi) = a(\delta^+(\pi))$$

因此，策略的大小就是其状态空间的大小 $|Q|$。当 $|Q| = 1$ 时，该策略是无记忆的。

B 第 3.1 节中关于部分可观测马尔可夫决策过程的证明

B.1 引理1的证明

回顾引理1：

引理1. 给定多环境马尔可夫决策过程 $\mathcal{N}$，存在获胜策略，当且仅当存在基于观测的策略 σ，使得 $\Pr_{\mathcal{N}_{\sqcup}}(T \mid \sigma) = 1$。

证明。$\mathcal{N}_{\sqcup}$ 的观测集合等于 $\mathcal{N}$、$O = S$ 的状态集合。因此，我们可以将 $\mathcal{N}$ 上的策略一一映射为 $\mathcal{N}_{\sqcup}$ 上的（基于观测的）策略，只需将状态视为观测即可，反之亦然。

($\Rightarrow$)：给定多环境马尔可夫决策过程 $\mathcal{N}$ 的获胜策略 σ，我们证明 σ 对并集部分可观测马尔可夫决策过程 $\mathcal{N}_{\sqcup}$ 也必胜。根据多环境马尔可夫决策过程获胜策略的定义，这意味着在每个环境中，到达目标状态的概率均为 1。根据 $\mathcal{N}_{\sqcup}$ 的构造，这也意味着从 $\mathcal{N}_{\sqcup}$ 的每个初始状态出发，到达目标状态的概率均为 1，因此 σ 在 $\mathcal{N}_{\sqcup}$ 中必胜。

($\Leftarrow$)：给定并集部分可观测马尔可夫决策过程 $\mathcal{N}_{\sqcup}$的获胜策略 σ，我们证明 σ 对于多环境马尔可夫决策过程 $\mathcal{N}$也必胜。我们知道，对于初始分布ι_{init}，σ 是获胜策略，这意味着从每个初始状态 $s_i \in \text{Supp}(\iota_{\text{init}})$出发，到达目标集合 T的概率均为 1。根据构造，每个初始状态 s_i 都属于 $\mathcal{N}_{\sqcup}$的一个互不相交部分，该部分与多环境马尔可夫决策过程中的一个环境相同，因此 σ 在所有环境中都能获胜，从而在该多环境马尔可夫决策过程中也能获胜。□

C 关于第 5 节中 BOMDP 的证明

C.1 引理 2 的证明

回顾引理 2:

引理 2. 对于 MEMDP $\mathcal{N}$及 BOMDP $\mathcal{G}_{\mathcal{N}}$中的路径 $\langle s_1, j, J_1\rangle a_1 \langle s_2, j, J_2\rangle \dots \langle s_n, j, J_n\rangle$，设 $j \in J_1$。则：MDP $\mathcal{N}_i$ 中存在路径 $s_1 a_1 \dots s_n$ 当且仅当 $i \in J_1 \cap J_n$。

证明。
(⇒)：路径 $s_1 \dots s_n$ 在 $\mathcal{N}_i$ 中的存在性告诉我们，对于所有 k：$p_i(s_k, a_k)(s_{k+1}) > 0$。正如引理所述，我们知道 $i \in J_1$。通过归纳地应用 J_k 的定义，我们确实得到，对于所有 $1 \le k \le n$：$i \in J_k$，从而得到$i \in J_n$。

(⇐)：根据信念更新函数 Up 的定义，我们有 $J_n \subseteq J_{n-1} \subseteq \cdots \subseteq J_1$。结合假设 $i \in J_n$，我们可知，对于所有 $1 \le k \le n$：$i \in J_k$。根据 J_k 的定义，我们得到 $p_i(s_k, a_k)(s_{k+1}) > 0$，这意味着 $s_1 \dots s_n$ 是 $\mathcal{N}_i$ 中的一条路径。□

C.2 引理 3 的证明

回顾引理 3:

引理 3. 对于 BOMDP 中任意一对路径 π, π'，我们有:

$$O(last(\pi)) = O(last(\pi')) \quad implies \quad \mathcal{B}(\pi) = \mathcal{B}(\pi').$$

证明。假设 $O(\text{last}(\pi)) = O(\text{last}(\pi'))$，我们需要证明 $\mathcal{B}(\pi) = \mathcal{B}(\pi')$。不失一般性，我们固定 $O(\text{last}(\pi)) = O(\text{last}(\pi')) = \langle s, J\rangle$ 以及 $\pi = \langle s_0, i, J_0\rangle a_0 \langle s_1, i, J_1\rangle \dots \langle s_k, i, J_k\rangle$ 和 $\pi' = \langle s'_0, i', J'_0\rangle a'_0 \langle s'_1, i', J'_1\rangle \dots \langle s'_l, i', J'_l\rangle$。这意味着 $J_k = J'_l = J$ 和 $s_k = s'_l = s$。

由于 π 和 π' 是 BOMDP 中的路径，我们知道

$$p(\langle s_k, i, J_k\rangle, a_k, \langle s_{k+1}, i, J_{k+1}\rangle) > 0$$

对所有 k，这意味着 $J_{k+1} = \text{Up}(J_k, s_k, a_k, s_{k+1})$。换句话说，$J$ 等于反复应用 Up，这正是 $\mathcal{B}(\pi)$ 的定义。□

C.3 定理 1 的证明

回顾定理 1:

定理 1. 多环境马尔可夫决策过程 $\mathcal{N}$ 存在以目标状态T 为目标的获胜策略，当且仅当 BOMDP $\mathcal{G}_{\mathcal{N}}$ 存在以目标状态 $T_{\mathcal{G}_{\mathcal{N}}}$为目标的获胜策略。

证明。根据 [41, 定理 20]中给出的一项更一般的结果即可得出这一结论。我们注意到，他们对 BOMDP 的定义略有不同，但两者等价，因为观测函数只揭示状态，这与多环境马尔可夫决策过程类似。该定理指出，对于给定状态 s，若某个策略在每个环境 $j \in J$ 中都能实现 Rabin 目标，则该策略在 BOMDP 状态 $\langle s, J\rangle$中也能获胜，反之亦然（应用必要的提升）。不失一般性，我们可以假设每个目标状态都是汇状态，这意味着可达性可直接化约为 Rabin 目标，因此我们可以将 [41, 定理 20] 应用于我们的情形。□

C.4 引理 4 的证明

回顾引理 4:

引理 4. 对于任意具有状态 S 和目标状态 T 的 BOMDP $\mathcal{G}_{\mathcal{N}}$，以及所有边界观测 $F \subseteq Z$，均有：$\text{Win}^{T}_{\mathcal{G}_{\mathcal{N}}|F} \subseteq \text{Win}^{T}_{\mathcal{G}_{\mathcal{N}}} \subseteq \text{Win}^{T\cup F}_{\mathcal{G}_{\mathcal{N}}|F}$。

Proof.
($\text{Win}^{T}_{\mathcal{G}_{\mathcal{N}}|F} \subseteq \text{Win}^{T}_{\mathcal{G}_{\mathcal{N}}}$)：假设 $s \in \text{Win}^{T}_{\mathcal{G}_{\mathcal{N}}|F}$，我们需要证明 $s \in \text{Win}^{T}_{\mathcal{G}_{\mathcal{N}}}$。设策略 σ 在 $\mathcal{G}_{\mathcal{N}}|F$ 中是必胜的，这意味着到达 F 中某个状态的概率为 0。因此，所有路径都留在一个与 $\mathcal{G}_{\mathcal{N}}$ 相等的“子 BOMDP”中，这表明对于 $\mathcal{G}_{\mathcal{N}}$ 中以 T 为目标集合的 s，σ 也同样是必胜的，即 $s \in \text{Win}^{T}_{\mathcal{G}_{\mathcal{N}}}$。

($\text{Win}^{T}_{\mathcal{G}_{\mathcal{N}}} \subseteq \text{Win}^{T\cup F}_{\mathcal{G}_{\mathcal{N}}|F}$)：假设 $s \in \text{Win}^{T}_{\mathcal{G}_{\mathcal{N}}}$，我们需要证明 $s \in \text{Win}^{T\cup F}_{\mathcal{G}_{\mathcal{N}}|F}$。设策略 σ 在以 T 为目标集合的 $\mathcal{G}_{\mathcal{N}}$ 中是必胜的。由 σ 产生的任意路径几乎必然会到达 T 中的某个状态。沿着这样的路径，我们可能会访问 F 中的某个状态，而该状态在以 $T \cup F$ 为目标集合的 $\mathcal{G}_{\mathcal{N}}|F$ 中是必胜的。$\mathcal{G}_{\mathcal{N}}|F$ 除了 F 中的状态外，与 $\mathcal{G}_{\mathcal{N}}$ 相等，因此，对于 $\mathcal{G}_{\mathcal{N}}|F$ 中以 $T \cup F$ 为目标集合的状态 s，σ 也同样是必胜的，即 $s \in \text{Win}^{T\cup F}_{\mathcal{G}_{\mathcal{N}}|F}$。□

C.5 引理 5 的证明

回顾引理 5:

引理 5. 对于具有环境图 GE 的多环境马尔可夫决策过程 $\mathcal{N}$，如果在 $\mathcal{G}_{\mathcal{N}}$ 中存在路径 $\pi = \pi_1 \dots \pi_n$，且每个 π_i 都是 J_i-局部的，则存在路径 $J_1 \dots J_n$。

证明。给定路径 $\pi = \pi_1 \dots \pi_n$ in $\mathcal{G}_{\mathcal{N}}$s.t. 每个 π_i 都是 J_i-局部的。我们固定 $\pi_i = \langle s_{i,\,0},\ j,\ J_i\rangle a_{i,\,0} \dots \langle s_{i,\,k},\ j,\ J_i\rangle$。我们需要证明存在一条路径 $J_1 \dots J_n$ in $\text{GE}_{\mathcal{N}}$。由于每个路径片段 π_i 都是 J_i-局部的，因此必有 $J_{i+1} = \text{Up}(J_i,\ s_{i,\,k},\ a_{i,\,k},\ s_{i+1,\,0})$，这意味着存在一条从 J_i 到 J_{i+1} 的转移，且该转移属于 $\text{GE}_{\mathcal{N}}$。□

C.6 引理 8 的证明

回顾引理 8:

引理 8。考虑一个多环境马尔可夫决策过程 $\mathcal{N}$ 和一个环境子集 J。

$$O\left(\mathsf{Win}^{T'_{\mathcal{G}_\mathcal{N}}}_{\mathrm{Loc}\mathcal{G}(J)}\right)\cap Z_J \;=\; O\left(\mathsf{Win}^{T_{\mathcal{G}_\mathcal{N}}}_{\mathcal{G}_\mathcal{N}}\right)\cap Z_J \quad \textit{with} \quad T'_{\mathcal{G}_\mathcal{N}} = T_{\mathcal{G}_\mathcal{N}} \cup (\mathsf{Win}^{T_{\mathcal{G}_\mathcal{N}}}_{\mathcal{G}_\mathcal{N}} \setminus S_J).$$

证明。我们证明，对于所有状态 s

$$\langle s, J\rangle \in O\left(\mathsf{Win}^{T'_{\mathcal{G}_\mathcal{N}}}_{\mathrm{Loc}\mathcal{G}(J)}\right) \Leftrightarrow \langle s, J\rangle \in O\left(\mathsf{Win}^{T_{\mathcal{G}_\mathcal{N}}}_{\mathcal{G}_\mathcal{N}}\right)$$

这便推出了该引理。
(⇒) 假设 $\langle s, J\rangle \in O\left(\mathsf{Win}^{T'_{\mathcal{G}_\mathcal{N}}}_{\mathrm{Loc}\mathcal{G}(J)}\right)$，我们知道存在一个基于观测的策略 σ，它对观测 $\langle s, J\rangle$ 是获胜的。这意味着，采用 σ 时，要么几乎必然到达 $T_{\mathcal{G}_\mathcal{N}}$ 中的某个状态 s'，要么到达 $\mathsf{Win}^{T_{\mathcal{G}_\mathcal{N}}}_{\mathcal{G}_\mathcal{N}} \setminus S_J$ 中的某个状态。如果路径到达 $T_{\mathcal{G}_\mathcal{N}}$ 中的状态，那么这条路径在 $\mathcal{G}_\mathcal{N}$ 中也会获胜。否则，这条路径会到达 $\mathsf{Win}^{T_{\mathcal{G}_\mathcal{N}}}_{\mathcal{G}_\mathcal{N}} \setminus S_J$ 中的某个状态 s'，而该状态在 $\mathcal{G}_\mathcal{N}$ 中有某个获胜策略 σ'。这表明，我们可以为 $\mathcal{G}_\mathcal{N}$ 构造一个在 s 中获胜的新策略：在前沿状态 s' 切换到 σ'。
(⇐) 假设 $\langle s, J\rangle \in O\left(\mathsf{Win}^{T_{\mathcal{G}_\mathcal{N}}}_{\mathcal{G}_\mathcal{N}}\right)$，我们知道存在一个基于观测的策略 σ，它对于观测 $\langle s, J\rangle$ 是必胜的。该策略对于目标集合为 $T'_{\mathcal{G}_\mathcal{N}}$ 的情形也必然是必胜的 $\mathrm{Loc}\mathcal{G}(J)$。如果该策略生成的路径始终处于 S_J 内，则它几乎必然会到达 $T_{\mathcal{G}_\mathcal{N}}$ 中的某个目标状态（因为该策略是必胜的）。如果路径离开 S_J，则说明前沿状态 $s' \in S \setminus S_J$ 必然是必胜的，即 $s' \in \mathsf{Win}^{T_{\mathcal{G}_\mathcal{N}}}_{\mathcal{G}_\mathcal{N}} \setminus S_J$。因此，$\sigma$ 在 $\mathrm{Loc}\mathcal{G}(J)$ 中几乎必然到达目标集合 $T'_{\mathcal{G}_\mathcal{N}}$。□

C.7 引理 9 的证明

回顾引理 9:

引理 9. 设 $\mathcal{N}$ 是一个具有状态 S 和动作 A 的多环境马尔可夫决策过程。$\mathrm{Loc}\mathcal{G}(J)$ 至多有 $\mathcal{O}(|S|^2 \cdot |A| \cdot |J|)$ 个状态和 $\mathcal{O}(|S|^2 \cdot |A| \cdot |J|^2)$ 个转移[9]。

证明。回顾局部BOMDP的定义，即（一个片段）

$$\mathrm{Loc}\mathcal{G}(J) = \mathsf{ReachFragment}(\mathcal{G}_{\mathcal{N}_{\downarrow J}} | F, \mathsf{unif}(S_J)) \quad \text{where} \quad F = Z \setminus Z_J\ .$$

$|S| \cdot |J|$ 在 S_J中有许多状态。在这些状态中，每个状态最多启用 $|A|$ 个动作。对于多环境马尔可夫决策过程中的每个状态-动作对，我们最多可以到达$|S|$ 个状态。特别是，对于每个 $\langle s、\ j、\ J\rangle$ 和动作 a，如果在局部BOMDP中，从 $\langle s、\ j、\ J\rangle$可以到达$\langle s'、\ j、\ J'\rangle$ 以及 $\langle s'、\ j、\ J''\rangle$，那么$J' = J''$。因此共有 $\mathcal{O}\left(|S|^2 \cdot |A| \cdot |J|\right)$ 个状态。然而，转移数量会重复计算，因此共有 $\mathcal{O}\left(|S|^2 \cdot |A| \cdot |J|^2\right)$ 个转移。□

[9] 转移的数量是 p 中非零条目的数量

D PSPACE 困难性证明（引理 11）

回顾引理 11：

引理 11. MEMDP 判定问题是 PSPACE 困难的。

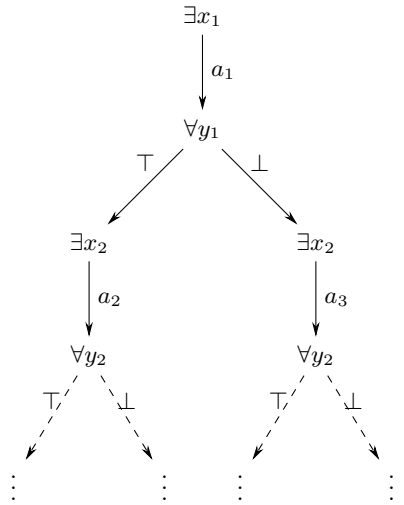

图 6： Ψ 的解树

全称量化布尔公式的解树。在给出证明之前，我们先指出，对于上述形式的每个量化布尔公式 Ψ，其真值可以递归地表示如下：

$$
\begin{aligned}
\Psi &= E_1([x_1 \mapsto \top]) \vee E_1([x_1 \mapsto \bot]) \text{ where} \\
E_1(m) &= F_1(m \cdot [y_1 \mapsto \top]) \wedge F_1(m \cdot [y_1 \mapsto \bot]) \text{ where} \\
F_1(m) &= E_2(m \cdot [x_2 \mapsto \top]) \vee E_2(m \cdot [x_2 \mapsto \bot]) \text{ where} \\
&\vdots \\
F_n(m) &= \Phi[m]
\end{aligned}
$$

其中，$\Phi[m]$ 表示应用替换 m 后 Φ 的真值。这表明，属于全称量化布尔公式的公式（例如 m）的见证呈树状结构。图 6 展示了这样一个见证的示例，我们称之为解树。这种树具有以下性质：

– 树中的任意路径都按照 Ψ 所示的顺序，为每个变量恰好赋予 $\top$ 或 $\bot$ 中的一个值。– 每个 $\forall$ 变量都必须允许两种赋值：$\top$ 和 $\bot$。

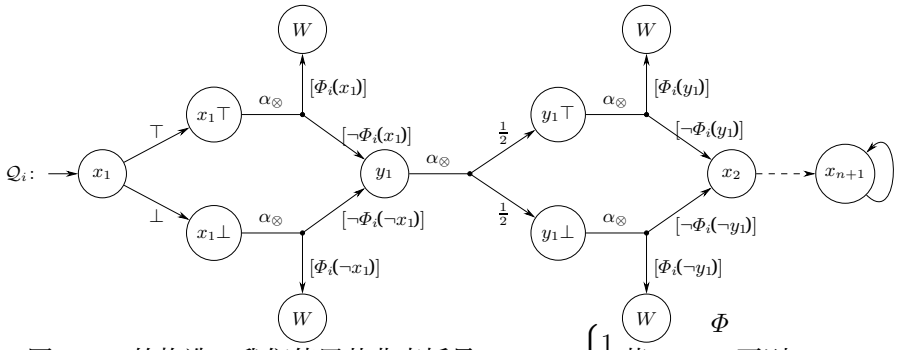

图 7: $\mathcal{Q}$的构造。我们使用艾弗森括号，$[\Phi] = \begin{cases} 1 & \text{若} \Phi \\ 0 & \text{否则} \end{cases}$

– 每个 $\exists$ 变量都必须允许 $\top$ 或 $\perp$ 中的一种赋值。– 任意路径上的赋值都满足 Φ。

构造 给定一个量化布尔公式

$$\Psi = \exists x_1 \forall y_1 \exists x_2 \forall y_2 \ldots \exists x_n \forall y_n \left[\Phi\right]$$

关于变量 $V = \{x_1, y_1, x_2, y_2, \ldots, x_n, y_n\}$，其中 Φ 是关于V的布尔公式。我们假设 Φ 处于合取范式，即它是正文字和负文字析取的合取。合取中的第 i 个子句记为 Φ_i。

为了判定 Φ 是否可满足，我们构造一个多环境马尔可夫决策过程，其中每个子句对应一个环境：更确切地说，如图7所示，多环境马尔可夫决策过程 $\mathcal{Q} = (S, A, s_\iota, \{p_i\}_{i\in[n]})$ 的状态空间为：

$$S = \{W, x_{n+1}\} \cup \left(\bigcup_{i\in[n]} \{x_i, x_i\top, x_i\perp, y_i, y_i\top, y_i\perp\} \right),$$

动作 $A = \{\top, \perp, \alpha_\otimes\}$ 和初始状态 $s_\iota = x_1$ 如正文所述。我们分部分定义 $\{p_i\}_{i\in[n]}$，对所有 $i, j \in [n], v \in \{x_j, y_j\}$：

- $p_i(x_j, \top) = [x_j\top \mapsto 1]$,
- $p_i(x_j, \perp) = [x_j\perp \mapsto 1]$,
- $p_i(y_j, \alpha_\otimes) = [y_j\top \mapsto 1/2, y_j\perp \mapsto 1/2]$,
- $p_i(v\top, \alpha_\otimes) = \left[W \mapsto [\Phi_i(v)], \mathsf{nextvar}(v) \mapsto [\neg\Phi_i(v)]\right]$,
- $p_i(v\perp, \alpha_\otimes) = \left[W \mapsto [\Phi_i(\neg v)], \mathsf{nextvar}(v) \mapsto [\neg\Phi_i(\neg v)]\right]$,
- $\ldots$ otherwise, $p_i(s, a) = [x_{n+1} \mapsto 1]$,

其中 $\mathsf{nextvar}(x_j) = y_j$ 和 $\mathsf{nextvar}(y_j) = x_{j+1}$。

该归约可在多项式时间内完成。

我们将目标集合固定为 $\{W\}$。现在证明以下命题，它与原定理等价： $\mathcal{Q}$ 具有获胜策略 $\Leftrightarrow$ Ψ 是一个真公式。

($\Rightarrow$) 假设我们能够得到一个获胜的确定性策略 $\sigma\colon (S\times A)^*\times S\to A$，则必须证明 Ψ 是一个真公式。现在我们说明， σ 可以转换为一棵见证解树。根据 $\mathcal{Q}$ 的定义以及获胜策略的定义， σ 和 $\mathcal{Q}$ 具有以下性质：

– 由于 σ 是必胜的，由 σ 产生的每条路径最终都会到达 W，因此在所选赋值下满足 Φ 。– 沿每条路径，对于每个变量 v，我们都会到达 $v\top$ 或 $v\bot$ 。– σ 无法阻止 $y\bot$ 或 $y\top$ ，因为 y 状态只有一个 $\alpha_\otimes$ 动作，它以相同的概率选择任一赋值。

换句话说，经过一些必要的转换后，由 σ 生成的所有路径构成一棵解树。将解树表示为一组赋值 m'_n:

$$
\begin{aligned}
m_1 &= \begin{cases} \{[x_1\mapsto\top]\} & \text{if } \sigma(x_1)=\top \\ \{[x_1\mapsto\bot]\} & \text{if } \sigma(x_1)=\bot \end{cases} \\
m'_1 &= \{m\cdot[y_1\mapsto\top \mid m\in m_1]\}\cup\{m\cdot[y_1\mapsto\bot]\mid m\in m_1\} \\
m_{i+1} &= \begin{cases} \{m\cdot[x_{i+1}\mapsto\top]\mid m\in m'_i\} & \text{if } \sigma(x_{i+1})=\top \\ \{m\cdot[x_{i+1}\mapsto\bot]\mid m\in m'_i\} & \text{if } \sigma(x_{i+1})=\bot \end{cases} \\
m'_{i+1} &= \{m\cdot[y_{i+1}\mapsto\top \mid m\in m_{i+1}]\}\cup\{m\cdot[y_{i+1}\mapsto\bot]\mid m\in m_{i+1}\}
\end{aligned}
$$

($\Leftarrow$) 假设 Ψ 是一个真公式，我们知道必定存在某棵见证解树 $m_\circ$。我们可以按如下方式将这棵解树转换为一个获胜策略 σ ：

$$
\sigma(x_1^{a_1}y_1^{a_2}\dots x_i) = \begin{cases} \top & \text{if } \exists p\colon [x_1\mapsto a_1, y_1\mapsto a_2,\dots,x_i\mapsto\top]\cdot p\in m \\ \bot & \text{if } \exists p\colon [x_1\mapsto a_1, y_1\mapsto a_2,\dots,x_i\mapsto\bot]\cdot p\in m \end{cases}
$$

$$
\text{otherwise}
$$

$$
\sigma(\pi)=\alpha_\otimes
$$

这是必胜的，因为根据解树的性质，由此生成的任意路径都满足 Φ 中的所有子句，这进而意味着在 Q 中必胜。

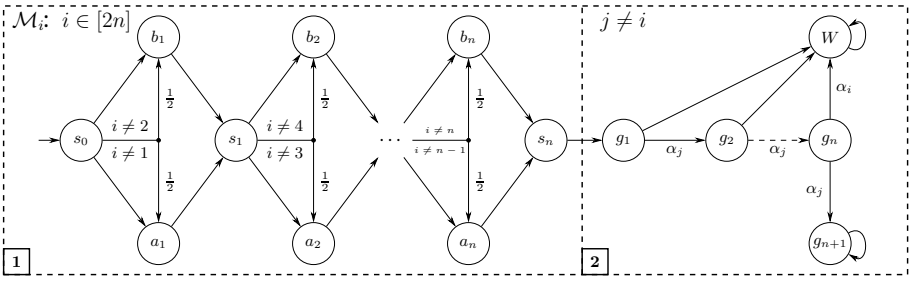

图8：获胜策略的指数级内存需求见证（为方便起见，重复展示图4）

E 指数级策略（定理 4 的证明）

回顾定理 4：

定理 4. 存在一族多环境马尔可夫决策过程 $\{\mathcal{N}^n\}_{n\geq 1}$，其中对于每个 n， $\mathcal{N}^n$ 都有 $2n$ 个环境和 $\mathcal{O}(n)$ 个状态，并且 $\mathcal{N}^n$ 的每个获胜策略都至少需要 2^n 个记忆状态。

证明。我们使用图 8 中给出的多环境马尔可夫决策过程族作为见证来证明这一点。我们如下定义该族的一个成员 $\mathcal{N}^n = (S, A, s_\iota, \{p_i\}_{i\in[2n]})$：

- $S = \{s_0, W, g_{n+1}\} \cup \bigcup_{i\in[n]}\{s_i, a_i, b_i\} \cup \{g_i \mid i \in [n]\}$
- $A = \{\alpha_\otimes\} \cup \{\alpha_i \mid i \in [n]\}$
- $s_\iota = s_0$
- 对于所有 i、$k \in [2n]$、$j \in [n]$，我们分部分定义 $\{p_i\}_{i\in[2n]}$：
  - $p_i(s_j, \alpha_\otimes) = \begin{cases} [a_{j+1} \mapsto 1] & \text{if } i = 2j+1 \\ [b_{j+1} \mapsto 1] & \text{if } i = 2j+2 \\ [a_{j+1} \mapsto 1/2, b_{j+1} \mapsto 1/2] & \text{otherwise} \end{cases}$
  - for all $s \in \{a_j, b_j\}$: $p_i(s, \alpha_\otimes) = [s_j \mapsto 1]$
  - $p_i(g_j, \alpha_k) = \begin{cases} [W \mapsto 1] & \text{if i=k} \\ [g_{j+1} \mapsto 1] & \text{otherwise} \end{cases}$

我们证明，任意获胜策略的规模都随 n 呈指数型增长。

观察 $p_i(s_j, \alpha_\otimes)$ 的定义可知，对于每个 s_j，大多数环境都会以相等的概率转移到 a_{j+1} 和 b_{j+1}。恰好有一个环境只允许转移到状态 a_{j+1} $(i = 2j+1)$，另一个环境只允许转移到状态 b_{j+1} $(i = 2j+2)$。因此，对于环境 i中的路径 π，我们有：

$$\left[a_j \in \pi \Rightarrow \pi \notin \textsc{Path}_{\mathcal{N}_{2j+2}}\right] \wedge \left[b_j \in \pi \Rightarrow \pi \notin \textsc{Path}_{\mathcal{N}_{2j+1}}\right] \tag{1}$$

对于任意路径 π，满足 $\mathsf{last}(\pi) = s_n$，我们都能得到一个可能的环境集合B_π（注意，MEMDP 的这一部分不存在非确定性）。

$$B_\pi = \{i \in [2n] \mid \forall \sigma.\ \Pr_{\mathcal{N}_i^n}(\pi \mid \sigma) > 0\}$$

根据方程（1），我们有 $|B_\pi| = n$，且从 π 到 B_π 的映射是一一对应的。因此，由于有 2^n 条可能的路径通向 s_n，也就有 2^n 个可能的环境集合对应于 s_n。

模型的最后一部分要求“猜测”出正确的环境。在环境 i 中，唯一的获胜动作是 α_i。如果在 n 次猜测次数内未能猜对，就会到达失败死锁状态。

因此，为了保证获胜，策略在观察到 π后执行的动作必须是 $\{\alpha_i \mid i \in B_\pi\}$的某种排列。现在反设存在某个策略 σ，它由有限状态控制器 $C = (Q, A, \delta, a, q_l)$表示，且具有 $|Q| < 2^n$。那么，由于有 2^n 条路径到达状态 s_n，必定存在两条路径 $\pi \neq \pi'$，使得 $\mathsf{last}(\pi) = \mathsf{last}(\pi') = s_n$，并且 $\delta^+(\pi) = \delta^+(\pi')$。由有限状态控制器策略的定义可知，对于所有 $\tau \in (A \times S)^*$，都有 $\sigma(\pi \cdot \tau) = \sigma(\pi' \cdot \tau)$。由于 $\pi \neq \pi'$，我们知道 $B_\pi \neq B_{\pi'}$。再结合 $|B_\pi| = |B_{\pi'}| = n$这一事实，我们知道 B_π 和 $B_{\pi'}$ 不可能互为排列，因此关于 σ 是获胜策略的假设不成立。

□

F 第5节的证明

F.1 引理13的证明

回顾引理13：

引理13. 若存在具有目标状态 T_Z的 BSG $\mathcal{B}_\mathcal{N}$ 中的获胜策略，则具有目标状态 T 的（无环）多环境马尔可夫决策过程 $\mathcal{N}$ 是获胜的。

证明。(⇒)：假设 $\mathcal{N}$ 是无环的，则 $\mathcal{B}_\mathcal{N}$ 也是无环的。如果我们有一个在 $\mathcal{N}$中必胜的策略 σ，我们证明它在 $\mathcal{B}_\mathcal{N}$中也必胜。σ 必胜意味着在每个环境中都是必胜的 $\mathcal{N}_i$。由于 $\mathcal{N}$，我们知道 σ 会在有限次转移后到达目标状态。现在反设 σ 对 $\mathcal{B}_\mathcal{N}$而言并非必胜，那么存在某条有限的轨迹$\pi = s_{0,\ 1}\ a_0\ s_{0,\ 2}\ i_0\ s_{1,\ 1}\ \ldots\ i_n\ s_{n,\ 1}$其中$s_{n,\ 1} \notin T$。由于信念支持集的单调性，我们知道玩家2选择的环境 i_n 可以为玩家2的每个状态选取，即 $i_k = i_n$ 对所有 $k \in \pi$均成立。这描述了根据定义， $\mathcal{B}_\mathcal{N}$ 也是 $\mathcal{N}_i$中存在的一条路径，这与假设 σ 是必胜的相矛盾。
（⇐）：假设存在一个针对 $\mathcal{B}_\mathcal{N}$的获胜策略 σ，我们可以将 σ 转换为一个适用于 $\mathcal{N}$ 的策略 σ'，方法是投影掉玩家2的状态。如果我们考虑 $\mathcal{B}_\mathcal{N}$ 中的玩家2策略，并且这些策略只选择一个环境，那么诱导出的马尔可夫决策过程恰好就是该环境。我们知道， σ 能够战胜每一种玩家2策略，因此它在每个环境中都是必胜的。□

F.2 引理14的证明

回顾引理 14：
引理 14。给定 $W \subseteq \mathsf{Win}_{\mathcal{B}_\mathcal{N}}^{T_{\mathcal{B}_\mathcal{N}}}$和 $L \subseteq S \setminus \mathsf{Win}_{\mathcal{B}_\mathcal{N}}^{T_{\mathcal{B}_\mathcal{N}}}$：$\mathsf{Win}_{\mathcal{B}_\mathcal{N}}^{T_{\mathcal{B}_\mathcal{N}}} = \mathsf{Win}_{\mathcal{B}_\mathcal{N}|W\cup L}^{T_{\mathcal{B}_\mathcal{N}}\cup W}$

证明。该定理等价于：$s \in \mathsf{Win}_{\mathcal{B}_\mathcal{N}}^{T_{\mathcal{B}_\mathcal{N}}} \Leftrightarrow s \in \mathsf{Win}_{\mathcal{B}_\mathcal{N}|W\cup L}^{T_{\mathcal{B}_\mathcal{N}}\cup W}$。我们将证明分为两部分。

(⇒)：假设 $s \in \mathsf{Win}_{\mathcal{B}_\mathcal{N}}^{T_{\mathcal{B}_\mathcal{N}}}$，我们证明 $s \in \mathsf{Win}_{\mathcal{B}_\mathcal{N}|W\cup L}^{T_{\mathcal{B}_\mathcal{N}}\cup W}$。根据我们的假设，存在一个在 $\mathcal{B}_\mathcal{N}$ 中针对 s 的获胜策略 σ。因此， σ 必然在以 $T_{\mathcal{B}_\mathcal{N}} \cup W$ 为目标集合的 $\mathcal{B}_\mathcal{N}|W \cup L$ 中必胜。σ 产生的所有路径在 $\mathcal{B}_\mathcal{N}|W \cup L$ 中都完全相同，除了那些包含 W 或 L 中某个状态的路径。如果一条路径到达 W 状态，那么它显然是必胜的。现在反设某条路径可以包含 L 中的状态。这意味着 σ 能够离开获胜区域，这与我们关于 σ 必胜的假设矛盾。(⇐)：假设 $s \in \mathsf{Win}_{\mathcal{B}_\mathcal{N}|W\cup L}^{T_{\mathcal{B}_\mathcal{N}}\cup W}$，我们证明 $s \in \mathsf{Win}_{\mathcal{B}_\mathcal{N}}^{T_{\mathcal{B}_\mathcal{N}}}$。我们采用与引理 8 证明类似的论证。根据我们的假设，在以 $T_{\mathcal{B}_\mathcal{N}} \cup W$ 为目标集合的 $\mathcal{B}_\mathcal{N}|W \cup L$ 中，存在一个获胜策略 σ。我们可以通过扩展 σ，构造一个在 $\mathcal{B}_\mathcal{N}$ 中获胜的新策略。我们遵循 σ；如果到达状态 $s' \in W$，就可以切换到一个在 $\mathcal{B}_\mathcal{N}$ 中针对 s' 获胜的策略。□

G 基准数据

G.1 模型

我们将简要介绍基准测试中使用的各类多环境马尔可夫决策过程。

指数型。与定理 4 的证明中使用的多环境马尔可夫决策过程族相同。对于多环境马尔可夫决策过程 $\mathcal{N}_n$，状态数量约为 $4 \cdot n$，并包含 $2 \cdot n$ 个环境。

青蛙过河。一个大小为 $N \times M$ 的网格，智能体从 (1, 1) 出发，尝试通过四个基本方向的移动到达 (N, M)。沿此路径有一辆来回行驶的汽车。该汽车可能的起始位置数量由环境数量表示。状态数量约为 $2 \cdot N^2 \cdot M$。

吃豆人。一个大小为 $N \times M$ 的网格，智能体从 (1, 1) 出发，尝试通过四个基本方向的移动到达 (N, M)。从 $(1, M)$ 开始，网格中有一个幽灵在移动。对于智能体移动的每个方向，幽灵也会沿一个固定但最初未知的方向移动。对于智能体移动的四个基本方向中的每一个，幽灵都可以以四个基本方向中的任意一个作为响应，这由环境编码。因此，多环境马尔可夫决策过程最多有 $4^4 = 256$ 个可能的环境。

抓幽灵。与吃豆人系列相同，但目标是幽灵。
Mastermind。这类多环境马尔可夫决策过程是 Mastermind棋盘游戏的简化版本。智能体尝试猜出由一串彩色球组成的密码。具体来说，智能体要在 G 次猜测内，猜出一个长度为 N 的序列，其中的球可以是 C 种颜色中的任意一种。每次猜测后，智能体都能观察到猜对的球数（颜色和位置都正确）。每个环境对应于 (C^N) 种可能密码中的一种。如果猜对密码，就会到达目标状态。如果最后一次猜测不正确，就会到达失败状态。

无限信念。用于展示定量方法可能会“碰壁”的虚拟多环境马尔可夫决策过程。这是因为由该多环境马尔可夫决策过程构造的部分可观测马尔可夫决策过程具有无限信念空间表示。

网格。一个大小为 $N \times M$ 的网格，智能体从 (1, 1) 出发，尝试通过四个基本方向移动到 (N, M)。该网格中有一个洞，但其位置并不确定。如果智能体走进洞里，就无法脱身，因此无法到达目标状态。每个环境对应一个不同的洞的位置。智能体在距离洞一个格子时可以感知到危险。

不可满足变体。上述一些（可满足的）多环境马尔可夫决策过程有对应的不可满足版本。青蛙过河多环境马尔可夫决策过程中的汽车更宽，导致无法穿过街道。吃豆人多环境马尔可夫决策过程的特定尺寸设置，使得模仿智能体移动的幽灵总能

阻止智能体到达目标状态。Mastermind 和指数级多环境马尔可夫决策过程允许的猜测次数，比获得获胜策略所需的次数少一次。

G.2 启发式方法

这里讨论基准测试中测试过的所有启发式方法。对于每种启发式方法，我们都测试了三种不同的变体。这些变体与我们在 PAGE 算法中计算下界、上界或两者有关。在表 2 至表 5 中，我们分别用 -、 + 和 -+ 表示这些变体。图 9 和图 10 展示了使用不同启发式方法的效果。

BFS。广度优先搜索启发式方法，按照首次发现博弈状态的顺序进行探索。DFS。深度优先搜索启发式方法，按照最后发现博弈状态的顺序进行探索。熵。正熵启发式方法，优先探索环境集合最小的博弈状态。负熵。负熵启发式方法，优先探索环境集合最大的博弈状态。

G.3 基准测试结果

	$\|I\|$	$\|S\|$	$\|A\|$	BFS-	BFS+	广度优先搜索-+	DFS-	DFS+	深度优先搜索-+	比较
Exponential	8	19	9	0.1	0.1	0.1	0.1	0.1	0.1	0.1
	12	27	13	0.2	0.2	0.2	0.2	0.2	0.2	0.2
	16	35	17	1.5	2.9	1.6	1.7	2.6	2.4	2.7
	20	43	21	36.1	39.2	41.6	40.1	129.6	129.1	39.4
	24	51 25		MO	MO	MO	MO	TO	TO	MO
Frogger	10	1200	4	0.2	0.2	0.2	0.2	0.2	0.2	20.0
	20	1200	4	0.4	0.4	0.4	0.4	0.4	0.4	MO
	29	1200	4	0.6	0.6	0.6	0.6	0.6	0.6	MO
	50	4000	4	2.7	2.7	2.7	2.6	2.6	2.6	MO
	80	4000	4	4.5	4.4	4.4	4.4	4.4	4.3	MO
	99	4000	4	6.1	6.0	6.2	6.0	5.9	5.9	MO
CMan	256	625	4	5.9	7.0	5.8	6.8	6.7	6.6	7.0
	256	6561	4	39.5	43.2	39.9	40.6	40.6	40.0	46.2
	256	14641	4	85.5	86.5	86.0	84.1	82.7	82.6	97.1
PMan	256	256	4	3.9	4.2	3.8	4.0	4.2	3.7	4.2
	256	1296	4	11.5	11.4	11.3	11.2	10.7	10.8	12.6
	256	4096	4	23.2	30.1	23.1	27.5	28.8	27.6	30.8
Grid	9	122	4	0.1	0.1	0.1	0.1	0.1	0.1	0.1
	19	132	4	0.6	1.2	0.6	0.5	1.1	0.5	1.0
	39	152	4	6.4	MO	9.0	119.5	MO	145.1	MO
	79	192	4	MO	MO	MO	MO	MO	MO	MO
	99	374	4	MO	MO	MO	MO	MO	MO	MO
	149	424	4	MO	MO	MO	MO	MO	MO	MO
	199	474	4	MO	MO	MO	MO	MO	MO	MO
Inf	5	4	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MMind	16	26	16	0.2	0.2	0.2	0.2	0.2	0.2	0.2
	27	21	27	0.5	0.9	0.5	0.4	0.7	0.3	0.8
	32	37	32	1.5	1.7	1.6	0.8	1.9	0.9	2.1
	81	31	81	50.1	166.6	49.5	8.2	TO	12.8	190.8

表 2：可满足基准测试结果（第 1 部分）

	$\|I\|$	$\|S\|$	$\|A\|$	Ent-	Ent+	熵-+	熵-	熵+	熵-+	SAT	信念
Exponential	8	19	9	0.1	0.1	0.1	0.0	0.1	0.1	75.9	0.1
	12	27	13	0.2	0.2	0.2	0.2	0.2	0.2	TO	0.2
	16	35	17	1.8	2.5	2.4	2.7	7.3	7.5	TO	4.1
	20	43	21	40.1	132.0	131.4	34.5	196.3	197.6	TO	217.6
	24	51 25		MO	TO	TO	MO	MO	MO	TO	MO
Frogger	10	1200	4	0.2	0.2	0.2	0.2	0.2	0.2	1.4	22.7
	20	1200	4	0.4	0.4	0.4	0.4	0.4	0.5	3.9	MO
	29	1200	4	0.6	0.6	0.6	0.6	0.6	0.7	9.1	MO
	50	4000	4	2.6	2.6	2.6	2.6	2.6	2.6	121.2	TO
	80	4000	4	4.3	4.4	4.4	4.3	4.3	4.4	597.3	TO
	99	4000	4	6.0	6.0	5.9	5.8	5.9	6.1	TO	TO
CMan	256	625	4	6.6	6.6	6.6	5.8	7.2	5.9	TO	3.8
	256	6561	4	40.4	40.7	40.1	32.6	43.7	32.6	TO	9.1
	256	14641	4	84.0	82.4	82.5	65.4	88.8	65.3	TO	16.2
PMan	256	256	4	4.0	4.2	3.9	3.5	4.2	3.8	11.2	4.6
	256	1296	4	10.8	10.9	10.9	9.1	11.4	9.1	375.7	19.5
	256	4096	4	27.2	28.7	27.3	21.0	30.9	21.1	TO	203.5
Grid	9	122	4	0.1	0.1	0.0	0.1	0.1	0.1	0.2	0.2
	19	132	4	0.2	1.1	0.2	0.2	1.3	0.2	3.7	0.6
	39	152	4	0.4	MO	0.4	1.5	MO	1.6	121.3	12.6
	79	192	4	1.2	MO	1.1	MO	MO	MO	TO	MO
	99	374	4	4.9	MO	4.9	MO	MO	MO	TO	MO
	149	424	4	6.2	MO	6.2	MO	MO	MO	TO	MO
	199	474	4	6.3	MO	6.4	MO	MO	MO	TO	MO
Inf	5	4	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	MO
MMind	16	26	16	0.2	0.2	0.2	0.2	0.1	0.2	0.4	0.2
	27	21	27	0.6	0.7	0.7	0.6	0.7	0.7	TO	0.7
	32	37	32	1.6	1.9	1.6	1.5	1.6	1.5	TO	1.5
	81	31	81	153.8	TO	653.5	128.9	126.7	128.1	TO	250.4

表3：可满足基准测试结果（第2部分）

	$\|I\|$	$\|S\|$	$\|A\|$	广度优先搜索-	广度优先搜索+	广度优先搜索-+	深度优先搜索-	深度优先搜索+	深度优先搜索-+	比较
Frogger	5	360	4	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	10	360	4	0.1	0.1	0.1	0.1	0.1	0.1	4.2
	15	360	4	0.2	0.1	0.1	0.2	0.1	0.1	150.8
	29	360	4	0.4	0.3	0.3	0.3	0.3	0.3	MO
Pacman	256	81	4	2.8	2.7	2.6	2.9	2.6	2.7	3.1
	256	625	4	5.6	5.3	5.3	5.7	5.4	5.5	7.2
	256	6561	4	30.4	30.4	30.4	30.8	30.5	30.4	46.2
MMind	16	21	16	0.1	0.2	0.1	0.2	0.1	0.2	0.1
	27	17	27	0.5	0.5	0.5	0.5	0.5	0.5	0.6
	32	25	32	1.0	1.0	1.0	1.1	0.6	0.7	1.1
	81	21	81	37.4	37.3	37.4	43.2	56.2	56.0	76.8
Exponential	12	26	13	0.2	0.2	0.2	0.2	0.1	0.1	0.2
	16	34	17	1.3	1.4	1.4	1.4	0.2	0.2	2.2
	20	42	21	32.3	34.9	37.9	33.8	0.3	0.3	33.3
	24	50	25	MO	MO	MO	MO	0.9	1.0	MO
	28	58	29	MO	MO	MO	MO	8.8	9.1	MO
	32	66	33	MO	MO	MO	MO	425.1	429.5	MO

表 4：不可满足基准测试结果（第 1 部分）

	$\|I\|$	$\|S\|$	$\|A\|$	Ent-	Ent+	Ent-+	NEnt-	NEnt+	NEnt-+	可满足性	信念
Frogger	5	360	4	0.0	0.0	0.0	0.0	0.0	0.0	至	0.5
	10	360	4	0.1	0.1	0.1	0.1	0.1	0.1	TO	10.6
	15	360	4	0.1	0.1	0.2	0.2	0.2	0.2	至	744.3
	29	360	4	0.3	0.3	0.3	0.3	0.3	0.3	TO	MO
Pacman	256	81	4	2.8	2.6	2.6	2.7	2.6	2.5	TO	3.2
	256	625	4	5.5	5.3	5.2	5.6	5.2	5.0	TO	11.0
	256	6561	4	30.5	30.3	30.2	30.7	30.1	29.9	TO	TO
MMind	16	21	16	0.1	0.1	0.1	0.1	0.1	0.2	TO	0.3
	27	17	27	0.5	0.5	0.5	0.5	0.5	0.5	TO	2.0
	32	25	32	1.1	0.7	0.6	0.9	0.9	0.9	TO	4.2
	81	21	81	41.9	47.1	41.1	39.8	38.7	38.6	至	MO
Exponential	12	26	13	0.2	0.1	0.1	0.2	0.2	0.2	TO	0.7
	16	34	17	1.4	0.2	0.2	1.3	1.3	1.4	TO	10.8
	20	42	21	33.8	0.8	0.8	31.0	163.5	173.8	TO	576.1
	24	50	25	MO	7.8	8.3	MO	MO	MO	TO	MO
	28	58	29	MO	329.7	334.3	MO	MO	MO	TO	TO
	32	66	33	MO	343.7	347.7	MO	MO	MO	TO	MO

表 5：不可满足基准测试结果（第 2 部分）

图表。

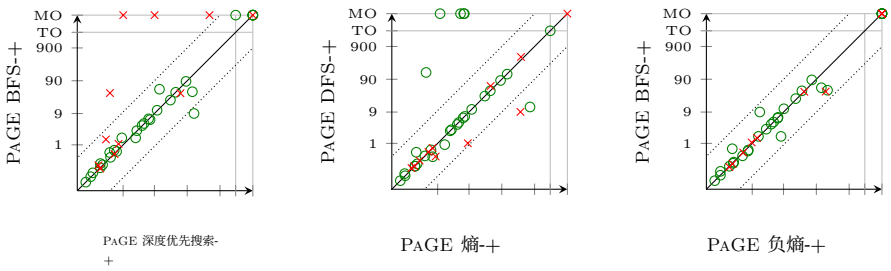

图 9： PAGE 启发式方法的相对性能

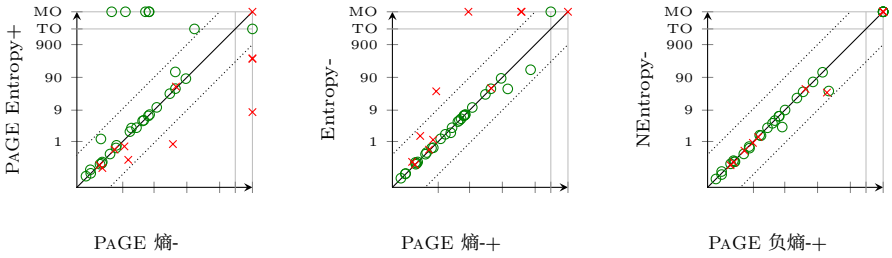

图10：使用下界、上界或两者的 PAGE 算法的相对性能