arXiv:2305.05739v1 [cs.LO] 9 May 2023

基于图的参数化与加权马尔可夫决策过程归约

Kasper Engelen、Guillermo A. Pérez 和 Shrisha Rao

安特卫普大学– Flanders Make，比利时 {kasper.engelen, guillermo.perez, shrisha.rao}@uantwerpen.be

摘要。 我们研究参数化马尔可夫决策过程中的加权可达性归约的计算复杂度。具体而言，如果对于多项式个不确定量的所有赋值，从 p 可达到的最大期望权重都大于从 q可达到的相同值，我们就称状态p 不劣于 q 。就计算复杂度而言，我们证明判定 p 是否不劣于 q 是 coETR 完全的。另一方面，我们给出一种多项式时间算法，用于计算我们所研究的序关系在马尔可夫链上的等价类。此外，我们描述并实现了两条推理规则，用于对不劣关系进行下近似，并通过实验证明，这些规则可作为一种高效的预处理步骤，用于分析大型马尔可夫决策过程。

关键词： 马尔可夫决策过程 · 敏感性分析 · 模型简化

1 引言

马尔可夫决策过程（简称 MDP）是有用的数学模型，可用于描述包含随机成分和离散（非确定性）选择的系统行为。在验证领域，它们被用作随机化算法、协议等的形式化模型进行研究。 [1,6] 在人工智能领域，MDP 为强化学习算法奠定了理论基础 [16,19]。

从验证领域来看，人们已经实现了用于分析 MDP 的高效工具，例如 PRISM [14], 、STORM [12], 和 Modest [9]。尽管这些工具能够处理规模更大、复杂度更高的 MDP，正如最新的《定量形式化模型分析工具比较》 [4], 所证实的那样，但它们仍受状态爆炸问题和最坏情况下值迭代的指数级复杂度所困扰。虽然大多数工具都能针对丰富的规范类别（线性时序逻辑、概率计算树逻辑等）检查 MDP，但它们通常会将任务简化为可达性分析，并使用某种形式的值迭代来求解 [10]。

在这项工作中，我们着重研究约简技术，以提升前述验证工具中所实现的马尔可夫决策过程分析算法的运行效率。我们尤其关注基于马尔可夫决策过程底层图结构的归约，

这类归约不使用具体的概率分布。这是因为，这种归约也适用于强化学习 [2] 中出现的部分已知马尔可夫决策过程。具有这些性质的若干图论和自动机理论约简技术已经存在。它们着重计算具有特定性质的状态，例如：极值状态（即到达目标状态集合的概率为 0 或 1 的状态）、极大末端分量（即其中的状态几乎必然能够相互到达的集合）以及本质状态 [7,5]（即在图论意义上构成某个给定状态与目标集之间分隔集的状态集合），等等。在 [15], 中，Le Roux 和 Pérez 在状态之间引入了一种偏序，称为 不劣关系（简称 NWR），它涵盖了上述所有概念。他们在该项工作中的主要成果是证明了 NWR 的自然判定问题——比较两个状态——是 NP的补类-完全的，并给出了一些推理规则，用于对完整的 NWR 进行下近似。

我们的贡献。在本文中，我们将从不更差关系扩展到了参数化马尔可夫决策过程中的加权可达性。形式上，我们为状态的目标集分配了权重，并允许将转移标记为多项式，而不仅仅是具体的概率值。于是，我们称状态 p 从不差于 q，如果对于不定元的所有赋值，从 p 可达的最大期望权重至少与从 q可达的相同值一样大。我们证明了判定 p 是否从不更差于 q 是**coETR**-完全的，因此它与判定实数存在理论中一个陈述的真值在多项式时间内可相互归约。在此过程中，我们证明了加权可达性的从不更差关系可以在多项式时间内归约为（布尔）可达性的从不更差关系。关于理论结果，我们进一步证明，对于一类特殊的参数化马尔可夫链，从不更差关系的等价类可以在多项式时间内计算出来；而对于参数化马尔可夫决策过程，这与从不更差关系判定问题同样困难。最后，我们显著改进了 Le Roux 和 Pérez 的推理规则（并纠正了其中的小错误），并提供了一个具体算法来使用这些规则，从而减小给定马尔可夫决策过程的规模。我们实现了该算法，并在定量验证基准集 [11]中的多个基准测试上评估了其性能。

2 预备知识

对于有向图 $\mathcal{G} = (V, E)$ 和顶点 $u \in V$，我们用 $uE = \{v \in V \mid (u, v) \in E\}$ 表示 u 的（直接）后继集合。汇点是满足 $uE = \emptyset$ 的顶点u 。

对于有限字母表 Σ， Σ^* 是由 Σ 构成的所有有限单词的集合，包括空单词 ε，而 Σ^+ 是所有非空单词的集合。

令 $X = \{x_1, \ldots, x_k\}$ 为一个有限变量集合， $p(x_1, \ldots, x_k)$ 为定义在 X 上、系数为有理数的多项式。对于变量的赋值 $\mathrm{val} : X \to \mathbb{R}$，我们用 $p[\mathrm{val}]$ 表示 p 在给定 $\mathrm{val}(X)$ 时的像，即 $p(\mathrm{val}(x_1) \ldots, \mathrm{val}(x_k))$。我们用 $p \equiv 0$ 表示 p 在语法上等于 0，并用 $p = 0$

表示其像在所有赋值下均为 0。最后，我们用 $\mathbb{Q}[X]$ 表示定义在 X 上的所有有理系数多项式的集合。

2.1 随机模型

有限集合 S 上的概率分布是一个函数 $f\colon S \to \mathbb{R}_{\geq 0}$，其中 $\sum_{s\in S} f(s) = 1$。

定义 1（马尔可夫链）。（加权）马尔可夫链是一个元组 $\mathcal{C} = (S, \mu, T, \rho)$，包含有限状态集合 S、目标状态集合 $T \subseteq S$、概率转移函数 $\mu : (S \setminus T) \times S \to \mathbb{R}_{\geq 0}$，其中 $\forall s \in S \setminus T\colon \sum_{s'\in S} \mu(s, s') = 1$（即 μ 将非目标状态映射到状态上的概率分布），以及权重函数 $\rho\colon T \to \mathbb{Q}$，它为目标状态分配权重。

注意，T 中的所有状态都是汇点：对于所有 $t \in T$、$s \in S$，$\mu(t, s)$ 均未定义。

由 $\mathcal{C}$ 构成的运行 是一个有限且非空的字 $\pi = s_0 \dots s_n$，其取自 S，并且对所有 $0 \leq i < n$，都有 $0 < \mu(s_i, s_{i+1})$。若 $s' = s_n$，则该运行 π 到达 $s' \in S$。与一次运行相关联的概率定义为 $\mathbb{P}_\mu(\pi) = \prod_{0\leq i<n} \mu(s_i, s_{i+1})$。

我们还定义到达一个状态集合的概率。下文中，我们用$\{\{x_1, \dots, x_n\}\}$ 表示由元素 x_1、...、x_n构成的多重集 。

定义 2（可达性）。给定一个马尔可夫链 $\mathcal{C} = (S, \mu, T, \rho)$、一个初始状态 $s_0 \in S$ 和一个状态集合 $B \subseteq S$，我们用 $\mathbb{P}^{s_0}_{\mathcal{C}}[\Diamond B]$ 表示从 s_0到达 B 的概率。若 $s_0 \in B$，则 $\mathbb{P}^{s_0}_{\mathcal{C}}[\Diamond B] = 1$；否则：

$$\mathbb{P}^{s_0}_{\mathcal{C}}[\Diamond B] = \sum \{\{\mathbb{P}_\mu(\pi) \mid \pi = s_0 \dots s_n \in (S \setminus B)^+ B\}\}.$$

为简洁起见，当 $B = \{a\}$ 时，我们写作 $\mathbb{P}^{s_0}_{\mathcal{C}}[\Diamond a]$，而非 $\mathbb{P}^{s_0}_{\mathcal{C}}[\Diamond\{a\}]$。

现在，我们使用目标状态的权重来定义加权可达性值。

定义 3（状态值）。给定一个马尔可夫链 $\mathcal{C} = (S, \mu, T, \rho)$ 和一个状态 $s_0 \in S$，s_0 的（期望奖励）值为 $\mathrm{Rew}^{s_0}_{\mathcal{C}} = \sum_{t\in T} \mathbb{P}^{s_0}_{\mathcal{C}}[\Diamond t] \cdot \rho(t)$。

最后，我们回顾参数化马尔可夫决策过程的定义。

定义 4（马尔可夫决策过程）。加权参数化马尔可夫决策过程（*wpMDP*）是一个元组 $\mathcal{M} = (S, A, X, \delta, T, \rho)$，其中包含有限的状态集合 S、动作集合 A、参数集合 X，以及目标状态集合 $T \subseteq S$。$\delta : (S \setminus T) \times A \times S \to \mathbb{Q}[X]$ 是一个将转移映射为多项式的概率转移函数，$\rho\colon T \to \mathbb{Q}$ 则是一个权重函数。

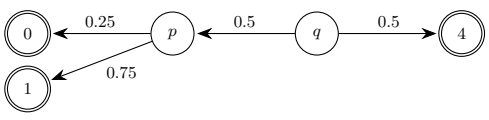


图 1。一个马尔可夫链的示例 $\mathcal{C}$，其中 $\mathrm{Rew}^p_{\mathcal{C}} = 0.75$ 和 $\mathrm{Rew}^q_{\mathcal{C}} = 2.375$。圆圈表示状态；箭头表示转移；双圆圈表示目标状态，其整数标签即为其权重。

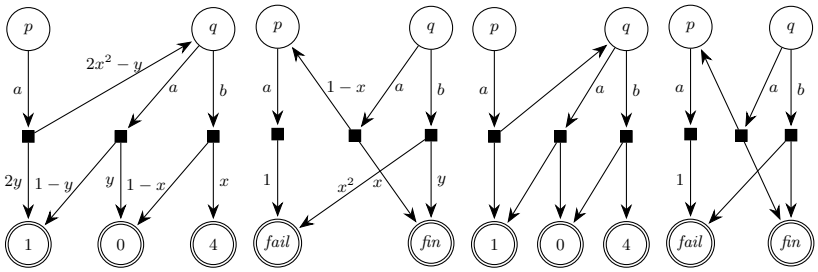
不失一般性，我们假设存在一个额外状态 fail $\in T$，且 ρ(fail) $= 0$。直观而言，这是一个我们希望避开的"坏"目标状态。在本文中，我们有时只指定部分概率转移函数。这种略微滥用记号的做法不失一般性，因为（就我们所有目的而言）可以令所有未指定的转移以概率 1 转移到 fail。I

n th本文中，我们将 ll work with the foll采用以下 bcl假设 f wpMDPs.

参数化马尔可夫决策过程，即无权重的参数化马尔可夫决策过程，是加权参数化马尔可夫决策过程的一个子类，其中在 T 中只有两个状态，分别是 $T = \{$终止状态和失败状态$\}$，其中 ρ(终止状态) $= 1$，且 ρ(失败状态) $= 0$。对于参数化马尔可夫决策过程，我们从定义它们的元组中省略 ρ，因为我们已知目标状态的权重。

加权˜参数化马尔可夫决策过程，即加权的平凡参数化 马尔可夫决策过程，对每个转移都具有一个唯一的变量作为多项式。也就是说，概率转移函数满足 $\delta(s,a,s') \equiv 0$ 或 $\delta(s,a,s') = x_{s,\ a,\ s'} \in X$，对于所有 $(s,a,s') \in (S\setminus T)\times A\times S$。由于这类加权参数化马尔可夫决策过程中的所有参数都是平凡的，因此我们省略它们。取而代之的是，一个加权˜参数化马尔可夫决策过程是一个元组 (S,A,Δ,T,ρ)，其中 $\Delta \subseteq (S\setminus T)\times A\times S$表示所有概率不为 0 的转移。

˜参数化马尔可夫决策过程 或（无权重）平凡参数化马尔可夫决策过程都既无权重又是平凡参数化的。因此，我们在它们的元组表示（S、 A、 Δ、 T）中省略 ρ 和 X，其中 Δ的定义与 w˜参数化马尔可夫决策过程相同。



(a) 一个加权参数化马尔可夫决策过程　(b) 参数化马尔可夫决策过程　(c) 一个 w˜参数化马尔可夫决策过程　(d) 一个 ˜参数化马尔可夫决策过程

图 2。我们所考虑的所有加权参数化马尔可夫决策过程子类的示例。实心方块表示状态-动作对（参见第 2.2 节）。

在本节余下部分，我们固定一个 wpMDP $\mathcal{M} = (S,A,X,\delta,T,\rho)$。

定义 5（图保持赋值，来源于 [13]）。如果对于所有 $s \in (S\setminus T)$ $s' \in S$， $a \in A$均满足以下条件，则赋值 val：$X \to \mathbb{R}$ 称为图保持赋值：

– 概率均为非负数： $\delta(s,a,s')[val] \geq 0$； – 出边概率构成一个分布：
$\sum_{s''\in S} \delta(s,a,s'')[val] = 1$； – $\delta(s,a,s') \not\equiv 0 \implies \delta(s,a,s')[val] \neq 0$。

所有此类保图赋值构成的集合记作 $\mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}$。

任意 wpMDP $\mathcal{M}$ 的一个保图赋值都会得到一个 MDP（具有实数概率！），我们将其称为 $\mathcal{M}[\mathsf{val}]$：具体而言，就是将每个参数替换为相应的实数，并计算每条转移上多项式的值。这样的 MDP 等价于文献 [1,15,13] 中出现的 MDP。

定义 6（策略）。一个（无记忆确定性）策略 σ 是一个函数 $\sigma\colon S\setminus T) \to A$，它将状态映射到动作。所有此类确定性无记忆策略的集合记为 $\Sigma^{\mathcal{M}}$。

带有赋值 val 和策略 σ 的加权参数化马尔可夫决策过程 $\mathcal{M}$ 诱导出一个马尔可夫链 $\mathcal{M}^{\sigma}_{\mathsf{val}} = (S, \mu, T, \rho)$，其中对于所有 $s \in (S \setminus T)$ $s' \in S$，$\mu(s,s') = \delta(s,\sigma(s)\ s')[\mathsf{val}]$。

注 *1*。注意，我们仅考虑无记忆确定性策略。这是因为我们关注的是针对给定赋值使期望奖励值最大化的策略。由于所有目标状态都是陷阱状态，因此可以证明，例如通过归约到定量可达性或期望平均收益 [1], ，对于所有赋值，都存在一个最优的无记忆确定性策略。

2.2 加权参数化马尔可夫决策过程的图

使用加权参数化马尔可夫决策过程的图表示很方便。在这方面，我们将状态 s（用圆形 ◯ 表示）和状态-动作对（s,a）（用方形 □ 表示）视为顶点。我们将 $S_N = (S \setminus T)$ 中的顶点称为 $\times A$ 自然顶点，并在全文中用它们表示状态-动作对。我们用 $V = S \cup S_N$ 表示加权参数化马尔可夫决策过程的顶点集合。因此，加权参数化马尔可夫决策过程的图 $\mathcal{G}(\mathcal{M}) = (V, E)$ 是一个二部图，其两个划分为 S 和 S_N。我们将边集合定义如下。

$$\begin{aligned} E = \{(s,(s,a)) \in S \times S_N \mid \exists s' \in S \setminus \{fail\} : \delta(s,a,s') \neq 0\} \\ \cup \{((s,a),s') \in S_N \times S \mid \delta(s,a,s') \neq 0\} \end{aligned}$$

也就是说，当自然顶点的元组的第一个分量是某个状态，且它能以非零概率到达除失败状态以外的某个状态时，该状态与此自然顶点之间存在一条边；如果对应转移上的多项式在语法上不为零，则自然顶点与该状态之间存在一条边。

设 $\mathsf{val} \in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}$ 为一个保图赋值，$\sigma \in \Sigma^{\mathcal{M}}$ 为一个策略。现在，我们定义相对于 val 和 σ，$\mathcal{G}(\mathcal{M})$ 中一个顶点的（奖励）值。

$$\mathrm{Rew}^{\sigma}_{\mathcal{M}[\mathsf{val}]}(v) = \begin{cases} \sum_{s' \in vE} \delta(s,a,s')[\mathsf{val}] \cdot \mathrm{Rew}^{\sigma}_{\mathcal{M}[\mathsf{val}]}(s') & \text{if } v = (s,a) \in S_N \\ \mathrm{Rew}^{v}_{\mathcal{M}^{\sigma}_{\mathsf{val}}} & \text{otherwise} \end{cases}$$

此外，我们用 $\mathrm{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(v)$ 表示值 $\max_{\sigma \in \Sigma^{\mathcal{M}}} \mathrm{Rew}^{\sigma}_{\mathcal{M}[\mathsf{val}]}(v)$，即选择使该顶点的值最大化的策略时的值。当上下文已明确所指的加权参数化马尔可夫决策过程时，我们使用 $\mathrm{Rew}^{\sigma}_{\mathsf{val}}$ 和 $\mathrm{Rew}^{*}_{\mathsf{val}}$。

注 *2.* 对于参数化马尔可夫决策过程和 ˜参数化马尔可夫决策过程，$\mathrm{Rew}^{\sigma}_{\mathcal{M}[\mathsf{val}]}(v) = \mathbb{P}^{v}_{\mathcal{M}^{\sigma}_{\mathsf{val}}}[\Diamond \text{终止}]$。读者只需查看奖励的定义，就能轻松验证这一点。

例 *1.* 在图 2a 中，赋值 $\mathsf{val} = \{x = 0.6, y = 0.28\}$ 是一个保持图结构的赋值。因此，我们得到 $\mathsf{Rew}^*_{\mathsf{val}}(q,a) = 1 - 0.28 = 0.72$ 且 $\mathsf{Rew}^*_{\mathsf{val}}(q,b) = 0.6(4) = 2.4$。因此，对于状态 q，最优策略将选择动作b，且 $\mathsf{Rew}^*_{\mathsf{val}}(q) = 2.4$。现在，我们得到 $\mathsf{Rew}^*_{\mathsf{val}}(p) = \mathsf{Rew}^*_{\mathsf{val}}(p,a) = 2(0.28) + (2(0.6)^2 - 0.28)\mathsf{Rew}^*_{\mathsf{val}}(q) = 1.616$。

2.3 不劣关系

设 $\mathcal{M} = (S, A, X, \delta, T, \rho)$ 是一个加权参数化马尔可夫决策过程，且 $\mathcal{G}(\mathcal{M}) = (V, E)$。现在，我们将采用不劣关系（NWR） [15]，并将其推广，使其考虑转移上的多项式以及加权目标状态。

定义 7（加权参数化马尔可夫决策过程的不劣关系）。若顶点的子集 $W \subseteq V$ 相对于顶点 $v \in V$ 不劣，记作 $v \trianglelefteq W$，当且仅当：

$$\forall\, \mathsf{val} \in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}, \exists\, w \in W : \mathrm{Rew}^*_{\mathsf{val}}(v) \leq \mathrm{Rew}^*_{\mathsf{val}}(w).$$

若 $v \trianglelefteq \{w\}$ 且 $w \trianglelefteq \{v\}$，我们记作 $v \sim w$；而 $\tilde{v}$ 表示 v 的等价类。

为简洁起见，我们用 $v \trianglelefteq w$ 代替 $v \trianglelefteq \{w\}$。对于两个子集 U、$W \subseteq V$，当且仅当 $U \neq \emptyset$ 且对于所有 $u \in U$ 都有 $u \trianglelefteq W$ 时，我们记作 $U \trianglelefteq W$。

例 *2.* 在图 2a 中，我们有 $p \trianglelefteq q$。这是因为，任何保持图结构的赋值上的约束 $2x^2 - y + 2y = 1$ 和 $0 < 2y < 1$ 都确保了$\frac{1}{2} < x$。因此，$\mathsf{Rew}^*_{\mathsf{val}}(q) = \mathsf{Rew}^*_{\mathsf{val}}(q, a) > 2$。现在，$p$ 的值将是 1 和 $\mathsf{Rew}^*_{\mathsf{val}}(q)$ 的凸组合。因此，其值至多为 q 的值。

3 从加权马尔可夫决策过程到无权重马尔可夫决策过程

在本节中，我们展示如何高效地将任意加权参数化马尔可夫决策过程转换为一个无权重的马尔可夫决策过程，使其具有状态的超集，并且保留 NWR。具体而言，我们有两种归约：一种用于将加权参数化马尔可夫决策过程转换为参数化马尔可夫决策过程，另一种用于将 w˜pMDP 转换为 ˜pMDP。前一种方法较为简单，但使用了标记转移的非平凡多项式；后一种方法则更为复杂。因此，在 [15] 中研究的关于 NWR 的算法和复杂度上界也可应用于平凡参数化加权情形。

注记*3*。 下面，我们将假设所有目标状态的权重都不同。也就是说，$T = \{t_0, \ldots, t_n\}$ with $0 = \rho(t_0) < \rho(t_1) < \cdots < \rho(t_n)$。这不会失去一般性。事实上，如果 t_i and t_j have $\rho(t_i) = \rho(t_j)$，那么由于所有目标状态都是陷阱状态，我们可以添加一条转移（针对任意动作 $a \in A$），使得 $\delta(t_i, a, t_j) = 1$，并从 T 中移除 t_i。这可以在对数空间内实现，不会添加新状态，而且显然会保留 NWR。

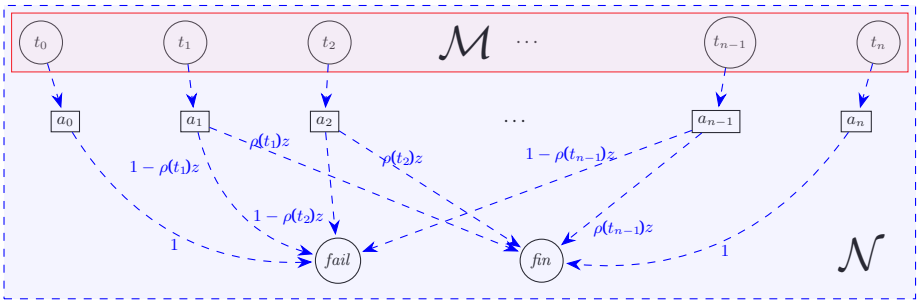


图3. 由 $\mathcal{M}$ 构造的参数化马尔可夫决策过程 $\mathcal{N}$，其中 $z = 1/\rho(t_n)$，而 a_i 表示状态-动作对 (t_i, a)。$\mathcal{M}$ 中所有目标状态在 $\mathcal{N}$ 中的奖励值比率与 $\mathcal{N}$ 中的相同。

3.1 从参数化马尔可夫决策过程中移除权重

设 $\mathcal{M}$ 是一个加权参数化马尔可夫决策过程。我们构造一个与 $\mathcal{M}$ 等价的参数化马尔可夫决策过程 $\mathcal{N}$，如图3所示。其思想是从目标状态向新添加的失败状态和终止状态添加转移，以保持原目标状态的值之间的比率。也就是说，对于所有状态 p、$q \in T \setminus \{t_0\}$，我们的构造保证如下：$\mathsf{Rew}^{p}_{\mathcal{M}^{\sigma}_{\mathsf{val}}}/\mathsf{Rew}^{q}_{\mathcal{M}^{\sigma}_{\mathsf{val}}} = \mathsf{Rew}^{p}_{\mathcal{N}^{\sigma_{\mathsf{val}}}}/\mathsf{Rew}^{q}_{\mathcal{N}^{\sigma_{\mathsf{val}}}}$，对于所有参数赋值 $\mathsf{val} \in \mathsf{Val}^{\mathsf{gp}}_{\mathcal{M}}$ 和所有策略 $\sigma \in \Sigma^{\mathcal{M}}$。根据值的定义，很容易看出，$\mathcal{G}(\mathcal{M})$ 中的所有其他顶点对——不包括保持其精确值的 t_0——也保持其值之间的比率。

定理 1。 给定一个加权参数化马尔可夫决策过程 $\mathcal{M}$，令 $\mathcal{G}(\mathcal{M}) = (V, E)$ 和 $\mathcal{N}$ 为通过上述构造得到的参数化马尔可夫决策过程。若 $v \in V$ 和 $W \subseteq V$，则 $v \trianglelefteq W$ 属于 $(\mathcal{M})$ 当且仅当 $v \trianglelefteq W$ 属于 $(\mathcal{N})$。

在下一节中，我们给出一种替代性归约，它保留平凡参数化这一性质。也就是说，它不会添加具有非平凡概率的转移。

3.2 从平凡参数化马尔可夫决策过程中移除权重

给定一个加权˜参数化马尔可夫决策过程 $\mathcal{M}$，我们通过添加 $n+3$个顶点和 $3n+2$ 条边来构造一个 ˜参数化马尔可夫决策过程 $\mathcal{N}$，如图4所示。为简化起见，我们假设任何没有通往终止状态的路径的自然顶点 $v \in S_N$ 都有一条通向 t_0的边。这不失一般性，因为这类顶点可以在多项式时间内检测出来（例如，参见 [1, 算法 46]），且此操作会保留其值。

与参数化情形不同，目标顶点的奖励值比率不会保留，但其序会保留：$\rho(t_i) < \rho(t_j)$ 在 $\mathcal{G}(\mathcal{M})$ 中成立，当且仅当$\mathbb{P}^{t_i}_{\mathcal{N}^{\sigma_{\mathsf{val}}}}[\lozenge \text{终止}] < \mathbb{P}^{t_j}_{\mathcal{N}^{\sigma_{\mathsf{val}}}}[\lozenge \text{终止}]$对所有赋值 $\mathsf{val} \in \mathsf{Val}^{\mathsf{gp}}_{\mathcal{N}}$ 和所有策略 $\sigma \in \Sigma^{\mathcal{N}}$ 均成立。

引理1。 对于无权重的 ˜参数化马尔可夫决策过程 $\mathcal{N}$，对所有赋值 $\mathsf{val} \in \mathsf{Val}^{\mathsf{gp}}_{\mathcal{N}}$ 以及所有策略 $\sigma \in \Sigma^{\mathcal{N}}$，都有 $0 = \mathbb{P}^{t_0}_{\mathcal{N}^{\sigma_{\mathsf{val}}}}[\lozenge \text{终止}] < \mathbb{P}^{t_1}_{\mathcal{N}^{\sigma_{\mathsf{val}}}}[\lozenge \text{终止}] < \cdots < \mathbb{P}^{t_n}_{\mathcal{N}^{\sigma_{\mathsf{val}}}}[\lozenge \text{终止}]$。

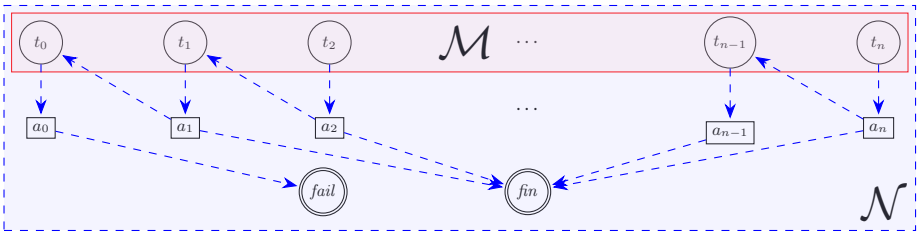


图4。由 $\mathcal{M}$ 构造的 ˜参数化马尔可夫决策过程 $\mathcal{N}$，其中 a_i 表示状态-动作对 (t_i, a)。顶点的奖励值序得以保留，但比率没有保留。

我们证明，尽管这一性质可能看起来不太直观，但它足以保留所有“绝不更差”关系。

定理 2。给定一个 *w˜pMDP* $\mathcal{M}$，令 $\mathcal{G}(\mathcal{M}) = (V, E)$ 和 $\mathcal{N}$ 为通过上述构造得到的 ˜*pMDP*。若 $v \in V$ 且 $W \subseteq V$，则 $v \trianglelefteq W$ 属于$(\mathcal{M})$ 当且仅当 $v \trianglelefteq W$ 属于 $(\mathcal{N})$。

证明（概要）。我们首先证明“若”方向的逆否命题：假设$v \not\trianglelefteq W$ 属于加权˜参数化马尔可夫决策过程 $\mathcal{M}$，那么存在一个赋值 val $\in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}$，使得对于所有 $w \in W$，$\mathsf{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(v) > \mathsf{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(w)$。现在我们可以构造一个赋值 $\mathsf{val}' \in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{N}}$，使得对于所有 $w \in W$，$\mathsf{Rew}^{*}_{\mathcal{N}[\mathsf{val}']}(v) > \mathsf{Rew}^{*}_{\mathcal{N}[\mathsf{val}']}(w)$。此类 val' 的存在将表明 $v \not\trianglelefteq W$ 属于 $\mathcal{N}$，从而完成这一方向的证明。更具体地说，val' 是由 val 得到的：为 $\mathcal{G}(\mathcal{N})$中的虚线边添加概率，使得 T中状态的奖励值之比与它们在 $\mathcal{M}$中的权重之比相同。

其次，我们证明“仅当”方向的逆否命题：假设在无权重的 ˜参数化马尔可夫决策过程 $\mathcal{N}$ 中，$v \not\trianglelefteq W$。这意味着存在一个赋值 val $\in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}$，使得对于所有 $w \in W$，$\mathsf{Rew}^{*}_{\mathcal{N}[\mathsf{val}]}(v) > \mathsf{Rew}^{*}_{\mathcal{N}[\mathsf{val}]}(w)$。然后，我们可以利用 $\mathcal{G}(\mathcal{N})$ 中各顶点的奖励值（对于这个固定的赋值 val），将顶点划分，使得同一划分中的所有顶点具有相同的奖励值。这个划分揭示了 $\mathcal{G}(\mathcal{N})$ 的结构所具有的一些良好性质，因此也揭示了 $\mathcal{G}(\mathcal{M})$ 的相关性质。我们利用这些性质，为原始的加权˜参数化马尔可夫决策过程 $\mathcal{M}$ 构造一个新的赋值 $\mathsf{val}' \in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}$，并证明对于所有 $w \in W$，$\mathsf{Rew}^{*}_{\mathcal{M}[\mathsf{val}']}(v) > \mathsf{Rew}^{*}_{\mathcal{M}[\mathsf{val}']}(w)$。□

4 判定 NWR 的复杂度

已有研究表明，在 [15] 中，对于平凡参数化马尔可夫决策过程，判定 NWR 是 NP的补类-完全问题，并且即使动作集合为单元素集合（即该马尔可夫决策过程本质上是一个马尔可夫链），甚至仅比较单元素集合时（即 $v \trianglelefteq w$），该问题仍然是 NP的补类-困难问题。我们已在上一节中证明，任何加权 ˜参数化马尔可夫决策过程都可以在多项式时间内约化为一个 ˜参数化马尔可夫决策过程，同时保留所有永不更差的.....

[1] 严格来说，Le Roux 和 Pérez 证明了该问题对于目标竞技场是困难的。在附录中，我们给出了目标竞技场与 ˜参数化马尔可夫决策过程之间的归约。

关系。注意，该归约不需要添加新动作。这些结果意味着，判定 w˜pMDPs 中的 NWR 是 NP的补类-完全问题。

定理 3. 给定一个加权参数化马尔可夫决策过程 $\mathcal{M}$，其中包含 $\mathcal{G}(\mathcal{M}) = (V, E)$ 以及 $v \in V$，$W \subseteq V$，判定 $v \trianglelefteq W$ 是否为 NP的补类-完全问题。此外，即使 W 和从 $\mathcal{M}$ 出发的动作集合 A 都是单元素集合，该问题仍然是 NP 的补类 - 困难的。

对于一般加权参数化马尔可夫决策过程，已知结果 [18,13] 表明，判定 NWR 是 **coETR**-困难的。实数存在理论 （ETR）是所有如下形式的真命题的集合：

$$\exists x_1 \ldots \exists x_n \varphi(x_1, \ldots, x_n) \tag{1}$$

其中，φ 是一个无量词（一级）公式，其谓词为不等式和等式，项为实系数多项式。复杂度类 **ETR** [17] 是指可在多项式时间内归约到判定类似于（1）的命题真值的问题集合，而 **coETR** 是指其补问题属于 **ETR** 的问题集合。已知 $\mathbf{NP} \subseteq \mathbf{ETR} \subseteq \mathbf{PSPACE}$。

为完整起见，我们给出一个自包含的证明，说明即使对于参数化马尔可夫决策过程，NWR 问题也是**coETR**-困难的。为此，我们从有界不等式合取问题（简称 BCon4Ineq）进行归约。该问题询问：给定四次多项式 $f_1, \ldots, f_m$，是否存在某个赋值 val : $X \to$ (0, 1)，使得 $\bigwedge_{i=0}^{m} f_i[\text{val}] < 0$? 根据引理 5], 该问题是**ETR**-困难的 [13, 。

定理 4. 给定一个参数化马尔可夫决策过程 $\mathcal{M}$，其中 $\mathcal{G}(\mathcal{M}) = (V, E)$ 以及 $v \in V$，$W \subseteq V$，判定 $v \trianglelefteq W$ 是否为**coETR**-完全问题。此外，即使 W 以及来自 $\mathcal{M}$ 的动作集合 A 均为单元素集合，该问题仍是**coETR**-困难的。

证明（概要）。为证明**coETR**-困难性，我们首先将 BCon4Ineq 归约为如下问题：判定给定参数化马尔可夫决策过程 $\mathcal{M}$ 是否存在保持图结构的赋值。设 $f_1, \cdots, f_m$ 为这些多项式。令 $k > k_i$，其中 k_i 是多项式 f_i 的系数绝对值之和。我们定义新的多项式 $f'_1, \ldots, f'_m$，其中 $f'_i = \frac{f_i}{k}$。注意，对于任意 val : $X \to$ (0, 1)，当且仅当 $\bigwedge_{i=0}^{m} f'_i[\text{val}] < 0$ 时，$\bigwedge_{i=0}^{m} f_i[\text{val}] < 0$ 成立；并且对任意这样的 val，都有 $-1 < f'_i[\text{val}] < 1$。现在，我们定义一个参数化马尔可夫决策过程，其中包含图 5 所示的状态和概率分布。从集合 $\{n_1, \ldots, n_m\}$ 到 {fail、fin} 的所有边确保对所有 $1 \le i \le m$ 都有 $0 < -f_i < 1$；从集合 $\{v_1, \ldots, v_n\}$ 到{fail、fin} 的边确保变量只能取自开集 (0, 1) 中的值。因此，当且仅当存在某个 val : $X \to$ (0, 1) 使得 $-1 < f'_i < 0$，该参数化马尔可夫决策过程才存在保持图结构的赋值。最后注意，当且仅当我们构造的参数化马尔可夫决策过程存在保持图结构的赋值时，fin $\not\trianglelefteq$ fail。注意，要得到完整的参数化马尔可夫决策过程，我们可以添加状态 $\{s_1, \ldots, s_m\}$，并为所有 i 添加从 s_i 到 n_i 的边。因此，动作集合 A 可以是单元素集合。

为了证明判定 NWR 属于 **coETR**，我们只需像公式 (1) 那样，将 NWR 的否定编码为一个公式，使得该公式属于 ETR 当且仅当 NWR 的否定成立。这种编码相当自然，可以将其视为用于编码马尔可夫决策过程值的经典线性规划的一种“符号化”版本。□

我们通过讨论 NWR 等价关系来结束本节。

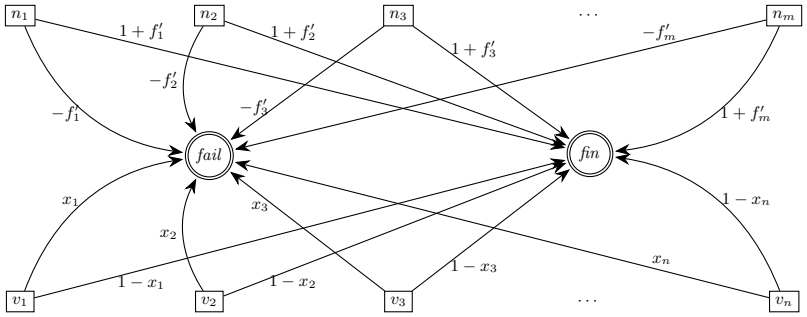


图 5. 用于证明判定参数化马尔可夫决策过程中是否存在保图赋值这一问题为 coETR-困难的参数化马尔可夫决策过程。

4.1 判定 NWR 等价关系的复杂度

遗憾的是，判定 NWR 等价关系与一般情况下判定 NWR 一样困难。为证明这一点，我们给出一个小型构件，它确保仅当某个 NWR 成立时，$v \sim w$ 才成立。

定理 5. 给定一个参数化马尔可夫决策过程 $\mathcal{M}$，其中包含 $\mathcal{G}(\mathcal{M}) = (V, E)$ 以及 $v, w \in V$，判定 $v \sim w$ 是否成立是 **coETR**-困难的。若 $\mathcal{M}$ 是一个 ˜参数化马尔可夫决策过程，则该问题一般而言是 NP的补类-困难的；若其动作集合 A 为单元素集合，则该问题属于 多项式时间复杂度类 。

为证明只有一个动作的 ˜参数化马尔可夫决策过程中的 NWR 等价关系可在多项式时间内判定，我们论证 $u \sim w$ 成立当且仅当存在唯一的$z \in V$使得 $\mathbb{P}^u_{\mathcal{C}}[\Diamond z] = \mathbb{P}^w_{\mathcal{C}}[\Diamond z] = 1$。这一刻画使我们能够根据每个状态几乎必然可从何处到达，计算出所有等价类。后者可通过 [1, 算法 45]在多项式时间内完成。

5 通过 NWR 剪枝动作

本节将介绍多种高效计算 NWR 下近似的方法，也就是求出该关系的子集，并利用这些下近似通过剪枝动作和合并加权参数化马尔可夫决策过程中的等价状态，来缩减加权参数化马尔可夫决策过程的规模。正如第 3 节所示，任何加权 w˜pMDP 都可以高效转换为（无权重的） ˜pMDP，同时保留原加权模型中的所有从不更差关系。因此，我们只介绍无权重模型的下近似方法（即 $T = \{\text{fin}, \text{fail}\}$）。此外，我们也只关注平凡参数化马尔可夫决策过程。请注意，如果 $U \trianglelefteq W$ in 一个 ˜ pMDP 中成立，那么在任何底层图与第一个 ˜pMDP 相同的 pMDP 中，$U \trianglelefteq W$ in 也成立。（也就是说，只要该模型存在保持图结构的赋值。）

今后，设 $\mathcal{M} = (S, A, \Delta, T)$ 为一个无权重的平凡参数化马尔可夫决策过程，并记 $\mathcal{G}(\mathcal{M}) = (V, E)$。我们假设所有 极值-值顶点 [15] 都已收缩：极值-值顶点 $v \in V$ 满足，对于所有 $\mathsf{val} \in \mathrm{Val}^{\mathrm{ep}}_{\mathcal{M}}$，$\mathrm{Rew}^*_{\mathsf{val}}(v) = 0$ 或 $\mathrm{Rew}^*_{\mathsf{val}}(v) = 1$ 成立。集合

所有此类顶点的集合都可以在多项式时间内计算出来 [1, 算法 45]。随后可以将所有值为 1 的顶点与终止以及所有值为 0 的顶点与失败进行收缩，而不会改变其他顶点的值。

5.1 下近似图

我们将通过一个（有向）下近似图 $\mathcal{U} = (N, R)$来表示当前对 NWR 的近似，使得 $N \subseteq 2^V$。对于 $U, W \in N$，令 $U \xrightarrow{\mathcal{U}} W$表示在 $\mathcal{U}$中存在一条从 U到 W的路径。整个算法过程中，我们将始终满足以下不变式。

$$\forall U, W \in N : U \xrightarrow{\mathcal{U}} W \implies U \trianglelefteq W \tag{2}$$

初始化图。为了得到对 NWR 的初始下近似，我们构造一个顶点集为 $N = \{\{v\}, vE \mid v \in V\}$ 的 $\mathcal{U}$。也就是说，它包含所有状态、所有状态-动作对，以及状态或状态-动作对的所有直接后继集合。对于边，我们按如下方式添加，以确保满足以下条件。

1. 对所有 $n \in N$， $R \ni (\{\text{失败}\}, n)$
2. 对所有 $n \in N$， $R \ni (n, \{\text{终止}\})$
3. 对所有 $v \in V$， $R \ni (\{v\}, vE)$
4. 对所有 $v \in S$， $R \ni (vE, \{v\})$

我们已经证明，这种初始化能够得到一个正确的下近似。

引理 2. 设 $\mathcal{U}$ 为上述定义的初始下近似图。那么，不变式 (2)成立。

下面，我们将介绍如何更新下近似图，从而得到对 NWR 越来越紧的下近似。

更新和查询图。 每当我们向某个下近似图 $\mathcal{U}$ 中添加一个顶点 U 时，也会添加:

– 一条从 $\{$ 到 U 的边，其中 $\}$ fail $U' \in N$，且 $U' \subset U$；– 一条从 U 到 $\{\text{fin}\}$ 的边，其中 $U'' \in N$，且 $U \subset U''$；以及– 一条从 U 到 W 的边，条件是对所有 $u \in U$， $\{u\} \xrightarrow{\mathcal{U}} W$ 均成立。

当我们向 $\mathcal{U}$ 添加一条边 (U, W) 时，如果顶点 U 和 W 尚未在图中，就先按照前文所述将它们添加进去。最后，要查询图并确定 $U \xrightarrow{\mathcal{U}} W$ 是否成立，只需按照定义中的建议，在图中搜索从U 出发的 W。

引理 3。 设 $\mathcal{U}'$ 是满足不变式 (2) 的下近似图，且$\mathcal{U}$ 是如上所述添加边 $(U, W) \in 2^V \times 2^V$ 后得到的下近似图。如果 $U \trianglelefteq W$，那么 $\mathcal{U}$ 满足不变式 (2)。

在深入探讨如何从 NWR 推断出成对元素以改进我们的近似图之前，我们需要回顾一下终端分量的概念。

5.2 末端分量与商化

若 $P \subseteq S$ 和 $B \subseteq \{(p,a) \mid (p,(p,a)) \in E\}$，则称 (P,B) 是 $\mathcal{M}$ 的一个子马尔可夫决策过程：
– 对于所有 $p \in P$，至少存在一个 $a \in A$，使得 $(p,a) \in B$；并且– 对于所有 $p' \in S$，以及满足 $((p,a)\ p') \in E$ 的 $(p,a) \in B$，都有 $p' \in P$。
可以将子马尔可夫决策过程视为原马尔可夫决策过程的连通分量，这些分量是通过从某些状态中移除一些动作而得到的。

如果由 $P \cup B$ 诱导的 $\mathcal{G}(\mathcal{M})$ 子图是强连通的，那么 $\mathcal{M}$ 的子马尔可夫决策过程 (P,B) 就是一个终端分量。若不存在其他终端分量 $(P',B') \neq (P,B)$ 满足 $B \subseteq B'$，则终端分量 (P,B) 是极大的（即一个极大末端分量）。极大末端分量的集合可以在多项式时间内通过 [1, 算法 47] 计算得到。

众所周知，末端分量是 NWR 的特殊情形。事实上，同一个末端分量中的所有顶点都是从不更差等价的 [15, 引理 3]。因此，正如我们对极值状态所做的那样，可以假设它们已经被收缩。在后续讨论中，明确说明实现这一收缩的商构造将会很有用。

极大末端分量商。令 $s \in S$，并用 $[\![s]\!]_{\mathrm{MEC}}$ 表示：如果 s 不属于任何极大末端分量，则为 $\{s\}$（P、B）；否则，为满足 $s \in P$ 的唯一极大末端分量 P、B。现在，我们将 $\mathcal{M}$ 的*MEC* 商记为 $\mathcal{M}_{/\mathrm{MEC}}$。也就是说，$\mathcal{M}_{/\mathrm{MEC}}$ 是 pMDP $\mathcal{M}_{/\mathrm{MEC}} = (S', A, \Delta', T')$，其中：
– $T' = \{[\![t]\!]_{\mathrm{MEC}} \mid \exists t \in T\}$，– $S' = \{[\![s]\!]_{\mathrm{MEC}} \mid \exists s \in S\}$，以及– $\Delta' = \{([\![s]\!]_{\mathrm{MEC}}, a, [\![s']\!]_{\mathrm{MEC}}) \mid \exists (s, a, s'') \in \Delta, [\![s]\!]_{\mathrm{MEC}} \neq [\![s'']\!]_{\mathrm{MEC}}\}$。

该构造确实保留了所有状态之间的不劣关系。
命题 1。对于所有 U，$W \subseteq S$ 我们都有：$U \trianglelefteq W$ 属于 $\mathcal{G}(\mathcal{M})$ 当且仅当 $\{[\![u]\!]_{\mathrm{MEC}} \mid \exists u \in U\} \trianglelefteq \{[\![w]\!]_{\mathrm{MEC}} \mid \exists w \in W\}$ 属于 $\mathcal{G}(\mathcal{M}_{/\mathrm{MEC}})$。
现在我们将说明如何通过 NWR 的推理规则更新近似图，同时剪枝非最优动作。

5.3 剪枝动作并推断 NWR

下一个定理指出，在取 MEC 商之后，我们可以剪枝那些会导向相对于 NWR 而言非最优顶点的动作。
定理 6。设 $\mathcal{M}$ 是一个 ˜参数化马尔可夫决策过程，使得 $\mathcal{M}$ 与 $\mathcal{M}_{/\mathrm{MEC}}$同构。那么，对于所有赋值 $\mathsf{val} \in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}$ 以及所有 $(s,a) \in S \times A$ 满足 $(s,a) \trianglelefteq (sE \setminus \{(s,a)\})$：

$$\max_{\sigma} \mathbb{P}^{s}_{\mathcal{M}^{\mathsf{val}}_{\sigma}}[\Diamond \mathit{fin}] = \max_{\sigma'} \mathbb{P}^{s}_{\mathcal{N}^{\mathsf{val}}_{\sigma'}}[\Diamond \mathit{fin}],$$

其中 $\mathcal{N} = (S,\ A,\ \Delta \setminus \{(s,\ a,\ s') \in \Delta \mid \exists s' \in S\},\ T)$。

接下来，我们将介绍推理规则，以基于近似图 $\mathcal{U}$推导出新的 NWR 对。以下定义将有所帮助。

本质状态与几乎必然可达性。如果对于从 U中的任意顶点出发的每条路径，集合 $W \subseteq V$ 都是[2] $U \subseteq V$的本质状态，记作 $U \sqsubseteq W$，且

[2] 此定义的灵感来源于 [7]，但与该论文中的定义并不完全相同。

以终止状态结尾的路径包含一个来自 W 的顶点。（直观地说，移除 W 会使 U 与终止状态断开。）我们称 $U \subseteq V$几乎必然到达 $W \subseteq V$，记作 $U \xrightarrow{\text{a.s.}} W$；如果将 T 替换为 $W \cup \{\text{失败}\}$，并令所有 $w \in W$ 的权重均为 1 后，所有状态 $u \in U$ 都变为值为 1 的状态，则称其满足此性质。回顾一下，对于所有保图赋值，值为 1 的状态确实如此，而且可以在多项式时间内计算出来。类似地，通过在移除 W 后得到的 $\mathcal{G}(\mathcal{M})$ 子图中进行搜索，可以在多项式时间内确定 W 是否是 U 的关键部分。我们还针对 $W \subseteq V$ 定义函数 $f_W \colon 2^S \to 2^S$:

$$f_W(D) = D \cup \left\{ z \in V \;\middle|\; \{z\} \xrightarrow{\mathcal{U}} W \cup D \right\}$$

并将其关于子集格的最小不动点记为 $\mu D.f_W(D)$。

推理规则。 对于所有 U、$W \subseteq V$，以下规则均成立。值得一提的是，这些规则是对 [15, 命题 1, 2] 的严格推广。

$$\frac{U \sqsubseteq \mu D.f_W(D)}{U \trianglelefteq W} \qquad \frac{\exists w \in W : \{w\} \xrightarrow{\text{a.s}} \{s \in S \mid U \xrightarrow{\mathcal{U}} \{s\}\}}{U \trianglelefteq W} \tag{3}$$

引理 4. 设 $\mathcal{U}$ 是满足 (2) 的近似图。那么，(3) 中的推理规则是正确的。

我们提出的动作剪枝方法如算法 1 所示。请注意，在第 15 行，我们会对 NWR 等价关系进行最终收缩。这些等价关系可以在位于同一环中的任意两个单元素集合 $\{x\}$ 和 $\{y\}$ 之间导出，该环属于 $\mathcal{U}$。该算法显然会终止，因为 NWR 是有限的，不存在无界递归，也不存在无界循环。最后，正确性由本节中的结果得出；对于前述 NWR 等价类的收缩，则由 [15, 定理 1] 得出。

6 实验

本节展示我们为评估前文所述技术的有效性而开展的实验结果。首先，我们阐述希望通过实验回答的研究问题。然后，介绍实验中使用的模型，包括所执行的预处理步骤。接下来，介绍将这些技术应用于所选模型和经过预处理的模型时采用的不同实验配置。最后，讨论实验中获得的结果。重新运行实验所需的一切内容都在 [8] 中。

我们开展实验时考虑了以下研究问题。

Q1. NWR 在实践中真的有用吗？Q2. 我们的下近似能否显著缩小某些模型的规模？这些归约与已知归约相比如何？Q3. 下近似能否在早期迭代中实现归约？

算法 1： 基于下近似的归约

输入： A ˜参数化马尔可夫决策过程 $\mathcal{M} = (S, A, \Delta, T)$

输出： 一个（希望更小的）˜参数化马尔可夫决策过程

初始化 $\mathcal{U}$;
重复
  对 $s \in S$ **执行**
    对 $(s,a) \in sE$ **执行**
      $W \leftarrow sE \setminus \{(s,a)\}$;
      如果 $\{(s,a)\} \xrightarrow{\mathcal{U}} W$ **则**
        剪枝所有 $\{(s,a,s') \in \Delta \mid \exists s' \in S\}$;
      否则如果 $\{(s,a)\} \sqsubseteq \mu D.f_W(D)$ **则**
        将边 $((s,a)\ W)$ 添加到 $\mathcal{U}$;
        剪枝所有 $\{(s,a,s') \in \Delta \mid \exists s' \in S\}$;
      否则如果 $\exists w \in W:\ \{w\} \xrightarrow{a} \{s' \in S \mid \{^{s}(s,a)\} \xrightarrow{\mathcal{U}} \{s'\}\}$ **则**
        将边 $((s,a)\ W)$ 添加到 $\mathcal{U}$;
        剪枝所有 $\{(s,a,s') \in \Delta \mid \exists s' \in S\}$;
直到 $\mathcal{U}$ 保持不变;
根据 $\mathcal{G}$中的环构造 NWR 等价关系;
返回新的 ˜参数化马尔可夫决策过程;

请注意，第一个问题（有意）比较模糊。因此，我们着重探讨后面两个更具体的问题。具体而言，对于下近似，我们想知道能否缩小常见基准测试的规模，此外，这种规模缩减是否可以通过少量应用我们的推理规则来实现。一方面，如果 Q2 的答案为肯定，就意味着 NWR 确实可以作为预处理步骤来缩小马尔可夫决策过程的规模。另一方面，如果 Q3 的答案为肯定，就意味着总值得尝试少量应用我们的推理规则。

6.1 基准测试与实验方案

在实验中，我们考虑了所有编码为 PRISM 模型的离散时间马尔可夫决策过程 [14] ，这些模型 [4], 来自定量验证基准集，且至少具有一个最大可达性属性。随后，我们按照 QComp 网站上的说明，筛除了状态数量未知或超过 64,000 个状态的模型，以限制实验的运行时间。

我们已将属性文件中的每个最大可达性性质都添加为模型文件中的标签。在某些情况下，出于好奇，或为了与其他已知归约进行比较，我们在模型中引入了新的标签作为终止状态。例如，在 wlan_dl 基准测试中，我们不检查两台站点都在截止期限内正确发送的最小概率，而是检查这样做的最大概率。在共识基准测试中，我们添加了新标签 fin_and_all_1 and fin_and_not_all_1，以便比较我们的归约

与 Bharadwaj 等人 [2] 所用的归约。下文表格和图表中使用的部分标签列示如下：

1. 模型：共识。标签：不一致、fin_and_all_1、fin_and_not_all_1。标签 disagree 是属性文件中的一个性质，另外两个标签用于与以往结果进行比较 [2]。2. 模型：zeroconf_dl。标签：deadline_min。在截止期限内未完成的最小概率看起来不像是最大可达概率问题，但结果表明，它等同于在截止期限内到达新 IP 地址的最大概率。3. 模型：zeroconf。标签：correct_max。4. 模型：crowds。标签：obs_send_all。5. 模型：brp。标签：no_succ_trans、rep_uncertainty、not_rec_but_sent。

随后，使用 STORM [12] 将添加了标签的 PRISM 文件转换为显式 JSON 格式。选择这种格式是因为它能在易于操作的文件格式中显式指定所有状态和状态-动作对。

随后，我们应用了两种预处理技术：合并所有极值状态和状态-动作对，以及合并所有最大内端组件。这两种技术及其与不劣关系的联系已在前文介绍。对于一些模型，合并所有极值状态后，模型便被完全归约，只剩下终止状态和失败状态。以下讨论不再包含这些马尔可夫决策过程，但它们列于 [8] 中。合并最大内端组件的效果则不那么显著，对所考察的模型没有带来额外归约。

6.2 不同的实验配置

在每项实验中，我们都会对每个模型反复应用算法 1。如果连续两次迭代中，合并等价状态（见第 15 行）均未产生任何影响，则终止实验。每次迭代开始时，我们都会从一个空的下近似图开始，以避免在合并过程中被移除的状态被纳入下近似图。算法 2 对此进行了说明。

算法 2： 实验设置

输入： 一个 ˜参数化马尔可夫决策过程 $\mathcal{M} = (S,\ A,\ \Delta,\ T)$

输出： 一个 ˜可分析的参数化马尔可夫决策过程

1 收缩极值状态；
2 收缩最大内端组件；
3 **重复**
4 　从一个空的下近似开始 $\mathcal{U}$；
5 　应用算法 1；
6 **直到**合并 *NWR* 等价关系没有影响；

为了缩短实验的运行时间，我们考虑将算法1第2行和第3行中指定的循环的最大迭代次数设为一个固定值。

算法 2。我们将算法 1 的循环称为“内循环”，将算法 2 的循环称为“外循环”。在实验期间，我们发现，在外循环的第一次迭代期间运行内循环耗时过长。因此，我们决定在外循环的第一次迭代期间跳过内循环。即便如此，对于某些模型，运行内循环仍然不切实际，从而得到以下两种实验配置：

1. 对于某些模型，我们允许外层循环最多运行 17 次。在我们的实验中，没有任何基准测试需要超过 7 次迭代。内层循环限制为 3 次迭代。2. 对于其他模型，我们运行了 3 次外层循环，但没有运行内层循环。

在我们的实验中，所考虑的马尔可夫链始终用于第二种实验设置，因为马尔可夫链是每个状态都只有一个动作的马尔可夫决策过程，因此动作剪枝内层循环是多余的。

备注 *4.* 每次向下近似图中添加节点 U 时，我们都会跳过“若 $\{u\} \xrightarrow{\mathcal{U}} W$ 对所有 $u \in U$都成立，则添加一条从 U 到 W 的边”这一步，因为我们发现这一步效率不高。

6.3 结果、表格和图

备注 *5.* 我们用术语选择项 表示马尔可夫决策过程中的一个状态-动作对。

表 1 展示了我们运行的一些基准测试中的状态空间缩减情况。经过预处理后，对于所有以“no_succ_trans”和“rep_uncertainty”为终止状态的 brp 基准测试，平均缩减了 61%；对于以“not_res_but_sent”为终止状态的基准测试，平均缩减了 40%。在 zeroconf_dl 基准测试中，平均还能进一步缩减 50%；在所有 zeroconf 基准测试中，缩减幅度约为 69%。对于以“不一致”为终止状态的共识基准测试，预处理仅移除了不到 16% 的状态，而之后又进一步缩减了约 37%。其余数据（包括更多基准测试）见 [8]。所有这些观察结果 [3] 都表明，Q2 的答案是肯定的。

在我们测试的所有基准测试中，只有 zeroconf、部分共识实例以及 zeroconf_dl 实例在算法 [8] 的第一次迭代之后有所改进。即便如此，大部分改进也发生在第一次迭代中，后续迭代的改进微乎其微，如图 6 所示，而且没有任何基准测试运行超过 7 次迭代。以“不一致”为终止状态的共识基准测试，在 (N =2, $K = 8$)、($N = 2, K = 16$) 上采用第二种配置，在 ($N = 2, K = 2$)、(N =2, $K = 4$) 上采用第一种配置。如表 1 所示，所有实例在预处理后都进一步缩减了约 36%，不过算法仅运行一次迭代的实例明显更快。这表明 Q3 的答案是肯定的。

[3] 我们无法将结果与 [2] 中的结果进行适当比较。该文报告的规模与 QComp 模型中的规模不一致。作者确认，他们的结果基于经过修改的模型，但相关数据和代码已遗失。

表 1。 表格展示了运行基准测试期间每个步骤中的状态数和选择数。最后一列给出了预处理后的运行时间。运行时间以（灰色）标示的实验采用了第一种实验设置。

基准测试	实例	原始规模 #st	原始规模 #ch	预处理 #st	预处理 #ch	下近似 #st #ch	time
共识 "不一致" (N,K)	(2,2)	274	400	232	344	148 260	78.17秒
	(2,4)	530	784	488	728	308 548 3	
	(2,8)	1042	1552	1000	1496	628 1124	19.38秒
	(2,16)	2066	3088	2024	3032	1268 2276	78.47秒
共识 "终止_和_所有_1" (N,K)	(2,2)	274	400	173	280	127 232 1	
	(2,4)	530	784	365	600	271 504 6	
	(2,8)	1042	1552	749	1240	561 1052	14.14秒
	(2,16)	2066	3088	1517	2520	1137 2140	57.78秒
共识 "终止_和_不_全部_1" (N,K)	(2,2)	274	400	165	268	123 226	61.15秒
	(2,4)	530	784	357	588	267 498 3	
	(2,8)	1042	1552	741	1228	555 1042	14.00秒
	(2,16)	2066	3088	1509	2508	1131 2130	58.24秒
zeroconf_dl "截止期限_min" (N,K,reset,截止期限)	(1000,1,true,10)	3837	4790	460	552	242 333	98.47s
	(1000,1,true,20)	7672	9775	2709	3351	1313 1945	79.49s
	(1000,1,true,30)	11607	14860	4999	6211	2413 3605	266.00s
	(1000,1,true,40)	15642	20045	7289	9071	3513 5265	560.82s
	(1000,1,true,50)	19777	25330	9579	11931	4613 6925	1013.18秒
	(1000,1,false,10)	12242	18200	460	552	242 333	95.69s
zeroconf "正确_最大值" (N,K,重置)	(1000,2,true)	672	814	388	481	118 150	29.13秒
	(1000,4,true)	1090	1342	674	855	206 260	76.06秒
	(1000,6,true)	1508	1870	960	1229	294 370	161.01秒
	(1000,8,true)	1926	2398	1246	1603	382 480	237.69秒
crowds "观测_发送_全部" (总运行次数,人群规模)	(3,5)	1200	1198	268	266	182 180	1.03秒
	(4,5)	3517	3515	1030	1028	632 630	11.93s
	(5,5)	8655	8653	2930	2928	1682 1680	87.76s
	(6,5)	18819	18817	6905	6903	3782 3780	467.08s
	(3,10)	6565	6563	833	831	662 660	13.12秒
	(4,10)	30072	30070	5430	5428	3962 3960	481.47秒
	(3,15)	19230	19228	1698	1696	1442 1440	69.90秒
brp "无_后继_转换" (N,最大值)	(64,2)	2695	2693	1982	1980	768 766	26.76秒
	(64,3)	3528	3526	2812	2810	1087 1085	53.80秒
	(64,4)	4361	4359	3642	3640	1406 1404	89.44秒
	(64,5)	5194	5192	4472	4470	1725 1723	134.57秒
brp "rep_不确定性" (N,最大值)	(64,2)	2695	2693	1982	1980	768 766	25.74秒
	(64,3)	3528	3526	2812	2810	1087 1085	52.20s
	(64,4)	4361	4359	3642	3640	1406 1404	87.62s
	(64,5)	5194	5192	4472	4470	1725 1723	133.96秒
brp “非_收_而是_发送” (N,最大值)	(64,2)	2695	2693	8	6	5 3	0.004秒
	(64,3)	3528	3526	10	8	6 4	0.01秒
	(64,4)	4361	4359	12	10	7 5	0.01秒
	(64,5)	5194	5192	14	12	8 6	0.01 秒

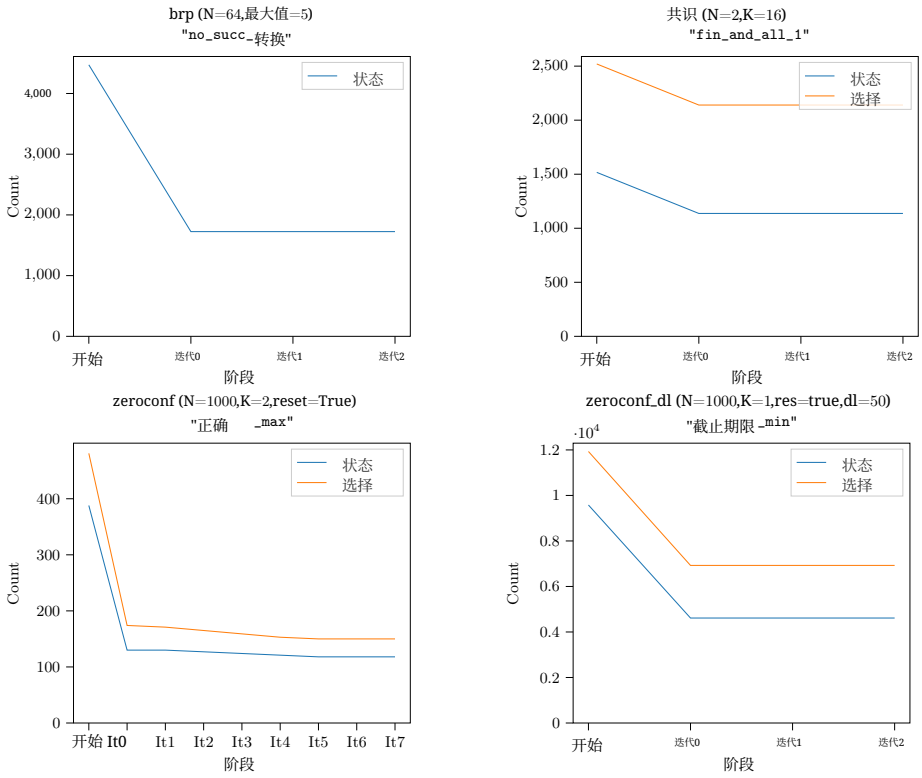


图 6. 一些展示每次迭代所获得归约的图表。

7 结论

我们将不劣关系扩展到了参数化马尔可夫决策过程上的定量可达性问题。尽管判定该关系的复杂度相对较高（**coETR**-完全），但高效的下近似似乎很有前景。我们认为，通过探索以即时方式计算此类下近似（例如采用即时 MEC 算法 [3]），可以使该关系得到更广泛的应用。以下问题也值得进一步开展（实证）研究：哪些类型的马尔可夫决策过程在 NWR 下具有较大的归约？约简后的马尔可夫决策过程是什么样的？NWR 下近似检测并移除了哪些类型的状态？

作为进一步的未来工作，我们认为值得研究当马尔可夫决策过程以简化有序二元决策图（二元决策图）编码，或以特定建模语言（如 PRISM 语言）描述时更高效的近似方法。针对后者，这方面的初步工作见 [20]。

致谢。 本研究得到了比利时 FWO “SAILor”（G030020N）项目和弗兰德斯跨大学（iBOF）“DESCARTES”项目的资助。

参考文献

1. Baier, C., Katoen, J.：模型检测原理。麻省理工学院出版社（2008）2. Bharadwaj, S., Roux, S.L., Pérez, G.A., Topcu, U.：马尔可夫决策过程模型检测与学习的约简技术。载于：Sierra, C.（编）《第二十六届国际人工智能联合会议论文集》，IJCAI 2017，澳大利亚墨尔本，2017年8月19—25日。第4273–4279页。ijcai.org（2017）。https://doi.org/10.24963/ijcai.2017/597, https://doi.org/10 597 3 24963 2017 . /ijcai. / . Brázdil, T., Chatterjee, K., Chmelik, M., Forejt, V., Kretínský, J., Kwiatkowska, M.Z., Parker, D., Ujma, M.：使用学习算法验证马尔可夫决策过程。载于：Cassez, F., Raskin, J.（编）《自动化技术验证与分析：第十二届国际研讨会，ATVA 2014》，澳大利亚新南威尔士州悉尼，2014年11月3—7日，论文集。《计算机科学讲义》，第8837卷，第98–114页。Springer（2014）。https://doi.org/10.1007/978-3-319-11936-6_8, https://doi.org/ 10 1007 978 3 319 11936 6_8 . / - - - - . 4 Budde, C.E., Hartmanns, A., Klauck, M., Kretínský, J., Parker, D., Quatmann, T., Turrini, A., Zhang, Z.：定量验证中的正确性、精度与性能：QComp 2020竞赛报告。载于：Margaria, T., Steffen, B.（编）《形式化方法、验证与确认的应用：工具与趋势——第九届形式化方法应用国际研讨会，ISoLA 2020》，2020年30日，希腊罗得岛，2020年10月20—24日，论文集，第四部分。《计算机科学讲义》，第216–241页。Springer（2020）。https://doi.org/10.1007/978-3-030-83723-5_15, https://doi.org/ 10 1007 978 3 030 83723 5_15 . / - - - - . Ciesinski, F., Baier, 5 C., Größer, M., Klein, J.：马尔可夫决策过程模型检测的约简技术。载于：第五届定量评估系统国际会议（QEST 2008），2008年9月14—17日，法国圣马洛。第45–54页。IEEE计算机学会（2008）。https://doi.org/10.1109/QEST.2008.45, https://doi.org/10 45 6 1109 2008 . /QEST. . . Clarke, E.M., Henzinger, T.A., Veith, H., Bloem, R.（编）：《模型检测手册》。Springer（2018）。https://doi.org/10.1007/978-3-319-10575-8, https://doi.org/10 8 7 1007 978 3 319 10575 . / - - - - . D'Argenio, P.R., Jeannet, B., Jensen, H.E., Larsen, K.G.：通过逐步细化分析概率系统的可达性。载于：de Alfaro, L., Gilmore, S.（编）《进程代数与概率方法、性能建模与验证：PAPM-PROBMIV 2001联合国际研讨会》，德国亚琛，2001年9月12—14日，论文集。《计算机科学讲义》，第2165卷，第39–56页。Springer（2001）。https://doi.org/10.1007/3-540-44804-7_3, https://doi.org/ 10 1007 3 540 44804 7_3 . / - - - . Engelen, K.：基于图的参数化与加权马尔可夫决策过程归约的代码。https://doi.org/10 7915828 9 5281. /zenodo. . Hartmanns, A., Hermanns, H.：Modest工具集：用于定量建模与验证的集成环境。载于：Ábrahám, E., Havelund, K.（编）《系统构造与分析工具及算法——第二十届国际会议，TACAS 2014》，作为欧洲软件理论与实践联合会议ETAPS 2014的一部分举行，法国格勒诺布尔，2014年4月5—13日，论文集。《计算机科学讲义》，第8413卷，第 8

593–598。Springer（2014）。https://doi.org/10.1007/978-3-642-54862-8_51, https://doi.org/10.1007/978-3-642-54862-8_5110. Hartmanns, A., Junges, S., Quatmann, T., Weininger, M.：《MDP模型检测算法实用指南》。CoRR **abs/2301.10197**（2023）。https://doi.org/10.48550/arXiv.2301.10197, https://doi.org/10.48550/arXiv.2301.10197 11. Hartmanns, A., Klauck, M., Parker, D., Quatmann, T., Ruijters, E.：《定量验证基准集合》。载于：Vojnar, T., Zhang, L.（编）：《系统构造与分析工具和算法——第25届国际会议，TACAS 2019，作为软件理论与实践欧洲联合会议ETAPS 2019的一部分举行，捷克共和国布拉格，2019年4月6日至11日，论文集》，第一部分。计算机科学讲义，第11427卷，第344–350页。Springer（2019）。https://doi.org/10.1007/978-3-030-17462-0_20, https://doi.org/10.1007/978-3-030-17462-0_2012. Hensel, C., Junges, S., Katoen, J., Quatmann, T., Volk, M.：《概率模型检测器STORM》。Int. J. Softw. Tools Technol. Transf.**24**（4），589–610（2022）。https://doi.org/10.1007/s10009-021-00633-z, https://doi.org/10.1007/s10009-021-00633-z 13. Junges, S., Katoen, J., Pérez, G.A., Winkler, T.：《参数化马尔可夫决策过程中可达性的复杂性》。J. Comput. Syst. Sci. **119**，183–210（2021）。https://doi.org/10.1016/j.jcss.2021.02.006, https://doi.org/10.1016/j.jcss.2021.02.006 14. Kwiatkowska, M.Z., Norman, G., Parker, D.：《PRISM 4.0：概率实时系统的验证》。载于：Gopalakrishnan, G., Qadeer, S.（编）：《计算机辅助验证——第23届国际会议，CAV 2011，美国犹他州斯诺鸟，2011年7月14日至20日，论文集》。计算机科学讲义，第6806卷，第585–591页。Springer（2011）。https://doi.org/10.1007/978-3-642-22110-1_47, https://doi.org/10.1007/978-3-642-22110-1_4715. Roux, S.L., Pérez, G.A.：《马尔可夫决策过程中基于图的可达性归约的复杂性》。载于：Baier, C., Lago, U.D.（编）：《软件科学与计算结构基础——第21届国际会议，FOSSACS 2018，作为软件理论与实践欧洲联合会议ETAPS 2018的一部分举行，希腊塞萨洛尼基，2018年4月14日至20日，论文集》。计算机科学讲义，第10803卷，第367–383页。Springer（2018）。https://doi.org/10.1007/978-3-319-89366-2_20, https://doi.org/10.1007/978-3-319-89366-2_2016. Russell, S., Norvig, P.：《人工智能：一种现代方法》（第4版）。Pearson（2020），http://aima.cs.berkeley.edu/ 17. Schaefer, M.：《某些几何与拓扑问题的复杂性》。载于：Eppstein, D., Gansner, E.R.（编）：《图绘制，第17届国际研讨会，GD 2009，美国伊利诺伊州芝加哥，2009年9月22日至25日，修订论文》。计算机科学讲义，第5849卷，第334–344页。Springer（2009）。https://doi.org/10.1007/978-3-642-11805-0_32, https://doi.org/10.1007/978-3-642-11805-0_3218. Spel, J., Junges, S., Katoen, J.：《参数化马尔可夫链具有单调性吗？》CoRR**abs/1907.08491**（2019），http://arxiv.org/abs/1907.08491 19. Sutton, R.S., Barto, A.G.：《强化学习——导论》。自适应计算与机器学习，麻省理工学院出版社（1998），https://www.worldcat.org/oclc/37293240

20. Winkler, T., Lehman, J., Katoen, J.：《失控：通过控制状态消除归约概率模型》。载于：Finkbeiner, B., Wies, T.（编）：《验证、模型检测与抽象解释——第23届国际会议，VMCAI 2022，美国宾夕法尼亚州费城，2022年1月16日至18日，论文集》。计算机科学讲义，第13182卷，第450–472页。Springer（2022）。
https://doi.org/10.1007/978-3-030-94583-1_22,
https://doi.org/10.1007/978-3-030-94583-1_22

A 定义与预备知识

我们通过使用贝尔曼方程，给出 $\mathrm{Rew}^*_{\mathsf{val}}(v)$ 的另一种等价定义，这在证明中会很有用：

引理 5。考虑一个加权参数化马尔可夫决策过程 $\mathcal{M}=(S,\ A,\ X,\ \delta,\ T,\ \rho)$。令 $\boldsymbol{Z}$ 表示所有无法到达 T的顶点所构成的集合。向量 $(x_v)_{v\in V}$，其中 $x_p=\mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(v)$ 是以下方程组的唯一解：

– 如果 $v\in T$ 则 $x_v=\rho(v)$；– 如果 $v\in \boldsymbol{Z}$ 则 $x_v=0$；– 如果 $v\in S_N\setminus(T\cup \boldsymbol{Z})$ 则 $x_v=\sum_{s\in vE}\delta(v,s)[\text{ 值 }]\cdot x_s$；– 如果 $v\in S\setminus(T\cup \boldsymbol{Z})$ 则 $x_v=$ 最大值 $_{a\in vE}\, x_a$。

以下引理揭示了状态和状态-动作对的最优奖励与其（直接）后继状态之间的关系。

引理6。对于加权参数化马尔可夫决策过程 $\mathcal{M}=(S,\ A,\ X,\ \delta,\ T,\ \rho)$ 以及赋值 $\mathsf{val}\in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}$，对于所有 $u\in S$，有

$$\mathrm{Rew}^*_{\mathsf{val}}(u)=\max_{a\in A}\mathrm{Rew}^*_{\mathsf{val}}(u,a),$$

并且对于所有 $u\in S_N$

$$\min_{(u,a)\in uE}\mathrm{Rew}^*_{\mathsf{val}}(u,a)\le \mathrm{Rew}^*_{\mathsf{val}}(u)\le \max_{(u,a')\in uE}\mathrm{Rew}^*_{\mathsf{val}}(u,a'),$$

其中，等式成立当且仅当 uE 中的所有顶点具有相同的值。

证明可直接由奖励值的等价定义得出。

B 移除参数化马尔可夫决策过程中的权重

令 $\mathcal{M}=(S,A,X,\delta,T,\rho)$ 为一个带有目标状态 $T=\{t_0,t_1,\cdots,t_n\}$的加权参数化马尔可夫决策过程，使得 $0=\rho(t_0)<\rho(t_1)<\cdots<\rho(t_n)$。现在我们将创建一个（无权重的）参数化马尔可夫决策过程 $\mathcal{M}'=(S',A',X',\delta',T')$，它与给定的加权参数化马尔可夫决策过程 $\mathcal{M}$具有相同的不劣关系。具体来说，我们添加一个新动作 $a\notin A$，以确保 $\mathcal{M}$ 中目标状态的权重比率与这些状态在 $\mathcal{N}$中的最优奖励值比率相同。形式化地：

$$\begin{aligned}
S' &= S\uplus\{\mathit{fail},\mathit{fin}\},\\
A' &= A\uplus\{a\},\\
\delta'(s,a',s') &= \delta(s,a',s')\text{ for }s,s'\in S\text{ and }a'\in A,\\
\delta'(t,a,\mathit{fin}) &= \tfrac{\rho(t)}{\rho(t_n)}\text{ for }t\in T,\\
\delta'(t,a,\mathit{fail}) &= 1-\tfrac{\rho(t)}{\rho(t_n)}\text{ for }t\in T,\\
X' &= X,\\
T' &= \{\mathit{fail},\mathit{fin}\}.
\end{aligned}$$

B.1 定理 1 的证明

令 $c = \frac{1}{\rho(t_n)}$。我们将证明，对于任意 $v \in V$，以及针对 $\mathcal{M}$ 和 $\mathcal{N}$ 的任意保图赋值 val，$\mathsf{Rew}^*_{\mathsf{val}}(v)_{\mathcal{M}} \cdot c = \mathsf{Rew}^*_{\mathsf{val}}(v)_{\mathcal{N}}$。证明这一点即可证明该定理。

首先，我们注意到 $\mathsf{Rew}^*_{\mathsf{val}}(t) = \rho(t) \cdot c$ 对于 $t \in T$。接下来，我们注意到，在 $\mathcal{N}$ 中，从任意状态 $s \in S$ 到终止的任何路径都必须经过 T。因此，对于任意状态 $s \in S$，我们有：

$$
\begin{aligned}
\mathrm{Rew}^*_{\mathcal{N}[\mathsf{val}]}(u) &= \max_{\sigma \in \Sigma^{\mathcal{N}}} \mathrm{Rew}^{\sigma}_{\mathcal{N}[\mathsf{val}]}(u) \\
&= \max_{\sigma \in \Sigma^{\mathcal{N}}} \mathrm{Rew}^{u}_{\mathcal{N}^{\sigma}_{\mathsf{val}}} \\
&= \max_{\sigma \in \Sigma^{\mathcal{N}}} \mathbb{P}^{u}_{\mathcal{N}^{\sigma}_{\mathsf{val}}}[\Diamond \mathit{fin}] \\
&= \max_{\sigma \in \Sigma^{\mathcal{N}}} \sum_{i=1}^{n} \mathrm{Rew}^{\sigma}_{\mathsf{val}}(t_i) \cdot \mathbb{P}^{u}_{\mathcal{N}^{\sigma}_{\mathsf{val}}}[\Diamond t_i] \qquad \text{using Lemma 7} \\
&= \max_{\sigma \in \Sigma^{\mathcal{N}}} \sum_{i=1}^{n} \mathrm{Rew}^{*}_{\mathsf{val}}(t_i) \cdot \mathbb{P}^{u}_{\mathcal{N}^{\sigma}_{\mathsf{val}}}[\Diamond t_i] \qquad \text{single choice for states in } T \\
&= \max_{\sigma \in \Sigma^{\mathcal{M}}} \sum_{i=1}^{n} \mathrm{Rew}^{*}_{\mathsf{val}}(t_i) \cdot \mathbb{P}^{u}_{\mathcal{M}^{\sigma}_{\mathsf{val}}}[\Diamond t_i] \qquad \text{by construction of } \mathcal{N} \\
&= \max_{\sigma \in \Sigma^{\mathcal{M}}} \sum_{i=1}^{n} \rho(t_i) \cdot c \cdot \mathbb{P}^{u}_{\mathcal{M}^{\sigma}_{\mathsf{val}}}[\Diamond t_i] \\
&= c \cdot \max_{\sigma \in \Sigma^{\mathcal{M}}} \sum_{i=1}^{n} \rho(t_i) \cdot \mathbb{P}^{u}_{\mathcal{M}^{\sigma}_{\mathsf{val}}}[\Diamond t_i] \\
&= c \cdot \max_{\sigma \in \Sigma^{\mathcal{M}}} \mathrm{Rew}^{u}_{\mathcal{M}^{\sigma}_{\mathsf{val}}} \\
&= c \cdot \mathrm{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(u).
\end{aligned}
$$

上文中，当我们使用引理 7 时，S、U、T 的值分别为 $\mathcal{N}^{\sigma}_{\mathsf{val}}$、$T$、终止。对于状态-动作对 $(s, a) \in S_N$，根据上述关于状态和奖励定义的结果，可直接得出结论。

C 从平凡参数化马尔可夫决策过程中移除权重

定义 8. 给定一个马尔可夫链 $\mathcal{C} = (S, \mu, T, \rho)$、一个初始状态 $s_0 \in S$以及状态集合 A，$B \subseteq S$，我们用 $\mathbb{P}^{s_0}_{\mathcal{C}}[A \mathsf{U}\, B]$ 表示从 s_0 出发，在最终到达 B 之前一直停留在 A 的概率。若 $s_0 \in B$，则 $\mathbb{P}^{s_0}_{\mathcal{C}}[A \mathsf{U}\, B] = 1$；否则：

$$\mathbb{P}^{s_0}_{\mathcal{C}}[A \mathsf{U}\, B] = \sum \{\{ \mathbb{P}_{\mu}(\pi) \mid \pi = s_0 \dots s_n \in A^{+} B \}\}.$$

当从 s_0 出发的所有运行在到达 T中的某个状态之前，都会经过 U 中的某个集合时，我们可以使用以下引理[15]分解到达 T 的概率。

引理 7（由 [15]）。设 $\mathcal{C} = (S, \mu)$ 为一个马尔可夫链，$U, T \subseteq S$ 为状态集合，$s_0 \in S \setminus U$ 为一个状态。若 $\mathbb{P}^{s_0}_{\mathcal{C}}[(S \setminus U) \cup T] = 0$，则：

$$\mathbb{P}^{s_0}_{\mathcal{C}}[\Diamond T] = \sum_{u \in U} \mathbb{P}^{s_0}_{\mathcal{C}}[(S \setminus U) \cup u] \cdot \mathbb{P}^{u}_{\mathcal{C}}[\Diamond T].$$

设 $\mathcal{M} = (S, A, \Delta, T, \rho)$ 是一个带权重的参数化马尔可夫决策过程，其目标状态为 $T = \{t_0, t_1, \cdots, t_n\}$，且 $0 = \rho(t_0) < \rho(t_1) < \cdots < \rho(t_n)$。现在，我们将构造一个无权重的参数化马尔可夫决策过程 $\mathcal{M}' = (S', A', \Delta', T')$，使其具有与给定的带权重参数化马尔可夫决策过程 $\mathcal{M}$ 相同的不劣关系。为此，我们添加一个新动作 $a \notin A$，以确保 $\mathcal{M}$ 中目标状态的权重序与这些状态在 $\mathcal{N}$ 中的奖励序相同。形式上：

$$\begin{aligned}
S' &= S \uplus \{\mathit{fail}, \mathit{fin}\},\\
A' &= A \uplus \{a\},\\
(s, a', s') &\in \Delta' \text{ iff } (s, a', s') \in \Delta \text{ for } s, s' \in S \text{ and } a' \in A,\\
(t, a, \mathit{fin}) &\in \Delta' \text{ for all } t \in \{t_1, \ldots, t_n\},\\
(t_i, a, t_{i-1}) &\in \Delta' \text{ for all } i \in \{1, \ldots, n\},\\
(t_0, a, \mathit{fail}) &\in \Delta',\\
T' &= \{\mathit{fail}, \mathit{fin}\}.
\end{aligned}$$

为方便起见，我们用 a_i 表示状态-动作对 (t_i, a)。

不失一般性，我们还假设任何没有通往终止的路径的状态-动作对 $v \in S_N$ 都有一条边通向 t_0。这样我们就可以避免以概率 1 陷入循环。给定一个加权参数化马尔可夫决策过程 $\mathcal{M}$，可以将赋值 $\mathsf{val} \in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}$ 视为一族具有全支撑的概率分布 $\mathsf{val} = (\mathsf{val}_u \in \mathbb{D}(uE))_{u \in S_N}$，其中 $\mathbb{D}(U)$ 是集合 U 上所有具有全支撑的概率分布构成的集合。我们将使用记号 $\mathsf{val}^{\mathcal{M}}_u(s)$ 表示由 val 在 $\mathcal{M}$ 中诱导的、从状态-动作对 u 到状态 s 的边上的概率。

注 *6.* 对于（加权或无权重的）平凡参数化马尔可夫决策过程，$\mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}$ 中的所有赋值恰好构成所有具有全支撑的概率分布族 $\mu = (\mu_u \in \mathbb{D}(uE))_{u \in S_N}$。

C.1 引理1的证明

很容易看出，$\mathbb{P}^{t_0}_{\mathcal{N}^\sigma_{\mathsf{val}}}[\Diamond \mathsf{fin}] = \mathbb{P}^{(t_0,a)_{\text{值}}}_{\mathcal{N}^\sigma}[\Diamond \mathsf{fin}] = 0$。我们还可以看出，$\mathbb{P}^{t_i}_{\mathcal{N}^\sigma_{\text{值}}}[\Diamond \mathsf{fin}] = \mathbb{P}^{(t_i,a)_{\text{值}}}_{\mathcal{N}^\sigma}[\Diamond \mathsf{fin}]$对于 $i \in \{1, \ldots, n\}$成立。现在，我们使用引理6以及 $\mathsf{Rew}^\sigma_{\mathsf{val}}(\mathsf{fin}) = 1$ 是可能的最大奖励这一事实，推导出$\mathbb{P}^{t_{i-1}}_{\mathcal{N}^\sigma_{\mathsf{val}}}[\Diamond \mathsf{fin}] \leq \mathbb{P}^{(t_i,a)_{\text{值}}}_{\mathcal{N}^\sigma}[\Diamond \mathsf{fin}] \leq 1$对于 $i \in \{1, \ldots, n\}$成立。现在我们将论证这些不等式实际上是严格的：假设对于某个 $i \in \{1, \ldots, n\}$，$\mathbb{P}^{(t_i,a)}_{\mathcal{N}^\sigma_{\mathsf{val}}}[\Diamond \mathsf{fin}] = 1$。这

根据引理 5，这将意味着 $\mathbb{P}^{t_{i-1}}_{\mathcal{N}^\sigma_{\mathsf{val}}}[\Diamond \mathsf{fin}] = \mathbb{P}^{(t_i,a_i)}_{\mathcal{N}^\sigma}{}_{\text{值}}[\Diamond \mathsf{fin}] = \mathbb{P}^{(t_{i-1},a_i)}_{\mathcal{N}^\sigma}{}_{\text{值}}[\Diamond \mathsf{fin}] = 1$。反复应用这一推理，我们得到 $\mathbb{P}^{t_0}_{\mathcal{N}^\sigma\text{值}}[\Diamond \mathsf{fin}] = \mathbb{P}^{(t_0,a_i)}{}_{\mathcal{N}^\sigma_{\mathsf{val}}}[\Diamond \mathsf{fin}] = 1$，这就产生了矛盾。因此，该不等式必须是严格的。证明至此完成。□

C.2 定理 2 的证明

我们证明如下性质足以保留 $\mathcal{M}$ 到 $\mathcal{N}$的所有不劣关系： $\mathcal{M}$ 中目标状态的权重序与 $\mathcal{N}$中这些状态的 $\mathsf{Rew}^*_{\mathcal{N}[\mathsf{val}]}$ 序相同。证明分为两部分，分别对应“若”和“仅若”。

第一种情况：若 $v \trianglelefteq W$ 属于$\mathcal{M}$，则 $v \trianglelefteq W$ 属于$\mathcal{N}$

我们证明其逆否命题：假设 $v \not\trianglelefteq W$ 属于加权 ˜参数化马尔可夫决策过程 $\mathcal{M}$，这意味着存在一个赋值 $\mathsf{val} \in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}$，使得对所有$w \in W$，$\mathsf{Rew}^*_{\mathcal{M}[\mathsf{val}]}(v) > \mathsf{Rew}^*_{\mathcal{M}[\mathsf{val}]}(w)$。现在我们将构造一个赋值 $\mathsf{val}' \in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{N}}$，使得对所有 $w \in W$，$\mathsf{Rew}^*_{\mathcal{N}[\mathsf{val}']}(v) > \mathsf{Rew}^*_{\mathcal{N}[\mathsf{val}']}(w)$。这样的 val'的存在将表明 $v \not\trianglelefteq W$属于 $\mathcal{N}$，从而完成这一方向的证明。

通过选择某个 $0 < \varepsilon < \rho^{\frac{1}{(t_n)}}$，可以轻松构造出 $\mathsf{Rew}^*_{\mathcal{N}[\mathsf{val}']}(t_i) = \mathsf{Rew}^*_{\mathcal{N}[\mathsf{val}']}(a_i) = \rho(t_i) \cdot \varepsilon$ 的概率分布 $\mathsf{val}'^{\mathcal{N}}_{a_i}$，其中 $a_i \in \{a_0, a_1, \cdots, a_n\}$。构造出用于 $a_i \in \{a_0, a_1, \cdots, a_n\}$ 的 $\mathsf{val}'^{\mathcal{N}}_{a_i}$ 后，我们便可定义赋值 val'，使得对于所有 $(u, v) \in (S_N \times S) \cap E$:

$$\mathsf{val}'^{\mathcal{N}}_{u}(v) = \begin{cases} \mathsf{val}^{\mathcal{M}}_{u}(v) & \text{if } u \in S_N, \\ \mathsf{val}'^{\mathcal{N}}_{a_i}(v) & \text{if } u = a_i \text{ for } 0 \le i \le n. \end{cases}$$

我们考虑这样一个事实：在 $\mathcal{N}$ 中，对于分布族 val'，我们有 $\mathsf{Rew}^*_{\mathcal{N}[\mathsf{val}']}(t_i) = \rho(t_i) \cdot \varepsilon$对于 $1 \le i \le n$。我们首先证明，对于任意 $u \in V$，都有 $\mathsf{Rew}^*_{\mathcal{N}[\mathsf{val}']}(u) = \varepsilon \cdot \mathsf{Rew}^*_{\mathcal{M}[\mathsf{val}]}(u)$。为证明这一点，我们首先注意到

在 $\mathcal{N}$ 中，从 u 到终止的任何路径总会经过 T 中的一个顶点。

$$
\begin{aligned}
\mathrm{Rew}^{*}_{\mathcal{N}[\mathsf{val}']}(u) &= \max_{\sigma\in\Sigma^{\mathcal{N}}} \mathbb{P}^{u}_{\mathcal{N}[\mathsf{val}']}[\Diamond \mathit{fin}]_{\mathcal{N}} \\
&= \max_{\sigma\in\Sigma^{\mathcal{N}}} \mathbb{P}^{u}_{\mathcal{N}^{\sigma}_{\mathsf{val}'}}[\Diamond \mathit{fin}] \\
&= \max_{\sigma\in\Sigma^{\mathcal{N}}} \sum_{i=1}^{n} \mathrm{Rew}^{\sigma}_{\mathcal{N}[\mathsf{val}']}(t_i)\cdot \mathbb{P}^{u}_{\mathcal{N}^{\sigma}_{\mathsf{val}'}}[\Diamond t_i] \quad \text{using Lemma 7} \\
&= \max_{\sigma\in\Sigma^{\mathcal{N}}} \sum_{i=1}^{n} \mathrm{Rew}^{*}_{\mathcal{N}[\mathsf{val}']}(t_i)\cdot \mathbb{P}^{u}_{\mathcal{N}^{\sigma}_{\mathsf{val}'}}[\Diamond t_i] \quad \text{single choice for targets} \\
&= \max_{\sigma\in\Sigma^{\mathcal{M}}} \sum_{i=1}^{n} \mathrm{Rew}^{*}_{\mathcal{N}[\mathsf{val}']}(t_i)\cdot \mathbb{P}^{u}_{\mathcal{M}^{\sigma}_{\mathsf{val}'}}[\Diamond t_i] \quad \text{by construction of } \mathcal{N} \\
&= \max_{\sigma\in\Sigma^{\mathcal{M}}} \sum_{i=1}^{n} \mathrm{Rew}^{*}_{\mathcal{N}[\mathsf{val}']}(t_i)\cdot \mathbb{P}^{u}_{\mathcal{M}^{\sigma}_{\mathsf{val}}}[\Diamond t_i] \quad \text{by construction of } \mathsf{val}' \\
&= \max_{\sigma\in\Sigma^{\mathcal{M}}} \sum_{i=1}^{n} \rho(t_i)\cdot \varepsilon \cdot \mathbb{P}^{u}_{\mathcal{M}^{\sigma}_{\mathsf{val}}}[\Diamond t_i] \\
&= \varepsilon \cdot \max_{\sigma\in\Sigma^{\mathcal{M}}} \sum_{i=1}^{n} \rho(t_i)\cdot \mathbb{P}^{u}_{\mathcal{M}^{\sigma}_{\mathsf{val}}}[\Diamond t_i] \\
&= \varepsilon \cdot \max_{\sigma\in\Sigma^{\mathcal{M}}} \mathrm{Rew}^{\sigma}_{\mathcal{M}[\mathsf{val}]}(u) \\
&= \varepsilon \cdot \mathrm{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(u).
\end{aligned}
$$

上文使用引理 7 时， S、U、T的值分别为 $\mathcal{N}^{\sigma}$值′、T、终止。

因此，如果对于 $\mathcal{M}$ 中的所有 $w\in W$，$\mathrm{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(v) > \mathrm{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(w)$，则有

$$
\begin{aligned}
\mathrm{Rew}^{*}_{\mathcal{N}[\mathsf{val}']}(v) &= \varepsilon \cdot \mathrm{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(v) \\
&> \varepsilon \cdot \mathrm{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(w) \\
&= \mathrm{Rew}^{*}_{\mathcal{N}[\mathsf{val}']}(w).
\end{aligned}
$$

第二种情况：仅当 $v \trianglelefteq W$ 属于 $\mathcal{M}$ 时，$v \trianglelefteq W$ 才属于 $\mathcal{N}$

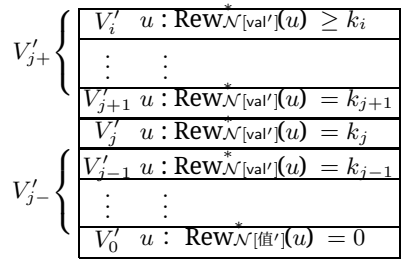
同样，我们证明其逆否命题：假设 $v \not\trianglelefteq W$ 属于（无权重）˜参数化马尔可夫决策过程 $\mathcal{N}$。这意味着存在一个分布族 val'，使得对于所有 $w\in W$，$\mathrm{Rew}^{*}_{\mathcal{N}[\mathsf{val}']}(v) > \mathrm{Rew}^{*}_{\mathcal{N}[\mathsf{val}']}(w)$。

令 $K=\{k_0=0,k_1,\cdots,k_m\}$ 为 $\mathcal{N}$ 中所有顶点取得的值按递增顺序组成的集合（即，当且仅当 $\mathrm{Rew}^{*}_{\mathcal{N}[\mathsf{val}']}(u)=k$ 对某个 $u\in V'$ 成立时，$k\in K$）。由于 $\mathrm{Rew}^{*}_{\mathcal{N}[\mathsf{val}']}(\mathit{fail})=0$，因此有 $k_0=0$。令 $\mathrm{Rew}^{*}_{\mathcal{N}[\mathsf{val}']}(v)=k_i$，其中 $1\le i\le m$（其中 v 是我们在逆否命题假设中的顶点）。我们将 V' 划分为 $\{V'_0,V'_1,\cdots,V'_i\}$，使得所有值为 $k_j<k_i$ 的顶点

分别属于对应的划分 V'_j，而值为 $k_j \geq k_i$ 的顶点属于 V'_i:

$$V'_i = \{u \in V' | \mathrm{Rew}^*_{\mathcal{N}[\mathsf{val}']}(u) \geq k_i\},$$
$$V'_{j<i} = \{u \in V' | \mathrm{Rew}^*_{\mathcal{N}[\mathsf{val}']}(u) = k_j\}.$$

此划分如以下图 7 所示。



Fi图 7. 节点按其 v 的 值。

我们将此划分的每个组成部分称为一"层"，其中 V'_0 是底层，V'_i是最顶层。为方便起见，对于 $0 \leq j \leq i$，我们定义 $V'_{j-} := V'_0 \cup V'_1 \cup \cdots \cup V'_{j-1}$ 以及 $V'_{j+} := V'_{j+1} \cup V'_{j+2} \cup \cdots \cup V'_i$。现在我们陈述该划分的若干性质:

1. $v \in V'_i$和 $W \subseteq V'_{i-}$ （即 v 位于顶层，且 W 中没有顶点位于该层）。2. $\{t_0, \cdots, t_p\} \subseteq V'_{i-}$ 和 $\{t_{p+1}, \cdots, t_n\} \subseteq V'_i$，其中 p 是满足 $\mathrm{Rew}^*_{\mathcal{N}[\mathsf{val}']}(t_p) < \mathrm{Rew}^*_{\mathcal{N}[\mathsf{val}']}(v) \leq \mathrm{Rew}^*_{\mathcal{N}[\mathsf{val}']}(t_{p+1})$ 的唯一数值。3. 对于 $0 \leq j < i$，$|V_j \cap T| \leq 1$（即除 V_i外，每层在 T中至多有一个状态）。4. 对于所有 $0 \leq j < i$和所有 $u \in V'_j \cap S'$，都有 $uE \cap V'_{j+} = \emptyset$（即状态-动作对所在的层不能高于对应状态所在的层）。5. 对于所有 $0 \leq j < i$、所有 $u \in V'_j \cap S'_N$和 $uE \cap V'_{j+} \neq \emptyset$，当且仅当 $uE \cap V'_{j-} \neq \emptyset$（即状态-动作对当且仅当存在一条边指向其下方层中的某个状态时，才存在一条边指向其上方层中的某个状态）。6. 对于所有 $0 \leq j \leq i$以及所有 $u \in V'_j \cap (S' \setminus T')$，都有 $uE \cap V'_j \neq \emptyset$（即每个非目标状态都必须有一条边指向同一层中的某个顶点）。

第一个性质源于以下事实： v 是顶层中值最小的顶点，且其值严格大于 W中每个顶点的值。第二个性质将 T 中的状态（即$\mathcal{M}$中的目标状态）划分为最顶层 V'_i 中的状态和不在其中的状态。第三个性质源于以下假设： T 中没有两个状态具有相同的权重。最后三个性质由引理6推出。

例*3.* 图8展示了由最后三个性质所限定的层间边的所有可能情况。

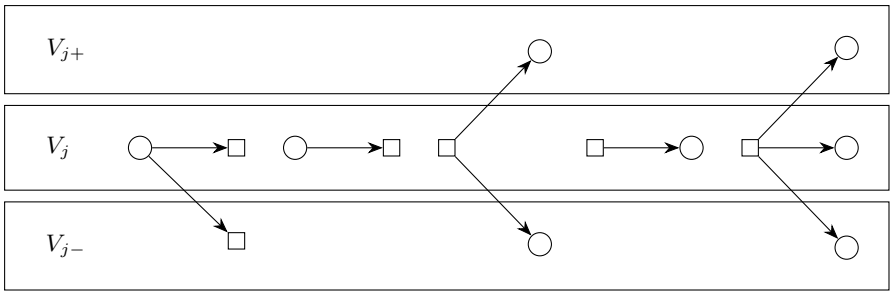


图8. 最后3条规则所限定的边的唯一可能形式

这一划分在 $\mathcal{N}$ 上诱导出 $\mathcal{M}$ 中的一个划分。我们将此划分称为 $\{V_0, V_1, \cdots, V_i\}$，其中 $V_j = V'_j \cap V$、V_{j+} 和 V_{j-} 对所有 $0 \le j \le i$ 均采用类似的定义。直观而言，我们移除了在构造 $\mathcal{N}$ 时添加的新顶点。请注意，从 $V' \setminus V$ 中移除这些顶点后，某些划分可能变为空集。

对于 V' 的划分，我们在上文陈述的六条性质对于 V 的这个划分也都成立。这是因为，除了 $T = \{t_1, t_2, \cdots, t_n\}$ 中的顶点外，没有其他顶点与 $V' \setminus V$ 中已被移除的（新）顶点之间存在边，而且 T 中的顶点在 $\mathcal{M}$ 中没有出边。

引理 8（来自 [15]）。在一个加权参数化马尔可夫决策过程 $\mathcal{M} = (S, A, X, \delta, T, \rho)$ 中，对于任意顶点$v \in V$和赋值 $\mathsf{val} \in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}$，如果顶点 $v \in V$ 能够到达某个目标状态，那么就存在一条从 v 到某个目标顶点的路径，且沿该路径的 $\mathrm{Rew}^*_{\mathsf{val}}$ 值非递减。

根据引理 8 以及 $\mathsf{Rew}^*_{\mathcal{N}[\mathsf{val}']}(v) = k_i > 0$ 这一事实，V'_i 中的每个顶点都有一条通往终止状态 fin 的路径，且该路径始终位于 V'_i 中。由于从 V 中任一顶点到终止状态 fin 的路径都必须经过 T 中的某个顶点，因此 V'_i 中的每个顶点都有一条通往 T 中某个顶点的路径，且该路径始终位于 V'_i 中。

为了证明 $v \not\preceq W$ 属于 $\mathcal{M}$，我们构造一个赋值（即一族分布）$\mathsf{val} = (\mathsf{val}_u \in \mathbb{D}(uE))_{u \in S_N}$，使得对于所有$w \in W$，$\mathsf{Rew}^*_{\mathcal{M}[\mathsf{val}]}(v) > \mathsf{Rew}^*_{\mathcal{M}[\mathsf{val}]}(w)$。

令 $0 < \varepsilon < 1$ 足够小，并令 m 为 V中的顶点数。令O：$v_1, v_2, \cdots, v_m$ 为这些顶点上的一种序，该序保留划分中的序（较低层中的顶点在序中先于较高层中的顶点出现）。形式化地，$v_{i_1} \in V_{j_1}$ 和 $v_{i_2} \in V_{j_2}$ 满足 $j_1 < j_2$ 即可推出 $i_1 < i_2$。

设 $u \in V_i \cap S_N$（即顶层中的一个状态-动作对）。我们考察从 u 到 T 中的某个目标状态、且始终停留在 V_i中的最短路径。将这组路径记为 P^T_u。定义 P^T_u中所有路径的长度为 ℓ_u。对于这样的顶点 $u \in V_i \cap S_N$ 和 $p \in uE$，我们如下定义 $\mathsf{val}^{\mathcal{M}}_u(p)$：

$$\mathsf{val}^{\mathcal{M}}_u(p) = \begin{cases} 1 - (|uE| - 1)\varepsilon & \text{if } p \text{ is a successor of } u \text{ in a path in } P^T_u \text{ and } p \text{ is} \\ & \text{the first one in } O \text{ among all such vertices} \\ \varepsilon & \text{else if } p \in uE. \end{cases}$$

通俗地说，可以认为 ν_u 确保以几乎为 1 的概率沿着从 u 到某个目标顶点的一条最短路径前进。回顾一下， m 是 V中的顶点数量， $\{t_{p+1}, \cdots, t_n\}$ 是最顶层（V_i）中的目标顶点，而 $\{t_0, \cdots, t_p\}$ 是 V_{i-}中的目标顶点。

引理 9. 上述为顶点 V_i 定义的（部分）概率分布族 ν 确保，对于任意顶点 $u \in V_i$:

$$\mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(u) \geq (1 - m\varepsilon)^{\ell_u} \cdot \rho(t_{p+1}).$$

证明. 我们将通过对路径 ℓ_u 的长度进行归纳，证明该引理对所有 $u \in V_i \cap S_N$ 均成立。根据引理6，顶点 $V_i \cap S$ 的情况也将自动成立。

基础情形： 如果 $\ell_u = 1$，那么根据定义，存在一个顶点 $t \in T$，使得$\nu_u(t) = 1 - (|uE| - 1)\varepsilon$。

$$\begin{aligned}
\mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(u) &\geq (1 - (|uE| - 1)\varepsilon) \cdot \rho(t) \\
&\geq (1 - m\varepsilon) \cdot \rho(t) \\
&\geq (1 - m\varepsilon) \cdot \rho(t_{p+1}). && \text{根据引理1和性质 2}
\end{aligned}$$

归纳步骤： 现在，假设该引理对所有满足 $\ell_u < \ell$ 的 $u \in V_i \cap S_N$ 均成立。设 $\ell_u = \ell$，并令 u、u_1、u_2、$\cdots$、$u_{\ell-1}$、t 为 P^T_u 中的一条路径，其中$\nu_u(u_1) = 1 - (|uE| - 1)\varepsilon$。我们知道这条路径属于 V_i，因此 $u_2 \in V_i \cap S_N$ 且 $\ell_{u_2} = \ell - 2$。于是，我们有 $\mathrm{Rew}^{\nu}(u_2) \geq (1 - m\varepsilon)^{\ell-2} \cdot \rho(t_{p+1})$，从而

$$\begin{aligned}
\mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(u) &= (1 - (|uE| - 1)\varepsilon) \cdot \mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(u_1) \\
&\geq (1 - m\varepsilon) \cdot \mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(u_1) \\
&\geq (1 - m\varepsilon) \cdot \mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(u_2) && \text{by Lemma 6} \\
&\geq (1 - m\varepsilon) \cdot (1 - m\varepsilon)^{\ell-2} \cdot \rho(t_{p+1}) \\
&= (1 - m\varepsilon)^{\ell-1} \cdot \rho(t_{p+1}) \\
&> (1 - m\varepsilon)^{\ell} \cdot \rho(t_{p+1}).
\end{aligned}$$

至此，该引理的证明完成。 □

下一个引理给出了子集内所有顶点的上界，前提是已知该子集外具有出边的顶点的奖励以及该子集内目标状态的权重的上界。

引理10。 在加权参数化马尔可夫决策过程 $\mathcal{M} = (S, A, X, \delta, T, \rho)$中，对于任意顶点子集$U \subseteq V$，如果所有顶点 $u \in U$ （满足 $uE \setminus U \neq \emptyset$）以及$U \cap T$ 中的所有顶点的值均不超过 x，那么 U 中所有顶点的值均不超过 x。

证明。 我们首先说明直观思路：由于集合 U 中每个有边离开的顶点的值都至多为 x，因此我们可以修改加权参数化马尔可夫决策过程，使所有这类顶点的值都为 x，并将其设为目标状态（没有出边）。显然，修改后的加权参数化马尔可夫决策过程中所有顶点的值都至少与原加权参数化马尔可夫决策过程中相应顶点的值相等。现在，修改后的加权参数化马尔可夫决策过程中 U 中的每个顶点都

具有至多为 x 的权重，因为所有目标状态的值都至多为 x，且没有边离开 U。形式化地说：

根据引理 5，我们知道，以下线性方程组有唯一解，该解描述了加权参数化马尔可夫决策过程中各顶点的 $\mathsf{Rew}^*_{\mathsf{val}}$ 值：

$$\mathrm{Rew}^*_{\mathsf{val}}(v) = \begin{cases} \rho(v) & \text{if } v \in T \\ 0 & \text{if } v \in Z \\ \max_{u \in vE} \mathrm{Rew}^*_{\mathsf{val}}(u) & \text{if } v \in S \setminus (T \cup Z) \\ \sum_{u \in vE} \delta(v,u)[\mathsf{val}] \cdot \mathrm{Rew}^*_{\mathsf{val}}(u) & \text{if } v \in S_N \setminus Z. \end{cases}$$

其中， Z 是所有无法通过路径到达 T的顶点集合。由于我们已知 U 中所有有边离开 U的顶点的 $\mathsf{Rew}^*_{\mathsf{val}}$ 值上界为 x，现在考虑以下关于 U中顶点的方程组：

$$\mathrm{Rew}^*_{\mathsf{val}}(v) = \begin{cases} x & \text{if } v \in (T \cap U) \text{ or } vE \setminus U \neq \emptyset \\ 0 & \text{if } v \in Z \\ \max_{u \in vE} \mathrm{Rew}^*_{\mathsf{val}}(u) & \text{if } v \in (S \cap U) \setminus (T \cup Z) \\ \sum_{u \in vE} \delta(v,u)[\mathsf{val}] \cdot \mathrm{Rew}^*_{\mathsf{val}}(u) & \text{if } v \in (S_N \cap U) \setminus Z. \end{cases}$$

如果将$(T \cap U)$中的所有顶点都设为权重为 x的 $\cup \{u \in U | uE \setminus U \neq \emptyset\}$ 目标状态，根据引理 5，这个修改后的加权参数化马尔可夫决策过程中各顶点的值就是上述方程组的唯一解（因为此时 U 中没有顶点能通过路径到达 U之外）。因此，我们得出结论：能够离开 U 的顶点和 U 中的目标状态的上界，意味着 U中所有顶点的上界。

□

现在我们将定义 V_{i-} 中各顶点的 val。令 $u \in V_j \cap S_N$，其中 $0 \leq j < i$。我们将如下定义从 u 出发的边的分布：

$$\mathsf{val}^{\mathcal{M}}_u(p) = \begin{cases} \frac{1}{|uE|} & \text{if } uE \cap V_{j-} = \emptyset \text{ and } p \in uE, \\ 1 - (|uE| - 1)\varepsilon & \text{else, if } p \in uE \cap V_{j-} \text{ and among all such} \\ & \text{vertices, } p \text{ is the first one in the ordering } O, \\ \varepsilon & \text{else, if } p \in uE. \end{cases}$$

非正式地说，对于足够小的 $0 < \varepsilon < 1$， ν_u 可确保在任何一次运行中，下一个顶点以几乎为 1 的概率位于较低的层中。这引出了以下引理。

引理 11. 对于所有满足 $0 \leq j < i$ 的顶点 $u \in V_j$，我们有：

$$\mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(u) \leq \rho(t_p) + jm\varepsilon\rho(t_n).$$

证明。 我们通过对顶点 $u \in V_j$ 所在的层进行归纳法来证明该结论。在归纳过程中，我们将先证明该结论适用于 S_N 中的顶点，然后

利用划分的性质4和引理10，证明该界也适用于 S 中其余的顶点。

基础情形： 令 $u \in V_0$，$\mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(u) = 0 \le \rho(t_p)$，因为 $\rho(t_0) = 0$ 且 $T \cap V_0 = \{t_0\}$（根据引理1以及我们构造划分的方式），并且由于不存在从该划分向外延伸的边 V_i（利用划分的性质4和性质5），因为该层以下没有其他层。

现在，假设对于$u \in V_{j-}$，我们有 $\mathrm{Rew}^*_{\mathsf{val}}(u) \le \rho(t_p) + (j-1)m\varepsilon\rho(t_n)$。

归纳步骤： 令 $u \in V_j \cap S_N$。若 $uE \cap V_{j+} \neq \emptyset$，则存在一个顶点 q（根据划分的性质5），它属于 V_{j-}，且满足 $\nu_u(q) = 1 - (|uE| - 1)\varepsilon$。因此我们有

$$
\begin{aligned}
&\mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(u) \\
&= (1 - (|uE| - 1)\varepsilon)\mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(q) \\
&+ \varepsilon \sum_{v \in uE \setminus \{p\}} \mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(v) \\
&\le \mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(q) + \varepsilon \sum_{v \in uE \setminus \{p\}} \mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(v) \\
&\le \mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(q) + (|uE| - 1)\varepsilon\rho(t_n) && \text{奖励不能超过 } \rho(t_n) \\
&\le \mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(q) + m\varepsilon\rho(t_n) \\
&\le (\rho(t_p) + (j-1)m\varepsilon n) + m\varepsilon\rho(t_n) && \text{根据归纳假设} \\
&= \mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(t_p) + jm\varepsilon\rho(t_n).
\end{aligned}
$$

我们考察顶点的子集 $V_{j-} \cup V_j$。在这个子集中，根据归纳假设，对于 V_{j-} 中的所有顶点以及满足 $uE \cap V_{j+} \neq \emptyset$ 的所有顶点 $u \in V_j$，我们都有上界 $\mathrm{Rew}^{\nu}(t_p) + jm\varepsilon\rho(t_n)$。简言之，对于所有可以离开集合 $V_{j-} \cup V_j$ 的顶点，这个上界都成立。现在我们利用引理10得出，V_j 中的所有顶点的值都必须有界于 $\mathrm{Rew}^{\nu}(t_p) + jm\varepsilon\rho(t_n)$。□

现在我们将完成定理 2 的证明。我们知道 $\rho(t_p) < \rho(t_{p+1})$。利用这一点以及引理 9 和 11，我们可以将 ε 取得足够小，使得对于所有 $u \in V_{i-}$ 和所有 $u' \in V_i$，都有 $\mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(u) < \mathrm{Rew}^*_{\mathcal{M}[\mathsf{val}]}(u')$。□

这一从加权定时自动机到非加权定时自动机的归约给出了以下结果。请注意，该结果略强于论文中的表述。事实上，它基本上说明，即使对于加权参数化马尔可夫链，NWR 也是困难的。

推论1. 给定一个加权˜参数化马尔可夫决策过程 $\mathcal{M} = (S, A, \Delta, T, \rho)$、一个非空顶点集合$W \subseteq V$ 以及一个顶点 $v \in V$，判定 $v \trianglelefteq W$ 是否为NP的补类-完全问题，即使对于所有 $u \in S$，$|uE| = 1$ 也成立。

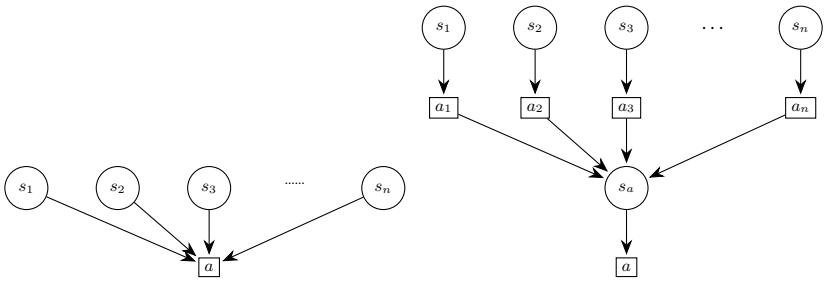
证明。 直接由 Le Roux 和 Pérez [15, 定理 4], 定理 2 以及定理 2 中的事实可得：如果 $uE \le 1$ 对于 $u \in S$ 中一个不是目标顶点的顶点成立，那么对于 $\mathcal{N}$ 中的此类顶点，$uE \le 1$ 也成立。□

D 定理 4 的证明

我们首先展示一种简单的方法，将目标竞技场 [15] 转换为 ˜参数化马尔可夫决策过程，从而使 NP的补类完全性证明仍然成立。

D.1 从目标竞技场到平凡参数化马尔可夫决策过程

只有当目标竞技场中存在多个玩家顶点，它们都连接到同一个自然顶点时，目标竞技场才不会是一个 ˜参数化马尔可夫决策过程（可以将玩家顶点视为状态，将自然顶点视为动作）。在马尔可夫决策过程中这是不可能的，因为每个动作恰好只有一条入边。假设玩家顶点 s_1、.....、 s_n 都连接到自然顶点 a。现在，我们将创建一个等价的目标竞技场，其中包含一个新的玩家顶点 s_a ，它只有一条出边，指向 a 和 n 这些新的自然顶点 a_1、.....、 a_n，使得对于 $0 < i \leq n$， a_i 有一条来自 a_i 的边，并有一条指向 s_a 的边。如图 9 所示。



（a）具有多个入边的自然顶点进入的边。　（b）修改后的竞技场。

图 9。 如何将动作具有多条进入边的目标竞技场转换为每个动作恰好具有一条进入边的竞技场的说明。

读者可以自行验证，这确实是一个 ˜参数化马尔可夫决策过程，并且可达概率与上述构造相比没有任何变化。

现在，我们开始证明定理 4。我们将证明分为两部分。首先，我们证明：给定一个加权参数化马尔可夫决策过程 $\mathcal{M} = (S, A, X, \delta, T, \rho)$、一个顶点 $v \in V$ 以及一个非空子集 $W \subseteq V$，判定 $v \trianglelefteq W$ 是否为 **coETR**-困难问题。接下来，我们将给出一个与询问上述问题的补问题等价的 **ETR** 公式。

D.2 判定 NWR 是否为 coETR-困难问题

定义 9（BCon4Ineq）。有界不等式合取问题：给定一组四次多项式 f_1, $\cdots$, f_m，是否存在某个赋值 $\mathsf{val}: X \to (0,\ 1)$，使得 $\bigwedge_{i=0}^{m} f_i(\mathsf{val}) < 0$?

已知 BCon4Ineq 是 **ETR**困难的 [13, 引理 5]。

为了证明 **coETR**-困难性，我们首先将 BCon4Ineq 归约为判定给定 $\mathcal{M}$ 是否存在保图赋值的问题。令$f_1, \cdots, f_m$为多项式。令 k为一个常数，使得 $k > \max_{1 \le i \le m}\{k_i\}$，其中 k_i 是多项式f_i的系数绝对值之和。我们定义新的多项式 $f'_1, \ldots, f'_m$，其中 $f'_i = \frac{f_i}{k}$。注意，对于任意 val : $X \to (0, 1)$，当且仅当 $\bigwedge_{i=0}^{m} f'_i[\text{val}] < 0$时，有 $\bigwedge_{i=0}^{m} f_i[\text{val}] < 0$；并且对于任意这样的 val，都有 $-1 < f'_i[\text{val}] < 1$。令 $X = \{x_1, \ldots, x_n\}$为 f_i 中出现的所有变量的集合，其中 $1 \le i \le m$。

我们现在定义参数化马尔可夫决策过程 $\mathcal{M} = (S, A, X, \delta, T)$ 如下：

$$
\begin{aligned}
S &= \{s1, \ldots, s_m\} \cup \{s'_1, \ldots, s'_n\} \cup \{fail, fin\},\\
A &= \{n_1, \ldots, n_m\} \cup \{v_1, \ldots, v_n\},\\
X &= \{x_1, \ldots, x_n\},\\
\delta(s_i, n_i, fail) &= -f'_i \text{ for } 1 \le i \le m,\\
\delta(s_i, n_i, fin) &= 1 + f'_i \text{ for } 1 \le i \le m,\\
\delta(s'_i, v_i, fail) &= x_i \text{ for } 1 \le i \le n,\\
\delta(s'_i, v_i, fin) &= 1 - x_i \text{ for } 1 \le i \le n,\\
T' &= \{fail, fin\}.
\end{aligned}
$$

从集合 $\{n_1, \ldots, n_m\}$到失败的所有边都确保 $0 < -f'_i < 1$对于

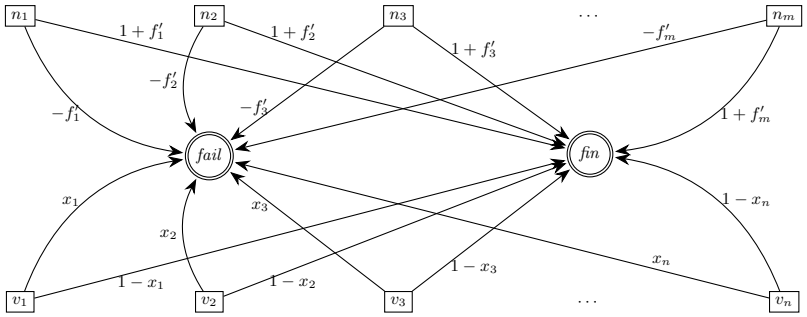


图 10。 用于证明确定参数化马尔可夫决策过程中是否存在保图赋值为 coETR-困难的参数化马尔可夫决策过程。

$1 \le i \le m$ 以及通向终止状态的边确保了 $0 < 1 + f'_i < 1$。这两点都说明了 $-1 < f'_i < 1$。从集合 $\{v_1, \ldots, v_n\}$到 {失败状态和终止状态}的边确保变量只能取自开集 (0, 1)。因此， $\mathcal{M}$ 具有保图赋值，当且仅当存在某个 val : $X \to (0, 1)$，使得 $0 < -f'_i, 1 + f'_i < 1$。但请注意，根据我们为 f'_i构造 $1 \le i \le m$的方式， f_i的取值范围始终包含于区间 $(-1, 1)$ 中。因此， $\mathcal{M}$ 具有保图赋值，当且仅当存在某个 val : $X \to (0, 1)$ 使得$f'_i < 0$，而这又当且仅当 $f_i < 0$ 成立。最后，我们注意到，终止状态 $\not\trianglelefteq$失败，当且仅当我们构造的参数化马尔可夫决策过程存在保图赋值。证明完毕。□

D.3 在 coETR 中判定 NWR

ETR 公式的长度定义为书写该公式所使用的符号数量。我们假设 δ 中的系数以及 ETR 公式中的常数采用二进制编码。此外，我们用 $|\delta|$ 表示某个多项式在其像中的最大表示规模：即其所有系数 c的 $\log_2 c$ 之和。

引理 12。考虑一个加权参数化马尔可夫决策过程 $\mathcal{M}=(S,\ A,\ X,\ \delta,\ T,\ \rho)$、一个顶点 $v \in V$和一个顶点子集 $W \subseteq V$。我们可以在多项式时间内构造一个长度为 $O(|S|^2|A|^2|\delta|)$ 的 *ETR* 句子$\Phi_{\neg wpNWR}$，使得当且仅当不存在图保持赋值 $\mathsf{val} \in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}$，满足对所有 $w \in W$ 都有$\mathrm{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(v) > \mathrm{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(w)$时，$\varphi$ 为真。

其思路是利用引理 5 中描述的等价记法，将奖励的概念编码为逻辑公式。

对于编码，对所有顶点 $v \in V$，我们都有一个变量 y_v，以及 X 中的所有变量。我们定义 $Y=\{y_v \mid v \in V\}$。我们将 V 划分为 4 个集合$\{T,Z,P,N\}$，其中 T 是目标状态集合，Z是不存在通往任何目标状态的路径的顶点集合，$P = S \setminus (T \cup Z)$ 和 $N = S_N \setminus Z$（回忆一下，S_N 是所有状态-动作对的集合）。该划分可在多项式时间内计算。我们首先在 ETR 中编码顶点的值。

首先，我们对目标状态顶点以及没有路径通往目标状态的顶点的值进行编码：

$$\Phi_T = \bigwedge_{v \in T} (y_v = \rho(v)) \quad \text{and} \quad \Phi_Z = \bigwedge_{v \in Z} (y_v = 0).$$

接下来，利用一个状态的值恰好等于其子节点值的最大值这一事实，对各状态的值进行编码：

$$\Phi_P = \bigwedge_{v \in S} \Big(\Big(\bigwedge_{u \in vE} y_v \geq y_u \Big) \wedge \Big(\bigvee_{u \in vE} y_v = y_u \Big) \Big).$$

现在，我们对状态-动作对的值进行编码；该值是其子节点值的凸组合，且组合方式由 δ 决定：

$$\Phi_N = \bigwedge_{v \in S_N} \Big(y_v = \sum_{u \in vE} \delta(v,u) \cdot y_u \Big).$$

因此，我们得到以下公式，用于对 $\mathcal{M}$ 中所有顶点的奖励进行编码：

$$\Phi_{\mathrm{Rew}^{*}_{\mathsf{val}}} = \Phi_T \wedge \Phi_Z \wedge \Phi_P \wedge \Phi_N.$$

读者可以利用引理 5 轻松验证，$\Phi_{\mathrm{Rew}^{*}_{\mathsf{val}}}$ 有唯一解；对于任意 $v \in V$，该解都将顶点 v 的最优奖励赋给 y_v。

现在，我们将对保持图结构条件进行编码。该条件规定，从一个状态-动作对出发的任何出边的概率都不能为 0，并且从 S_N 中某个顶点出发的所有边上的概率之和必须为 1。

$$\Phi_{gp} = \Big(\bigwedge_{\substack{v \in S_N \\ u \in vE}} \delta(v,u) > 0 \Big) \wedge \Big(\bigwedge_{v \in S_N} \sum_{u \in vE} \delta(v,u) = 1 \Big).$$

最后需要编码的是不劣关系的补集：

$$\Phi_{v \not\trianglelefteq W} = \bigwedge_{w \in W} y_v > y_w.$$

现在，NWR 的补集在 ETR 中的最终编码为：

$$\Phi_{\neg \mathrm{wpNWR}} = \exists X \exists Y : \Phi_{\mathrm{Rew}^*_{\mathsf{val}}} \wedge \Phi_{gp} \wedge \Phi_{v \not\trianglelefteq W}.$$

这为我们提供了一种 ETR 编码，用于判定 $W \subseteq V$ 是否并非永不劣于 $v \in V$，这表明检查永不劣于关系属于**coETR**。

E 定理 5 的证明

我们首先证明，在 pMDP 和 ˜pMDP 中判定等价关系分别是**coETR**-完全和NP的补类-完全的。

E.1 将 NWR 判定问题归约为等价性判定问题

令 $\mathcal{M} = (S, A, X, \delta, T)$ 为任意一个参数化马尔可夫决策过程，令 u 和 $v \in S$ 为该参数化马尔可夫决策过程中的任意两个状态。现在我们将构造另一个参数化马尔可夫决策过程 $\mathcal{M}' = (S', A', X, \delta', T)$，其中 $S' = S \cup \{p, q\}$ 和 $A' = A \cup \{b, c\}$，使得 $p \sim q$ 在 $\mathcal{M}'$ 中当且仅当 $u \trianglelefteq v$ 在 $\mathcal{M}$ 中。我们通过如下方式定义 δ'：

$$\delta'(s,a,s') = \begin{cases} \delta(s,a,s') & \text{if } s, s' \in S \text{ and } a \in A, \\ 1 & \text{if } s \in \{p,q\},\ a = b \text{ and } s' = v, \\ 1 & \text{if } s = p,\ a = c \text{ and } s' = u, \\ 0 & \text{otherwise.} \end{cases}$$

这在图 11 中有所描述。

证明基于以下事实：$q \sim v$；并且 $v \sim p$ 成立，当且仅当不存在某个赋值使得 u 的奖励大于 v 的奖励，即 $u \trianglelefteq v$。

E.2 在多项式时间复杂度类中寻找马尔可夫链的等价类

我们首先将 参数化马尔可夫链 $\mathcal{M} = (S, A, X, \delta, T)$ 定义为一种参数化马尔可夫决策过程，其中每个非目标状态恰好有一条边指向一个动作。

为简化记号，我们将状态与其唯一的动作合并，以得到与参数化马尔可夫链等价的模型。从此处起，我们将这些模型称为参数化马尔可夫链。也就是说，现在我们只有带有指向其他状态的边的状态，而边上标有多项式。

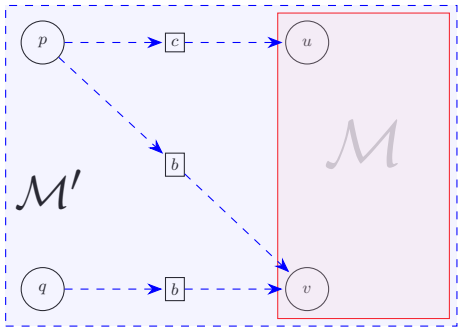


图 11. 马尔可夫决策过程 $\mathcal{M}'$ 其中 $p \sim q$ 当且仅当 $u \trianglelefteq V$ 在 $\mathcal{M}$.

定义 10（平凡参数化马尔可夫链）。一个平凡参数化的马尔可夫链 $\mathcal{M} = (S, E, T)$ 是一个参数化马尔可夫链，其中每条边上的变量都是唯一的。

注意，我们省略变量集合 X，因为每条边都有一个唯一的变量，因此，对概率的唯一限制是：从任意状态出发的出边所诱导的分布必须具有满支撑。因此，我们只关注该链的底层图，即状态集合 S、边集合 E 以及目标状态集合 $T = \{$fail、fin$\}$。

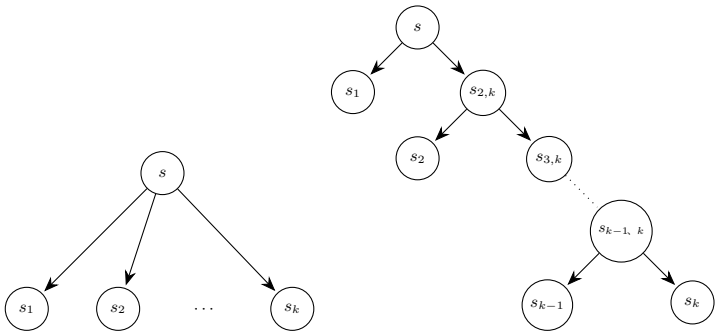
请注意，针对参数化马尔可夫决策过程 ˜所示的类似规约并不适用于马尔可夫链，因为该规约使用了具有多个选择的状态。

在给出用于计算／判定等价关系的多项式时间算法之前，我们先介绍一些预备知识。本节中，我们固定将 $\mathcal{M} = (S, E, T)$ 作为所使用的马尔可夫链。设 $\mathcal{M}$ 中的 NWR 等价类数量为 n。为简化证明，我们作出若干假设。不失一般性，我们假设与 fin 和 fail 等价的顶点分别只有 fin 和 fail。接下来，我们假设不存在自环，因为如果状态 s 具有概率为 x 的自环，我们可以将 s 的所有其他出边上的概率乘以 $\frac{1}{1-x}$（这样的边至少存在一条，因为 s 存在一条通往 fin 的路径）；读者可以轻易验证，新的赋值保持图结构，并且与原赋值具有相同的可达概率。我们还假设每个非目标状态至少有两条出边，因为如果某个状态没有出边，那么它没有通往 fin 的路径，因此根据前述假设，它会与 fail 合并；如果它恰好有一条出边，那么它与其唯一后继状态等价，因此我们可以将它与该后继状态合并。请注意，合并可能会产生新的自环，但这一过程最终会终止，因为状态数量有限。

我们证明，对于任意马尔可夫链，都可以构造一个等价的马尔可夫链，使得在新马尔可夫链中，每个非目标状态都恰好有 2 条出边。

引理 13. 对于任意给定的马尔可夫链 $\mathcal{M}$，我们可以构造另一个马尔可夫链 $\mathcal{N} = (S', E', T)$，使得 $sE' = 2$ 对于所有 $s \in S' \backslash T'$、$S \subseteq S'$，对于所有值 $\in$ 值域$^{\text{gp}}_{\mathcal{M}}$，都存在值$' \in$ 值域$^{\text{gp}}_{\mathcal{N}}$，使得对于任意 $s \in S$、$\mathbb{P}^{s}_{\mathcal{M}^{\text{val}}}[\lozenge \text{终止}] = \mathbb{P}^{s}_{\mathcal{N}^{\text{值}'}}[\lozenge \text{终止}]$，反之亦然。

证明。 图 12 描述了我们如何利用新状态确保每个状态恰好有两个后继状态。



(a) 一个具有 $sE = k$ 的状态。 (b) 具有 $k-2$ 个新状态的新马尔可夫链。

图 12. 说明如何通过添加 $|sE| - 2$ 个新状态，将一个拥有两个以上出边的状态转换为恰好拥有两条出边的状态。

令 val$(s, s_i) = x_i$ for $1 \leq i \leq k$. 对所有 i，我们都有 $x_i > 0$ for，且 $val \in \text{Val}^{\text{gp}}_{\mathcal{M}}$ 时也有 $\sum_{1 \leq i \leq k} s_i = 1$。现在我们为图 12b 中的所有边定义 val′： val′$(s_{i,k}, s_i) = \frac{x_i}{1-\sum_{j<i} x_j}$，以及 val′$(s_{i,k}, s_{i+1}, k) = \frac{1-\sum_{j<i} x_j}{1-\sum_{j<i} x_j}$，其中我们考虑 $s_{1,k} = s$ 和 $s_{k,k} = s_k$。现在读者可以轻松验证，对于任意 $1 \leq i \leq k$，从 s 到 s_i 的唯一路径上的概率恰好为 x_i。

对于反方向，令 val′ $\in \text{Val}^{\text{gp}}_{\mathcal{N}}$。我们将 val$(s, s_i)$ 定义为由 val′ 诱导的路径 $s, s_{2,k}, \ldots, s_{i,k}, s_i$ 上各边概率的乘积。容易验证，如果 val′ 是一个图保持赋值，那么 val 也是。

现在，由于我们已经证明，可以从 $\mathcal{M}$ 或 $\mathcal{N}$ 上给定的赋值构造出新的赋值，并且这些赋值能保持在子图中从 s 到达 s_i 的概率不变，因此可以推断，新赋值下到达终止的概率必定与旧赋值下相同。□

引理 14. 如果一个顶点 $s \in S$ 几乎必然能够到达一个等价类 A，那么 $s \in A$。

证明。 由于 $\mathcal{M}$ 是一条马尔可夫链， s 几乎必然到达 A 意味着 A 必须是 s 的本质状态。令 $v \in A$ 为 A 中任意一个状态，且 val $\in \text{Val}^{\text{gp}}_{\mathcal{M}}$。利用

引理 7，我们得到

$$\begin{aligned}\mathbb{P}^{s}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathit{fin}] &= \sum_{u\in A} \mathbb{P}^{s}_{\mathcal{M}^{\mathsf{val}}}[(S\setminus A)\cup u]\cdot \mathbb{P}^{u}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathit{fin}],\\ &= \mathbb{P}^{v}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathit{fin}]\cdot \sum_{u\in A} \mathbb{P}^{s}_{\mathcal{M}^{\mathsf{val}}}[(S\setminus A)\cup u],\\ &= \mathbb{P}^{v}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathit{fin}].\end{aligned}$$

由于 val 是任意选取的，这意味着 s 与 v 等价，因此属于等价类 A。□

引理 15. 对于所有状态 $s \notin T$ 以及所有赋值 $\mathsf{val} \in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}$，有 $0 < \mathbb{P}^{s}_{\mathcal{M}^{\mathsf{val}}}[\Diamond$终止$] < 1$。

证明。 我们假设 fin 和 fail 是各自等价类中仅有的状态。这意味着每个其他顶点都必须有一条通往 fin 的路径（否则它们的值总是为 0），也必须有一条通往 fail 的路径（否则它们几乎必然会到达 fin，因而属于同一个等价类）。由于 val 保持图结构，这些通往 fin 和 fail 的路径的概率都必须非零，因此 $\mathbb{P}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathsf{fin}]$ 既不能为 0，也不能为 1。□

现在我们将给出两种证明。第一种证明简短而简单，第二种证明较长，但很精彩。

第一个证明 假设每个等价类中恰好有一个状态，其边指向该等价类之外，那么该等价类中的所有其他状态几乎必然都会到达该状态。再结合几乎必然到达蕴含状态等价这一事实，我们便可以通过检查 S 中所有状态对之间的几乎必然可达性，轻松设计算法来求出马尔可夫链中的等价类。

若满足 $s \in U$ 且 $sE \not\subseteq U$，则状态 s 是集合 $U \subseteq S$ 的出口。

引理 16. 若 u 是等价类 U 的出口，则 u 的两个后继状态都必须位于 U 之外，即 $uE \cap U = \emptyset$。

证明。 我们知道，u 必须有一个后继状态位于 U 之外，设为 a。为导出矛盾，我们假设另一个后继状态（设为 b）位于 U 内。现在，对于任意给定的赋值 $\mathsf{val} \in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}$，我们有

$$\begin{aligned}\mathbb{P}^{u}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathit{fin}] &= \mathsf{val}(u,a)\mathbb{P}^{a}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathit{fin}] + \mathsf{val}(u,b)\mathbb{P}^{b}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathit{fin}]\\ &= \mathsf{val}(u,a)\mathbb{P}^{a}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathit{fin}] + (1-\mathsf{val}(u,a))\mathbb{P}^{b}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathit{fin}] \quad \text{since } \mathsf{val} \in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}} \text{ and } |uE| = 2\\ &= \mathsf{val}(u,a)\mathbb{P}^{a}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathit{fin}] + (1-\mathsf{val}(u,a))\mathbb{P}^{u}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathit{fin}] \qquad \text{since } u \text{ and } b \text{ are equivalent}\\ &= \mathbb{P}^{a}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathit{fin}].\end{aligned}$$

由于 val 的选取是任意的，我们得到 a 必定属于 U，这就产生了矛盾。□

定理 7。 任何非平凡等价类（即不包含终止 或 失败的等价类）都恰好有一个出口。

证明。 首先，我们注意到，每个状态都有一条通往终止的路径，因此一个等价类至少有一个出口。现在，我们将证明它不可能有多个出口。证明采用反证法。假设存在一个等价类$U \subseteq S$，它至少有两个出口，例如 a 和 b。由于 a 是一个出口，根据引理 14 和引理 16，存在两个状态c 和 d，使得等价类 $\tilde{c}$ 和 d 彼此不相交，且都与 U 不相交。由于 c 与 a并非永不更差等价，因此必定存在一个赋值 $\mathsf{val} \in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}$，使得$\mathbb{P}^{c}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathsf{fin}] > \mathbb{P}^{a}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathsf{fin}]$，或者 $\mathbb{P}^{c}_{\mathcal{M}^{\mathsf{val}}}[\Diamond\ ] < \mathbb{P}^{a}_{\mathcal{M}^{\mathsf{val}}}[\Diamond\ ] < \mathbb{P}^{d}_{\mathcal{M}^{\mathsf{val}}}[\Diamond\ ]$ 或 $\mathbb{P}^{c}_{\mathcal{M}^{\mathsf{val}}}[\Diamond\ ] > \mathbb{P}^{a}_{\mathcal{M}^{\mathsf{val}}}[\Diamond\ ] > \mathbb{P}^{d}_{\mathcal{M}^{\mathsf{val}}}[\Diamond\ ]$。若前者成立，则 $\mathsf{fin}\ \mathsf{fin}\ \mathsf{fin}$；若后者成立，则 $\mathsf{fin}\ \mathsf{fin}\ \mathsf{fin}$，因为 a 的奖励值是 c 和 d的奖励值的凸组合。不失一般性，假设$\mathbb{P}^{c}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathsf{fin}] < \mathbb{P}^{a}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathsf{fin}] < \mathbb{P}^{d}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathsf{fin}]$。令 $m = |\{\mathsf{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}u | u \in S\}|$。现在，我们使用 val 构造一个划分 $\{S_0, \ldots, S_m\}$，使得 $\mathsf{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}p < \mathsf{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}q$ 当且仅当 $p \in S_i$ 和 $q \in S_j$ 满足 $i < j$。在这个划分中，与定理 2 的证明类似，我们将每个 S_i 称为“层”，其中$S_0 = \{\mathsf{fail}\}$ 是底层，而 $S_m = \{\mathsf{fin}\}$ 是顶层。这个划分具有以下一些性质：

– 每一层恰好由以 val 到达 fin 的概率相同的状态组成；到达概率越高的层，位置越高。– 每个等价类必须完全包含在某一层中。– 除最顶层和最底层外，每一层都必须至少有一个状态离开该层（因为每个状态都必须有一条通往 fin 的路径）。– 每个这样的出口都必须有一条边通往上方的一层，另一条边通往下方的一层（因为从一个状态到达 fin 的概率是其后继状态对应概率的凸组合）。– 从任意状态出发，都必须存在一条通往离开该层的状态的路径，且该路径完全包含在该层内。– 从任意离开某层的状态出发，都存在一条通往 fin 的路径，使路径中的第二个状态的值严格更大，且路径上各状态的值单调不减（由前两条性质可得）。– 从任意离开某层的状态出发，都存在一条通往 fail 的路径，使路径中的第二个状态的值严格更小，且路径上各状态的值单调不增（理由与上一条性质相同）。

~根据我们的假设，可以看出 $\tilde{c}$ 位于包含 U 的层之下，而 d 位于 U 之上。设 S_i 为包含 U 的层。

情况 1： b 也会离开 S_i。根据划分的最后两个性质，我们可以观察到，存在从 a 到终止状态以及从 b 到失败状态的顶点不相交路径。由于这些路径的顶点不相交，我们可以定义一个新的赋值，使这条路径上的每条边都具有概率$1-\varepsilon$（对于某个较小的 ε）。这样就得到一个保持图结构的赋值，在该赋值下，从 a 到达终止状态的概率接近 1，而从 b 到达终止状态的概率接近 0（因为 b 以接近 1 的概率到达失败状态）。这与 $a \sim b$矛盾。因此，这种情况不可能发生。

情况 2： b 不会离开 S_i。由于 b 是 U的一个出口，这意味着存在一个状态 $e \in bE \cap （S_i \backslash U）$。我们声称， e 必须存在一条通往另一个出口$f \in S_i \backslash U$ 的路径，且该路径不会经过 U中的任何状态。如果情况并非如此，就意味着从 e 到终止状态的每条路径都必须经过 U中的一个状态。现在，我们使用引理 7 得到：

$$\begin{aligned}\mathbb{P}^e_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathit{fin}] &= \sum_{u \in U} \mathbb{P}^e_{\mathcal{M}}[(S \setminus U) \cup u] \cdot \mathbb{P}^u_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathit{fin}] \\ &= \sum_{u \in U} \mathbb{P}^e_{\mathcal{M}}[(S \setminus U) \cup u] \cdot \mathbb{P}^e_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathit{fin}] \quad e \text{ is in the same layer as } U \\ &= \mathbb{P}^e_{\mathcal{M}^{\mathsf{val}}}[\Diamond \mathit{fin}] \cdot \sum_{u \in U} \mathbb{P}^e_{\mathcal{M}}[(S \setminus U) \cup u]\end{aligned}$$

$\sum_{u \in U} \mathbb{P}^e_{\mathcal{M}}[(S \setminus U) \cup u] = 1.$

这意味着 e 几乎必然会到达 U，而根据引理 14， $e \in U$，这与我们的假设矛盾。因此， b 有一条路径通向同一层中的另一个状态（例如 f），并从该层离开；该路径中除 b 外，没有其他状态属于 U。现在，我们有从 a 到终止以及从 f 到失败的状态不相交路径，因此也有从b 到失败的路径。我们可以定义一个新的赋值，使这条路径中的每条边都具有概率 $1-\varepsilon$ （其中 ε 足够小）。由此得到一个图保持赋值，使得从 a 到达终止的概率接近 1，而从 b 到达终止的概率接近 0。这再次与 $a \sim b$ 的事实矛盾。

由于这两种情况不可能同时发生，我们可以得出结论：一个等价类不可能有多个出口。□

既然我们已经证明每个等价类恰有一个出口，便可立即得出，该等价类中的每个状态几乎必然都会到达这个唯一的出口。

第二个证明 回想一下，对于状态 $s \in S$， $\tilde{s}$ 表示 s的 NWR 等价类。

定义 11（等价划分）。马尔可夫链 $\mathcal{M}$ 的等价划分是一个划分 $\{S_0, \ldots, S_m\}$，满足以下条件：

- $S_0 = \{$失败$\}$、 $S_m = \{$终止$\}$，且对于任意 $u \in S$、 $\tilde{u} \subseteq S_i$，存在某个 $0 < i < m$；- 任何 $u \in S_j$ 要么与某些 S_i 和 S_k 之间有边，其中 $i < j < k$，要么它只有指向 S_j内部的边；- S_i 对所有 $0 \le i \le m$都至少有一个顶点

这种划分的各个元素称为层，其中 S_0 是底层， S_m 是顶层（类似于定理 2 证明中的划分）。每一层由若干等价类的并集组成，且该并集非空。

我们将“高于”层 S_i 的所有顶点组成的集合记为 $S_{i+} = S_{i+1} \cup \cdots \cup S_m$，将“低于” S_i 的所有顶点组成的集合记为 $S_{i-} = S_0 \ldots S_{i-1}$。

定义 12（出口）。对于任意状态划分，如果状态 s 所在的顶点至少有一条边指向另一个划分，则称该状态退出该划分。

注意，除分别包含终止和失败状态的顶层与底层外，每一层至少有一个出口，因为每个状态都可以到达终止状态。

以下引理将引理10推广到马尔可夫链中的状态：只要这些状态存在一条通往至少一个出口的路径，就可以为其给出下界。

引理17。考虑一个马尔可夫链 $\mathcal{M}$，并令 $U \subseteq S \backslash T$、值 $\in$ 值$^{\text{gp}}_{\mathcal{M}}$ 和 V 表示满足 $sE \setminus U \neq \emptyset$ 的状态集合 s。若 $|V| \neq 0$，且对所有 $s \in V$ 都有 $x \leq \mathbb{P}^{s}_{\mathcal{M}^{\mathsf{val}}}[\Diamond$ 终止 $] \leq y$，那么对所有 $u \in U$，都有 $x \leq \mathbb{P}^{u}_{\mathcal{M}^{\mathsf{val}}}[\Diamond$终止$] \leq y$。

证明。上界的证明与引理10的证明相同。我们用矛盾法证明下界。假设存在一个状态 $u \in U$，使得 $x_u = \mathbb{P}^{u}_{\mathcal{M}^{\mathsf{val}}}[\Diamond$终止$] < x$。令 u 为 U 中值最小的顶点。由于$uE = 2$，u 的值必定是其两个后继状态（设为 u' 和 u''）的凸组合。但是，根据假设，u 在 U 中的值最小，且 $u \notin V$，可知 u'、u'' 的值必须与 u 相同，因此它们也都不在 V 中。我们对 u' 和 u'' 继续这一过程。由于状态数量有限，该过程必定会停止。将可通过u 到达的这些状态组成的集合记为 S_u。S_u 中每个状态的奖励值都等于 x_u。根据假设，U 中的每个状态都有一条通往终止状态的路径，因此有终止状态 $\in S_u$，这与 $\mathbb{P}^{fin}_{\mathcal{M}^{\mathsf{val}}}[\Diamond$终止$] = 1 > x_u$ 矛盾。□

引理1对于任意等价划分 $\{S_0, \ldots, S_m\}$，存在一个赋值 $\mathsf{val} \in \mathsf{Val}^{\text{gp}}_{\mathcal{M}}$，使得对于任意 $s_i \in S_i$ 和 $s_j \in S_j$，若有 $i < j$，则有 $\mathbb{P}^{s_i}_{\mathcal{M}^{\mathsf{val}}}[\Diamond$终止$] < \mathbb{P}^{s_j}_{\mathcal{M}^{\mathsf{val}}}[\Diamond$终止$]$。

证明。令 $0 < \varepsilon_{m-1} < \cdots < \varepsilon_1 < 1$ 足够小。令 $\{E_1, \ldots, \mathbf{E}_{m-1}\}$ 分别为集合 S_1、.....、S_{m-1} 中所有出口构成的集合。我们定义赋值val 如下：

$$\mathsf{val}(s, s') = \begin{cases} 0.5 & \text{if } s, s' \in S_i \text{ for some } 0 < i < m; \\ 1 - \varepsilon_i & \text{if } s \in S_i \text{ and } s' \in S_{i+} \text{ for some } 0 < i < m; \\ \varepsilon_i & \text{if } s \in S_i \text{ and } s' \in S_{i-} \text{ for some } 0 < i < m. \end{cases}$$

也就是说，同一层内的所有边都被赋予概率 0.5，从层 S_i 离开的“向上”边被赋予概率 $1 - \varepsilon_i$，而“向下”边被赋予概率 ε。现在，我们来证明，以这种方式赋予概率可以为一层中的顶点给出一些有趣的上界和下界。

命题 2. 令 val 如上所定义。那么对于任意 $s \in S_i$，存在某个$0 < i < m$

$$(1 - \varepsilon_{m-1}) \cdot \cdots \cdot (1 - \varepsilon_i) \leq \mathbb{P}^{s}_{\mathcal{M}^{\mathsf{val}}}[\Diamond fin] \leq 1 - \varepsilon_1 \cdot \cdots \cdot \varepsilon_i.$$

证明。 这两个界都通过归纳法证明。首先，我们从顶层 S_{m-1} 开始归纳，证明下界。在这一层中，任何出口都有一条通往终止的边，其概率为 $1-\varepsilon_{m-1}$（另一条边可能通往失败）。因此，根据引理17， S_{m-1} 中的所有顶点都满足该界。现在，假设对于任意满足 $j>i$ 的 $s'\in S_j$，界 $(1-\varepsilon_{m-1})\cdot\dots\cdot(1-\varepsilon_j)\leq \mathbb{P}^{s'}_{\mathcal{M}^{\text{val}}}[\Diamond \text{终止}]$ 均成立。令 s 为 S_i 中的一个出口。我们有

$$\begin{aligned}\mathbb{P}^{s}_{\mathcal{M}^{\text{val}}}[\Diamond \mathit{fin}] &\geq (1-\varepsilon_i)\min_{j>i}\{\mathbb{P}^{s'}_{\mathcal{M}^{\text{val}}}[\Diamond \mathit{fin}] | s'\in S_j\}\\ &\geq (1-\varepsilon_i)\min_{j>i}\{(1-\varepsilon_{m-1})\cdot\dots\cdot(1-\varepsilon_j)\}\\ &= (1-\varepsilon_i)(1-\varepsilon_{m-1})\cdot\dots\cdot(1-\varepsilon_{i+1}).\end{aligned}$$

因此，该层中所有出口都满足下界；根据引理17，该层中的所有顶点也都满足下界。

现在，为了证明上界，我们从底层 S_1开始进行归纳。该层中的任何出口都会以概率 ε_1 转移到失败状态，因此其值至多为 $1-\varepsilon_1$。根据引理17，这个界对 S_1中的所有顶点都成立。现在，假设对于任意满足 $j<i$的 $s'\in S_j$，界 $\mathbb{P}^{s'}_{\mathcal{M}^{\text{val}}}[\Diamond \text{fin}]\leq 1-\varepsilon_1\cdot\dots\cdot\varepsilon_j$ 都成立。令 s 为 S_i中的一个出口。在最佳情形下，它可以以概率 $1-\varepsilon_i$ 有一条边通向终止状态，并以概率 ϵ_i通向 S_{i-} 中的“最佳”顶点。我们有

$$\begin{aligned}\mathbb{P}^{s}_{\mathcal{M}^{\text{val}}}[\Diamond \mathit{fin}] &\leq 1-\varepsilon_i+\varepsilon_i\cdot\max_{j<i}\{\mathbb{P}^{s'}_{\mathcal{M}^{\text{val}}}[\Diamond \mathit{fin}] | s'\in S_j\}\\ &\leq 1-\varepsilon_i+\varepsilon_i\cdot\max_{j<i}\{1-\varepsilon_1\cdot\dots\cdot\varepsilon_j\}\\ &= 1-\varepsilon_i+\varepsilon_i\cdot(1-\varepsilon_1\cdot\dots\cdot\varepsilon_{i-1})\\ &= 1-\varepsilon_1\cdot\dots\cdot\varepsilon_i.\end{aligned}$$

因此，所有出口都满足上界，根据引理17，S_i 中的所有状态也满足该上界。□

回到引理的证明，只需证明存在 $0<\varepsilon_{m-1}<\dots<\varepsilon_1<1$，使得任意层 S_i 的下界都大于层 S_{i-1}的上界。更准确地说，我们要确保对于任意 $1<i<m$，都有 $(1-\varepsilon_{m-1})\cdot\dots\cdot(1-\varepsilon_i)>1-\varepsilon_1\cdot\dots\cdot\varepsilon_{i-1}$。我们有

$$\begin{aligned}(1-\varepsilon_{m-1})\cdot\dots\cdot(1-\varepsilon_i) &> (1-\varepsilon_i)^{m-i} && \varepsilon_i>\varepsilon_j \text{ for } j>i\\ &> (1-\varepsilon_i)^n\end{aligned}$$

and

$$\begin{aligned}1-\varepsilon_1\cdot\dots\cdot\varepsilon_{i-1} &< 1-\varepsilon_{i-1}^{i-1} && \varepsilon_{i-1}<\varepsilon_j \text{ for } j<i\\ &< 1-\varepsilon_{i-1}^{n}.\end{aligned}$$

现在，我们可以轻松地选择 ε_i 小得多的 ε_{i-1}，使得 $(1-\varepsilon_i)^n>1-\varepsilon_{i-1}^n$成立。因此，这样的 $0<\varepsilon_{m-1}<\dots<\varepsilon_1<1$ 序列存在，从而完成证明。□

引理19。 任意赋值$\mathsf{val} \in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}$ 都会诱导出一个等价划分 $\{S_0, \ldots, S_m\}$，使得两个状态 s， $s' \in S$ 处于同一层，当且仅当$\mathbb{P}^{s}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \text{终止}] = \mathbb{P}^{s'}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \text{终止}]$。

我们将奖励相同的所有顶点归入同一层，从而定义该划分；随后，按照各层中状态的奖励对这些层进行排序。也就是说，奖励较高的层位于奖励较低的层之上。很容易验证，这确实是一个等价划分。

定义 13（最细赋值）。若保持图结构的赋值 val 对于所有 $u, v \in S$, $\mathbb{P}^{u}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \text{终止}] = \mathbb{P}^{v}_{\mathcal{M}^{\mathsf{val}}}[\Diamond \text{终止}]$ 当且仅当 $u \sim v$，则称其为最细赋值。

定义 14（最细划分）。最细划分是一个等价划分，其层数恰好等于等价类的数量。

引理 20。马尔可夫链总是存在最细赋值。

证明。 假设不存在最细赋值。设 val 是一个赋值，不存在其他赋值使其诱导的等价划分（引理 19）比 val 所诱导的划分具有更多的层。令$\{S_0, \ldots, S_m\}$ 为 val 所诱导的等价划分。由于 val 不是最细赋值，因此存在一层，比如说 S_i，其中至少有两个等价类。现在我们注意到，这些等价类中至少有两个包含会离开 S_i 的状态（否则，如果只有一个等价类包含出口，那么该层上的所有其他等价类几乎必然会到达这个等价类；根据引理 14，这些等价类都将与包含出口的等价类等价，这是不可能的）。现在，我们将其中一个包含出口的等价类移至一个新层 $S_{i'}$ ，该层位于 S_i 和 S_{i+1}之间。我们将移除 $S_{i'}$ 后得到的层称为 $S_{i''}$。注意，在一个等价划分中，每个出口都有一条边向上，一条边向下。再结合任意状态都几乎必然能够到达所有出口组成的集合这一事实，可以保证：若某个等价划分中的状态 u 所在的层高于 v 所在的层，则存在从 u 到 fin 以及从 v 到 fail 的状态不相交路径。考虑新的划分 $\{S_0, \ldots, S_{i''}, S_{i'}, S_{i+1}, \ldots, S_m\}$（它不一定是等价划分）。令 s 为 $S_{i'}$ 中的一个出口，它在原划分中也是 S_i 中的一个出口。也就是说， s 有一条边通向 $S_{i'+}$ ，还有一条边通向 $S_{i''-}$。我们声称，s 是从 $S_{i'}$ （即其等价类）有出边的唯一顶点。假设还有另一个顶点 s' 会离开 $S_{i'}$，那么它必定有一条边通向 $S_{i''}$ 或更低的层。因此，存在从 s 到 fin 以及从 s' 到 fail 的不相交路径。通过使这些路径上的概率接近 1，我们可以使 s 的奖励接近 1，并使 s' 的奖励接近 0。这与二者在 NWR 意义下等价这一事实矛盾。因此， $S_{i'}$ 是等价划分中的一个“有效”层。我们可以用类似的论证说明，S_i 中任何有出口离开 S_i 的等价类，都必须有唯一一个顶点拥有通向该等价类之外的边。但是，原始 S_i 中可能存在一些状态，它们有一条边通向 $S_{i'}$中的某个状态。现在，我们将迭代地构造层，使得每一步中，除 S_i''外的所有层仍然都是有效层。令 $S_{i1} \subset S_{i''}$ 为其中一个有边通向

an edge to $S_{i'}$. We move S_{i1} between $S_{i''}$ and $S_{i'}$. Using the argument from above, this state going up to $S_{i'}$ and down to $S_{i''}$ is the unique exit of this layer. We continue doing this until there are no more new exits (apart from the exits originally in S_i) in $S_{i''}$. This gives us an equivalence partition which is finer than the original one. Now, using Le mma 18, we can get a valuation val′ finer than val which is a contradiction to the assumption. □

现在，我们使用这个最精细的赋值来得到一个最精细的划分，并用它证明每个等价类都有一个唯一的状态，其边指向该等价类之外。

引理 21. 对于任意等价类 $U \subset S \backslash T$，存在唯一的状态 $s \in S$，使得 $sE \not\subseteq U$。

证明。令 val 为一个最精细的赋值。令 $\{S_0, \ldots, S_n\}$ 为一个最精细的划分。因此，每个等价类都有其自身的层。令 $U = S_i$，其中 $0 < i < n$。假设 S_i 有两个出口 s 和 s'。现在，我们可以构造从 s 到 fin （仅向上移动）以及从 s' 到 fail （向下移动）的状态不相交路径。现在，我们可以将这些路径上的边的概率调得接近 1，使得 s 的奖励接近 1，而 s' 的奖励接近 0。这与它们属于同一个等价类这一事实矛盾。□

现在，利用引理21和引理17，我们得到主要结果：马尔可夫链中等价类的所有顶点几乎必然都能到达唯一的出口。

通过检查顶点对之间的几乎必然可达性，可以很容易地为此给出一个多项式时间算法。

定理 8. 计算并合并马尔可夫链的 *NWR* 等价类可以在多项式时间内完成。

F 引理 2 的证明

我们首先陈述以下引理，该引理指出不劣关系具有传递性：

引理 22。 对于任意加权参数化马尔可夫决策过程 $\mathcal{M} = (S, A, X, \delta, T, \rho)$，对于任意 $P, Q, R \subseteq V$，如果$P \trianglelefteq Q$ 且 $Q \trianglelefteq R$，那么 $P \trianglelefteq R$。
读者可以根据 NWR 的定义轻松验证上述引理成立。

现在，为了证明引理 2 并说明式 (2) 成立，根据引理 22，只需证明对于任意 U、$W \in N$，若 $U \to W$ 通过一条边，则 $U \trianglelefteq W$。我们针对以下四种边类型逐一证明：

1. 我们知道，对于任意 $W \subseteq V$，fail $\trianglelefteq W$。因此，从 fail 指向任意节点的边都满足 NWR。

2. 我们知道，对于任意 $W \subseteq V$， $W \trianglelefteq \mathsf{fin}$。因此，从任意节点指向 fin 的边都满足 NWR。3. 根据引理6，对于任意赋值 $\mathsf{val} \in \mathrm{Val}^{\mathrm{sp}}_{\mathcal{M}}$ 和顶点$v \in V$，$\mathrm{Rew}^{*}_{\mathsf{val}}(v) \leq$ 最大值$_{u \in vE}\, \mathrm{Rew}^{*}_{\mathsf{val}}(u)$。因此，从 v 指向 vE 的边满足 NWR。4. 同样，根据引理6，我们知道，对于赋值 $\mathsf{val} \in \mathrm{Val}^{\mathrm{sp}}_{\mathcal{M}}$ 和状态$v \in S$，$\mathrm{Rew}^{*}_{\mathsf{val}}(v) =$ 最大值$_{u \in vE} \mathrm{Rew}^{*}_{\mathsf{val}} u$。因此， v 永远不劣于 vE 中的每个状态-动作对，因此该边满足 NWR。

证明至此结束。 □

G 引理 3 的证明

根据引理 22，如果 $\mathcal{U}'$ 满足不变式 (2)，那么要证明 $\mathcal{U}$ 也满足不变式 (2)，只需证明所有新添加的边都满足 NWR。从 U 和 W 的子集（超集）添加的边满足 NWR，因为根据定义，对于任何 $P \subseteq P' \subseteq V$， $P \trianglelefteq P'$ 都成立。已知 $U \trianglelefteq W$，因此，添加边 $U \to W$（如果不存在从 U 到 W 的路径）也应满足 NWR。

H 定理 6 的证明

证明（定理 *6* 的证明）。为简单起见，我们将状态-动作对 (s, a) 重命名为 x。证明分为两种情况：一种考虑移除该边后 s 可能与终止状态断开的情形；另一种考虑移除该边后 s 仍有一条通往终止状态的路径的情形。

情况 1，即移除边 (s, x) 后， s 没有通往终止的路径：这意味着移除该边后，$sE\backslash\{x\}$ 中的任何顶点都没有通往终止的路径。令$u \in sE\backslash\{x\}$。因此，对于任意 σ，我们有 $\mathbb{P}^{v}_{\mathcal{M}^{\mathsf{val}}_{\sigma}}[\lozenge \text{终止}] = \mathbb{P}^{v}_{\mathcal{M}^{\mathsf{val}}_{\sigma}}[\lozenge s] \cdot \mathbb{P}^{s}_{\mathcal{M}^{\mathsf{val}}_{\sigma}}[\lozenge \text{终止}]$ ，对所有 $v \in uE$ 均成立（根据引理 7）。如果对所有 $v \in uE$都有 $\mathbb{P}^{v}_{\mathcal{M}^{\mathsf{val}}_{\sigma}}[\lozenge s] = 1$ ，那么这意味着 v 和 s 处于同一个极大末端分量中，因此根据我们的假设，即 $\mathcal{M}$ 与 $\mathcal{M}_{/\mathrm{MEC}}$同构，可推出 $v = s$ ，进而得到 $uE = \{s\}$。但这不可能，因为若$uE = \{s\}$，我们就不会有从 s 到 u 的边。这意味着至少存在一个顶点 $v \in uE$ ，使得$\mathbb{P}^{v}_{\mathcal{M}^{\mathsf{val}}_{\sigma}}[\lozenge s] < 1$。因此，我们有 $\mathbb{P}^{u}_{\mathcal{M}^{\mathsf{val}}_{\sigma}}[\lozenge \text{终止}] < \mathbb{P}^{s}_{\mathcal{M}^{\mathsf{val}}_{\sigma}}[\lozenge \text{终止}]$。由于 u 是从 $sE\backslash\{x\}$中任意选取的，并且因为$\mathbb{P}^{u}_{\mathcal{M}^{\mathsf{val}}_{\sigma}}[\lozenge \text{终止}] < \mathbb{P}^{s}_{\mathcal{M}^{\mathsf{val}}_{\sigma}}[\lozenge \text{终止}],$ ，我们得到

$$
\begin{aligned}
\mathrm{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(s) &= \max\{\mathrm{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(u) \mid u \in sE\} \\
&= \max\left\{\left(\max_{u \in sE\backslash\{x\}} \mathrm{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(u)\right) \cup \mathrm{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(x)\right\} \\
&= \mathrm{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(x).
\end{aligned}
$$

现在，我们得到了矛盾：对于 $u \in sE\backslash\{x\}$ ，$\mathrm{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(x) = \mathrm{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(s) > \mathrm{Rew}^{*}_{\mathcal{M}[\mathsf{val}]}(u)$，因为我们的假设是 $x \trianglelefteq (sE \setminus \{x\})$。因此，情况 1 不可能发生。

情况 2，即移除$s \to x$ 边后， s 到终止的连通性没有变化：对于任意赋值 $\mathsf{val} \in \mathsf{Val}^{\mathrm{gp}}_{\mathcal{M}}$，根据引理 5，我们知道对于以下线性方程组

$$\text{For all } s \in S,\ y(s) = \begin{cases} 1 & \text{if } s = \mathit{fin}; \\ 0 & \text{if } s \in \mathbf{Z}_{\mathit{fin}}; \\ \max_{v \in sE} \sum_{s' \in vE} \delta(v, s')[\mathsf{val}] \cdot y(s') & \text{otherwise} \end{cases}$$

有唯一解

$$y(u) = \max_{\sigma} \mathbb{P}^{u}_{\mathcal{M}^{\mathsf{val}}_{\sigma}}[\Diamond \mathit{fin}].$$

由于我们有 $x \trianglelefteq (sE \setminus \{x\})$，因此断言以下方程组具有相同的唯一解：

$$\text{For all } u \in S,\ y(u) = \begin{cases} 1 & \text{if } u = \mathit{fin}; \\ 0 & \text{if } u \in \mathbf{Z}_{\mathit{fin}}; \\ \max_{v \in uE \setminus \{x\}} \sum_{s' \in vE} \delta(v, s') \cdot y(s') & \text{if } u = s; \\ \max_{v \in uE} \sum_{s' \in vE} \mu_v(s') \cdot y(s') & \text{otherwise.} \end{cases}$$

我们已经证明，**情况 1** 不可能发生。也就是说，从 s 到 x 移除边不会改变集合 $\mathbf{Z}_{\mathit{fin}}$。因此，这个修改后的方程组的解具有唯一解：

$$y(u) = \max_{\sigma} \mathbb{P}^{u}_{\mathcal{N}^{\mathsf{val}}_{\sigma}}[\Diamond \mathit{fin}].$$

因此，即使移除边$s \to x$，各顶点的值也保持不变。□