并发随机博弈的计算复杂性：精确值与近似

Ali Asadi 奥地利科学技术研究院 (ISTA)
ali.asadi@ist.ac.at

Krishnendu Chatterjee 奥地利科学技术研究院 (ISTA)krishnendu.chatterjee@ist.ac.at

arXiv:2611.30447v1 [cs.GT] 23 Nov 2026

摘要

我们研究在有限状态空间上进行、持续无限多个阶段的双人零和并发随机博弈。每个阶段中，两位玩家同时且独立地选择一个动作。随后，根据当前状态和所选动作，按照随机转移函数选出后继状态。我们研究极限平均目标，其中玩家1要最大化分配给每次转移的奖励的长期平均值。从初始状态出发的值，是玩家1能够确保的最大期望值。我们研究两个基本计算问题：(a) 在任意给定的加性误差范围内计算各状态值的近似问题；(b) 判断某个状态的值是否至少达到给定阈值的精确问题。文献中的已有结果表明：(a) 近似问题属于 TFNP[NP],；(b) 精确问题属于 PSPACE，可通过实数理论求解，并且是 SQRT-SUM 困难。

我们的贡献如下。(a) 主要结果表明，近似问题属于 TFNP，且相关的间隙问题或承诺问题属于 NP∩coNP。(b) 我们还证明，精确问题属于 $\exists\mathbb{R} \cap$co-$\exists\mathbb{R}$，也就是说，它既可归约到实数的存在理论，也可归约到实数的全称理论，并且它是 PosSLP-困难的。

目录

1 引言 1
2 预备知识 3
3 结果与技术概述：价值近似 5
3.1 已有结果 5
3.2 我们的结果与技术 6

4 结果与技术概述：价值判定 9
4.1 已有结果 9
4.2 我们的结果与技术 9
5 用于CsgApproxValue 的 TFNP 算法 10
5.1 幂级数表示及其性质 11
5.2 见证与验证器 14

6 存在可由幂级数表示的见证 17
6.1 可用幂级数表示的 ε-最优策略 18
6.2 区间界的失败自然候选方案 22
6.3 马尔可夫决策过程中的可用幂级数表示的 ε-紧收缩区间界 24
6.4 定理 5.8 的证明 29
7 极限平均与可达性目标 30
8 价值判定 32
8.1 下界 32
8.2 上界 35

9 扩展 39

10 结论 40

1 引言

本文针对具有极限平均目标的并发随机博弈的价值计算与近似，提出了改进的复杂度结果。我们将非正式地介绍模型、目标、计算问题、已有结果以及我们的贡献。

并发随机博弈。并发随机博弈 [Sha53, Gil57, Eve57] 是定义在有限状态图上的双人零和博弈，进行无限多个阶段。在每个阶段，每位玩家同时且独立于另一位玩家选择一个动作。当前状态以及玩家选择的动作共同决定后继状态的概率分布。无限多个这样的阶段构成一场对局，它是由状态和动作组成的无限序列。并发随机博弈是博弈论中最基本的模型之一。Shapley [Sha53]最初将其作为停时博弈引入；等价地说，这类博弈是以概率 1 终止的折扣和博弈。Gillette [Gil57]和 Everett [Eve57]研究了将其推广到无限时长博弈这一基础性扩展。从数学角度 [Sha53, Gil57, Eve57, MN81, Mar98, FV97],，以及从算法与计算复杂性角度，并发随机博弈都得到了广泛研究；相关工作包括可达性目标的计算复杂性 [DAHK98, EY08, HKM09, FM13], ，极限平均目标的算法与复杂性 [CMH08, HKL+11, OB21, BIJT24], ，对 ω-正则目标的定性与定量分析 [DAH00, DAM01, CDAH06, ACSS24], ，以及时序逻辑 [AHK02]。特别是在反应式系统分析中，对于组件同步交互的反应式系统，并发博弈是恰当的模型 [AHK02, DAHM00, DAHM01]。

目标。 目标为每一场对局指定一个收益。玩家1旨在最大化其期望值，而玩家 2 则采取对抗策略。我们考虑以下经典目标。在 极限平均 目标中（也称为均值收益目标），每次转移都被赋予一个奖励，而一场对局的收益是奖励的长期平均值 [Gil57、MN81]。这些目标推广了可达性目标，其中玩家1的目标是最终到达某个目标状态。对于我们的技术结果，我们使用折扣和 目标，其收益是奖励的折扣和 [Sha53]。

策略、值与计算问题。 策略是规定玩家如何选择动作的规则。它们是一些函数，根据迄今为止访问过的状态序列和选择过的动作，返回一个动作上的概率分布。给定一个并发随机博弈和一个目标，玩家1在状态 s 的值 val(s)，是玩家1能够保证对抗玩家 2 的所有策略时，该目标的最大期望值。给定一个并发随机博弈和一个目标，主要的计算问题如下：(a) 值判定问题：给定状态 s 和阈值 α，询问 val(s) $\geq \alpha$ 是否成立；(b) 值间隙 问题：给定状态 s、阈值 α 和误差 $\varepsilon \in (0, 1)$，询问 val(s) $\geq \alpha + \varepsilon$ 或 val(s) $\leq \alpha - \varepsilon$ 是否成立；若 val(s) 严格介于 $\alpha - \varepsilon$ 和 $\alpha + \varepsilon$ 之间，则任何答案都正确；(c) 值近似问题：给定误差 $\varepsilon \in (0, 1)$，要求计算所有状态的值，误差不超过加性误差 ε。

已有结果。 我们讨论关于精确问题和近似问题的已有结果。
价值判定问题。 并发随机博弈的价值判定问题属于 co-$\exists\mathbb{R}$，即对于可达性目标可归约为实数全称理论 [EY08] ，对于极限平均目标则属于 PSPACE [CMH08， CIJ15], 通过实数理论。此外，对于可达性目标，它是 SQRT-SUM 困难的 [EY08], ；由于极限平均目标是可达性目标的推广，这一困难性也适用于极限平均目标。SQRT-SUM 问题是计算几何中的经典问题 [GGJ76]，而 SQRT-SUM 是否属于 NP 一直是一个悬而未决的问题 [ABKPM09]。一个有趣的开放问题是，对于具有极限平均目标和可达性目标的并发随机博弈，能否改进上界和下界。

近似问题　　题。 并发随机　　t

博弈的价值近似问题的复杂度一直是一个引人入胜且具有挑战性的开放问题，即使对于可达性目标也是如此。研究随机博弈的经典方法是猜测近似最优策略，然后验证这一猜测。该方法的复杂度取决于猜测的规模以及验证过程的复杂度。对于可达性目标， [CMJ04], 使用上述经典方法声称，对于每个固定误差 $\varepsilon \in (0, 1)$，值间隙问题属于 NP $\cap$ coNP。然而， [EY08] 指出， [CMJ04] 中的这一证明是错误的。后来， [HKM09、CHI17]表明，两个玩家的近似最优策略都需要双指数小数，因此其显式表示需要指数空间。这排除了显式猜测策略并进行验证的经典方法。随后采用的方法是对策略进行符号化猜测。研究表明，存在一种近似最优策略的符号化多项式规模表示，使得验证过程需要 NP 预言机 [FM13]。这为可达性目标的价值近似问题带来了 TFNP[NP]上界。对于极限平均目标，近似最优策略可能需要无限记忆 [BF68], ，因此无法直接猜测。相反，极限平均目标的值是折扣和目标的值的极限，即折扣因子趋近于 1 时的极限 [MN81]。此外，当折扣因子与 1 双指数级接近时，折扣和目标的值可以近似这些值 [OB21]。TFNP[NP] 上界后来扩展到了极限平均目标 [BIJT24]。对于遍历博弈这一子类，其中所有状态以概率 1 被无限次访问，极限平均目标的价值近似问题属于 FNP [CIJ14]。对于一般博弈，能否获得 TFNP 上界一直是一个重要的开放问题 [FM13、CHI17、ACSS24、GKW25], ，即使对于可达性目标也是如此。需要注意的是，即使对于并发随机博弈的特殊情形，例如轮流制随机博弈（又称简单随机博弈（SSG）） [Con92], ，可达性目标的价值近似问题也属于 TFNP，而多项式时间算法的存在性仍是一个长期悬而未决的重要问题。我们在表1中总结了此前的复杂度结果。

目标	CsgDecision	CsgApproxValue	CsgGapValue
可达性	共同-$\exists\mathbb{R}$ [EY08]	TFNP[NP] [FM13]	NP[NP] $\cap$ coNP[NP]†
	SQRT-SUM 困难 [EY08]	SSG 困难 [Con92]	SSG 困难 [Con92]
极限平均	PSPACE [CMH08, CIJ15]	TFNP[NP] [双射24]	NP[NP] $\cap$ coNP[NP]†
	SQRT-SUM 困难 [EY08]	SSG 困难 [Con92]	SSG 困难 [Con92]

表1：并发随机博弈已有的复杂度界。在每个单元格中，第一行是上界，第二行是下界。SSG 困难表示其难度至少不低于求解简单随机博弈。$\dagger$这是 CsgApproxValue 上界的直接推论。

我们的贡献。 在本文中，我们解决了这些未决问题。

- 近似问题。 我们的主要结果表明，对于具有极限平均和可达性目标的并发随机博弈，价值近似问题属于 TFNP，见定理 3.1。此外，对于所有这些目标，值间隙问题属于 NP $\cap$ coNP。作为推导过程的一部分，我们还针对折扣因子双指数级接近 1 的折扣和目标得到了相同的结果，见定理 3.2，这一结果本身也可能具有独立的意义。
- 价值判定问题。 我们还证明，对于极限平均和可达性目标，价值判定问题是 PosSLP-困难的，并且属于 $\exists\mathbb{R} \cap$ co-$\exists\mathbb{R}$，即它可归约到实数的存在理论和实数全称理论，见定理 4.1。

我们在表2中总结了我们的结果。一方面，我们的结果改进了针对一个基础博弈论

模型；另一方面，我们讨论研究结果的进一步扩展（例如，对于 ω-正则目标）见第9节。

目标	CsgDecision	CsgApproxValue	CsgGapValue
可达性	$\exists\mathbb{R}\cap$ $\exists\mathbb{R}$co-co-co- **PosSLP-困难**	**TFNP** SSG 困难 [Con92]	**NP** $\cap$ **coNP** SSG 困难 [Con92]
极限平均 $\exists\mathbb{R}\cap$	$\exists\mathbb{R}$co-co-co- **PosSLP-困难**	**TFNP** SSG 困难 [并发92]	**NP** $\cap$ **coNP** SSG 困难 [并发92]

表2：并发随机博弈的复杂度界，包括以粗体显示的我们的结果（定理 3.1 和定理 4.1），采用表1中的约定。

2 预备知识

我们给出了并发随机博弈的标准记号和定义，与 [CHI17、EY08、ACSS24]。

基本记号。给定有限集 X， X 上的概率分布是一个函数 $\mu\colon X \to [0,1]$，满足 $\sum_{x\in X}\mu(x) = 1$。 X 上的所有概率分布构成的集合记为 $\Delta(X)$。对于$\mu \in \Delta(X)$， μ的支撑集，即集合 $\{x \mid \mu(x) > 0\}$，记为 $\mathrm{supp}(\mu)$。对于两个函数f、$g\colon X \to \mathbb{R}$，如果对所有 $x \in X$，都有 $f(x) \le g(x)$，则称 $f \preceq g$。对于函数 $\rho\colon X \to \mathbb{R}$，其 L_1 范数和 L_∞ 范数分别记为 $\|\rho\|_1 \coloneqq \sum_{x\in X}|\rho(x)|$ 和 $\|\rho\|_\infty \coloneqq \max_{x\in X}|\rho(x)|$。对于函数 $f\colon X \to Y$和 $g\colon Y \to Z$，它们的复合记为 $g \circ f$，即对所有 $x \in X$，$(g \circ f)(x) \coloneqq g(f(x))$。正整数集记为 $\mathbb{N}$，非负整数集 $\mathbb{N}_0 \coloneqq \mathbb{N} \cup \{0\}$。对于 $k \in \mathbb{N}$，记 $[k] \coloneqq \{i \in \mathbb{N} \mid i \le k\}$。

并发随机博弈。 并发随机博弈（CSG）是一种双人有限博弈$G = (S, \mathcal{A}, \mathcal{B}\ \delta)$，由以下部分组成：

- 状态集合 S，大小为 n；
- 玩家1和玩家2的动作集$\mathcal{A}$ 和$\mathcal{B}$，每个动作集至多包含 m个动作；以及
- 随机转移函数 $\delta\colon S \times \mathcal{A} \times \mathcal{B} \to \Delta(S)$。

阶段。 给定初始状态 $s_0 \in S$，博弈按如下方式进行。在每个阶段，两个玩家同时且独立地选择一个动作，分别为 $a \in \mathcal{A}$ 和 $b \in \mathcal{B}$。根据两个动作（a， b）和当前状态 s，下一个状态 $s' \in S$按照概率分布 $\delta(s, a, b)$抽取。

历史和对局。CSG中的一段历史是状态和动作构成的有限序列$s_0, a_0, b_0, s_1, a_1, b_1, \ldots, s_k$，其中 $a_t \in$ A且 $b_t \in$ B，并且对任意阶段 $0 \le t < k$，均有$s_{t+1} \in \mathrm{supp}(\delta(s_t, a_t, b_t))$。所有历史的集合记为H。类似地，CSG中的一场对局是一个无限序列 $\omega = s_0, a_0, b_0, s_1, a_1, b_1, \ldots$，并且对任意阶段 $t \ge 0$，均有$s_{t+1} \in \mathrm{supp}(\delta(s_t, a_t, b_t))$。所有对局的集合记为Ω。对于任意状态 s，所有从 s开始的对局构成的集合，即 $\omega = s_0, a_0, b_0, s_1, \ldots$，其中 $s_0 = s$，记为Ω_s。

目标。 目标是一个可测函数 $\phi\colon \Omega \to [0,1]$，为每个对局分配收益。玩家1旨在最大化目标的期望，而玩家2旨在将其最小化。在本文中，我们考虑以下目标，其中 $\omega = \langle s_0, a_0, b_0, s_1, a_1, b_1, \ldots\rangle$ 表示一个对局。

- （可达性） 给定一个目标集 $T \subseteq S$，可达性目标要求最终到达 T，即，$\mathrm{Reach}(T)(\omega) \coloneqq 1$ 如果 $s_t \in T$ 对某个 $t \ge 0$成立，则为 $\mathrm{Reach}(T)(\omega) \coloneqq 0$ 否则。
- *(折扣和)* 给定奖励函数 $r\colon S\times\mathcal{A}\times\mathcal{B} \to [0,1]$ 和折扣因子 $\lambda \in [0,1)$，折扣和目标是奖励的折扣和，即：

$$\mathrm{Disc}(r,\lambda)(\omega) \coloneqq (1-\lambda)\sum_{t\ge 0}\lambda^t\, r(s_t, a_t, b_t)\,.$$

- *(极限平均)* 给定奖励函数 $r\colon S \times \mathcal{A} \times \mathcal{B} \to [0,1]$，极限平均目标是奖励的长期平均值，即：

$$\mathrm{LimAvg}(r)(\omega) \coloneqq \liminf_{k\to\infty} \frac{1}{k} \sum_{t=0}^{k-1} r(s_t, a_t, b_t)\,.$$

请注意，可达性目标是极限平均目标的特殊情况，参见第7节。

策略。玩家1的策略是一个函数 $\sigma\colon \mathcal{H} \to \Delta(\mathcal{A})$，它为每个历史指定一个动作的概率分布。类似地，玩家 2 的策略是一个函数 $\tau\colon \mathcal{H} \to \Delta(\mathcal{B})$。给定策略 σ 和 τ，博弈按如下方式进行。在每个阶段，给定当前历史 h，玩家1（分别为玩家 2）根据分布 $\sigma(h)$（分别为 $\tau(h)$）独立于另一位玩家选择一个动作。玩家1和玩家 2 的所有策略集合分别记为 Σ 和 Γ。平稳策略只通过当前状态依赖于历史，即它是一个为每个状态指定动作概率分布的函数；玩家1的平稳策略记为$\sigma\colon S \to \Delta(\mathcal{A})$（玩家 2 的平稳策略记为 $\tau\colon S \to \Delta(\mathcal{B})$）。玩家1和玩家 2 的所有平稳策略集合分别记为 $\Sigma^{\mathcal{S}}$ 和 $\Gamma^{\mathcal{S}}$。如果平稳策略为每个状态指定一个狄拉克分布，则称其为纯策略。

马尔可夫决策过程和马尔可夫链。对于 $\mathfrak{p} \in \{1, 2\}$，玩家-$\mathfrak{p}$ 马尔可夫决策过程（玩家-$\mathfrak{p}$ 马尔可夫决策过程）是一类特殊的随机博弈，其中另一名玩家只有一个动作，记为 $P = (S, \mathcal{A}, \delta\colon S \times \mathcal{A} \to \Delta(S))$。马尔可夫链（MC）是一类特殊的马尔可夫决策过程，其中两名玩家都只有一个动作，记为 $C = (S, \delta\colon S \to \Delta(S))$。

概率空间。初始状态 $s \in S$ 与一对策略 (σ, τ) 在 Ω_s 上诱导出唯一的概率测度，并配备由有限历史生成的标准σ-代数。我们分别用 $\mathbb{P}^{\sigma,\tau}_s$ 和 $\mathbb{E}^{\sigma,\tau}_s$ 表示概率和期望。下面陈述具有上述目标的随机博弈的基本确定性。

定理2.1（确定性 [Eve57, Sha53, MN81]）。对于所有随机博弈、状态 s，以及可达性、折扣和或极限平均目标 ϕ，我们有

$$\sup_{\sigma\in\Sigma} \inf_{\tau\in\Gamma} \mathbb{E}^{\sigma,\tau}_s[\phi] = \inf_{\tau\in\Gamma} \sup_{\sigma\in\Sigma} \mathbb{E}^{\sigma,\tau}_s[\phi]\,.$$

值，以及 ε-最优策略。上述确定性结果表明，交换策略量词的顺序不会产生差异，因此可以得到唯一的值概念。对于状态 s，其定义如下：

$$\mathrm{val}^G_\phi(s) \coloneqq \sup_{\sigma\in\Sigma} \inf_{\tau\in\Gamma} \mathbb{E}^{\sigma,\tau}_s[\phi]\,.$$

如果上下文明确，我们省略 G。此外，虽然我们将极限平均定义为下极限平均，但在此定义中用上极限替换下极限不会改变值 [MN81]。对于上述目标，我们分别用 val_T、val_λ 和 val 表示 Reach(T)、Disc(r, λ) 和 LimAvg(r) 的值，其中奖励函数 r 可由上下文确定。对于 $\varepsilon \geq 0$，如果玩家1的策略 σ 满足对所有状态 $s \in S$均有以下性质，则称其为ε-最优：

$$\inf_{\tau\in\Gamma} \mathbb{E}^{\sigma,\tau}_s[\phi] \geq \mathrm{val}_\phi(s) - \varepsilon\,.$$

我们称该策略是最优的，当且仅当 $\varepsilon = 0$。对于玩家 ε-最优策略的概念也类似地定义。对于可达性目标，对于每个 $\varepsilon \in (0,1)$，每位玩家都存在一个平稳的 ε-最优策略 [Eve57]。此外，玩家 2 具有最优平稳策略 [HPRV76, Par71]。对于折扣和目标，两位玩家都具有最优平稳策略 [Sha53]。相比之下，对于极限平均目标，通常而言，ε-最优策略需要无限记忆 [BF68]。

通过固定策略得到的马尔可夫决策过程。给定博弈中玩家1的一个平稳策略 σ，以及奖励函数 r，通过固定 σ，我们得到玩家 2 马尔可夫决策过程 $G_\sigma = (S, \mathsf{B}\ \delta_\sigma)$，其奖励函数为 r_σ，其中对于所有状态 s、$s' \in S$ 和动作 $b \in \mathsf{B}$，有 $\delta_\sigma(s,b)(s') := \sum_{a\in\mathsf{A}} \sigma(s)(a) \cdot \delta(s,a,b)(s')$，以及 $r_\sigma(s,b) := \sum_{a\in\mathsf{A}} \sigma(s)(a) \cdot r(s,a,b)$。类似地，通过固定玩家 2 的一个平稳策略 τ，我们得到玩家 1 马尔可夫决策过程 G_τ。

贝尔曼算子 Ψ。对于折扣和目标，马尔可夫决策过程的值由贝尔曼算子刻画。对于 $\mathfrak{p} \in \{1、2\}$，考虑一个玩家-$\mathfrak{p}$ 马尔可夫决策过程 P，其奖励函数为 r，折扣因子为 λ。P的贝尔曼算子 将每个函数 $\rho\colon S \to \mathbb{R}$映射为函数 $\Psi_P(\rho)\colon S \to \mathbb{R}$，定义为

$$\Psi_P(\rho)(s) := \operatorname*{op}_{a\in\mathcal{A}} \left((1-\lambda)\, r(s,a) + \lambda \sum_{s'\in S} \delta(s,a)(s')\rho(s') \right),$$

其中 op 对于 玩家 1 马尔可夫决策过程取 最大值（相应地，最小值）对于玩家 2 马尔可夫决策过程。由于折扣因子小于 1，贝尔曼算子是压缩映射，其唯一不动点是 val^P_λ [Sha53]。

计算问题。计算问题的一个实例由一个 CSG G 组成，其转移概率为有理数，并且使用至多 B_{inp} 位二进制位表示；还包括一个目标ϕ，该目标可以是由一组状态给出的可达性目标，也可以是由奖励函数给出的极限平均目标，其中奖励函数的值为有理数，并且使用至多 B_{inp} 位二进制位表示。对于整数 $B_\varepsilon \geq 1$，加性误差 $\varepsilon = 2^{-B_\varepsilon} \in (0,\ 1)$ 以二进制形式给出，即使用 $B_\varepsilon + 1$ 位二进制位表示。我们考虑以下三个计算问题。

CsgDecision。给定一个实例 (G, ϕ)、一个状态 s，以及一个至多用 B_{inp} 位表示的有理数阈值 α，判断 $\mathsf{val}_\phi(s) \geq \alpha$ 是否成立。

CsgGapValue。给定一个实例 (G, ϕ)、一个状态 s、一个至多用 B_{inp} 位表示的有理数阈值 α，以及二进制表示的 $\varepsilon = 2^{-B_\varepsilon}$，判断 $\mathsf{val}_\phi(s) \geq \alpha + \varepsilon$ 还是 $\mathsf{val}_\phi(s) \leq \alpha - \varepsilon$；若 $\mathsf{val}_\phi(s)$ 严格位于 $\alpha - \varepsilon$ 和 $\alpha + \varepsilon$ 之间，则任意回答均正确。

CsgApproxValue。给定一个实例 (G, ϕ) 和二进制表示的 $\varepsilon = 2^{-B_\varepsilon}$，计算一个函数 $\hat{v}\colon S \to \mathbb{Q}$，使得对于所有状态 $s \in S$，$|\hat{v}(s) - \mathsf{val}_\phi(s)| \leq \varepsilon$。

关于 CsgGapValue 问题的复杂度论述假定满足如下承诺：$\mathsf{val}_\phi(s)$ 不严格位于 $\alpha - \varepsilon$ 和 $\alpha + \varepsilon$ 之间。

3 结果与技术概述：价值近似

本节聚焦于 CsgApproxValue 和 CsgGapValue 问题，它们是本文的主要研究对象。我们回顾相关的已有结果，陈述本文的结果，并介绍用于得出这些结果的技术。

3.1 已有结果

对于可达性目标，由于总是存在 ε-最优平稳策略，一种自然的方法是猜测此类策略并验证其 ε-最优性。因此，计算复杂性在很大程度上取决于 ε-最优平稳策略的复杂性。用于刻画与平稳策略相关联的分布复杂性的数学定义，称为耐心度。

策略的耐心度。对于概率分布 $\mu \in \Delta(X)$，其耐心度是最小非零概率的倒数，即 $\text{Patience}(\mu) := \max\left\{ \frac{1}{\mu(x)} \mid x \in \text{supp}(\mu) \right\}$。对于任一玩家的平稳策略 π，其耐心度是其在所有状态 $s \in S$下的概率分布 π 的耐心度最大值，即 $\text{Patience}(\pi) := \max\{\text{Patience}(\pi(s)) \mid s \in S\}$。对于可达性目标，双方玩家的 ε-最优平稳策略都需要双指数耐心度 [HKM09, CHI17]，且这也是充分的 [HKL+11, 完整版，推论 42]。因此，显式猜测一个ε-最优策略需要指数空间，必须采用不同的平稳策略表示方式。

符号表示。文献 [FM13] 提出了一种称为浮点表示的策略符号表示。他们证明，对于可达性玩家，ε-最优策略在 $\varepsilon \in (0,1)$内具有多项式大小的浮点表示。然而，给定这样一个猜测策略后，验证过程需要 NP 预言机。具体而言，验证猜测策略 σ 是否为 ε-最优策略，可归约为近似求解固定 σ 后得到的马尔可夫决策过程的值；而验证过程需要对该马尔可夫决策过程中的对手纯平稳策略进行量化。因此，他们为可达性目标的 CsgApproxValue 问题得到了 TFNP[NP] 上界。一个尚未解决的问题是如何获得多项式时间验证过程。

推广到极限平均目标。对于极限平均目标，ε-最优策略可能需要无限记忆，因此无法直接猜测。相反，可以用折扣和目标的值来近似其值，其中折扣因子与 1 双指数级接近 [OB21], ，且在这种情况下，双方都存在最优平稳策略。文献[NP]上界 [FM13]中的 TFNP [BIJT24]将文献FM中的TFNPNP上界推广到了极限平均目标。

3.2 我们的结果与技术

Ou主要结果改进了 表1中从 TFNP[NP]到 TFNP 的上界 .

定理 3.1（价值近似）。对于极限平均和可达性目标，*(i)* CsgAp-proxValue 问题属于 *TFNP*；*(ii)* CsgGapValue 问题属于 *NP* ∩ *coNP*。

现在，我们介绍所采用的技术，并将其与以往的方法进行比较。

归约到折扣和目标。我们通过折扣和值近似极限平均值。所需的折扣因子与 1 双指数级接近，但可以选择为 $\lambda = 1 - 2^{-2^{B_\lambda}}$，其中 $B_\lambda \in \mathbb{N}_0$ 是输入规模的多项式 [OB21，定理 3]。我们将这样的折扣因子称为 简洁。因此，定理 3.1由以下结果推出。

定理 3.2（折扣和目标的价值近似）。对于具有简洁折扣因子的折扣和目标，*(i)* CsgApproxValue 问题属于 *TFNP*；*(ii)* CsgGapValue 问题属于 *NP* ∩ *coNP*。

备注 3.3（以二进制形式给出的折扣因子）。对于以二进制形式给出的折扣因子，CsgAp-proxValue 问题属于 UEOPL [BFG+25]，这是 TFNP 的一个子类。该结果不适用于简洁折扣因子，因为它们的二进制表示包含指数级数量的比特，而博弈 x 的贝尔曼算子 Ψ 的近似不动点 $\|x - \Psi(x)\|_\infty \leq \varepsilon$ 只能保证 $\|x - \text{val}_\lambda\|_\infty \leq \frac{\varepsilon}{1-\lambda}$，因此要近似值就需要精度 $\varepsilon(1-\lambda)$，而该精度同样包含指数级数量的比特。

为证明定理 3.2，我们猜测一个见证，它由两位玩家的平稳策略和值的区间界组成，并在多项式时间内对其进行验证。最优策略的概率可能双指数级小 [HKM09、CHI17], ，而且简洁折扣因子与 1 双指数级接近。因此，我们首先需要一种支持对这类数进行精确算术的数值表示。

幂级数表示。FM的浮点表示使用多项式数量的比特表示这类数，但它不支持精确算术，因为一次运算的结果可能需要指数数量的比特。例如， 2^{-1} 和 2^{-2^n} 的浮点表示都只需要多项式数量的比特，而它们的差 $2^{-1} - 2^{-2^n}$ 具有 $2^n - 1$ 有效二进制位。我们提出一种新的符号表示，称为幂级数表示（第 5.1 节），其中一个数是多项带符号的二的幂之和，其指数以二进制形式书写。这种表示支持在多项式时间内进行精确加法、减法、乘法和比较（定理 5.1），并且能够精确表示简洁折扣因子（第 5.1 节）。

见证与验证。我们猜测 ε-最优平稳策略 σ 和 τ，它们现在可用幂级数表示。难点在于验证它们的 ε-最优性。以往的方法会近似固定这些策略后所得马尔可夫决策过程 G_σ 和 G_τ 的值；这需要对对手的策略进行量化，因此需要 NP 预言机 [FM13]。我们的关键想法是将这些值的证书纳入见证中。

- *(见证)*除了策略之外，见证还包含区间界 (L_σ, U_σ) 和 (L_τ, U_τ)，分别对应 G_σ和 G_τ的值；这些区间须满足 $\frac{\varepsilon}{2}$-接近，即区间长度至多为 $\frac{\varepsilon}{2}$，且须收缩，即 $L_\sigma \preceq \Psi_{G_\sigma}(L_\sigma)$ 和 $\Psi_{G_\sigma}(U_\sigma) \preceq U_\sigma$，对 G_τ也同样如此。
- *(可靠性)*由于贝尔曼算子是单调的，且其唯一不动点就是值，收缩区间界会包含该值（定理 5.5）。因此， L_σ 和 U_τ 能够证明ε-最优性，适用于 σ 和 τ (定理 5.7)。
- （多项式时间验证）我们的验证器（算法 1）通过对每个状态和动作应用一次贝尔曼算子，检查区间界是否收缩；该过程可在确定性多项式时间内完成（定理 5.9）。

关键挑战在于见证的完备性，即多项式规模见证的存在性（定理 5.8），这包含两个方面：(a) 多项式规模策略，可通过标准技术获得；(b) 多项式规模的 ε-接近的收缩区间界，这是主要挑战，因为这些值之间的距离可能小至双指数级。

多项式规模策略的存在性。与 [FM13、ACSS24], 类似，我们分两步，从一个最优平稳策略中获得幂级数表示下、多项式规模的 ε-最优平稳策略。首先，我们去除小于某个阈值的概率，该阈值取决于 ε 和折扣因子；这限制了幂级数表示中的指数，并得到一个 ε-最优策略（定理 6.1 和6.2）。其次，将其余每个概率截断为具有多项式数量有效二进制位的数（定理 5.4）。为了界定由此产生的误差，且使其不依赖于折扣因子，我们将折扣和目标解释为可达性目标（定理 6.3），并应用 [Sol03]关于转移概率相对扰动的连续性结果（定理 6.4和6.5）。

多项式规模区间界的存在性。 现在，我们为马尔可夫决策过程 ε构造多项式规模的 G_σ 和 G_τ (定理 6.10) 接近的收缩区间界。这是我们方法面临的主要技术挑战。考虑一个马尔可夫决策过程 P，其值函数为 val_λ。一方面，这些值通常不是幂级数数，例如 $_{1\,3}$，因此区间界必然会偏离这些值。另一方面，区间界又必须近似这些值，因此必须与这些值接近。此外，不同取值之间可能双指数级接近。由于收缩界的不等式会将每个状态的区间界与其后继状态区间界的 λ加权平均进行比较，其中 λ 与 1 双指数级接近，因此这些挑战会导致许多自然的候选方案失败（第 6.2 节）。相反，我们按如下方式构造这些区间界。

- *(凸变换)* 我们通过将非递减凸函数 $f\colon [0,1] \to \mathbb{R}$（满足 $f(x) \leq x$）应用于值函数，得到下界 $L := f \circ \mathrm{val}_\lambda$。对于每个状态 s，如果 P 是玩家1马尔可夫决策过程，则令a 为 s 处的最优动作；否则，令其为任意动作。于是，

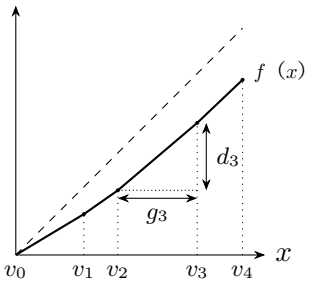


图1：凸分段线性函数 f，其断点位于值 $v_0 < \cdots < v_4$ 处，并位于恒等函数（虚线）下方。 f 在v_{j-1} 和v_j 之间的斜率 d_j/g_j 是增量 d_j 与间隙 g_j的比值，并且随 j增大。值为 v_j 的状态的下界为 $f(v_j)$。

我们有

$$f(\mathrm{val}_\lambda(s)) \le f\Big((1-\lambda)\, r(s,a) + \lambda \sum_{s'} \delta(s,a)(s')\, \mathrm{val}_\lambda(s')\Big) \qquad (f \text{ is non-decreasing})$$
$$\le (1-\lambda)\, r(s,a) + \lambda \sum_{s'} \delta(s,a)(s')\, f(\mathrm{val}_\lambda(s')) \qquad (f \text{ is convex, } f(x) \le x)\,.$$

右端是动作 a 在 $\Psi_P(L)(s)$定义中的项。对于玩家1马尔可夫决策过程， $\Psi_P(L)(s)$ 是所有动作对应这些项的最大值，因此至少等于右端。对于玩家 2 马尔可夫决策过程，该不等式对每个动作都成立，所以对所有动作的最小值也成立，即 $\Psi_P(L)(s)$。因此，对于每个这样的函数 $L \preceq \Psi_P(L)$，不等式 f (定理6.8)成立，接下来只需选择 f 使得 L 具有多项式规模。

- （函数的选择）令 $0 = v_0 < v_1 < \cdots < v_J = 1$ 为 P 各状态的不同取值，再加上 0 和 1。我们选择一个分段线性函数 f，使其满足 $f(0) = 0$，并在 $v_0, \ldots, v_J$处设置断点（图1）。其增量 $d_j := f(v_j) - f(v_{j-1})$ 是间隙 $g_j := v_j - v_{j-1}$乘以一个略小于 1 的因子后，再截断为多项式数量的有效二进制位。截断会使每个斜率 d_j/g_j 略微减小，但各因子从左到右的增幅大于这一减小量，因此斜率递增，且 f 是凸函数。此外，所有斜率都不超过 1，这意味着 $f(x) \le x$；同时，所有斜率至少为 $1 - \frac{\varepsilon}{2}$，这意味着对于所有状态 s，$\mathrm{val}_\lambda(s) - \frac{\varepsilon}{2} \le L(s) \le \mathrm{val}_\lambda(s)$（见定理 6.10 的证明）。

- （多项式规模）每个界都是至多 $n+1$ 个增量之和，并且无论间隙多么小，将其截断至多项式数量的有效二进制位，只需幂级数表示中的少量项（定理 5.3）。此外，根据克拉默法则和哈达玛不等式，间隙至少是双指数级小（定理 6.9），因此幂级数表示中的指数只需多项式数量的二进制位。
- （上界）我们将相同的构造应用于奖励为 $1 - r$ 且由对手玩家行动的马尔可夫决策过程，从而得到上界 U；该过程的值为 $1 - \mathrm{val}_\lambda(s)$。等价地，$U = g \circ \mathrm{val}_\lambda$，其中 g 是一个单调不减的凹函数。

新颖之处。我们方法的主要新颖之处在于，针对具有简洁折扣因子的马尔可夫决策过程，其值存在多项式规模的证书，且可在确定性多项式时间内验证。以往的方法能得到 ε-最优策略的多项式规模表示，但验证其 ε-最优性需要 NP 预言机 [FM13, BIJT24]。尽管涉及贝尔曼算子的不等式可以为马尔可夫决策过程的值提供界的证书 [CQS+25]，但在我们的设定中，主要挑战在于证明，即使折扣因子与 1 双指数级接近，这类证书仍具有多项式规模。我们通过以下两个要素克服了这一挑战。

- *(幂级数表示)* 我们将数表示为多项带符号的二的幂之和，其指数以二进制形式书写。与浮点数 [FM13], 相比，这种表示法支持在多项式时间内，对双指数小数进行精确加法、减法、乘法和比较。
- （多项式规模的 ε-接近收缩区间界）我们的主要技术贡献是

构造多项式规模的下界和上界，使其精确满足涉及贝尔曼算子的不等式。我们通过对值函数应用凸和凹的分段线性变换来得到这些界。其增量是连续值之间的间隙，按递增因子缩放并以固定相对精度截断；这既保持了凸性，也确保了表示具有多项式规模，即使不同取值之间双指数级接近亦是如此。几种自然的此类界候选方案均不成立（第 6.2 节），主要技术结果见定理6.8、定理 6.9和定理 6.10 (第 6.3 节)。

结合多项式规模的 ε-最优策略表示，这些界消除了对 NP 预言机的需求，并给出了 CsgApproxValue 问题的 TFNP 上界。此外，同样的见证也给出了 CsgGapValue 问题的 NP $\cap$ coNP 上界，因为满足 $L_\sigma(s) \geq \alpha$ （分别满足 $U_\tau(s) \leq \alpha$）的见证可以证明 $\mathrm{val}_\lambda(s) \geq \alpha$ （分别证明 $\mathrm{val}_\lambda(s) \leq \alpha$），且当值至少为 $\alpha + \varepsilon$ （分别至多为 $\alpha - \varepsilon$）时，这样的见证必定存在。

4 结果与技术概述：价值判定

本节聚焦于 CsgDecision 问题。我们回顾相关的已有结果，陈述本文的结果，并介绍用于获得这些结果的技术。

4.1 已有结果

实数的存在理论与全称理论。实数存在理论（ETR）是判定如下形式的句子是否在实数域上成立的问题： $\exists x_1 \cdots \exists x_k \, \varphi(x_1, \ldots, x_k)$，其中 φ 是由具有整数系数的多项式方程和不等式构成的无量词布尔组合。类似地，实数全称理论（UTR）是判定如下形式的句子是否在实数域上成立的问题： $\forall x_1 \cdots \forall x_k \, \varphi(x_1, \ldots, x_k)$。复杂性类 $\exists\mathbb{R}$ 和 co-$\exists\mathbb{R}$ 分别由可在多项式时间内归约到 ETR 和 UTR 的问题组成 [Sch10, SŠ17]。我们有 NP $\subseteq \exists\mathbb{R} \subseteq$ PSPACE，且 coNP $\subseteq$ co-$\exists\mathbb{R} \subseteq$ PSPACE [Can88]。

直线程序。一个直线程序 是一系列门 1、.....、k，其中每个门i 要么是常量 $z_i \in \{0、1\}$，要么计算 $z_i := z_j \circ z_{j'}$ 对应门 j、$j' < i$ 以及一个运算 $\circ \in \{+、 -、 \times\}$。该程序计算整数 z_k，其位数可能呈指数级增长。 *PosSLP* 问题要求给定一个直线程序，判断其计算出的整数是否为正数 [ABKPM09]。SQRT-SUM 问题可通过多项式时间图灵归约归约到 PosSLP，而 PosSLP 属于计数层级 [ABKPM09]。

下界。 CsgDecision问题对于可达性目标是 SQRT-SUM 困难 [EY08]，因此对于极限平均目标也同样如此。此前已知递归马尔可夫链 [EY09] 以及马尔可夫链中的成本问题 [HK15], 具有 PosSLP 困难性，但据我们所知，对于具有有限个状态的并发随机博弈，尚无这一结论。

上界。 对于可达性目标，CsgDecision 问题属于 PSPACE，事实上也属于 co-$\exists\mathbb{R}$，因为 [EY08] 将值小于某个阈值这一问题归约为实数存在理论。对于折扣和目标，利用平稳策略以及值函数的一个界作为见证，可以将值大于某个阈值这一问题归约为实数存在理论 [Cha07、CMH08]。对于极限平均目标， [CMH08] 将 CsgDecision 问题归约为具有常数次量词交替的实数理论，因此该问题属于 PSPACE [CIJ15]。

4.2 我们的结果与技术

对于CsgDecision 问题，我们证明了其具有 PosSLP 困难性，这意味着在多项式时间图灵归约下，SQRT-SUM 困难 [ABKPM09],并将 PSPACE 上界改进为

到 $\exists\mathbb{R} \cap$co-$\exists\mathbb{R}$。

定理 4.1（价值判定）。 对于极限平均和可达性目标，以下断言成立。

1.（下界）在多项式时间多对一归约下，CsgDecision 问题是 *PosSLP*-困难的，即使阈值为 $\frac{1}{2}$，且对于其中每次转移都是确定性的、两名玩家各有两个动作，并且每个对局都会到达两个吸收态之一的博弈也是如此。*2.*（上界）CsgDecision 问题属于 $\exists\mathbb{R} \cap$*co*-$\exists\mathbb{R}$。

下界。为证明 PosSLP 困难性，我们利用具有可达性目标的博弈来模拟直线程序（第 8.1 节）。我们考虑具有两个吸收态 $\top$ 和 $\bot$的博弈，每个对局都会到达其中一个吸收态，且目标集为 $T := \{\top\}$。对于每个状态 s，我们使用其中心化值 $2\mathrm{val}_T(s) - 1 \in [-1, 1]$，即当两名玩家均采用最优策略时，到达 $\top$ 的概率减去到达 $\bot$ 的概率。我们构造三个构件：(i) 某个状态的对偶会使其中心化值取反，这类似于 [EY09] 针对递归马尔可夫链的归约中的互补分量；(ii) 平均状态对两个中心化值取平均；(iii) 乘法状态是我们主要的新构造，在该状态中两名玩家都进行随机化，并将两个中心化值相乘，结果相差因子 1 8。门所计算的整数可能非常大，达到双指数级，而中心化值属于 $[-1, 1]$。因此，对于每个门，我们构造一个分子状态和一个分母状态，使得该门计算出的整数等于二者中心化值之比。不同于 [EY09], 的归约——该归约将同一深度上的所有整数按相同因子缩放——分母状态允许每个门使用不同的因子。最后，通过将输出门的分子状态的值与阈值 1 2 进行比较，确定输出的符号。由于所有转移概率都是 1 4 的整数倍，我们可以在不改变值的情况下，用确定性转移替代随机转移 [EY08, CDGH15]。

上界。上界将标准技术与值和有理数阈值的分离相结合。折扣和目标的值与实数阈值之间的严格比较，可由一个平稳策略以及值函数的下界或上界作为见证 [Cha07, CMH08]（定理 8.2）。极限平均目标的值是一个代数数，其次数和系数受到界定 [OB21],，因此，如果它不同于某个有理数阈值 α，那么它与 α 至少相差一个双指数小量 η（定理 8.3 和 8.4）。由于具有简洁折扣因子的折扣和价值与极限平均值的近似误差至多为 $\frac{\eta}{4}$（定理 7.2），我们将其与平移阈值$\alpha - \frac{\eta}{2}$进行严格比较，并通过反复平方定义双指数级小常数 [IJ26, BIJT24]。这将比较 $\mathrm{val}(s) \geq \alpha$ 和 $\mathrm{val}(s) < \alpha$ 都归约为实数存在理论。

新颖性。 我们关于值判定结果的主要新颖之处在于下界。据我们所知，这是并发随机博弈中的首个 PosSLP 困难性结果，并且强化了 [EY08],的 SQRT-SUM 困难性，即使对于可达性目标、确定性转移以及每个玩家只有两个动作的情形也是如此。不同于递归马尔可夫链中的 PosSLP 困难性结果 [EY09],所依赖的递归机制，我们通过一个新的乘法状态来模拟乘法；在该状态中，两个玩家都进行随机化，并结合分子状态和分母状态，分别对每个门的整数进行缩放。

5 用于 CsgApproxValue 的 TFNP 算法

在本节中，我们证明定理 3.2，其中输入由一个博弈、一个值为有理数且定义时最多使用 B_{inp} 位的奖励函数，以及一个简洁折扣因子组成。为此，我们首先介绍数的幂级数表示（第 5.1 节）。然后，我们定义见证，它由可用幂级数表示的策略和 $\frac{\varepsilon}{2}$-接近的收缩区间

界，并为它们提供一个多项式时间验证器（第 5.2 节）。我们将见证完备性的证明，即多项式大小见证存在这一结论（定理 5.8），留待第 6 节。

5.1 幂级数表示及其性质

本小节介绍数的幂级数表示，并阐述其性质。如果某个对象的指定数值以幂级数表示，我们就称该对象可用幂级数表示，具体而言，这些数值包括：(a) 区间界的端点；(b) 某个分布的概率（见下文$\mathcal{D}(\ell, B)$）；以及 (c) 平稳策略分配给各状态的分布中的概率。

幂级数表示。 设$\mathcal{F}(\ell, B)$是一个有理数子集，其中的元素是至多ℓ 个带符号的2的幂之和，指数非正且大于 -2^B，即其绝对值用二进制表示时占 B 位。形式化地，

$$\mathcal{F}(\ell, B) := \left\{ \sum_{i=1}^{k} \pm 2^{-a_i} \quad \middle| \quad k \le \ell, \quad \forall i \le k \quad a_i \in \mathbb{N}_0, \quad a_i < 2^B \right\}.$$

元素$x = \pm 2^{-a_1} \pm \cdots \pm 2^{-a_k} \in \mathcal{F}(\ell, B)$的幂级数表示最多使用 $\ell(B+1)$位，即每一项都使用一位符号位和 B 位指数。注意，指数相对于输入规模可能呈指数级增长，但由于指数以二进制表示，表示长度仍为多项式。我们用$\mathcal{F}_+(\ell, B) := \mathcal{F}(\ell, B) \cap [0, \infty)$表示非负的可用幂级数表示的数。

与浮点数的比较。 浮点数 (m, x) 表示 $m2^x$。两个浮点数的加法/减法不一定能表示为多项式规模的浮点数。例如，两个浮点数 2^{-1} 和 2^{-2^n} 可以用 $O(n)$ 位浮点数表示，但它们的减法 $2^{-1} - 2^{-2^n}$ 需要 $2^n - 1$ 位。相比之下，幂级数表示支持多项式时间的加法、减法、乘法和比较。

算术运算。 幂级数数上的基本运算具有以下性质。

引理5.1。 若 $x \in F(\ell_1, B_1)$ 且 $y \in F(\ell_2, B_2)$，则

- 加法和减法：$x + y, x - y \in F(\ell_1 + \ell_2, \max\{B_1, B_2\})$
- 乘法：$x \cdot y \in F(\ell_1 \cdot \ell_2, \max\{B_1, B_2\} + 1)$，并且所有运算以及比较 $x < y$，均可在关于输入规模的多项式时间内完成。

证明。 考虑 $x = \sum_{i=1}^{k} (-1)^{b_i} \cdot 2^{-a_i}$ 和 $y = \sum_{i=1}^{k} (-1)^{b_i} \cdot 2^{-\tilde{a}_i}$，并记 $B := \max\{B_1, B_2\}$。

- （加法）值

$$x + y = \sum_{i=1}^{k} (-1)^{b_i} \cdot 2^{-a_i} + \sum_{i=1}^{\tilde{k}} (-1)^{\tilde{b}_i} \cdot 2^{-\tilde{a}_i} \in \mathcal{F}(\ell_1 + \ell_2, B),$$

且可在$\mathcal{O}((\ell_1 + \ell_2)(B+1))$时间内计算。

- （减法）值

$$x - y = \sum_{i=1}^{k} (-1)^{b_i} \cdot 2^{-a_i} + \sum_{i=1}^{\tilde{k}} (-1)^{1-\tilde{b}_i} \cdot 2^{-\tilde{a}_i} \in \mathcal{F}(\ell_1 + \ell_2, B),$$

并且可以在$\mathcal{O}((\ell_1 + \ell_2)(B+1))$时间内计算得出。

- *(乘法)* 该值

$$x \cdot y = \sum_{i=1}^{k} \sum_{j=1}^{\tilde{k}} (-1)^{b_i + \tilde{b}_j} \cdot 2^{-(a_i + \tilde{a}_j)} \in \mathcal{F}(\ell_1 \cdot \ell_2, B+1)\,,$$

由于 $a_i + \tilde{a}_j < 2^{B_1} + 2^{B_2} \le 2^{B+1}$，并且可以在 $\mathcal{O}(\ell_1 \ell_2 (B+1))$ 时间内计算。

- （比较）只需确定 $z := x - y$ 的符号。通过减法，我们有 $z = \sum_{i=1}^{\hat{k}} (-1)^{b_i} 2^{-a_i}$，其中 $\hat{k} \le \ell_1 + \ell_2$。首先，我们在不改变其值的情况下重写该和，使所有指数两两不同。当两个项具有相同的指数 a 时，我们按如下方式替换它们：如果它们符号相反，则将二者同时消去，因为 $2^{-a} - 2^{-a} = 0$；否则，它们具有相同的符号（-1）^{b}，我们将它们替换为单个项（-1）$^b 2^{-(a-1)}$，因为 $2^{-a} + 2^{-a} = 2^{-(a-1)}$。每次替换都会减少项的数量，因此最多进行 $\hat{k}$ 次替换。此外，每次替换至多使一个指数减少一，因此所有指数仍为绝对值至多为 $\hat{k} + 2^B$ 的整数。如果没有剩余项，则$x = y$。否则，我们得到 $z = \sum_{i=1}^{k^*} (-1)^{b_i^*} 2^{-a_i^*}$，其指数 $a_1^* < a_2^* < \cdots < a_{k^*}^*$ 两两不同。其次，第一项支配其余各项，因为指数 a_2^*，……，$a_{k^*}^*$ 是大于 a_1^*的不同整数，因此

$$\sum_{i=2}^{k^*} 2^{-a_i^*} < \sum_{j > a_1^*} 2^{-j} = 2^{-a_1^*}\,.$$

因此，z 的符号就是其首项的符号（-1）$^{b_1^*}$。最后，保持各项按指数排序，每次替换都需要$\mathcal{O}\left(\hat{k}(\log \hat{k} + B + 1)\right)$ 时间，因此整个过程需要$\mathcal{O}\left((\ell_1 + \ell_2)^2 (\log(\ell_1 + \ell_2) + B + 1)\right)$ 时间。

所需结果由此得证。 □

有理系数。 输入中的奖励和转移概率都是有理数，通常不是幂级数数，例如$\frac{1}{3}$；而幂级数数在除法下不封闭。尽管如此，以下结果表明，含有有理系数的幂级数数的线性组合，其符号可以在多项式时间内精确判定。

引理 5.2（有理系数）。 设 $x_1, \ldots, x_k \in F(\ell, B)$，且设 $c_1, \ldots, c_k$ 为至多用 B_{inp} 位表示的有理数。那么，$\sum_{i=1}^{k} c_i x_i$ 的符号可以在关于 k、ℓ、B 和 B_{inp} 的多项式时间内判定。

证明。令 Q 为 c_1、……、c_k 的分母之积，并令 $B_Q := kB_{\text{inp}}$。由于 $Q \cdot 2^{-B_Q} > 0$，因此 $\sum_{i=1}^{k} c_i x_i$ 与下式同号

$$Q \cdot 2^{-B_Q} \sum_{i=1}^{k} c_i x_i = \sum_{i=1}^{k} \left(Q c_i 2^{-B_Q}\right) x_i\,.$$

对于每个 $i \in [k]$，数 Qc_i 是 c_i 的分子与其他系数的分母之积，因此它是一个具有 $|Qc_i| < 2^{B_Q}$ 的整数，其计算可在多项式时间内完成。令 $|Qc_i| = \sum_{j=0}^{B_Q - 1} z_j 2^j$ 为 $|Qc_i|$的二进制表示，其中 $z_j \in \{0, 1\}$。于是，我们有

$$Qc_i 2^{-B_Q} = \pm \sum_{j=0}^{B_Q - 1} z_j 2^{-(B_Q - j)} \in \mathcal{F}(B_Q, B_Q)\,,$$

其中，符号与 c_i 的符号相同，且该成员关系成立是因为 $1 \le B_Q - j \le B_Q < 2^{B_Q}$。因此，根据定理 5.1，对于每个 $i \in [k]$，有 $(Qc_i 2^{-B_Q})\ x_i \in \mathsf{F}\,(B_Q \ell, \max\{B, B_Q\} + 1)$，并且右端属于 $\mathsf{F}\,(kB_Q\ell, \max\{B, B_Q\} + 1)$。最后，由于 $B_Q = kB_{\mathrm{inp}}$，根据定理 5.1，右端可在关于 k、ℓ、B 和 B_{inp} 的多项式时间内计算并与 0 比较。 □

近似。 我们将两个数 $x, x' \in (0, 1]$ 之间的相对距离定义为

$$\mathrm{rel}(x, x') \coloneqq \max\left\{\frac{x}{x'}, \frac{x'}{x}\right\} - 1 = \inf\{\gamma \ge 0 \ \mid \ x \le (1+\gamma)x', \quad x' \le (1+\gamma)x\}\,,$$

即，x 和 x' 相差一个乘法因子 $1 + \mathrm{rel}(x, x')$。此外，我们设 $\mathrm{rel}(0,0) \coloneqq 0$ 且 $\mathrm{rel}(x, 0) = \mathrm{rel}(0, x) \coloneqq \infty$，其中 $x \in (0, 1]$。集合 $\mathcal{F}(\ell, B)$ 以如下方式近似区间 $(0, 1]$。

引理 5.3（截断）。设 $2^B \ge \ell \ge 1$，且 $x \in (2^{\ell - 2^B}, 1]$。将 $\mathrm{Trunc}_\ell(x)$ 记作通过保留 x 的二进制展开中从第一个非零位开始的前 ℓ 位所得的数；若 x 存在有限二进制展开，则使用该展开。于是，$\mathrm{Trunc}_\ell(x) \in \mathcal{F}_+(\ell, B)$，$\mathrm{Trunc}_\ell(x) \le x$，且 $\mathrm{rel}(x, \mathrm{Trunc}_\ell(x)) \le 2^{1-\ell}$。特别地，$(1 - 2^{1-\ell})\, x \le \mathrm{Trunc}_\ell(x) \le x$。

证明。 考虑一个二进制序列 $(x_i)_{i \ge 1}$，使得 $x = \sum_{i \ge 1} 2^{-i} x_i$（如果 x 有两个这样的序列，我们选择有限的那个），并将其第一个非零元素的索引记为 $i_0 \coloneqq \min\{i \mid x_i = 1\}$。注意，由于 $2^{\ell - 2^B} < x \le \sum_{i \ge i_0} 2^{-i} = 2^{1 - i_0}$，索引 i_0 有如下上界

$$i_0 < 2^B - \ell + 1\,.$$

将近似 x' 定义为以下截断的二进制展开：从 i_0 开始，最多包含 ℓ 项

$$x' \coloneqq \mathrm{Trunc}_\ell(x) \coloneqq \sum_{i=i_0}^{i_0+\ell-1} 2^{-i} x_i \in \mathcal{F}_+(\ell, B)\,,$$

其中，包含关系成立是因为 $i_0 + \ell - 1 < 2^B$。显然，$x' \le x$。于是，它们之间的距离有如下界

$$\begin{aligned}
\mathrm{rel}(x, x') &= \frac{x}{x'} - 1 && (x' \le x) \\
&= \frac{1}{x'}(x - x') \\
&\le 2^{i_0} \left(\sum_{i \ge i_0 + \ell} 2^{-i} x_i \right) && (x' \ge 2^{-i_0}) \\
&\le 2^{i_0} 2^{-i_0 - \ell + 1} \\
&= 2^{1-\ell}\,.
\end{aligned}$$

最后，相对误差 $\mathrm{rel}(x, x') \le 2^{1-\ell}$ 意味着 $x' \ge x/(1 + 2^{1-\ell}) \ge (1 - 2^{1-\ell})x$，由此完成证明。 □

幂级数概率分布表示。 我们将概率为幂级数数的概率分布集合记为

$$\mathcal{D}(\ell, B) \coloneqq \{\mu \in \Delta([t]) \mid \mu(i) \in \mathcal{F}_+(\ell, B) \text{ for all } i \in [t]\}\ .$$

将相对距离 rel 扩展到概率分布上，定义为

$$\mathrm{rel}(\mu, \mu') \coloneqq \max\{\mathrm{rel}(\mu(i), \mu'(i)) \mid i \in [t]\}\,,$$

我们有如下结果。

引理 5.4（分布近似）。设 $t \geq 1$ 和 $2^B \geq \ell \geq 2$，且 $\mu \in \Delta([t])$ 是一个具有耐心度（μ）$< 2^{2^B-\ell}$ 的分布。那么，存在一个分布 $\mu' \in D(t\ell, B)$，使得 $supp(\mu') = supp(\mu)$，且 $rel(\mu, \mu') \leq t \cdot 2^{1-\ell}$。

证明。设 $i^* \in [t]$使得 $\mu(i^*)$ 取最大值。因此，我们有 $\mu(i^*) \geq 1/t$，从而$1 - \mu(i^*) \leq (t-1)\mu(i^*)$。对于每个 $i \in [t]$，我们定义

$$\mu'(i) \coloneqq \begin{cases} 1 - \sum_{j \neq i^*} \mu'(j) & i = i^*, \\ \mathrm{Trunc}_\ell(\mu(i)) & i \neq i^* \text{ and } \mu(i) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

对于所有满足 $i \neq i^*$ 的 $i \in \mathrm{supp}(\mu)$，由于 $\mu(i) \geq 1/$耐心度$(\mu) > 2^{\ell-2^B}$，由定理 5.3 可得$\mu'(i) \in \mathrm{F}_+(\ell, B)$，$\mu'(i) \leq \mu(i)$，且相对误差$(\mu(i) \| \mu'(i)) \leq 2^{1-\ell}$。根据定义，我们有 $\sum_{i \in [t]} \mu'(i) = 1$，并且由定理 5.1 得 $\mu'(i^*) \in \mathrm{F}(t\ell, B)$。此外，我们有

$$0 \leq \mu'(i^*) - \mu(i^*) = \sum_{i \neq i^*} (\mu(i) - \mu'(i)) \leq 2^{1-\ell}(1 - \mu(i^*)) \leq (t-1)\, 2^{1-\ell} \mu(i^*).$$

因此，我们 有 $\mu'(i^*) > 0$，且 $\mathrm{rel}(\mu(i^*) \| \mu'(i^*)) \leq (t-1)\, 2^{1-\ell}$，由此完成证明 f. □

简洁折扣因子。简洁折扣因子 $\lambda = 1 - 2^{-2^{B_\lambda}}$ 以幂级数表示法精确表示，即 $\lambda \in \mathcal{F}(2, B_\lambda + 1)$ 和 $1 - \lambda \in \mathcal{F}(1, B_\lambda + 1)$。

5.2 见证与验证器

在本小节中，我们首先展示贝尔曼算子的性质（定理 5.5），并定义 ε-紧的收缩区间界。随后，我们定义见证，并证明其可靠性（定理 5.7）和完备性（定理 5.8），相关证明将在第 6 节中给出。接下来，我们介绍一个用于见证的验证器，并证明其可在多项式时间内运行（定理 5.9）。最后，我们证明定理 3.2。

引理 5.5（贝尔曼算子的性质）。对于 $\mathfrak{p} \in \{1,2\}$，设 P 是一个具有奖励函数和折扣因子 λ 的玩家-$\mathfrak{p}$ 马尔可夫决策过程，并设 $\rho, \rho' \colon S \to \mathbb{R}$。则以下断言成立。

1.（单调性）如果 $\rho \preceq \rho'$，那么 $\Psi_P(\rho) \preceq \Psi_P(\rho')$。*2.*（下界）如果 $\rho \preceq \Psi_P(\rho)$，那么 $\rho \preceq \mathrm{val}_\lambda$。*3.*（上界）如果 $\Psi_P(\rho) \preceq \rho$，那么 $\mathrm{val}_\lambda \preceq \rho$。

证明。我们给出这三项的证明。

- （单调性）对于所有函数 $\rho \preceq \rho'$，由于定义 $\Psi_P(\rho)$ 时 ρ 的系数均非负，且 op 是单调的，因此有 $\Psi_P(\rho) \preceq \Psi_P(\rho')$。
- （下界）首先，我们证明贝尔曼算子关于L_∞范数是一个压缩映射。对于所有函数 $\rho, \rho' \colon S \to \mathbb{R}$ 以及状态 $s \in S$，有

$$\begin{aligned} \left|\Psi_P(\rho)(s) - \Psi_P(\rho')(s)\right| &\leq \max_{a \in \mathcal{A}} \left|\lambda \sum_{s' \in S} \delta(s,a)(s')(\rho(s') - \rho'(s'))\right| && (\text{op is 1-Lipschitz}) \\ &\leq \lambda \|\rho - \rho'\|_\infty && (\delta(s,a) \in \Delta(S)). \end{aligned}$$

对所有状态 s 取最大值，可得 $\|\Psi_P(\rho) - \Psi_P(\rho')\|_\infty \leq \lambda \|\rho - \rho'\|_\infty$，其中$\lambda < 1$。因此，根据巴拿赫不动点定理，$\Psi_P$ 具有唯一的不动点，即 val_λ [Sha53]，；对于任意函数 ρ，迭代序列 $\Psi_P^k(\rho)$ 都会在 $k \to \infty$ 时收敛于 val_λ。其次，假设 $\rho \preceq \Psi_P(\rho)$。我们通过归纳法证明，对于所有 $k \geq 1$，都有 $\rho \preceq \Psi_P^k(\rho)$。基础情形 $k = 1$ 由假设成立。若 $\rho \preceq \Psi_P^k(\rho)$，则根据单调性，有 $\Psi_P(\rho) \preceq \Psi_P^{k+1}(\rho)$，因此 $\rho \preceq \Psi_P(\rho) \preceq \Psi_P^{k+1}(\rho)$。最后，由于 $\preceq$在逐点极限下保持不变，取极限 $k \to \infty$ 可得 $\rho \preceq \mathrm{val}_\lambda$。

- *(上界)* 假设 $\Psi_P(\rho) \preceq \rho$。类似于下界，根据归纳法和单调性，对于所有 $k \geq 1$，我们有 $\Psi_P^k(\rho) \preceq \rho$。取极限 $k \to \infty$，可得 $\mathrm{val}_\lambda \preceq \rho$，证明完毕。

□

区间界。区间界是函数 L、U 的一对 $I = (L, U)$: $S \to \mathbb{R}$ 使得 $L \preceq U$。对于区间界 $I = (L, U)$ 和 $I' = (L', U')$，如果对所有状态 $s \in S$ 都有 $L \preceq L'$ 和 $U' \preceq U$，即 $I' \subseteq I$if $[L'(s)\ U'(s)] \subseteq [L(s)\ U(s)]$，我们记为 $I' \subseteq I$。对于 $\mathfrak{p} \in \{1, 2\}$ 和一个玩家-$\mathfrak{p}$ 马尔可夫决策过程 P，如果 $(\Psi_P(L)\ \Psi_P(U)) \subseteq I$，则称 I 是收缩的。对于 $\varepsilon \in (0, 1)$，如果对所有状态 $s \in S$ 都有 $U(s) - L(s) \leq \varepsilon$，则称 I 是 ε-近似的。

备注 5.6（收缩区间界）。 换言之，如果对下界和上界应用一次贝尔曼算子后得到的区间包含于原区间中，则称区间界是收缩的。类似的区间界出现在马尔可夫决策过程的区间值迭代中 [HM18、HK20]以及并发随机博弈中 [GKW25], ，在这些场景中，它们充当可靠的停止准则；它们还出现在马尔可夫决策过程的值的证书中 [CQS+25]。我们将其用作并发随机博弈见证中的证书。这些见证可以在多项式时间内得到验证，并取代以往方法中的 NP 预言机。

见证。 考虑一个博弈 G、奖励函数 r（其值为有理数，且每个数至多用 B_{inp} 位表示）、一个简洁折扣因子 $\lambda = 1 - 2^{-2^{B_\lambda}}$，以及一个加性误差 $\varepsilon = 2^{-B_\varepsilon}$。一个见证 是满足以下条件的元组 $(\sigma,\ \tau,\ I_\sigma,\ I_\tau)$:

1. （见证策略） σ 和 τ 分别是玩家1和玩家 2的可用幂级数表示的平稳策略。2. （$\frac{\varepsilon}{2}$-紧的收缩区间） I_σ（分别为 I_τ）是 G_σ （分别为 G_τ）的可用幂级数表示的 $\frac{\varepsilon}{2}$-紧收缩区间界。3. （区间交）¯$I_\sigma \cap I_\tau \neq \varnothing$，即对于所有状态 $s \in S$，$[L_\sigma(s)\ U_\sigma(s)] \cap [L_\tau(s)\ U_\tau(s)] \neq \varnothing$。¯以下结果说明了见证的可靠性，即每个见证都能给出值的一个 ε-近似以及 ε-最优策略。

引理 *5.7*（见证的可靠性）。考虑一个博弈 G，其奖励函数为 r、简洁折扣因子为 λ，加性误差为 $\varepsilon = 2^{-B_\varepsilon}$，并设 $(\sigma, \tau, I_\sigma, I_\tau)$ 是一个见证。那么，(L_σ, U_τ) 是 val_λ 的 ε-近似，即，对于所有状态 $s \in S$，$val_\lambda(s) \in [L_\sigma(s)\ U_\tau(s)]$，且 $U_\tau(s) - L_\sigma(s) \leq \varepsilon$。此外， σ 和 τ 是 ε-最优的。

证明。区间是收缩的。因此，根据定理 5.5，我们有 $L_\sigma \preceq \mathrm{val}_\lambda^{G_\sigma} \preceq U_\sigma$ 和$L_\tau \preceq \mathrm{val}_\lambda^{G_\tau} \preceq U_\tau$。特别地， L_σ 是 G 的值的下界，因为 σ 保证至少达到 L_σ。类似地， U_τ 是 G 的值的上界，因为 τ 保证至多达到 U_τ。由于 I_σ 和 I_τ 相交，且两个区间的长度都至多为 $\varepsilon/2$，因此我们有 $U_\tau - L_\sigma \preceq \varepsilon \cdot \vec{1}$。因此，$(L_\sigma, U_\tau)$ 是该值的 ε- 近似。最后，对于所有状态 $s \in S$，我们有 $\mathrm{val}_\lambda^{G_\sigma}(s) \geq L_\sigma(s) \geq U_\tau(s) - \varepsilon \geq \mathrm{val}_\lambda(s) - \varepsilon$，即， σ 是 ε-最优的；对称地， τ 是 ε-最优的。 □

见证的大小 是其幂级数表示的比特数。以下结果表明见证的完备性，即每个博弈都有一个多项式规模的见证。

定理 5.8 (多项式规模见证的存在性)。 考虑一个博弈 G，其奖励函数为 r，简洁折扣因子为 λ，加性误差为 $\varepsilon = 2^{-B_\varepsilon}$。那么，存在一个见证 $(\sigma, \tau, I_\sigma, I_\tau)$，其大小关于 n、 m、 B_{inp}、 B_ε 和 B_λ 为多项式。

定理 5.8 是我们的主要技术成果。我们将在第 6 节证明该定理。

验证器。 给定一个博弈 G、奖励函数、折扣因子、加性误差 ε以及一个元组$(\sigma,\ \tau,\ I_\sigma,\ I_\tau)$，验证器 会检查该元组是否满足成为见证的所有条件。算法 1描述了这样一个验证器。定理 5.9 指出，算法 1 当且仅当一个元组是见证时才接受它，并且其运行时间为多项式时间。

算法 1 CsgApproxValue 的多项式验证器

输入： 博弈 G、奖励函数 r、折扣因子 λ、加性误差 ε、元组（$\sigma,\ \tau,\ I_\sigma,\ I_\tau$）
输出： 当且仅当该元组是有效见证时返回 true

过程 VERIFIER（$G,\ r,\ \lambda,\ \varepsilon,\ \ (\sigma,\ \tau,\ I_\sigma,\ I_\tau)$）
    检查 σ 和 τ 是平稳策略（概率为非负且总和为一），并且所有条目均按要求可用幂级数表示
    检查 I_σ 和 I_τ 是区间界 , i.e., $L_\sigma \preceq U_\sigma$ 和 $L_\tau \preceq U_\tau$
    $I'_\sigma \leftarrow (\Psi_{G_\sigma}(L_\sigma)\ \Psi_{G_\sigma}(U_\sigma))$
    $I'_\tau \leftarrow (\Psi_{G_\tau}(L_\tau)\ \Psi_{G_\tau}(U_\tau))$
    如果 $I'_\sigma \not\subseteq I_\sigma$ 或 $I'_\tau \not\subseteq I_\tau$ **then** ▷收缩性质
        返回 假
    如果 $U_\sigma - L_\sigma \not\preceq \frac{\varepsilon}{2}\cdot\vec{1}$ 或 $U_\tau - L_\tau \not\preceq \frac{\varepsilon}{2}\cdot\vec{1}$ **则** ▷ $\frac{\varepsilon}{2}$-接近性性质
        返回 假
    如果 $I_\sigma \cap I_\tau \neq \varnothing$ **则** ▷ 区间交
        返回 真
    else
        返回 假

定理 5.9（多项式时间验证器）。 考虑一个博弈 G，其奖励函数为 r，折扣因子为 λ，加性误差为 $\varepsilon = 2^{-B_\varepsilon}$，且有一个元组 $(\sigma, \tau, I_\sigma, I_\tau)$。则以下断言成立。

1.（正确性）算法 *1* 当且仅当该元组是一个见证时返回 *true*。*2.*（运行时间）算法 *1* 的运行时间关于输入规模为多项式时间。

证明。 我们给出这两项的证明。

- （正确性）算法 1 检查见证定义中的条件。由于使用了幂级数数和有理数进行精确算术运算，其所有计算都是精确的（定理 5.1 和 5.2）。因此，当且仅当该元组是一个见证时，它返回 true。

- （运行时间）我们证明，验证器执行多项式数量的检查，并且每项检查都判断输入中一个线性组合的符号。该线性组合由多项式数量的幂级数数组成，其规模为多项式规模，且系数为有理数。因此，根据定理 5.2，– （幂级数数）元组中的数都是幂级数数，其规模至多为元组的规模。此外， λ 和 $1-\lambda$ 是至多含两项的幂级数数（第 5.1 节），而 $\frac{\varepsilon}{2} = 2^{-(B_\varepsilon+1)}$ 是仅含一项的幂级数数。根据定理 5.1，这些数的常数个乘积仍是多项式规模的幂级数数。–（基本检查）检查元组中的数是否为幂级数数属于语法检查。– 关于策略、区间界、 $\frac{\varepsilon}{2}$ - 接近性以及交集的其余条件，都是对元组中的数之和与$\mathcal{O}(nm)$进行的比较。 $\frac{\varepsilon}{2}$。– （收缩性质）考虑玩家 2 马尔可夫决策过程 G_σ 和一个状态 s。根据以下定义

–

–

Ψ_{G_σ} 和 G_σ，我们有 $\Psi_{G_\sigma}(L_\sigma)(s) \geq L_\sigma(s)$ 当且仅当，对于所有动作 $b \in \mathcal{B}$，

$$L_\sigma(s) \leq \sum_{a \in \mathcal{A}} \sigma(s)(a)\Big((1-\lambda)\, r(s,a,b) + \lambda \sum_{s' \in S} \delta(s,a,b)(s')\, L_\sigma(s')\Big)\,. \tag{1}$$

同样，$\Psi_{G_\sigma}(U_\sigma)(s) \leq U_\sigma(s)$ 当且仅当对于某个动作 $b \in \mathcal{B}$，将 U_σ 替换为 L_σ 并将 $\geq$ 替换为 $\leq$后，相应的不等式成立。对于玩家1马尔可夫决策过程 G_τ，则“对所有”和“对某个”的角色互换。因此，收缩性质至多包含 $4nm$ 个形如 方程（1）的不等式。

– (Li 近似组合）展开乘积后，方程（1）等价于 o

$$\sum_{a \in \mathcal{A}} r(s,a,b)\Big((1-\lambda)\,\sigma(s)(a)\Big) + \sum_{a \in \mathcal{A}} \sum_{s' \in S} \delta(s,a,b)(s')\Big(\lambda\,\sigma(s)(a)\, L_\sigma(s')\Big) - L_\sigma(s) \geq 0\,.$$

左侧是 $(n+1)m+1$ 个至多包含三个幂级数数的乘积的线性组合，其中输入中的奖励和转移概率作为系数。因此，根据第一项，它是输入中具有有理系数的、多项式规模的多项式数量幂级数数的线性组合，从而完成证明。

□

见证的可靠性和完备性（定理 5.7和 定理 5.8）以及多项式时间验证器（定理 5.9）给出了 定理 3.2。形式化地，我们证明 CsgApproxValue 可在多项式时间内归约为寻找见证的搜索问题，该问题属于 TFNP [MP91]。为此，我们在多项式时间内根据见证计算值的一个 ε近似值 $\hat{v}$。

定理 定理 *3.2* 的证明。考虑一个博弈 G、一个奖励函数、一个简洁折扣因子以及一个加性误差 $\varepsilon = 2^{-B_\varepsilon}$。根据 定理 5.8，存在一个多项式 p，使得存在一个针对加性误差 $\frac{\varepsilon}{2}$、大小至多为 $p(n, m, B_{\text{inp}}, B_\varepsilon, B_\lambda)$ 的见证。根据 定理 5.9，此类见证可在多项式时间内验证，因此寻找一个见证属于 TFNP。此外，根据 定理 5.7，每个此类见证都满足：对于所有状态 $s \in S$，

$$L_\sigma(s) \leq \mathrm{val}_\lambda(s) \leq U_\tau(s) \leq L_\sigma(s) + \frac{\varepsilon}{2}\,.$$

价值近似。对于每个状态 s，令 $\hat{v}(s)$ 为 $2^{-(B_\varepsilon+2)}$ 的最大整数倍，且该值属于 $[-1,1]$ 并且不超过 $L_\sigma(s)$；该值通过二分搜索计算（定理 5.1）。于是，$|\hat{v}(s) - \text{值}_\lambda(s)| \leq \frac{\varepsilon}{4} + \frac{\varepsilon}{2} \leq \varepsilon$。因此，CsgApproxValue 可在多项式时间内归约为寻找此类见证。

间隙问题。考虑状态 s 和阈值 α。若 $\mathrm{val}_\lambda(s) \geq \alpha + \varepsilon$，则 $L_\sigma(s) \geq \mathrm{val}_\lambda(s) - \frac{\varepsilon}{2} > \alpha$；若 $\mathrm{val}_\lambda(s) \leq \alpha - \varepsilon$，则 $U_\tau(s) \leq \mathrm{val}_\lambda(s) + \frac{\varepsilon}{2} < \alpha$。因此，猜测这样的见证，并且当且仅当 $L_\sigma(s) \geq \alpha$（分别地，当且仅当 $U_\tau(s) \leq \alpha$）时接受，即可证明该问题属于 NP（分别属于 coNP）；其中，与 α 的比较可根据定理 5.2 在多项式时间内完成。 □

6 存在可由幂级数表示的见证

本节中，我们证明见证的完备性（定理 5.8）。首先，我们证明双方玩家都拥有可用幂级数表示的 ε-最优平稳策略（第 6.1 节）。其次，我们讨论可用幂级数表示的 ε-近似收缩区间界的自然候选方案，但这些方案并不成立（第 6.2 节）。第三，我们证明每个马尔可夫决策过程都有可用幂级数表示的 ε-近似收缩区间界（第 6.3 节）。最后，我们结合这些结果证明定理 5.8（第 6.4 节）。

6.1 可用幂级数表示的 ε-最优策略

本小节中，我们证明双方玩家都拥有可用幂级数表示的 ε-最优平稳策略（定理 6.5）。我们分两步，从最优平稳策略构造出这些策略。

- （移除小概率）我们移除低于某个阈值的概率，从而限制幂级数表示中的指数。我们利用马尔可夫决策过程的绝对连续性性质（定理 6.1 和 6.2）来界定由此产生的误差。

- （截断）我们以固定的相对精度截断其余概率，从而限制项数。我们将折扣和目标解释为可达性目标，并应用马尔可夫链的相对连续性性质（定理 6.3至6.5），由此界定所得误差，且该误差与折扣因子无关。

马尔可夫决策过程的马尔可夫链。对于 $\mathfrak{p} \in \{1, 2\}$，考虑一个玩家-$\mathfrak{p}$ 马尔可夫决策过程 P，其奖励函数为 r，折扣因子为 λ。玩家 $\mathfrak{p}$ 的纯平稳策略 π 为每个状态 s 指定一个动作 $\pi(s) \in \mathsf{A}$。固定 π 后，我们得到马尔可夫链 P_π，其转移函数为 $\delta(s, \pi(s))$，奖励函数为$r(s, \pi(s))$，且折扣因子相同。由于玩家 $\mathfrak{p}$ 在P [Sha53 中存在最优纯平稳策略，对于所有状态 $s \in S$，FV97], ，我们有

$$\mathrm{val}_\lambda^P(s) = \operatorname*{op}_\pi \mathrm{val}_\lambda^{P_\pi}(s)\,, \tag{2}$$

其中，对于玩家 1（分别为玩家 2）的马尔可夫决策过程，op 为 max（分别为 min），且 π 遍历玩家 $\mathfrak{p}$的纯平稳策略。

绝对连续性。马尔可夫决策过程的以下连续性性质推广了 [FV97，公式（4.19）]以及 [ACSS24，引理 36]。

引理 6.1（连续性）。对于 $\mathfrak{p} \in \{1,2\}$，设 P和 P 是玩家 $\mathfrak{p}$ 的马尔可夫决策过程，它们具有相同的状态集和动作集以及折扣因子 λ，其转移函数分别为 δ 和 δ，奖励函数分别为 r 和 $\tilde{r}$。令

$$\eta \coloneqq \max_{s\in S, a\in\mathcal{A}} \max\left\{ |r(s,a) - \tilde{r}(s,a)|\,,\ \sum_{s'\in S} |\delta(s,a)(s') - \tilde{\delta}(s,a)(s')| \right\}.$$

那么，对于所有状态 $s \in S$，均有 $|\text{值}_\lambda^P(s) - \text{值}_\lambda^P(s)| \le \eta/(1-\lambda)$。

证明。注意到 val_λ^P和 val_λ^P取值于 $[0,1]$。对于每个状态 $s \in S$，我们有

$$\begin{aligned}
|\mathrm{val}_\lambda^P(s) - \mathrm{val}_\lambda^{\tilde{P}}(s)| &= |\Psi_P(\mathrm{val}_\lambda^P)(s) - \Psi_{\tilde{P}}(\mathrm{val}_\lambda^{\tilde{P}})(s)| \\
&\le \max_{a\in\mathcal{A}} \Big| (1-\lambda)(r(s,a) - \tilde{r}(s,a)) \\
&\qquad + \lambda \sum_{s'\in S} \left(\delta(s,a)(s')\,\mathrm{val}_\lambda^P(s') - \tilde{\delta}(s,a)(s')\,\mathrm{val}_\lambda^{\tilde{P}}(s')\right)\Big| \\
&\le \max_{a\in\mathcal{A}} \Big((1-\lambda)|r(s,a) - \tilde{r}(s,a)| \\
&\qquad + \lambda \sum_{s'\in S} |\delta(s,a)(s') - \tilde{\delta}(s,a)(s')|\,\mathrm{val}_\lambda^P(s') \\
&\qquad + \lambda \sum_{s'\in S} \tilde{\delta}(s,a)(s')\,|\mathrm{val}_\lambda^P(s') - \mathrm{val}_\lambda^{\tilde{P}}(s')|\Big) \\
&\le (1-\lambda)\eta + \lambda\eta + \lambda\|\mathrm{val}_\lambda^P - \mathrm{val}_\lambda^{\tilde{P}}\|_\infty \\
&= \eta + \lambda\|\mathrm{val}_\lambda^P - \mathrm{val}_\lambda^{\tilde{P}}\|_\infty\,,
\end{aligned}$$

其中，第一个等式由 $\mathsf{val}_\lambda^P = \Psi_P(\mathsf{val}_\lambda^P)$ 和 $\mathsf{val}_\lambda^P = \Psi_{\tilde{P}}(\mathsf{val}_\lambda^P)$ 得出，第一个不等式源于 op 是 1-Lipschitz 的事实，第二个不等式源于三角不等式，第三个不等式则由 η、$\mathsf{val}_\lambda^P \preceq \vec{1}$ 和 $\tilde{\delta}(s,\ a) \in \Delta(S)$ 的定义得出。对所有状态 s 取最大值可得 $\|\mathsf{val}_\lambda^P - \mathsf{val}_\lambda^P\|_\infty \leq \eta + \lambda\|\mathsf{val}_\lambda^P - \mathsf{val}_\lambda^P\|_\infty$，即 $\|\mathsf{val}_\lambda^P - \mathsf{val}_\lambda^P\|_\infty \leq \eta/(1-\lambda)$。

□

引理 6.2 (有界耐心)。考虑一个博弈 G，其奖励函数为 r，折扣因子为 λ。设 $\varepsilon \in (0,1]$ 和 $p_{\min} \coloneqq (1-\lambda)\varepsilon/(2m)$。那么，存在 ε-最优平稳策略σ 和 τ，使得它们所有非零概率都至少为 $p_{\min}$。特别地，它们的耐心度至多为 $1/p_{\min}$。

证明。我们给出玩家1的证明，玩家 2 的证明与之对称。令 σ^* 为一个最优平稳策略 [Sha53]。对于每个状态 s，令 $a_s^* \in \mathcal{A}$ 满足 $\sigma^*(s)(a_s^*)$ 取最大值。因此，我们有 $\sigma^*(s)(a_s^*) \geq 1/m$。对于所有状态 s 和动作 a，定义

$$\sigma(s)(a) \coloneqq \begin{cases} 1 - \sum_{a' \neq a_s^*} \sigma(s)(a') & a = a_s^*\,, \\ \sigma^*(s)(a) & a \neq a_s^* \text{ and } \sigma^*(s)(a) \geq p_{\min}\,, \\ 0 & \text{otherwise.} \end{cases}$$

首先，我们证明 σ 是一个平稳策略，其非零概率至少为 $p_{\min}$。根据定义，对于所有状态 s 和动作 $a \neq a_s^*$，我们有 $\sigma(s)(a) \leq \sigma^*(s)(a)$。因此，对于每个状态 s，我们有

$$\sigma(s)(a_s^*) = 1 - \sum_{a \neq a_s^*} \sigma(s)(a) \geq 1 - \sum_{a \neq a_s^*} \sigma^*(s)(a) = \sigma^*(s)(a_s^*) \geq \frac{1}{m} \geq p_{\min}\,.$$

此外，根据定义，我们有 $\sum_{a\in\mathcal{A}} \sigma(s)(a) = 1$，且 $\sigma(s)$ 的其他非零概率至少为 $p_{\min}$。其次，我们对 L_1 distance between $\sigma(s)$ 和 $\sigma^*(s)$ 之间的距离进行界定。对于每个状态 s，我们有

$$\sum_{a\in\mathcal{A}} |\sigma(s)(a) - \sigma^*(s)(a)| = 2 \sum_{a \neq a_s^*} (\sigma^*(s)(a) - \sigma(s)(a)) \leq 2m\, p_{\min} = (1-\lambda)\varepsilon\,,$$

其中，等式由对所有动作 $a \neq a_s^*$，$\sum_{a\in\mathcal{A}} \sigma(s)(a) = \sum_{a\in\mathcal{A}} \sigma^*(s)(a) = 1$可得 $\sigma(s)(a) \leq \sigma^*(s)(a)$ 推出，而不等式则由求和中的每个非零项都是 $\sigma^*(s)(a) < p_{\min}$ 这一事实推出。最后，我们证明 σ 是 ε- 最优的。对于所有状态 s 和动作 $b \in \mathcal{B}$，我们有

$$|r_\sigma(s,b) - r_{\sigma^*}(s,b)| \leq \sum_{a\in\mathcal{A}} |\sigma(s)(a) - \sigma^*(s)(a)|\, r(s,a,b) \leq (1-\lambda)\varepsilon\,,$$

$$\sum_{s'\in S} |\delta_\sigma(s,b)(s') - \delta_{\sigma^*}(s,b)(s')| \leq \sum_{a\in\mathcal{A}} |\sigma(s)(a) - \sigma^*(s)(a)| \sum_{s'\in S} \delta(s,a,b)(s') \leq (1-\lambda)\varepsilon\,,$$

其中，第一个不等式由 G_σ 的定义和三角不等式推出，第二个不等式由 $r(s,a,b) \leq 1$, $\delta(s,a,b) \in \Delta(S)$ 以及前一步推出。因此，G_σ 和 G_{σ^*} 满足定理 6.1 的条件，其中 $\eta \leq (1-\lambda)\varepsilon$。因此，对于所有状态 s，我们有 $\mathsf{val}_\lambda^{G_\sigma}(s) \geq \mathsf{val}_\lambda^{G_{\sigma^*}}(s) - \varepsilon = \mathsf{val}_\lambda(s) - \varepsilon$，即 σ 是 ε-最优的，至此证明完成。

□

相对连续性。为了以固定的相对精度界定截断误差，我们将折扣和目标解释为马尔可夫链中的可达性目标（定理 6.3）。如果从每个状态出发都以概率 1 到达某个吸收态，则称马尔可夫链为吸收的 。

引理 6.3（折扣和作为可达性）。考虑一个马尔可夫链 $C = (S, \delta)$，其奖励函数为$r\colon\ S \to [0,1]$，折扣因子为 λ。令 C^{abs} 为状态集合为 $S \cup \{\top, \bot\}$ 的马尔可夫链，其中 $\top$ 和 $\bot$ 为吸收态，并且对于所有状态 $s, s' \in S$，

$$\delta^{\mathrm{abs}}(s)(s') \coloneqq \lambda\delta(s)(s')\,, \qquad \delta^{\mathrm{abs}}(s)(\top) \coloneqq (1-\lambda)r(s)\,, \qquad \delta^{\mathrm{abs}}(s)(\bot) \coloneqq (1-\lambda)(1-r(s))\,.$$

那么，C^{abs} 是吸收的，并且对于所有状态 $s \in S$，我们有 $值_\lambda^C(s) = 值_\top^{C^{\mathrm{abs}}}(s)$。

证明。首先，从每个状态 $s \in S$出发，马尔可夫链 C^{abs} 都会以概率 $1-\lambda > 0$转移到 $\{\top, \bot\}$。因此，它以概率 1 到达一个吸收态，即 C^{abs} 是吸收的。其次，根据第一步进行条件化，我们有值 $值_\top^{C^{\mathrm{abs}}}(s) = (1-\lambda)r(s) + \lambda\sum_{s'\in S}\delta(s)(s')值_\top^{C^{\mathrm{abs}}}(s')$，对所有状态 $s \in S$均成立。因此，值$值_\top^{C^{\mathrm{abs}}}$ 在 S 上的限制是马尔可夫链 C的贝尔曼算子的不动点，其唯一不动点为值$值_\lambda^C$。 □

~引理 *6.4*（[*Sol03*，定理 6], 另见 [*ACSS24*，引理 27]）。考虑具有相同有限状态集 S'、转移函数 δ 和 δ以及吸收态 $\top$的吸收马尔可夫链 C 和 $\tilde{C}$。令$\eta \coloneqq \max_{s,s'\in S'}$ 相对误差$(\delta(s)(s')\ \tilde{\delta}(s)(s'))$。那么，对于所有状态 $s \in S'$，我们有 $|值_\top^C(s) - 值_\top^{\tilde{C}}(s)| \le 4|S'|\eta$。

引理 6.5 (多项式规模的 ε-最优策略)。考虑一个博弈 G，其奖励函数为 r，简洁折扣因子为 $\lambda = 1 - 2^{-2^{B_\lambda}}$。令 $\varepsilon = 2^{-B_\varepsilon}$，并令

$$\ell_{\mathrm{str}} \coloneqq B_\varepsilon + \lceil\log(n+2)\rceil + \lceil\log m\rceil + 5 \qquad \textit{and} \qquad B_{\mathrm{str}} \coloneqq B_\lambda + B_\varepsilon + \lceil\log m\rceil + \ell_{\mathrm{str}} + 3\,.$$

那么，存在 ε-最优平稳策略 σ 和 τ，使得对所有状态 $s \in S$，$\sigma(s)\ \tau(s) \in \mathcal{D}(m\ell_{\mathrm{str}}, B_{\mathrm{str}})$。

证明。我们给出玩家1的证明，玩家 2 的证明类似。

- *(有界耐心)* 令 σ 为由 $\frac{\varepsilon}{2}$ 给出的一个ε-最优平稳策略 定理 6.2。那么，我们有

$$\begin{aligned}
\mathrm{Patience}(\sigma) &\le \frac{4m}{(1-\lambda)\varepsilon} && (\text{定理 6.2 对于 } \tfrac{\varepsilon}{2}) \\
&= m \cdot 2^{2^{B_\lambda}+B_\varepsilon+2} && (1-\lambda = 2^{-2^{B_\lambda}},\quad \varepsilon = 2^{-B_\varepsilon}) \\
&\le 2^{2^{B_\lambda}+B_\varepsilon+\lceil\log m\rceil+2} && (m \le 2^{\lceil\log m\rceil}) \\
&= 2^{2^{B_\lambda}+(B_{\mathrm{str}}-B_\lambda)-\ell_{\mathrm{str}}-1} && (\ B_{\mathrm{str}}\text{的定义}) \\
&\le 2^{2^{B_{\mathrm{str}}}-\ell_{\mathrm{str}}-1} && (2^a + b \le 2^{a+b} \text{ 对于 } a, b \in \mathbb{N}_0) \\
&< 2^{2^{B_{\mathrm{str}}}-\ell_{\mathrm{str}}}\,.
\end{aligned}$$

- （截断）由定理 5.4 应用于每个分布 $\sigma(s)$，可得存在一个平稳策略 $\sigma\prime$，使得对于所有状态 $s \in S$，均有 $\sigma\prime(s) \in \mathcal{D}(m\ell_{\mathrm{str}}\ B_{\mathrm{str}})$，并且

$$\begin{aligned}
\mathrm{rel}(\sigma(s), \sigma'(s)) &\le m \cdot 2^{1-\ell_{\mathrm{str}}} && (\text{Theorem 5.4}) \\
&\le 2^{\lceil\log m\rceil+1-\ell_{\mathrm{str}}} && (m \le 2^{\lceil\log m\rceil}) \\
&= 2^{-(B_\varepsilon+\lceil\log(n+2)\rceil+4)} && (\text{definition of } \ell_{\mathrm{str}}) \\
&= \frac{\varepsilon}{16 \cdot 2^{\lceil\log(n+2)\rceil}} && (\varepsilon = 2^{-B_\varepsilon}) \\
&\le \frac{\varepsilon}{16(n+2)} && (n+2 \le 2^{\lceil\log(n+2)\rceil})\,.
\end{aligned}$$

- （截断误差）考虑玩家 2 的一个纯平稳策略 π，以及吸收马尔可夫链

$$C \coloneqq ((G_\sigma)_\pi)^{\mathrm{abs}} \quad \text{and} \quad \tilde{C} \coloneqq ((G_{\sigma'})_\pi)^{\mathrm{abs}},$$

其具有状态 $n+2$ $S \cup \{\top, \bot\}$，并分别将它们的转移函数记为 δ_C 和 $\delta_{\tilde{C}}$。那么，对于所有状态 $s \in S$ 和 $s' \in S \cup \{\top, \bot\}$，有

$$\delta_C(s)(s') = \begin{cases} \lambda \sum_{a \in \mathcal{A}} \sigma(s)(a)\, \delta(s,a,\pi(s))(s') & s' \in S, \\ (1-\lambda) \sum_{a \in \mathcal{A}} \sigma(s)(a)\, r(s,a,\pi(s)) & s' = \top, \\ (1-\lambda) \sum_{a \in \mathcal{A}} \sigma(s)(a)(1 - r(s,a,\pi(s))) & s' = \bot, \end{cases}$$

以及 $\delta_{\tilde{C}}(s)(s')$ 由将 σ 替换为 σ' 得到。因此，对于所有状态 $s, s' \in S$，有

$$\begin{aligned} \delta_{\tilde{C}}(s)(s') &= \lambda \sum_{a \in \mathcal{A}} \sigma'(s)(a)\, \delta(s,a,\pi(s))(s') && \text{(definition of } \delta_{\tilde{C}}) \\ &\leq \lambda \sum_{a \in \mathcal{A}} \Big(1 + \frac{\varepsilon}{16(n+2)}\Big) \sigma(s)(a)\, \delta(s,a,\pi(s))(s') && (\mathrm{rel}(\sigma(s), \sigma'(s)) \leq \tfrac{\varepsilon}{16(n+2)}) \\ &= \Big(1 + \frac{\varepsilon}{16(n+2)}\Big) \delta_C(s)(s') && \text{(definition of } \delta_C). \end{aligned}$$

情形 $s' = \top$ 和 $s' = \bot$ 类似，因为 $r(s,a,\pi(s))$ 和 $1 - r(s,a,\pi(s))$ 均非负。对称地，我们有 $\delta_C(s)(s') \leq (1 + \frac{\varepsilon}{16(n+2)})\delta_{\tilde{C}}(s)(s')$。因此，根据相对误差的定义，对于所有状态 $s \in S$，有

$$\mathrm{rel}(\delta_C(s), \delta_{\tilde{C}}(s)) \leq \frac{\varepsilon}{16(n+2)}. \tag{3}$$

˜此外，$\top$ 和 $\bot$ 的转移在 C 和 C 中完全相同。因此，对于所有状态$s \in S$，有

$$\begin{aligned} |\mathrm{val}_\lambda^{(G_\sigma)_\pi}(s) - \mathrm{val}_\lambda^{(G_{\sigma'})_\pi}(s)| &= |\mathrm{val}_\top^{C}(s) - \mathrm{val}_\top^{\tilde{C}}(s)| && \text{(Theorem 6.3)} \\ &\leq 4(n+2) \cdot \frac{\varepsilon}{16(n+2)} && \text{(Theorem 6.4, Equation (3))} \\ &= \frac{\varepsilon}{4}. && (4) \end{aligned}$$

最后，对于所有状态 $s \in S$，我们有

$$\begin{aligned} |\mathrm{val}_\lambda^{G_\sigma}(s) - \mathrm{val}_\lambda^{G_{\sigma'}}(s)| &= |\min_\pi \mathrm{val}_\lambda^{(G_\sigma)_\pi}(s) - \min_\pi \mathrm{val}_\lambda^{(G_{\sigma'})_\pi}(s)| && \text{(Equation (2))} \\ &\leq \max_\pi |\mathrm{val}_\lambda^{(G_\sigma)_\pi}(s) - \mathrm{val}_\lambda^{(G_{\sigma'})_\pi}(s)| && (\min \text{ is 1-Lipschitz}) \\ &\leq \frac{\varepsilon}{4} && \text{(Equation (4))}, \end{aligned}$$

其中 π 遍历玩家 2 的纯平稳策略。
因此，由于 σ 是 $\frac{\varepsilon}{2}$-最优的，σ' 是 ε-最优的，从而证明完毕。 □

还需证明马尔可夫决策过程 G_σ 和 G_τ 具有可用幂级数表示的 ε-接近收缩区间界。我们首先讨论几种自然的候选方案，说明它们为何在 第 6.2 节中行不通；然后在第 6.3 节中证明，每个马尔可夫决策过程都存在这样的界。

6.2 区间界的失败的自然候选方案

本小节讨论三种自然的候选方案，用于构造可用幂级数表示的 ε-接近收缩区间界，分别是：(1)独立截断、(2)按阶平移，以及(3)贪心构造，并通过示例说明它们为何行不通。

设定。 在本小节中，我们考虑一个玩家 2 马尔可夫决策过程 P，其奖励函数为 r，简洁折扣因子为 $\lambda = 1 - 2^{-2^{B_\lambda}}$，加性误差为 $\varepsilon \in (0, 1/4)$。我们着重研究 ε-接近的收缩区间界的下界，即研究函数 $L: S \to \mathbb{R}$ 满足 $L \preceq \Psi_P(L)$ 且 $\mathrm{val}_\lambda - \varepsilon \cdot \vec{1} \preceq L \preceq \mathrm{val}_\lambda$ 的函数，因为收缩区间界包含该值（定理 5.5 ）。

挑战。 存在可用幂级数表示的下界颇具挑战。一方面，值通常不是幂级数数，例如 $1/3$，因此下界必须偏离这些值。另一方面，下界必须近似这些值，因此必须与这些值接近。此外，不同取值之间可能双指数级接近。由于下界的不等式会将每个状态的界与其后继状态的界的 λ-加权平均值进行比较，其中 λ 与 1 双指数级接近，这些挑战导致许多自然候选方案失败。我们通过以下两个例子说明这一点，见图2。

示例 6.6。 考虑图2（a）中的玩家 2 马尔可夫决策过程，其中有三个状态 s_1、 s_2和 s_3，每个状态都只有一个动作。状态 s_1 以奖励 1 转移到 s_2 ，状态 s_2 以奖励 0 转移到 s_3 。状态 s_3 是吸收的，其自环的奖励为$1/3$。各状态的值为

$$\begin{aligned}
\mathrm{val}_\lambda(s_1) &= (1-\lambda) + \lambda \mathrm{val}_\lambda(s_2) = (1-\lambda) + \frac{\lambda^2}{3}\,, \\
\mathrm{val}_\lambda(s_2) &= \lambda \mathrm{val}_\lambda(s_3) = \frac{\lambda}{3}\,, \\
\mathrm{val}_\lambda(s_3) &= \frac{1}{3}\,.
\end{aligned}$$

示例 6.7。 考虑图2 （b）中的玩家 2 马尔可夫决策过程，其中有三个状态 s_1、 s_2和 s_3，每个状态都只有一个动作。状态 s_1分别以$3/5$和$2/5$的概率转移到 s_2和 s_3，奖励为$2/3$。状态 s_2和 s_3是吸收的，其自环的奖励分别为 1 和$1/6$。各状态的值为

$$\begin{aligned}
\mathrm{val}_\lambda(s_1) &= (1-\lambda)\frac{2}{3} + \lambda\Big(\frac{3}{5}\mathrm{val}_\lambda(s_2) + \frac{2}{5}\mathrm{val}_\lambda(s_3)\Big) = \frac{2}{3}\,, \\
\mathrm{val}_\lambda(s_2) &= 1\,, \\
\mathrm{val}_\lambda(s_3) &= \frac{1}{6}\,.
\end{aligned}$$

候选方案1：独立截断。一种自然的候选方案是将每个值独立截断为项数较少的幂级数数，并将所有值平移一个共同的量，以获得下界。形式化地，对于 $\ell \in \mathbb{N}$ 和 $c \in \mathbb{R}$，候选下界为 $L := \mathrm{Trunc}_\ell \circ \mathrm{val}_\lambda - c \cdot \vec{1}$，其中 Trunc_ℓ 是定理 5.3 中的截断。该候选方案在定理 6.6 中失败了，因为它将 s_2 和 s_3 的双指数级接近的值截断为同一个数，而收缩界的不等式要求它们的界之间存在正差值。事实上，任何满足 $L \preceq \Psi_P(L)$ 的函数 L 都满足 $L(s_2) \le \Psi_P(L)(s_2) = \lambda L(s_3)$，即，

$$L(s_3) - L(s_2) \ge (1-\lambda)\, L(s_3)\,.$$



(a) (b)

图2： 定理 6.6 和 6.7的马尔可夫决策过程。每个动作均标有其奖励。在 (b) 中，动作 s_1 从黑点移动到 s_2 和 s_3，其概率如给定值所示。

然而，由于 $1-\lambda = 2^{-2^{B_\lambda}}$，$\mathrm{val}_\lambda(s_2)$ 和 $\mathrm{val}_\lambda(s_3)$ 的二进制展开在前 $2^{B_\lambda}-2$ 位上相同。此外，在 s_3 处的接近性，即 $L(s_3) \geq \mathrm{val}_\lambda(s_3) - \varepsilon$，意味着 $c \leq \varepsilon < \frac{1}{4}$。因此，对于所有 $\ell \leq 2^{B_\lambda}-3$，该候选方案满足 $L(s_2) = L(s_3) > 0$，从而违反了上述不等式。这表明，不同状态的界应当联合选取。

候选方案 2：按阶平移。为了联合选择不同状态的界，一个自然的候选方案是：将值向下平移不同的量，且值越小，平移量越大；这样可以保持值的顺序，并在平移所产生的松弛量 $\Psi_P(L) - L$内对平移后的值进行截断。形式化地，对于 $\eta > 0$和 c： $S \to \mathbb{R}_{>0}$使得对于所有状态 s、s'，若 $\mathrm{val}_\lambda(s') < \mathrm{val}_\lambda(s)$，则有 $c(s') \geq (1+\eta)\, c(s)$，候选下界通过截断 $\mathrm{val}_\lambda - c$得到。该候选方案在定理 6.6 中失败，因为 s_1 处的松弛量相对于平移量是双指数级小的，因此该候选方案要么违反 $L \preceq \Psi_P(L)$，要么需要以绝对精度进行截断。事实上，考虑 L： 截断前的$= \mathrm{val}_\lambda - c$。由于 $\mathrm{val}_\lambda(s_2) < \mathrm{val}_\lambda(s_1)$，我们有 $c(s_2) \geq (1+\eta)\, c(s_1)$，因此

$$\begin{aligned}
\Psi_P(L)(s_1) - L(s_1) &= (1-\lambda) + \lambda L(s_2) - L(s_1) && \text{(definition of } \Psi_P) \\
&= c(s_1) - \lambda\, c(s_2) && (L = \mathrm{val}_\lambda - c,\ \mathrm{val}_\lambda = \Psi_P(\mathrm{val}_\lambda)) \\
&\leq \big(1 - \lambda(1+\eta)\big)\, c(s_1) && (c(s_2) \geq (1+\eta)\, c(s_1))\,.
\end{aligned}$$

因此，若 $\eta > \frac{1-\lambda}{\lambda}$，例如，若 $\eta > 2^{1-2^{B_\lambda}}$，则 L在截断之前就违反了 $L \preceq \Psi_P(L)$。否则， s_1 处的松弛量至多为 $(1-\lambda)\, c(s_1) = 2^{-2^{B_\lambda}} c(s_1)$，因此截断需要 $2^{-2^{B_\lambda}} c(s_1)$的绝对精度，而项数较少的幂级数数只能保证相对精度（定理 5.3）。这表明，界应由连续值之间的间隙构造，并以相对精度进行截断。

候选方案 3：贪心构造。根据这些观察，一个自然的候选方案是按不同取值从大到小的顺序处理，并通过从前一个值的界中减去两个值之间的间隔（按固定相对精度截断），来得到每个值的界。形式化地，令 $v_1 > v_2 > \cdots > v_J$为 P的不同取值；对于 $j \geq 2$，令 $g_j := v_{j-1} - v_j$ 为连续值之间的间隔，并令 $d_j := \mathrm{Trunc}_\ell(g_j)$为其截断，其中 Trunc_ℓ 是定理 5.3 中的截断。那么，对于 $c \in \mathbb{R}$，候选下界为$L(s) := v_1 - c - \sum_{i=2}^{j} d_i$，对所有满足 $\mathrm{val}_\lambda(s) = v_j$的状态 s 均成立。这些界具有多项式规模，因为每个间隔都按相对精度截断，且间隔至少是双指数级小（定理 6.9）。定理 6.7 表明该候选方案不成立，因为各间隔的截断相对误差不同，因此下界并非值的凸函数。事实上，这些值为 $v_1 = 1$、 $v_2 = \frac{2}{3}$ 和 $v_3 = \frac{1}{6}$。第一个间隔 $g_2 = \frac{1}{3}$ 不是幂级数数，因此对于某个 $\rho \geq \frac{1}{3} 2^{-(\ell+1)}$，$d_2 = \frac{1}{3} - \rho$；而第二个间隔 $g_3 = \frac{1}{2}$被精确截断，即 $d_3 = \frac{1}{2}$。于是，我们有 $L(s_2) = 1 - c$、 $L(s_1) = \frac{2}{3} - c + \rho$，以及 $L(s_3) = \frac{1}{6} - c + \rho$，因此

$$L(s_1) - \Psi_P(L)(s_1) = \frac{2}{3} - c + \rho - (1-\lambda)\frac{2}{3} - \lambda\Big(\frac{3}{5}(1-c) + \frac{2}{5}\Big(\frac{1}{6} - c + \rho\Big)\Big) = \Big(1 - \frac{2}{5}\lambda\Big)\rho - (1-\lambda)\, c\,.$$

此外，在 s_2处的接近性，即 $1-\varepsilon \le L(s_2) \le 1$，意味着 $0 \le c \le \varepsilon < 1$。因此，由于 $1-\frac{2}{5}\lambda \ge \frac{3}{5}$，对于所有 $\ell \le 2^{B_\lambda} - 5$，该候选方案都违反了 $L \preceq \Psi_P(L)$。请注意，将界视为值的函数时，它们在 $1\,6$ 与 $2\,3$ 之间的斜率为 1，而在 $2\,3$ 与 1 之间的斜率较小，为 $1-3\rho$。

我们的构造。我们在第 6.3 节中的构造同样基于连续值之间经截断的间隔来构造下界，但在截断之前，会用不断增大的因子乘以这些间隔。因此，斜率不断增大，即下界是关于这些值的凸函数。于是，由詹森不等式可得 $L \preceq \Psi_P(L)$ (定理6.8)。

6.3 马尔可夫决策过程中的可用幂级数表示的 ε-近收缩区间界

在本小节中，我们证明，任何具有折扣和目标的马尔可夫决策过程都具有可用幂级数表示的 ε-近收缩区间界（定理 6.10）。为此，我们将一个凸分段线性函数作用于值函数以得到下界，并对称地得到上界。

- *(凸变换)* 我们证明，对于每个非递减凸函数 f 且满足$f(x) \le x$，函数 $L := f \circ \mathrm{val}_\lambda$satisfies $L \preceq \Psi_P(L)$ (定理 6.8)。
- （值的分离）我们证明，马尔可夫决策过程的两个不同取值之差至少为一个双指数级小量（定理 6.9）。
- （幂级数可表示的区间界）我们选择 f 以各个值为断点的分段线性函数。其连续值之间的增量，即这些值之间的间隔，乘以递增因子后，再按固定的相对精度进行截断，这使得 f为凸函数（定理 6.10）。无论间隔多么小，截断所需的项数都很少，而定理 6.9 给出了指数的界，因此下界是可用幂级数表示的。

与第6.2节中的候选方案不同，这些界无论值彼此多么接近，都严格满足收缩界的不等式。

引理6.8 （凸变换）。对于 $\mathfrak{p} \in \{1,2\}$，设 P 为一个具有奖励函数和折扣因子 λ 的玩家-$\mathfrak{p}$ 马尔可夫决策过程，并设 $f\colon [0,1] \to \mathbb{R}$ 为满足以下条件的单调不减凸函数：对所有 $x \in [0,1]$，有 $f(x) \le x$。则有 $f \circ \mathrm{val}_\lambda \preceq \Psi_P(f \circ \mathrm{val}_\lambda)$。

证明。 注意，val_λ取值于 $[0,1],$，因为奖励属于 $[0,1]$。考虑状态 $s \in S$。若$\mathfrak{p} = 1$，令 $a \in \mathcal{A}$ 为在 $\Psi_P(\mathrm{val}_\lambda)(s)$的定义中取得最大值的动作；若$\mathfrak{p} = 2$，则令 $a \in \mathcal{A}$ 为任意动作。由于 $\mathrm{val}_\lambda = \Psi_P(\mathrm{val}_\lambda)$，在这两种情况下，我们都有

$$\mathrm{val}_\lambda(s) \le (1-\lambda)\, r(s,a) + \lambda \sum_{s' \in S} \delta(s,a)(s')\, \mathrm{val}_\lambda(s')\,.$$

因此，我们有

$$\begin{aligned}
f(\mathrm{val}_\lambda(s)) &\le f\Big((1-\lambda)\, r(s,a) + \lambda \sum_{s' \in S} \delta(s,a)(s')\, \mathrm{val}_\lambda(s')\Big) && (f \text{ is non-decreasing}) \\
&\le (1-\lambda)\, f(r(s,a)) + \lambda \sum_{s' \in S} \delta(s,a)(s')\, f(\mathrm{val}_\lambda(s')) && (f \text{ is convex},\ \delta(s,a) \in \Delta(S)) \\
&\le (1-\lambda)\, r(s,a) + \lambda \sum_{s' \in S} \delta(s,a)(s')\, f(\mathrm{val}_\lambda(s')) && (f(x) \le x)\,. \qquad (5)
\end{aligned}$$

接下来只需证明　比较公式（5）的右端与 $\Psi_P(f \circ \mathrm{val}_\lambda)(s)$。如果 $\mathfrak{p} = 1$，那么我们

have

$$
\begin{aligned}
f(\mathrm{val}_\lambda(s)) &\le (1-\lambda)\, r(s,a) + \lambda \sum_{s'\in S} \delta(s,a)(s')\, f(\mathrm{val}_\lambda(s')) && \text{(Equation (5))}\\
&\le \max_{a'\in\mathcal{A}} \Big((1-\lambda)\, r(s,a') + \lambda \sum_{s'\in S} \delta(s,a')(s')\, f(\mathrm{val}_\lambda(s')) \Big) && (a\in\mathcal{A})\\
&= \Psi_P(f\circ \mathrm{val}_\lambda)(s) && \text{(definition of } \Psi_P,\ \mathsf{op} = \max)。
\end{aligned}
$$

如果 $\mathfrak{p}=2$，那么公式（5）对每个动作 $a\in\mathcal{A}$都成立，因为 a 是任意的，因此我们有

$$
\begin{aligned}
f(\mathrm{val}_\lambda(s)) &\le \min_{a\in\mathcal{A}} \Big((1-\lambda)\, r(s,a) + \lambda \sum_{s'\in S} \delta(s,a)(s')\, f(\mathrm{val}_\lambda(s')) \Big) && \text{(Equation (5))}\\
&= \Psi_P(f\circ \mathrm{val}_\lambda)(s) && \text{(definition of } \Psi_P,\ \mathrm{op}=\min)\,,
\end{aligned}
$$

这就完成了证明。 □

为了界定区间界表示的大小，我们利用马尔可夫决策过程的不同取值不可能任意接近这一事实。这类界是标准结果，例如参见 [Con92] 和 [ABdMS21，定理 23]。为完整起见，我们给出一个关于简洁折扣因子的显式界的证明。

引理 6.9（值的分离）。对于 $\mathfrak{p}\in\{1,2\}$，令 P 为一个玩家-$\mathfrak{p}$ 马尔可夫决策过程，其奖励函数为r，且具有简洁折扣因子 $\lambda = 1-2^{-2^{B_\lambda}}$。令 $B_P\in\mathbb{N}_0$ 满足：P 的所有奖励和转移概率的公分母至多为 $2^{2^{B_P}}$，并令

$$
B_{\mathrm{sep}} := B_\lambda + B_P + 2\lceil \log n\rceil + 1\,.
$$

那么，$\{\mathrm{val}_\lambda(s) \mid s\in S\}\cup\{0,1\}$中的任意两个不同元素之差至少为 $2^{-2^{B_{\mathrm{sep}}}}$。

证明。令 π 为 P [Sha53 中的 FV97]，的最优纯平稳策略，即 $\mathrm{val}_\lambda^P = \mathrm{val}_\lambda^{P_\pi}$。令$Q\le 2^{2^{B_P}}$ 为 P 中所有奖励和转移概率的公分母。首先，将 val_λ 写成一个具有整数系数的线性方程组的解。对于每个状态 $s\in S$，有

$$
\begin{aligned}
2^{2^{B_\lambda}} Q\, \mathrm{val}_\lambda(s) &= 2^{2^{B_\lambda}} Q \Big((1-\lambda)\, r(s,\pi(s)) + \lambda \sum_{s'\in S} \delta(s,\pi(s))(s')\, \mathrm{val}_\lambda(s') \Big) && (\mathrm{val}_\lambda = \Psi_{P_\pi}(\mathrm{val}_\lambda))\\
&= Q\, r(s,\pi(s)) + (2^{2^{B_\lambda}}-1) \sum_{s'\in S} Q\, \delta(s,\pi(s))(s')\, \mathrm{val}_\lambda(s') && (\lambda = 1-2^{-2^{B_\lambda}})\,.
\end{aligned}
$$

因此，向量 $x=\mathrm{val}_\lambda$ 是线性方程组 $Ax=c$ 的解，其中每个状态 $s\in S$ 对应以下方程

$$
2^{2^{B_\lambda}} Q\, x_s - (2^{2^{B_\lambda}}-1) \sum_{s'\in S} Q\, \delta(s,\pi(s))(s')\, x_{s'} = Q\, r(s,\pi(s))\,. \tag{6}
$$

根据 Q 的选择，矩阵 A 和向量 c 的元素均为整数。此外，将方程 (6) 除以 $2^{2^{B_\lambda}}Q$，即可得到 P_π 的贝尔曼算子的不动点方程，其唯一解为 val_λ。因此，val_λ 是 $Ax=c$ 的唯一解，故 A 可逆。此外，由于 $Q\,r(s,\pi(s))$ 和 $Q\,\delta(s,\pi(s))(s')$ 分别属于 $[0,Q]$ 和 $Q\le 2^{2^{B_P}}$，A 和 c 的所有元素的绝对值均不超过 $2^{2^{B_\lambda}+2^{B_P}}$。其次，我们对该解的公分母进行界定。根据克拉默法则，对于所有状态 $s\in S$，有 $x_s = \det(A_s)/\det(A)$，其中 A_s 是通过将 A 的列替换为 A 得到的

列。两个行列式都是整数。由于 A 的每一行有 n 个元素，因此有

$$\begin{aligned}
|\det(A)| &\le \left(\sqrt{n}\cdot 2^{2^{B_\lambda}+2^{B_P}}\right)^n && \text{(Hadamard's inequality)}\\
&\le 2^{n(2^{B_\lambda}+2^{B_P}+\lceil\log n\rceil)} && (\sqrt{n}\le 2^{\lceil\log n\rceil})\\
&\le 2^{2^{\lceil\log n\rceil}(2^{B_\lambda}+B_P+1}+\lceil\log n\rceil) && (n\le 2^{\lceil\log n\rceil},\ 2^a+2^b\le 2^{a+b+1})\\
&\le 2^{2^{\lceil\log n\rceil}\cdot 2^{B_\lambda}+B_P+1+\lceil\log n\rceil}} && (2^a+b\le 2^{a+b})\\
&= 2^{2^{B_{\text{sep}}}} && (\text{definition of } B_{\text{sep}})\,.
\end{aligned}$$

最后，定义 $q := |\det(A)| \le 2^{2^{B_{\text{sep}}}}$。那么，对于所有状态 $s\in S$， $q\cdot \mathsf{val}_\lambda(s) = \pm\det(A_s)$ 都是整数。对于 $\{\mathsf{val}_\lambda(s) \mid s\in S\}\cup\{0,1\}$ 中两个不同的元素 v,v'，数值 $q\cdot v$ 和 $q\cdot v'$ 是不同的整数，因此 $|v-v'|\ge 1/q \ge 2^{-2^{B_{\text{sep}}}}$，证明完成。 □

现在可以陈述本小节的主要结果了。

引理 6.10 (马尔可夫决策过程中可用幂级数表示的区间界)。 对于 $\mathfrak{p}\in\{1,2\}$，设 P 是一个玩家-$\mathfrak{p}$ 马尔可夫决策过程，具有奖励函数和简洁折扣因子 $\lambda = 1-2^{-2^{B_\lambda}}$，并设 $B_P\in\mathbb{N}_0$和 B_{sep} 如 定理 6.9中所定义。设 $\varepsilon = 2^{-B_\varepsilon}\in(0,1)$，并设

$$\ell_{\text{bnd}} := (n+1)(B_\varepsilon+\lceil\log(n+1)\rceil+8)+1 \qquad \textit{and} \qquad B_{\text{bnd}} := B_{\text{sep}}+B_\varepsilon+\lceil\log(n+1)\rceil+10\,.$$

那么，存在一个 ε-接近的收缩区间界 $I=(L,U)$，使得对于所有状态 $s\in S$， $L(s)\ U(s)\in F(\ell_{\text{bnd}},B_{\text{bnd}})$。

证明。 我们通过将一个凸函数 f 应用于值函数来构造下界 L，并将相同的构造应用于对偶马尔可夫决策过程来构造上界 U。函数 f 是分段线性的，其断点是 P的值，也就是说， f 在每两个连续值之间都是线性的。因此， f 由 $f(0)=0$ 及其增量决定，也就是 f在 P的连续值处的值之差。证明分为以下四个步骤。

- *(增量)* 我们将相应连续值之间的间隙乘以一个因子，再截断为幂级数数，以此选取每个增量。要使 f 为凸函数，其斜率必须从左向右递增。截断为幂级数数会使每个斜率略微减小。因此，在截断之前，我们将每个间隙乘以一个略小于一的因子，并使这些因子从左向右递增，且增幅大于截断造成的斜率减小幅度。因此， f 的斜率接近一、不超过一且递增，由此可知 f 是凸函数。

- （下界）我们证明 $L := f\circ\mathsf{val}_\lambda$ 满足 $L\preceq\Psi_P(L)$，以至多 $\frac{\varepsilon}{2}$ 的误差逼近 val_λ，并且可用幂级数表示。
- （上界）我们将该构造应用于奖励为 $1-r$ 的马尔可夫决策过程以及对手玩家，其值函数为 $\vec{1}-\mathsf{val}_\lambda$，由此得到上界 U。
- （区间界）我们证明 (L,U) 是一个 ε-接近的收缩区间界。

现在我们证明每个步骤。

- （增量）设 $0=v_0<v_1<\cdots<v_J=1$ 为 $\{\mathsf{val}_\lambda(s)\mid s\in S\}\cup\{0,1\}$ 的互异元素，并且对于 $j\in[J]$，令 $g_j := v_j - v_{j-1}$ 为连续值之间的间隙。注意 $J\le n+1$，并且由 定理 6.9， $g_j\ge 2^{-2^{B_{\text{sep}}}}$ 对所有 $j\in[J]$ 成立。令 $\ell_{\text{inc}} := B_\varepsilon+\lceil\log(n+1)\rceil+8$

为幂级数表示中每个增量的项数。于是，我们有

$$\begin{aligned} 2^{1-\ell_{\mathrm{inc}}} &= \varepsilon \cdot 2^{-\lceil \log(n+1) \rceil - 7} && (\text{definition of } \ell_{\mathrm{inc}},\ \varepsilon = 2^{-B_\varepsilon}) \\ &\le \frac{\varepsilon}{128(n+1)} && (n+1 \le 2^{\lceil \log(n+1) \rceil}) \\ &\le \frac{\varepsilon}{128J} && (J \le n+1)\,. \end{aligned} \tag{7}$$

对于 $j \in [J]$，定义因子 $\gamma_j := 1 - \frac{\varepsilon}{4} + 2^{3-\ell_{\mathrm{inc}}} j$，用于在截断前乘以间隙 g_j。于是，我们有 $\gamma_j \ge 1 - \frac{\varepsilon}{4} > \frac{1}{2}$，并且

$$\begin{aligned} \gamma_j &\le \gamma_J && (j \le J) \\ &= 1 - \frac{\varepsilon}{4} + 4 \cdot 2^{1-\ell_{\mathrm{inc}}} J && (\gamma_J \text{ 的定义}) \\ &\le 1 - \frac{\varepsilon}{4} + \frac{\varepsilon}{32} && (\text{方程（7）}) \\ &< 1\,. \end{aligned}$$

由于 $\frac{1}{2} < \gamma_j < 1$ 和 $2^{-2^{B_{\mathrm{sep}}}} \le g_j \le 1$，我们有 $\gamma_j g_j \in (2^{-2^{B_{\mathrm{sep}}}-1}, 1]$。此外，由于对于所有 a，$B_{\mathrm{bnd}} = B_{\mathrm{sep}} + \ell_{\mathrm{inc}} + 2$ 和 $2^{a+b} \ge 2^a + b$，$b \in \mathbb{N}_0$，我们有 $2^{B_{\mathrm{bnd}}} \ge 2^{B_{\mathrm{sep}}} + \ell_{\mathrm{inc}} + 2$，因此 $\gamma_j g_j \in (2^{\ell_{\mathrm{inc}} - 2^{B_{\mathrm{bnd}}}}, 1]$。根据定理 5.3，增量 $d_j := \mathsf{Trunc}_{\ell_{\mathrm{inc}}}(\gamma_j g_j)$ 满足 $d_j \in \mathsf{F}_+(\ell_{\mathrm{inc}}, B_{\mathrm{bnd}})$，并且

$$(1 - 2^{1-\ell_{\mathrm{inc}}})\, \gamma_j g_j \le d_j \le \gamma_j g_j\,. \tag{8}$$

我们现在证明，比值 d_j/g_j 接近 1，且随 j 增大。一方面，对于所有 $j \in [J]$，我们有 $d_j/g_j \le \gamma_j < 1$，并且

$$\begin{aligned} \frac{d_j}{g_j} &\ge (1 - 2^{1-\ell_{\mathrm{inc}}})\, \gamma_j && (\text{Equation (8)}) \\ &\ge (1 - 2^{1-\ell_{\mathrm{inc}}})\Big(1 - \frac{\varepsilon}{4}\Big) && (\gamma_j \ge 1 - \tfrac{\varepsilon}{4}) \\ &\ge 1 - \frac{\varepsilon}{4} - 2^{1-\ell_{\mathrm{inc}}} \\ &\ge 1 - \frac{\varepsilon}{2} && (\text{Equation (7)})\,. \end{aligned}$$

另一方面，对于所有 $j \in [J-1]$，我们有

$$\begin{aligned} \frac{d_j}{g_j} &\le \gamma_j && (\text{Equation (8)}) \\ &= (1 - 2^{1-\ell_{\mathrm{inc}}})\, \gamma_{j+1} - 2^{1-\ell_{\mathrm{inc}}}(4 - \gamma_{j+1}) && (\gamma_{j+1} - \gamma_j = 4 \cdot 2^{1-\ell_{\mathrm{inc}}}) \\ &< (1 - 2^{1-\ell_{\mathrm{inc}}})\, \gamma_{j+1} && (\gamma_{j+1} < 1) \\ &\le \frac{d_{j+1}}{g_{j+1}} && (\text{Equation (8)})\,. \end{aligned}$$

因此，我们有

$$1 - \frac{\varepsilon}{2} \le \frac{d_1}{g_1} < \frac{d_2}{g_2} < \cdots < \frac{d_J}{g_J} \le 1\,. \tag{9}$$

- *(下界)* 令 $f\colon [0,1] \to \mathbb{R}$ 为连续分段线性函数，它在$[v_{j-1}, v_j]$上为线性函数，对所有 $j \in [J]$ 均成立，并且对所有 $j \in \{0, 1, \ldots, J\}$均满足，

$$f(v_j) = \sum_{l=1}^{j} d_l\,.$$

那么， f 在 $[v_{j-1}, v_j]$ 上的斜率为 d_j/g_j。由公式（9）可知，斜率为正且递增，因此 f 单调不减且为凸函数。此外，由于 $f(0) = 0$ 且所有斜率均不超过 1，因此对所有 $x \in [0,1]$，均有 $f(x) \le x$。定义 $L := f \circ \mathrm{val}_\lambda$。首先，根据定理6.8，我们有 $L \preceq \Psi_P(L)$。其次，对每个状态 s，令 j 满足 $\mathrm{val}_\lambda(s) = v_j$。那么，我们有

$$\begin{aligned}
\mathrm{val}_\lambda(s) - L(s) &= \sum_{l=1}^{j} g_l - \sum_{l=1}^{j} d_l && (v_j = \textstyle\sum_{l=1}^{j} g_l, \text{ definition of } f) \\
&= \sum_{l=1}^{j} \Big(1 - \frac{d_l}{g_l}\Big) g_l \\
&\le \frac{\varepsilon}{2} \sum_{l=1}^{j} g_l && (\text{Equation (9)}) \\
&= \frac{\varepsilon}{2}\, v_j \\
&\le \frac{\varepsilon}{2} && (v_j \le 1)\,.
\end{aligned}$$

此外，由公式（9），我们有 $\sum_{l=1}^{j}(1 - d_l/g_l)\, g_l \ge 0$，即 $L(s) \le \mathrm{val}_\lambda(s)$。最后， $L(s)$ 是 $\mathsf{F}^+(\ell_{\mathrm{inc}}, B_{\mathrm{bnd}})$ 中至多 $J \le n+1$ 个元素之和。因此，由定理 5.1 可得 $L(s) \in \mathsf{F}\,((n+1)\ell_{\mathrm{inc}}, B_{\mathrm{bnd}}) \subseteq \mathsf{F}(\ell_{\mathrm{bnd}}, B_{\mathrm{bnd}})$。

- （上界）考虑玩家-3 – 𝔭马尔可夫决策过程 $P\prime$，它与 P 具有相同的状态、动作、转移函数和折扣因子，奖励函数为 $1 - r$。将 $P\prime$ 的算子记为 op，即若 $\mathrm{op} = \max$，则 $\mathrm{op} = \min$；若 $\mathrm{op} = \min$，则 $\overline{\mathrm{op}} = \max$。首先，我们联系 $\overline{P}$ 与 $P\prime$ 的贝尔曼算子。对于每个函数 $\rho\colon\ S \to \mathbb{R}$ 以及状态 $s \in S$，我们有

$$\begin{aligned}
\Psi_{P'}(\rho)(s) &= \operatorname*{\overline{op}}_{a \in \mathcal{A}} \Big((1-\lambda)(1 - r(s,a)) + \lambda \sum_{s' \in S} \delta(s,a)(s')\, \rho(s') \Big) && (\text{definition of } \Psi_{P'}) \\
&= \operatorname*{\overline{op}}_{a \in \mathcal{A}} \Big(1 - (1-\lambda)\, r(s,a) - \lambda \sum_{s' \in S} \delta(s,a)(s')(1 - \rho(s')) \Big) && (\delta(s,a) \in \Delta(S)) \\
&= 1 - \operatorname*{op}_{a \in \mathcal{A}} \Big((1-\lambda)\, r(s,a) + \lambda \sum_{s' \in S} \delta(s,a)(s')(1 - \rho(s')) \Big) && (\text{definition of } \overline{\mathrm{op}}) \\
&= 1 - \Psi_P(\vec{1} - \rho)(s) && (\text{definition of } \Psi_P)\,.
\end{aligned}$$

因此，对于每个函数 $\rho\colon\ S \to \mathbb{R}$，我们有

$$\Psi_{P'}(\rho) = \vec{1} - \Psi_P(\vec{1} - \rho)\,. \tag{10}$$

特别地， $\vec{1} - \mathrm{val}_\lambda^P$是 $\Psi_{P\prime}$的不动点，即 $\mathrm{val}_\lambda^{P'} = \vec{1} - \mathrm{val}_\lambda^P$。此外， P 的所有奖励和转移概率的一个公分母，也是 $P\prime$的所有奖励和转移概率的一个公分母，因此 P' 满足引理的假设，且具有相同的 B_P。其次，将前述条目应用于 P'，可得存在一个函数$L'\colon\ S \to \mathbb{R}$，使得 $L' \preceq \Psi_{P\prime}(L')$、$\mathrm{val}^{P}{}_{\lambda} - \frac{\varepsilon}{2} \cdot \vec{1} \preceq L' \preceq \mathrm{val}^{P}{}_{\lambda}$，并且对于所有状态 s， $L'(s) \in \mathcal{F}((n+1)\ell_{\mathrm{inc}}, B_{\mathrm{bnd}})$。定义 $U := \vec{1} - L'$。于是，我们有

$$\begin{aligned}
\Psi_P(U) &= \Psi_P(\vec{1} - L') && (\text{definition of } U) \\
&= \vec{1} - \Psi_{P'}(L') && (\text{公式（10）}) \\
&\preceq \vec{1} - L' && (L' \preceq \Psi_{P\prime}(L')) \\
&= U\,.
\end{aligned}$$

此外，由于 $U - \mathrm{val}_\lambda^P = \mathrm{val}_\lambda^{P'} - L'$，我们有 $\mathrm{val}_\lambda^P \preceq U \preceq \mathrm{val}_\lambda^P + \frac{\varepsilon}{2} \cdot \vec{1}$。最后，对于每个状态 s，由于 $U(s) = 2^{-0} - L'(s)$，根据定理 5.1，我们有 $U(s) \in \mathrm{F}((n+1)\ell_{\mathrm{inc}} + 1, B_{\mathrm{bnd}}) = \mathrm{F}(\ell_{\mathrm{bnd}}, B_{\mathrm{bnd}})$。

- （区间界）根据前述条目，我们有 $L \preceq \mathrm{val}_\lambda \preceq U$，因此 $I = (L, U)$ 是一个区间界。此外，I 是收缩的，因为 $L \preceq \Psi_P(L)$ 和 $\Psi_P(U) \preceq U$。最后，I 是 ε-接近的，因为对于所有状态 $s \in S$，$U(s) - L(s) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$，至此证明完成。 □

6.4 定理 5.8的证明

在本小节中，我们通过结合定理 5.8、定理 6.5、6.9和 6.10来证明。

定理 *5.8*的证明。我们分四步构造见证。

- （策略）设 ℓ_{str} 和 B_{str} 分别为定理 6.5中关于加性误差 $\frac{\varepsilon}{4} = 2^{-(B_\varepsilon + 2)}$ 的常数，即:

$$\ell_{\mathrm{str}} := B_\varepsilon + \lceil \log(n+2) \rceil + \lceil \log m \rceil + 7 \qquad \text{and} \qquad B_{\mathrm{str}} := B_\lambda + B_\varepsilon + \lceil \log m \rceil + \ell_{\mathrm{str}} + 5 .$$

根据定理 6.5，存在 $\frac{\varepsilon}{4}$-最优平稳策略 σ 和 τ，使得 $\sigma(s)\ \tau(s) \in \mathcal{D}(m\ell_{\mathrm{str}}, B_{\mathrm{str}})$ 对于所有状态 $s \in S$均成立。由于 σ 和 τ 都是 $\frac{\varepsilon}{4}$-最优的，对于所有状态 $s \in S$，我们有

$$\mathrm{val}_\lambda(s) - \frac{\varepsilon}{4} \leq \mathrm{val}_\lambda^{G_\sigma}(s) \leq \mathrm{val}_\lambda(s) \leq \mathrm{val}_\lambda^{G_\tau}(s) \leq \mathrm{val}_\lambda(s) + \frac{\varepsilon}{4} . \tag{11}$$

- （公分母）令 $Q \leq 2^{B_{\mathrm{inp}} nm^2(n+1)}$ 为 G 的所有奖励和转移概率的分母之积，并令 $B_P := B_{\mathrm{str}} + B_{\mathrm{inp}} nm^2(n+1)$。对于所有状态$s$ 和动作 a，$\sigma(s)(a) \in \mathrm{F}(m\ell_{\mathrm{str}}, B_{\mathrm{str}})$ 的指数均小于 $2^{B_{\mathrm{str}}}$，因此$\sigma(s)(a)$ 是 $2^{-2^{B_{\mathrm{str}}}}$ 的整数倍。此外，G 的所有奖励和转移概率都是 $1/Q$的整数倍。根据 G_σ 的定义，其奖励和转移概率是这些数的乘积之和，例如，$r_\sigma(s,b) = \sum_{a \in A} \sigma(s)(a)\ r(s,a,b)$。因此，$G_\sigma$ 的所有奖励和转移概率都是 $1/(2^{2^{B_{\mathrm{str}}}} Q)$ 的整数倍，其中

$$\begin{aligned} 2^{2^{B_{\mathrm{str}}}} Q &\leq 2^{2^{B_{\mathrm{str}}} + B_{\mathrm{inp}} nm^2(n+1)} && (Q \leq 2^{B_{\mathrm{inp}} nm^2(n+1)}) \\ &\leq 2^{2^{B_{\mathrm{str}} + B_{\mathrm{inp}} nm^2(n+1)}} && (2^a + b \leq 2^{a+b}) \\ &= 2^{2^{B_P}} && (\text{definition of } B_P) . \end{aligned}$$

对于 G_τ，结论同样成立。因此，G_σ和 G_τ满足定理 6.9 的假设，其中 B_P，并且 $B_{\mathrm{sep}} = B_\lambda + B_P + 2\lceil \log n \rceil + 1$ 如该定理中所定义。

- *(*区间界*)* 令 ℓ_{bnd} 和 B_{bnd} 为定理 6.10 中加性误差 $\frac{\varepsilon}{4}$ 对应的常数，即

$$\ell_{\mathrm{bnd}} := (n+1)(B_\varepsilon + \lceil \log(n+1) \rceil + 10) + 1 \quad \text{and} \quad B_{\mathrm{bnd}} := B_{\mathrm{sep}} + B_\varepsilon + \lceil \log(n+1) \rceil + 12 .$$

根据 定理 6.10，存在 $\frac{\varepsilon}{4}$-接近的收缩区间界 (L_σ, U_σ)，其对应于 G_σ 以及 (L_τ, U_τ)，其对应于 G_τ，其各项均属于 $\mathcal{F}(\ell_{\mathrm{bnd}}, B_{\mathrm{bnd}})$。由于这些区间界是收缩的，根据定理 5.5，对于所有状态 $s \in S$，有

$$L_\sigma(s) \leq \mathrm{val}_\lambda^{G_\sigma}(s) \leq U_\sigma(s) \qquad \text{and} \qquad L_\tau(s) \leq \mathrm{val}_\lambda^{G_\tau}(s) \leq U_\tau(s) . \tag{12}$$

- （见证）我们将区间界扩大 $\frac{\varepsilon}{4}$，即定义

$$I_\sigma := \left(L_\sigma, U_\sigma + \frac{\varepsilon}{4} \cdot \vec{1} \right) \qquad \text{and} \qquad I_\tau := \left(L_\tau - \frac{\varepsilon}{4} \cdot \vec{1}, U_\tau \right) .$$

由于 $L_\sigma \preceq U_\sigma \preceq U_\sigma + \frac{\varepsilon}{4}\cdot\vec{1}$ 且 $L_\tau - \frac{\varepsilon}{4}\cdot\vec{1} \preceq L_\tau \preceq U_\tau$，因此这些对 I_σ 和 I_τ 是区间界。我们证明 $(\sigma, \tau, I_\sigma, I_\tau)$ 是一个多项式规模的见证。首先，我们证明 I_σ 和 I_τ 是收缩的。对于任意马尔可夫决策过程 P、函数 $\rho\colon S \to \mathbb{R}$ 以及常数 $c \in \mathbb{R}$，由于从每个状态和动作出发的转移概率之和为一，因此有

$$\Psi_P(\rho + c\cdot\vec{1}) = \Psi_P(\rho) + \lambda c\cdot\vec{1}\,. \tag{13}$$

因此，我们有

$$\begin{aligned}
\Psi_{G_\sigma}\left(U_\sigma + \frac{\varepsilon}{4}\cdot\vec{1}\right) &= \Psi_{G_\sigma}(U_\sigma) + \lambda\frac{\varepsilon}{4}\cdot\vec{1} && \text{(Equation (13))}\\
&\preceq U_\sigma + \lambda\frac{\varepsilon}{4}\cdot\vec{1} && ((L_\sigma, U_\sigma) \text{ is shrinking})\\
&\preceq U_\sigma + \frac{\varepsilon}{4}\cdot\vec{1} && (\lambda < 1)\,.
\end{aligned}$$

此外，由于 (L_σ, U_σ) 是收缩的，我们有 $L_\sigma \preceq \Psi_{G_\sigma}(L_\sigma)$，因此 I_σ 是收缩的。对称地，I_τ 也是收缩的。其次，I_σ 和 I_τ 是 $\frac{\varepsilon}{2}$-接近的，因为 (L_σ, U_σ) 和 (L_τ, U_τ) 是 $\frac{\varepsilon}{4}$-接近的，而扩张部分增加了 $\frac{\varepsilon}{4}$。第三，我们证明 I_σ 和 I_τ 有交集。对于每个状态 $s \in S$，我们有

$$\begin{aligned}
L_\sigma(s) &\le \mathrm{val}_\lambda^{G_\sigma}(s) && \text{(方程（12）)}\\
&\le \mathrm{val}_\lambda(s) && \text{(方程（11）)}\\
&\le \mathrm{val}_\lambda^{G_\sigma}(s) + \frac{\varepsilon}{4} && \text{(方程（11）)}\\
&\le U_\sigma(s) + \frac{\varepsilon}{4} && \text{(方程（12）)},
\end{aligned}$$

也就是说，$\mathrm{val}_\lambda(s)$ 位于 s 处 I_σ 的区间内。对称地，$\mathrm{val}_\lambda(s)$ 位于 s 处 I_τ 的区间内，因此 I_σ 与 I_τ 相交。最后，我们给出见证大小的界。由于 $\frac{\varepsilon}{4} = 2^{-(B_\varepsilon+2)}$ 且 $2^{B_{\mathrm{bnd}}} > B_\varepsilon + 2$，根据定理 5.1，$I_\sigma$ 和 I_τ 的各项都属于 $\mathsf{F}(\ell_{\mathrm{bnd}}+1, B_{\mathrm{bnd}})$。策略至多包含 $2nm$ 个属于 $\mathsf{F}(m\ell_{\mathrm{str}}, B_{\mathrm{str}})$ 的概率，而区间界由 $4n$ 个属于 $\mathsf{F}(\ell_{\mathrm{bnd}}+1, B_{\mathrm{bnd}})$ 的数构成。因此，见证的大小为 $\mathsf{O}(nm(m\ell_{\mathrm{str}}B_{\mathrm{str}} + \ell_{\mathrm{bnd}}B_{\mathrm{bnd}}))$。由于 ℓ_{str}、B_{str}、B_P、B_{sep}、ℓ_{bnd} 和 B_{bnd} 都是这些参数的多项式，所以该大小关于 n、m、B_{inp}、B_ε 和 B_λ 是多项式级的，由此完成证明。 □

7 极限平均与可达性目标

在本节中，我们证明定理 3.1。为此，我们首先将可达性目标归约为极限平均目标。然后，我们用具有简洁折扣因子的折扣和目标的值来近似极限平均目标的值，并应用定理 3.2。

引理 7.1 (将可达性表示为极限平均)。 考虑一个博弈 G，其转移函数为 δ，目标集为 T。设 G_T 为从 G 得到的博弈，其中每个目标状态都变为吸收状态，即其转移函数为 δ_T，并设 r_T 为奖励函数，对所有状态 s 和动作 a、b，定义如下：

$$\delta_T(s,a,b) \coloneqq \begin{cases} \mathbb{1}[s] & s \in T\,,\\ \delta(s,a,b) & \textit{otherwise,}\end{cases} \qquad \textit{and} \qquad r_T(s,a,b) \coloneqq \begin{cases} 1 & s \in T\,,\\ 0 & \textit{otherwise,}\end{cases}$$

其中 $\mathbb{1}[s]$ 是 s 上的狄拉克分布。于是，对于所有状态 $s \in S$，都有 $\mathrm{val}_T^G(s) = \mathrm{val}^{G_T}(s)$，其中 val^{G_T} 是 G_T 中 $\mathit{LimAvg}(r_T)$ 的值。特别地，如果每个目标状态在 G 中都是吸收的，那么 $G_T = G$。

证明。 G_T 的每一场对局要么永远不会访问 T，此时其所有奖励均为 0；要么访问 T 并永远停留在那里，此时其极限平均为 1。此外， G 和 G_T 的对局在首次访问 T 之前完全相同。因此，对于所有状态 $s \in S$，都有 $\mathsf{val}_T^G(s) = \mathsf{val}^{G_T}(s)$。最后，如果每个目标状态在 G 中都是吸收的，那么 $\delta_T = \delta$，即 $G_T = G$。 □

极限平均目标的值是折扣因子趋近于 1 时折扣和目标的值的极限 [MN81]。下面的结果界定了折扣因子需要多接近 1，才能使这两个值接近。该结果由 [OB21, Theorem 3],推出；其中，我们用上界 m^n替换玩家的纯平稳策略的最小数量 $\min\{|A|^n, |B|^n\}$，并分别用其上界 $\lceil \log n \rceil + 1$、$n\lceil \log m \rceil + 1$ 和 $B_Q + 1$替换 n、 m^n 和 Q 的比特数，这只会增大 B_λ。

引理 7.2 ([OB21, 定理 3])。 设 G 是一个博弈，且 r 是一个奖励函数，使得所有转移概率和奖励都是 $1/Q$的整数倍，其中 $Q \leq 2^{B_Q}$为正整数。设 $\varepsilon = 2^{-B_\varepsilon}$，并设

$$B_\lambda := \lceil \log\left(4nm^n\left(\lceil \log n \rceil + n\lceil \log m \rceil + B_Q + 3\right) + B_\varepsilon\, n\, m^n\right) \rceil\ .$$

那么，对于每个折扣因子 $\lambda \in [1 - 2^{-2^{B_\lambda}}, 1)$ 和状态 $s \in S$，我们有 $|\text{值}_\lambda(s) - \text{值}(s)| \leq \varepsilon$。

现在我们可以证明主要结果了。

定理 *3.1* 的证明。 根据定理 7.1，由于 G_T和 r_T可在多项式时间内计算，因此只需对极限平均目标证明该结果。考虑一个实例 $(G,\ \phi)$，其极限平均目标为 $\phi = \text{极限平均}(r)$，加性误差为 $\varepsilon = 2^{-B_\varepsilon}$。我们分三步进行。

- *(折扣因子)* 我们用多项式时间构造一个简洁折扣因子 λ，使得 $\mathsf{Disc}(r, \lambda)$ 的值近似于 ϕ的值，误差不超过 $\frac{\varepsilon}{4}$。令 Q 为 G所有转移概率和奖励的分母的乘积。由于这类数共有 $nm^2(n+1)$ 个，且每个数的分母至多为 $2^{B_{\text{inp}}}$，因此对于 $B_Q := B_{\text{inp}} nm^2(n+1)$，我们有 $Q \leq 2^{B_Q}$。令 B_λ 为定理 7.2 中针对加性误差 $\frac{\varepsilon}{4} = 2^{-(B_\varepsilon+2)}$的常数，即，

$$B_\lambda := \lceil \log\left(4nm^n\left(\lceil \log n \rceil + n\lceil \log m \rceil + B_Q + 3\right) + (B_\varepsilon + 2)\, n\, m^n\right) \rceil\ ,$$

  并令 $\lambda := 1 - 2^{-2^{B_\lambda}}$。由于 $\log(nm^n) = \log n + n \log m$，且 B_Q是关于 n的多项式， m、B_{inp}和 B_λ 的位数是关于 n、 m、 B_{inp}和 B_ε的多项式，并且可在多项式时间内计算。根据 定理 7.2，对于所有状态 $s \in S$，有

$$|\text{val}_\lambda(s) - \text{val}(s)| \leq \frac{\varepsilon}{4}\,. \tag{14}$$

- （价值近似）根据 定理 3.2，折扣总和目标 DiscCsgApproxValue 问题(r, λ) 和加性误差 $\frac{\varepsilon}{2}$ 可在多项式时间内归约为 TFNP 中的一个搜索问题。设 $\hat{v}$ 是该问题的一个解，即 $|\hat{v}(s) - \text{val}_\lambda(s)| \leq \frac{\varepsilon}{2}$对所有状态均成立$s \in S$。那么，对于所有状态 $s \in S$，有

$$\begin{aligned} |\hat{v}(s) - \text{val}(s)| &\leq |\hat{v}(s) - \text{val}_\lambda(s)| + |\text{val}_\lambda(s) - \text{val}(s)| && \text{(triangle inequality)} \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{4} && \text{(choice of } \hat{v},\ \text{Equation (14)}) \\ &\leq \varepsilon\,. \end{aligned}$$

  因此， ϕ 的 CsgApproxValue 问题属于 TFNP。

- （间隙问题）考虑一个状态 s 和一个阈值 α，并假设 $\mathrm{val}(s) \geq \alpha + \varepsilon$ 或 $\mathrm{val}(s) \leq \alpha - \varepsilon$，因为否则任意答案都是正确的。如果 $\mathrm{val}(s) \geq \alpha + \varepsilon$，那么我们有

$$\begin{aligned} \mathrm{val}_\lambda(s) &\geq \mathrm{val}(s) - \frac{\varepsilon}{4} && \text{（公式（14））} \\ &\geq \alpha + \frac{\varepsilon}{2} && (\mathrm{val}(s) \geq \alpha + \varepsilon)。 \end{aligned}$$

  对称地，如果 $\mathrm{val}(s) \leq \alpha - \varepsilon$，那么我们有 $\mathrm{val}_\lambda(s) \leq \alpha - \frac{\varepsilon}{2}$。因此，该实例的答案与折扣和目标 Disc 的CsgGapValue问题的实例 (r, λ) 相同，且该实例具有相同的状态、相同的阈值以及加性误差 $\frac{\varepsilon}{2}$，这些内容均可在多项式时间内计算。根据定理 3.2，CsgGapValue 问题 ϕ 属于 NP∩ coNP，从而完成证明。 □

8 价值判定

在本节中，我们证明定理 4.1，即CsgDecision 问题对于极限平均目标和可达性目标是 PosSLP-困难且属于$\exists\mathbb{R} \cap$co-$\exists\mathbb{R}$。我们将在第 8.1 节证明下界，并在第 8.2 节证明上界。

8.1 下界

在本小节中，我们证明定理 4.1的下界。为此，我们通过博弈模拟直线程序，其中心化值编码程序所计算的整数。

定理 4.1 下界的证明。我们将 PosSLP 归约到具有可达性目标的 CsgDecision 问题。根据定理 7.1，可达性目标可在多项式时间内归约为具有相同值的极限平均目标，并且该归约保留目标状态为吸收态的博弈中的转移。因此，对于极限平均目标同样成立。我们构造一个博弈 G，其动作集为 $\mathrm{A} = \{a_1, a_2\}$ 和 $\mathrm{B} = \{b_1, b_2\}$，并包含两个吸收态 $\top$和 $\bot$；我们考虑目标集 $T := \{\top\}$。我们逐个添加状态，并且每个新状态只转移到之前添加的状态。因此，每个对局至多经过 n 步就会到达 $\top$ 或 $\bot$。对于每个状态 $s \notin \{\top, \bot\}$，令 M_s 为一个 2×2 矩阵，其元素为 $M_s(a,b) := \sum_{s' \in S} \delta(s,a,b)(s')\,\mathrm{val}_T(s')$。根据逆向归纳法，$\mathrm{val}_T(s)$ 是矩阵博弈 M_s 的值，其中玩家1选择一行，玩家 2 选择一列。对于每个状态 s，我们将其中心化值记为 $\chi(s) := 2\mathrm{val}_T(s) - 1 \in [-1, 1]$。证明包含以下步骤。

- *(构件)* 我们构造一些状态，用于对中心化值取反、求平均和相乘。
- *(直线程序)* 我们将每个门对应的整数表示为两个状态的中心化值之比。
- *(归约)* 我们将输出门对应状态的值与阈值 $\frac{1}{2}$进行比较。
- *(确定性转移)* 我们将每个随机转移替换为保持值不变的确定性转移。

现在我们证明每个步骤。

- *(构件)* 首先，对于每个新状态 s，我们添加一个 对偶 状态 $\bar{s}$，交换玩家的角色，并将每个后继状态替换为其对偶；也就是说，在状态 $\bar{s}$ 下，动作对 (a_i, b_j) 会导向动作对 (a_j, b_i) 在状态 s 下对应后继状态的对偶。形式化地，对于所有 $i, j \in \{1,2\}$ 以及状态 $s' \in S$，我们定义

$$\delta(\bar{s}, a_i, b_j)(\overline{s'}) := \delta(s, a_j, b_i)(s'),$$

  其中 $\overline{\top} := \bot$ 和 $\overline{\bot} := \top$，而 $\bar{s}$ 的对偶是 s。通过对构造进行归纳法 n,

对于 $\bar{s}$的所有后继状态 s'，$_T(s') = 1 - _T(s')$，因此，对于所有 $i,\ j \in \{1,\ 2\}$，我们有

$$\begin{aligned} M_{\bar{s}}(a_i, b_j) &= \sum_{s' \in S} \delta(\bar{s}, a_i, b_j)(\overline{s'})\, \mathrm{val}_T(\overline{s'}) && \text{(definition of } M_{\bar{s}}) \\ &= \sum_{s' \in S} \delta(s, a_j, b_i)(s')\, (1 - \mathrm{val}_T(s')) && \text{(definition of } \bar{s},\ \text{induction hypothesis}) \\ &= 1 - M_s(a_j, b_i) && \text{(definition of } M_s)\,. \end{aligned}$$

在矩阵博弈 $M_{\bar{s}}$中，玩家1扮演 M_s中玩家 2的角色，因此

$$\mathrm{val}_T(\bar{s}) = 1 - \mathrm{val}_T(s)\,, \qquad \text{i.e.,} \qquad \chi(\bar{s}) = -\chi(s)\,. \tag{15}$$

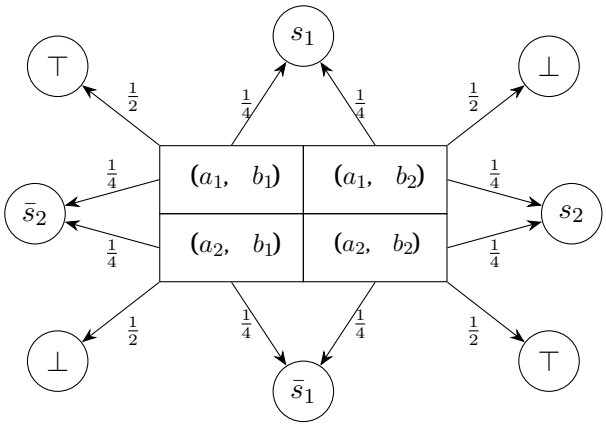
其次，对于所有状态 s_1 和 s_2，令 Avg(s_1，s_2) 表示一个状态，它以相同的概率$_1$ $_2$ 分别转移到 s_1 和 s_2 $\bar{}$。于是，我们有

$$\chi(\mathrm{Avg}(s_1, s_2)) = \frac{\chi(s_1) + \chi(s_2)}{2}\,. \tag{16}$$

特别地，由于 $\chi(\top) = 1$ 且 $\chi(\bot) = -1$，我们有 χ(Avg($\top$，$\bot$)) $= 0$。第三，对于所有状态s_1 和 s_2，令 Mul(s_1，s_2) 表示状态 s，其转移为

$$\begin{aligned} \delta(s, a_1, b_1) &:= \tfrac{1}{2}\mathbb{1}[\top] + \tfrac{1}{4}\mathbb{1}[s_1] + \tfrac{1}{4}\mathbb{1}[\bar{s}_2]\,, & \delta(s, a_1, b_2) &:= \tfrac{1}{2}\mathbb{1}[\bot] + \tfrac{1}{4}\mathbb{1}[s_1] + \tfrac{1}{4}\mathbb{1}[s_2]\,, \\ \delta(s, a_2, b_1) &:= \tfrac{1}{2}\mathbb{1}[\bot] + \tfrac{1}{4}\mathbb{1}[\bar{s}_1] + \tfrac{1}{4}\mathbb{1}[\bar{s}_2]\,, & \delta(s, a_2, b_2) &:= \tfrac{1}{2}\mathbb{1}[\top] + \tfrac{1}{4}\mathbb{1}[\bar{s}_1] + \tfrac{1}{4}\mathbb{1}[s_2]\,. \end{aligned}$$

这些转移如下面的图所示，其中每个单元格都是一个动作对，并且$\top$和 $\bot$各画了两次，以避免箭头交叉。



设 $x :=\chi(s_1)$ 以及 $y :=\chi(s_2)$。根据 公式（15），可得

$$M_{\mathrm{Mul}(s_1, s_2)} = \begin{pmatrix} \frac{3}{4} + \frac{x-y}{8} & \frac{1}{4} + \frac{x+y}{8} \\ \frac{1}{4} - \frac{x+y}{8} & \frac{3}{4} - \frac{x-y}{8} \end{pmatrix}.$$

如果玩家1以概率 $\frac{1}{2} + \frac{y}{4}$ 选择 a_1，那么无论面对 b_1 还是 b_2，期望收益都是 $\frac{1}{2} + \frac{xy}{16}$。类似地，如果玩家 2 以概率 $\frac{1}{2} - \frac{x}{4}$ 选择 b_1，那么无论面对 a_1 还是 a_2，期望收益都是 $\frac{1}{2} + \frac{xy}{16}$。由于 $x, y \in [-1, 1]$，这两个数都在 $[0, 1]$, 中，因此这些混合策略分别保证玩家1和玩家 2 获得期望收益 $\frac{1}{2} + \frac{xy}{16}$。因此，我们有 val_T(Mul(s_1, s_2)) $=$val($M_{\overline{\mathrm{Mul}(s_1, s_2)}}$) $= \frac{1}{2} + \frac{xy}{16}$，从而

$$\chi(\mathrm{Mul}(s_1, s_2)) = \frac{\chi(s_1)\,\chi(s_2)}{8}\,. \tag{17}$$

- （直线程序）考虑一个包含门 $1,\ldots,k$ 的直线程序，并回忆一下， z_i 是门 i 计算出的整数。整数 z_i 可能大到双指数级，而中心化值属于 $[-1,1]$。因此，我们将 z_i 表示为两个状态的中心化值之比，也就是说，对于每个门 i，我们构造一个分子 状态 $s_{\mathrm{num}i}$ 和一个分母 状态s_{den} i ，使得

$$\chi(s_i^{\mathrm{den}}) > 0 \qquad \text{and} \qquad \chi(s_i^{\mathrm{num}}) = z_i\,\chi(s_i^{\mathrm{den}})\,. \tag{18}$$

我们通过对门进行归纳来构造这些状态，并证明方程（18）对每一种门类型都成立。

– （常数）如果 $z_i = 0$，则令 $s_i^{\mathrm{num}} := \mathrm{Avg}(\top,\bot)$ 和 $s_i^{\mathrm{den}} := \top$。如果 $z_i = 1$，则令$s_i^{\mathrm{num}} := \top$ 和 $s_i^{\mathrm{den}} := \top$。由于 $\chi(\mathrm{Avg}(\top,\bot)) = 0$ 且 $\chi(\top) = 1$，方程（18）成立。– （乘法）如果 $z_i = z_j \times z_{j'}$，则令

$$s_i^{\mathrm{num}} := \mathrm{Mul}(s_j^{\mathrm{num}}, s_{j'}^{\mathrm{num}}) \qquad \text{and} \qquad s_i^{\mathrm{den}} := \mathrm{Mul}(s_j^{\mathrm{den}}, s_{j'}^{\mathrm{den}})\,.$$

根据方程（17）和归纳假设，我们有 $\chi(s_i^{\mathrm{den}}) = \frac{1}{8}\chi(s_j^{\mathrm{den}})\,\chi(s_{j'}^{\mathrm{den}}) > 0$以及

$$\begin{aligned}
\chi(s_i^{\mathrm{num}}) &= \frac{\chi(s_j^{\mathrm{num}})\,\chi(s_{j'}^{\mathrm{num}})}{8} && \text{(Equation (17))}\\
&= z_j z_{j'}\,\frac{\chi(s_j^{\mathrm{den}})\,\chi(s_{j'}^{\mathrm{den}})}{8} && \text{(induction hypothesis)}\\
&= z_i\,\chi(s_i^{\mathrm{den}}) && \text{(Equation (17))}\,.
\end{aligned}$$

– （加法）如果 $z_i = z_j + z_{j'}$，则令

$$\begin{aligned}
s_i^{\mathrm{num}} &:= \mathrm{Avg}\big(\mathrm{Mul}(s_j^{\mathrm{num}}, s_{j'}^{\mathrm{den}}),\, \mathrm{Mul}(s_{j'}^{\mathrm{num}}, s_j^{\mathrm{den}})\big)\,,\\
s_i^{\mathrm{den}} &:= \mathrm{Avg}\big(\mathrm{Mul}(s_j^{\mathrm{den}}, s_{j'}^{\mathrm{den}}),\, \mathrm{Avg}(\top,\bot)\big)\,.
\end{aligned}$$

根据方程（16）和（17）以及归纳假设，我们有 $\chi(s_i^{\mathrm{den}}) = \frac{1}{16}\chi(s_j^{\mathrm{den}})\,\chi(s_{j'}^{\mathrm{den}}) > 0$ 以及

$$\begin{aligned}
\chi(s_i^{\mathrm{num}}) &= \frac{\chi(s_j^{\mathrm{num}})\,\chi(s_{j'}^{\mathrm{den}}) + \chi(s_{j'}^{\mathrm{num}})\,\chi(s_j^{\mathrm{den}})}{16} && \text{(Equations (16) and (17))}\\
&= (z_j + z_{j'})\,\frac{\chi(s_j^{\mathrm{den}})\,\chi(s_{j'}^{\mathrm{den}})}{16} && \text{(induction hypothesis)}\\
&= z_i\,\chi(s_i^{\mathrm{den}}) && \text{(Equations (16) and (17))}\,.
\end{aligned}$$

– （减法）如果 $z_i = z_j - z_{j'}$，则令 s_i^{den} 的定义与加法相同，并令

$$s_i^{\mathrm{num}} := \mathrm{Avg}\big(\mathrm{Mul}(s_j^{\mathrm{num}}, s_{j'}^{\mathrm{den}}),\, \overline{\mathrm{Mul}(s_{j'}^{\mathrm{num}}, s_j^{\mathrm{den}})}\big)\,.$$

至于加法，我们有 $\chi(s_i^{\mathrm{den}}) = \frac{1}{16}\chi(s_j^{\mathrm{den}})\,\chi(s_{j'}^{\mathrm{den}}) > 0$ 以及

$$\begin{aligned}
\chi(s_i^{\mathrm{num}}) &= \frac{\chi(s_j^{\mathrm{num}})\,\chi(s_{j'}^{\mathrm{den}}) - \chi(s_{j'}^{\mathrm{num}})\,\chi(s_j^{\mathrm{den}})}{16} && \text{(Equations (15) to (17))}\\
&= (z_j - z_{j'})\,\frac{\chi(s_j^{\mathrm{den}})\,\chi(s_{j'}^{\mathrm{den}})}{16} && \text{(induction hypothesis)}\\
&= z_i\,\chi(s_i^{\mathrm{den}}) && \text{(Equations (16) and (17))}\,.
\end{aligned}$$

该构造为每个门增加固定数量的状态，并且不会计算整数 z_i。

- *(归约)* 我们在程序中追加常数个门，用以计算 $z^\star := 2z_k - 1$。由于 $z^\star$ 是奇整数，因此我们有 $z^\star \neq 0$，并且 $z_k > 0$ 当且仅当 $z^\star > 0$。令 $s_{\text{num}\star}$ 和 $s_{\text{den}\star}$ 分别为输出门的状态。根据方程（18），我们有

$$\text{val}_T(s_\star^{\text{num}}) = \frac{1 + \chi(s_\star^{\text{num}})}{2} = \frac{1}{2} + \frac{z^\star \chi(s_\star^{\text{den}})}{2},$$

其中 $\chi(s_\star^{\text{den}}) > 0$。因此，我们有 $z_k > 0$ 当且仅当 $\text{val}_T(s_\star^{\text{num}}) \geq \frac{1}{2}$。

- *(确定性转移)* 在并发博弈中，随机转移可以通过确定性转移来模拟；也就是说，存在一个多项式时间归约，可将其归约为具有确定性转移的并发博弈，同时保持可达性目标 [EY08，命题 2。1],另见 [CDGH15]。该归约将每个概率选择替换为新的状态，在这些状态中，双方玩家各有两个动作，后继状态仅取决于双方动作是否匹配。由于上述所有转移概率都是 $1\ 4$ 的整数倍，这些新状态不会引入环，因此每场对局仍会到达 $\top$ 或 $\bot$，证明至此完成。

□

注8.1（双指数级接近的值）。构造的博弈中的值可能与阈值 $1\ 2$双指数级接近。例如，如果每个门 $i \geq 2$ 都将前一个门的值平方，并从常数 1开始，那么 $\chi(s_i^{\text{num}}) = 2^{-3(2^{i-1}-1)}$。因此，定理 4.1的下界并不与定理 3.1矛盾，因为其中的加性误差是以二进制表示的。

8.2 上界

在本小节中，我们证明定理 4.1的上界。首先，我们证明，折扣和目标的值与实数阈值之间的严格比较，可以由平稳策略以及值函数的界作为见证（定理 8.2）。其次，我们证明，除非极限平均目标的值等于有理数阈值，否则二者至少相差一个双指数级小量（定理 8.3和8.4）。最后，我们结合这些结果，在多项式时间内构造实数域上的存在句，以表示 $\text{val}(s) \geq \alpha$ 和 $\text{val}(s) < \alpha$，由此证明CsgDecision 问题属于 $\exists\mathbb{R} \cap \text{co-}\exists\mathbb{R}$。

实数见证。以下结果表明，对于折扣和目标的值与实数阈值之间的严格比较，可以由一个平稳策略以及值函数的下界或上界作为见证。类似的见证也用于 [Cha07，定理 39]和 [CMH08], ，其中单个存在句包含双方玩家的策略，并同时证明值的下界和上界。我们的见证与第 5.2 节中的见证类似，但其策略和值界是任意实数，而非幂级数数。由于这些条件是多项式不等式，我们用它们构造实数域上的存在句。为确保完备性，我们给出证明。

引理 8.2 (实数见证，参见 [Cha07，定理 39])。设 G 是一个具有奖励函数和折扣因子 λ的博弈，设 $s \in S$ 是一个状态，且 $c \in \mathbb{R}$。那么，以下断言成立。

1.（下界见证）当且仅当存在一个玩家*1*的平稳策略 σ 和一个函数 $L\colon S \to \mathbb{R}$，使得 $L \preceq \Psi_{G_\sigma}(L)$ 且 $L(s) > c$，此时 $val_\lambda(s) > c$。*2.*（上界见证）当且仅当存在一个玩家 *2* 的平稳策略 τ 和一个函数 $U\colon S \to \mathbb{R}$，使得 $\Psi_{G_\tau}(U) \preceq U$ 且 $U(s) < c$，此时 $val_\lambda(s) < c$。

证明。我们给出下见证的证明，上见证的证明与之对称。首先，设 σ 和 L 满足 $L \preceq \Psi_{G_\sigma}(L)$ 且 $L(s) > c$。那么，我们有

$$\begin{aligned} \mathrm{val}_\lambda(s) &\geq \mathrm{val}_\lambda^{G_\sigma}(s) && (\sigma \text{ 是玩家1的策略}) \\ &\geq L(s) && (\text{定理 5.5}) \\ &> c\,. \end{aligned}$$

其次，假设 $\mathrm{val}_\lambda(s) > c$。设 σ 是玩家 1 [Sha53], 的最优平稳策略，且 $L \coloneqq \mathrm{val}_\lambda^{G_\sigma}$。由于 L 是 Ψ_{G_σ} 的不动点，我们有 $L = \Psi_{G_\sigma}(L)$。此外，由于 σ 是最优的，我们有 $L(s) = \mathrm{val}_\lambda(s) > c$，由此完成证明。 □

与阈值的间隔。 为了将极限平均目标的值与有理数阈值进行比较，我们证明：如果两者不同，那么它们至少相差一个双指数小量。为此，我们利用值是一个代数数，即某个非零整数系数多项式的根，且该多项式的次数和系数有如下界。该结果由 [OB21，命题 3 得出。3], 其中，我们将玩家的纯平稳策略的最小数量替换为其上界 m^n，并将 n、m^n 和 Q 的位数替换为它们的上界$\lceil \log n \rceil + 1$，$n\lceil \log m \rceil + 1$ 和 $B_Q + 1$，这只会增大这些界。

引理8.3 ([OB21，命题3。3])。设 G 是一个博弈，且 r 是一个奖励函数，其中所有转移概率和奖励都是 $1/Q$的整数倍，$Q \leq 2^{B_Q}$是一个正整数，并设

$$B_{\mathrm{coef}} \coloneqq \lceil \log\left(4nm^n\left(\lceil \log n \rceil + n\lceil \log m \rceil + B_Q + 3\right)\right) \rceil\,.$$

于是，对于每个状态 $s \in S$，都存在整数 $e_0, \ldots, e_d$，满足 $e_d \neq 0$，使得 $\sum_{i=0}^{d} e_i 值(s)^i = 0$，$d \leq m^n$，且对于所有 $i \in \{0, \ldots, d\}$，均有 $|e_i| < 2^{2^{B_{\mathrm{coef}}}}$。

引理8.4 (与阈值的间隔)。设 G 是一个博弈，且 r 是一个奖励函数，其中所有转移概率和奖励都是 $1/Q$的整数倍，$Q \leq 2^{B_Q}$是一个正整数，并设 $\alpha = p/q \in [0,1]$ 是一个有理数，其定义最多使用 B_{inp} 位，$0 \leq p \leq q$是整数，且满足 $q \geq 1$。设 B_{coef} 如 定理*8.3*中所定义，并设

$$B_\alpha \coloneqq B_{\mathrm{coef}} + \lceil \log\left(m^n\left(B_{\mathrm{inp}} + 3\right)\right) \rceil + 1\,.$$

那么，对于每个状态 $s \in S$ 使得 $\mathrm{val}(s) \neq \alpha$，都有 $|\mathrm{val}(s) - \alpha| \geq 2^{-2^{B_\alpha}}$。

证明。考虑一个满足 $s \neq \alpha$ 的状态 $s \in S$，并令 $\beta \coloneqq \mathrm{val}(s) - \alpha \neq 0$。首先，我们将代数界转化为 β。根据定理8.3，存在整数 $e_0, \ldots, e_d$，满足 $e_d \neq 0$，使得 $\sum_{i=0}^{d} e_i \mathrm{val}(s)^i = 0$、$d \leq m^n$，并且对所有 $i \in \{0, \ldots, d\}$ 都有 $|e_i| < 2^{2^{B_{\mathrm{coef}}}}$。将 $\mathrm{val}(s) = \frac{p}{q} + \beta$ 代入并乘以 q^d，可得

$$\sum_{j=0}^{d} e'_j \beta^j = 0\,, \qquad \text{where} \qquad e'_j \coloneqq \sum_{i=j}^{d} e_i\, q^{d-i} \binom{i}{j} p^{i-j} q^j\,. \tag{19}$$

这些数 e'_j 是整数，且 $e'_d = e_d q^d \neq 0$。此外，对于所有 $j \in \{0, \dots, d\}$，我们有

$$\begin{aligned}
|e'_j| &\leq \sum_{i=j}^{d} 2^{2^{B_{\text{coef}}}} q^d \binom{i}{j} && (|e_i| < 2^{2^{B_{\text{coef}}}},\ p \leq q) \\
&\leq (d+1)\, 2^{2^{B_{\text{coef}}}+d} q^d && \left(\tbinom{i}{j} \leq 2^d\right) \\
&\leq 2^{2^{B_{\text{coef}}}+2m^n+m^n B_{\text{inp}}} && (d+1 \leq 2^d,\ d \leq m^n,\ q \leq 2^{B_{\text{inp}}}) \\
&\leq 2^{2^{B_{\text{coef}}}+m^n(B_{\text{inp}}+3)-1} && (m^n \geq 1) \\
&\leq 2^{2^{B_\alpha}-1} && (\text{definition of } B_\alpha,\ 2^a + 2^b \leq 2^{a+b+1})\,.
\end{aligned}$$

其次，我们给出 $|\beta|$ 的下界。令 j_0 为满足 $e'_{j_0} \neq 0$的最小下标；这样的下标存在，因为$e'_d \neq 0$。反设 $|\beta| < 2^{-2^{B_\alpha}}$。那么，我们有

$$\begin{aligned}
1 &\leq |e'_{j_0}| && (e'_{j_0} \text{ 是一个非零整数}) \\
&= \Big| \sum_{j=j_0+1}^{d} e'_j \beta^{j-j_0} \Big| && (\text{方程（19）}\quad \beta \neq 0) \\
&\leq 2^{2^{B_\alpha}-1} \sum_{j \geq 1} |\beta|^j && (|e'_j| \leq 2^{2^{B_\alpha}-1}) \\
&= 2^{2^{B_\alpha}-1} \frac{|\beta|}{1-|\beta|} && (|\beta| < 1) \\
&< 2^{2^{B_\alpha}} |\beta| && \left(|\beta| < \tfrac{1}{2}\right) \\
&< 1 && (|\beta| < 2^{-2^{B_\alpha}})\,,
\end{aligned}$$

这就产生了矛盾。因此，我们有 $|\text{val}(s) - \alpha| = |\beta| \geq 2^{-2^{B_\alpha}}$，证明至此完成。

□

现在我们可以证明定理 4.1的上界了。

定理 *4.1*上界的证明。 根据定理 7.1，可达性目标可在多项式时间内归约为具有相同值的极限平均目标。因此，只需考虑一个实例(G, ϕ)，其中 $\phi =$ 极限平均(r) 是一个极限平均目标，一个状态 s，以及一个有理数阈值$\alpha = p/q$。由于值$(s) \in [0, 1]$，当 $\alpha \leq 0$ 时答案为肯定，当 $\alpha > 1$时答案为否定，因此我们假设 $\alpha \in (0, 1]$。我们分五步进行。

- (分离) 令 Q 为 G的所有转移概率和奖励的分母之积。如定理 3.1 的证明所示，对于 B_Q，有 $Q \leq 2^{B_Q} := B\text{inp}nm^2(n+1)$。令 B_α 为定理 8.4 中定义的量，并令 $\eta := 2^{-2^{B_\alpha}}$。根据定理 8.4，如果 $\text{val}(s) \neq \alpha$，则 $|\text{val}(s) - \alpha| \geq \eta$。由于 $\log(m^n) = n\log\ m$，且 B_Q 是输入规模的多项式，因此 B_α 是输入规模的多项式，并且可在多项式时间内计算得到。
- *(折扣因子)* 设 B_λ 为定理 7.2 中加性误差为 $\frac{\eta}{4} = 2^{-(2^{B_\alpha}+2)}$时的常数，即，

$$B_\lambda := \left\lceil \log\left(4nm^n\left(\lceil \log n \rceil + n\lceil \log m \rceil + B_Q + 3\right) + \left(2^{B_\alpha}+2\right) n\, m^n\right)\right\rceil\,,$$

并令 $\lambda := 1 - 2^{-2^{B_\lambda}}$。由于 B_α 是关于输入规模的多项式， B_λ 是关于输入规模的多项式，且可在多项式时间内计算。根据 定理 7.2，我们有

$$|\text{val}_\lambda(s) - \text{val}(s)| \leq \frac{\eta}{4}\,. \tag{20}$$

- *(平移阈值)* 设 $c := \alpha - \frac{\eta}{2}$。若 $\mathsf{val}(s) \geq \alpha$，则有

$$\begin{aligned} \mathrm{val}_\lambda(s) &\geq \mathrm{val}(s) - \frac{\eta}{4} && \text{（公式（20））} \\ &\geq \alpha - \frac{\eta}{4} && (\mathsf{val}(s) \geq \alpha) \\ &> c && \text{（}c\text{ 的定义）。} \end{aligned}$$

  如果 $\mathsf{val}(s) < \alpha$，则有

$$\begin{aligned} \mathrm{val}_\lambda(s) &\leq \mathrm{val}(s) + \frac{\eta}{4} && \text{（公式（20））} \\ &\leq \alpha - \eta + \frac{\eta}{4} && \text{（定理 8.4）} \\ &< c && \text{（}c\text{的定义）。} \end{aligned}$$

  因此，我们有

$$\mathrm{val}(s) \geq \alpha \iff \mathrm{val}_\lambda(s) > c \qquad \text{and} \qquad \mathrm{val}(s) < \alpha \iff \mathrm{val}_\lambda(s) < c\,. \tag{21}$$

- *(存在句)* 我们根据以下三个公式构造该句子。
  - *(常数)* 令 $B_{\max} := \max\{B_\lambda, B_\alpha\}$，并定义关于实变量 λ、c 和 $x = (x_0, \ldots, x_{B_{\max}})$ 的如下公式：

$$\mathrm{Const}(\lambda, c, x) := 2x_0 = 1 \;\wedge\; \bigwedge_{i=0}^{B_{\max}-1} x_{i+1} = x_i^2 \;\wedge\; \lambda = 1 - x_{B_\lambda} \;\wedge\; 2qc = 2p - q x_{B_\alpha}\,.$$

    对 i 归纳可知，前两个合取项成立，当且仅当对于所有 $i \in \{0, \ldots, B_{\max}\}$，$x_i = 2^{-2^i}$。因此，$\mathsf{Const}(\lambda, c, x)$ 成立，当且仅当此外还有 $\lambda = 1 - 2^{-2^{B_\lambda}}$ 和 $c = \alpha - \frac{\eta}{2}$，即 λ 和 c 是前面步骤中的常数。 –
  - *(策略)* 我们定义关于实变量 $\sigma(s')(a)$ 的如下公式，其中 $s' \in S$ 和 $a \in \mathcal{A}$：

$$\mathrm{Strat}(\sigma) := \bigwedge_{s' \in S} \Big(\bigwedge_{a \in \mathcal{A}} \sigma(s')(a) \geq 0 \;\wedge\; \sum_{a \in \mathcal{A}} \sigma(s')(a) = 1 \Big)\,.$$

    当且仅当 σ 是玩家1的平稳策略时，该条件成立。
  - （下界证书）我们定义如下公式，其中 $L(s')$（对于 $s' \in S$）是额外的实变量：

$$\begin{aligned} &\mathrm{LowerCert}(\sigma, L, \lambda, c) := \\ &\quad \bigwedge_{s' \in S} \bigwedge_{b \in \mathcal{B}} L(s') \leq \sum_{a \in \mathcal{A}} \sigma(s')(a) \Big((1-\lambda)\, r(s', a, b) + \lambda \sum_{s'' \in S} \delta(s', a, b)(s'')\, L(s'') \Big) \\ &\quad \wedge\; L(s) > c\,. \end{aligned}$$

    当且仅当 $L \preceq \Psi_{G_\sigma}(L)$ 和 $L(s) > c$ 时，该条件成立。将其不等式乘以 Q 后可得到整数系数，因为所有奖励和转移概率都是$1/Q$的整数倍。

根据公式 (21)和定理 8.2的下界见证，我们有 $\mathsf{val}(s) \geq \alpha$ 当且仅当

$$\exists \sigma\; \exists L\; \exists \lambda\; \exists c\; \exists x\colon\; \mathrm{Const}(\lambda, c, x) \;\wedge\; \mathrm{Strat}(\sigma) \;\wedge\; \mathrm{LowerCert}(\sigma, L, \lambda, c)\,, \tag{22}$$

其中，$\exists \sigma$、$\exists L$ 和 $\exists x$ 对实变量 $\sigma(s')(a)$、$L(s')$ 以及 x_i进行量化，其中 $s' \in S$、$a \in \mathsf{A}$ 和 $i \in \{0, \ldots, B_{\max}\}$。由于 $B_{\max}$ 关于输入规模是多项式规模，公式 (22) 是一个具有多项式规模、可在多项式时间内计算的实数上的存在性语句，因此 CsgDecision 问题属于 $\exists\mathbb{R}$。

- *(全称语句)* 对称地，令 Strat(τ) 表示由 Strat(σ) 得到的公式：将 σ 和 $\mathcal{A}$ 分别替换为 τ 和 $\mathcal{B}$，并定义

$$\begin{aligned}&\mathrm{UpperCert}(\tau, U, \lambda, c) \coloneqq \\ &\quad \bigwedge_{s' \in S} \bigwedge_{a \in \mathcal{A}} U(s') \geq \sum_{b \in \mathcal{B}} \tau(s')(b) \Big((1-\lambda)\, r(s', a, b) + \lambda \sum_{s'' \in S} \delta(s', a, b)(s'')\, U(s'') \Big) \\ &\quad \wedge\ U(s) < c\,,\end{aligned}$$

当且仅当 $\Psi_{G_\tau}(U) \preceq U$ 且 $U(s) < c$时成立。根据公式 (21) 和定理 8.2 的上界见证，我们有 val(s) $< \alpha$ 当且仅当 $\exists \tau\ \exists U\ \exists \lambda\ \exists c\ \exists x$: Const($\lambda, c, x$) $\wedge$Strat(τ) $\wedge$UpperCert(τ, U, λ, c)。因此，我们有 val(s) $\geq \alpha$ 当且仅当

$$\forall \tau\ \forall U\ \forall \lambda\ \forall c\ \forall x\colon\ \neg\big(\mathrm{Const}(\lambda, c, x)\ \wedge\ \mathrm{Strat}(\tau)\ \wedge\ \mathrm{UpperCert}(\tau, U, \lambda, c)\big)\,. \tag{23}$$

与之前一样，公式 (23) 是一个关于实数的全称语句，其规模为多项式，并且可在多项式时间内计算，因此 CsgDecision 问题属于 co-$\exists\mathbb{R}$，由此完成证明。

□

备注 8.5（推论）。 两篇非常新的预印本声称 $\exists\mathbb{R}$ 包含于计数层级 [Mei26、Ope26]。如果这些说法成立，那么我们的结果意味着，对于极限平均和可达性目标，CsgDecision 问题属于计数层级。

9 扩展

在本节中，我们讨论结果的扩展。

奇偶目标。奇偶目标由一个优先级函数给出 $p\colon\ S \to \{0, 1, \ldots\ldots, d\}$；若无限次出现的最小优先级为偶数，则玩家1赢得该对局。所有 ω-正则目标都可以通过确定性奇偶自动机表示，这类目标源自系统的形式化分析；LTL（线性时序逻辑）等时序逻辑也可以转换为确定性奇偶自动机。将博弈图与确定性奇偶自动机作乘积，得到一个具有奇偶目标的博弈，因此奇偶目标是ω-正则目标的规范表示。奇偶目标的值是状态折扣目标的值的迭代极限，其中奖励在具有偶数优先级的状态下为 1 ，在其他状态下为0 ；每个优先级 i 都与一个折扣因子 λ_i相关联，并且折扣因子按照优先级顺序趋近于 1 [DAHM03，GZ05]。此外，奇偶目标的值可以近似到

ε 通过这种带有简洁折扣因子的状态折扣目标的值 $\lambda_i = 1 - 2^{-2^{B_{\lambda_i}}}$，其中每个 $B_{\lambda_i} \in \mathbb{N}_0$ 都是输入规模的多项式，因为 [ACSS24的定理 18]适用于所有足够接近 1，且其与 1 的距离随优先级提高而足够快地减小的折扣因子。我们的定理 3.2 的证明可推广到这类目标，其中马尔可夫决策过程的贝尔曼算子 P 使用当前状态优先级对应的折扣因子，即对于所有函数 $\rho\colon\ S \to \mathbb{R}$ 和状态 s，

$$\Psi_P(\rho)(s) = \operatorname*{op}_{a \in \mathcal{A}} \Big((1 - \lambda_{p(s)})\, r(s, a) + \lambda_{p(s)} \sum_{s' \in S} \delta(s, a)(s')\, \rho(s') \Big)\,.$$

首先，折扣因子是含两项的幂级数数，而且这个贝尔曼算子具有单调性和压缩性，因此，收缩区间界包含该值，并且与之前一样，可在多项式时间内验证。其次，多项式规模的 ε-最优策略的存在性可仿照

第 6.1 节中的证明，其中我们保留的最小概率取决于 $\min_i(1-\lambda_i)$，而非 $1-\lambda$；相对连续性适用，因为有状态折扣目标也是博弈中的可达性目标，并且该博弈在每个状态 s以概率 $1-\lambda_{p(s)}$ 终止。第三，多项式规模的ε-近似收缩区间界的存在性可仿照第 6.3 节中的证明。事实上，贝尔曼算子在每个状态仍然取奖励与后继状态的值的凸组合，因此，詹森不等式推出定理 6.8，而克拉默法则和哈达玛不等式在以 $\max_i B_{\lambda_i}$ 替换 B_λ 后推出定理6.9。因此，对于奇偶目标，CsgApproxValue 问题属于 TFNP，而 CsgGapValue 问题属于 NP ∩ coNP，这改进了 [ACSS24] 给出的 TFNP[NP] 上界。

受限策略下的部分信息博弈。 虽然我们的工作聚焦于完全信息并发随机博弈，在这类博弈中，策略能够观察历史中的状态，但我们的结果也适用于部分信息博弈的特殊情形。对于部分信息博弈，可以通过猜测有界记忆策略的多项式大小表示来近似这些策略所能保证的值；这些策略的概率可能小到双指数级，并通过 NP 预言机 [BIJT24]进行验证。一般而言，NP 预言机不可避免，因为即使对于无记忆策略，近似这些值也已是 NP-难和 coNP-难的 [BIJT24]。然而，如果对手能够同时观察状态和猜测策略的记忆，那么固定该策略后，就会得到一个定义在状态与记忆的乘积空间上的马尔可夫决策过程。我们首先用一个具有简洁折扣因子的折扣和目标的值来近似这个马尔可夫决策过程的极限平均值（定理 7.2），其中 B_λ 是多项式，因为这个马尔可夫决策过程的转移概率具有公分母 $Q \le 2^{B_Q}$，其 $\log B_Q$ 为多项式，正如 定理 5.8的证明中所示。然后，我们的 ε-近似收缩区间界可在多项式时间内验证猜测的策略。

平稳纳什均衡。尽管我们关注的是零和博弈，但我们的方法同样适用于多人轮流随机博弈中的近似约束平稳纳什均衡问题，其中带有终端奖励，且永远未到达终端顶点的对局收益为零 [ABCT25]。这是一个承诺问题：如果存在满足给定收益约束的平稳纳什均衡，那么目标是计算一个满足约束且误差不超过 ε的-Nash均衡 ε。在这一承诺下，存在一个策略具有双指数耐心度 [ABCT25]的此类均衡。固定除某一玩家外所有玩家的策略后，会得到一个具有双指数级小转移概率的马尔可夫决策过程。如定理 7.1所示，其中的终端奖励目标是极限平均目标；我们首先使用简洁折扣因子将其近似为折扣和目标（定理 7.2），这与部分信息博弈中的做法相同。对于所得到的折扣马尔可夫决策过程，以及通过固定所有策略得到的马尔可夫链，我们的 ε-接近收缩区间界可以在多项式时间内证明：在考虑折扣因子的近似误差后，通过检查其最佳响应收益的上界超过策略组合下收益的下界不超过 ε，可知该玩家无法将其收益提高超过 ε。相同的界还可验证收益约束在 ε以内得到满足，并取代了 [ABCT25]中用于此验证的 NP 预言机。

总之，在若干场景中会出现具有双指数级小转移概率的马尔可夫决策过程，而它们可由具有简洁折扣因子的马尔可夫决策过程表示。以往的方法借助 NP 预言机验证此类马尔可夫决策过程的见证，而我们的 ε 近似收缩区间界则可在多项式时间内完成验证。

10 结论

我们研究并发随机博弈中带有极限平均目标的价值近似这一基本问题，并将其复杂度从 TFNP[NP]降低至 TFNP。作为一个特例，可达性目标也有同样的改进。我们的方法通过折扣和来近似极限平均

值，其中折扣因子与 1 双指数级接近，并以幂级数的形式简洁表示。我们的方法还通过有状态折扣目标扩展到奇偶目标 [DAHM03，ACSS24]（第 9 节）。我们还证明，极限平均目标和可达性目标的价值判定问题属于 $\exists\mathbb{R} \cap \text{co-}\exists\mathbb{R}$，并且是 PosSLP-困难的。

未来研究中仍有许多值得进一步探索的开放问题。对于近似问题，将 TFNP 上界改进至多项式时间仍是一项长期存在的重要开放问题，即使对于轮流制随机博弈也是如此。对于价值判定问题，$\exists\mathbb{R}$ 最近有人声称其属于计数层级的较低层级 [Mei26，Ope26]，如果这些说法成立，那么我们的 $\exists\mathbb{R} \cap \text{co-}\exists\mathbb{R}$ 上界便将并发随机博弈的价值判定问题置于计数层级的较低层级。此外，我们证明的 PosSLP 困难性使上界与下界彼此接近，而为该问题建立紧确的复杂度界仍是一个值得研究的开放问题。

致谢

我们感谢 Raimundo Saona 和 Jakub Svoboda 在本研究早期阶段进行的有益讨论。本研究部分由奥地利科学基金（FWF）10.55776/COE12 和 ERC CoG 863818（ForM-SMArt）项目资助。

人工智能方法

本项目始于数年前，其主要思想早于我们开始使用人工智能工具。具体而言，幂级数表示、由平稳策略和与 ε 相近的收缩区间界构成的见证，以及由此得到的多项式时间验证器，均为我们原创。我们使用人工智能工具的情况如下：(a) 对于与 ε 相近的收缩区间相关的计算，我们使用了 GPT-5.6（OpenAI）；(b) 为编辑论文文本，我们使用了 Claude Opus 5.5（Anthropic）和 GPT-6 Astra；(c) 为检查证明，我们使用了 GPT-6 Astra。作者已核实所有内容（包括参考文献）的正确性和原创性，并对本文承担全部责任。

参考文献

[ABCT25] Ali Asadi、Léonard Brice、Krishnendu Chatterjee 和 K. S. Thejaswini。多玩家随机图博弈中的 ε-平稳纳什均衡。载于 *FSTTCS*，*LIPIcs* 第 360 卷，第 9:1–9:17 页，2025 年。完整版本：arXiv:2508.15356。[ABdMS21] David Auger、Xavier Badin de Montjoye 和 Yann Strozecki。简单随机博弈的通用策略改进方法。载于 *MFCS*，*LIPIcs* 第 202 卷，第 12:1–12:22 页，2021 年。[ABKPM09] Eric Allender、Peter Bürgisser、Johan Kjeldgaard-Pedersen 和 Peter Bro Miltersen。关于数值分析的复杂度。*SIAM*计算杂志，38(5):1987–2006，2009 年。[ACSS24] Ali Asadi、Krishnendu Chatterjee、Raimundo Saona 和 Jakub Svoboda。具有有状态折扣目标与奇偶目标的并发随机博弈：复杂度与算法。载于 *FSTTCS*，*LIPIcs* 第 323 卷，第 5:1–5:17 页，2024 年。完整版本：arXiv:2405.02486。[AHK02] Rajeev Alur、Thomas A Henzinger 和 Orna Kupferman。交替时间时序逻辑。《*ACM*杂志》，49(5):672–713，2002 年。

[BF68] David Blackwell 和 Thomas S. Ferguson。大匹配。《数理统计年鉴》，39(1):159–163，1968 年。[BFG+25] Eleni Batziou、John Fearnley、Spencer Gordon、Ruta Mehta 和 Rahul Savani。单调压缩映射。载于 STOC，第 507–517 页，2025 年。[BIJT24] Sougata Bose、Rasmus Ibsen-Jensen 和 Patrick Totzke。部分信息博弈中的有界记忆策略。载于 LICS，第 1–14 页，2024 年。[Can88] John Canny。PSPACE 中的一些代数与几何计算。载于 STOC，第 460–467 页，1988 年。[CDAH06] Krishnendu Chatterjee、Luca De Alfaro 和 Thomas A. Henzinger。定量并发奇偶博弈的复杂度。载于 SODA，第 678–687 页，2006 年。[CDGH15] Krishnendu Chatterjee、Laurent Doyen、Hugo Gimbert 和 Thomas A. Henzinger。免费的随机性。《信息与计算》，245:3–16，2015 年。[Cha07] Krishnendu Chatterjee。随机 ω-正则博弈。博士论文，加州大学伯克利分校，2007 年。技术报告 UCB/EECS-2007-122。[CHI17] Krishnendu Chatterjee、Kristoffer Arnsfelt Hansen 和 Rasmus Ibsen-Jensen。并发安全博弈的策略复杂度。载于 MFCS，第 55:1–55:13 页，2017 年。[CIJ14] Krishnendu Chatterjee 和 Rasmus Ibsen-Jensen。遍历平均收益博弈的复杂度。载于 ICALP，第 122–133 页，2014 年。[CIJ15] Krishnendu Chatterjee 和 Rasmus Ibsen-Jensen。并发平均收益博弈中有限记忆策略下的值为 1 问题。载于 SODA，第 1018–1029 页，2015 年。[CMH08] Krishnendu Chatterjee、Rupak Majumdar 和 Thomas A. Henzinger。随机极限平均博弈属于 EXPTIME。《国际博弈论杂志》，37(2):219–234，2008 年。[CMJ04] Krishnendu Chatterjee、Rupak Majumdar 和 Marcin Jurdziński。随机博弈中的纳什均衡。载于 CSL，第 26–40 页，2004 年。[Con92] Anne Condon。随机博弈的复杂度。《信息与计算》，96(2):203–224，1992 年。[CQS+25] Krishnendu Chatterjee、Tim Quatmann、Maximilian Schäffeler、Maximilian Weininger、Tobias Winkler 和 Daniel Zilken。马尔可夫决策过程中可达性与期望奖励的不动点证书。载于 TACAS，第 130–151 页，2025 年。[DAH00] L. De Alfaro 和 T.A. Henzinger。并发 ω-正则博弈。载于 LICS，第 141–154 页，2000 年。[DAHK98] L. De Alfaro、T.A. Henzinger 和 O. Kupferman。并发可达性博弈。载于 FOCS，第 564–575 页，1998 年。[DAHM00] Luca De Alfaro、Thomas A Henzinger 和 Freddy YC Mang。同步系统的控制。载于 CONCUR，第 458–473 页，2000 年。

[DAHM01] Luca De Alfaro、Thomas A Henzinger 和 Freddy YC Mang。同步系统的控制，第二部分。载于 CONCUR，第 566–581 页，2001 年。[DAHM03] Luca De Alfaro、Thomas A. Henzinger 和 Rupak Majumdar。系统理论中的未来折扣。载于 ICALP，LNCS 第 2719 卷，第 1022–1037 页，2003 年。[DAM01] Luca De Alfaro 和 Rupak Majumdar。ω-正则博弈的定量求解。载于 STOC，第 675–683 页，2001 年。[Eve57] Hugh Everett。递归博弈。载于《博弈理论贡献 III》，第 47–78 页，1957 年。[EY08] Kousha Etessami 和 Mihalis Yannakakis。递归并发随机博弈。《计算机科学逻辑方法》，4(4):7，2008 年。[EY09] Kousha Etessami 和 Mihalis Yannakakis。递归马尔可夫链、随机文法与非线性方程单调系统。ACM 学报，56(1):1:1–1:66，2009 年。[FM13] Søren Kristoffer Stiil Frederiksen 和 Peter Bro Miltersen。在多项式时间层次中近似并发可达性博弈的值。载于 ISAAC，第 457–467 页，2013 年。[FV97] Jerzy Filar 和 Koos Vrieze。竞争性马尔可夫决策过程。Springer，1997 年。[GGJ76] M. R. Garey、Ronald L. Graham 和 David S. Johnson。一些 NP 完全几何问题。载于 STOC，第 10–22 页，1976 年。[Gil57] Dean Gillette。具有零停止概率的随机博弈。载于《博弈理论贡献 III》，第 179–187 页，1957 年。[GKW25] Marta Grobelna、Jan Křetínský 和 Maximilian Weininger。并发随机可达性博弈和安全博弈中值迭代的停止准则。载于 LICS，第 568–580 页，2025 年。完整版本：arXiv:2505.21087。[GZ05] Hugo Gimbert 和 Wiesław Zielonka。如何以及为何对无限博弈进行折扣？《理论计算机科学电子札记》，119(1):3–9，2005 年。[HK15] Christoph Haase 和 Stefan Kiefer。保持预算不超支的概率。载于 ICALP，第 234–246 页，2015 年。[HK20] Arnd Hartmanns 和 Benjamin Lucien Kaminski。乐观值迭代。载于 CAV，第 12225 卷，第 488–511 页，2020 年。[HKL+11] Kristoffer Arnsfelt Hansen、Michal Koucký、Niels Lauritzen、Peter Bro Miltersen 和 Elias P. Tsigaridas。求解随机博弈的精确算法：扩展摘要。载于 STOC，第 205–214 页，2011 年。完整版本：arXiv:1202.3898。[HKM09] K. A. Hansen、M. Koucky 和 P. B. Miltersen。赢得并发可达性博弈需要双指数耐心度。载于 LICS，第 332–341 页，2009 年。[HM18] Serge Haddad 和 Benjamin Monmege。用于马尔可夫决策过程和区间马尔可夫决策过程的区间迭代算法。《理论计算机科学》，735:111–131，2018 年。

[HPRV76] C. J. Himmelberg、T. Parthasarathy、T. E. S. Raghavan 和 F. S. Van Vleck。随机博弈中 p-均衡与最优平稳策略的存在性。《美国数学会会刊》，60(1):245–251，1976 年。[IJ26] Rasmus Ibsen-Jensen。并发博弈。载于 Nathanaël Fijalkow（编），《图上的博弈：从逻辑和自动机到算法》，第 8 章，第 273–297 页。剑桥大学出版社，2026 年。arXiv:2305.10546。[Mar98] Donald A. Martin。Blackwell 博弈的确定性。《符号逻辑杂志》，63(4):1565–1581，1998 年。[Mei26] Alex Meiburg。 $\exists\mathbb{R} \subseteq \mathrm{CH}$。arXiv:2610.10514，2026 年。[MN81] Jean-François Mertens 和 Abraham Neyman。随机博弈。《国际博弈论杂志》，10(2):53–66，1981 年。[MP91] Nimrod Megiddo 和 Christos H. Papadimitriou。关于全函数、存在性定理与计算复杂性。《理论计算机科学》，81(2):317–324，1991 年。[OB21] Miquel Oliu-Barton。求解零和随机博弈的新算法。《数学运筹研究》，46(1):255–267，2021 年。[Ope26] OpenAI。计数层级中的存在–全称实数句。OpenAI 数学发布预印本，2026年10月4日版本。网址：https://github.com/openai/math/blob/adc7f1241b42e322a6451854ab7e4b4c146bf78a/preprints/Existential-universal-real-sentences-in-the-counting-hierarchy-October-4-2026/etr-counting-hierarchy.pdf。[Par71] T. Parthasarathy。折扣型与正随机博弈。《美国数学会通报》，77(1):134–136，1971 年。[Sch10] Marcus Schaefer。一些几何与拓扑问题的复杂性。载于 GD 2009，第 334–344 页，2010 年。[Sha53] Lloyd S Shapley。随机博弈。《美国国家科学院院刊》，39(10):1095–1100，1953 年。[Sol03] Eilon Solan。竞争性马尔可夫决策过程的值的连续性。《理论概率杂志》，16(4):831–845，2003 年。[SŠ17] Marcus Schaefer 和 Daniel Štefankovič。不动点、纳什均衡与实数存在理论。《计算系统理论》，60(2):172–193，2017 年。