arXiv:2302.12940v1 [cs.LG] 25 Feb 2023

线性函数逼近强化学习的指数级困难性

Daniel Kane*加利福尼亚大学圣迭戈分校 dakane@eng.ucsd.edu

Sihan Liu 加利福尼亚大学圣迭戈分校 sil046@ucsd.edu

Shachar Lovett†加利福尼亚大学圣迭戈分校 slovett@cs.ucsd.edu

Gaurav Mahajan 加利福尼亚大学圣迭戈分校 gmahajan@eng.ucsd.edu

Csaba Szepesv´ari‡DeepMind，英国伦敦 阿尔伯塔大学，加拿大埃德蒙顿 szepesva@ualberta.ca

Gell´ert Weisz DeepMind，英国伦敦 伦敦大学学院，英国伦敦 gellert@deepmind.com

2023年2月28日

摘要

强化学习理论中的一个基本问题是：假设最优价值函数关于给定特征是线性的，我们能否高效地学习它们？该问题在监督学习中的对应问题——线性回归——在统计和计算方面都可以高效求解。因此，近期一项研究[KLLM22] 揭示了线性强化学习中的计算-统计鸿沟：尽管存在样本复杂度为多项式的算法，但除非 NP =RP，否则这一设定不存在多项式时间算法。

在这项工作中，我们基于他们的结果，在随机指数时间假设下，为线性函数近似的强化学习证明了一个计算下界，该下界关于特征维度和时域呈指数增长。为证明这一点，我们构造了一个分轮游戏，在每一轮中，学习者都要在单位超立方体中搜索一个未知向量。我们对该游戏中的奖励作了特殊设计，使得如果学习者获得了较高奖励，那么学习者的动作就可以用来模拟求解 3-SAT 的一个变体。在这个变体中：(a) 每个变量出现在数量有界的子句中；(b) 如果某个实例无解，那么也不存在满足超过 (1-ϵ) 比例子句的解。我们利用标准归约证明，这个 3-SAT 变体的近似求解难度与 3-SAT 一样高。最后，我们还证明了一个针对时域依赖性优化的下界，它几乎达到了目前已知的 $\exp(\sqrt{H})$ 上界。

*本研究由美国国家科学基金会 CCF-1553288 奖项（CAREER）和斯隆研究奖学金资助。

†本研究由美国国家科学基金会 DMS-1953928 和 CCF-2006443 奖项资助。

‡由加拿大自然科学与工程研究委员会、阿尔伯塔机器智能研究所和加拿大人工智能研究主席计划资助。

目录

1 引言 **3**

2 我们的贡献 **3**

2.1 预备知识.2.2 LINEAR-3-RL 的指数下界. 4
5

3 证明概述 **6**

4 下界构造 **9**

4.1 从 3-CNF 公式到三动作马尔可夫决策过程.94.2 线性价值函数.114.3 从强化学习算法到SAT算法.16

带间隙和少量子句的近似 SAT 难度**22**B 上界**24**

1 引言

在状态空间较大的环境中进行高效探索和规划，是强化学习中的一个核心问题。近年来，将函数逼近应用于经典强化学习算法取得了许多成功，在各种实际应用中带来了最先进的成果。

这也促使强化学习（RL）理论界越来越关注如何为大状态空间情形设计和分析高效算法。在这种情形下，目标是设计复杂度不以状态空间大小呈多项式依赖的算法。由于在不对环境作任何假设的情况下这是不可能实现的，因此人们投入了大量精力来寻找最小化假设，使得能够高效地找到最优策略：状态聚合[Li09、DRZ20], 线性 q^π [DKWY20、LSW20、YHAY+22、WGKS22], 线性马尔可夫决策过程 [YW19、JYWJ20],线性混合马尔可夫决策过程 [MJTS20、AJS+20、ZHG21], 反应式部分可观测马尔可夫决策过程 [KAL16], 块状马尔可夫决策过程 [DKJ+19]、FLAMBE[AKKS20], 反应式预测状态表示 [LSS01], 线性贝尔曼完备性 [Mun05、ZLKB20], 贝尔曼秩[JKA+16]、见证秩 [SJK+19]、双线性函数类 [DKL+21]、贝尔曼 Eluder 维度 [JLM21]以及决策-估计系数[FKQR21]。

这条研究脉络中提出的一项最小假设是采用线性函数逼近的强化学习：最优价值函数（ Q^*、V^*，或两者）可以表示为有限个已知基函数的线性组合。当两个最优价值函数 Q^* 和 V^* 都满足这一假设（以下称为线性 Q^*&V^* ）时，文献中有两种样本高效算法，其样本复杂度关于基函数数量 d 和时域 H呈多项式关系。首先， [DKL+21] 提出的算法还假设整个状态-动作空间上的基函数取值已知且经过预处理。其次，TensorPlan [WAJ+21、WSG22]则将这一假设替换为一个隐含假设，即动作数量是较小的常数（因为其样本复杂度关于动作数量呈指数增长）。[WAS21、WSG22] 给出了关于 $\min(d, H)$ 呈指数增长的样本复杂度下界，这表明，当动作数量关于 d 呈多项式增长，且基函数取值仅在采样状态上揭示时，寻找近最优策略在统计上是困难的。这说明，要实现样本高效算法，前述两项额外假设至少有一项是必需的。然而，即使满足这两项额外假设，这些工作仍将如何为这一问题设置找到计算高效算法留作一个重要的开放问题。

近期的一项工作 [KLLM22] 在这一问题上取得了进展，证明了在线性函数逼近的强化学习中存在计算-统计鸿沟：除非 NP=RP，否则即使在最简单的设定下，也不存在多项式时间算法；该设定采用线性 Q^*&V^*、确定性转移、随机奖励和 2 个动作。这一结果令人意外，因为如果进一步假设奖励是确定性的，那么该问题可以在$O(dH)$时间内求解 [WVR17]。因此， [KLLM22] 的结果表明，奖励中加入噪声可能导致计算不可解性（若动作数量不受限，样本复杂度也会出现类似的转变 [WAS21, WSG22]）。不过， [KLLM22] 给出的下界并不紧：他们证明了一个关于 d 的准多项式下界，而目前已知的最佳上界则是关于 $\min(d, H)$ 的指数级上界[DLMW20]。

2 我们的贡献

在这项工作中，我们为采用线性函数近似的强化学习给出了几乎匹配的指数级计算下界。在陈述主要结果之前，我们先介绍本文中始终使用的一些关键定义。

2.1 预备知识

马尔可夫决策过程（MDP）。我们首先定义强化学习的框架——马尔可夫决策过程（MDP）。我们将确定性转移马尔可夫决策过程定义为一个元组 $M = (\mathcal{S}, \mathcal{A}, R, P)$，其中 $\mathcal{S}$ 是状态空间，$\mathcal{A}$ 是动作空间，$R\colon \mathcal{S} \times \mathcal{A} \mapsto \Delta([0,1])$ 是随机奖励函数，[1]而 $P\colon \mathcal{S} \times \mathcal{A} \mapsto \mathcal{S}$ 是确定性转移函数。这样的 MDP M 会形成一个离散时间的序贯决策过程：智能体从起始状态 $S_0 \in \mathcal{S}$出发。随后，在每个时刻 t，处于某个当前状态 S_t的智能体采取动作 A_t，获得奖励 $R_t \sim R(S_t, A_t)$，并转移到下一状态 $S_{t+1} = P(S_t, A_t)$。这一过程持续进行，直至智能体到达终止状态 $\perp \in \mathcal{S}$。在 H-时域问题中，从起始状态 s_0 到终止状态 $\perp$的每条此类轨迹/路径长度至多为 H，并且在 $0 \le t \le H$条件下，经过 t 步（采取任意动作）后可到达的状态集合 $\mathcal{S}_t$ 彼此不相交。决策智能体的目标是最大化沿此类轨迹获得的期望奖励总和。事实证明，无论初始状态如何，都可以通过遵循某个确定性的平稳策略来获得期望奖励总和；该策略由映射 $\pi\colon \mathcal{S} \mapsto \mathcal{A}$给出，遵循 π 意味着：在第t 步，如果状态为 S_t，则采取的动作是 $A_t = \pi(S_t)$。给定策略 π 和状态-动作对 $(s,a) \in \mathcal{S} \times \mathcal{A}$，我们令

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\tau-1} R(S_t, A_t) \mid S_0 = s, \pi\right], \quad Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\tau-1} R(S_t, A_t) \mid S_0 = s, A_0 = a, \pi\right]$$

表示总期望奖励，其中 S_1、A_1、$\ldots S_{\tau-1}$、$A_{\tau-1}$ 是通过在马尔可夫决策过程 M 中执行策略 π 得到的，而 τ 是策略 π 到达终止状态 $\perp$的首次时刻，即 $S_\tau = \perp$，其中始终有 $\tau \le H$。我们用 Q^* 和 V^* 表示最优价值函数

$$V^*(s) = \sup_\pi V^\pi(s), \quad Q^*(s,a) = \sup_\pi Q^\pi(s,a), \quad s \in \mathcal{S}, a \in \mathcal{A}$$

我们称最优价值函数 V^* 和 Q^* 可以表示为关于 d维特征 ψ的线性函数：$\mathcal{S} \sqcup (\mathcal{S} \times \mathcal{A}) \to \mathbb{R}^d$，如果对于所有状态 s和动作 a，都有 $V^*(s) = \langle \theta, \psi(s) \rangle$，且 $Q^*(s,a) = \langle \theta, \psi(s,a) \rangle$，其中某个固定的 $\theta \in \mathbb{R}^d$ 与 s 和 a无关。[2] 在我们的构造中，对于$\psi(s,a) = \psi(P(s,a))$，线性 V^* 意味着线性 Q^*，因为：(i) 在确定性转移马尔可夫决策过程中，$Q^*(s,a) = r(s,a) + V^*(P(s,a))$；(ii) 在我们的构造中，除叶节点外，所有位置的奖励均为 0；(iii) 叶节点处的奖励不依赖于动作。

计算问题。接下来我们介绍 3-SAT，这是一种针对 3-CNF 公式的可满足性问题。在 3-SAT 问题中，输入是一个含有 φ 个变量和 $O(v)$个子句的 3-CNF 公式 v，我们的目标是判断 φ是否可满足。

复杂性问题 3-SAT
输入：一个含有 v 个变量和 $O(v)$个子句的 3-CNF 公式 φ 目标：
判断该公式是否可满足。

本文聚焦于计算强化学习问题 LINEAR-k-RL。在特征维度为 d 的 LINEAR-k-RL 问题中，我们可以访问一个确定性马尔可夫决策过程 M，其动作数量为 k，时域为$H = O(d)$，且最优价值函数 Q^* 和 V^* 均可表示为 d 维特征 ψ 的线性函数。我们的目标是输出一个良好的策略，我们将其定义为满足$V^\pi > V^* - 1/8$的任意策略 π，其中 V^π 和 V^* 分别指策略 π 和最优策略在固定起始状态下的价值，且始终位于 $[0, H]$ [3]。从现在起，我们始终假设动作数量为

[1] $\Delta([0,1])$ 表示区间 $[0,1]$ 上所有分布的集合。[2]在上文中，$\sqcup$ 表示对各参数取不交并。[3]在我们的构造中，我们满足更严格的条件，即 $V^* \in [0,1]$。

$k = 3$。

复杂性问题	LINEAR-k-RL
预言机：	一个确定性马尔可夫决策过程 M，具有 k 个动作，其最优价值函数 V^* 和 Q^* 在线性于 d 维度特征 ψ、时域为 H，且状态空间大小至多为 exp(poly(d))。
目标：	寻找策略 π 使得 $V^\pi > V^* - 1/8$。

现在我们描述算法如何与马尔可夫决策过程交互。我们假设算法可以访问状态空间和动作空间（它们可视为整数的子集），并能随机访问相关的 (i) 奖励函数 R、(ii) 转移函数 P 和 (iii) 特征 ψ。对于上述所有函数，算法提供一个状态 s 和动作 a（如有需要），并从分布$R(s, a)$中接收一个随机样本（针对奖励函数）、状态 $P(s, a)$（针对转移函数）、特征 $\psi(s)$ 以及 $\psi(s, a)$（针对特征）。我们假设每次调用的运行时间均为常数，且这些函数的输入和输出大小均为特征维度 d 的多项式。

我们通常会讨论以时间 t 解决某个问题、错误概率为 p的随机化算法 A。这意味着：(i) A 的运行时间为 $O(t)$；(ii) 对于可满足性问题，它以至少 $1 - p$ 的概率在正输入实例上返回 YES，并在负输入实例上以概率 1 返回 NO；(iii) 对于强化学习问题，它以至少 $1 - p$的概率返回一个良好的策略。

2.2 LINEAR-3-RL 的指数下界

本文在强化版的 NP $\neq$ RP 猜想——随机化指数时间假设（rETH） [DHM+14]的基础上，给出了计算下界。该假设断言，概率算法无法在次指数时间内判定一个给定的、含有 v 个变量和 $O(v)$ 个子句的 3-SAT 问题是否可满足。

定义 2.1（随机化指数时间假设（rETH））。存在一个常数 $c > 0$，使得不存在随机化算法能够在时间 2^{cv} 内以错误概率 $1/3$ 判定含有 v 个变量的 3-SAT。

随机化指数时间假设，以及由指数时间假设启发的诸多变体，[IP01] 在发现各类问题的困难性结果方面发挥了重要作用，参见例如[CFK+15、Wil19]。在随机化指数时间假设下，我们的主要结果是：对于具有线性最优价值函数的确定性马尔可夫决策过程，学习良好策略存在指数级计算下界。

定理 2.2（关于时域和维度的指数级下界）。在 rETH 下，不存在任何随机化算法，能够以至少 $9/10$ 的概率，在 exp($\tilde{O}$(min($d^{1/4}, H^{1/4}$))) 时间内求解特征维度为 d、时域为 H 的 LINEAR-3-RL，其中 $\tilde{O}$ 隐含着 polylog(d) 和 polylog(H) 因子。

有几点说明。首先，$\min(\cdot)$ 是这里正确的复杂度度量。为说明这一点，我们注意到，该问题可以在 $\exp(\tilde{O}(\min(d, \sqrt{H})))$ 时间内求解（我们在附录 B 中证明了这些上界），因此，如果维度 d 或时域 H 为常数，我们就可以针对另一个参数高效地求解该问题。其次，这是该设定下首个指数级计算下界；此前已知的最佳结果 [KLLM22] 即使假设 rETH 成立，也至多给出准多项式下界。

就时域 H 而言，定理 2.2 给出了 $\exp(\tilde{O}(H^{1/4}))$ 下界，以及$\exp(\tilde{O}(\sqrt{H}))$ 上界。接下来，我们将证明一个针对时域 H 优化的下界，它几乎匹配该上界。

定理 2.3（近乎匹配的时域下界）。在 rETH 假设下，不存在能够解决时域为 H、特征维度为 $d \geq H^{\log H}$ 的线性-3-强化学习问题的随机化算法，使其运行时间为 exp($\tilde{O}(\sqrt{H})$)，且以至少 $9/10$ 的概率成功，其中 O 隐藏了多对数(H)因子。

现在我们讨论一些开放问题。尽管定理 2.3 中的下界在时域方面几乎匹配上界 H，但它要求特征维度至少达到关于 H的准多项式量级。我们将上述结果在 $d = \text{poly}(H)$ 时是否仍成立作为一个开放问题。另一个重要方向是理解关于维度 d 的复杂度，即针对维度 d优化的下界。我们对定理 2.2的证明可以修改为得到 $\exp(d)$ 下界，适用于 $H = \exp(d)$。这一结果对 $H = \text{poly}(d)$是否也成立?

相关工作。 我们已经讨论了大量在各种假设下为强化学习提供统计高效算法的工作。与之互补的是，研究线性函数逼近下强化学习统计下界的工作，其中动作数量不断增加。具体而言， [WAS21、WSG22、WWK21] 的研究证明了样本复杂度下界，其增长速度关于 $\min(d, H)$ 呈指数级；当动作数量随基函数数量增长，且基函数的取值仅在采样状态上揭示时，这些下界意味着寻找近最优策略在统计上具有困难性。此外，近期有研究[GMR22a、GMR22b、USL+22] 致力于设计准多项式时间的端到端算法，以在“可观测”的部分可观测马尔可夫决策过程中进行学习（我们的下界结果否定了在线性 Q^* 及 V^* 假设下存在类似准多项式时间算法的可能性。）

本文其余部分。在第3节，我们简要概述下界构造中的主要技术思路。在第4节，我们详细介绍指数下界构造，并证明主要定理。在附录A 中，我们使用标准归约证明，在随机化指数时间假设下，归约中使用的间隙版 SAT 在计算上是困难的。在附录B 中，我们给出强化学习算法，其运行时间关于$\min(d, \sqrt{H})$呈指数级，这表明我们的下界接近最优。

3 证明概述

此前由[KLLM22] 提出的下界，其高层次思路如下。作者设计了一个马尔可夫决策过程，迫使学习者在 w^* 中搜索一个未知向量 $\{0,1\}^v$，该向量构成给定 SAT 公式 ϕ的一个满足解。具体而言，马尔可夫决策过程中的每个状态都对应一个赋值；在该状态下，学习者可以翻转公式中第一个不满足子句里出现的一个变量（假设子句按某种规范顺序排列）。当学习者找到满足赋值或到达时域末尾时，就会获得奖励。奖励的设计方式使得：（i）学习者有动力快速找到 w^*；但（ii）学习者无法利用奖励中的太多信息来加速搜索过程。因此，该任务与求解原始 SAT 问题一样困难。

上述方法的一个瓶颈在于，只有当算法进行博弈的次数少于准多项式次数时，奖励才不会提供信息。此后，算法有相当大的概率可以从奖励结构中获取额外信息，而这些信息可能会显著简化任务。我们沿用了将困难的 SAT 实例变体嵌入线性强化学习问题这一总体思路。然而，我们对马尔可夫决策过程的转移和奖励结构进行了重大修改，使得除非算法进行指数次数的博弈，否则几乎无法从奖励中获取任何有用信息。

本质上，我们将时域结束时给出的奖励设为均值呈指数级小的伯努利变量，从而确保这些奖励不包含有用信息。这样一来，学习者高概率只会在

复杂性问题(b, ϵ)- GAP-3-SAT

输入：一个包含 φ 个变量和 O 个子句（v）的 3-CNF 公式 v，并满足以下承诺：(1) 每个变量最多出现在 b 个子句中；(2) 公式 φ 可满足，或者对于任何赋值，至少有 ϵ 比例的子句不满足。目标：判断该公式是否可满足。

除非它多次进行该博弈，否则.....作为热身，可以设想一个马尔可夫决策过程，其动作和转移与[KLLM22]中的相同。不过，我们将奖励修改为$\exp(-$ 迄今为止的步数 $-$距离$(w, w^*))$，并将其设定在任意终止状态 w。这样可以确保在时域结束时给出的（期望）奖励始终呈指数级小。不幸的是，由此产生的价值函数也会具有相同的指数函数形式，因此无法表示为仅依赖于状态的一些低维特征的线性函数。

分轮游戏。一种解决方法是将博弈改为分轮游戏。我们将搜索划分为若干轮，在每一轮中，学习者依次看到各个变量，并决定是否翻转该变量。然后，如果学习者在第 n 轮终止，我们将奖励函数设为大致等于 $\prod_{i=1}^{n} g_i$(第 i 轮中翻转的次数)$\cdot g_{n+1}$(距离(w, w^*))，其中 g_i 是经过精心选择的低次单调递减多项式。现在，考虑一种贪心策略，它会在可能时尽量缩短与 w^* 的距离。由于贪心策略总能在一整轮内到达 w^*，因此，该策略在第 i 轮开始时的价值函数将为 $\prod_{j<i} g_j$(本轮翻转的次数) $* g_i$(距离(w, w^*))。由于只有最后一项依赖于 w 和 w^*，我们可以得出，价值函数本质上是关于 w 和 w^* 的低次多项式，确实可以表示为某些依赖于状态的低维特征向量的线性函数。详情请参见引理 4.5。

然而，为了确保这是最优策略，我们需要非常精确地定义 g_i，使得在当前轮次进行一次翻转，总是比将其推迟到后续轮次更有利。从本质上说，这意味着，对于g_j，其对数导数应小于 g_i 在 $j > i$处的对数导数。理想情况下，我们希望令 $g_i(x) = \exp(-c_i\, x)$，其中 c_i是某个递增序列，这样上述性质便显然成立。然而，由于 g_i 必须是多项式，我们只能将它设为该指数函数在 $x = 0$附近的泰勒近似。只要我们能使此泰勒近似的误差相对于$\exp(-c_i \cdot x)$的对数导数之差足够小，智能体在较早轮次采取额外步骤仍然有利。幸运的是，使用低次数泰勒近似确实可以做到这一点。论证详情见 Claim 4.7和 Lemma 4.8。

翻转强制机制。虽然分轮游戏确实保证了价值函数的线性性，但时域结束时给出的奖励不一定很小。由于 $g_i(x)$ 被视为 $\exp(-c_i \cdot x)$ 在 $x = 0$ 处的泰勒近似，因此 $g_i(0)$ 将为 1。于是，如果学习者选择不进行任何翻转，最终可能会获得巨额奖励，使算法能够从奖励结构中提取信息。

为了防止这种情况，我们将在每轮的第一步向学习者提供一个变量组，使其必须翻转其中一个给定变量。这样做需要注意的是，我们希望至少有一个变量确实是错误的，这样翻转它就能使其真值赋值正确，因此贪心策略仍然是定义良好且最优的。幸运的是，只需提供出现在任一未满足子句中的变量，就能保证这一点。

这使我们能够迫使算法至少进行一次翻转。为了让奖励以更快的速度递减，我们进一步采用这一思路：不断向学习者呈现包含

尚未翻转的变量的子句。只有在这类子句用尽后，我们才开始遍历其余变量，并让学习者选择是否跳过翻转。

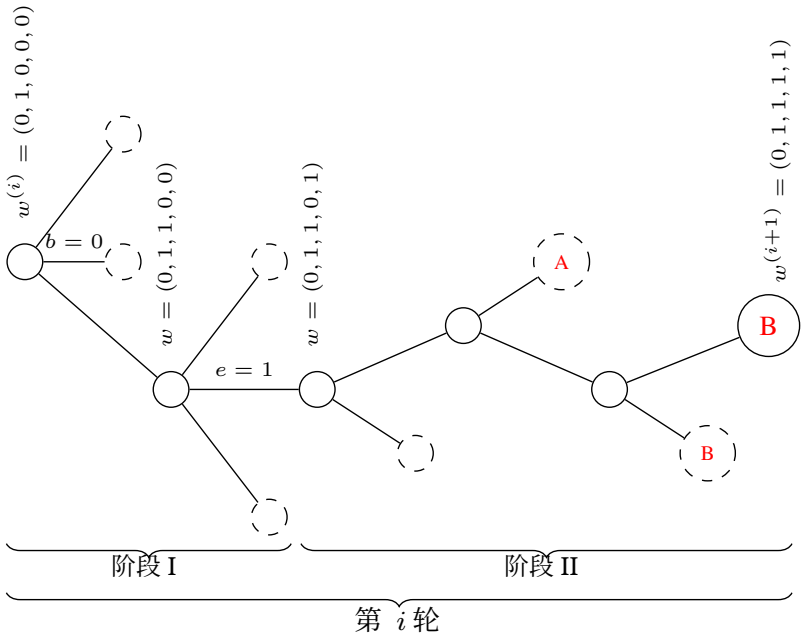

图 1：第 i 轮马尔可夫决策过程（MDP）的机制示例。该 MDP 由 h 个连续轮次组成，此处仅展示第 i 轮。节点表示状态，其赋值 w 会在发生变化时标出；边表示动作，标签代表对某个变量的设定。可满足性问题为 $(a \vee \neg b \vee c) \wedge (c \vee d \vee e) \wedge (a \vee d \vee e) \wedge (a \vee \neg b \vee \neg c) \wedge (a \vee \neg b \vee \neg e)$，变量 a 至 e 在该轮开始时的赋值为 $w^{(i)}$。为便于说明，请注意此问题不属于 (b, ϵ)- GAP-3-SAT。前两个步骤构成阶段 I，因为存在一个仅由自由变量组成的未满足子句。第二阶段允许逐一更改其余任意自由变量。转移是确定性的。除终止条件 A 和 B 外，奖励始终为零；在这两个条件下，奖励服从伯努利分布。A：赋值满足的子句比例至少为 $(1 - \epsilon)$。B：仅当 i 是最后一轮时，博弈才会在该轮结束时终止。

我们希望要求能够找到许多这样的子句。当然，在一般的 SAT 实例中无法保证这一点。不过，我们证明存在一类特殊的 3-SAT 实例，在其中寻找这样的赋值——使得这些未满足子句会很快耗尽——在计算上是困难的。具体而言，我们使用一些标准归约，证明 3-SAT 的近似难度与我们称为 GAP-3-SAT 的问题大致相当；GAP-3-SAT 满足：(a) 每个变量出现在有界数量的子句中；(b) 如果不存在解，那么也不存在能满足占子句比例为 (1-ϵ) 的解。注意，(b) 意味着找到任何满足占子句比例为 (1-ϵ) 的赋值都很困难，而 (a) 则说明，翻转一个变量最多只能让常数个未满足子句不再纳入考虑。具体而言，如果每个变量最多出现在 b 个子句中，那么任何计算高效算法在前 $\epsilon *$（子句总数/b）步内都不会耗尽未满足子句。这确保了时域上的奖励相对于轮次数呈指数级小。

4 下界构造

本节将在 rETH 假设下证明 LINEAR-3-RL 的以下计算下界。

命题 4.1。令 $v \in \mathbb{Z}^+$ 足够大。假设 d、$H \in \mathbb{Z}^+$ 满足以下任一条件

1. $d = v^4 \cdot \text{polylog}(v)$ 且 $H = \Theta(v^4)$，或 2. $d = \exp\left(\log^2 v \cdot \text{polyloglog}(v)\right)$ 且 $H = \Theta(v^2)$。

那么，在 rETH 假设下，不存在随机化算法能够以错误概率 $1/10$，在 $\exp(v/\text{polylog}(v))$ 时间内求解具有特征维度 d、时间范围 H 的 LINEAR-3-RL。

我们的主要定理，定理 2.2 和定理 2.3，可由命题 4.1 通过将 v 表示为 d和 H的函数而得出。

4.1 从 3-CNF公式到三动作马尔可夫决策过程

回顾一下，在 (b, ϵ)-GAP-3-SAT 中，输入是一个关于 v 个变量的 3-CNF 公式 φ，其中：（1）保证每个变量至多出现在 b 个子句中；（2）该公式要么可满足，要么任意赋值至多满足比例为 $(1-\epsilon)$ 的子句（保证该公式属于这两种情况之一）。此外，我们可以假设子句数量至少为 v。根据命题 4.11，当 b、ϵ 被设为两个绝对常数时，在 rETH 假设下，判断 φ 是否可满足所需的时间必须关于 v 呈指数增长。我们的目标是构造一个由 φ 参数化的马尔可夫决策过程，使得为该马尔可夫决策过程学习近最优策略至少与求解 (b, ϵ)-GAP-3-SAT 一样困难。当该公式可满足时，该马尔可夫决策过程还将具有一个额外参数 w^*，其取值为该公式的任意满足赋值。

为整合命题4.1中的两个结果，在我们的归约中，我们构造的马尔可夫决策过程带有两个额外的“度参数” p，$q \in \mathbb{Z}^+$。该马尔可夫决策过程的特征维度为 $d = 2 \cdot v^{2p}$，时域为 $H = \alpha \cdot v^q$，其中 α 是一个足够小的绝对常数，稍后将确定其取值。具体而言，第一个结果中的困难实例通过设置 $p = 2$，$q = 4$ 得到，而第二个结果中的困难实例通过设置 $p = 2\log v$，$q = 2$得到。

状态-动作转移。时间步被划分为 $h := H/v$ 轮，每轮包含v 步。每轮中，马尔可夫决策过程维护一个“已使用变量”集合，并在每轮开始时将其初始化为空集。我们将未使用的变量称为“自由变量”。每轮又进一步分为以下两个阶段：

1. 在第一阶段的每一步中，智能体会得到一个仅包含自由变量的未满足 3-SAT子句，并被要求指出应翻转三个变量中的哪一个。算法选择的变量随后会被加入已使用变量集合。当仅包含自由变量的未满足子句耗尽时，第一阶段结束（第二阶段开始）。

2. 在第二阶段，马尔可夫决策过程会依次向智能体呈现每个剩余的自由变量，并询问是否应翻转该变量。每一步之后，无论算法决定是否翻转该变量，所呈现的变量都会被加入已使用变量集合。

请注意，每轮恰好有 v 步，因为公式中有 v 个变量，并且每一步都会将其中一个标记为已使用。

终止条件。若马尔可夫决策过程到达最后一层，或超过 $(1-\epsilon)$ 比例的子句已满足，则马尔可夫决策过程终止。关于终止条件，我们有几点观察。首先，如果马尔可夫决策过程在到达最后一层之前终止，那么算法实际上已经解决了底层的 GAP-3-SAT问题，因为这意味着存在一个赋值，能够满足超过 $(1-\epsilon)$ 比例的子句。其次，终止条件确保在每轮开始时，至少有 ϵ 比例的未满足子句。由于每个变量最多出现在 b 个子句中，因此在前 $\epsilon \cdot$（子句总数$/b$）步中，我们不会遇到未满足子句中只剩自由变量的情况。

状态空间大小。对于一轮和一个赋值，所需的状态数最多为 O（3^v）：一轮内的转移构成一棵分支因子至多为3 、高度至多为 v 的树（见图1）。共有 $h = H/v$ 轮，整个转移结构是一棵树，因此状态总数至多为O（（3^v）^{h}）$=$ exp（poly（v，H））。

奖励。只有在马尔可夫决策过程终止时才会给出奖励，且奖励值因公式是否可满足而异。若公式不可满足，则所有状态的奖励均为 0。在接下来的讨论中，我们将赋值视为 $\{-1,1\}^v$ 中的向量。若公式可满足，我们需要跟踪每轮开始时的赋值，并将它们记为 $w^{(1)}$、$w^{(2)}\ldots$、$w^{(n)}$（注意，算法从赋值为 $w^{(1)}$ 的第一个状态开始），最终奖励取决于这些赋值。奖励仅取决于历史 $w^{(1)}$、$w^{(2)}\ldots$、$w^{(n)}$、当前赋值 w 和最优赋值 w^*，其定义为 $Ber(r(w^{(1)}, w^{(2)}\ldots, w^{(n)}, w))$，其中 $r(\cdot)$ 是期望奖励函数，伯努利分布 $Ber(\rho)$ 以 ρ 的概率取值为 1，以 $1-\rho$ 的概率取值为 0。在具体说明期望奖励函数 $r(\cdot)$ 之前，我们先介绍扩展赋值的概念。

定义4.2（扩展赋值）。设 S 为自由变量集合。那么，w 在 S下的扩展赋值，记作$\text{ext}(w, S)$，定义为：当 $i \in S$时，$\text{ext}(w,S)_i = w^*_i$；当 $i \notin S$时，$\text{ext}(w,S)_i = w_i$。

通俗地说，扩展赋值是从 w 导出的赋值，其中所有自由变量均被校正为与 w^*一致。我们注意到，奖励函数对扩展赋值的依赖至关重要，它确保了与贪心策略相关联的价值函数是线性函数，这一点将在后文变得重要。现在，我们准备根据智能体在过去轮次结束时达到的历史赋值 $w^{(1)}$、$\cdots$、$w^{(n)}$，智能体达到的当前赋值 w，以及马尔可夫决策过程终止时的自由变量集合 S，来定义期望奖励函数 $r(\cdot)$。

定义4.3（期望奖励）。设 p、$q \in \mathbb{Z}^+$ 为两个度参数。设 T_p：$\mathbb{R} \mapsto \mathbb{R}^+$ 为指数函数$\exp(\cdot)$在零点处的p 阶泰勒近似：

$$T_p(x) = \sum_{i=0}^{p} \frac{x^i}{i!}.$$

然后，我们将期望奖励函数 $r(w^{(1)}, w^{(2)}\ldots, w^{(n)}, w, S)$ 定义为

$$\left(\prod_{i=1}^{n-1} g_i(dist(w^{(i)}, w^{(i+1)}))\right) \cdot g_n(dist(w^{(n)}, \text{ext}(w,S))) \cdot g_{n+1}(dist(\text{ext}(w,S), w^*)), \tag{1}$$

其中，轮次 i 的多项式 g_i：$\mathbb{R} \mapsto \mathbb{R}$ 定义为

$$g_i(x) = T_p\left(-\frac{x}{v^{q-1}\cdot(3-i/h)}\right). \tag{2}$$

如证明概述中所述，多项式 g_i 的选取旨在确保最优策略尽可能快速地朝着 w^* 前进，而使用低阶泰勒近似则确保最优策略的价值函数可以表示为低维特征的线性函数。

4.2 线性价值函数

当底层公式不可满足时，任意策略都是最优的，因为奖励始终为 0。当公式可满足时，我们将证明“贪心策略”是最优的。

定义 4.4（贪心策略）。如果一个策略在每个状态下，只要可能就选择任何能减小到 w^* 的距离的动作；否则，它会尝试不增加到 w^*的距离，那么我们称该策略为贪心策略。

请注意，根据我们对马尔可夫决策过程的设置，贪心策略是存在的：在一轮的第一阶段，算法会得到一个未满足子句，因此子句中至少有一个变量可以翻转，从而减小当前赋值到 w^*的距离；在第二阶段，算法会逐个得到变量，并且如果当前赋值在该变量上已经与w^* 一致，它总可以选择不翻转该变量。

我们首先讨论与贪心策略 π相关联的价值函数 V^π 。给定一个具有当前赋值 w 和自由变量集合 S 的状态，我们定义以下概念，这些概念在接下来的讨论中会用到。令 $m(S) \in \{0,1\}^v$ 为掩码向量，使得如果第 i个变量属于 S，则 $m(S)_i = 1$，否则 $m(S)_i = 0$。此外，令 $\mathbf{1}$ 表示全一向量，令 $\circ$表示逐元素乘法算子。然后，我们定义

$$\begin{aligned}\text{dist}_{S,\text{free}}(w, w^*) &= \text{dist}(w \circ m(S), w^* \circ m(S)) \\ \text{dist}_{S,\text{used}}(w, w^*) &= \text{dist}(w \circ (\mathbf{1} - m(S)), w^* \circ (\mathbf{1} - m(S)))\end{aligned}$$

换句话说，$\text{dist}_{S,\text{used}}(w, w^*)$ 和 $\text{dist}_{S,\text{free}}(w, w^*)$ 分别表示当前赋值与 w^* 不同的已使用变量和自由变量的数量。注意，$\text{dist}_{S,\text{used}}(w, w^*) + \text{dist}_{S,\text{free}}(w, w^*) = \text{dist}(w, w^*)$。

此外，由于两个向量 $a, b \in \{-1, 1\}^v$ 的汉明距离 $\text{dist}(a, b)$ 关于 a 和 b 都是线性的（因为 $\text{dist}(a, b) = (v - \langle a, b\rangle)/2$），这意味着 $\text{dist}_{S,\text{free}}(w, w^*)$ 和 $\text{dist}_{S,\text{used}}(w, w^*)$ 可以表示为 w^* 的线性函数，以及一些仅取决于当前赋值 w 和自由变量集合 S 的状态特定参数。这使我们能够证明，贪心策略的价值函数也可以表示为 w^* 的线性函数，以及一些状态特定参数。

引理 4.5。当 φ 可满足时，在状态 s、轮次历史 $w^{(1)}, \ldots, w^{(n)}$、当前赋值 w 以及自由变量集合 S下，贪心策略的价值为

$$V^\pi(s) = \prod_{i=1}^{n-1} g_i(\mathit{dist}(w^{(i)}, w^{(i+1)})) \cdot g_n(\mathit{dist}(w^{(n)}, w) + \mathit{dist}_{S,\mathit{free}}(w, w^*)) \cdot g_{n+1}(\mathit{dist}_{S,\mathit{used}}(w, w^*)). \quad (3)$$

因此，存在特征 $\psi(s)$ $\psi(s, a) \in \mathbb{R}^d$，其特征维度为 $d \le 2v^{2p}$，且仅取决于状态 s 和动作 a；还存在 $\theta \in \mathbb{R}^d$，其仅取决于 w^*，使得 V^π 和 Q^π 都可以表示为特征 ψ的线性函数，即 $V^\pi(s) = \langle \theta, \psi(s)\rangle$ 和 $Q^\pi(s, a) = \langle \theta, \psi(s, a)\rangle$。

证明。 第一个结论源于这样一个事实：只要存在能够缩短当前赋值与马尔可夫决策过程所用的最优赋值之间距离的动作，贪心策略就会选择该动作 w^*。因此，从状态 s出发，它会在当前轮翻转所有满足 w 与w^* 不同的自由变量，然后在下一轮翻转所有满足 w 与 w^* 不同的已使用变量。到达 w^*后，最终获得的奖励恰好等于方程（3），且没有中间奖励。

遵循贪心策略可能无法到达w^*。这种情况唯一可能发生的原因是马尔可夫决策过程提前终止：当满足的子句比例超过$(1-\epsilon)$时，或者在最后一轮到达最终状态时。在这种情况下，收到的奖励取决于终止状态的扩展赋值。根据定义4.2 和4.3可知，收到的奖励与马尔可夫决策过程未在该时刻终止时本应获得的奖励相同

。因此，收到的奖励仍与方程（3）一致。

为证明第二个论断，我们采用与 [KLLM22] 中命题10的证明类似的方法。具体而言，我们将证明 $V^\pi(s)$ 可以写成关于 w^* 的次数至多为 $2p$ 的多项式。为说明这一点为何足够，我们令 θ 表示 w^* 中次数至多为 $2p$ 的所有单项式。也就是说，θ 的每个坐标都对应于一个大小为 $|S| \le 2p$ 的多重集 $S \subset [v]$，其值为 $\theta_S = \prod_{i \in S} w_i^*$。我们令 $\psi(s)$ 表示多项式 V^π 中对应的系数。于是，我们可以写成 $V^\pi(s) = \langle \theta, \psi(s) \rangle$。由于系数最多有 $\sum_{i=0}^{2p} v^i \le 2v^{2p}$ 个，我们可以将特征维度设为 $d = 2v^{2p}$。

最后，我们证明 $V^\pi(s)$ 可以写成关于 w^* 的次数至多为 $2p$ 的多项式。首先回顾一下，$\mathrm{dist}_{S,\mathrm{free}}(w, w^*)$ 和 $\mathrm{dist}_{S,\mathrm{used}}(w, w^*)$ 可以表示为 w^* 以及一些特定于状态的参数 $w^{(n)}$、w 和 S 的线性函数。此外，$\mathrm{dist}(w^{(n)}, w)$ 与 w^* 无关，只取决于 $w^{(n)}$ 和 w。然后，只需注意表达式中的每个 $g_i(\cdot)$ 均满足以下条件即可证明该事实：(i) 对于 $i < n$，它与 w^* 无关；(ii) 对于 $i \in \{n, n+1\}$，它是 $\mathrm{dist}(w^{(n)}, w)$、$\mathrm{dist}_{S,\mathrm{free}}(w, w^*)$ 和 $\mathrm{dist}_{S,\mathrm{used}}(w, w^*)$ 的次数为 p 的多项式。

最后请注意，在确定性马尔可夫决策过程中，对于 $\psi(s, a) = \psi(P(s, a))$，线性 V^π 意味着线性 Q^π；这是因为根据定义，在具有确定性转移的马尔可夫决策过程中，$Q^\pi(s, a) = r(s, a) + V^\pi(P(s, a))$，而我们的马尔可夫决策过程中，除最后一个阶段外，奖励均为零；最后一个阶段的奖励则与动作无关。 □

现在我们来证明多项式 g_i 的一些结构性质。首先，我们将证明，如果一个策略在一轮中进行的翻转次数“合理”，那么价值函数会按乘法因子递减。这是因为 g_i 是 $\exp(\cdot)$ 函数在零点附近的泰勒近似（次数适当）。

引理 4.6. 方程 g_i 中定义的多项式 (2) 有界：

$$\frac{1}{4} \le g_i(x) \le 1 - \frac{\epsilon}{6bv^{q-2}}$$

对于所有 $\frac{\epsilon}{b} \cdot v \le x \le v$ 和 $i \in \{1, \ldots, h\}$ 均成立。此外，g_i 单调递减。

证明。为简单起见，令

$$z = \frac{x}{v^{q-1} \cdot (3 - i/h)}.$$

对于我们感兴趣的 x 取值范围，由于 $q \ge 2$，因此可得 $z \le 1/2$。接着，利用 g_i 是泰勒近似这一事实，我们可以将 g_i 的上界估计为

$$\begin{aligned} g_i(x) = T_p(-z) = \sum_{j=0}^{p} \frac{(-z)^j}{j!} &\le 1 - \frac{z}{2} && \text{(as } p \ge 2 \text{ and } z \le 1/2\text{)} \\ &\le 1 - \frac{\epsilon}{6bv^{q-2}}. && \text{(as } x \ge \tfrac{\epsilon}{b} \cdot v\text{)} \end{aligned}$$

另一方面，我们可以如下给出 $g_i(x)$ 的下界：

$$g_i(x) = \exp(-z) - \sum_{j=p+1}^{\infty} \frac{(-z)^j}{j!} \ge \exp\left(-\frac{1}{2}\right) - \frac{1}{2^p(p+1)!} \ge \frac{1}{4}, \tag{4}$$

其中，第一个不等式同样由 $z \le 1/2$ 以及几何级数的求和得到，而最后一个不等式在 $p \ge 1$成立时成立。

接下来，我们证明 $g_i(x)$ 单调递减。为此，我们证明 $g_i(x)$ 的导数为负。对此，我们计算

$$\frac{d}{dx} g_i(x) = \left(\frac{d}{dx} z\right) \cdot \left(\frac{d}{dz} T_p(-z)\right) = -\sum_{j=0}^{p-1} \left(\frac{(-z)^j}{j!}\right).$$

与方程（4）类似，我们有

$$\sum_{j=0}^{p-1}\left(\frac{(-z)^j}{j!}\right) \geq \exp\left(-\frac{1}{2}\right) - \frac{1}{2^{p-1}p!} > 0$$

无论 r $p \geq 2$。因此，$\frac{d}{dx}g_i(x) < 0$，这意味着 $g_i(x)$ 单调递减 . □

接下来，我们将证明这些多项式的设计使得在第 i 轮校正变量（ w 和 w^*不同之处）总是优于在第 $i+1$ 轮校正变量。具体而言，假设我们在第 i轮翻转了 c位，在第 $(i+1)$轮翻转了 d位。我们要证明，对于任意 $1 \leq c \leq v$和 $0 \leq d \leq v$，都有 $g_i(c) \cdot g_{i+1}(d) \geq g_i(c-1) \cdot g_{i+1}(d+1)$。为此，我们需要证明，依赖于 p 选择的泰勒近似误差相对较小。

命题 4.7。 对于任意两个多项式 g_i， g_{i+1} 由方程 (2)定义，令

$$f_{i,c,d}(x) = g_i(c+x) \cdot g_{i+1}(d-x).$$

其中 $i \in \{1,\ldots,h\}$、 $0 \leq c$、 $d \leq v$和 $x = \{1,2,\ldots,d\}$。那么，当 v足够大时 ，

$$f_{i,c,d}(x) \geq f_{i,c,d}(x-1).$$

ˆ 证明。考虑函数 $f_{i,c,d}(\cdot)$，其定义为

$$\hat{f}_{i,c,d}(x) = \exp\left(-\frac{c+x}{v^{q-1}\cdot(3-i/h)}\right) \cdot \exp\left(-\frac{d-x}{v^{q-1}\cdot(3-(i+1)/h)}\right). \tag{5}$$

为证明我们的结论，我们将证明

$$\hat{f}_{i,c,d}(x) - \hat{f}_{i,c,d}(x-1) \geq \Omega\left(\frac{1}{\alpha\cdot v^{2q-2}}\right), \tag{6}$$

$$|f_{i,c,d}(x) - \hat{f}_{i,c,d}(x)| = O\left(\frac{1}{v^{2q-2}}\right), \tag{7}$$

其中，回顾一下，方程 (6) 中的 α 是时间范围因子中的参数，即 $H = \alpha \cdot v^q$ 和 $h = H/v$。因此，只要将 α 设为足够小的常数，上述不等式即可推出我们的结论。ˆ 我们首先证明方程 (6)。为此，我们将证明 $f_{i,c,d}(x)$ 的导数不仅为正，而且有下界 $\Omega(1/\alpha \cdot v^{2q-2})$。ˆ $f(\cdot)$ 的导数为

$$\hat{f}'_{i,c,d}(x) = \hat{f}_{i,c,d}(x) \cdot \frac{h}{v^{q-1}\cdot(3h-i)\cdot(3h-i-1)}. \tag{8}$$

注意，我们始终有

$$\hat{f}_{i,c,d}(x) \geq \exp\left(-\frac{2v}{v^{q-1}}\right)\cdot\exp\left(-\frac{v}{v^{q-1}}\right) = \exp\left(-\frac{3}{v^{q-2}}\right) \geq \Omega(1),$$

其中第一步由 $0 \leq c$， d， $x \leq v$以及 $i \in \{1,\ldots,h\}$，最后一步由 $q \geq 2$得出。因此我们有 ˆ$f_{i,c,d}(x) \geq \Omega(1)$。结合 $h := \alpha \cdot v^{q-1}$ 和方程（8），我们可以得到导数的下界

$$\hat{f}'_{i,c,d}(x) \geq \Omega\left(\frac{1}{\alpha\cdot v^{2q-2}}\right).$$

ˆ由于 $f_{i,\ c,\ d}$ 是凸函数，这就证明了方程（6）。

接下来，我们证明方程（7）。回顾一下，

$$g_i(y) = T_p\left(\frac{-y}{v^{q-1}\cdot(3-i/h)}\right)$$

其中， T_p 是指数函数的p 阶泰勒近似。于是，对于 $0 \le y \le 2v$ ，我们有

$$\left|g_i(y) - \exp\left(-\frac{y}{v^{q-1}\cdot(3-i/h)}\right)\right| \le O\left(\left(\frac{y}{v^{q-1}\cdot(3-i/h)}\right)^{p+1}\right) \le O\left(\frac{1}{v^{(q-2)\cdot(p+1)}\cdot 2^{p+1}}\right). \quad (9)$$

此外，对于 $y \ge 0$ ，我们有

$$\exp(-y/(v^{q-1}\cdot(3-i/h)) < 1. \quad (10)$$

将方程（9）和（10）代入方程（5），得到

$$\begin{aligned} f_{i,c,d}(x) &= \left(\exp\left(-\frac{c+x}{v^{q-1}\cdot(3-i/h)}\right) \pm O\left(\frac{1}{v^{(q-2)\cdot(p+1)}\cdot 2^{p+1}}\right)\right)\cdot \\ &\quad \left(\exp\left(-\frac{d-x}{v^{q-1}\cdot(3-(i+1)/h)}\right) \pm O\left(\frac{1}{v^{(q-2)\cdot(p+1)}\cdot 2^{p+1}}\right)\right) \\ &= \hat{f}_{i,c,d}(x) \pm O\left(\frac{1}{v^{(q-2)\cdot(p+1)}\cdot 2^{p+1}}\right). \end{aligned}$$

对于我们考虑的 p和 q这两种设置， $p=2$；对于第一个结果， $q=4$；对于第二个结果， $p=2$对数 v；$q=2$，这意味着

$$|f_{i,c,d}(x) - \hat{f}_{i,c,d}(x)| \le O\left(\frac{1}{v^{(q-2)\cdot(p+1)}\cdot 2^{p+1}}\right) = O\left(\frac{1}{v^{2q-2}}\right).$$

□

递归地应用上述结论，可以证明，在我们的马尔可夫决策过程中，任何贪心策略都是最优策略。这一点很重要，因为结合引理4.5 可知，最优价值函数 V^* 和 Q^*可以表示为仅依赖于状态 s 和动作 a的某些特征的线性函数。

引理 4.8。 定义 *4.4* 中定义的任何贪心策略都是最优的。

证明。设 π 为贪心策略，并考虑另一个策略 $\tilde{\pi}$。我们证明，从任意状态 s_{curr} 出发，在赋值 w_{curr}、自由变量 S_{curr} 和轮次历史 $w^{(1)},\ldots,w^{(n)}$ 的情况下，贪心策略 π 获得的奖励不低于 $\tilde{\pi}$。注意， $\tilde{\pi}$ 和 π 的最终奖励中都会包含项 $G := \prod_{i=1}^{n-1} g_i(\text{dist}(w^{(i)}, w^{(i+1)}))$。我们将该项简记为 G ，以便专注于比较其余各项。

假设 $\tilde{\pi}$ 在第（$n+l$）轮终止。具体而言，假设它在状态 $\tilde{s}$ 终止，其轮次历史为 $w^{(1)},\ldots,w^{(n)}$, $\tilde{w}^{(n+1)},\ldots,\tilde{w}^{(n+l)}$，自由变量为 $\tilde{S}$，终止赋值为 $\tilde{w}$。为方便记号，我们将 $\tilde{w}^{(n+l+1)} = \text{ext}\left(\tilde{w},\tilde{S}\right)$. [4]。那么，从 w_{curr} 开始时 $\tilde{\pi}$ 的值（也就是状态 $\tilde{s}$ 的奖励）可以写为

$$V^{\tilde{\pi}}(s_{\text{curr}}) = G\cdot\prod_{i=n}^{n+l} g_i\left(\text{dist}(\tilde{w}^{(i)},\tilde{w}^{(i+1)})\right)\cdot g_{n+l+1}\left(\text{dist}\left(\tilde{w}^{(n+l+1)},w^*\right)\right). \quad (11)$$

[4]注意，终止状态 $\tilde{s}$ 可能与 s_{curr} 处于同一轮次。在这种情况下，我们有 $l=0$。

首先，我们说明在第（$n+1$）轮之后终止对 $\tilde{\pi}$ 从来都没有益处。更形式化地说，我们将证明

$$\prod_{i=n}^{n+l} g_i(\text{dist}(\tilde{w}^{(i)}, \tilde{w}^{(i+1)})) \cdot g_{n+l+1}(\text{dist}(\tilde{w}^{(n+l+1)}, w^*)) \leq g_n\left(\tilde{w}^{(n)}, \tilde{w}^{(n+1)}\right) \cdot g_{n+1}\left(\tilde{w}^{(n+1)}, w^*\right). \quad (12)$$

根据命题 4.7，我们有

$$\begin{aligned}
&g_{n+l}\left(\text{dist}(\tilde{w}^{(n+l)}, \tilde{w}^{(n+l+1)})\right) \cdot g_{n+l+1}\left(\text{dist}(\tilde{w}^{(n+l+1)}, w^*)\right) \\
&\leq g_{n+l}\left(\text{dist}(\tilde{w}^{(n+l)}, \tilde{w}^{(n+l+1)}) + \text{dist}(\tilde{w}^{(n+l+1)}, w^*)\right) \cdot g_{n+l+1}(0) \leq g_{n+l}\left(\text{dist}(\tilde{w}^{(n+l)}, w^*)\right),
\end{aligned}$$

其中，最后一个不等式由 $g_{n+l+1}(0) = 1$、对 dist 使用三角不等式$(\cdot)$ 以及$g_{n+l}(\cdot)$ 单调递减（命题 4.6）得出。由此可得

$$\prod_{i=n}^{n+l} g_i(\text{dist}(\tilde{w}^{(i)}, \tilde{w}^{(i+1)}))\cdot g_{n+l+1}(\text{dist}(\tilde{w}^{(n+l+1)}, w^*)) \leq \prod_{i=n}^{n+l-1} g_i(\text{dist}(\tilde{w}^{(i)}, \tilde{w}^{(i+1)}))\cdot g_{n+l}(\text{dist}(\tilde{w}^{(n+l)}, w^*)).$$

接下来，我们对 l 进行归纳，从而得到方程 (12)。将方程 (12) 代入方程 (11)，即可得到

$$V^{\tilde{\pi}}(s) \leq G \cdot g_n\left(\text{dist}\left(w^{(n)}, \tilde{w}^{(n+1)}\right)\right) \cdot g_{n+1}\left(\text{dist}\left(\tilde{w}^{(n+1)}, w^*\right)\right). \quad (13)$$

接下来，我们证明上面的表达式以 V^{π} 为上界。注意，根据三角不等式，对于任意 $\tilde{w}^{(n+1)}$，都有

$$\text{dist}(w_{\text{curr}}, \tilde{w}^{(n+1)}) + \text{dist}(\tilde{w}^{(n+1)}, w^*) \geq \text{dist}(w_{\text{curr}}, w^*).$$

另一方面，我们总有

$$\text{dist}(w_{\text{curr}}, w^*) = \text{dist}_{S_{\text{curr}},\text{used}}(w_{\text{curr}}, w^*) + \text{dist}_{S_{\text{curr}},\text{free}}(w_{\text{curr}}, w^*).$$

将两者结合并重新整理各项，可得

$$\text{dist}(w_{\text{curr}}, \tilde{w}^{(n+1)}) + \text{dist}(\tilde{w}^{(n+1)}, w^*) - \text{dist}_{S_{\text{curr}},\text{free}}(w_{\text{curr}}, w^*) \geq \text{dist}_{S_{\text{curr}},\text{used}}(w_{\text{curr}}, w^*) \quad (14)$$

现在，我们将根据 dist(w_{curr}, $\tilde{w}^{(n+1)}$) 和 dist$_{S_{\text{curr}},}$free(w_{curr}, w^*) 的相对大小进行分类讨论。我们首先考虑 dist(w_{curr}, $\tilde{w}^{(n+1)}$) $\leq$ dist$_{S_{\text{curr}},}$free(w_{curr}, w^*) 的情形。在这种情形下，我们有

$$\begin{aligned}
&g_n\left(\text{dist}(w^{(n)}, \tilde{w}^{(n+1)})\right) \cdot g_{n+1}\left(\text{dist}(\tilde{w}^{(n+1)}, w^*)\right) \\
&= g_n\left(\text{dist}(w^{(n)}, w_{\text{curr}}) + \text{dist}(w_{\text{curr}}, \tilde{w}^{(n+1)})\right) \cdot g_{n+1}\left(\text{dist}(\tilde{w}^{(n+1)}, w^*)\right) \\
&\leq g_n\left(\text{dist}(w^{(n)}, w_{\text{curr}}) + \text{dist}_{S_{\text{curr}},\text{free}}(w_{\text{curr}}, w^*)\right) \\
&\cdot g_{n+1}\left(\text{dist}(\tilde{w}^{(n+1)}, w^*) + \text{dist}(w_{\text{curr}}, \tilde{w}^{(n+1)}) - \text{dist}_{S_{\text{curr}},\text{free}}(w_{\text{curr}}, w^*)\right) \\
&\leq g_n\left(\text{dist}(w^{(n)}, w_{\text{curr}}) + \text{dist}_{S_{\text{curr}},\text{free}}(w_{\text{curr}}, w^*)\right) \cdot g_{n+1}\left(\text{dist}_{S_{\text{curr}},\text{used}}(w_{\text{curr}}, w^*)\right),
\end{aligned}$$

其中，第一个不等式由命题 4.7得出，第二个不等式由 g_n 是一个单调递减函数（命题 4.6）以及方程（14）得出。

现在，我们考虑另一个剩余情形：当 $\text{dist}(w_{\text{curr}}, \tilde{w}^{(n+1)}) > \text{dist}_{S_{\text{curr}}, \text{free}}(w_{\text{curr}}, w^*)$ 时。将 M 记为这样一组自由变量： w_{curr} 与 w^* 在这些变量上取值相同，而 w_{curr} 与 $\tilde{w}^{(n+1)}$ 在这些变量上取值不同。换言之，这些变量是在从 w_{curr} 到 $\tilde{w}^{(n+1)}$的路径上被策略 $\tilde{\pi}$ 错误翻转的变量。由于 $\tilde{w}^{(n+1)}$与 w^*在这些变量上的取值不同，因此这些变量必定会在从 $\tilde{w}^{(n+1)}$ 到 w^*的路径上再次被翻转。接着，我们可以考虑 $\bar{w}$ 的另一条路径 $w_{\text{curr}} \to \bar{w} \to w^*$，满足 $\bar{w}_i \neq \tilde{w}_i^{(n+1)}$ 对于 $i \in M$ 成立，且 $\bar{w}_i = \tilde{w}_i^{(n+1)}$ 对于 $i \notin M$成立。于是，容易看出

$$
\begin{aligned}
& g_n(\text{dist}(w^{(n)}, w_{\text{curr}}) + \text{dist}(w_{\text{curr}}, \tilde{w}^{(n+1)})) \cdot g_{n+1}(\text{dist}(\tilde{w}^{(n+1)}, w^*)) \\
& \leq g_n(\text{dist}(w^{(n)}, w_{\text{curr}}) + \text{dist}(w_{\text{curr}}, \bar{w})) \cdot g_{n+1}(\text{dist}(\bar{w}, w^*))
\end{aligned}
$$

由于 $g_n(\cdot)$ 是单调递减的。此外，现在我们有 $\text{dist}(w_{\text{curr}}, \bar{w}) \leq \text{dist}_{S_{\text{curr}}, \text{free}}(w_{\text{curr}}, w^*)$，因为被翻转的变量被限制为那些 w_{curr} 与 w^* 不一致的变量。因此，证明归结为第一种情况。

□

4.3 将强化学习算法转换为 SAT 算法

沿用先前下界证明中采用的方法 [KLLM22], ，我们现在利用一个用于强化学习问题的随机化算法 $\mathcal{A}_{RL}$，构造一个用于 3-SAT 的随机化算法$\mathcal{A}_{SAT}$。具体而言，我们为 MDP 预言机 M_φ 构造一个“近似”模拟器 M_φ。就与 MDP M_φ 相关的转移函数和特征而言，该模拟器 M_φ 与 MDP M_φ 完全相同，但在最后一层的奖励函数方面有所不同：对于模拟器 M_φ，该函数始终为 0。通过这一修改，我们可以在 poly(d) 时间内执行对模拟器 M_φ 的每次调用¯。

算法。输入 3-CNF 公式 φ, $\mathcal{A}_{SAT}$后，运行算法 $\mathcal{A}_{RL}$，将每次对 MDP 预言机M_φ 的调用替换为对模拟器 M_φ的相应调用。回想一下，在我们的设定（确定性转移马尔可夫决策过程）中，强化学习算法的输出是一系列动作。如果 $\mathcal{A}_{RL}$ 返回的动作序列最终到达一个状态，且该状态的赋值 w 满足超过 $(1-\epsilon)$ 比例的子句，则 $\mathcal{A}_{SAT}$ 立即终止模拟并输出 YES。如果在整个模拟过程中， $\mathcal{A}_{RL}$ 始终没有找到与此类赋值关联的任何状态，则 $\mathcal{A}_{SAT}$ 输出 NO。

正确性。为完成我们的归约，我们将证明以下内容：

(i) 如果算法 $\mathcal{A}_{RL}$ 输出一个策略 π，使得 $V^\pi > V^* - 1/8$，那么对于 3-CNF 公式 φ， $\mathcal{A}_{SAT}$ 会在 φ 可满足时输出 YES，否则输出 NO。(ii) 如果 $\mathcal{A}_{RL}$ 在能够访问 MDP 预言机 M_φ 的情况下输出一个策略 π，使得 $V^\pi > V^* - 1/8$ 的错误概率为¯ $1/10$，那么 $\mathcal{A}_{RL}$ 在能够访问模拟器 M_φ 的情况下会输出一个策略 π，使得 $V^\pi > V^* - 1/8$相对于 M_φ 的错误概率为 $1/8$ （也就是说，尽管 $\mathcal{A}_{RL}$ 正在与¯模拟器 M_φ交互，返回的策略仍保证在真实的马尔可夫决策过程 M_φ上表现良好）。

回想一下，如果 ϕ 不可满足，那么任意策略都是最优的。上述两个论断由此证明， $\mathcal{A}_{SAT}$ 以错误概率 $\leq 1/8$ 求解 GAP-3-SAT。我们首先证明，如果 $\mathcal{A}_{RL}$ 在马尔可夫决策过程 M_φ上成功，那么$\mathcal{A}_{SAT}$就在 3-CNF 公式 φ上成功。这是因为，马尔可夫决策过程 M_φ 中的任意优良策略都必须到达一个赋值为 w^*的状态； w^*是构造 M_φ时任意选取的满足赋值。

命题 4.9。假设 α、b、ϵ 是常数，且 v 足够大。那么，如果 φ 可满足，并且 $\mathcal{A}_{RL}$ 在 M_φ 上运行后返回策略 π，满足 $V^\pi > V^* - 1/8$，那么 π 最终会得到一个赋值，该赋值至少满足比例为 $(1-\epsilon)$ 的子句。

证明。取一个可满足公式 φ。此时的最优值至少为 $1/4$。事实上，根据引理 4.8，贪心策略是最优的，其值为 g_1(距离$(w,\ w^*)$)，因此根据推论 4.6，

$$V^* = g_1(\text{dist}(w, w^*)) \geq \frac{1}{4}.$$

现在我们通过逆否命题进行论证：假设 π 最终得到的赋值满足的子句比例不足 $(1-\epsilon)$。令 $w^{(1)}$、.....、 $w^{(h)}$、 $\bar{w}$ 表示通过 π:$w^{(1)} = w$ 得到的赋值序列，其中 $w^{(i+1)}$ 是第 $1 \leq i \leq h-1$ 轮结束时的赋值， $\bar{w}$ 是最终赋值。回顾一下，每一轮中，马尔可夫决策过程有两个阶段。在第一阶段，智能体会看到仅由自由变量构成的未满足子句。根据我们的构造，第一阶段的长度至少为 $\epsilon v/b$。因此， dist$(w^{(i)}, w^{(i+1)}) \geq \epsilon v/b$，因为策略不允许撤销任何翻转。于是，我们可以给最终获得的奖励一个上界：

$$\prod_{i=1}^{h} g_i(\text{dist}(w^{(i)}, w^{(i+1)})) \cdot g_{h+1}(\text{dist}(\bar{w}, w^*)) \leq \prod_{i=1}^{h} g_i(\epsilon v/b) \leq \left(1 - \frac{\epsilon}{6bv^{q-2}}\right)^h \leq \exp(-cv),$$

其中 $c = \Theta(\alpha\epsilon/b)$，第一个不等式由 dist$(w^{(i)}, w^{(i+1)}) \geq \epsilon v/b$得出，第二个不等式来自命题 4.6，第三个不等式则由对所有 $1 - x \leq e^{-x}$ 均成立的 x 以及我们对 $h = \alpha v^{q-1}$的选择得出。因此，如果 $V^\pi > V^* - 1/8 \geq 1/8$，且 v 足够大，使得 $\exp(-cv) < 1/8$，那么策略 π 最终得到的状态必须至少满足 $(1-\epsilon)$比例的子句。 □

接下来，我们将证明，即使在模拟器 M_φ上运行， $\mathcal{A}_{RL}$ 的行为也大致相同。具体而言，给定 $\mathcal{A}_{RL}$ 在次指数时间内运行并在 M_φ上成功，我们可以认为， $\mathcal{A}_{RL}$ 在 M_φ 和 M_φ 上执行时将获得大致相同的信息，因此即使面对模拟器 M_φ 的输出，也会以较小的常数概率在上成功。

命题 4.10。假设 $\mathcal{A}_{RL}$ 可访问 MDP 预言机 M_φ，在时间 T内运行，并输出策略 π，使得 $V^\pi > V^* - 1/8$ 的错误概率为 $1/10$。此外，假设 M_φ 在最后一层的期望奖励上界为 $1/(5T)$。那么， $\mathcal{A}_{RL}$ 在可访问模拟器 M_φ的情况下，仍在时间 T内运行，并输出策略 π，使得相对于 M_φ， $V^\pi > V^* - 1/8$ 的错误概率为 $1/8$。

证明。令 $\Pr^{M_\varphi}$ 和 $\Pr_{\bar{M}_\varphi}$ 分别表示在算法 $\mathcal{A}_{RL}$ 使用 MDP 预言机 M_φ 和模拟器 M_φ 时，所诱导的观测奖励与输出策略上的分布。令 i 表示第 N 条轨迹结束时在最后一层获得的奖励，令 $\mathcal{A}_{RL}$ 表示算法 $\mathcal{A}_{RL}$ 使用 MDP 预言机 M_φ 时采样的轨迹总数。根据我们的假设，$\mathcal{A}_{RL}$ 的运行时间为 T，因此 $N \leq T$。

我们指出，如果算法 $\mathcal{A}_{RL}$ 曾到达满足赋值， $\mathcal{A}_{SAT}$ 会立即终止模拟并返回 YES。在到达满足赋值之前， $\mathcal{A}_{RL}$ 只能从最后一层获得奖励。根据我们的假设，MDP M_φ 中最后一层的期望奖励以$1/(5T)$为上界，并且该算法在最后一层至多访问 $N \leq T$ 个状态，因此由并集界可知，以高概率，最后一层上的所有奖励均为零。更准确地说，我们有

$$\Pr_{M_\varphi}[R_i = 0\ \forall i \in [N]] \geq 1 - T/(5T) \geq \frac{4}{5}.$$

我们称 $\mathcal{A}_{RL}$ 在可访问 M_φ （或 M_φ）的情况下取得成功，是指输出策略 π 在运行时间不超过 T 后满足 $V^\pi > V^* - 1/8$，且该结论相对于 M_φ 而言。根据上述推理，并假设 $\mathcal{A}_{RL}$在可访问 MDP 预言机 M_φ 的情况下以概率 $9/10$ 取得成功，则可推出

$$\Pr_{M_\varphi}[\mathcal{A}_{RL} \text{ succeeds with access to } M_\varphi \mid R_i = 0\ \forall i \in [N]] \geq \frac{\frac{9}{10} - \frac{1}{5}}{\frac{4}{5}} = \frac{7}{8}.$$

注意，在给定 $R_i = 0\ \forall i \in [N]$ 的条件下，边缘分布 $\Pr^{M_\varphi}$ 和 $\Pr_{\bar{M}_\varphi}$ 完全相同，因为在 $\mathcal{A}_{RL}$ 达到满足赋值之前，MDP 预言机 M_φ 与模拟器 M_φ 的差异可能仅体现在最后一层奖励上。这意味着

$$\begin{aligned}&\Pr_{\bar{M}_\varphi}\left[\mathcal{A}_{RL}\text{ succeeds with access to }\bar{M}_\varphi \mid R_i = 0\ \forall i \in [N]\right]\\&= \Pr_{M_\varphi}\left[\mathcal{A}_{RL}\text{ succeeds with access to }M_\varphi \mid R_i = 0\ \forall i \in [N]\right]\end{aligned}$$

由于 $\Pr_{\bar{M}_\varphi}[R_i = 0\ \forall i \in [N]] = 1$，我们得出

$$\Pr_{\bar{M}_\varphi}\left[\mathcal{A}_{RL}\text{ succeeds with access to }\bar{M}_\varphi\right] \geq \frac{7}{8}.$$

□

接下来我们证明使用标准归约，(b, ϵ)- GAP-3-SAT 的近似求解难度与 3-SAT 大致相当。

命题 4.11。在 rETH 假设下，存在常数 b、ϵ、$c > 0$，使得不存在随机化算法能够在时间 $\exp(cv/\text{polylog}(v))$ 内，以错误概率 $1/8$ 求解带有 v 个变量的 (b, ϵ)- GAP-3-SAT。

我们在附录 A 中给出了证明。现在，我们准备证明主要结果——命题 4.1。为此，我们说明如何将一个 (b, ϵ)- GAP-3-SAT 实例归约为一个马尔可夫决策过程实例。

命题 4.1 的证明。令 $p = 2$、$q = 4$ 或 $p = 2\log v$，并令 $q = 2$。对于任意 $v \in \mathbb{Z}^+$，假设存在一个算法 $\mathcal{A}_{RL}$，能够以错误概率 $1/10$ 求解特征维度为 $d = \Theta\left(v^{2p}\right)$ 且 $H = \Theta\left(v^q\right)$ 的 LINEAR-3-RL，并且当 $c_1 < \min(1/2, c/2)$ 时，其运行时间为 $\exp(c_1 \cdot v/\text{polylog}(v))$，其中 c 是命题 4.11 中的常数。那么，我们声称可以构造另一个算法 $\mathcal{A}_{SAT}$，以错误概率 $1/8$ 在时间 $\exp(cv/\text{polylog}(v))$ 内求解 (b, ϵ)- GAP-3-SAT。注意到，在 rETH 假设下，这将与命题 4.11 矛盾，从而证明我们的命题。

设 φ 为一个（b，ϵ）-GAP-3-SAT 实例的 3-CNF 公式，其中包含 v 个变量和至少 v 个子句。根据定义，每个变量最多出现在 b 个子句中。此外，可以保证 φ 要么是可满足的，要么在任意赋值下至少有 ϵ 比例的子句不可满足。为了在这两种情况下作出判断，我们首先按照第 4.1 节中的描述，构造一个马尔可夫决策过程 M_φ（由两个正整数 p、q 参数化）。具体而言，该马尔可夫决策过程被设计为包含 αv^{q-1} 轮次，并且多项式 g_i 将是公式（2）中规定的 p 阶泰勒近似。由于 α、b、ϵ 是绝对常数，下面忽略对它们的依赖关系。

接下来，我们将分别界定 M_φ 的时域和特征维度。由于每轮包含 v 个步骤，时域为 $H = \Theta\left(v^q\right)$。此外，根据引理 4.5，贪心策略的价值函数可以表示为一个特征向量的线性函数，该特征向量的大小为 $\Theta\left(v^{2p}\right)$。根据引理 4.8，贪心策略是最优的。因此，该马尔可夫决策过程的特征维度为 $d = \Theta\left(v^{2p}\right)$。

接下来，正如命题 4.9 的证明中所指出的，对于任何在最后一层终止的策略 π，其期望奖励始终以 $\exp(-v)$ 为上界。令 M_φ 为一个与 M_φ 仅在时域结束时获得的奖励方面不同的马尔可夫决策过程（M_φ 的奖励始终为 0）。那么，根据命题 4.10 以及上文提到的较小的 $\exp(-v)$ 奖励，我们知道，当 $\mathcal{A}_{RL}$ 在 M_φ 的模拟器上以 $c_1 < 1/2$ 的方式运行至多 $\exp(c_1 \cdot v/\text{polylog}(v))$ 时间时，它仍会以至少 $7/8$ 的概率输出一个相对于 M_φ 的良好策略 $\tilde{\pi}$。

根据命题 4.9，如果 $\mathcal{A}_{RL}$ 成功且 φ 可满足，那么策略 $\tilde{\pi}$ 将在一个满足该条件的赋值处终止。因此，我们只需检查运行策略 $\tilde{\pi}$ 所得到的路径，即可判断 φ是否可满足，这至多需要 $\text{poly}(v)$ 时间。

因此，若存在这样一种算法 $\mathcal{A}_{RL}$，其运行时间至多为 $\exp(c_1 \cdot v/\text{polylog}(v))$，则意味着存在另一种算法，能够在 $\exp(c_1 v/\text{polylog}(v)) + \text{poly}(v) \leq \exp(cv/\text{polylog}(v))$ 时间内解决 (b, ϵ)- GAP-3-SAT问题。

□

参考文献

[AJS+20] Alex Ayoub、Zeyu Jia、Csaba Szepesvari、Mengdi Wang 和 Lin F Yang。基于模型的强化学习与价值目标回归。arXiv:2006.01107，2020年。[AKKS20] Alekh Agarwal、Sham Kakade、Akshay Krishnamurthy 和 Wen Sun。FLAMBE：低秩马尔可夫决策过程的结构复杂度与表征学习。arXiv 预印本 arXiv:2006.10814，2020年。[CFK+15] Marek Cygan、Fedor Fomin、Lukasz Kowalik、Daniel Lokshtanov、D´aniel Marx、Marcin Pilipczuk、Michał Pilipczuk 和 Saket Saurabh。基于指数时间假设的下界，第467–521页。施普林格，2015年07月。[DHM+14] Holger Dell、Thore Husfeldt、D´aniel Marx、Nina Taslaman 和 Martin Wahl´en。积和式与 Tutte 多项式的指数时间复杂度。《ACM算法汇刊》，10(4)，2014年。[DKJ+19] Simon S Du、Akshay Krishnamurthy、Nan Jiang、Alekh Agarwal、Miroslav Dud´ık 和 John Langford。通过潜在状态解码，实现基于丰富观测的可证明高效强化学习。载于《国际机器学习会议》，2019年。[DKL+21] Simon Du、Sham Kakade、Jason Lee、Shachar Lovett、Gaurav Mahajan、Wen Sun 和 Ruosong Wang。双线性函数类：强化学习中可证明泛化的结构框架。载于《第38届国际机器学习大会论文集》，第139卷，第2826–2836页，2021年7月18–24日。[DKWY20] Simon S Du、Sham M Kakade、Ruosong Wang 和 Lin F Yang。良好的表征是否足以实现样本高效的强化学习？载于《国际学习表征会议》，2020年。[DLMW20] Simon S Du、Jason D Lee、Gaurav Mahajan 和 Ruosong Wang。确定性系统中带函数逼近的无先验 Q 学习：逼近误差与样本复杂度的紧界。载于《神经信息处理系统进展》，2020年。[DRZ20] Shi Dong、Benjamin Van Roy 和 Zhengyuan Zhou。基于聚合状态的可证明高效强化学习，2020年。[FKQR21] Dylan J Foster、Sham M Kakade、Jian Qian 和 Alexander Rakhlin。交互式决策的统计复杂度。arXiv 预印本 arXiv:2112.13487，2021年。[GMR22a] Noah Golowich、Ankur Moitra 和 Dhruv Rohatgi。无需计算上难以处理的预言机，在可观测部分可观测马尔可夫决策过程中学习。arXiv 预印本 arXiv:2206.03446，2022年。[GMR22b] Noah Golowich、Ankur Moitra 和 Dhruv Rohatgi。在准多项式时间内规划可观测部分可观测马尔可夫决策过程。arXiv 预印本 arXiv:2201.04735，2022年。[IP01] Russell Impagliazzo 和 Ramamohan Paturi。论 k-SAT 的复杂度。《计算机系统科学杂志》，62(2)：367–375，2001年。[JKA+16] Nan Jiang、Akshay Krishnamurthy、Alekh Agarwal、John Langford 和 Robert E. Schapire。具有低贝尔曼秩的上下文决策过程可 PAC 学习，2016年。

[JLM21] Chi Jin、Qinghua Liu 和 Sobhan Miryoosefi。贝尔曼 Eluder 维度：新型丰富的强化学习问题类别与样本高效算法。arXiv 预印本 arXiv:2102.00815，2021年。[JYWJ20] Chi Jin、Zhuoran Yang、Zhaoran Wang 和 Michael I Jordan。具有线性函数逼近的可证明高效强化学习。载于学习理论会议，2020年。[KAL16] Akshay Krishnamurthy、Alekh Agarwal 和 John Langford。具有丰富观测的 PAC 强化学习。载于第30届神经信息处理系统国际会议论文集，第1848–1856页，2016年。[KLLM22] Daniel Kane、Sihan Liu、Shachar Lovett 和 Gaurav Mahajan。强化学习中的计算与统计差距。arXiv 预印本 arXiv:2202.05444，2022年。[Li09] Lihong Li。计算强化学习理论的统一框架。罗格斯大学，美国，博士论文，2009年。AAI3386797。[LSS01] Michael L Littman、Richard S Sutton 和 Satinder P Singh。状态的预测表示。载于神经信息处理系统大会，第14卷，第30页，2001年。[LSW20] Tor Lattimore、Csaba Szepesvari 和 Gellert Weisz。在多臂老虎机以及具有生成模型的强化学习中利用良好特征表示进行学习。载于国际机器学习会议，2020年。[MJTS20] Aditya Modi、Nan Jiang、Ambuj Tewari 和 Satinder Singh。使用线性组合模型集成的强化学习的样本复杂度。载于人工智能与统计会议，2020年。[MR08] Dana Moshkovitz 和 Ran Raz。具有次常数误差的双查询概率可检验证明。《美国计算机学会杂志》（JACM），57(5):1–29，2008年。[Mun05] R´emi Munos。近似值迭代的误差界。载于全国人工智能大会论文集，第20卷，第1006页。Menlo Park, CA；Cambridge, MA；伦敦；AAAI Press；MIT Press；1999年，2005年。[PY91] Christos H Papadimitriou 和 Mihalis Yannakakis。优化、近似与复杂度类。《计算机与系统科学杂志》，43(3):425–440，1991年。[SJK+19] Wen Sun、Nan Jiang、Akshay Krishnamurthy、Alekh Agarwal 和 John Langford。上下文决策过程中的基于模型的强化学习：PAC 界及相对于无模型方法的指数级改进。载于学习理论会议，2019年。[USL+22] Masatoshi Uehara、Ayush Sekhari、Jason D Lee、Nathan Kallus 和 Wen Sun。具有潜在确定性和条件嵌入的部分可观测马尔可夫决策过程中的计算高效 PAC 强化学习。arXiv 预印本 arXiv:2206.12081，2022年。[WAJ+21] Gell´ert Weisz、Philip Amortila、Barnab´as Janzer、Yasin Abbasi-Yadkori、Nan Jiang 和 Csaba Szepesv´ari。在线性可实现最优状态价值函数条件下马尔可夫决策过程中的查询高效规划，2021年。[WAS21] Gell´ert Weisz、Philip Amortila 和 Csaba Szepesv´ari。具有线性可实现最优动作价值函数的马尔可夫决策过程中的规划指数下界。载于算法学习理论，第1237–1264页。PMLR，2021年。

[WGKS22] Gell´ert Weisz、Andr´as Gy¨orgy、Tadashi Kozuno 和 Csaba Szepesv´ari。用于在 q^{π}-可实现马尔可夫决策过程中进行高效局部规划的置信近似策略迭代。载于神经信息处理系统进展，2022年。[Wil19] Virginia Vassilevska Williams。关于算法与复杂度中的一些细粒度问题。国际数学家大会（ICM 2018）论文集，2019年。[WSG22] Gell´ert Weisz、Csaba Szepesv´ari 和 Andr´as Gy¨orgy。在最优价值函数线性可实现条件下马尔可夫决策过程规划的 TensorPlan 与少动作下界。载于国际算法学习理论会议，第1097–1137页。PMLR，2022年。[WVR17] Zheng Wen 和 Benjamin Van Roy。具有价值函数泛化的确定性系统中的高效强化学习。《运筹学数学》，42(3):762–782，2017年。[WWK21] Yuanhao Wang、Ruosong Wang 和 Sham M. Kakade。具有常数次优差距的线性可实现马尔可夫决策过程的指数下界，2021年。[YHAY+22] Dong Yin、Botao Hao、Yasin Abbasi-Yadkori、Nevena Lazi´c 和 Csaba Szepesv´ari。采用线性函数逼近的高效局部规划。载于国际算法学习理论会议，第1165–1192页。PMLR，2022年。[YW19] Lin Yang 和 Mengdi Wang。使用线性可加特征的样本最优参数化Q学习。载于国际机器学习会议，2019年。[ZHG21] Dongruo Zhou、Jiafan He 和 Quanquan Gu。具有特征映射的折扣马尔可夫决策过程的可证明高效强化学习。载于国际机器学习会议，第12793–12802页，2021年。[ZLKB20] Andrea Zanette、Alessandro Lazaric、Mykel Kochenderfer 和 Emma Brunskill。在低固有贝尔曼误差下学习近最优策略，2020年。

带间隙且子句数少的近似 SAT 的困难性

本节将证明以下结论：

命题 A.1. 在 rETH 假设下，存在常数 b、ϵ、$c > 0$，使得不存在随机化算法能在 $\exp(cv/\text{polylog}(v))$ 时间内，以错误概率 $1/8$ 求解具有 v 个变量的 (b,ϵ)- GAP-3-SAT。

为证明这一点，我们将考察另一个问题：ϵ- GAP-3-SAT。它与(b,ϵ)- GAP-3-SAT 类似，不过它不限制一个变量可以出现在多少个子句中。利用标准技术，可以证明 ϵ- GAP-3-SAT 同样是困难的。具体而言，其困难性在[MR08] 中得到证明，并依赖于概率可检验证明（PCP）定理的某个版本。

定理 A.2（从 3-SAT 到 GAP-3-SAT 的归约）。对规模为 n 的输入求解 3-SAT，可以归约为区分以下两种情况：一个规模为 $n \cdot \text{polylog}(n)$ 的 3-CNF 公式可满足；或者只有其 $1-\epsilon$ 比例的子句可满足，其中某个常数为 $\epsilon > 0$。

为完整起见，我们为上述定理提供证明。我们首先回顾一些关于概率可检验证明定理的基本概念。给定一个陈述（例如，某个 SAT 实例是否可满足），概率可检验证明验证器可以查询针对该陈述、在字母表 Σ上构造的证明，并需要判断该陈述是否为真。概率可检验证明验证器有几个重要参数。

- 完备性 c：验证器接受正确证明的最小概率。
- 可靠性 ϵ：验证器接受错误定理的证明的最大概率。
- 查询 q：验证器对证明发起的查询次数。
- 大小 m：证明的长度。
- 随机性 r：验证器使用的随机比特数。
- 字母表 Σ：证明中使用的字母表。

我们用 $\text{PCP}_{c,\epsilon}[r,q]_\Sigma$ 表示这样一类语言：它们具有一个 PCP 验证器，其完备性为 c，可靠性为s，随机性为 r，并且 q 对字母表 Σ上的证明进行查询。此外，PCP 验证器只允许执行双查询投影测试。在双查询投影测试中，验证器只允许进行两次查询。看到第一次查询的答案后，验证器要么立即拒绝，要么唯一确定第二次查询的答案，并在该答案下接受。我们的出发点是以下定理，出自[MR08]。

定理 A.3（ [**MR08**] 中的定理 7）。存在一个常数 $\epsilon > 0$ 和一个大小为常数的字母表 Σ，使得 3-SAT$\in PCP_{1,\ 1-\epsilon}[\log\ n + O(\log\log\ n)\ 2]_\Sigma$。

复杂性问题 ϵ- GAP-3-SAT

输入：给定间隙参数 $\epsilon > 0$ 和一个 3-CNF 公式 φ，其中有 v 个变量和 $O(v)$ 个子句，并且满足以下条件之一：(i) φ 可满足；或 (ii) 对于任意赋值，至少有 an ϵ比例的子句不满足，其中 $\epsilon > 0$。目标：判断该公式是否可满足。

定理 A.2 的证明。给定一个大小为 n 的 3-CNF 公式 ψ，目标是利用上述定理中的验证器构造一个不同的 3-SAT 实例 ϕ，其大小为 $O(n$多对数 $n)$，使得：(i) 若 ψ可满足，则 ϕ 可满足；且 (ii) 若 ψ 不可满足，则对于任意赋值，ϕ 中至少有 an ϵ 比例的子句不满足。

注意，如果假设验证器还将 $r := \log\ n + O(\log\log\ n)$ 个随机比特作为输入，那么不失一般性，可以假设验证器是确定性的。固定一个随机比特串后，验证器最多读取证明中的 2 个字符。由于不同的随机比特串至多有 $2^r = n\text{polylog}\ n$ 个，不失一般性，可以假设证明的大小至多为 $T := 2n\text{polylog}\ n$。构造的第一步是创建 T 个变量 $\{x_1, x_2, \cdots, x_T\}$，其中 $x_i \in \Sigma$ 表示提供给验证器的查询响应。我们将为每个 2^r 随机比特串创建一个 SAT 公式，最终的构造只需用“AND”逻辑运算符连接所有 SAT 公式。固定一个任意的随机字符串 $q \in \{0,1\}^r$。接下来，我们可以计算验证器将读取的第一个位置，将其记为 $l_1(q)$。存在一组值的子集 $R(q) \subseteq \Sigma$，验证器会对这些值立即拒绝 $x_{l_1(q)} \in R(q)$。如果验证器没有立即拒绝，就可能根据 $x_{l_1(q)}$ 的值分支执行不同操作。假设 $x_{l_1(q)} = \sigma \in \Sigma \backslash R(q)$。接下来，我们可以计算验证器将读取的第二个位置，记为 $l_2(q,\sigma)$，以及验证器所期望的“正确”字符，记为 $f(q,\sigma)$。于是我们知道，当且仅当由 $x_1, \cdots, x_T$ 表示的证明满足以下条件时，验证器才会接受

$$\bigcup_{\sigma \in \Sigma \backslash R} \left(x_{l_1(q)} = \sigma\right) \wedge \left(x_{l_2(q,\sigma)} = f(q,\sigma)\right).$$

由于 Σ 的规模是常数，很容易看出，可以对 Σ 使用二进制编码，并将上述陈述转化为一个规模为常数的 3-CNF 公式。此外，至多有 n 多对数 n 个二进制字符串 q。因此，整体的 3-CNF 公式 ϕ 的规模为 $O(n$ 多对数 $n)$。我们知道，如果原始 SAT 实例 ψ 不可满足，验证器以至少 ϵ 的概率拒绝。因此，在任何赋值下，ϕ 的至少 ϵ 比例的子公式都无法满足（这可以解释为给定证明的二进制编码）。另一方面，如果 ψ 可满足，那么 ϕ 的每个子公式都可满足，因为验证器在使用“正确”的证明时总是接受。

如果存在一种算法，能够区分在任何赋值下 ϕ 中有 ϵ 比例无法满足的情形与 ϕ 可满足的情形，那么就可以判定 ψ 的可满足性。 □

我们关注 $(b,\epsilon)-$ GAP-3-SAT，它是 ϵ- GAP-3-SAT 的一个受限版本，其中承诺每个变量在某个常数 b 下至多出现在 b 个子句中。通过 [PY91] 中给出的归约，可以证明，近似求解 $(b,\epsilon)-$ GAP-3-SAT 同样是困难的。

命题 A.4（改编自 **[PY91]** 中定理 2 的证明）。对于某个常数整数 $b = O(1)$，存在一个多项式时间变换，将一个 3-CNF 公式 ϕ 映射为另一个定义在相同变量集上的 3-CNF 公式 ψ，使得

1. 在 ψ 中，每个变量至多出现在 b 个子句中。
2. 如果 ϕ可满足，那么 ψ 也可满足。
3. 设 $|\phi|$、$|\psi|$ 分别表示 ϕ、ψ 中的子句数。那么，$|\phi| \leq |\psi| \leq O(1) \cdot |\phi|$。
4. 令 $\max(\phi)$、$\max(\psi)$ 分别表示在 ϕ 和 ψ 中最多可满足的子句数。成立 $\max(\psi) \leq \max(\phi) + |\psi| - |\phi|$。

命题 A.1 的证明。命题 A.4 指出，存在一种高效算法，可将一个包含 m 个子句的 ϵ- GAP-3- SAT 实例转换为一个 $(b, \alpha \cdot \epsilon)$- GAP-3-SAT 实例，其中 $\alpha \in (0, 1)$ 是某个常数。因此，如果对于某个常数 $\epsilon_1 \in (0, 1)$，计算问题 ϵ_1- GAP-3-SAT 不存在亚指数时间算法，那么对于某个常数b 和 $\epsilon_2 \in (0, \epsilon_1)$，$(b, \epsilon_2)$- GAP-3-SAT 也不存在亚指数时间算法。结合定理 A.2 和命题 A.4，即可证明我们的主张。 □

B 上界

在这两个上界中，我们的算法计算出的最终策略具有以下形式：在˜ ˜状态 s 下，对于每个 $a \in \mathcal{A}$，我们都有一些估计值 $Q(s,a)$，使得 $|Q(s,a) - Q^*(s,a)| \le \epsilon$，并且策略˜始终选择动作 $a = \text{argmax}_a Q(s,a)$。我们声称，只要 ϵ 足够小，该策略就是近似最优的。正式表述如下。

˜引理 B.1. 对任意状态-动作对 (s,a)，设 $Q(s,a)$ 是 $Q^*(s,a)$ 的一个近似，且满足˛$|Q(s,a) - Q^*(s,a)| \le \epsilon/(2H)$。接下来，考虑策略 π，该策略在状态 s下始终选择˜动作 $a = \text{argmax}_a Q(s,a)$。那么，对于任意状态，都有 $V^\pi(s) \ge V^*(s) - \epsilon$。

证明。我们声称 π 是一个策略，满足对于马尔可夫决策过程中的任意状态 s，只要还剩下 h 步，就有 $V^\pi(s) \ge V^*(s) - \epsilon \cdot h/H$。我们通过对剩余步数进行归纳法来证明这一点。设 s 是最后一步之前的一个状态。那么，对于 $a^* = \text{argmax}_a Q^*(s,a^*)$，有 $V^*(s) = Q^*(s,a^*)$；并且对于 $a' = \text{argmax}_a Q(s,a)$，有 ˜ ˛$V^\pi(s) = Q^*(s,a')$。根据我们的假设，有 $V^\pi(s) \ge Q(s,a') - \epsilon/(2H)$˜，又由于 $a' = \text{argmax}_a Q(s,a)$，因此我们进一步得到

$$V^\pi(s) \ge \tilde{Q}(s,a^*) - \epsilon/(2H) \ge Q^*(s,a^*) - \epsilon/H = V^*(s) - \epsilon/H.$$

˜现在，考虑一个状态 s，其剩余步数为 $(h+1)$。仍令 $a' = \text{argmax}_a Q(s,a)$ 且$a^* = \text{argmax}_a Q^*(s,a^*)$。此外，令 $s' = P(s,a')$ 表示应用 a' 后的下一状态。于是有$V^*(s) = Q^*(s,a^*)$ 且 $V^\pi(s) = \mathbb{E}[R(s,a')] + V^\pi(s')$。于是有

$$\begin{aligned}
V^\pi(s) &= \mathbb{E}\left[R(s,a')\right] + V^\pi(s') && \text{(策略的定义 } \pi\text{)} \\
&\ge \mathbb{E}\left[R(s,a')\right] + V^*(s') - \epsilon h/H && \text{(归纳假设)} \\
&= Q^*(s,a') - \epsilon h/H && \text{（} Q^*\text{、} V^*\text{的定义）} \\
&\ge \tilde{Q}\text{（}s\text{、}a'\text{）} - \epsilon h/H - \epsilon/\text{（}2H\text{）} && \text{(Assumption about } \tilde{Q}\text{)} \\
&\ge \tilde{Q}\text{（}s\text{、}a^*\text{）} - \epsilon h/H - \epsilon/\text{（}2H\text{）} && \text{(Choice of } a' = \operatorname*{argmax}_a \tilde{Q}(s,a)\text{)} \\
&\ge Q^*\text{（}s\text{、}a^*\text{）} - \epsilon h/H - \epsilon/H && \text{（关于 } \tilde{Q}\text{)} \\
&= V^*\text{（}s\text{）} - \epsilon\text{（}h+1\text{）}/H && \text{（选择 } a^* = \operatorname*{argmax}_a Q^*(s,a)\text{以及 } Q^*, V^*\text{的定义）。}
\end{aligned}$$

由于马尔可夫决策过程中总共有 H 个步骤，因此对于任意状态，这就给出了 $V^\pi(s) \ge V^*(s) - \epsilon$。 □

我们首先证明一个关于计算的上界，该上界关于特征维度 d呈指数级增长。从较高层面来说，我们对 θ^* 可能所在的参数空间进行离散化，以构建策略覆盖集；通过估计每个策略的价值，该覆盖集使我们能够搜索其中的最优策略。

命题 B.2. 假设线性马尔可夫决策过程的动作数量为常数，特征维度为 d，时域为 H。此外，假设对于所有状态-动作对，特征向量满足 $\|\psi(s,a)\|_2 \le 1$，且最优参数 θ^*满足 $\|\theta^*\|_2 \le 1$。令 $\epsilon \in (0,1)$。存在一种算法，运行时间为 $\exp(c \cdot d \cdot \log(Hd/\epsilon))$，其中 c为充分大的常数；该算法能够找到一个策略 π，使得$V^\pi(s) \ge V^*(s) - \varepsilon$，其成立概率为 $9/10$。

证明。令 θ^* 表示最优 Q^* 函数的未知参数，即 $Q^*(s,a) = \langle \theta^*, \psi(s,a) \rangle$。假设我们能够找到满足 $\|\theta - \theta^*\|_2 \le \varepsilon/(2H \cdot \sqrt{d})$ 的此类 $\theta \in \mathbb{R}^d$。注意，这意味着

$$|\langle \theta, \psi(s,a) \rangle - Q^*(s,a)| \le \varepsilon/(2H). \tag{15}$$

然后，考虑策略 $\pi(\theta)$，使得在状态 s 下，它总是选择动作 $a = \text{argmax}_a \langle \theta, \psi(s,a) \rangle$. 根据引理 B.1，对任意状态 s，均有 $V^{\pi(\theta)}(s) \geq V^*(s) - \epsilon$。

现在，令 $S \in \mathbb{R}^d$ 为构成 d 维单位球面的 $\epsilon/(2H\sqrt{d})$ -覆盖的向量集合，即对于任意满足 $\|\theta^*\|_2 \leq 1$ 的 θ^*，都有 $\min_{\theta \in S} \|\theta - \theta^*\|_2 \leq \epsilon/(2H\sqrt{d})$。通过标准的组合构造，对于某个足够大的常数 c，存在这样一个覆盖 S，其大小为 $|S| \leq \exp(c \cdot d \cdot \log(Hd/\epsilon))$。根据上述论证，我们知道必定存在某个 $\theta \in S$，使得 $\pi(\theta)$ 近似最优，即对于任意状态 s，$V^{\pi(\theta)}(s) \geq V^*(s) - \epsilon$。

我们的策略很简单：我们将以蛮力方式尝试所有 $\theta \in S$ 的 $\pi(\theta)$，并将诱导轨迹的期望奖励估计到精度 ϵ。注意，任何轨迹所获得的最大奖励至多为 H。因此，如果我们访问同一轨迹 $\text{poly}(H, 1/\epsilon) \cdot \log(1/\delta)$ 次，就能以至少 $1-\delta$ 的概率将其期望奖励估计到精度 ϵ。我们可以取 $\delta = \frac{1}{|S|}$，这样根据并集界，对于所有 $\theta \in S$，我们对 $V^{\pi(\theta)}(s_0)$ 的估计都能以至少 $9/10$ 的概率达到 ϵ 的误差范围。在此条件下，我们可以选择使 $V^{\pi(\theta)}(s_0)$ 的经验估计最大化的 θ。随后，很容易看出，这样的 $\pi(\theta)$ 必然满足 $V^{\pi(\theta)}(s_0) \geq V^*(s_0) - 2\epsilon$。现在，由于模拟一条轨迹的交互至多需要 $\text{poly}(d, H)$ 的时间，因此总运行时间上界为

$$\text{poly}(d,H) \cdot \text{poly}(H, 1/\epsilon) \cdot \log(1/|S|) \cdot |S| \leq \exp\left(c \cdot d \cdot \log\left(Hd/\epsilon\right)\right)$$

对于某个足够大的常数 c。 □

为了证明时域上界，我们基于先前工作的结果 [DLMW20]。这一上界最初由 Ruosong Wang 亲自告知作者。我们在此给出这一结果，仅为使论述完整。首先，我们概述一下其中的差异。证明几乎完全相同，只是现在我们将马尔可夫决策过程的步骤划分为 $\sqrt{H}$ “轮次”。我们将在这些轮次中通过穷举搜索来寻找最优策略，并使用先前工作 [DLMW20] 中构造的基，确保误差仅增长 $\sqrt{d}$倍。下面我们将更详细地证明这一点。

命题 B.3（Ruosong Wang，个人交流）。假设线性马尔可夫决策过程的动作数量为常数，特征维度为 d，时域为 H。此外，假设对于所有状态-动作对，特征向量满足 $\|\psi(s,a)\|_2 \leq 1$，且最优参数满足 $\|\theta^*\|_2 \leq 1$ θ^*。令 $\epsilon \in (0,1)$。存在一种算法，运行时间为 $\exp\left(c \cdot \sqrt{H} \log d\right)/\epsilon^{-2}$，其中某个常数 c 足够大；该算法能够找到一个策略 π，使得 $V^{\pi}(s) \geq V^*(s) - \varepsilon$。

证明。给定任意状态 s，假设存在一个过程，其运行时间为

$$T := \exp\left(c \cdot \sqrt{H} \log d\right)/\epsilon^{-2} \cdot \log(1/\delta)$$

~ ~，并对每个动作 $a \in \mathcal{A}$ 计算 $Q(s,a)$ 的一个估计值，使得以至少 $1-\delta$ 的概率有 $|Q(s,a) - Q^*(s,a)| \leq \epsilon/(2H)$。那么，我们声称可以设计一种算法，输出一个策略 π，使得对于初始状态 s_0，$V^{\pi}(s_0) \geq V^*(s_0) - \epsilon$ 以至少 $9/10$ 的概率成立。从状态 $s = s_0$ 开始，我们将迭代执行以下步骤：

~1. 对当前状态 s，计算估计值 $Q(s,a)$。

~2. 选择 $a = \text{argmax}_a Q(s,a)$，然后更新 s，使其成为执行动作 a 后的下一状态。

~上述过程至多进行 H 轮次。因此，如果设置 $Q(s, a)$，则我们的估计值在所有轮次中均准确的概率至少为 $9/10$。根据引理B.1，由此可知，从初始状态 s_0开始得到的策略近似最优。此外，算法的运行时间为 $H \cdot O(T)$，符合所需的运行时间。

˜为完成证明，下面我们描述计算状态 s下估计值 $\{Q(s,a)|a\in\mathcal{A}\}$ 的过程。我们只描述初始状态 s_0 下的过程，因为其他状态的估计值也可以类似地计算。为此，我们将马尔可夫决策过程的时间步划分为 $\sqrt{H}$ 轮次。对于每一轮$h\in\sqrt{H}$，我们构造一个向量集合 B_h，它们对应于某个更大的特征向量集合$\psi(s,a)$的“基”，其中 s 是第 $h\cdot\sqrt{H}$ 层的状态，即存在一条轨迹，在 $h\cdot\sqrt{H}$ 个时间步内从 s_0 到达 s。该步骤类似于先前的工作 [DLMW20]，具体如下。令 $B_0=\{\psi(s_0,a):\ a\in A\}$¯，其中 a 是所有动作的集合。然后，我们根据 $B_{h-1}=\{\psi(s_i,a_i)\}$递归地构造 B_h：令 $B_h=$_{$\psi(P(s_i,a_i)\ a):\ a\in A$，以及 $\psi(s_i,a_i)\in B_{h-1}\}$。注意，$|B_h|\le|\mathcal{A}|\cdot|B_{h-1}|$。接下来，我们令 $B_h\subset B_h$¯为 B_h中任意一个由线性无关向量组成的极大子集。注意，此处 $|B_h|\le d$。此外，由于根据假设有 $\|\psi(s,a)\|_2\le 1$，因此任意 $\psi(s,a)\in\bar{B}_h$ 都可以写成 $\sum_{i=1}^{|B_h|}\alpha_i\cdot\psi(s_i,a_i)$，满足 $\|\alpha\|_2\le\sqrt{d}$，其中$\psi(s_i,a_i)$是 B_h中的基向量。

我们声称，对于任意 $h\in[\sqrt{H}]$，我们都可以基于 B_h 学习 Q^*，使其精度达到 $(2d)^{-h}\epsilon$，且最多使用 $O\left(\exp(\sqrt{H}\cdot\log d)\cdot\epsilon^{-2}\right)$ 时间。我们通过对 h 使用归纳法来证明这一点。注意，对于最后一层（距离终止仅 1 步）的任意状态 s 和动作 $a\in\mathcal{A}$，函数 $Q^*(s,a)$ 就是期望奖励 $\mathbb{E}[R(s,a)]$，因为马尔可夫决策过程随后立即终止。因此，我们可以沿着相同的¯轨迹，从 $R(s,a)$ 中多次采样，并计算经验均值 $\bar{R}(s,a)$。假设我们从 $R(s,a)$ 中抽取 $C\cdot\varepsilon^{-2}\log(H\cdot|B_H|/\delta)\cdot(2d)^{\sqrt{H}}$ 个样本，其中常数 C 足够大。由标准集中不等式可知，$|\bar{R}(s,a)-\mathbb{E}[R(s,a)]|\le(2d)^{-\sqrt{H}}\epsilon$ 的概率至少为 $1-\delta/(H\cdot|B_H|)$。由并集界可知，这对所有 $\phi(s,a)\in B_H$ 同时成立的概率至少为$1-\delta/H$。因此，对于所有 $\phi(s,a)\in B_{\sqrt{H}}$，我们都可以为 $Q^*(s,a)$ 计算一个估计量，其精度为 $(2d)^{-\sqrt{H}}$，所需时间至多为

$$\left(C\cdot\varepsilon^{-2}\log(H\cdot|B_H|/\delta)\cdot(2d)^{\sqrt{H}}\right)\cdot|B_H|\cdot\mathrm{poly}(d,H)\le\exp\left(c\cdot\sqrt{H}\cdot\log d\right)\epsilon^{-2}\log(1/\delta)$$

对于某个足够大的常数 c。

假设我们已经在基 B_h 上以精度 $d^{-h}\epsilon$学习了 Q^*。下面我们将看到如何利用这些信息，在基 B_{h-1} 上以精度 $d^{-h+1}\epsilon$ 估计 Q^*。仍然考虑一个状态 - 动作对 (s,a)，满足 $\phi(s,a)\in B_{h-1}$。令 $\mathcal{R}_{s,a,\sqrt{H}}$为在第一步为 a的条件下，从 s出发经过至多 $\sqrt{H}$个¯步可到达的状态集合（注意，根据我们的构造，$\mathcal{R}_{s,a,\sqrt{H}}$是 B_h的子集）。换言之，$\mathcal{R}_{s,a,\sqrt{H}}$中的每个状态 s'都是第 $(h\cdot\sqrt{H})$ 层中的一个状态，且存在一条从 s到 s'、以动作 a开头的轨迹。不失一般性，我们假设每个状态 s' 都有一条从 s出发的唯一轨迹：如果有两条不同的轨迹到达同一个状态 s'，我们可以创建两个 s' 的副本，并根据到达它们的唯一轨迹为其编号。

我们知道，必然存在某个状态 $s^*\in\mathcal{R}_{s、a、\sqrt{H}}$和 $a^*\in\mathcal{A}$，使得 $Q^*(s,a)$ 等于从 s到 s^*的轨迹以及 $Q(s^*,a^*)$ 所收集的期望奖励之和。我们将用 $\kappa(s,s')$ 表示从 s 到 s'、针对 $s'\in\mathcal{R}_{s、a、\sqrt{H}}$的路径所收集的期望奖励。我们的目标是计算：(i) 对每个 $\kappa(s,s')$ 的估计，其中 $s'\in\mathcal{R}_{s、a、\sqrt{H}}$；以及 (ii) 对每个 $Q^*(s',a')$ 的估计，其中$s'\in\mathcal{R}_{s、a、\sqrt{H}}$和 $a'\in A$。很容易看出，如果我们能够以精度 $d^{-h+1}\cdot 2^{-h}$ 计算 (i) 和 (ii)，那么便可取 s'、a' 的最优组合，从而得到对 $Q^*(s,a)$ 的估计，其精度为 $(2d)^{-h+1}$。

为了估计 $\kappa(s,s')$（即从一条轨迹中获得的期望奖励），我们只需多次遍历该轨迹。由于 $\kappa(s,s')\le\sqrt{H}$，根据标准集中不等式可知，如果我们遍历该轨迹

$$C\cdot H\cdot\varepsilon^{-2}\log(H\cdot|\mathcal{R}_{s,a,\sqrt{H}}|/\delta)\cdot(2d)^h$$

许多次，其中 C 是足够大的常数，那么我们就能以高概率将所有 $\kappa(s,s')$ 估计到所需的精度。$\mathcal{R}_{s,a,\sqrt{H}}$ 中至多有 $|\mathcal{A}|^{\sqrt{H}}=\exp(\log|\mathcal{A}|\cdot\sqrt{H})$ 个状态。

由于我们访问一条轨迹 $C \cdot H \cdot \varepsilon^{-2} \log(H \cdot |\mathcal{R}_{s,a,\sqrt{H}}|/\delta) \cdot (2d)^h$ 次，估计每个 $\kappa(s, s')$ 所需的时间至多为

$$\begin{aligned}
&\exp(\log|\mathcal{A}| \cdot \sqrt{H}) \cdot \left(C \cdot H \cdot \varepsilon^{-2} \log(H \cdot |\mathcal{R}_{s,a,\sqrt{H}}|/\delta) \cdot (2d)^h\right) \cdot \mathrm{poly}(d, H) \\
&\leq \exp\left(c \cdot \sqrt{H} \cdot \log d\right) \epsilon^{-2} \log(1/\delta)
\end{aligned}$$

对于某个足够大的常数 c。

为了估计 $Q^*(s', a')$，其中 $s' \in \mathcal{R}_{s,\ a,\ \sqrt{H}}$ 和 $a' \in A$，我们将利用这样一个事实：我们已经在 B_h 中获得了基上 Q^* 的估计值。具体而言，我们可以将 $\psi(s', a')$ 表示为 $=\sum_{i=1}^{|B_h|} \alpha_i \cdot \psi(s_i, a_i)$，其中 $\psi(s_i, a_i)$ 是 B_h 中的基。由线性性可得

$$Q^*(s', a') = \sum_{i=1}^{|B_h|} \alpha_i \cdot Q^*(s_i, a_i). \tag{16}$$

一方面，我们有 $\|\alpha\|_2 \leq \sqrt{d}$。另一方面，根据归纳假设，对于每个 $Q^*(s_i, a_i)$，我们都有精度达到 $(2d)^{-h} \cdot \epsilon$ 的估计值。如果我们直接将对 $Q^*(s_i, a_i)$ 的估计代入方程 (16)，以计算对 $Q^*(s', a')$ 的估计，那么根据柯西-施瓦茨不等式，误差至多为 $d \cdot (2d)^{-h} \cdot \epsilon \leq 2^{-h} \cdot d^{-h+1} \cdot \epsilon$，这正是所需的界。计算一个 $Q^*(s', a')$ 的估计值需要 $\mathrm{poly}(d)$ 时间。由于 (s', a') 至多有 $\exp(\log\ |\mathcal{A}| \cdot \sqrt{H})\ \cdot |\mathcal{A}|$ 个这样的配对，这部分所需时间至多为 $\exp\left(c \cdot \left(\sqrt{H} + \log\ d\right)\right)$，其中 c 是某个足够大的常数。

通过归纳法，这就为我们提供了一种方法，可以以高概率将 Q^* on $B_0 = \{\phi(s_0, a)$ 近似到精度 ϵ，其中 $a \in A\}$。此外，整个过程的运行时间为 $\exp\left(c \cdot \sqrt{H} \cdot \log\ d\right) \epsilon^{-2} \log(1/\delta)$，其中 c 是某个足够大的常数。

□