具有子模奖励的多智能体强化学习

Wenjing Chen[1] Chengyuan Qian[1] ShuoXing[1] Yi Zhou[1]Victoria G.Crawford[11]计算机科学与工程系，德克萨斯农工大学 {jj9754, cyqian, shuoxing, yi.zhou, vcrawford}@tamu.edu

arXiv:2603.06810v1 [cs.LG] 6 Mar 2026

摘要

本文研究联合奖励具有子模性的合作式多智能体强化学习（MARL）。子模性是一种自然属性，用于刻画向团队中增加智能体时边际收益递减的现象。不同于采用加性奖励的标准多智能体强化学习，子模奖励能够建模智能体贡献存在重叠的现实场景（例如，多无人机监视和协作探索）。我们首次为这一设置提供了形式化框架，并开发了在样本效率和遗憾界方面具有可证明保证的算法。对于已知动力学，我们的贪心策略优化以多项式复杂度在智能体数量 K 下实现了 $1/2$ 近似，克服了联合策略优化中固有的指数级维数灾难。对于未知动力学，我们提出了一种基于UCB的学习算法，在 T 个回合内实现了 $1/2$-遗憾 $O(H^2KS\sqrt{AT})$。

1 引言

合作式多智能体强化学习（多智能体强化学习）已成为在复杂环境中协调交互式智能体团队的有力范式，其应用范围从机器人系统 [Hüttenrauch等人，2017年，Ismail等人， 2018] 到网络优化 [Zhang和Lesser， 2011] 以及资源分配 [Rauch等人，2024年，Tilahun等人， 2023]。在合作式多智能体强化学习中，智能体在随机环境中协调其动作，以最大化期望累积奖励，该过程被形式化为多智能体马尔可夫决策过程（MDP）。

一个常见的简化假设是，在每个时间步观测到的联合奖励是单个智能体贡献的线性（加性）函数 [Li等人，2021年，Hsu等人， 2024]。虽然该假设便于分解与分析，但它从根本上限制了奖励模型的表达能力。在许多现实世界的协作任务中，智能体对全局目标的贡献关键取决于其他智能体的动作，从而产生重叠、冗余或饱和等交互效应，而这些效应无法被加性奖励结构充分刻画。

这类现象自然会出现在各种协作决策问题中。例如，在多机器人地图探索中，一组机器人探索未知环境，以构建全面的地图，并以最大化信息增益为目标 [Ma 等， 2020]。每个机器人获取的信息可能与其他机器人的信息重叠；随着投入搜索任务的机器人增多，信息增益还会呈现边际收益递减的特性。同样，在多无人机监视与跟踪中，一组无人机监视大片地理区域，目标是最大化受监视的不同目标数量 [Zhang 和 Lesser， 2011]。在此过程中，不同无人机实现的覆盖范围往往会相互重叠，随着分配给任务的智能体增多，同样会导致边际收益递减。若忽略这类重叠效应，智能体可能会收敛到冗余行为，从而妨碍高效探索并削弱有效合作（见图1）。

为了对这类设定进行建模，我们研究一类自然的合作多智能体强化学习问题，其中联合奖励表现出子模性，这是组合优化中的一个基本性质，刻画了边际收益递减 [Hassidim and Singer, 2017年, Iyer et al., 2021]。子模函数在机器学习与优化中无处不在 Chen 等人 [2025a], Yue 和 Guestrin [2011], , 涵盖集合覆盖、熵、互信息以及设施选址目标等重要情形。这一性质自然地描述了合作多智能体强化学习任务：智能体的贡献存在重叠、智能体行为的多样性很重要，或者覆盖范围增益趋于饱和，并且将一个智能体加入较小团队所带来的边际收益至少与将其加入较大团队一样大 [Liet al., 2021]。值得注意的是，多无人机监视中的覆盖范围目标函数是单调次模的 [Zhang and Lesser, 2011], , 而协作探索中的信息增益可以通过熵或互信息来量化，二者都是子模函数 [Ma et al., 2020]。

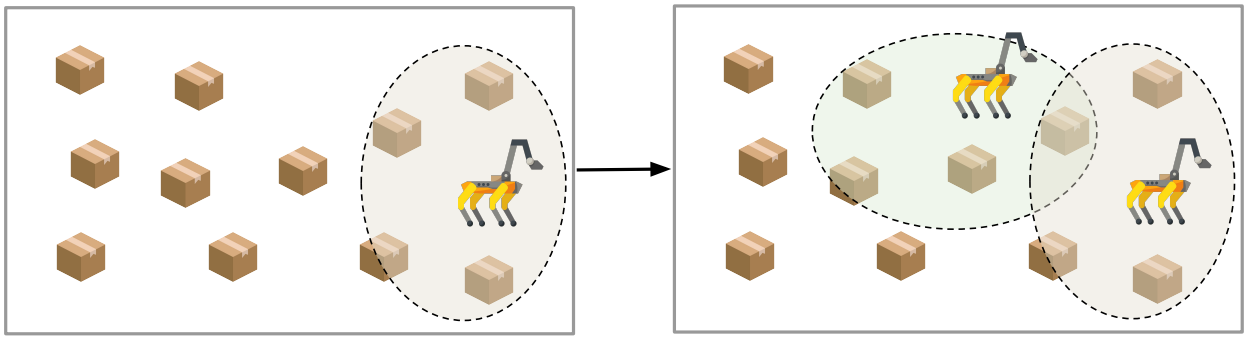

图1：多智能体协作示例。

基于这些考虑，我们提出具有子模奖励的多智能体强化学习（MARLS），这是一种新颖的协作式多智能体强化学习框架，其中智能体在每个时间步共同影响一个单调次模奖励函数。我们考虑具有共享状态与动作空间的回合制设定，目标是最大化总期望累积回报。尽管单调性与子模性带来了结构上的规律性，MARLS 仍带来显著的算法与计算挑战。首先，即便在单调性与子模性假设下，在 MARLS 中寻找最优策略在计算上仍然难以处理：正如我们在第4.1.1节中所说明的，单步情形可归约为分区拟阵约束下的子模最大化这一 NP难问题。其次，尽管经典的贝尔曼最优性方程（引理 2） [Bellman, 1957, Sutton等人，1998] 为最优策略提供了递归刻画，直接评估给定随机策略的价值函数所需的时间与内存随智能体数量呈指数增长 K。最后，尽管子模最大化在静态设定中已得到广泛研究，将其扩展到序贯决策设定会带来实质性困难，这些困难源于随机转移动态，以及每个智能体的边际奖励贡献依赖于先前已优化智能体的策略与随机性。当转移动态未知时，这一困难会进一步加剧。

具体而言，本文的贡献如下：
1. 我们在第3.1节提出具有子模奖励的多智能体强化学习（MARLS）设定。我们刻画了多智能体强化学习固有的计算挑战，并证明即使是单步 MARLS 问题也是 NP 困难的（第4.1.1节）。为应对这些挑战，我们考虑因子化策略，并提出边际价值分解方法。这促使我们为 MARLS 开发易于处理的近似算法。

2. 针对转移动态已知的情况，我们在第4.2节提出贪心策略优化。贪心策略优化利用奖励函数的单调性和子模性，为 MARLS 提供了一种易于处理的解决方案。我们的关键见解是，子模结构能够支持高效的贪心序贯优化，从而找到一种可分解策略；这种策略具有多项式规模表示，并支持分布式执行（见第4.1.2节）。尽管存在这一结构性限制，我们证明，贪心策略优化相对于最优（可能不可分解的）联合策略，能够达到 $1/2$ 近似保证。

3. 针对转移动态未知的情况，我们在第 4.3 节提出并分析了一种基于上置信界的强化学习算法 UCB-GVI，该算法将乐观探索与贪心子模最大化相结合。UCB-GVI 在 T 个回合内实现了 $O(S^2AH^3K^2\log$ —— $T + H^2KS\sqrt{AT})$ 的$_{1\,2}$-近似遗憾（定理 2），¯这是针对 MARLS 的首个次线性遗憾保证。我们在第 4.3 节介绍了分析中的主要技术创新和贡献。

2 相关工作

2.1 多智能体强化学习

合作式多智能体强化学习（MARL）的一项根本挑战是，联合状态-动作空间会随着智能体数量的增加而呈指数级增长 K，这往往使传统的单智能体强化学习（RL）方法难以实际应用。为应对这种复杂性，通常采用去中心化策略 [Sukhbaatar et al., 2016, Foerster 等，2016, Chen 等人， 2022], 其中每个智能体仅根据自身的局部动作-观测历史来选择动作。尽管策略执行过程是

去中心化的，训练过程仍可采用集中式方法。去中心化方法着重于提升通信效率和学习协调协议。另一类工作则通过价值分解方法应对这一可扩展性挑战，这些方法对联合动作价值函数施加结构约束。 [Sunehag et al., 2017, Rashid et al., 2020, Ma et al., 2021, Dou et al., 2022] 通过施加结构约束来解决可扩展性问题，使得各智能体动作价值上的局部最优对应于联合动作价值上的全局最优。

多智能体强化学习领域的一些已有研究通过鼓励智能体之间采取多样化行为，解决了现实世界协作任务中捕捉依赖结构的需求 [Mahajan et al., 2019, Wang et al., 2020, Li et al., 2021, Jiang and Lu, 2021]。Li 等人提出了一种新颖的信息论目标函数，该函数基于智能体身份与轨迹之间的互信息，是我们方法的一个特例。然而，大多数现有方法都隐含地假设智能体贡献以简单的方式组合，要么采用加法组合，要么通过单调混合进行组合，并未利用具有子模性的奖励函数所固有的结构。

2.2 拟阵约束下的子模最大化

带拟阵约束的单调子模最大化（MSMM）是一个 NP 困难的组合优化问题[1]，也是子模优化中研究最广泛的问题之一 [Feldman et al., 2025, Calinescu et al., 2011, Banihashem et al., 2024]。鉴于 MSMM 的计算复杂性，研究工作主要致力于开发和分析近似算法。标准贪心算法可达到 $\frac{1}{2}$-近似比 [Fisher et al., 1978]，而采用连续扩展（多线性松弛）的更复杂算法则可达到 $\left(1-\frac{1}{e}\right)$-近似 [Calinescu et al., 2007, Badanidiyuru and Vondrák, 2014]。然而，与相对简单的贪心方法相比，后者通常会带来显著更高的查询复杂度。如第 3 节所述，MSMM 固有的难度在很大程度上造成了本文所研究场景中的计算挑战。具体而言，当马尔可夫决策过程被限制为单步时，我们的问题可归约为 MSMM。

其他相关研究方向包括具有子模奖励结构的组合多臂赌博机，以及具有可证明保证和遗憾分析的强化学习方法。附录 A 对这些工作进行了更详细的讨论。

3 预备知识

在本节中，我们介绍具有子模奖励的多智能体强化学习（MARLS）的背景与形式化。我们首先给出关键符号，然后呈现核心问题的形式化。

3.1 具有子模奖励的多智能体马尔可夫决策过程

我们考虑一个包含 K 个智能体的合作式回合制多智能体强化学习设定。环境被建模为表格型马尔可夫决策过程（MDP），其中智能体在每个回合内通过 H 个离散时间步进行交互。智能体共享相同的局部状态空间 $\mathcal{S}$ 和动作空间 $\mathcal{A}$。它们的交互由一个具有子模 全局奖励函数的多智能体马尔可夫决策过程建模。形式上，我们将具有子模奖励的多智能体马尔可夫决策过程（MAMDP-SR）定义如下：

定义 1（具有子模奖励的多智能体马尔可夫决策过程（MAMDP-SR））。*MAMDP-SR* 由元组 $M=(\mathcal{S}, \mathcal{A}, \{P_{i,h}\}_{i\in[K], h\in[H]}, r, K, H)$ 表征，其中：

1. H 是回合长度，*2.* K 是智能体数量，*3.* $\mathcal{S}$ 是所有智能体共有的有限状态空间，*4.* $\mathcal{A}$ 是所有智能体共有的有限动作空间，*5.* $P_{i,h}$：$\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$ 是智能体 i 在时间步 h 的转移概率函数，其中 $P_{i,h}(s' \mid s,a)$ 表示智能体 i 在第 h 步处于状态 s 并采取动作 a 后转移到状态 s' 的概率。*6.* r：$\mathcal{S}\times\mathcal{A})^K\to[0,1]$ 是依赖于所有智能体的联合状态空间和动作空间的全局奖励函数。此外，假设 r 是单调且子模的（见第*3.3*节）。

[1]已证明，除非 NP 存在 $n^{\mathcal{O}(\log(\log(n)))}$-时间确定性算法，否则 MSMM 是 NP 困难的 [Wolsey, 1982, Feige, 1998]。

3.2 回合动态

在回合开始时（$h=1$），智能体被初始化于联合状态

$$\bar{\mathbf{s}}_1 = (\bar{s}_1^1, \ldots, \bar{s}_1^K).$$

然后，在每个时间步 $h \in [H]$，每个智能体 $i \in [K]$ 都会观察其状态 $s_h^i \in \mathcal{S}$，并从动作空间 $\mathcal{A}$中选择一个动作 a_h^i。我们将联合状态向量和动作向量分别记为 $\mathbf{s}_h = (s_h^1, \ldots, s_h^K)$ 和 $\mathbf{a}_h = (a_h^1, \ldots, a_h^K)$，其中 $s_h^i \in \mathcal{S}$ 是第 i 个智能体的状态，而 a_h^i 是智能体 i 在时间步 h选择的动作。随后，环境会发出全局奖励 $r(\mathbf{s}_h, \mathbf{a}_h)$，并且每个智能体都会独立转移到其下一个状态 s_{h+1}^i；该状态在时间步 $h+1$从分布 $P_{i,h}(\cdot|s_h^i, a_h^i)$中抽取。为简化符号，我们用 $P_h(\mathbf{s}_{h+1}|\mathbf{s}_h, \mathbf{a}_h)$ 表示在时间步 h从联合状态-动作对 $(\mathbf{s}_h, \mathbf{a}_h)$ 转移到联合状态 $\mathbf{s}_{h+1}$ 的概率，并用 P 表示所有智能体在所有时间步的整体转移模型。我们考虑每个智能体的转移动态彼此独立的情形。也就是说，$P_h(\mathbf{s}_{h+1}|\mathbf{s}_h, \mathbf{a}_h) = \prod_{i=1}^K P_{i,h}(s_{h+1}^i|s_h^i, a_h^i)$，其中 $s_h^i \in \mathcal{S}$是第 i 个智能体的状态，而 a_h^i 是智能体 i 在时间步 h选择的动作。

步 h处的（随机）策略是一个映射 π_h： $\mathcal{S}^K \times \mathcal{A}^K \to [0,1]$ 其中，$\pi_h(\mathbf{s}、\mathbf{a})$ 表示在联合状态 $\mathbf{s}$ 下采取联合动作 $\mathbf{a}$ 的概率。因此，对于任意 $h \in [H]$ 和 $\mathbf{s} \in \mathcal{S}^K$，可得 $\sum_{\mathbf{a}:\ \mathbf{a}\in\mathcal{A}^K} \pi_h(\mathbf{s}、\mathbf{a}) = 1$。为简化记号，我们使用 $\pi := \{\pi_h\}_{h\in[H]}$ 表示所有时间步上的策略。

3.3 子模奖励

我们模型的关键特征在于，全局奖励并非在智能体之间具有加和性，而是体现了边际收益递减——这一性质由子模性刻画。形式化地，我们假设奖励由一个单调子模集合函数 f： $2^{\mathcal{S}\times\mathcal{A}} \to [0,1]$ 决定，该函数作用于当前智能体状态-动作对的集合：

$$r(\mathbf{s}, \mathbf{a}) = f\big(\{(s^1, a^1), (s^2, a^2), \ldots, (s^K, a^K)\}\big). \tag{1}$$

其中，$\mathbf{s} = (s^1、\ s^2、\ \ldots\ldots、\ s^K)$ 是联合状态，而 $\mathbf{a} = (a^1、\ a^2、\ \ldots\ldots、\ a^K)$ 是联合动作。这里我们使用记号$\{x^1, x^2, \ldots x^n\}$ 表示 $x^1、\ x^2、\ \ldots\ldots、\ x^n$ 中各元素组成的集合。该定义表明，奖励由每个智能体的状态-动作对决定。此外，我们假设对于任意状态-动作对的子集，都存在预言机访问 f。即，对于任意 $S \in 2^{\mathcal{S}\times\mathcal{A}}$，我们都可以访问 $f(S)$ 的预言机。

假设 1（单调次模奖励）。 我们假设，对于状态-动作对 $A \subseteq B \subseteq 2^{\mathcal{S}\times\mathcal{A}}$ 的任意子集，以及满足 $x \in 2^{\mathcal{S}\times\mathcal{A}}/B$ 的状态-动作对 x，都有

$$f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B)$$
$$f(A) \leq f(B).$$

我们将向状态-动作对集合 A 添加新的状态-动作对 x 所产生的边际增益定义为 $\Delta f(A,\ x)$，即 $\Delta f(A,\ x) = f(A \cup x) - f(A)$。

接下来，我们定义问题的目标，即最大化一个回合中获得的总奖励。为此，我们将价值函数 V 和 Q 定义如下。给定 $\mathbf{s}_h = \mathbf{s}$，且$\mathbf{a}_h = \mathbf{a}$ 时，$V_h^\pi(\mathbf{s},\ \mathbf{a})$ 被定义为从步 h 到最终时间步 H 的期望总奖励。

$$V_h^\pi(\mathbf{s}) = \mathbb{E}_\pi\Big[\sum_{t=h}^H r(\mathbf{s}_t, \mathbf{a}_t)|\mathbf{s}_h = \mathbf{s}\Big].$$

动作价值函数 Q_h^π： $\mathcal{S}^K \times \mathcal{A}^K \to \mathbb{R}$ 的定义类似，但条件是在采取联合动作 $\mathbf{a}_h = \mathbf{a}$ 的情况下：

$$Q_h^\pi(\mathbf{s}, \mathbf{a}) := \mathbb{E}_\pi\Big[\sum_{t=h}^H r(\mathbf{s}_t, \mathbf{a}_t)\ \Big|\ \mathbf{s}_h = \mathbf{s}, \mathbf{a}_h = \mathbf{a}\Big].$$

继续　是找到一个最优策略 π^*，使其最大化从初始状态出发的期望总奖励　　e:

$$\pi^* \in \text{argmax}_\pi\ V_1^\pi(\bar{\mathbf{s}}_1).$$

接下来，我们以多机器人协同与跟踪为例，说明上述问题设定（Xu 等人 [2023]）。在此设定中，由 K 架无人机组成的团队协同作业，每架无人机都被建模为一个独立的智能体。目标是确定这些无人机的全局策略，以尽可能多地捕获物体。状态空间 $\mathcal{S}$ 是无人机的空间位置，动作空间 $\mathcal{A}$规定了每架无人机可执行的移动动作。因此，不同无人机的转移动态相互独立。时刻 t 的奖励是无人机团队覆盖的不同物体数量。若存在 M 个物体，我们可以定义 $r_t(\mathbf{s},\ \mathbf{a}) = \sum_{i=1}^{M} \mathbb{I}$（智能体在状态 $\mathbf{s}$ 下捕获了物体 i），其中 $\mathbb{I}$（智能体在状态 $\mathbf{s}$ 下捕获了物体 i）是一个指示函数，用于判断物体 i 是否在联合状态 $\mathbf{s}$ 下被无人机覆盖。这是一个有界于 $[0, M]$ 的次模集合覆盖函数。[2]

4 主要结果

本节介绍我们针对 MARLS 的主要算法和理论贡献。我们首先阐明该问题固有的计算障碍，然后针对两种设置提出基于序贯贪心策略优化的可处理解决方案，并利用子模性：规划设置（转移动态已知）和在线学习设置（转移动态未知）。

4.1 计算挑战与价值函数分解

4.1.1 计算挑战

多智能体马尔可夫决策过程的经典贝尔曼最优性方程为一般策略和最优策略提供了递归刻画（参见附录 B.1 中的引理 2）。然而，直接求解贝尔曼方程和最优策略（如引理 2 的公式 (3) 和 (4) 所述）在计算上不可行：对于任意给定的随机策略 π，按照公式 (3) 计算其价值函数 V 和 Q 需要 $O(|\mathcal{S}|^{2K}|\mathcal{A}|^{K}H)$ 的时间和 $O(|\mathcal{S}|^{K}|\mathcal{A}|^{K}H)$ 的内存，二者均随智能体数量 K 呈指数增长。这种指数级复杂度源于两个根本性挑战：

1. 一个通用的联合policy $\pi = \{\pi_h(\mathbf{s}, \mathbf{a})\}_{h\in[H]}$ requires $O(|\mathcal{S}|^{K}|\mathcal{A}|^{K}H)$ space to store, which is exponential in K
2. 贝尔曼方程更新还需要计算复杂度和内存复杂度，而这两者相对于智能体数量 K 都呈指数级增长。

为规避这一维数灾难，我们利用奖励函数的子模结构。我们的目标是提出一种算法，该算法输出具有较高期望总奖励的策略，并且其计算复杂度和内存复杂度相对于 K均为多项式复杂度。

然而，下面我们将说明，即使在子模性假设下，在 MARLS 中以相对于 K 为多项式复杂度寻找最优策略仍然在计算上不可行。为说明这一点，考虑单步情形（$H = 1$）。价值函数可化简为

$$\begin{aligned} Q_1^{\pi}(\mathbf{s}, \mathbf{a}) &= \mathbb{E}_{\pi}[r(\mathbf{s}_1, \mathbf{a}_1)|\mathbf{s}_1 = \mathbf{s}, \mathbf{a}_1 = \mathbf{a}] \\ &= r(\mathbf{s}, \mathbf{a}). \end{aligned}$$

根据贝尔曼最优性方程，最优策略满足

$$\begin{aligned} \pi_1^*(\bar{\mathbf{s}}_1) &= \arg\max_{\mathbf{a}\in\mathcal{A}^K} r(\bar{\mathbf{s}}_1, \mathbf{a}) \\ &= \arg\max_{a^1,\ldots,a^K} f\big(\{(\bar{s}^1, a^1), \ldots, (\bar{s}^K, a^K)\}\big). \end{aligned}$$

这等价于拟阵约束下的子模最大化问题，其中每个智能体的动作空间构成一个分区，而该约束要求每个分区至多选择一个动作（见附录 B.2）。由于分区拟阵是一般拟阵约束的一种特殊情形，并且拟阵约束下的子模最大化是 NP 困难的，除非 NP $\subseteq n^{\mathcal{O}(\log\log n)}$-时间 [Feige, 1998], 即使对于 $H = 1$，在 MARLS 中寻找最优策略在计算上也是不可行的。

幸运的是，针对单调子模最大化的经典贪心算法能够实现 $1/2$-近似[Wolsey, 1982]。受此启发，我们提出了一种贪心序贯策略优化方法，利用子模结构实现多项式复杂度，并获得可证明的近似保证。

[2]在本篇 在本文中，我们假设函数 f 在 $[0, 1]$, 中有界，并且通过缩放，我们的结果可以扩展到 $[0, M]$ 情形。

4.1.2 联合策略与奖励分解

为克服图 1 所示的指数级内存复杂度这一挑战，我们定义另一类可因式分解并可分解为局部智能体策略的策略，这类策略也称为可分解策略：

定义2。 若存在局部策略 $\{\pi_{i,h}\}_{i\in[K],h\in[H]}$，使得策略 π 可分解，则称其为可分解策略。

$$\pi_h(\mathbf{s},\mathbf{a}) = \prod_i \pi_{i,h}(s^i,a^i).$$

这里，$\pi_{i,h}(s,a)$ 表示在时间步 h，第 i 个智能体在状态 s 下选择动作 a 的概率，其中 $\sum_a \pi_{i,h}(s,a) = 1$ 对任意 $i \in [K]$、 $h \in [H]$和 $s \in \mathcal{S}$均成立。

沿用前述约定，我们用 $\pi_i := \{\pi_{i,h}\}_{h\in[H]}$ 表示第 i 个智能体在所有时间步上的策略。智能体的联合策略由 $[i] = \{1,\dots,i\}$ 编号，记作 $\pi_{[i]}$。更具体地说，$\pi_{[i],h}(s^{[i]},a^{[i]}) = \prod_{l=1}^{i}\pi_{l,h}(s^l,a^l)$，其中 $s^{[i]} = (s^1,s^2,...,s^i)$ $a^{[i]} = (a^1,a^2,...a^i)$ 表示部分状态向量和动作向量。注意，可分解策略具有多项式规模表示（仅需 $O(K|\mathcal{S}||\mathcal{A}|H)$ 内存），并支持分布式执行，因为每个智能体都可以根据其局部状态独立选择动作。然而，限制为可分解策略会引入近似误差，因为最优策略 π^* 可能不可分解。

为了能够有效地优化全局奖励，我们将子模函数 f 分解为边际贡献。定义由索引集合 $[i]$ 中的智能体在状态 $\mathbf{s}$ 和动作 $\mathbf{a}$ 下产生的部分奖励为

$$r(\mathbf{s},\mathbf{a};[i]) \triangleq f(\{(s^l,a^l)\}_{l\in[i]}),$$

然后，添加智能体 i所带来的边际增益为

$$\Delta r_i(\mathbf{s},\mathbf{a}) \triangleq r(\mathbf{s},\mathbf{a};[i]) - r(\mathbf{s},\mathbf{a};[i-1]).$$

接着，通过望远镜求和，总奖励可分解为

$$r(\mathbf{s},\mathbf{a}) = \sum_{i=1}^{K}\Delta r_i(\mathbf{s},\mathbf{a}).$$

4.1.3 边际价值与 Q-函数

给定前 $i-1$ 个智能体的固定策略 $\pi_{[i-1]} = \{\pi_{j,h}\}_{j\in[i-1],h\in[H]}$ 以及初始状态 $\bar{\mathbf{s}}_1$, , 前 $i-1$ 个智能体的状态-动作轨迹分布 $\{(s_h^j,a_h^j)\}_{h\in[H],j\in[i-1]}$ 便确定了。因此，添加第 i个智能体所带来的期望边际增益也随之固定。从这个角度看，智能体 i 可以被视为在由智能体 $1,\dots,i-1$的固定行为所诱导的环境中运行，其奖励对应于相对于现有团队所贡献的边际增益。因此，固定前 $i-1$ 个智能体的策略，就为第 i个智能体诱导出一个定义明确的奖励函数。对于给定策略 $\pi_{[i-1]}$，我们将智能体 i 在状态 s 下于第 h 步执行动作 a 时的期望边际奖励 定义为

$$R_{i,h}(s,a \mid \pi_{[i-1]}) = \mathbb{E}\Big[\Delta r_i(\mathbf{s}_h,\mathbf{a}_h)\Big| s_h^i = s, a_h^i = a, \pi_{[i-1]}\Big], \quad (2)$$

其中，期望是针对由$\pi_{[i-1]}$诱导的前 $i-1$ 个智能体的状态-动作对分布（$s_h^{[i-1]}$， $a_h^{[i-1]}$）计算的。对于前 $i-1$ 个智能体的固定策略$\pi_{[i-1]}$，我们将智能体 i采用策略 π_i 时的边际价值函数定义为：

$$V_h^{\pi_i}(s;\pi_{[i-1]}) = \mathbb{E}_{\pi_i}\Big[\sum_{t=h}^{H}\Delta r_i(\mathbf{s}_t,\mathbf{a}_t)|s_h^i = s,\pi_{[i-1]}\Big]$$

$$Q_h^{\pi_i}(s,a;\pi_{[i-1]}) = \mathbb{E}_{\pi_i}\Big[\sum_{t=h}^{H}\Delta r_i(\mathbf{s}_t,\mathbf{a}_t)|s_h^i = s,a_h^i = a,\pi_{[i-1]}\Big].$$

[3]请注意， π_i 和 π_h 可能具有相似的形式。下标用于区分它们的含义：字母 i 或 j 表示智能体在所有步上的策略，而 h 表示某一步的联合策略。

这些边际价值函数衡量将智能体 i 加入团队 $[i-1]$后所贡献的预期总边际奖励。关键的是，以下引理证明，一旦前序智能体的策略固定，智能体 i就面临一个具有时变奖励$R_{i,\ h}(s,\ a \mid \pi_{[i-1]})$的单智能体马尔可夫决策过程。

Le引理 1（边际收益的贝尔曼方程）。 对于任意策略 π_i， $h \in [H]$， $s \in \mathcal{S}$，以及 $a \in \mathcal{A}$，

$$\begin{aligned} V_h^{\pi_i}(s;\pi_{[i-1]}) &= \sum_{a\in\mathcal{A}} \pi_{i,h}(s,a) Q_h^{\pi_i}(s,a;\pi_{[i-1]}) \\ Q_h^{\pi_i}(s,a;\pi_{[i-1]}) &= R_{i,h}(s,a|\pi_{[i-1]}) \\ &\quad + \sum_{s'\in\mathcal{S}} P_{i,h}(s'|s,a) V_{h+1}^{\pi_i}(s';\pi_{[i-1]}). \end{aligned}$$

此外，令 $\widetilde{\pi}_i = \arg\max_{\pi_i} \mathbb{E}_{\pi_i}[\sum_{t=1}^{H} \Delta r_i\ (\mathbf{s}_t,\ \mathbf{a}_t)\ |s_1^i = \bar{s}_1^i,\ \pi_{[i-1]}]$。那么

$$\begin{aligned} V_h^{\widetilde{\pi}_i}(s;\pi_{[i-1]}) &= \max_a Q_h^{\widetilde{\pi}_i}(s,a;\pi_{[i-1]}) \\ Q_h^{\widetilde{\pi}_i}(s,a;\pi_{[i-1]}) &= R_{i,h}(s,a|\pi_{[i-1]}) \\ &\quad + \sum_{s'\in\mathcal{S}} P_{i,h}(s'|s,a) V_{h+1}^{\widetilde{\pi}_i}(s';\pi_{[i-1]}). \end{aligned}$$

证明概要推迟至附录 B.3。接下来，我们讨论转移矩阵 P已知时的策略优化算法。

4.2 已知 P 时的贪心策略优化

在转移动力学 P 已知的情况下，我们提出算法“贪心策略优化”，该算法通过贪心地执行策略优化来实现，如算法 1 所述。该算法以迭代、贪心的方式确定各智能体的策略。在从第 2 行到第 20 行的外循环中，算法利用奖励函数的次模性质，依次贪心地确定不同智能体的策略。在从第 3 行到第 16 行的内循环中，算法通过从第 H 步到第 1 步进行逆向归纳，递归地优化策略，从而求出智能体 i 的策略。确定第 i 个智能体的策略后，算法在第 17 行至第 19 行从该智能体采样 $N := \frac{1}{2\epsilon^2}\log\frac{2K|\mathcal{S}||\mathcal{A}|^H}{\delta}$ 条轨迹，用于估计第 8 行中的边际奖励 $R_{i,\ h}(s,\ a|\pi_{[i-1]})$。这一步对于实现关于 K 的多项式计算复杂度非常重要，因为精确计算$R_{i,\ h}(s,\ a|\pi_{[i-1]})$需要对智能体 1，…，$i$ 的所有联合配置进行边缘化，导致指数级复杂度 $O(|\mathcal{S}|^i|\mathcal{A}|^i)$。相反，第 8 行采用采样进行估计。这将计算复杂度从指数级降低到多项式级： $O(K\cdot H\cdot|\mathcal{S}||\mathcal{A}|\cdot\max\{|\mathcal{S}|,\epsilon^{-2}\log(K|\mathcal{S}||\mathcal{A}|H/\delta)\})$，该复杂度关于所有参数均为多项式。下面，我们给出所提算法的理论保证。

定理1. 我们将最优策略记为 π^*。当算法 *1*中的样本数 N 设为$N = \frac{1}{2\epsilon^2}\log\frac{2K|\mathcal{S}||\mathcal{A}|^H}{\delta}$ 时，以至少 $1-\delta$的概率，算法 *1*所产生的策略π 满足

$$V_1^{\pi}(\bar{\mathbf{s}}_1) \geq \frac{V_1^{\pi^*}(\bar{\mathbf{s}}_1)}{2} - \epsilon KH.$$

定理1保证了一个 $1/2$近似，并具有加性误差 ϵKH，这与单调子模最大化中贪心算法的近似比相匹配 [Wolsey, 1982]。参数 ϵ和 δ 通过 N的取值控制精度与复杂度之间的权衡。值得注意的是，尽管我们将策略限制为可分解形式，这意味着各个智能体的执行过程不依赖于其他智能体，但输出策略相对于最优联合策略仍具有竞争力；后者可能不可分解，且通常需要指数级内存才能表示和执行。这一结果表明，在 MARLS 中，可分解性并不会从根本上限制近似质量，而且在易于处理且可实际实现的策略类别中，也能实现近乎最优的性能。

算法 1贪心策略优化

1: $N := \frac{1}{2\epsilon^2}\log \frac{2K|\mathcal{S}||\mathcal{A}|H}{\delta}$
2: **对于** $i = 1$**到** K **执行**
3: **对于** $h = H$ **到**1**执行**
4: **对于**(s, a) **在** $\mathcal{S} \times \mathcal{A}$ **中执行**
5: **如果** $i = 1$ **则**
6: $\widehat{R}_{1,h}(s, a) = f((s, a))$
7: **else**
8: 使用前 $i-1$ 个智能体的采样轨迹计算平均边际奖励：$\widehat{R}_{i,h}(s, a|\pi_{[i-1]}) = \frac{1}{N}\sum_{l=1}^{N}(f(\{(s^j_{h,l}, a^j_{h,l})\}_{j\in[i-1]} \cup \{(s, a)\}) - f(\{(s^j_{h,l}, a^j_{h,l})_{j\in[i-1]})\})$
9: **结束 if**
10: $\widehat{Q}^i_h(s, a) = R_{i,h}(s, a|\pi_{[i-1]}) + \sum_{s'\in\mathcal{S}} P_{i,h}(s'|s, a)\,\widehat{V}_{i,h+1}(s')$
11: **结束循环**
12: **对** s **中的** $\mathcal{S}$ **执行**
13: $\widehat{V}^i_h(s) =$ 最大值$_{a\in\mathcal{A}}\ \widehat{Q}^i_h(s, a)$
14: $\pi_{i,h}(s) =$ 参数最大值$_{a\in\mathcal{A}}\ \widehat{Q}^i_h(s, a)$
15: **结束循环**
16: **结束循环**
17: **从** $l = 1$**到** N **执行**
18: 采样智能体 i的第 l条轨迹：$\tau_{i,l} = \{(s^i_{1,l}, a^i_{1,l}), (s^i_{2,l}, a^i_{2,l}) \ldots (s^i_{H,l}, a^i_{H,l})\}$
19: **结束 for**
20: **结束 for**
21: **返回**策略 π

4.3 未知转移动态下基于UCB的贪心价值迭代

在第4.2节中，我们在转移动态已知的假设下提出了贪心策略优化算法。然而，在许多实际应用中，转移矩阵 P是未知的，必须通过与环境交互来学习。在这种情况下，我们的目标转为最小化遗憾，即在探索过程中，最优策略的表现与学习得到的策略的表现之间的累积差值。

为应对这一挑战，我们提出了UCB-GVI（上置信界贪心价值迭代），这是一种结合乐观探索与子模贪心最大化的基于模型的强化学习算法。完整流程见算法2。

算法描述。UCB-GVI在 T 个回合中运行，每个回合 k 包含三个阶段：通过乐观价值迭代计算策略、通过轨迹采样估计边际奖励，以及在真实环境中执行策略。

首先，策略计算阶段按顺序针对每个智能体 $i = 1, \ldots, K$ 进行，从第 $h = H$ 步到第 $h = 1$步执行反向价值迭代。对于在先前回合中访问过的每个状态-动作对（s, a）（即$N_{i,h}$（s, a）> 0），算法首先根据观测到的转移构建经验转移模型$P_{i,h}$（$s'|s, a$）。然后，通过在经验动态 P_i 下采样轨迹，并结合先前计算出的智能体$1, \ldots, i-1$ 的策略$\pi_{[i-1]}$，估计边际奖励$R_{i,h}$（$s, a|\pi_{[i-1]}$）（第10行）。这种基于采样的方法能够捕捉智能体 i 在状态 s选择动作 a 所带来的预期边际贡献，同时考虑转移动态以及顺序排列中先前智能体的策略所引发的随机行为。

使用这些估计，算法计算出融入探索奖励的乐观Q值估计：

$$\begin{aligned}\widehat{Q}_{i,h}(s, a) = \widehat{R}_{i,h}(s, a|\pi_{[i-1]}) \\ + \sum_{s'\in\mathcal{S}} \widehat{P}_{i,h}(s'|s, a)\widehat{V}_{i,h+1}(s') \\ + b(N_{i,h}(s, a)) + \frac{\epsilon}{KH},\end{aligned}$$

算法2上置信界贪心价值迭代（UCB-GVI）

输入： 回合 T，置信度 δ，近似参数 ϵ
初始化： $\mathcal{D} \leftarrow \emptyset$, $N_{i,h}(s,a) \leftarrow 0$, $N_{i,h}(s,a,s') \leftarrow 0$ 对所有 i,h,s,a,s'
设置 $N \leftarrow \frac{K^2H^2}{2\epsilon^2} \log\frac{6K|\mathcal{S}|\mathcal{A}|H}{\delta}$
对于回合 $k=1$ 到 T **执行**
  对于智能体 $i=1$ 到 K **执行**
    对于 步 $h=H$ 递减至 **1执行**
      对于 $(s,a) \in \mathcal{S}\times\mathcal{A}$ **执行**
        如果 $N_{i,h}(s,a) > 0$ **则**
          $\widehat{P}_{i,h}(s'|s,a) \leftarrow \frac{N_{i,h}(s,a,s')}{N_{i,h}(s,a)}$ 对所有 $s' \in \mathcal{S}$
          $\widehat{R}_{i,h}(s,a|\pi_{[i-1]}) \leftarrow \frac{1}{N}\sum_{l=1}^{N}\Delta f(\{(\tilde{s}^m_{h,l}, \tilde{a}^m_{h,l})\}_{m\in[i-1]}, (s,a))$
          $\widehat{Q}_{i,h}(s,a) \leftarrow \widehat{R}_{i,h}(s,a|\pi_{[i-1]}) + \sum_{s'\in\mathcal{S}}\widehat{P}_{i,h}(s'|s,a)\widehat{V}_{i,h+1}(s') + b(N_{i,h}(s,a)) + \frac{\epsilon}{KH}$
        else
          $\widehat{Q}_{i,h}(s,a) \leftarrow H$
        结束 if
      结束 for
      对 $s \in \mathcal{S}$ **执行**
        $\widehat{V}_{i,h}(s) \leftarrow$最大$_{a\in\mathcal{A}}\ \widehat{Q}_{i,h}(s,a)$
        $\pi_{i,h}(s) \leftarrow$ 最大值对应的参数$_{a\in\mathcal{A}}\ \widehat{Q}_{i,h}(s,a)$
      结束循环
    结束 for
    for $l=1$ 到 N **执行**
      采样轨迹 $\tau_{i,l} = \{(\tilde{s}^i_{h,l}, \tilde{a}^i_{h,l})\}_{h=1}^{H}$ 使用 π_i 和 $\widehat{P}_i$
    结束 for
  结束循环
  从初始状态 $\bar{s}_1^1, \ldots, \bar{s}_1^K$ 开始
  对步 $h=1$至 H **执行**
    对智能体 $i=1$ 至 K **执行**
      执行 $a^i_{h,k} \leftarrow \pi_{i,h}(s^i_{h,k})$ ； 观察 $s^i_{h+1,k} \sim P_{i,h}(\cdot|s^i_{h,k}, a^i_{h,k})$
      $N_{i,h}(s^i_{h,k}, a^i_{h,k}) \leftarrow N_{i,h}(s^i_{h,k}, a^i_{h,k}) + 1$
      $N_{i,h}(s^i_{h,k}, a^i_{h,k}, s^i_{h+1,k}) \leftarrow N_{i,h}(s^i_{h,k}, a^i_{h,k}, s^i_{h+1,k}) + 1$
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{(h, s^i_{h,k}, a^i_{h,k}, s^i_{h+1,k})\}$
    结束循环
  结束循环
结束循环
返回： 策略 $\{\pi^k\}_{k=1}^{T}$

其中，探索奖励 $b(N) = H\sqrt{\frac{2S\iota}{N}} + \frac{3HS\iota}{N}$ with $\iota = \log(6S^2ATHK/\delta)$ 为估计误差提供了一个置信上界。$N_{i,h}(s, a, s')$ 表示在所有先前回合中，智能体 i 在步 h 从 (s, a) 转移到 s' 的次数。对于尚未访问过的状态-动作对，算法采用乐观初始化 $\widehat{Q}_h^i(s, a) = H$。

接下来，轨迹采样阶段会生成合成数据，用于后续回合中的边际奖励估计。对于每个智能体 i，算法通过在经验转移模型 P_i下执行策略 π_i，采样 $N = \frac{K^2H^2}{2\epsilon^2}\log\frac{6K|\mathcal{S}|\mathcal{A}|H}{\delta}$ 条模拟轨迹。最后，在第 26 行至第 33 行的策略执行阶段，算法在真实环境中部署计算得到的联合策略 $\pi = (\pi_1, \ldots, \pi_K)$。

遗憾定义。为形式化刻画算法的性能保证，我们引入 α-遗憾的概念，该概念考虑了子模最大化的多项式时间贪心算法所固有的近似因子。

定义3（α-遗憾）。令 $\pi^* = \mathrm{argmax}_\pi V_1^\pi(\bar{\mathrm{s}}_1)$ 表示全局最优策略，并令 π^k 表示在回合 k中执行的策略。算法在 T 个回合中的 α-遗憾为

$$\mathcal{R}_{T,\alpha} = \sum_{k=1}^{T}\left[\alpha V_1^{\pi^*}(\bar{\mathrm{s}}_1) - V_1^{\pi^k}(\bar{\mathrm{s}}_1)\right].$$

参数 $\alpha \in (0, 1]$ 刻画了近似保证：$\alpha = 1$ 对应标准遗憾概念，而 $\alpha < 1$ 则反映了计算约束。对于我们的问题，贪心算法在经验模型下可达到相对于最优策略的 $1/2$-近似。具体而言，我们的主要理论贡献是为 UCB-GVI 建立了近乎最优的遗憾界。在给出定理 2 的结果之前，我们先阐述证明中的技术新意。

技术概述 我们的遗憾界证明（附录 D）需要采用全新的技术，超越单智能体强化学习中标准的 UCB 类分析方法 [Azar 等人（2017年），Jin 等人， 2018]。虽然我们的方法建立在乐观价值估计之上，但 MARLS 的多智能体和子模结构带来了若干根本性挑战，需要借助新的分析工具。具体而言，我们的分析在三个方面不同于现有工作：（i）在经验转移动力学下建立近似保证；（ii）在不引入对智能体数量的指数级依赖的前提下，控制多智能体转移估计误差 K；（iii）在学习循环中通过额外的采样过程估计边际奖励。

我们的分析包含两个主要部分。

第一部分：经验转移动态下的近似比。我们证明，贪心序贯策略构造在辅助MDP $\widehat{M}$（其转移动态为 $\widehat{\mathrm{P}}^k$，增广奖励为 $r(\mathrm{s}, \mathrm{a}) + \sum_{i=1}^{K} b(N_i^k, h(s^i, a^i))$）上可达到 1/2 近似（引理 8）。核心难点在于，最优策略 π^* 可能不可分解，这意味着其价值函数无法表示为各个智能体边际增益之和。为解决这一问题，我们提出了一种新颖的证明技巧，对比两个假想的智能体群组：一个遵循贪心策略 π^k，另一个遵循最优策略 π^*。我们仔细定义了量 G_i^k，用于刻画从 π^* 组中加入智能体 i 所带来的边际贡献（定义见式 (14)）。利用子模性和单调性，我们证明了 $\sum_{i=1}^{K} G_i^k \geq \bar{V}^{\pi^*,k}_{1}(\bar{\mathrm{s}}_1) - \bar{V}^{\pi^k,k}_{1}(\bar{\mathrm{s}}_1)$，同时对每个 $G_i^k \leq \bar{V}^{\pi_i^k,k}_{1}(\bar{s}_1^i;\, \pi^k_{[i-1]}) + O(\epsilon/K)$ 进行界定。随后对各个智能体进行望远镜求和，即可得到所需的近似比。

第二部分：通过多智能体望远镜分解与误差隔离进行遗憾分解。我们通过从 $h = H$ 向后递归展开价值差 $\bar{V}^{\pi^k,k}_{h}(\mathrm{s}_h^k) - V_h^{\pi^k}(\mathrm{s}_h^k)$，得到最终的遗憾界。关键的技术挑战在于，如何界定引理 10 和引理 9 中给出的非负函数在经验转移与真实转移下期望值之差。这两个界都利用转移的乘积结构 $\mathbf{P}_h = \prod_{i=1}^{K} P_{i,h}$进行望远镜分解，从而隔离单个智能体误差，避免对 K 的指数依赖。引理 9 采用了更复杂、精细的分析：首先通过校准后的集中不等式建立粗略的递归关系 $Y_i - Y_{i-1} \leq \frac{Y_{i-1}}{2HK} + O(\frac{\iota}{N})$，再将其细化为 $Y_i \leq (1 + \frac{1}{K})Y_{i-1} + O(\frac{\iota}{N})$，以得到紧确界。这两个界相互配合——利用引理 10 控制确定性探索成本，并利用引理 9 支持自归一化递归分析——最终得到 $\tilde{O}(S^2AH^3K^2\sqrt{T})$ 的遗憾率。

贪心策略优化的理论保证见定理 2。

定理 2（遗憾　　的遗憾界UCB-GVI）。 以至少 $1-\delta$ 的概率，算法2 达到遗憾界 o　　f

$$\mathcal{R}_{T,1/2} = O\Big(S^2AH^3K^2\log(SATHK/\delta)\log T + H^2KS\sqrt{AT\log(SATHK/\delta)}\Big),$$

其中 $O(\cdot)$ 省略了多对数因子。

定理的证明推迟至附录 D。根据该结果，当 $K=1$ 时，我们的上界可化简为 $O(H^2S\sqrt{AT})$（忽略对数因子），这恢复了单智能体设定。与标准分回合强化学习的下界以及已知的最佳上界 $\Omega(H\sqrt{SAHT})$ 相比 [Domingues 等人， 2021], Azar 等人 [2017], 我们的上界多出一个 $\sqrt{HS}$ 因子，这反映了多智能体协同和边际奖励估计所面临的挑战。

重要的是，遗憾仅随智能体数量 K 呈多项式增长，这表明尽管联合动作空间规模呈指数级增长，在这一设定下仍能实现有意义的学习。具体而言，主导项$O(H^2KS\sqrt{AT})$ 随 K 线性增长，这意味着每个智能体的学习成本与独立的单智能体学习相当，尽管智能体之间通过共享的子模奖励进行交互。对 K 的二次依赖所带来的额外对数项，也进一步体现了子模最大化在拟阵约束下的多项式时间算法中固有的计算与近似之间的权衡。

影响声明

本文旨在推动机器学习领域的发展。我们的工作可能产生许多潜在的社会影响，但我们认为没有哪一项需要在此特别强调。

参考文献

Maximilian Hüttenrauch、Adrian Šoši´c与Gerhard Neumann。面向群体系统的引导式深度强化学习。*arXiv*预印本 *arXiv:1709.06011*，2017年。Zool Hilmi Ismail、Nohaidda Sariff与E Gorrostieta Hurtado。协作多智能体机器人系统的综述与分析：挑战与方向。移动机器人应用，5:8–14，2018年。Chongjie Zhang与Victor Lesser。网络化分布式部分可观测马尔可夫决策过程中的协同多智能体强化学习。载于*AAAI*人工智能会议论文集，第25卷，第764–770页，2011年。Robert Rauch、Zdenek Becvar、Pavel Mach与Juraj Gazda。面向车载边缘计算中动态任务执行与资源分配的协作多智能体深度强化学习。*IEEE*车载技术汇刊，2024年。

Fitsum Debebe Tilahun、Ameha Tsegaye Abebe与Chung G Kang。面向无蜂窝大规模MIMO赋能的移动边缘计算网络中分布式资源分配的多智能体强化学习。*IEEE*车载技术汇刊，72(12):16454–16468，2023年。

Chenghao Li、Tonghan Wang、Chengjie Wu、Qianchuan Zhao、Jun Yang 和 Chongjie Zhang。《共享多智能体强化学习中的多样性颂扬》。《神经信息处理系统进展》，34：3991–4002，2021年。

Hao-Lun Hsu、Weixin Wang、Miroslav Pajic 和 Pan Xu。《合作式多智能体强化学习中的随机化探索》。《*AdvancesinNeural Information Processing Systems*》，37：74617–74689，2024年。

Aaron Ma、Michael Ouimet 和 Jorge Cortés。《通过动态子空间搜索实现用于多智能体规划的分层强化学习》。《自主机器人》，44(3)：485–503，2020年。

Avinatan Hassidim 和 Yaron Singer。《噪声下的子模优化》。载于学习理论会议，第1069–1122页。PMLR，2017年。

Rishabh Iyer、Ninad Khargoankar、Jeff Bilmes 和 Himanshu Asanani。《次模组合信息测度及其在机器学习中的应用》。载于算法学习理论，第722–754页。PMLR，2021年。

Wenjing Chen、Shuo Xing、Samson Zhou 和 Victoria G Crawford。公平次模覆盖。载于《第十三届国际学习表征会议》，2025a。

Yisong Yue与Carlos Guestrin。线性次模赌博机及其在多样化检索中的应用。《神经信息处理系统进展》，第24卷，2011年。Richard Bellman。马尔可夫决策过程。《数学与力学杂志》，第679–684页，1957。Richard S Sutton、Andrew G Barto等。《强化学习：导论》，第1卷。麻省理工学院出版社（剑桥），1998。Sainbayar Sukhbaatar、Rob Fergus等。利用反向传播学习多智能体通信。《神经信息处理系统进展》，第29卷，2016年。Jakob Foerster、Ioannis Alexandros Assael、Nando De Freitas与Shimon Whiteson。利用深度多智能体强化学习学习通信。《神经信息处理系统进展》，第29卷，2016年。Ziyi Chen、Yi Zhou、Rong-Rong Chen与Shaofeng Zou。具有有限时间分析的样本高效且通信高效的去中心化演员-评论家算法。载于国际机器学习会议，第3794–3834页。PMLR，2022年。

Peter Sunehag、Guy Lever、Audrunas Gruslys、Wojciech Marian Czarnecki、Vinicius Zambaldi、Max Jaderberg、Marc Lanctot、Nicolas Sonnerat、Joel Z Leibo、Karl Tuyls等。面向协作多智能体学习的价值分解网络。*arXiv*预印本 *arXiv:1706.05296*，2017年。Tabish Rashid、Mikayel Samvelyan、Christian Schroeder De Witt、Gregory Farquhar、Jakob Foerster与Shimon Whiteson。面向深度多智能体强化学习的单调价值函数分解。《机器学习研究杂志》，21(178):1–51，2020年。Xiaoteng Ma、Yiqin Yang、Chenghao Li、Yiwen Lu、Qianchuan Zhao与Jun Yang。合作式多智能体强化学习中智能体间交互的建模。载于第*20*届国际自主智能体与多智能体系统会议论文集，第853–861页，2021年。Zehao Dou、Jakub Grudzien Kuba与Yaodong Yang。理解深度合作式多智能体强化学习中的价值分解算法。*ArXiv:2202.04868*，2022年。Anuj Mahajan、Tabish Rashid、Mikayel Samvelyan与Shimon Whiteson。Maven：多智能体变分探索。《神经信息处理系统进展》，第32卷，2019年。Tonghan Wang、Jianhao Wang、Yi Wu与Chongjie Zhang。基于影响力的多智能体探索。载于第*8*届国际学习表征会议，*ICLR 2020*，2020年。Jiechuan Jiang与Zongqing Lu。个体性的涌现。载于国际机器学习会议，第4992–5001页。PMLR，2021年。Laurence A Wolsey。子模集合覆盖问题贪心算法的分析。《组合学》，2(4):385–393，1982。Uriel Feige。近似集合覆盖的ln(n)阈值。《美国计算机学会学报》（*JACM*），45(4):634–652，1998。Moran Feldman、Paul Liu、Ashkan Norouzi-Fard、Ola Svensson与Rico Zenklusen。拟阵约束下的流式子模最大化。《运筹学数学》，2025年。Gruia Calinescu、Chandra Chekuri、Martin Pal与Jan Vondrák。在拟阵约束下最大化单调子模函数。《*SIAM*计算杂志》，40(6):1740–1766，2011年。Kiarash Banihashem、Leyla Biabani、Samira Goudarzi、MohammadTaghi Hajiaghayi、Peyman Jabbarzade与Morteza Monemizadeh。拟阵子模最大化的动态算法。载于*2024*年*ACM-SIAM*离散算法年会（*SODA*）论文集，第3485–3533页。SIAM，2024年。Marshall L Fisher、George L Nemhauser与Laurence A Wolsey。最大化子模集合函数的近似分析——*II*。Springer，1978年。Gruia Calinescu、Chandra Chekuri、Martin Pál与Jan Vondrák。在拟阵约束下最大化子模集合函数。载于国际整数规划与组合优化会议，第182–196页。Springer，2007年。Ashwinkumar Badanidiyuru与Jan Vondrák。最大化子模函数的快速算法。载于第*25*届*ACM-SIAM*离散算法年会论文集，第1497–1514页。SIAM，2014年。Zirui Xu、Xiaofeng Lin与Vasileios Tzoumas。不可预测且部分可观测环境中多机器人协同的老虎机子模最大化。*arXiv*预印本 *arXiv:2305.12795*，2023年。Mohammad Ghavamzadeh Azar、Ian Osband与Rémi Munos。强化学习的极小极大遗憾界。载于国际机器学习会议（*ICML*），第263–272页，2017年。Chi Jin、Zeyuan Allen-Zhu、Sebastien Bubeck与Michael I Jordan。Q 学习是否可证明高效？《神经信息处理系统进展》，第31卷，2018年。Omar Darwiche Domingues、Pierre Ménard、Emilie Kaufmann与Michal Valko。有限马尔可夫决策过程中的分回合强化学习：极小极大下界再探。载于算法学习理论，第578–598页。PMLR，2021年。

Wenjing Chen、Shuo Xing 和 Victoria G Crawford。次模赌博机中用于纯探索的自适应阈值采样。载于第*41*届人工智能不确定性会议，2025b。Sho Takemori、Masahiro Sato、Takashi Sonoda、Janmajay Singh 和 Tomoko Ohkuma。多重约束下的次模赌博机问题。载于人工智能不确定性会议，第191–200页。PMLR，2020年。Wenjing Chen 和 Victoria G Crawford。带赌博机反馈的线性次模最大化。*arXiv*预印本 *arXiv:2407.02601*，2024年。Adish Singla、Sebastian Tschiatschek 和 Andreas Krause。基于自适应采样的含噪次模最大化及其在众包图像集摘要中的应用。载于*AAAI*人工智能会议论文集，第30卷，2016年。Fares Fourati、Vaneet Aggarwal、Christopher Quinn 和 Mohamed-Slim Alouini。全赌博机反馈下非单调随机次模最大化的随机贪心学习。载于人工智能与统计国际会议，第7455–7471页。PMLR，2023年。Fares Fourati、Christopher John Quinn、Mohamed-Slim Alouini 和 Vaneet Aggarwal。组合随机贪心赌博机。载于*AAAI*人工智能会议论文集，第38卷，第12052–12060页，2024年。Ian Osband、Benjamin Van Roy 和 Zheng Wen。通过随机价值函数实现泛化与探索。载于国际机器学习会议（*ICML*），第2377–2386页，2016年。Zihan Zhang、Xiangyang Ji 和 Simon Du。强化学习是否比赌博机更困难？一种摆脱时域诅咒的近最优算法。载于学习理论会议，第4528–4531页。PMLR，2021年。Christoph Dann、Teodor Vanislavov Marinov、Mehryar Mohri 和 Julian Zimmert。超越价值函数差距：分回合强化学习中改进的依赖实例的遗憾界。《神经信息处理系统进展》，34:1–12，2021年。Priyank Agrawal 和 Shipra Agrawal。面向平均奖励与分回合强化学习的乐观Q学习扩展摘要。载于第三十八届学习理论年会，第1–1页。PMLR，2025年。Gergely Neu 和 Ciara Pike-Burke。分回合强化学习中乐观主义的统一视角。《神经信息处理系统进展》，33:1392–1403，2020年。Andrea Zanette 和 Emma Brunskill。在不具备领域知识的强化学习中，利用价值函数界得到更紧的依赖问题的遗憾界。载于国际机器学习会议（*ICML*），第7304–7313页，2019年。Zihan Zhang、Yuan Zhou 和 Xiangyang Ji。基于参考优势分解的近乎最优无模型强化学习。《神经信息处理系统进展》，33:15198–15207，2020年。

附录

A 其他相关工作

A.1 具有子模奖励的组合多臂老虎机问题

子模奖励已被引入多臂组合赌博机设定 [Chen 等人，2025b，Takemori 等人，2020年，Chen 和 Crawford，2024年，Singla 等人， 2016]。该领域主要有两条研究路线。第一条聚焦于在线遗憾最小化设定，其目标是与未知环境顺序交互并最小化累积遗憾Takemori 等人 [2020], Fourati 等人 [2023, 2024], Yue 和 Guestrin[2011]。第二条研究纯探索设定，其目标是以尽可能少的含噪样本找到高质量解Singla 等人 [2016], Chen 和 Crawford [2024]。尽管这些问题设定相互关联，但它们在问题复杂度上存在根本差异：赌博机设定缺乏马尔可夫状态动态，且仅考虑单步回合，其中智能体选择臂的子集并立即观测随机奖励。相比之下，我们的框架纳入了完整的 MDP 动态，包括状态转移、长度为 H的多步规划时域，以及多个智能体之间的协调，其联合动作同时影响即时奖励和未来状态分布。

A.2 表格型强化学习中的遗憾分析

在分回合马尔可夫决策过程（MDP）的遗憾最小化研究中，强化学习理论有着悠久而丰富的历史 Osband 等人 [2016], Jin 等人 [2018], Zhang 等人 [2021], Dann 等人 [2021], Agrawal 和 Agrawal [2025], Neu 和 Pike-Burke [2020]。对于时域为 H、具有 S 个状态、 A 个动作，且总交互步数为~~T 的~~分回合 MDP，Azar 等人 [2017] 建立了 $\Omega(\sqrt{H^2SAT})$ 的近极小极大遗憾下界，并提出了基于模型的算法 UCBVI，其上界与该下界相匹配（相差对数因子）。Zanette 和 Brunskill[2019] 提出了 EULER 算法，该算法达到了极小极大速率，并通过利用伯恩斯坦型集中界和价值函数的方差结构，得出了更紧的依赖实例的界。与这些基于模型的方法互为补充，Jin 等人 [2018] 通过结合~~UCB~~ 探索奖励的 Q 学习，开发了首个具有可证明效率的无模型算法，实现了 $\tilde{O}(\sqrt{H^3SAT})$ 遗憾。后续研究通过改进探索策略降低了对时域的依赖 [Zhang 等人， 2020] 并得到了与近极小极大遗憾下界相匹配的界。

B 第4.1节附录

在本附录部分，我们给出正文第4.1节中缺失的细节与证明。我们首先在B.1节和B.2节中给出第4.1.1节中缺失的内容，然后在B.3节中给出引理 1 被省略的证明。

B.1 贝尔曼最优性

Lemma 2 (贝尔曼最优性 Bellman [1957],Sutton等人 [1998])。 对于任意策略 π，我们有 t

$$V_h^\pi(\mathbf{s}) = \sum_{a^1,a^2,\ldots,a^K \in \mathcal{A}} \pi_h(\mathbf{s},\mathbf{a}) Q_h^\pi(\mathbf{s},\mathbf{a}),$$

$$Q_h^\pi(\mathbf{s},\mathbf{a}) = r(\mathbf{s},\mathbf{a}) + \sum_{s'^1,s'^2,\ldots,s'^K \in \mathcal{S}} \prod_{i=1}^{K} P_{i,h}(s'^i|s^i,a^i) V_{h+1}^\pi(\mathbf{s}',\mathbf{a}). \tag{3}$$

此外，存在一个确定性最优策略 π^*，使得 $\pi_h^*(\mathbf{s},\mathbf{a}) = 1$ 或 0。特别地，我们将最优价值函数和 Q 函数分别记为 V^* 和 Q^*。那么，最优解满足

$$V_h^*(\mathbf{s}) = \max_{\mathbf{a}\in\mathcal{A}^K} Q_h^*(\mathbf{s},\mathbf{a}),$$
$$Q_h^*(\mathbf{s},\mathbf{a}) = r(\mathbf{s},\mathbf{a}) + \sum_{s'^1,s'^2,\ldots,s'^K\in\mathcal{S}} \prod_{i=1}^{K} P_{i,h}(s'^i|s^i,a^i) V_{h+1}^*(\mathbf{s}'). \tag{4}$$

B.2 $H = 1$的特殊情形

在本附录部分中，我们将证明，当 $H = 1$ 时，通过将其归约为分区拟阵约束下的子模最大化问题，在 MARLS 中寻找最优策略是 NP 困难的。具体而言，我们有以下引理。

引理 3（等价于分区拟阵次模最大化）。设 $f\colon\ 2^{\mathcal{S}\times\mathcal{A}} \to [0,1]$ 为单调子模函数。当 $H = 1$ 时，寻找最优策略的问题

$$\max_{a^1,\ldots,a^K} f\big(\{(\bar{s}^1,a^1),\ldots,(\bar{s}^K,a^K)\}\big),$$

等价于满足分区拟阵约束的单调次模最大化问题。

证明。我们通过构造一个分区拟阵及相应的单调子模函数来建立这种等价性。对于每个智能体 $i\in[K]$，定义 $\mathcal{A}_i = \{(i,\ a)\colon a\in\mathcal{A}\}$ 为动作空间的带标签副本。令基集为 $\mathcal{U} = \bigcup_{i=1}^K \mathcal{A}_i$，则 $\{A_i\}_{i\in[K]}$ 构成了 $\mathcal{U}$的一个划分，因为对于任意 $i\neq j$，A_i 和 A_j 互不相交。我们定义分区拟阵 $\mathcal{M} = (\mathcal{U},\ \mathcal{I})$，其独立集为

$$\mathcal{I} = \{X\subseteq\mathcal{U} : |X\cap\mathcal{A}_i|\le 1 \text{ for all } i\in[K]\}.$$

T这一约束确保每个智能体的动作空间中至多只能选择一个动作。

对于任意 $X\subseteq\mathcal{U}$，令 $X_i = X\cap\mathcal{A}_i$ 表示 X 中对应于智能体 i的元素。定义 $F\colon\ 2^{\mathcal{U}}\to[0,1]$ 如下:

$$F(X) = f\big(\{(\bar{s}^i,a) : (i,a)\in X\}\big).$$

若 $X_i = \emptyset$，则智能体 i 不会向 f的自变量贡献任何状态-动作对。

我们首先验证 F 是单调子模的。对于单调性，若 $X\subseteq Y\subseteq\mathcal{U}$，则 $\{(\bar{s}^i,a) : (i,a)\in X\}\subseteq\{(\bar{s}^i,a) : (i,a)\in Y\}$，因此根据 f的单调性，$F(X)\le F(Y)$。对于子模性，考虑 $X\subseteq Y\subseteq\mathcal{U}$和$e = (j,a)\in\mathcal{U}\setminus Y$。定义 $S_X = \{(\bar{s}^i,a') : (i,a')\in X\}$ 以及 $S_Y = \{(\bar{s}^i,a') : (i,a')\in Y\}$。于是 $S_X\subseteq S_Y$且 $(\bar{s}^j,a)\notin S_Y$。根据 f的子模性，

$$\begin{aligned} F(X\cup\{e\}) - F(X) &= f(S_X\cup\{(\bar{s}^j,a)\}) - f(S_X) \\ &\ge f(S_Y\cup\{(\bar{s}^j,a)\}) - f(S_Y) = F(Y\cup\{e\}) - F(Y). \end{aligned}$$

接下来，我们证明该等价性。由于 F 是单调的，问题 $\max_{X\in\mathcal{I}} F(X)$ 至少存在一个最优解 X^*，其形式为 $X^* = \{(1,a^1)\ldots,(K,a^K)\}$，即从每个划分$\mathcal{A}_i$中恰好包含一个元素。因此，

$$\max_{X\in\mathcal{I}} F(X) = \max_{X\subseteq\mathcal{U}:|X\cap\mathcal{A}_i|=1 \text{ for all } i} F(X).$$

由于优化问题 $\max_{a^1,\ldots,a^K\in\mathcal{A}} f\big(\{(\bar{s}^i,a^i)\colon\ i\in[K]\}\big)$ 等价于 $\max_{X\subseteq\mathcal{U}:|X\cap\mathcal{A}_i|=1 \text{ 对所有 } i} F(X)$，我们得出

$$\max_{a^1,\ldots,a^K\in\mathcal{A}} f\big(\{(\bar{s}^i,a^i) : i\in[K]\}\big) = \max_{X\in\mathcal{I}} F(X),$$

这是一个带分区拟阵约束的单调次模最大化问题。 □

B.3 引理 1 的证明

本节中，我们将证明引理 1 中省略的证明概要。

证明概要。 给定固定的 $\pi_{[i-1]}$，对于第一部分，根据 $Q_h^{\pi_i}$的定义和期望的塔式性质，

$$\begin{aligned} Q_h^{\pi_i}(s,a;\pi_{[i-1]}) &= \mathbb{E}_{\pi_i}[\Delta r_i(\mathbf{s}_h,\mathbf{a}_h) + \sum_{t=h+1}^{H} \Delta r_i(\mathbf{s}_t,\mathbf{a}_t) \mid s_h^i = s, a_h^i = a, \pi_{[i-1]}] \\ &= R_{i,h}(s,a|\pi_{[i-1]}) + \mathbb{E}_{s_{h+1}^i \sim P_{i,h}(\cdot|s,a)}[V_{h+1}^{\pi_i}(s_{h+1}^i;\pi_{[i-1]})] \\ &= R_{i,h}(s,a|\pi_{[i-1]}) + \sum_{s'\in\mathcal{S}} P_{i,h}(s'|s,a) V_{h+1}^{\pi_i}(s';\pi_{[i-1]}). \end{aligned}$$

$V_h^{\pi_i}$的等式由全期望定律得出。

我们为智能体 i 定义一个增广马尔可夫决策过程 M_i，其状态空间为 $\mathcal{S}$、动作空间为 $\mathcal{A}$、转移核为 $P_{i,h}$，时变奖励为 $R_{i,h}(\cdot,\cdot|\pi_{[i-1]})$。关键观察是，$R_{i,h}(s,a|\pi_{[i-1]})$ 是确定性的（作为s,a,h的函数），因为式 (2) 中的期望是针对固定的随机策略 $\pi_{[i-1]}$ 和已知的转移动态求取的。由上述结果可知，$Q_h^{\pi_i}(s,a;\pi_{[i-1]})$ 和 $V_h^{\pi_i}(s;\pi_{[i-1]})$ 是在策略 π_i 下 M_i 的价值函数。根据 $\widetilde{\pi}_i$的定义，我们知道 $\widetilde{\pi}_i$ 是马尔可夫决策过程 M_i 的最优策略。根据贝尔曼最优性原理，$\widetilde{\ }$ $\widetilde{\ }$最优价值函数满足 $V_h^{\pi_i}(s;\pi_{[i-1]}) = \max_a Q_h^{\pi_i}(s,a;\pi_{[i-1]})$，而最优Q函数满足贝尔曼方程。通过从 $h=H$ 到 $h=1$ 的逆向归纳可以验证这一点。 □

C 第4.2节附录

本节给出定理1中主要结果的证明。

首先，我们给出以下引理，它保证了贪心策略优化执行过程中的高概率事件。

引理4. 定义事件

$$\mathcal{E} := \left\{ \forall (s,a) \in \mathcal{S}\times\mathcal{A}, \forall i \in [K], h \in [H] : |R_{i,h}(s,a|\pi_{[i-1]}) - \widehat{R}_{i,h}(s,a|\pi_{[i-1]})| \le \epsilon \right\}.$$

则 $\mathcal{E}$ 以概率 $1-\delta$成立。

证明。 对于任意 $i \in [K]$、$h \in [H]$和 $l \in [N]$，我们考虑由第 l 条轨迹 $\{(s_{h,l}^j, a_{h,l}^j)\}_{h\in[H],j\in[K]}$ 得到的以下随机变量。

$$f(\{(s_{h,l}^j, a_{h,l}^j)\}_{j\in[i-1]} \cup \{(s,a)\}) - f(\{(s_{h,l}^j, a_{h,l}^j)\}_{j\in[i-1]}) = \Delta f(\{(s_{h,l}^j, a_{h,l}^j)\}_{j\in[i-1]}, (s,a))$$

由此可得

$$\begin{aligned} \mathbb{E}[f(\{(s_{h,l}^j, a_{h,l}^j)\}_{j\in[i-1]} \cup \{(s,a)\}) - f(\{(s_{h,l}^j, a_{h,l}^j)\}_{j\in[i-1]})] &= \mathbb{E}[\Delta f(\{(s_{h,l}^j, a_{h,l}^j)\}_{j\in[i-1]}, (s,a))] \\ &= R_{i,h}(s,a|\pi_{[i-1]}) \end{aligned}$$

根据霍夫丁不等式，对于任意固定的 $(s,a) \in \mathcal{S}\times\mathcal{A}$ 和 $h\in[H]$，以下不等式以至少 $1-\frac{\delta}{K|\mathcal{S}||\mathcal{A}|H}$ 的概率成立：

$$|R_{i,h}(s,a|\pi_{[i-1]}) - \widehat{R}_{i,h}(s,a|\pi_{[i-1]})| \le \epsilon$$

对所有 $i\in[K]$、$h\in[H]$、$(s,a)\in\mathcal{S}\times\mathcal{A}$ 应用并集界可得，事件 $\mathcal{E}$ 以至少 $1-\delta$ 的概率成立。

□

接下来，我们给出定理1的主要证明。

证明。 假设我们有两组 K 智能体，第一组 K 智能体遵循算法 1 输出的策略。我们将轨迹记为 $\{(s_h^i,\ a_h^i)\}_{h\in[H],\ i\in[K]}$。第二组 K 智能体遵循最优策略 π^*，我们将第二组智能体的轨迹记为 $\{(\tilde{s}_h^i,\ \tilde{a}_h^i)\}_{h\in[H],\ i\in[K]}$。为简化符号，我们定义

$$G_i = \mathbb{E}[\sum_{h=1}^{H} \Delta f(\{(s_h^j, a_h^j)\}_{j\in[K]} \cup \{(\tilde{s}_h^m, \tilde{a}_h^m)\}_{m\in[i-1]}, (\tilde{s}_h^i, \tilde{a}_h^i))|(\mathbf{s}, \mathbf{a}) \sim \pi, (\tilde{\mathbf{s}}, \tilde{\mathbf{a}}) \sim \pi^*],$$

其中 $(\mathbf{s},\ \mathbf{a}) \sim \pi$，$(\tilde{\mathbf{s}}\ \tilde{\mathbf{a}}) \sim \pi^*$ 表示 $\{(s_h^i,\ a_h^i)\}_{h\in[H],\ i\in[K]}$ 按照策略 π进行采样，而 $\{(\tilde{s}_h^i,\ \tilde{a}_h^i)\}_{h\in[H],\ i\in[K]}$根据最优策略 π^*进行采样。根据定义，可得

$$G_i = \mathbb{E}[\sum_{h=1}^{H} f(\{(s_h^j, a_h^j)\}_{j\in[K]} \cup \{(\tilde{s}_h^m, \tilde{a}_h^m)\}_{m\in[i]}) - f(\{(s_h^j, a_h^j)\}_{j\in[K]} \cup \{(\tilde{s}_h^m, \tilde{a}_h^m)\}_{m\in[i-1]})|(\mathbf{s}, \mathbf{a}) \sim \pi, (\tilde{\mathbf{s}}, \tilde{\mathbf{a}}) \sim \pi^*].$$

由此可得

$$\begin{aligned}
\sum_{i=1}^{K} G_i &= \mathbb{E}[\sum_{h=1}^{H} f(\{(s_h^j, a_h^j)\}_{j\in[K]} \cup \{(\tilde{s}_h^m, \tilde{a}_h^m)\}_{m\in[K]}) - f(\{(s_h^j, a_h^j)\}_{j\in[K]})|(\mathbf{s}, \mathbf{a}) \sim \pi, (\tilde{\mathbf{s}}, \tilde{\mathbf{a}}) \sim \pi^*] \\
&\geq \mathbb{E}[\sum_{h=1}^{H} f(\{(\tilde{s}_h^m, \tilde{a}_h^m)\}_{m\in[K]}) - f(\{(s_h^j, a_h^j)\}_{j\in[K]})|(\mathbf{s}, \mathbf{a}) \sim \pi, (\tilde{\mathbf{s}}, \tilde{\mathbf{a}}) \sim \pi^*] \\
&= \mathbb{E}[\sum_{h=1}^{H} f(\{(\tilde{s}_h^m, \tilde{a}_h^m)\}_{m\in[K]})|(\tilde{\mathbf{s}}, \tilde{\mathbf{a}}) \sim \pi^*] - \mathbb{E}[\sum_{h=1}^{H} f(\{(s_h^j, a_h^j)\}_{j\in[K]})|(\mathbf{s}, \mathbf{a}) \sim \pi] \\
&= V_1^*(\bar{\mathbf{s}}_1) - V_1^{\pi}(\bar{\mathbf{s}}_1),
\end{aligned} \tag{5}$$

其中，不等式成立是因为 f是单调的。接下来，我们对每个 i界定 G_i。由于 $\{(s_h^j,\ a_h^j)\}_{j\in[i-1]} \subseteq \{(s_h^j,\ a_h^j)\}_{j\in[K]} \cup \{(\tilde{s}_h^m,\ \tilde{a}_h^m)\}_{m\in[i]}$，根据子模性，可得

$$G_i \leq \mathbb{E}[\sum_{h=1}^{H} \Delta f(\{(s_h^j, a_h^j)\}_{j\in[i-1]}, (\tilde{s}_h^i, \tilde{a}_h^i))|(\mathbf{s}, \mathbf{a}) \sim \pi, (\tilde{\mathbf{s}}, \tilde{\mathbf{a}}) \sim \pi^*]. \tag{6}$$

请注意，无法保证最优策略可分解。令策略 $\pi_{i,\ h}^*$ 表示第 i个智能体在时间步 h 的边际策略，而全局策略为 π^*。那么，智能体 i 的边际策略应通过对其他智能体的状态-动作对进行积分来计算。具体而言，

$$\pi_{i,h}^*(s, a) = \sum_{\substack{\mathbf{s},\mathbf{a} \\ s^i=s,\, a^i=a}} \pi_h^*(\mathbf{s}, \mathbf{a}).$$

为简化符号，本文也用 $\pi_i^* = \{\pi_{i,h}^*\}_{h\in[H]}$ 表示智能体 i 在整个回合中的边际策略。因此，

$$\begin{aligned}
&\mathbb{E}[\sum_{h=1}^{H} \Delta f(\{(s_h^j, a_h^j)\}_{j\in[i-1]}, (\tilde{s}_h^i, \tilde{a}_h^i))|(\mathbf{s}, \mathbf{a}) \sim \pi, (\tilde{\mathbf{s}}, \tilde{\mathbf{a}}) \sim \pi^*] \\
=&\mathbb{E}[\sum_{h=1}^{H} \Delta f(\{(s_h^j, a_h^j)\}_{j\in[i-1]}, (\tilde{s}_h^i, \tilde{a}_h^i))|(\mathbf{s}, \mathbf{a}) \sim \pi, (\tilde{s}^i, \tilde{a}^i) \sim \pi_i^*] \\
=&V_1^{\pi_i^*}(\bar{s}_1; \pi_{[i-1]}).
\end{aligned}$$

将上述不等式与式 (6) 结合，可得

$$G_i \leq V_1^{\pi_i^*}(\bar{s}_1; \pi_{[i-1]}). \tag{7}$$

接下来，我们介绍 述并解释以下引理 5。该引理的证明将在本节后面给出 n.

引理 5 **.** 以至少 $1-\delta$ 的概率，对于任意时间步 $h \in [H]$ 和任意智能体 $i \in [K,$ $],$

$$V_h^{\pi_i^*}(s;\pi_{[i-1]}) \le \widehat{V}_h^i(s) + \epsilon(H-h+1) \le V_h^{\pi_i}(s;\pi_{[i-1]}) + 2\epsilon(H-h+1)$$
$$Q_h^{\pi_i^*}(s,a;\pi_{[i-1]}) \le \widehat{Q}_h^i(s,a) + \epsilon(H-h+1) \le Q_h^{\pi_i}(s,a;\pi_{[i-1]}) + 2\epsilon(H-h+1).$$

该引理表明，价值函数 $\widehat{V}_h^i(s)$ 和 $\widehat{Q}_h^i(s,\ a)$ 的估计值确实介于最优边缘策略 π_i^* 和 π_i 的真实价值函数之间，并且输出策略 π_i 近似优于 π_i^*。将引理中的结果与 (7) 结合，可得

$$G_i \le V_1^{\pi_i}(\bar{s}_1^i;\pi_{[i-1]}) + 2\epsilon H$$

根据 (5)，由于 $\sum_{i=1}^K G_i \ge V_1^*(\bar{\mathbf{s}}_1) - V_1^{\pi}(\bar{\mathbf{s}}_1)$，因此可得

$$\begin{aligned} V_1^*(\bar{\mathbf{s}}_1) - V_1^{\pi}(\bar{\mathbf{s}}_1) &\le \sum_{i=1}^K \{V_1^{\pi_i}(\bar{s}_1^i;\pi_{[i-1]}) + 2\epsilon H\} \\ &\le V_1^{\pi}(\bar{\mathbf{s}}_1) + 2\epsilon K H. \end{aligned}$$

通过重新整理上述不等式，我们可以证明定理中的结论。

□

接下来，我们给出引理 5 中省略的证明，重述如下。

引理 5。 以至少 $1-\delta$ 的概率，对于任意时间步 $h \in [H]$ 和任意智能体 $i \in [K],$

$$V_h^{\pi_i^*}(s;\pi_{[i-1]}) \le \widehat{V}_h^i(s) + \epsilon(H-h+1) \le V_h^{\pi_i}(s;\pi_{[i-1]}) + 2\epsilon(H-h+1)$$
$$Q_h^{\pi_i^*}(s,a;\pi_{[i-1]}) \le \widehat{Q}_h^i(s,a) + \epsilon(H-h+1) \le Q_h^{\pi_i}(s,a;\pi_{[i-1]}) + 2\epsilon(H-h+1)$$

证明。 我们通过归纳法证明该引理。首先，当 $h = H$ 时，根据定义，我们有

$$Q_H^{\pi_i^*}(s,a;\pi_{[i-1]}) = Q_H^{\pi_i}(s,a;\pi_{[i-1]}) = R_{i,H}(s,a|\pi_{[i-1]}),$$

and

$$\widehat{Q}_H^i(s,a) = \widehat{R}_{i,H}(s,a|\pi_{[i-1]})$$

根据引理 4，我们有，以至少 $1-\delta$ 的概率，

$$Q_H^{\pi_i^*}(s,a;\pi_{[i-1]}) \le \widehat{Q}_H^i(s,a) + \epsilon \le Q_H^{\pi_i}(s,a;\pi_{[i-1]}) + 2\epsilon. \tag{8}$$

根据引理 1 中的贝尔曼方程，我们有 $V_H^{\pi_i^*}(s;\pi_{[i-1]}) = \sum_{a\in\mathcal{A}} \pi_{i,H}^*(s,a) Q_H^{\pi_i^*}(s,a;\pi_{[i-1]})$。由于 $\sum_{a\in\mathcal{A}} \pi_{i,h}^*(s,a) = 1$，因此我们有

$$\begin{aligned} V_H^{\pi_i^*}(s;\pi_{[i-1]}) &\le \max_{a\in\mathcal{A}} Q_H^{\pi_i^*}(s,a;\pi_{[i-1]}) \\ &\le \max_{a\in\mathcal{A}} \widehat{Q}_H^i(s,a) + \epsilon \\ &= \widehat{V}_H^i(s) + \epsilon, \end{aligned}$$

其中，第二个不等式由式 (8) 中的不等式得出，最后一个等式则由 $\widehat{V}_H^i(s)$ 的定义得出。根据输出策略 π 的定义，我们有 $\widehat{V}_H^i(s) = \widehat{Q}_H^i(s,\ \pi_{i,\ H}(s))$，且 $V_H^{\pi_i}(s;\pi_{[i-1]}) = Q_H^{\pi_i}(s,\ \pi_{i,\ H}(s);\pi_{[i-1]})$。因此可得

$$\begin{aligned} \widehat{V}_H^i(s) &= \widehat{Q}_H^i(s,\pi_{i,h}(s)) \\ &\le Q_H^{\pi_i}(s,\pi_{i,h}(s);\pi_{[i-1]}) + \epsilon \\ &= V_H^{\pi_i}(s;\pi_{[i-1]}) + \epsilon \end{aligned}$$

其中，最后一个不等式由引理 1 中的贝尔曼方程得出。根据式 (8) 的结果，我们有该引理对 $h = H$成立。

假设该引理在步 $h+1$ 的情形下成立。那么对于时间步 h，根据引理 1 中的贝尔曼方程，我们有

$$\begin{aligned}
Q_h^{\pi_i^*}(s,a;\pi_{[i-1]}) &= R_{i,h}(s,a|\pi_{[i-1]}) + \sum_{s'\in\mathcal{S}} P_{i,h}(s'|s,a) V_{h+1}^{\pi_i^*}(s';\pi_{[i-1]}) \\
&\le R_{i,h}(s,a|\pi_{[i-1]}) + \sum_{s'\in\mathcal{S}} P_{i,h}(s'|s,a)\{\widehat{V}_{h+1}^{i}(s') + \epsilon(H-h)\} \\
&\le \widehat{R}_{i,h}(s,a|\pi_{[i-1]}) + \epsilon + \sum_{s'\in\mathcal{S}} P_{i,h}(s'|s,a)\{\widehat{V}_{h+1}^{i}(s') + \epsilon(H-h)\} \\
&\le \widehat{Q}_h^i(s,a) + \epsilon(H-h+1)
\end{aligned}$$

其中，第一个等式由引理 1 中的贝尔曼方程得出，此时第 i 个智能体的策略为 π_i^*。第二个不等式由事件 $\mathcal{E}$ 得出，最后一个不等式则由以下事实得出：对于任意 $(s,a) \in \mathcal{S}\times\mathcal{A}, h\in[H]$ 和 $i\in[K]$, $\sum_{s'\in\mathcal{S}} P_{i,h}(s'|s,a) = 1$。此外，根据算法 1 第 10 行的算法描述，我们有

$$\widehat{Q}_h^i(s,a) = \widehat{R}_{i,h}(s,a|\pi_{[i-1]}) + \sum_{s'\in\mathcal{S}} P_{i,h}(s'|s,a)\widehat{V}_{h+1}^{i}(s').$$

由于根据引理 1，我们有 $Q_h^{\pi_i}(s,a;\pi_{[i-1]}) = R_{i,h}(s,a|\pi_{[i-1]}) + \sum_{s'\in\mathcal{S}} P_{i,h}(s'|s,a) V_{h+1}^{\pi_i}(s';\pi_{[i-1]})$。根据结果对 $h+1$ 成立这一假设，我们有

$$V_{h+1}^{\pi_i^*}(s;\pi_{[i-1]}) \le \widehat{V}_{h+1}^{i}(s) + \epsilon(H-h) \le V_{h+1}^{\pi_i}(s;\pi_{[i-1]}) + 2\epsilon(H-h).$$

Then

$$\begin{aligned}
\widehat{Q}_h^i(s,a) &= \widehat{R}_{i,h}(s,a|\pi_{[i-1]}) + \sum_{s'\in\mathcal{S}} P_{i,h}(s'|s,a)\widehat{V}_{h+1}^{i}(s') \\
&\le R_{i,h}(s,a|\pi_{[i-1]}) + \sum_{s'\in\mathcal{S}} P_{i,h}(s'|s,a)\{V_{h+1}^{\pi_i}(s';\pi_{[i-1]}) + \epsilon(H-h)\} + \epsilon \\
&\le Q_h^{\pi_i}(s,a;\pi_{[i-1]}) + \epsilon(H-h+1).
\end{aligned}$$

接下来，根据引理 1，我们有

$$\begin{aligned}
V_h^{\pi_i^*}(s;\pi_{[i-1]}) &= \sum_{a\in\mathcal{A}} \pi_{i,h}^*(s,a) Q_h^{\pi_i^*}(s,a;\pi_{[i-1]}) \\
&\le \max_{a\in\mathcal{A}} Q_h^{\pi_i^*}(s,a;\pi_{[i-1]}) \\
&\le \max_{a\in\mathcal{A}} \widehat{Q}_h^i(s,a) + \epsilon(H-h+1) \\
&= \widehat{V}_h^i(s,a) + \epsilon(H-h+1).
\end{aligned}$$

由于

$$\begin{aligned}
\widehat{V}_h^i(s) &= \widehat{Q}_h^i(s,\pi_{i,h}(s)) \\
&\le Q_h^{\pi_i}(s,\pi_{i,h}(s);\pi_{[i-1]}) + \epsilon(H-h+1) \\
&= V_h^{\pi_i}(s;\pi_{[i-1]}) + \epsilon(H-h+1)
\end{aligned}$$

因此，我们可以证明该引理。 □

D 第 4.3 节的附录

本节将给出第 4.3 节中省略的证明和讨论。具体而言，我们将给出定理 2 中省略的证明。首先，我们定义并证明高概率事件。

D.1 高概率事件

在证明该定理之前，我们先介绍几个定义。首先，将智能体 i 在第 k 个回合的时间步 h 执行的策略定义为 $\pi_{i,h}^k$。然后，将智能体 i 在回合 k 的完整策略记为 $\pi_i^k := \{\pi_{i,h}^k\}_{h\in[H]}$。此外，将转移矩阵的经验估计记为 $\widehat{P}_{i,h}^k$。令 $\widehat{R}_{i,h}^k(s,a|\pi_{[i-1]}^k)$ 表示在算法2的第 k 个回合中第 10 行计算得到的边际奖励估计值。该估计值通过使用经验转移模型 $\widehat{P}^k$ 从策略 π^k 中采样轨迹获得。定义 $R_{i,h}^k(s,a|\pi_{[i-1]}^k) = \mathbb{E}_{\widehat{P}^k}[\Delta r_i(s_h,a_h)|s_h^i = s, a_h^i = a, \pi_{[i-1]}^k]$ 为在 $\widehat{P}^k$ 和 $\pi_{[i-1]}^k$ 下的期望边际奖励。根据构造，$\mathbb{E}[\widehat{R}_{i,h}^k(s,a|\pi_{[i-1]}^k)] = R_{i,h}^k(s,a|\pi_{[i-1]}^k)$，其中期望是针对用于计算 $\widehat{R}_{i,h}^k$ 的轨迹样本而言的。

我们定义高概率事件 $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2$ 满足 $\iota := \log(6S^2ATHK/\delta)$：

$$\mathcal{E}_1 := \Bigg\{ \forall (s,a) \in \mathcal{S}\times\mathcal{A}, \forall i \in [K], \forall h \in [H] \text{ with } k \in [T] : \\ |\widehat{R}_{i,h}^k(s,a|\pi_{[i-1]}^k) - R_{i,h}^k(s,a|\pi_{[i-1]}^k)| \le \frac{\epsilon}{KH} \Bigg\} \tag{9}$$

$$\mathcal{E}_2 := \Bigg\{ \forall (s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}, \forall i \in [K], \forall h \in [H] \text{ with } N_{i,h}^k(s,a) \ge 1 : \\ |P_{i,h}(s'|s,a) - \widehat{P}_{i,h}^k(s'|s,a)| \le \sqrt{\frac{2P_{i,h}(s'|s,a)\iota}{N_{i,h}^k(s,a)}} + \frac{3\iota}{N_{i,h}^k(s,a)} \Bigg\} \tag{10}$$

引理 6. 事件 $\mathcal{E}$ 以概率 $1-\frac{2\delta}{3}$ 成立。

证明。对于任意 $i \in [K]$、$h \in [H]$ 和 $l \in [N]$，我们考虑由第 l 条轨迹 $\{(s_{h,l}^j, a_{h,l}^j)\}_{h\in[H],j\in[K]}$ 在回合 k 中得到的以下随机变量。

$$f(\{(s_{h,l}^j,a_{h,l}^j)\}_{j\in[i-1]} \cup \{(s,a)\}) - f(\{(s_{h,l}^j,a_{h,l}^j)\}_{j\in[i-1]}) = \Delta f(\{(s_{h,l}^j,a_{h,l}^j)\}_{j\in[i-1]},(s,a))$$

由此可得

$$\begin{aligned}\mathbb{E}[f(\{(s_{h,l}^j,a_{h,l}^j)\}_{j\in[i-1]} \cup \{(s,a)\}) - f(\{(s_{h,l}^j,a_{h,l}^j)\}_{j\in[i-1]})] &= \mathbb{E}[\Delta f(\{(s_{h,l}^j,a_{h,l}^j)\}_{j\in[i-1]},(s,a))] \\ &= R_{i,h}^k(s,a|\pi_{[i-1]})\end{aligned}$$

根据霍夫丁不等式，对于任意固定的 $(s,a) \in \mathcal{S}\times\mathcal{A}$ 和$h \in [H]$，以至少 $1-\frac{\delta}{K|\mathcal{S}||\mathcal{A}|H}$ 的概率，有

$$|R_{i,h}^k(s,a|\pi_{[i-1]}) - \widehat{R}_{i,h}^k(s,a|\pi_{[i-1]})| \le \frac{\epsilon}{KH}$$

通过对所有 $i \in [K], h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$ 取并集界，我们得到事件 $\mathcal{E}_1$ 以至少 $1-\delta/3$ 的概率成立。

接下来，由于

$$\begin{aligned}\widehat{P}_{i,h}^k(\cdot|s,a) &= \frac{\sum_{e=1}^k \mathbb{I}(s_{h,e}^i = s, a_{h,e}^i = a, s_{h+1,e}^i = s')}{\sum_{e=1}^k \mathbb{I}(s_{h,e}^i = s, a_{h,e}^i = a)} \\ &= \frac{\sum_{e=1}^k \mathbb{I}(s_{h,e}^i = s, a_{h,e}^i = a)\cdot \mathbb{I}(s_{h+1,e}^i = s')}{\sum_{e=1}^k \mathbb{I}(s_{h,e}^i = s, a_{h,e}^i = a)}\end{aligned}$$

注意，在访问 (s,a) 的条件下，变量 $\mathbb{I}(s_{h+1,e}^i = s')$ 是一个二元随机变量，满足 $\mathbb{E}[\mathbb{I}(s_{h+1,e}^i = s' | s_{h,e}^i = s, a_{h,e}^i = a] = P_{i,h}(s'|s,a)$ $Var[\mathbb{I}(s_{h+1,e}^i = s' | s_{h,e}^i = s, a_{h,e}^i = a] = P_{i,h}(s'|s,a)(1-P_{i,h}(s'|s,a)) \le P_{i,h}(s'|s,a)$。

通过应用引理中的标准 Bernstein 不等式，并对 $(s,a), i \in [K], k \in [T], h \in [H]$ 取并集界，我们可以得到事件 $\mathcal{E}_2$ 以概率 $1-\frac{\delta}{3}$ 成立。

由于 $\mathcal{E}$ 是上述事件的交集，我们可以通过再次对这两个事件取并集界来证明该引理。

□

D.2 关键引理

对于给定策略 π，我们将回合 k 在估计转移动态 $\widehat{P}^k$ 下的价值函数定义为 $\bar{V}^{k}$:

$$\bar{V}_h^{\pi,k}(\mathbf{s}) = \mathbb{E}_{\widehat{P}^k,\pi}\left[\sum_{t=h}^{H}\left(r(\mathbf{s}_t,\mathbf{a}_t) + \sum_{i=1}^{K} b(N_{i,t}^k(s_t^i,a_t^i))\right)\middle|\mathbf{s}_h = \mathbf{s}\right] \tag{11}$$

and

$$\bar{Q}_h^{\pi,k}(\mathbf{s},\mathbf{a}) = \mathbb{E}_{\widehat{P}^k,\pi}\left[\sum_{t=h}^{H}\left(r(\mathbf{s}_t,\mathbf{a}_t) + \sum_{i=1}^{K} b(N_{i,t}^k(s_t^i,a_t^i))\right)\middle|\mathbf{s}_h = \mathbf{s},\mathbf{a}_h = \mathbf{a}\right],$$

其中 $b(N) = \sqrt{\frac{2SH^2\iota}{N}} + \frac{3SH\iota}{N}$ 是奖励项。当联合策略可分解，且智能体 i 的策略记为 π_i时，我们可以定义在 $\widehat{P}^k$ 下的以下边际价值函数:

$$\bar{V}_h^{\pi_i,k}(s;\pi_{[i-1]}^k) = \mathbb{E}_{\widehat{P}^k}\left[\sum_{t=h}^{H}\Delta r_i(\mathbf{s}_t,\mathbf{a}_t) + b(N_{i,t}^k(s_t^i,a_t^i))\middle|(\mathbf{s}^{[i-1]},\mathbf{a}^{[i-1]}) \sim \pi_{[i-1]}^k, s_h^i = s\right]$$

$$\bar{Q}_h^{\pi_i,k}(s,a;\pi_{[i-1]}^k) = \mathbb{E}_{\widehat{P}^k}\left[\sum_{t=h}^{H}\Delta r_i(\mathbf{s}_t,\mathbf{a}_t) + b(N_{i,t}^k(s_t^i,a_t^i))\middle|(\mathbf{s}^{[i-1]},\mathbf{a}^{[i-1]}) \sim \pi_{[i-1]}^k, s_h^i = s, a_h^i = a\right].$$

通过望远镜求和，我们有:

$$\bar{V}_1^{\pi,k}(\bar{\mathbf{s}}_1) = \sum_{i=1}^{K}\bar{V}_1^{\pi_i,k}(\bar{s}_1^i;\pi_{[i-1]}^k). \tag{12}$$

令 $\widehat{V}_{i,h}^{k}(s)$ 表示算法2中回合 k 的第 h 步时智能体 i 的估计价值函数 $\widehat{V}_{i,h}(s)$，并令 π_i^*表示全局最优策略 π^*下智能体 i的边际策略。

引理 7. 在引理 6 中事件 $\mathcal{E}$ 发生的条件下，对于所有 $i \in [K], h \in [H], s \in \mathcal{S}, a \in \mathcal{A}$ ：

$$\bar{V}_h^{\pi_i^*,k}(s;\pi_{[i-1]}^k) \le \widehat{V}_{i,h}^k(s) \le \bar{V}_h^{\pi_i^k,k}(s;\pi_{[i-1]}^k) + \frac{2\epsilon(H-h+1)}{KH}$$

and

$$\bar{Q}_h^{\pi_i^*,k}(s,a;\pi_{[i-1]}^k) \le \widehat{Q}_{i,h}^k(s,a) \le \bar{Q}_h^{\pi_i^k,k}(s,a;\pi_{[i-1]}^k) + \frac{2\epsilon(H-h+1)}{KH}.$$

证明。 应用转移动态 $\widehat{P}^k$ 下的贝尔曼方程（引理 1），我们有

$$\bar{V}_h^{\pi_i,k}(s;\pi_{[i-1]}^k) = \sum_{a\in\mathcal{A}}\pi_{i,h}(s,a)\bar{Q}_h^{\pi_i,k}(s,a;\pi_{[i-1]}^k)$$

$$\bar{Q}_h^{\pi_i,k}(s,a;\pi_{[i-1]}^k) = R_{i,h}^k(s,a|\pi_{[i-1]}^k) + b(N_{i,h}^k(s,a)) + \sum_{s'\in\mathcal{S}}\widehat{P}_{i,h}^k(s'|s,a)\bar{V}_{h+1}^{\pi_i,k}(s';\pi_{[i-1]}^k).$$

第一部分：对 h 进行归纳法证明（从 H 向后推至 1）。

基本情形（$h = H$）：根据定义，

$$\bar{Q}_H^{\pi_i^*,k}(s,a;\pi_{[i-1]}^k) = \bar{Q}_H^{\pi_i^k,k}(s,a;\pi_{[i-1]}^k) = R_{i,H}^k(s,a|\pi_{[i-1]}^k) + b(N_{i,H}^k(s,a))$$

and

$$\widehat{Q}_{i,H}^k(s,a) = \widehat{R}_{i,H}^k(s,a|\pi_{[i-1]}^k) + b(N_{i,H}^k(s,a)) + \frac{\epsilon}{KH}.$$

在事件条件下 $\mathcal{E}_1$ 在引理 6 中，我们有

$$\bar{Q}_H^{\pi_i^*,k}(s,a;\pi_{[i-1]}^k) \le \widehat{Q}_{i,H}^k(s,a) \le \bar{Q}_H^{\pi_i^k,k}(s,a;\pi_{[i-1]}^k) + \frac{2\epsilon}{KH}. \tag{13}$$

对于价值函数，由贝尔曼方程我们有 $\bar{V}^{\pi_i^*,k}_H(s;\pi^k_{[i-1]}) = \sum_{a\in\mathcal{A}} \pi^*_{i,H}(s,a)\,\bar{Q}^{\pi_i^*,k}_H(s,a;\pi^k_{[i-1]})$。由于 $\sum_{a\in\mathcal{A}} \pi^*_{i,H}(s,a) = 1$,

$$\begin{aligned}
\bar{V}^{\pi_i^*,k}_H(s;\pi^k_{[i-1]}) &\le \max_{a\in\mathcal{A}} \bar{Q}^{\pi_i^*,k}_H(s,a;\pi^k_{[i-1]}) \\
&\le \max_{a\in\mathcal{A}} \widehat{Q}^k_{i,H}(s,a) && \text{(by (13))} \\
&= \widehat{V}^k_{i,H}(s). && \text{(by definition)}
\end{aligned}$$

根据输出策略的定义，$\widehat{V}^k_{i,H}(s) = \widehat{Q}^k_{i,H}(s,\pi^k_{i,H}(s))$ 和 $\bar{V}^{\pi^k_i,k}_H(s;\pi^k_{[i-1]}) = \bar{Q}^{\pi^k_i,k}_H(s,\pi^k_{i,H}(s);\pi^k_{[i-1]})$。因此

$$\begin{aligned}
\widehat{V}^k_{i,H}(s) &= \widehat{Q}^k_{i,H}(s,\pi^k_{i,H}(s)) \\
&\le \bar{Q}^{\pi^k_i,k}_H(s,\pi^k_{i,H}(s);\pi^k_{[i-1]}) + \frac{2\epsilon}{KH} && \text{(by (13))} \\
&= \bar{V}^{\pi^k_i,k}_H(s;\pi^k_{[i-1]}) + \frac{2\epsilon}{KH}.
\end{aligned}$$

这就证明了基本情形。

归纳步骤：假设结果对第 $h+1$ 步成立。对于第 h 步，根据贝尔曼方程,

$$\begin{aligned}
\bar{Q}^{\pi_i^*,k}_h(s,a;\pi^k_{[i-1]}) &= R^k_{i,h}(s,a|\pi^k_{[i-1]}) + b(N^k_{i,h}(s,a)) + \sum_{s'\in\mathcal{S}} \widehat{P}^k_{i,h}(s'|s,a)\bar{V}^{\pi_i^*,k}_{h+1}(s';\pi^k_{[i-1]}) \\
&\le R^k_{i,h}(s,a|\pi^k_{[i-1]}) + b(N^k_{i,h}(s,a)) + \sum_{s'\in\mathcal{S}} \widehat{P}^k_{i,h}(s'|s,a)\widehat{V}^k_{i,h+1}(s') \\
&\le \widehat{R}^k_{i,h}(s,a|\pi^k_{[i-1]}) + b(N^k_{i,h}(s,a)) + \frac{\epsilon}{KH} + \sum_{s'\in\mathcal{S}} \widehat{P}^k_{i,h}(s'|s,a)\widehat{V}^k_{i,h+1}(s') && \text{(by } \mathcal{E}_1\text{)} \\
&= \widehat{Q}^k_{i,h}(s,a), && \text{（根据定义）}
\end{aligned}$$

其中，第一个不等式由结果对 $h+1$ 成立这一假设得出。对于上界，根据归纳假设和贝尔曼方程,

$$\begin{aligned}
\widehat{Q}^k_{i,h}(s,a) &= \widehat{R}^k_{i,h}(s,a|\pi^k_{[i-1]}) + b(N^k_{i,h}(s,a)) + \frac{\epsilon}{KH} + \sum_{s'\in\mathcal{S}} \widehat{P}^k_{i,h}(s'|s,a)\widehat{V}^k_{i,h+1}(s') \\
&\le R^k_{i,h}(s,a|\pi^k_{[i-1]}) + b(N^k_{i,h}(s,a)) + \frac{2\epsilon}{KH} \\
&\quad + \sum_{s'\in\mathcal{S}} \widehat{P}^k_{i,h}(s'|s,a)\left\{\bar{V}^{\pi_i^k,k}_{h+1}(s';\pi^k_{[i-1]}) + \frac{2\epsilon(H-h)}{KH}\right\} && \text{(by IH and } \mathcal{E}_1\text{)} \\
&= \bar{Q}^{\pi_i^k,k}_h(s,a;\pi^k_{[i-1]}) + \frac{2\epsilon}{KH} + \frac{2\epsilon(H-h)}{KH} \\
&= \bar{Q}^{\pi_i^k,k}_h(s,a;\pi^k_{[i-1]}) + \frac{2\epsilon(H-h+1)}{KH}.
\end{aligned}$$

对于价值函数，根据贝尔曼方程,

$$\begin{aligned}
\bar{V}^{\pi_i^*,k}_h(s;\pi^k_{[i-1]}) &= \sum_{a\in\mathcal{A}} \pi^*_{i,h}(s,a)\bar{Q}^{\pi_i^*,k}_h(s,a;\pi^k_{[i-1]}) \\
&\le \max_{a\in\mathcal{A}} \bar{Q}^{\pi_i^*,k}_h(s,a;\pi^k_{[i-1]}) \\
&\le \max_{a\in\mathcal{A}} \widehat{Q}^k_{i,h}(s,a) = \widehat{V}^k_{i,h}(s).
\end{aligned}$$

由于 $\widehat{V}^k_{i,h}(s) = \widehat{Q}^k_{i,h}(s, \pi^k_{i,h}(s))$，

$$\begin{aligned}\widehat{V}^k_{i,h}(s) &= \widehat{Q}^k_{i,h}(s, \pi^k_{i,h}(s)) \\ &\le \bar{Q}^{\pi^k_i, k}_h(s, \pi^k_{i,h}(s); \pi^k_{[i-1]}) + \frac{2\epsilon(H-h+1)}{KH} \\ &= \bar{V}^{\pi^k_i, k}_h(s; \pi^k_{[i-1]}) + \frac{2\epsilon(H-h+1)}{KH}.\end{aligned}$$

至此，证明完成。 □

接下来，我们证明以下引理，它保证输出策略相对于经验转移动力学可近似达到 1/2-近似比。

引理 8. 在算法2的执行过程中，对于任意 $k \in [T]$，$\bar{V}^{\pi^k}_{1,k}(\bar{\mathbf{s}}_1) \ge \bar{V}^{\pi^*}_{1,k}(\bar{\mathbf{s}}_1)/2 - \epsilon$.

证明。假设我们有两组 K 个智能体，第一组的 K 个智能体遵循算法2在回合 k 中执行的策略 π^k 和转移动态 $\hat{P}^k$。我们用 $\{(s^i_h, a^i_h)\}_{h\in[H], i\in[K]}$ 表示其轨迹。第二组的 K 个智能体遵循最优策略 π^* 和转移动态 $\hat{P}^k$，我们用 $\{(\tilde{s}^i_h, \tilde{a}^i_h)\}_{h\in[H], i\in[K]}$ 表示第二组智能体的轨迹。此外，我们定义函数 $\widehat{f}^k\colon 2^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}$，使得 $\widehat{f}^k(S) = f(S) + \sum_{(s,a)\in S} b(N^k_i, h(s,a))$。由于 $\sum_{(s,a)\in S} b(N^k_{i,h}(s,a))$ 是线性函数，且 $b(N)$ 对每个 $N \ge 0$ 都为正，因此 $\widehat{f}^k(S)$ 也是单调且子模的。为简化记号，我们定义

$$G^k_i = \mathbb{E}_{\hat{P}^k}[\sum_{h=1}^{H} \Delta\widehat{f}^k(\{(s^j_h, a^j_h)\}_{j\in[K]} \cup \{(\tilde{s}^m_h, \tilde{a}^m_h)\}_{m\in[i-1]}, (\tilde{s}^i_h, \tilde{a}^i_h))|(\mathbf{s}, \mathbf{a}) \sim \pi^k, (\tilde{\mathbf{s}}, \tilde{\mathbf{a}}) \sim \pi^*], \tag{14}$$

其中，$(\mathbf{s}, \mathbf{a}) \sim \pi^k$，$(\tilde{\mathbf{s}}, \tilde{\mathbf{a}}) \sim \pi^*$ 表示 $\{(s^i_h, a^i_h)\}_{h\in[H], i\in[K]}$ 是根据策略 π^k 采样得到的，而 $\{(\tilde{s}^i_h, \tilde{a}^i_h)\}_{h\in[H], i\in[K]}$ 是根据最优策略 π^* 采样得到的。根据定义，我们有

$$G^k_i = \mathbb{E}_{\hat{P}^k}[\sum_{h=1}^{H} \widehat{f}^k\{(s^j_h, a^j_h)\}_{j\in[K]} \cup \{(\tilde{s}^m_h, \tilde{a}^m_h)\}_{m\in[i]}) - \widehat{f}^k\{(s^j_h, a^j_h)\}_{j\in[K]} \cup \{(\tilde{s}^m_h, \tilde{a}^m_h)\}_{m\in[i-1]})|(\mathbf{s}, \mathbf{a}) \sim \pi^k, (\tilde{\mathbf{s}}, \tilde{\mathbf{a}}) \sim \pi^*].$$

由此可得

$$\begin{aligned}\sum_{i=1}^{K} G^k_i &= \mathbb{E}_{\hat{P}^k}[\sum_{h=1}^{H} \widehat{f}^k\{(s^j_h, a^j_h)\}_{j\in[K]} \cup \{(\tilde{s}^m_h, \tilde{a}^m_h)\}_{m\in[K]}) - \widehat{f}^k\{(s^j_h, a^j_h)\}_{j\in[K]})|(\mathbf{s}, \mathbf{a}) \sim \pi^k, (\tilde{\mathbf{s}}, \tilde{\mathbf{a}}) \sim \pi^*] \\ &\ge \mathbb{E}_{\hat{P}^k}[\sum_{h=1}^{H} \widehat{f}^k\{(\tilde{s}^m_h, \tilde{a}^m_h)\}_{m\in[K]}) - \widehat{f}^k\{(s^j_h, a^j_h)\}_{j\in[K]})|(\mathbf{s}, \mathbf{a}) \sim \pi^k, (\tilde{\mathbf{s}}, \tilde{\mathbf{a}}) \sim \pi^*] \\ &= \mathbb{E}_{\hat{P}^k}[\sum_{h=1}^{H} \widehat{f}^k\{(\tilde{s}^m_h, \tilde{a}^m_h)\}_{m\in[K]})|(\tilde{\mathbf{s}}, \tilde{\mathbf{a}}) \sim \pi^*] - \mathbb{E}_{\hat{P}^k}[\sum_{h=1}^{H} \widehat{f}^k\{(s^j_h, a^j_h)\}_{j\in[K]})|(\mathbf{s}, \mathbf{a}) \sim \pi^k] \\ &= \bar{V}^{\pi^*, k}_1(\bar{\mathbf{s}}_1) - \bar{V}^{\pi^k, k}_1(\bar{\mathbf{s}}_1),\end{aligned} \tag{15}$$

其中，不等式成立是因为 f 是单调的，而最后一个不等式则根据式（11）中给出的 $\bar{V}^{\pi, k}_1(\bar{\mathbf{s}}_1)$ 的定义成立。接下来，我们对每个 i 界定 G^k_i。由于 $\{(s^l_h, a^l_h)\}_{l\in[i-1]} \subseteq \{(s^j_h, a^j_h)\}_{j\in[K]} \cup \{(\tilde{s}^m_h, \tilde{a}^m_h)\}_{m\in[i]}$，根据子模性，我们有

$$G^k_i \le \mathbb{E}[\sum_{h=1}^{H} \Delta\widehat{f}^k(\{(s^l_h, a^l_h)\}_{l\in[i-1]}, (\tilde{s}^i_h, \tilde{a}^i_h))|(\mathbf{s}, \mathbf{a}) \sim \pi^k, (\tilde{\mathbf{s}}, \tilde{\mathbf{a}}) \sim \pi^*]. \tag{16}$$

令策略 $\pi^*_{i,h}$ 表示全局策略为 π^* 时第 i 个智能体在时间步 h 的边际策略。为简化记法，下文也用 $\pi^*_i = \{\pi^*_{i,h}\}_{h\in[H]}$ 表示智能体 i 在整个

回合中的边际策略。因此，

$$\begin{aligned}&\mathbb{E}[\sum_{h=1}^{H}\Delta\widehat{f}^k(\{(s_h^l,a_h^l)\}_{l\in[i-1]},(\tilde{s}_h^i,\tilde{a}_h^i))|(\mathbf{s},\mathbf{a})\sim\pi^k,(\tilde{\mathbf{s}},\tilde{\mathbf{a}})\sim\pi^*]\\=&\mathbb{E}[\sum_{h=1}^{H}\Delta\widehat{f}^k(\{(s_h^l,a_h^l)\}_{l\in[i-1]},(\tilde{s}_h^i,\tilde{a}_h^i))|(\mathbf{s},\mathbf{a})\sim\pi^k,(\tilde{s}^i,\tilde{a}^i)\sim\pi_i^*]\\=&\bar{V}_1^{\pi_i^*,k}(\bar{s}_1;\pi_{[i-1]}).\end{aligned}$$

将上述不等式与（16）结合，可得

$$G_i^k\leq\bar{V}_1^{\pi_i^*,k}(\bar{s}_1;\pi_{[i-1]}).\tag{17}$$

接下来，根据引理 7 的结果，可得

$$G_i^k\leq\bar{V}_1^{\pi_i^k,k}(\bar{s}_1^i;\pi_{[i-1]})+2\frac{\epsilon}{K}$$

由于根据（15），$\sum_{i=1}^K G_i^k\geq\bar{V}^{\pi^*,k}_1(\bar{\mathbf{s}}_1)-\bar{V}^{\pi^k,k}_1(\bar{\mathbf{s}}_1)$，因此可得

$$\begin{aligned}\bar{V}_1^{\pi^*,k}(\bar{\mathbf{s}}_1)-\bar{V}_1^{\pi^k,k}(\bar{\mathbf{s}}_1)&\leq\sum_{i=1}^{K}\{\bar{V}_1^{\pi_i^k,k}(\bar{s}_1^i;\pi_{[i-1]})+\frac{2\epsilon}{K}\}\\&=\bar{V}_1^{\pi^k,k}(\bar{\mathbf{s}}_1)+2\epsilon.\end{aligned}$$

最后一个等号由（12）中的价值分解得出。重新整理上述不等式，即可证明该定理中的结论。

□

接下来，我们证明以下分解引理。

引理 9。 对于任意联合状态-动作对$(\mathbf{s},\mathbf{a})\in\mathcal{S}^K\times\mathcal{A}^K$以及满足条件 $0\leq G(\mathbf{s})\leq H$的任意函数 $G:\ \mathcal{S}^K\to\mathbb{R}$，对于所有$\mathbf{s}\in\mathcal{S}^K$，我们有

$$\sum_{\mathbf{s}'\in\mathcal{S}^K}[\widehat{\mathbf{P}}_h^k(\mathbf{s}'|\mathbf{s},\mathbf{a})-\mathbf{P}_h(\mathbf{s}'|\mathbf{s},\mathbf{a})]G(\mathbf{s}')\leq\frac{1}{H}\sum_{\mathbf{s}'\in\mathcal{S}^K}\mathbf{P}_h(\mathbf{s}'|\mathbf{s},\mathbf{a})G(\mathbf{s}')+\sum_{i=1}^{K}\frac{10SH^2K\iota}{N_{i,h}^k(s^i,a^i)}.$$

证明。 该证明利用多智能体转移动态的乘积结构，通过对智能体进行望远镜求和论证，并结合经过仔细校准的集中不等式应用。

第 1 步：望远镜分解。 由智能体转移的独立性，联合转移概率可分解为：

$$\widehat{\mathbf{P}}_h^k(\mathbf{s}'|\mathbf{s},\mathbf{a})=\prod_{i=1}^{K}\widehat{P}_{i,h}^k(s'^i|s^i,a^i),\quad\mathbf{P}_h(\mathbf{s}'|\mathbf{s},\mathbf{a})=\prod_{i=1}^{K}P_{i,h}(s'^i|s^i,a^i).$$

我们的目标量变为：

$$\begin{aligned}&\sum_{\mathbf{s}'\in\mathcal{S}^K}\widehat{\mathbf{P}}_h^k(\mathbf{s}'|\mathbf{s},\mathbf{a})G(\mathbf{s}')-\sum_{\mathbf{s}'\in\mathcal{S}^K}\mathbf{P}_h(\mathbf{s}'|\mathbf{s},\mathbf{a})G(\mathbf{s}')\\&=\sum_{\mathbf{s}'\in\mathcal{S}^K}\left[\prod_{i=1}^{K}\widehat{P}_{i,h}^k(s'^i|s^i,a^i)-\prod_{i=1}^{K}P_{i,h}(s'^i|s^i,a^i)\right]G(\mathbf{s}').\end{aligned}\tag{18}$$

为分解这一乘积之差，我们引入一系列中间项，逐步从全部使用真实概率过渡到全部使用经验概率。对于$i\in\{0,1,\ldots,K\}$，定义：

$$Y_i:=\sum_{\mathbf{s}'\in\mathcal{S}^K}\left[\prod_{l=1}^{i}\widehat{P}_{l,h}^k(s'^l|s^l,a^l)\right]\left[\prod_{m=i+1}^{K}P_{m,h}(s'^m|s^m,a^m)\right]G(\mathbf{s}'),\tag{19}$$

其中，我们约定空乘积等于 1。通过这一构造，我们可以一次改变一个智能体的转移，从而弥合经验期望与真实期望之间的差距。利用望远镜求和，我们可以将总差异写成一系列增量变化之和：

$$\sum_{\mathbf{s}'\in\mathcal{S}^K}\widehat{\mathbf{P}}_h^k(\mathbf{s}'|\mathbf{s},\mathbf{a})G(\mathbf{s}')-\sum_{\mathbf{s}'\in\mathcal{S}^K}\mathbf{P}_h(\mathbf{s}'|\mathbf{s},\mathbf{a})G(\mathbf{s}')=Y_K-Y_0=\sum_{i=1}^{K}(Y_i-Y_{i-1}). \tag{20}$$

每个差值 Y_i-Y_{i-1} 刻画了在保持所有其他智能体的转移不变的情况下，将智能体 i的真实转移概率替换为其经验估计所产生的影响。这隔离了单个智能体误差，并使我们能够对每个智能体分别应用集中不等式。

第2步：界定单个智能体贡献。根据定义(19)，项 Y_i 与 Y_{i-1} 仅在智能体 i的转移概率上有所不同：

$$\begin{aligned}
Y_i-Y_{i-1}&=\sum_{\mathbf{s}'\in\mathcal{S}^K}\prod_{l=1}^{i-1}\widehat{P}_{l,h}^k(s'^l|s^l,a^l)\prod_{m=i+1}^{K}P_{m,h}(s'^m|s^m,a^m)G(\mathbf{s}')(\widehat{P}_{i,h}^k(s'^i|s^i,a^i)-P_{i,h}(s'^i|s^i,a^i))\\
&\le\sum_{\mathbf{s}'\in\mathcal{S}^K}\prod_{l=1}^{i-1}\widehat{P}_{l,h}^k(s'^l|s^l,a^l)\prod_{m=i+1}^{K}P_{m,h}(s'^m|s^m,a^m)G(\mathbf{s}')|\widehat{P}_{i,h}^k(s'^i|s^i,a^i)-P_{i,h}(s'^i|s^i,a^i)|\\
&=\sum_{s'\in\mathcal{S}}|\widehat{P}_{i,h}^k(s'|s^i,a^i)-P_{i,h}(s'|s^i,a^i)|\sum_{\mathbf{s}':s'^i=s'}\prod_{l=1}^{i-1}\widehat{P}_{l,h}^k(s'^l|s^l,a^l)\prod_{m=i+1}^{K}P_{m,h}(s'^m|s^m,a^m)G(\mathbf{s}').
\end{aligned}$$

此外，我们记 $Z_i(s')=\sum_{\mathbf{s}':s'^i=s'}\prod_{l=1}^{i-1}\widehat{P}_{l,h}^k(s'^l|s^l,a^l)\prod_{m=i+1}^{K}P_{m,h}(s'^m|s^m,a^m)G(\mathbf{s}')$。则

$$Y_i-Y_{i-1}\le\sum_{s'\in\mathcal{S}}|\widehat{P}_{i,h}^k(s'|s^i,a^i)-P_{i,h}(s'|s^i,a^i)|Z_i(s').$$

根据引理 6中的事件 $\mathcal{E}_2$，我们有

$$\begin{aligned}
\sum_{s'\in\mathcal{S}}|\widehat{P}_{i,h}^k(s'|s^i,a^i)-P_{i,h}(s'|s^i,a^i)|Z_i(s')&\le\sum_{s'\in\mathcal{S}}\left(\sqrt{\frac{2P_{i,h}(s'|s^i,a^i)\iota}{N_{i,h}^k(s^i,a^i)}}+\frac{3\iota}{N_{i,h}^k(s^i,a^i)}\right)Z_i(s')\\
&\le\sum_{s'\in\mathcal{S}}\left(\frac{P_{i,h}(s'|s^i,a^i)}{2HK}+\frac{HK\iota}{N_{i,h}^k(s^i,a^i)}+\frac{3\iota}{N_{i,h}^k(s^i,a^i)}\right)Z_i(s')\\
&\le\frac{\sum_{s'\in\mathcal{S}}P_{i,h}(s'|s^i,a^i)Z_i(s')}{2HK}+\frac{4HK\iota}{N_{i,h}^k(s^i,a^i)}\sum_{s'\in\mathcal{S}}Z_i(s'),
\end{aligned} \tag{21}$$

其中第二个不等式由 $\sqrt{ab}\le\frac{a+b}{2}$ 得出，且 $a=\frac{P_{i,h}(s'|s^i,a^i)}{2HK}$ 以及 $b=\frac{2HK\iota}{N_{i,h}^k(s^i,a^i)}$。

由于根据假设有 $G(\mathbf{s}')\le H$，我们可以通过提出因子 H来界定 $Z_i(s')$：

$$Z_i(s')\le H\sum_{\mathbf{s}':s'^i=s'}\prod_{l=1}^{i-1}\widehat{P}_{l,h}^k(s'^l|s^l,a^l)\prod_{m=i+1}^{K}P_{m,h}(s'^m|s^m,a^m)=H,$$

根据定义，我们有

$$\begin{aligned}
\sum_{s'\in\mathcal{S}}P_{i,h}(s'|s^i,a^i)Z_i(s')&=\sum_{\mathbf{s}'\in\mathcal{S}^K}\left[\prod_{l=1}^{i-1}\widehat{P}_{l,h}^k(s'^l|s^l,a^l)\right]\left[\sum_{s'^i\in\mathcal{S}}P_{i,h}(s'^i|s^i,a^i)\right]\left[\prod_{m=i+1}^{K}P_{m,h}(s'^m|s^m,a^m)\right]G(\mathbf{s}')\\
&=\sum_{\mathbf{s}'\in\mathcal{S}^K}\left[\prod_{l=1}^{i-1}\widehat{P}_{l,h}^k(s'^l|s^l,a^l)\right]\left[\prod_{m=i}^{K}P_{m,h}(s'^m|s^m,a^m)\right]G(\mathbf{s}')=Y_{i-1},
\end{aligned} \tag{22}$$

结合（21），我们可以得到

$$\begin{aligned} Y_i - Y_{i-1} &\le \frac{Y_{i-1}}{2HK} + \frac{4HK\iota}{N_{i,h}^k(s^i,a^i)} \sum_{s'\in\mathcal{S}} Z_i(s') \\ &\le \frac{Y_{i-1}}{2HK} + \frac{4H^2SK\iota}{N_{i,h}^k(s^i,a^i)}. \end{aligned} \tag{23}$$

由 (20) 和 (23)，我们得到

$$\sum_{\mathbf{s}'} \widehat{\mathbf{P}}_h^k(\mathbf{s}'|\mathbf{s},\mathbf{a})G(\mathbf{s}') - \sum_{\mathbf{s}'} \mathbf{P}_h(\mathbf{s}'|\mathbf{s},\mathbf{a})G(\mathbf{s}') \le \frac{1}{2HK}\sum_{i=1}^K Y_{i-1} + \sum_{i=1}^K \frac{4SH^2K\iota}{N_{i,h}^k(s^i,a^i)}. \tag{24}$$

为了界定 $\sum_{i=1}^K Y_{i-1}$，我们需要对每个 Y_i 给出一个上界。

步 3：关于 Y_i 的精细递归界。接下来，我们递归地界定 Y_i。使用与 $a = P_{i,h}(s'|s^i,a^i)$ 和 $b = \frac{K\iota}{N_{i,h}^k(s^i,a^i)}$ 类似的论证，我们得到

$$\sum_{s'\in\mathcal{S}} |\widehat{P}_{i,h}^k(s'|s^i,a^i) - P_{i,h}(s'|s^i,a^i)| \le \frac{\sum_{s'\in\mathcal{S}} P_{i,h}(s'|s^i,a^i)}{K} + \frac{4K\iota}{N_{i,h}^k(s^i,a^i)}.$$

与步 2 类似，我们可以证明

$$\begin{aligned} Y_i - Y_{i-1} &\le \sum_{s'\in\mathcal{S}} |\widehat{P}_{i,h}^k(s'|s^i,a^i) - P_{i,h}(s'|s^i,a^i)| Z_i(s') \\ &\le \frac{\sum_{s'\in\mathcal{S}} P_{i,h}(s'|s^i,a^i) Z_i(s')}{K} + \frac{4HK\iota}{N_{i,h}^k(s^i,a^i)} \sum_{s'\in\mathcal{S}} Z_i(s') \\ &\le \frac{1}{K} Y_{i-1} + \frac{4HSK\iota}{N_{i,h}^k(s^i,a^i)}, \end{aligned}$$

由此得到

$$Y_i \le (1+\frac{1}{K})Y_{i-1} + \frac{4HSK\iota}{N_{i,h}^k(s^i,a^i)}.$$

通过递归，我们得到

$$Y_i \le (1+\frac{1}{K})^i Y_0 + \sum_{m=1}^i \frac{4SHK\iota}{N_{m,h}^k(s^m,a^m)} (1+\frac{1}{K})^{i-m}.$$

因此，

$$\begin{aligned} \sum_{i=1}^K Y_{i-1} &\le \sum_{i=1}^K (1+\frac{1}{K})^{i-1} Y_0 + \sum_{i=2}^K \sum_{m=1}^{i-1} \frac{4SHK\iota}{N_{m,h}^k(s^m,a^m)} (1+\frac{1}{K})^{i-1-m} \\ &\le \frac{(1+\frac{1}{K})^K - 1}{\frac{1}{K}} Y_0 + \sum_{m=1}^{K-1} \sum_{i=m+1}^K \frac{4SHK\iota}{N_{m,h}^k(s^m,a^m)} (1+\frac{1}{K})^{i-1-m} \\ &\le K[(1+\frac{1}{K})^K - 1] Y_0 + \sum_{m=1}^{K-1} \frac{4SHK\iota}{N_{m,h}^k(s^m,a^m)} \sum_{j=0}^{K-m-1} (1+\frac{1}{K})^j \\ &\le K(e-1)Y_0 + \sum_{m=1}^{K-1} \frac{4SHK^2\iota}{N_{m,h}^k(s^m,a^m)} e, \end{aligned} \tag{25}$$

其中我们使用了 $(1+\frac{1}{K})^K \le e$ 和 $\sum_{j=0}^K (1+\frac{1}{K})^j \le K\Big((1+\frac{1}{K})^K - 1\Big)$。

步骤 4：最终界。将（25）代入（24）：

$$\sum_{\mathbf{s}'} \widehat{\mathbf{P}}_h^k(\mathbf{s}'|\mathbf{s},\mathbf{a})G(\mathbf{s}') - \sum_{\mathbf{s}'} \mathbf{P}_h(\mathbf{s}'|\mathbf{s},\mathbf{a})G(\mathbf{s}')$$
$$\leq \frac{e-1}{2H} Y_0 + 2SKe\iota \sum_{m=1}^{K-1} \frac{1}{N_{m,h}^k(s^m,a^m)} + \sum_{i=1}^{K} \frac{4SH^2K\iota}{N_{i,h}^k(s^i,a^i)}.$$

由于 $e<3$，we 有 $\frac{e-1}{2H} < \frac{1}{H}$ 和 $2SKe\iota < 6SK\iota$。将第一个求和式扩展为包含 $i=K$，并结合 g:

$$\sum_{\mathbf{s}'} \widehat{\mathbf{P}}_h^k(\mathbf{s}'|\mathbf{s},\mathbf{a})G(\mathbf{s}') - \sum_{\mathbf{s}'} \mathbf{P}_h(\mathbf{s}'|\mathbf{s},\mathbf{a})G(\mathbf{s}')$$
$$\leq \frac{1}{H} Y_0 + \sum_{i=1}^{K} \frac{6SK\iota + 4SH^2K\iota}{N_{i,h}^k(s^i,a^i)}$$
$$\leq \frac{1}{H} Y_0 + \sum_{i=1}^{K} \frac{10SH^2K\iota}{N_{i,h}^k(s^i,a^i)}.$$

由此完成证明。 □

D.3 乐观估计

回顾 $\bar{V}^{\pi,k}_h(\mathbf{s})$ 的定义：

$$\bar{V}_h^{\pi,k}(\mathbf{s}) = \mathbb{E}_{\widehat{P}^k,\pi}\left[\sum_{t=h}^{H}\left(r(\mathbf{s}_t,\mathbf{a}_t) + \sum_{i=1}^{K} b(N_{i,t}^k(s_t^i,a_t^i))\right)\middle|\mathbf{s}_h = \mathbf{s}\right].$$

本节将证明，对于任意策略 π，$\bar{V}^{\pi,k}_h(\mathbf{s})$ 确实是真实价值函数 V_h^π 的乐观估计。首先，我们证明以下引理。

引理 10. 在事件 $\mathcal{E}$（引理 6）发生的条件下，对于任意 $\mathbf{s} \in \mathcal{S}^K$、$\mathbf{a} \in \mathcal{A}^K$, $h \in [H]$, $k \in [T]$，以及任意满足 $V(\mathbf{s}) \in [0,H]$ 的函数V: $\mathcal{S}^K \to \mathbb{R}$，

$$\left|\sum_{\mathbf{s}'} \widehat{\mathbf{P}}_h^k(\mathbf{s}'|\mathbf{s},\mathbf{a})V(\mathbf{s}') - \sum_{\mathbf{s}'} \mathbf{P}_h(\mathbf{s}'|\mathbf{s},\mathbf{a})V(\mathbf{s}')\right| \leq \sum_{i=1}^{K} b(N_{i,h}^k(s^i,a^i)),$$

其中 $b(N) = H\sqrt{\frac{2S\iota}{N}} + \frac{3SH\iota}{N}$，且 $\iota = \log(6S^2ATHK/\delta)$。

证明. 该引理的证明与引理 9 的证明类似。由智能体转移的独立性可知，$\mathbf{P}_h(\mathbf{s}'|\mathbf{s},\mathbf{a}) = \prod_{i=1}^{K} P_{i,h}(s'^i|s^i,a^i)$，对于 $\widehat{\mathbf{P}}_h^k$ 也同样成立。定义

$$X_i := \sum_{\mathbf{s}'} \prod_{l=1}^{i} \widehat{P}_{l,h}^k(s'^l|s^l,a^l) \prod_{m=i+1}^{K} P_{m,h}(s'^m|s^m,a^m)V(\mathbf{s}'), \quad i \in \{0,\ldots,K\},$$

其中空乘积 等于 1。注意，X_0 使用所有真实转移，而 X_K 使用所有经验转移。通过望远镜求和，

$$\left|\sum_{\mathbf{s}'} \widehat{\mathbf{P}}_h^k V(\mathbf{s}') - \sum_{\mathbf{s}'} \mathbf{P}_h V(\mathbf{s}')\right| = |X_K - X_0| \leq \sum_{i=1}^{K} |X_i - X_{i-1}|. \tag{26}$$

对于每个 $i \in [K]$，X_i 和 X_{i-1}，它们仅在智能体 i 的转移上有所不同。重新排列得，

$$X_i - X_{i-1} = \sum_{s'^i} \left(\widehat{P}_{i,h}^k(s'^i|s^i,a^i) - P_{i,h}(s'^i|s^i,a^i)\right) \cdot W_i(s'^i),$$

其中 $W_i(s'^i) := \sum_{\mathbf{s}':\ s'^i \text{ 固定}} \prod_{l\neq i} \mathbb{P}_{l,\ h}(s'^l|s^l,\ a^l)V(\mathbf{s}') \in [0,H]$ 和 $\mathbb{P}_{l,\ h}$ 表示 $\widehat{P}$ 对于 $l<i$，以及 P 对于 $l>i$。由于 $|W_i(s'^i)| \le H$（概率之和为 1），

$$|X_i - X_{i-1}| \le H \sum_{s'\in\mathcal{S}} |\widehat{P}^k_{i,h}(s'|s^i,a^i) - P_{i,h}(s'|s^i,a^i)|. \tag{27}$$

由事件 $\mathcal{E}_2$（引理 6），

$$|\widehat{P}^k_{i,h}(s'|s^i,a^i) - P_{i,h}(s'|s^i,a^i)| \le \sqrt{\frac{2P_{i,h}(s'|s^i,a^i)\iota}{N^k_{i,h}(s^i,a^i)}} + \frac{3\iota}{N^k_{i,h}(s^i,a^i)}.$$

对 $s'\in\mathcal{S}$ 求和并应用柯西-施瓦茨不等式，

$$\begin{aligned}\sum_{s'} |\widehat{P}^k_{i,h}(s'|s^i,a^i) - P_{i,h}(s'|s^i,a^i)| &\le \sqrt{\frac{2\iota}{N^k_{i,h}(s^i,a^i)}} \underbrace{\sum_{s'} \sqrt{P_{i,h}(s'|s^i,a^i)}}_{\le\sqrt{S}} + \frac{3S\iota}{N^k_{i,h}(s^i,a^i)} \\ &\le \sqrt{\frac{2S\iota}{N^k_{i,h}(s^i,a^i)}} + \frac{3S\iota}{N^k_{i,h}(s^i,a^i)}.\end{aligned}$$

与 (27) 结合可得 $|X_i - X_{i-1}| \le b(N^k_{i,h}(s^i,a^i))$。该结果由 (26) 得出。 □

接下来，我们证明以下引理。

引理 11. 在引理 *6* 中的事件 $\mathcal{E}$ 条件下，我们有：在整个算法*2*中，对任意$\mathbf{s}\in\mathcal{S}$、$\mathbf{a}\in\mathcal{A}^K$以及任意策略 π，成立

$$\bar{V}^{\pi,k}_h(\mathbf{s}) \ge V^\pi_h(\mathbf{s})$$

and

$$\bar{Q}^{\pi,k}_h(\mathbf{s},\mathbf{a}) \ge Q^\pi_h(\mathbf{s},\mathbf{a})$$

证明。首先，考虑在转移动态 $\widehat{\mathbf{P}}^k$ 下、奖励为 $r(\mathbf{s},\mathbf{a}) + \sum_{i=1}^K b(N^k_{i,h}(s^i,a^i))$ 的修正马尔可夫决策过程 $\widehat{M}$，其中 $b(\cdot)\ge 0$ 是奖金函数。根据定义，$\bar{V}^{\pi,k}$ 和 $\bar{Q}^{\pi,k}$ 是该修正 MDP 中策略 π 的价值函数。通过对 M应用贝尔曼方程，我们可以得到，对任意 $\mathbf{s}=(s^1,s^2,...,s^K)\in\mathcal{S}^K$ 以及 $\mathbf{a}=(a^1,a^2,...,a^K)\in\mathcal{A}^K$，

$$\begin{aligned}\bar{V}^{\pi,k}_h(\mathbf{s}) &= \sum_{\mathbf{a}\in\mathcal{A}^K} \pi(\mathbf{s},\mathbf{a})\bar{Q}^{\pi,k}_h(\mathbf{s},\mathbf{a}) \\ \bar{Q}^{\pi,k}_h(\mathbf{s},\mathbf{a}) &= r(\mathbf{s},\mathbf{a}) + \sum_{i=1}^K b(N^k_{i,h}(s^i,a^i)) + \sum_{\mathbf{s}'\in\mathcal{S}^K} \widehat{\mathbf{P}}^k_h(\mathbf{s}'|\mathbf{s},\mathbf{a})\bar{V}^{\pi,k}_{h+1}(\mathbf{s}')\end{aligned} \tag{28}$$

其中 $\widehat{\mathbf{P}}^k_h(\mathbf{s}'|\mathbf{s},\ \mathbf{a}) = \prod_{i=1}^K \widehat{P}^k_{i,\ h}(s'^i|s^i,\ a^i)$ 表示从 $(\mathbf{s},\ \mathbf{a})$ 到 $\mathbf{s}$ 的联合转移概率 $\mathbf{s}'$。

同理，对于真实价值函数：

$$\begin{aligned}V^\pi_h(\mathbf{s}) &= \sum_{\mathbf{a}\in\mathcal{A}^K} \pi(\mathbf{s},\mathbf{a})Q^\pi_h(\mathbf{s},\mathbf{a}) \\ Q^\pi_h(\mathbf{s},\mathbf{a}) &= r(\mathbf{s},\mathbf{a}) + \sum_{\mathbf{s}'\in\mathcal{S}^K} \mathbf{P}_h(\mathbf{s}'|\mathbf{s},\mathbf{a})V^\pi_{h+1}(\mathbf{s}')\end{aligned} \tag{29}$$

其中 $\mathbf{P}_h(\mathbf{s}'|\mathbf{s},\ \mathbf{a}) = \prod_{i=1}^K P_{i,\ h}(s'^i|s^i,\ a^i)$。

接下来，我们用归纳法证明该结论。首先，当 $h = H$ 时，我们有

$$\begin{aligned}\bar{Q}_H^{\pi,k}(\mathbf{s},\mathbf{a}) &= r(\mathbf{s},\mathbf{a}) + \sum_{i=1}^{K} b(N_{i,H}^k(s^i,a^i)) \\ &\geq r(\mathbf{s},\mathbf{a}) = Q_H^\pi(\mathbf{s},\mathbf{a}).\end{aligned}$$

由于 $\bar{V}^{\pi,k}_h(\mathbf{s}) = \sum_{\mathbf{a}\in\mathcal{A}^K} \pi(\mathbf{s},\mathbf{a})\bar{Q}^{\pi,k}_h(\mathbf{s},\mathbf{a})$ 且 $V_h^\pi(\mathbf{s}) = \sum_{\mathbf{a}\in\mathcal{A}^K} \pi(\mathbf{s},\mathbf{a})Q_h^\pi(\mathbf{s},\mathbf{a})$，因此结果对 $h = H$ 成立。

接下来，假设结果对 $h+1$ 成立。由式 (28) 中的贝尔曼方程可得

$$\begin{aligned}\bar{Q}_h^{\pi,k}(\mathbf{s},\mathbf{a}) &= r(\mathbf{s},\mathbf{a}) + \sum_{\mathbf{s}'\in\mathcal{S}^K} \widehat{\mathbf{P}}_h^k(\mathbf{s}'|\mathbf{s},\mathbf{a})\bar{V}_{h+1}^{\pi,k}(\mathbf{s}') + \sum_{i=1}^{K} b(N_{i,h}^k(s^i,a^i)) \\ &\geq r(\mathbf{s},\mathbf{a}) + \sum_{\mathbf{s}'\in\mathcal{S}^K} \widehat{\mathbf{P}}_h^k(\mathbf{s}'|\mathbf{s},\mathbf{a})V_{h+1}^{\pi}(\mathbf{s}') + \sum_{i=1}^{K} b(N_{i,h}^k(s^i,a^i)).\end{aligned}$$

由引理 10 可得

$$\begin{aligned}\bar{Q}_h^{\pi,k}(\mathbf{s},\mathbf{a}) &\geq r(\mathbf{s},\mathbf{a}) + \sum_{\mathbf{s}'\in\mathcal{S}^K} \mathbf{P}(\mathbf{s}'|\mathbf{s},\mathbf{a})V_{h+1}^{\pi}(\mathbf{s}') \\ &\geq Q_h^\pi(\mathbf{s},\mathbf{a})\end{aligned}$$

由于 $\bar{V}^{\pi,k}_h(\mathbf{s}) = \sum_{\mathbf{a}\in\mathcal{A}^K} \pi(\mathbf{s},\mathbf{a})\bar{Q}^{\pi,k}_h(\mathbf{s},\mathbf{a}) \geq \sum_{\mathbf{a}\in\mathcal{A}^K} \pi(\mathbf{s},\mathbf{a})Q_h^\pi(\mathbf{s},\mathbf{a}) = V_h^\pi(\mathbf{s})$，该结果对 h 的情形成立。因此，证明得证。 □

D.4 主定理证明

我们现在建立主要的遗憾界。该证明通过将遗憾与乐观价值估计和真实价值之间的差距联系起来，经由贝尔曼式递归分解该差距，并利用集中不等式界定所得各项来进行。

D.4.1 将遗憾与价值估计联系起来

根据定义，

$$\mathcal{R}_{T,\frac{1}{2}} = \sum_{k=1}^{T}\left[\frac{1}{2}V_1^{\pi^*}(\bar{\mathbf{s}}_1) - V_1^{\pi^k}(\bar{\mathbf{s}}_1)\right].$$

由引理 11，我们有 $\bar{V}^{\pi^*,k}_1(\bar{\mathbf{s}}_1) \geq V_1^{\pi^*}(\bar{\mathbf{s}}_1)$ 对所有 k 成立，从而得到

$$\mathcal{R}_{T,\frac{1}{2}} \leq \sum_{k=1}^{T}\left[\frac{1}{2}\bar{V}_1^{\pi^*,k}(\bar{\mathbf{s}}_1) - V_1^{\pi^k}(\bar{\mathbf{s}}_1)\right]. \tag{30}$$

由引理 8，我们有 $\bar{V}^{\pi^k,k}_1(\bar{\mathbf{s}}_1) \geq \bar{V}^{\pi^*,k}_1(\bar{\mathbf{s}}_1) - \epsilon$，将其整理为 $\frac{1}{2}\bar{V}^{\pi^*,k}_1(\bar{\mathbf{s}}_1) \leq \bar{V}^{\pi^k,k}_1(\bar{\mathbf{s}}_1) + \epsilon$。代入 (30)：

$$\begin{aligned}\mathcal{R}_{T,\frac{1}{2}} &\leq \sum_{k=1}^{T}\left[\bar{V}_1^{\pi^k,k}(\bar{\mathbf{s}}_1) - V_1^{\pi^k}(\bar{\mathbf{s}}_1) + \epsilon\right] \\ &= \sum_{k=1}^{T}\left[\bar{V}_1^{\pi^k,k}(\bar{\mathbf{s}}_1) - V_1^{\pi^k}(\bar{\mathbf{s}}_1)\right] + T\epsilon.\end{aligned} \tag{31}$$

证明的其余部分着重于对第一项进行界定。

D.4.2 ValueGap 的递归分解

对于回合 k 中的每一步 h，由于策略 π^k 是确定性的，并且在状态 $\mathbf{s}_h^k$ 下选择 $\mathbf{a}_h^k$，我们有

$$\bar{V}_h^{\pi^k,k}(\mathbf{s}_h^k) - V_h^{\pi^k}(\mathbf{s}_h^k) = \bar{Q}_h^{\pi^k,k}(\mathbf{s}_h^k, \mathbf{a}_h^k) - Q_h^{\pi^k}(\mathbf{s}_h^k, \mathbf{a}_h^k). \tag{32}$$

应用贝尔曼方程 (28) 和 (29)，可得

$$\begin{aligned}
\bar{Q}_h^{\pi^k,k}(\mathbf{s}, \mathbf{a}) &= r(\mathbf{s}, \mathbf{a}) + \sum_{i=1}^{K} b(N_{i,h}^k(s^i, a^i)) + \sum_{\mathbf{s}'} \widehat{\mathbf{P}}_h^k(\mathbf{s}'|\mathbf{s}, \mathbf{a}) \bar{V}_{h+1}^{\pi^k,k}(\mathbf{s}'), \\
Q_h^{\pi^k}(\mathbf{s}, \mathbf{a}) &= r(\mathbf{s}, \mathbf{a}) + \sum_{\mathbf{s}'} \mathbf{P}_h(\mathbf{s}'|\mathbf{s}, \mathbf{a}) V_{h+1}^{\pi^k}(\mathbf{s}'),
\end{aligned}$$

其中 $b(N) = H\sqrt{\frac{2S\iota}{N}} + \frac{3HS\iota}{N}$。计算 $(\mathbf{s}_h^k, \mathbf{a}_h^k)$ 处的差值，可得

$$\begin{aligned}
&\bar{Q}_h^{\pi^k,k}(\mathbf{s}_h^k, \mathbf{a}_h^k) - Q_h^{\pi^k}(\mathbf{s}_h^k, \mathbf{a}_h^k) \\
&= \sum_{\mathbf{s}'} \widehat{\mathbf{P}}_h^k(\mathbf{s}'|\mathbf{s}_h^k, \mathbf{a}_h^k) \bar{V}_{h+1}^{\pi^k,k}(\mathbf{s}') - \sum_{\mathbf{s}'} \mathbf{P}_h(\mathbf{s}'|\mathbf{s}_h^k, \mathbf{a}_h^k) V_{h+1}^{\pi^k}(\mathbf{s}') + \sum_{i=1}^{K} b(N_{i,h}^k(s_h^{i,k}, a_h^{i,k})) \\
&= \sum_{\mathbf{s}'} [\widehat{\mathbf{P}}_h^k - \mathbf{P}_h](\mathbf{s}'|\mathbf{s}_h^k, \mathbf{a}_h^k) \bar{V}_{h+1}^{\pi^k,k}(\mathbf{s}') + \sum_{\mathbf{s}'} \mathbf{P}_h(\mathbf{s}'|\mathbf{s}_h^k, \mathbf{a}_h^k) [\bar{V}_{h+1}^{\pi^k,k}(\mathbf{s}') - V_{h+1}^{\pi^k}(\mathbf{s}')] \\
&\quad + \sum_{i=1}^{K} b(N_{i,h}^k(s_h^{i,k}, a_h^{i,k})).
\end{aligned} \tag{33}$$

我们进一步通过加上和减去 $\sum_{\mathbf{s}'} [\widehat{\mathbf{P}}_h^k - \mathbf{P}_h](\mathbf{s}'|\mathbf{s}_h^k, \mathbf{a}_h^k) V_{h+1}^{\pi^k}(\mathbf{s}')$ 来细化第一项：

$$\begin{aligned}
&\sum_{\mathbf{s}'} [\widehat{\mathbf{P}}_h^k - \mathbf{P}_h](\mathbf{s}'|\mathbf{s}_h^k, \mathbf{a}_h^k) \bar{V}_{h+1}^{\pi^k,k}(\mathbf{s}') \\
&= \sum_{\mathbf{s}'} [\widehat{\mathbf{P}}_h^k - \mathbf{P}_h](\mathbf{s}'|\mathbf{s}_h^k, \mathbf{a}_h^k) V_{h+1}^{\pi^k}(\mathbf{s}') + \sum_{\mathbf{s}'} [\widehat{\mathbf{P}}_h^k - \mathbf{P}_h](\mathbf{s}'|\mathbf{s}_h^k, \mathbf{a}_h^k) [\bar{V}_{h+1}^{\pi^k,k}(\mathbf{s}') - V_{h+1}^{\pi^k}(\mathbf{s}')].
\end{aligned} \tag{34}$$

D.4.3 对分解项进行界定

由于对所有 $\mathbf{s}'$，$V_{h+1}^{\pi^k}(\mathbf{s}') \in [0, H]$，根据引理 10：

$$\left| \sum_{\mathbf{s}'} [\widehat{\mathbf{P}}_h^k - \mathbf{P}_h](\mathbf{s}'|\mathbf{s}_h^k, \mathbf{a}_h^k) V_{h+1}^{\pi^k}(\mathbf{s}') \right| \le \sum_{i=1}^{K} b(N_{i,h}^k(s_h^{i,k}, a_h^{i,k})). \tag{35}$$

根据引理 11，我们有 $\bar{V}^{\pi^k,\,k}_{h+1}(\mathbf{s}') - V_{h+1}^{\pi^k}(\mathbf{s}') \ge 0$ 对所有 $\mathbf{s}'$均成立。应用引理 9：

$$\begin{aligned}
&\sum_{\mathbf{s}'} [\widehat{\mathbf{P}}_h^k - \mathbf{P}_h](\mathbf{s}'|\mathbf{s}_h^k, \mathbf{a}_h^k) [\bar{V}_{h+1}^{\pi^k,k}(\mathbf{s}') - V_{h+1}^{\pi^k}(\mathbf{s}')] \\
&\quad \le \frac{1}{H} \sum_{\mathbf{s}'} \mathbf{P}_h(\mathbf{s}'|\mathbf{s}_h^k, \mathbf{a}_h^k) [\bar{V}_{h+1}^{\pi^k,k}(\mathbf{s}') - V_{h+1}^{\pi^k}(\mathbf{s}')] + \sum_{i=1}^{K} \frac{10SH^2K\iota}{N_{i,h}^k(s_h^{i,k}, a_h^{i,k})}.
\end{aligned} \tag{36}$$

结合式 (35)、(36)、(33) 和 (34)：

$$\begin{aligned}
\bar{V}_h^{\pi^k,k}(\mathbf{s}_h^k) - V_h^{\pi^k}(\mathbf{s}_h^k) &\le \left(1 + \frac{1}{H}\right) \sum_{\mathbf{s}'} \mathbf{P}_h(\mathbf{s}'|\mathbf{s}_h^k, \mathbf{a}_h^k) [\bar{V}_{h+1}^{\pi^k,k}(\mathbf{s}') - V_{h+1}^{\pi^k}(\mathbf{s}')] \\
&\quad + 2\sum_{i=1}^{K} b(N_{i,h}^k(s_h^{i,k}, a_h^{i,k})) + \sum_{i=1}^{K} \frac{10SH^2K\iota}{N_{i,h}^k(s_h^{i,k}, a_h^{i,k})}.
\end{aligned} \tag{37}$$

D.4.4 鞅结构的引入

将差分项 ξ_h^k 定义为

$$\xi_h^k := \sum_{\mathbf{s}'} \mathbf{P}_h(\mathbf{s}'|\mathbf{s}_h^k, \mathbf{a}_h^k)[\bar{V}_{h+1}^{\pi^k,k}(\mathbf{s}') - V_{h+1}^{\pi^k}(\mathbf{s}')] - [\bar{V}_{h+1}^{\pi^k,k}(\bar{\mathbf{s}}_{h+1}^k) - V_{h+1}^{\pi^k}(\mathbf{s}_{h+1}^k)],$$

以及确定性误差项:

$$\eta_h^k := 2\sum_{i=1}^{K} b(N_{i,h}^k(s_h^{i,k}, a_h^{i,k})) + \sum_{i=1}^{K} \frac{10SH^2K\iota}{N_{i,h}^k(s_h^{i,k}, a_h^{i,k})}.$$

根据这些定义，(37) 变为:

$$\bar{V}_h^{\pi^k,k}(\mathbf{s}_h^k) - V_h^{\pi^k}(\mathbf{s}_h^k) \le \left(1+\frac{1}{H}\right)[\bar{V}_{h+1}^{\pi^k,k}(\bar{\mathbf{s}}_{h+1}^k) - V_{h+1}^{\pi^k}(\mathbf{s}_{h+1}^k)] + \left(1+\frac{1}{H}\right)\xi_h^k + \eta_h^k. \tag{38}$$

D.4.5 递归展开

从 $h=1$ 到 $h=H$ 递归应用 (38):

$$\begin{aligned}
&\bar{V}_1^{\pi^k,k}(\bar{\mathbf{s}}_1^k) - V_1^{\pi^k}(\bar{\mathbf{s}}_1^k) \\
&\le \left(1+\frac{1}{H}\right)^H [\bar{V}_{H+1}^{\pi^k,k}(\bar{\mathbf{s}}_{H+1}^k) - V_{H+1}^{\pi^k}(\mathbf{s}_{H+1}^k)] + \sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{H-h+1}\xi_h^k + \sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{H-h}\eta_h^k.
\end{aligned} \tag{39}$$

由于 $\bar{V}_{H+1}^{\pi^k,k}(\bar{\mathbf{s}}_{H+1}^k) = V_{H+1}^{\pi^k}(\mathbf{s}_{H+1}^k) = 0$ 以及 $(1+\frac{1}{H})^H \le e$:

$$\bar{V}_1^{\pi^k,k}(\bar{\mathbf{s}}_1^k) - V_1^{\pi^k}(\bar{\mathbf{s}}_1^k) \le e\sum_{h=1}^{H}\xi_h^k + e\sum_{h=1}^{H}\eta_h^k. \tag{40}$$

对所有 T 回合求和，可得

$$\begin{aligned}
\mathcal{R}_{T,\frac{1}{2}} &\le \sum_{k=1}^{T}[\bar{V}_1^{\pi^k,k}(\bar{\mathbf{s}}_1) - V_1^{\pi^k}(\bar{\mathbf{s}}_1)] + T\epsilon \\
&\le e\sum_{k=1}^{T}\sum_{h=1}^{H}\xi_h^k + e\sum_{k=1}^{T}\sum_{h=1}^{H}\eta_h^k + T\epsilon.
\end{aligned} \tag{41}$$

D.4.6 确定性探索成本的界定

利用 $b(N) = H\sqrt{\frac{2S\iota}{N} + \frac{3HS\iota}{N}}$ 展开 η_h^k 的定义:

$$\sum_{k=1}^{T}\sum_{h=1}^{H}\eta_h^k \le \sum_{k=1}^{T}\sum_{h=1}^{H}\left(\sum_{i=1}^{K}\frac{16SH^2K\iota}{N_{i,h}^k(s_h^{i,k}, a_h^{i,k})} + 2H\sum_{i=1}^{K}\sqrt{\frac{2S\iota}{N_{i,h}^k(s_h^{i,k}, a_h^{i,k})}}\right),$$

其中我们用到了 $10SH^2K\iota + 2\cdot 3HS\iota \le 16SH^2K\iota$。

对于第一项，重新排列求和顺序并按状态-动作对分组:

$$\begin{aligned}
&\sum_{k=1}^{T}\sum_{h=1}^{H}\sum_{i=1}^{K}\frac{16SH^2K\iota}{N_{i,h}^k(s_h^{i,k}, a_h^{i,k})} = 16SH^2K\iota\sum_{i=1}^{K}\sum_{h=1}^{H}\sum_{(s,a)}\sum_{\ell=1}^{N_{i,h}^T(s,a)}\frac{1}{\ell} \\
&\le 16SH^2K\iota\sum_{i=1}^{K}\sum_{h=1}^{H}\sum_{(s,a)}(1+\log N_{i,h}^T(s,a)) \le 32S^2AH^3K^2\iota\log T,
\end{aligned}$$

其中我们使用了 $\sum_{\ell=1}^{n} \frac{1}{\ell} \leq 1 + \log n$ and $N_{i,h}^{T}(s, a) \leq T$。

对于第二项，使用 $\sum_{\ell=1}^{n} \frac{1}{\sqrt{\ell}} \leq 1 + 2\sqrt{n}$：

$$
\begin{aligned}
&2H \sum_{k=1}^{T} \sum_{h=1}^{H} \sum_{i=1}^{K} \sqrt{\frac{2S\iota}{N_{i,h}^{k}(s_h^{i,k}, a_h^{i,k})}} = 2H\sqrt{2S\iota} \sum_{i=1}^{K} \sum_{h=1}^{H} \sum_{(s,a)} \sum_{\ell=1}^{N_{i,h}^{T}(s,a)} \frac{1}{\sqrt{\ell}} \\
&\leq 2H\sqrt{2S\iota} \sum_{i=1}^{K} \sum_{h=1}^{H} \sum_{(s,a)} (1 + 2\sqrt{N_{i,h}^{T}(s,a)}) \\
&\leq 2\sqrt{2S\iota} H^2 KSA + 2H\sqrt{2S\iota} \sum_{i=1}^{K} \sum_{h=1}^{H} \sum_{(s,a)} 2\sqrt{N_{i,h}^{T}(s,a)} \\
&\leq 2\sqrt{2S\iota} H^2 KSA + 4H^2 KS\sqrt{2AT\iota}.
\end{aligned}
$$

合并可得：

$$
\sum_{k=1}^{T} \sum_{h=1}^{H} \eta_h^k \leq 32S^2AH^3K^2\iota \log T + 4H^2KS\sqrt{2AT\iota} + 2\sqrt{2S\iota}H^2KSA. \tag{42}
$$

D.4.7 对鞅项进行界定

令 $\mathcal{F}_h^k$ 表示由回合 k 的第 h 步及之前的所有随机性生成的 σ-代数。对于任意 $\mathbf{s}$，价值差 $\bar{V}^{\pi^k, k}_{h+1}(\mathbf{s}) - V^{\pi^k}_{h+1}(\mathbf{s})$ 是 $\mathcal{F}_h^k$-可测的（因为它只依赖于截至回合 $k-1$ 的数据），而 $\mathbf{s}_{h+1}^k$ 是 $\mathcal{F}_{h+1}^k$-可测的，但不是 $\mathcal{F}_h^k$-可测的。

计算条件期望：

$$
\mathbb{E}[\xi_h^k | \mathcal{F}_h^k] = \sum_{\mathbf{s}'} \mathbf{P}_h(\mathbf{s}' | \mathbf{s}_h^k, \mathbf{a}_h^k)[\bar{V}_{h+1}^{\pi^k,k}(\mathbf{s}') - V_{h+1}^{\pi^k}(\mathbf{s}')] - \mathbb{E}\left[\bar{V}_{h+1}^{\pi^k,k}(\bar{\mathbf{s}}_{h+1}^k) - V_{h+1}^{\pi^k}(\mathbf{s}_{h+1}^k) \middle| \mathcal{F}_h^k\right].
$$

由于在给定 $\mathcal{F}_h^k$ 的条件下，$\mathbf{s}_{h+1}^k \sim \mathbf{P}_h(\cdot | \mathbf{s}_h^k, \mathbf{a}_h^k)$：

$$
\mathbb{E}\left[\bar{V}_{h+1}^{\pi^k,k}(\bar{\mathbf{s}}_{h+1}^k) - V_{h+1}^{\pi^k}(\mathbf{s}_{h+1}^k) \middle| \mathcal{F}_h^k\right] = \sum_{\mathbf{s}'} \mathbf{P}_h(\mathbf{s}' | \mathbf{s}_h^k, \mathbf{a}_h^k)[\bar{V}_{h+1}^{\pi^k,k}(\mathbf{s}') - V_{h+1}^{\pi^k}(\mathbf{s}')].
$$

因此：

$$
\mathbb{E}[\xi_h^k | \mathcal{F}_h^k] = 0, \tag{43}
$$

这证实了 $\{\xi_h^k\}$ 是一个鞅差序列。

由于 $0 \leq \bar{V}^{\pi^k, k}_{h+1}(\mathbf{s}') - V^{\pi^k}_{h+1}(\mathbf{s}') \leq H$：

$$
|\xi_h^k| \leq \sum_{\mathbf{s}'} \mathbf{P}_h(\mathbf{s}' | \mathbf{s}_h^k, \mathbf{a}_h^k) |\bar{V}_{h+1}^{\pi^k,k}(\mathbf{s}') - V_{h+1}^{\pi^k}(\mathbf{s}')| + |\bar{V}_{h+1}^{\pi^k,k}(\bar{\mathbf{s}}_{h+1}^k) - V_{h+1}^{\pi^k}(\mathbf{s}_{h+1}^k)| \leq 4H.
$$

应用 Azuma-Hoeffding 不等式，其中 TH 个鞅差均有界于 $4H$，其概率至少为 $1 - \frac{\delta}{3}$：

$$
\sum_{k=1}^{T} \sum_{h=1}^{H} \xi_h^k \leq \sqrt{2TH \cdot (4H)^2 \cdot \log(3/\delta)} = 4H\sqrt{2TH \log(3/\delta)}.
$$

由于 $\iota = \log(10S^2AT/\delta) \geq \log(3/\delta)$：

$$
\sum_{k=1}^{T} \sum_{h=1}^{H} \xi_h^k \leq 4H\sqrt{2TH\iota}. \tag{44}
$$

D.4.8 最终汇总

结合 (41)、(42) 和 (44)，并对引理 6 中的事件 $\mathcal{E}$ 以及用于界定 $\sum_{k=1}^{T}\sum_{h=1}^{H}\xi_h^k$ 的 Azuma 不等式应用并集界，我们得到该结论以至少 $1-\delta$ 的概率成立：

$$\begin{aligned}\mathcal{R}_{T,\frac{1}{2}} &\le 4eH\sqrt{2TH\iota} + e(32S^2AH^3K^2\iota\log T + 4H^2KS\sqrt{2AT\iota} + 2\sqrt{2S\iota}H^2KSA) + T\epsilon \\ &\le 4eH\sqrt{2TH\iota} + 96S^2AH^3K^2\iota\log T + 4eH^2KS\sqrt{2AT\iota} + T\epsilon.\end{aligned}$$

取 $\epsilon = O(1/T)$，并将常数吸收到 O 记号中：

$$\mathcal{R}_{T,\frac{1}{2}} = O\left(S^2AH^3K^2\iota\log T + H^2KS\sqrt{AT\iota}\right).$$

由于 $\iota := \log(6S^2ATHK/\delta)$，我们可以得到

$$\mathcal{R}_{T,\frac{1}{2}} = O\left(S^2AH^3K^2\log(SATHK/\delta)\log T + H^2KS\sqrt{AT\log(SATHK/\delta)}\right).$$

这就得到了所需的 $O(\sqrt{T})$ 遗憾率，证明完毕。

E 技术引理

引理 12 (Bernstein 不等式). 设 $x_1, ..., x_n$ 是相互独立的有界随机变量，满足 $\mathbb{E}[x_i] = 0$ 和 $|x_i| \le \xi$，其概率为 1。设 $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}[x_i]$，则以概率 $1-\delta$ 有

$$\frac{1}{n}\sum_{i=1}^{n} x_i \le \sqrt{\frac{2\sigma^2\cdot\log(1/\delta)}{n}} + \frac{2\xi}{3n}\log(1/\delta)$$

引理 13 （霍夫丁不等式）。设 $X_1, ..., X_N$ 是相互独立的随机变量，满足 X_i 是 R-次高斯的，且对所有 i 均有 $\mathbb{E}[X_i] = \mu$。设 $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$。那么对于任意 $t > 0$,

$$P(|\overline{X} - \mu| \ge t) \le 2\exp\{-\frac{Nt^2}{2R^2}\}.$$