arXiv:2302.06606v2 [cs.LG] 2 Mar 2023

打破多主体诅咒：基于函数逼近、可证明高效的去中心化多智能体强化学习

Yuanhao Wang[†*] Qinghua Liu[†*] Yu Bai[‡§] Chi Jin[†§]

2023年3月3日

摘要

多智能体强化学习（MARL）面临的一项独特挑战是多主体诅咒：博弈的描述长度以及许多现有学习算法的复杂度都会随智能体数量呈指数级增长。尽管近期研究在表格型马尔可夫博弈模型下成功应对了这一挑战，但其机制严重依赖于状态数量有限且较少这一条件，无法推广到状态空间巨大、必须使用函数逼近来近似价值函数或策略的实际场景。

本文提出了首批能够在函数逼近条件下以可证明的方式破解多主体诅咒的多智能体强化学习算法。我们设计了一种新的去中心化算法——带策略回放的 V 学习，它首次为在去中心化线性函数逼近条件下学习马尔可夫博弈的近似粗相关均衡提供了多项式样本复杂度保证。该算法始终输出马尔可夫粗相关均衡，并以最优速率 $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ 找到 ε-最优解。此外，在表格情形下，我们的结果在寻找马尔可夫粗相关均衡方面优于当前最佳去中心化结果 $\widetilde{\mathcal{O}}(\varepsilon^{-3})$。我们还提出了一种替代算法——去中心化乐观策略镜像下降，它只需多项式数量的样本即可找到策略类受限的粗相关均衡。作为学习较弱版本粗相关均衡的交换，该算法可应用于更广泛的、采用通用函数逼近的问题，例如线性二次博弈，以及具有较低“边际”Eluder 维度的多智能体强化学习问题。

1 引言

多智能体强化学习（MARL）研究这样一类问题：智能体通过与未知环境以及其他可能采取策略并自适应的智能体交互，学习最大化自身效用。现代多智能体强化学习系统已在广泛的挑战性任务中取得显著成功，包括围棋（Silver 等，2016年、2017年）、扑克（Brown 和 Sandholm，2018年、2019年）、策略博弈（Vinyals 等，2019年；OpenAI，2018年；Bakhtin 等，2022年；Wurman 等，2022年）、去中心化控制（Brambilla 等，2013年）、自动驾驶（Shalev-Shwartz 等，2016年），以及捉迷藏等复杂社会场景（Baker 等，2020年）。与理论研究丰富的单智能体强化学习相比，多智能体强化学习带来了一系列新的博弈论挑战，其中许多仍有待解决。

多智能体强化学习中的一个独特挑战是多主体诅咒：博弈的描述长度（尤其是联合动作空间的大小）会随着智能体数量呈指数级增长。因此，任何试图对整个博弈进行建模（例如建模转移概率或联合Q值）的学习算法，都会面临呈指数级增长的样本复杂度或计算复杂度（Bai 等人，2020年；Liu 等人，2021年）。即使对于规模相当小的多智能体应用，这些算法在实践中也难以运行。为应对这一挑战，实践者提倡设计去中心化算法（综述请参见，例如Zhang 等人（2021a）），让智能体只从各自的局部视角学习博弈中相关的部分，例如个体策略、V值或边际Q值

[†]普林斯顿大学。电子邮箱：{yuanhao,qinghual,chij}@princeton.edu
[‡]Salesforce Research。电子邮箱：yu.bai@salesforce.com
* 和 § 表示贡献相同。

（参见第2节中的定义）。去中心化算法还使每个智能体能够几乎独立地学习，智能体之间只需极少通信，甚至无需通信，从而为算法的实现带来灵活性和优势。

多智能体强化学习中的多主体诅咒，近期已通过一系列理论研究得到可证明地解决（Song 等人，2021年；Jin 等人，2021b；Mao 和 Bas¸ar，2022年），这些研究采用了 V 学习算法（Bai 等人，2020年）。然而，这些结果仅适用于基础的表格型马尔可夫博弈设置（Shapley，1953年），其中状态和动作的数量有限且较少。此外，这些方法的机制严重依赖表格设置所带来的协同作用，即：（1）逐状态无悔算法；（2）增量价值更新；以及（3）乐观性。这使得它们无法直接扩展到具有大规模状态空间的实际场景。这与现代多智能体强化学习的实践形成对比：现代实践通常涉及状态数量极其庞大的问题，因此必须使用函数逼近（通常采用深度神经网络）来逼近价值函数或策略（Sutton 和 Barto，2018年）。由此自然引出了以下开放问题：

> 即使在函数逼近下，我们能否设计出打破多主体诅咒的去中心化多智能体强化学习算法?

本文通过设计能够在一般函数逼近条件下求得近似粗相关均衡（CCEs）的算法，对上述问题给出了肯定的回答；这些算法的样本复杂度关于问题参数（包括智能体数量）为多项式级。具体而言，

- 我们设计了一种新的去中心化多智能体强化学习元算法——带策略回放的 V 学习 (VLPR)，以及它的加速版本 AVLPR (第 3.1 节)。这两种算法将标准的 V 学习算法（Bai 等人，2020年；Jin 等人，2021b）与新的策略回放机制相结合，以输出马尔可夫粗相关均衡，并促进函数逼近条件下的学习。VLPR 是完全去中心化的（假设参与者之间具有共享随机性），而 AVLPR需要最小通信量（见第 3.2 节）。在具备高效子程序的情况下，两者都能在多项式时间内运行。

- 我们的元算法VLPR 和 AVLPR 需要调用抽象子程序来：(1) 估计每个智能体的 V 值（而非联合 Q 值）；(2) 通过无悔算法计算逐阶段粗相关均衡策略。我们证明，在对子程序施加温和条件的情况下，这两种元算法都能以多项式数量的样本高效地求得近似粗相关均衡（第 3 节）。这些温和条件在线性与表格型设置下均成立（第 4 节）。

- 我们在AVLPR 中采用去中心化线性函数逼近设定，从而得到首个能够在该设定下从理论上打破多主体诅咒的去中心化多智能体强化学习算法（第4.1节）。我们的算法以最优速率 $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ 寻找 ε-最优解。对于表格型马尔可夫博弈，目前寻找马尔可夫粗相关均衡的最佳去中心化算法需要 $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ 个样本（Daskalakis 等人，2022年）。AVLPR在 $1/\varepsilon$（状态数量）以及时域的依赖关系方面改进了这一结果（第4.2节）。

- 我们提供了另一种算法——去中心化乐观策略镜像下降（DOPMD），该算法能够找到策略类受限的粗相关均衡——这是一种弱于标准定义的粗相关均衡概念——且其样本复杂度打破了多主体诅咒（第5节）。作为采用较弱粗相关均衡概念的交换，DOPMD 适用于更广泛的一般函数逼近问题，并且这些问题具有有界的 Bellman-Eluder维数。这些问题包括线性二次博弈，以及具有较低“边际”Eluder 维度或贝尔曼秩的博弈。

1.1 相关工作

本节回顾了基于马尔可夫博弈模型的多智能体强化学习理论研究（Shapley，1953年；Littman，1994年）。我们也注意到，近年来关于经验性多智能体强化学习或基于其他数学模型的研究成果十分丰富，但这些内容超出了本文的讨论范围。

集中式多智能体强化学习 近年来，大量研究广泛探讨了马尔可夫博弈的样本高效学习（Brafman 和 Tennenholtz，2002年；Wei 等人，2017年；Jia 等人，2019年；Sidford 等人，2020年；Bai 和 Jin，2020年；Xie 等人，2020年；Bai 等人，2020年；Zhang 等人，2020年；Tian 等人，2021年；Liu 等人，2021年；Bai 等人，2021年；Huang 等人，2021年；Jin 等人，2022年；Chen 等人，2022b年）。这些方法大多本质上是集中式的，因为它们会估计一些量（例如转移模型或联合 Q 函数），而这些量的参数数量会随着参与者数量呈指数级增长，因此其样本复杂度会受到多主体诅咒的影响。

去中心化多智能体强化学习 打破马尔可夫博弈中多主体诅咒的去中心化方法，始于 V 学习算法。该算法最初由 Bai 等（2020年）在零和设定中提出，随后被扩展到一般和设定（Song 等人，2021年；Jin 等，2021b年；Mao 和 Bas¸ar，2022年；Mao 等人，2022年；Cui 和 Du，2022年；Zhang 等人，2022年）。在这一设定下，该算法能够学习博弈的近似相关均衡（CE）或粗相关均衡（CCE），其样本复杂度相对于智能体数量呈多项式增长。之后，Daskalakis 等（2022年）提出的 SPoCMAR 算法进一步学习近似粗相关均衡（CCEs），并保证其为马尔可夫型，但样本复杂度的多项式增长幅度略高。这两种算法都只适用于表格型马尔可夫博弈，无法处理函数逼近。我们的算法 VLPR 和 AVLPR 可视为将 V 学习算法扩展到函数逼近设置，并且还能输出马尔可夫策略。此外，我们的算法在表格设置下的特例，在学习马尔可夫粗相关均衡方面，比 Daskalakis 等（2022年）具有更优的样本复杂度。

用于学习相关均衡/粗相关均衡的去中心化算法，在标准式博弈（NFGs）（Stoltz，2005；Cesa-Bianchi 和 Lugosi，2006）和扩展式博弈（EFGs）（Kozuno 等人，2021；Bai 等人，2022b,a；Song 等人，2022；Fiegel 等人，2022）等其他博弈中也已成熟：这些算法让每个智能体运行一种即使面对对抗性对手也有效的无悔算法。然而，由于在马尔可夫博弈中针对对抗性对手进行学习存在根本性难题，这一成功无法推广至马尔可夫博弈：最坏情况下，遗憾下界随时域呈指数增长（Liu 等人，2022）。最后，马尔可夫势博弈（Zhang 等人，2021b；Leonardos 等人，2021；Song 等，2021；Ding 等人，2022）这一马尔可夫博弈的子类中也已建立去中心化算法，但这类算法严重依赖其特殊的势结构。

函数逼近下的多智能体强化学习 近期有一些研究通过改编单智能体设置中的技术（Jiang 等人，2017年；Jin 等，2020；Zhou 等人，2021年；Du 等人，2021年；Jin 等人，2021a；Foster 等人，2021年），探讨了线性（Xie 等人，2020年；Chen 等人，2022b年）和一般（Jin 等人，2022年；Huang 等人，2021年；Zhan 等（2022）；Xiong 等人，2022年；Chen 等人，2022年a；Ni 等人，2022年）函数逼近下的马尔可夫博弈学习。这些研究都要求使用集中式函数类，在专门应用于表格设置时会遭遇多主体诅咒。我们的 DOPMD 算法不同于 Liu 等（2022）和 Zhan 等（2022）的相关算法：我们新的内循环支持去中心化函数类，因此可应用于更广泛的场景。

从技术上讲，我们算法中使用的策略重放机制（尤其是通过AVLPR 中的倍增技巧实现的机制）与Zanette 和 Wainwright (2022年) 的方法类似；他们使用该机制，在单智能体设置中为线性函数逼近设计一种 Q 学习风格的算法。然而，除此之外，我们的方法截然不同，尤其是在值更新方式上：他们采用 Q 学习风格的增量更新，而我们的算法采用带批量更新的分阶段学习（其思想与值迭代相似）。

与独立工作的比较（Cui 等人，2023年）在本工作进行期间，Cui 等人（2023年）也研究了在线性函数逼近下的马尔可夫博弈中打破多主体诅咒的问题，并且在基本表格设置中找到了马尔可夫粗相关均衡所需的样本复杂度也有所改进，达到了相同水平。除了算法设计方面显而易见的差异外，我们在此强调两项工作在线性设置中的几个关键区别：（1）在假设方面，两项工作都假设相对于特定策略类满足贝尔曼完备性（例如，见假设 4）。我们指出，限制策略类的表达能力至关重要，否则博弈就会“本质上是表格型的”（见附录 D）。我们只要求相对于线性 argmax 策略满足完备性，而 Cui 等人（2023年）则要求相对于一个策略类 Π^{estimate} 满足完备性；该策略类由其算法和所使用的无悔学习预言机隐式定义，通常包含比线性 argmax 策略更复杂的策略。[2] （2）在样本复杂度方面，本文达到 $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ 速率，其对误差 ε具有最优的统计依赖关系，而 Cui 等人（2023年）达到的速率为 $\widetilde{\mathcal{O}}(\varepsilon^{-4})$。值得指出的是，Cui 等人（2023年）对动作数量 A的依赖关系更优，而我们的结果则对

[1]相比之下，V 学习所学习的策略通常是非马尔可夫、依赖历史的策略。[2]我们指出，由于两篇论文在算法设计和底层机制方面存在若干关键差异，关于其 Π^{estimate} 的这一说法成立，与其算法中采用何种全信息无悔学习预言机无关：无论是指数权重算法（Cui 等人 (2023年) 所采用的选择），还是期望扰动跟随领袖算法（Hazan 和 Minasyan，2020年）（本文所采用的选择）。更多细节请参见附录D.1 。

维度 d和时域 H的依赖关系更优。 A、 d、 H方面依赖关系的差异，源于算法技术和假设两方面的不同（极小极大最优速率可能因此有所不同）。

除了上述差异之外，Cui 等人(2023年) 还进一步给出了在一定程度的模型设定错误下进行学习、学习近似相关均衡（CE），以及学习线性马尔可夫势博弈的结果，这些问题均未在本文中考虑。另一方面，本文还给出了线性函数逼近设定之外的结果：VLPR 和AVLPR 都是通用元算法，只要子程序满足所需条件，就能为任意函数类提供保证。我们还针对一般函数逼近设计了一种新算法，能够在较弱条件下学习策略类受限的粗相关均衡（第5节）。

2 预备知识

马尔可夫博弈 我们考虑包含 m 个参与者的一般和分幕式马尔可夫博弈，可表示为元组 $\mathrm{MG}(H, \mathcal{S}, \{\mathcal{A}_i\}_{i\in[m]}, \mathbb{P}, \{r_i\}_{i\in[m]})$。这里， H 是时域长度， $\mathcal{S}$ 是状态空间， $\mathcal{A}_i$ 是第 i 个参与者的动作空间，其中 $|\mathcal{A}_i| = A_i$[3]；我们用 $\mathbf{a} = (a_1, \ldots, a_m) \in \prod_{i=1}^m \mathcal{A}_i =: \mathcal{A}$ 表示所有参与者的联合动作， $\mathbb{P} = (\mathbb{P}_h)_{h\in[H]}$ 是转移概率，其中 $\mathbb{P}_h(\cdot|s, \mathbf{a}) \in \Delta(\mathcal{S})$ 表示在第 h 步当前状态-动作 $(s, \mathbf{a})$ 下下一状态的概率分布； $r_i = (r_{i,h})_{h\in[H]}$ 是参与者 i 的奖励函数，其中每个 $r_{i,h}\colon \mathcal{S}\times\mathcal{A} \to [0,1]$ 都是一个函数，它将任意状态-动作 $(s, \mathbf{a})$ 映射为确定性[4] 奖励。在每个回合中，博弈从固定的初始状态 s_1 开始。在第 $h \in [H]$ 步、状态为 s_h 时，每个参与者采取自己的动作 $a_{i,h} \in \mathcal{A}_i$，获得自己的奖励 $r_{i,h} = r_{i,h}(s_h, \mathbf{a}_h)$，其中 $\mathbf{a}_h = (a_{1,h}, \ldots, a_{m,h})$，随后博弈按照马尔可夫方式转移到下一状态$s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, \mathbf{a}_h)$。

第 i 位玩家的马尔可夫策略记作 $\pi_i = \{\pi_{i,h}(\cdot|s) \in \Delta(\mathcal{A}_i)\}_{(s,h)\in\mathcal{S}\times[H]}$，它规定了在任意 (s, h) 下，第 i 位玩家的动作分布 $\pi_{i,h}(\cdot|s) \in \Delta(\mathcal{A}_i)$。这里，我们用 $\Delta(\mathcal{A}_i)$ 表示动作集 $\mathcal{A}_i$ 上的概率单纯形。马尔可夫联合策略 $\pi = \{\pi_h(\cdot|s) \in \Delta(\mathcal{A})\}_{(s,h)\in\mathcal{S}\times[H]}$ 是所有玩家的联合策略，它规定了联合动作上的分布；一般而言，不同玩家的随机性可以相关。马尔可夫联合策略的一种特殊情形是乘积策略 $\pi = \{\pi_i\}_{i\in[m]}$，其中每个智能体独立地执行 π_i。对于任意联合策略 π，我们分别将其在任意 $(i, h) \in [m]\times[H]$ 下的 V 值函数和（联合）Q 值函数定义为 $V_{i,h}^{\pi}(s) := \mathbb{E}_{\pi}[\sum_{h'=h}^{H} r_{i,h'}(s_{h'}, \mathbf{a}_{h'}) \mid s_h = s]$ 和 $Q_{i,h}^{\pi}(s, \mathbf{a}) := \mathbb{E}_{\pi}[\sum_{h'=h}^{H} r_{i,h'}(s_{h'}, \mathbf{a}_{h'}) \mid (s_h, \mathbf{a}_h) = (s, \mathbf{a})]$。此外，稍微扩展一下符号的用法，我们将玩家 $i \in [m]$ 的边缘 Q 函数以及任意 $h \in [H]$ 定义为

$$Q_{i,h}^{\pi}(s, a_i) := \mathbb{E}_{\pi}\left[\sum_{h'=h}^{H} r_{i,h'}(s_{h'}, \mathbf{a}_{h'}) \mid (s_h, a_{i,h}) = (s, a_i)\right],$$

它衡量玩家 i 在给定状态和自身动作时的 Q 值，并根据策略 π对对手的动作进行边缘化。为简化记号，我们定义算子 $\mathbb{P}_h$ 和 $\mathbb{D}_{\pi}$ 为 $[\mathbb{P}_h V](s, \mathbf{a}) := \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,\ \mathbf{a})}[V(s')]$ 和 $\mathbb{D}_{\pi}[Q](s) := \mathbb{E}_{\mathbf{a}\sim\pi(\cdot|s)}[Q(s, \mathbf{a})]$。我们还用 π_{-i} 表示由 π指定的除第 i位玩家之外所有玩家的联合策略。对于任意马尔可夫乘积策略 $\pi = \pi_i \times \pi_{-i}$，玩家 i 在第 h 步的贝尔曼算子 $\mathcal{T}_{i,h}^{\pi}$，h 是定义在第 i位玩家的边缘 Q 函数空间 $(\mathcal{S}\times\mathcal{A}_i \to \mathbb{R})$ 上的自映射，定义如下

$$[\mathcal{T}_{i,h}^{\pi} f](s, a_i) := \mathbb{E}_{\mathbf{a}_{-i}\sim\pi_{-i,h}(\cdot|s), s'\sim\mathbb{P}_h(\cdot|s,\mathbf{a}), a_i'\sim\pi_{i,h+1}(s')}[r_{i,h}(s, \mathbf{a}) + f(s', a_i')].$$

粗相关均衡 我们的目标是找到马尔可夫博弈的近似均衡，即一个联合策略，使每个玩家的自身策略在某种意义上对其对手而言近乎最优。在我们的多玩家一般和设定中，标准的纳什均衡概念不仅在计算上是 PPAD 困难（Daskalakis，2013年），在统计上也难以处理，需要 $\exp(\Omega(m))$ 个样本（Rubinstein，2017年）。我们专注于学习粗相关均衡（CCE），这是一般和马尔可夫博弈中常见的一种放宽的均衡概念（Liu 等人，2021年），它不存在上述困难，并且在基础表格设置中确实可以通过多项式时间和多项式数量的样本来学习（Song 等人，2021年；Jin 等人，2021b；Mao and Başar，2022年）。

[3]第 3、5 节中的结果不要求 A_i 为有限集合。 [4]我们的结果可以直接推广到随机奖励的情形。

算法1 带策略回放的 V 学习（VLPR）

1：将 π^1 初始化为均匀策略：$\pi_{i,\,h}^1(\cdot|s)$ ← 均匀分布$(\mathcal{A}_i)$，对所有 (i, s, h) 均成立。2：对于迭代 $t = 1, \ldots, T$，执行 3：设置回放策略 $\overline{\pi}^t$ ← 均匀分布$(\{\pi^\tau\}_{\tau\in[t]})$ 和 $\overline{V}_{i,\,H+1}^{t+1} \leftarrow 0$。4：对于 $h = H, \ldots, 1$，执行 5：计算 $\pi_h^{t+1} \leftarrow$ CCE-APPROX$_h(\overline{\pi}^t, \{\overline{V}_{i,h+1}^{t+1}\}_{i\in[m]}, t)$。6：计算 $\{\overline{V}_{i,h}^{t+1}\}_{i\in[m]} \leftarrow$ V-APPROX$_h(\overline{\pi}^t, \pi_h^{t+1}, \{\overline{V}_{i,h+1}^{t+1}\}_{i\in[m]}, t)$。输出：从 $\{\pi^t\}_{t\in[T]}$ 中均匀随机抽取的 π^{out} 样本。

对于任意 $\varepsilon > 0$，如果联合策略 π 是该博弈的 ε-近似相关均衡，则称其为

$$\text{CCEGap}(\pi) := \max_{i\in[m]}(\max_{\pi_i^\dagger} V_{1,i}^{\pi_i^\dagger,\pi_{-i}}(s_1) - V_{1,i}^{\pi}(s_1)) \le \varepsilon,$$

这里，最大化者 $\pi_i^\dagger$ 也称为最佳响应。我们记作 $V^{\dagger,\,i\,h}{}^{\pi_{-i}}{}_{,} := \text{最大值}_{\pi_i^\dagger} V^{\pi_i^\dagger}{}_{h,\,i}{}^{\pi_{-i}}{}_{,}$。

我们考虑基于老虎机反馈进行 PAC 学习的标准设定：智能体在多个回合中反复与底层马尔可夫博弈交互，并在每个回合中观察轨迹 $(s_1, \mathbf{a}_1, \mathbf{r}_1, \ldots, s_H, \mathbf{a}_H, \mathbf{r}_H)$（其中 $\mathbf{r}_h := (r_{i,\,h})_{i\in[m]}$）。目标是用尽可能少的博弈回合找到该博弈的一个 ε-近似相关均衡 $\widehat{\pi}$。

2.1 基于函数逼近的去中心化多智能体强化学习

为了支持大状态空间下的去中心化多智能体强化学习，本文考虑函数逼近设定，其中每位玩家$i \in [m]$ 都有自己的边际 Q 值函数类 $\mathcal{F}_i$。具体而言，我们令每位玩家 $i \in [m]$ 配备一个有限的[5]函数类 $\mathcal{F}_i = \mathcal{F}_{i,\,1} \times \cdots \times \mathcal{F}_{i,\,H}$，其中每个 $f_{i,\,h} \in \mathcal{F}_{i,\,h} \subset (\mathcal{S} \times \mathcal{A}_i \to \mathbb{R})$ 对应于第 $h \in [H]$ 步的边缘 Q 函数。[6]

在对 $\mathcal{F}_i$ 和博弈（将在下文介绍）作出适当假设后，我们希望找到一个近似相关均衡，其样本复杂度能够避免多智能体维数灾难（Jin 等人，2021b；Song 等人，2021年），即关于$\max^{i\in[m]} \log|\mathcal{F}_i|$、参与者人数 m以及所有其他问题参数呈多项式增长。

3 去中心化 基于策略重放的多智能体强化学习：元算法与理论保证

算法 我们的首个主要算法——带策略回放的 V 学习（VLPR；算法 1）——是一种用于函数逼近的去中心化多智能体强化学习元算法。总体而言，VLPR 采用了策略重放机制（第 3 行），在第 t 次迭代中，将滚入策略 $\overline{\pi}^t = \text{Unif}(\{\pi^\tau\}_{\tau\in[t]})$ 设为此前学得的所有策略的均匀混合。随后，它利用该滚入策略，通过分阶段学习递归计算近似 CCE 策略和 V值，从 $h = H$ 到1，并使用两个子程序：

- CCE-APPROX$_h$（算法2）接收价值估计 $\{\overline{V}_{i,h+1}^{t+1}\}_{i\in[m]}$，并计算第 h 步的近似相关均衡 π_h^{t+1}。它需要两个要素：(1) 一组有序的探索策略和活跃玩家 $(\widetilde{\pi}, P) \in \Gamma$explore$(\overline{\pi}, \mu_h^k)$ ($P \subseteq [m]$ 是索引集)，其中每轮都会执行每个这样的 $\widetilde{\pi}$ 以观察一条轨迹，并将观测 $(s_h, a_{i,h}, r_{i,h} + \overline{V}_{i,h+1}(s_{h+1}))$ 添加到第 i 个玩家的数据集 $\mathcal{D}^{k,i}$sample 中，当且仅当 $i \in P$。(2) 然后，每个玩家使用收集到的数据运行无悔算法 NO-REGRET-ALG。我们要求 NO-REGRET-ALG 相对较强：在老虎机反馈下，即使面对大规模状态空间（在适当意义上），也能实现较小的逐状态遗憾（参见条件 (1A)）；稍后将对此进行讨论。

[5]通过标准的覆盖论证，我们的结果可直接扩展到函数类无限的情况。[6]虽然我们侧重于 Q 型函数逼近，但我们的元算法也可以扩展到 V 型函数逼近，不过这两种类型可能涵盖相当不同的问题结构；讨论请见附录 B。

算法2 CCE-近似$_h$($\overline{\pi}$, $\{\overline{V}_{i,h+1}\}_{i\in[m]}$, K)

要求：探索策略映射 Γ_{explore}；子程序 NO-REGRET-ALG。1：执行 $\overline{\pi}$ 个回合以收集 $\{\mathcal{D}^i_{\text{init}}\}_{i\in[m]}$。对所有 $(i,k) \in [m]\times[K]$，初始化 $\mathcal{D}^k{}_{,i}$ sample $\leftarrow \{\}$。2：对于 $k = 1,\ldots,K$，执行以下操作。3：对于 $(\widetilde{\pi}, P) \in \Gamma_{\text{explore}}(\overline{\pi}, \mu^k_h)$，执行以下操作。4：执行 $\widetilde{\pi}$ 以收集一条轨迹 $(s_1, \mathsf{a}_1, \mathsf{r}1, \ldots, s_H, \mathsf{a}_H, \mathsf{r}_H)$。5：对所有 $i \in P$，更新 $\mathcal{D}^k{}_{,i}$ sample $\leftarrow \mathcal{D}^k{}_{,i}$ sample $\cup \{(s_h, a_{i,h}, r_{i,h} + \overline{V}_{i,h+1}(s_{h+1}))\}$。6：对所有 $i \in [m]$，更新 $\mu_i^{k+1}{}_{,h} \leftarrow$ NO-REGRET-ALG($\mu^k_{i,h}, \mathcal{D}^k{}_{,i}$ sample $\mathcal{D}^i_{\text{init}}$)。输出：$\pi^{\text{out}}_h$: $= \frac{1}{K}\sum_{k=1}^K \mu^k_h$，其中 $\mu^k_h = \mu^k_{1,h} \times \cdots \times \mu^k_{m,h}$。

算法3 V-近似$_h$($\overline{\pi}$, π_h, $\{\overline{V}_{i,h+1}\}_{i\in[m]}$, K)

要求：探索策略映射 Γ_{explore}；子程序 OPTIMISTIC-REGRESS。1：对所有 $i \in [m]$ 初始化 $\mathcal{D}^i_{\text{reg}} \leftarrow \{\}$。2：对于 $k = 1,\ldots,K$，执行 3：对于 $(\widetilde{\pi}, P) \in \Gamma_{\text{explore}}(\overline{\pi}, \pi_h)$，执行 4：执行 $\widetilde{\pi}$，以收集轨迹 $(s_1, \mathsf{a}_1, \mathsf{r}_1, \ldots, s_H, \mathsf{a}_H, \mathsf{r}_H)$。5：对所有 $i \in P$，将 $(s_h, a_{i,h}, r_{i,h} + \overline{V}_{i,h+1}(s_{h+1}))$ 加入 $\mathcal{D}^i_{\text{reg}}$。6：对所有 $i \in [m]$，$\overline{V}_{i,h} \leftarrow$ OPTIMISTIC-REGRESS($\pi_{i,h}, \mathcal{D}^i_{\text{reg}}$)。输出：$\{\overline{V}_{i,h}\}_{i\in[m]}$。

- V-近似$_h$（算法 3）接收新的策略 π^{t+1}_h 和价值估计 $\{\overline{V}^{t+1}_{i,h+1}\}_{i\in[m]}$，并通过回归算法 $\{\overline{V}^{t+1}_{i,h}\}_{i\in[m]}$ 为第 h 步生成估计，该算法为 乐观-回归，其实现小误差的乐观估计是必要的（参见条件 （1B））。

值得注意的是，VLPR 结合了策略重放机制和 V-近似 子程序，该子程序在每次迭代中以 批量 方式重新学习一个新的价值函数。这一机制不同于标准的 V 学习算法：后者在每次迭代中直接执行新学习的策略而不进行重放，但采用增量价值更新。该机制实际上学习了一个隐式“输出策略”（即“认证策略”）的价值，该策略不同于之前执行过的策略（Bai 等人，2020年；Jin 等人，2021b；Song 等人，2021年；Mao 和 Bas‑ar，2022年）。然而，在存在函数逼近的情况下，VLPR 中的批量学习更为理想，并且正是由策略重放机制实现的，因为否则很难明确如何将增量价值更新方法推广到一般函数类的情形。

条件与保证 VLPR 是一种通用元算法。一旦子程序满足特定要求，该元算法就能保证学习到博弈的近似粗相关均衡。

条件1 （VLPR所需条件）。 存在 奖励函数 $G_{i,\ h}(s,\ \overline{\pi},\ K,\ \delta)$ 对于每个 $(i,\ h) \in [m]\times[H]$，执行算法 *1*时满足以下条件。

（1A）逐状态无遗憾：子程序 $\pi = \text{CCE-APPROX}_h(\overline{\pi},\ \{\overline{V}_{i,h+1}\}_{i\in[m]},\ K)$（算法 *2*）满足：其以至少 $1-\delta$ 的概率，对所有 $(i,\ s) \in [m]\times\mathcal{S}$:

$$\max_{\mu_{i,h}\in\Delta(\mathcal{A}_i)} \left(\mathbb{D}_{\mu_{i,h}\times\pi_{-i,h}} - \mathbb{D}_{\pi_h}\right)\left[r_{i,h} + \mathbb{P}_{h+1}\overline{V}_{i,h+1}\right](s) \le G_{i,h}(s,\overline{\pi},K,\delta).$$

(1B) 乐观V估计：子程序$\overline{V}_{i,\ h} = \text{V-APPROX}_h(\overline{\pi},\ \pi_h,\ \{\overline{V}_{i,h+1}\}_{i\in[m]},\ K)$（算法*3*）满足：以至少 $1-\delta$的概率，对所有$(i,\ s) \in [m]\times\mathcal{S}$:

$$\begin{cases} \overline{V}_{i,h}(s) \ge \min\left\{\mathbb{D}_{\pi_h}\left[r_{i,h} + \mathbb{P}_{h+1}\overline{V}_{i,h+1}\right](s) + G_{i,h}(s,\overline{\pi},K,\delta), H-h+1\right\}, \\ \overline{V}_{i,h}(s) \le \quad \mathbb{D}_{\pi_h}\left[r_{i,h} + \mathbb{P}_{h+1}\overline{V}_{i,h+1}\right](s) + 2G_{i,h}(s,\overline{\pi},K,\delta). \end{cases}$$

(1C) 鸽巢条件：存在一个绝对复杂度度量 $L \in \mathbb{R}^+$，使得对于任意$(i,h) \in [m] \times [H]$、$(T,\delta) \in \mathbb{N} \times (0,1)$以及任意策略序列 $\{\pi^1, \ldots, \pi^T\}$，

$$\sum_{t=1}^{T} \mathbb{E}_{s_h \sim \pi^{t+1}} \left[G_{i,h}(s_h, \mathrm{Unif}(\{\pi^\tau\}_{\tau \in [t]}), t, \delta) \right] \le \sqrt{LT \log^2(T/\delta)}.$$

条件 (1A) 要求子程序 CCE-APPROX (它调用 NO-REGRET-ALG) 实现 逐状态 低遗憾（回想一下，在算法 2 中，输出策略是所采用策略的均匀混合）。这比标准 上下文 老虎机问题中的固定状态分布遗憾界更严格（ Lattimore and Szepesv´ari，2020年），但这对于学习粗相关均衡至关重要，因为这要求学到的策略能够很好地泛化到多种滚入分布。

条件 (1B) 要求子程序 V-APPROX（它调用 OPTIMISTIC-REGRESS）为策略 π_h 生成乐观且准确的价值估计，具体而言，估计值 $\overline{V}_{i,\,h}$ 与真实值 $\mathbb{D}_{\pi_h}\left[r_{i,h} + \mathbb{P}_{h+1}\overline{V}_{i,h+1}\right]$ 之间的差值（经截断后）被限制在 $[1,2]$ 倍奖励函数 $G_{i,\,h}$范围内。

条件 (1C) 与鸽巢原理有相似之处，用于确保期望奖励之和达到 $\widetilde{\mathcal{O}}(\sqrt{T})$，正如 UCB风格算法中所采用的那样，例如 Azar 等人 (2017年)； Jin 等人 (2020年)。

现在我们可以陈述 VLPR 的主要保证了。

定理 2 （VLPR 的“遗憾”保证）。 假设条件 *1* 对算法 *1*成立。则以至少 $1 - 3\delta$的概率，有

$$\mathrm{CCEReg}(T) := \max_{i \in [m]} \sum_{t=1}^{T} \left[V_{i,1}^{\dagger, \pi_{-i}^t}(s_1) - V_{i,1}^{\pi^t}(s_1) \right]_+ \le \widetilde{\mathcal{O}}(\sqrt{H^2 LT}). \tag{1}$$

推论3（样本复杂度）。选择 $T = \widetilde{\mathcal{O}}(H^2 L/\varepsilon^2)$ 可确保算法 *1*的输出策略 π^{out} 满足 $\mathrm{CCEGap}(\pi^{\mathrm{out}}) \le \varepsilon$，并且以至少[7] 0.99 的概率，回合总数至多为 $\widetilde{\mathcal{O}}\big(H^5 L^2 \overline{\Gamma}/\varepsilon^4\big)$，其中 $\overline{\Gamma} := \max_{\overline{\pi}, \pi'} |\Gamma_{\mathrm{explore}}(\overline{\pi}, \pi')|$。

定理 2 和推论 3 断言，只要算法 ε在 $\mathrm{poly}(H, L, \overline{\Gamma}, 1/\varepsilon)$ 个样本内满足条件 1 中的所有子程序均满足条件 1，即可找到一个 ε-近似相关均衡。证明（见附录 C.1）在这些条件下相对直接：它利用性能差异论证，并结合条件 (1A) & (1B)，通过奖励项对 CCE遗憾(T) 给出上界，再利用条件 (1C) 进一步界定 $t \in [T]$上的奖励总和。

˜3.1 通过低频策略更新加速 $\mathcal{O}(1/\varepsilon^2)$ 算法

定理 $\widetilde{\mathcal{O}}(1/\varepsilon^4)$ 中得到的速率2 慢于标准的 $1/\varepsilon^2$ 速率。这是因为 VLPR 采用回放机制，并在每次迭代时更新策略 $t \in [T]$，从而导致 $T \times T = \widetilde{\mathcal{O}}(1/\varepsilon^4)$ 速率。然而，如果回放策略诱导的滚入分布$\{\overline{\pi}^t\}_{t \ge 1}$ 随时间变化不快，那么如此频繁地更新策略可能并无必要 t。

为此，我们设计了一种名为 AVLPR （算法 5）的加速算法，在满足附加条件（条件 17）的情况下，将该速率提升至 $1/\varepsilon^2$；该条件使算法能够通过 低频策略更新——更准确地说，是在 $\mathcal{O}$（log T） 次更新——在 T 次迭代内表现良好（定理 18）。我们将通过 倍增技巧实现这一条件。详情请参见附录C.2 。

3.2 去中心化执行

到目前为止，我们对算法 VLPR 和 AVLPR 的描述都是针对所有参与者共同进行的。尽管如此，这两种算法都可以用去中心化方式实现。严格来说，我们考虑这样的设置：每位参与者只能

[7]通过对粗相关均衡差距进行乐观评估并结合提升，可以进一步将成功概率提高至 $1 - \delta$，对于任意小的 $\delta > 0$，样本复杂度至多增加一个 $\log(1/\delta)$ 因子。

查看共享状态以及自己的动作和奖励。也就是说，如果没有通信，他们就无法获知其他参与者的动作或奖励。我们将证明，使用一些简单的协议，VLPR 可以在完全去中心化的方式下执行，且无需任何通信（假设参与者之间具有共享随机性）；而 AVLPR 则只需在检查触发条件时进行 $\mathcal{O}(\log T)$ 轮次的极少量通信（算法 5 中的第 4 行）。详细论证见附录 C.4。

4 线性与表格型设置中的实例化

现在，我们具体地在两种设置中实例化 AVLPR： 去中心化线性函数逼近（一个新的设置），以及在表格型马尔可夫博弈中学习马尔可夫粗相关均衡。这里我们重点关注样本复杂度；这两种实例化在计算上也都高效（参见附录 E.1 & G.1）。

4.1 去中心化线性函数逼近

我们考虑具有去中心化线性函数逼近的马尔可夫博弈，其中每个 $\mathcal{F}_{i,h} = \{f_{i,h}(\cdot,\cdot) = \phi_i(\cdot,\cdot)^\top \theta_h : \|\theta_h\|_2 \le B_\theta := H\sqrt{d}\}$ 都是关于已知的 d 维特征映射[8] $\phi_i : \mathcal{S} \times \mathcal{A}_i \to \mathbb{R}^d$ 的线性函数类。我们考虑由特征映射 $\mathcal{F}_i$ 诱导的线性 argmax 策略类

$$\Pi^{\text{lin}}_{i,h} := \left\{\pi_{i,h}(\cdot|s) = \arg\max_{a_i \in \mathcal{A}_i} \phi_i(s,a_i)^\top w_{i,h},\ \forall\, s \in \mathcal{S} \mid w_{i,h} \in \mathbb{R}^d\right\}. \tag{2}$$

并记作 $\Pi^{\text{lin}}_i = \bigtimes_{h\in[H]} \Pi^{\text{lin}}_{i,h}$ 和 $\Pi^{\text{lin}} := \bigtimes_{i\in[m]} \Pi^{\text{lin}}_i$。为确保特征映射具有足够的信息量，我们作出如下假设。

假设 4（Π^{lin}-完备性）。对于任意 $(i,\ h) \in [m] \times [H]$、任意 $f_{i,\ h+1} : \mathcal{S} \times \mathcal{A}_i \to [0, H]$、任意 $\pi \in \Pi^{\text{lin}}$，我们有 $\mathcal{T}^\pi_{i,\ h} f_{i,\ h+1} \in \mathcal{F}_{i,\ h}$。

在 $m = 1$（单智能体设置）中，假设 4 严格弱于线性 MDP 假设（Jin 等人，2020年），但强于线性完备性假设（Zanette 等人，2020年）；这两者都是采用线性函数逼近的强化学习中常见的假设。对于 $m \ge 2$，假设 4 可以看作线性 MDP 假设的去中心化多智能体推广，该假设要求，对于每个玩家 $i \in [m]$，任意 $\overline{V}_{i,\ h+1}$ 相对于任意线性 argmax 策略 π_{-i} 的贝尔曼回传都包含于 $\mathcal{F}_{i,\ h}$ 中（因此关于 ϕ_i（s，a_i）是线性的）。

我们指出，在假设4中，仅要求受限策略类 Π^{lin} 满足完备性至关重要：如果要求所有马尔可夫策略都满足完备性，那么该博弈就会在某种意义上“本质上是表格型”的，即非平凡状态的数量必须很少（参见附录D）。

主要结果针对去中心化线性函数逼近，我们具体化AVLPR，从而得到以下保证。算法细节和证明见附录E。

定理 5(用于去中心化线性函数逼近的AVLPR)。假设去中心化线性函数逼近满足假设4。那么，适当实例化后的 *AVLPR* 可在 ε-粗相关均衡所需的 $\widetilde{\mathcal{O}}\left(d^4H^6m^2(\max_{i\in[m]} A_i^5)/\varepsilon^2\right)$个博弈回合内找到该均衡。

定理5 实现了 $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ 样本复杂度，其对$(d, H, m, \max_{i\in[m]} A_i)$呈多项式依赖，从而避免了多主体诅咒。据我们所知，这是关于通过去中心化线性函数逼近学习马尔可夫博弈的首个此类结果。

技术概述 建立定理5 需要在O-Regret-Alg 和 Optimistic-Regress子程序中进行实例化，使其适用于线性函数逼近AVLPR 设置，并满足条件 (1A)-(1C) & 17 。我们选择 Optimistic-Regress 采用标准岭回归，根据假设(1B) 确保条件 4。

[8]不失一般性，我们假设特征有界：对于所有 $i \in [m]$，$\sup_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i} \|\phi_i(s,a_i)\|_2 \le B_\phi := 1$。

更具挑战性的任务是选择满足条件 O-REGRET-ALG (1A)，粗略而言，该条件要求：（1）在所有 $s \in \mathcal{S}$ 中提供逐状态遗憾保证；（2）策略 $\{\mu_h^k\}_{k\in[K]}$ 属于 Π^{lin}。或许有悖直觉的是，这排除了两种做法：一是在每个状态下分别运行一个线性对抗老虎机算法，因为这违反了（2）；二是运行对抗性上下文 线性老虎机算法，例如 LINEXP3（Neu 和 Olkhovskaya，2020年），因为这违反了（1）。我们通过利用 $\mathcal{S}$ 的特殊结构，将问题转换为 Π^{lin}并行的在线线性优化问题，然后应用期望扰动跟随领袖算法（Hazan 和 Minasyan，2020年），在 Π^{lin} 中生成一组同时解决所有 $\mathcal{S}$ 问题的迭代点（速率不依赖于任何 $|\mathcal{S}|$），从而满足这两个要求。

选定这些子程序后，我们证明条件 1 在采用以下奖励函数时成立

$$G_{i,h}(s,\overline{\pi},K,\delta) := \widetilde{\Theta}\left(\max_{a_i\in\mathcal{A}_i}\|\phi_i(s,a_i)\|_{(\Sigma_{i,h}^{\overline{\pi}}+\lambda I)^{-1}} \times d(\max_i A_i^{1.5})H/\sqrt{K}+K^{-1}\right), \tag{3}$$
$$\text{where}\quad \Sigma_{i,h}^{\overline{\pi}} := \mathbb{E}_{s_h\sim\overline{\pi}}\,\mathbb{E}_{a_{i,h}\sim\text{Unif}(\mathcal{A}_i)}\left[\phi_i(s_h,a_{i,h})\phi_i(s_h,a_{i,h})^\top\right],\quad \lambda=\widetilde{\Theta}(d(\max_i A_i)/K).$$

4.2 在表格型马尔可夫博弈中学习马尔可夫粗相关均衡

我们还将 AVLPR 应用于表格型马尔可夫博弈（其中 $\mathcal{F}_i$ 是所有可能边际 Q 函数的类），并得到以下结果（算法细节和证明见附录 G）。

定理 6（表格型马尔可夫博弈）。对于具有 S 个状态的表格型马尔可夫博弈，适当实例化的 *AVLPR* 能在 $\widetilde{\mathcal{O}}\left(H^6S^2(\max_{i\in[m]}A_i)/\varepsilon^2\right)$ 个博弈回合内找到一个马尔可夫 ε-粗相关均衡。

目前，唯一一种能够避免多主体诅咒、用于学习马尔可夫粗相关均衡的算法是 Daskalakis 等人 （2022年）提出的 SPoCMAR 算法，其样本复杂度为 $\widetilde{\mathcal{O}}\left(H^{10}S^3(\max_i A_i)/\varepsilon^3\right)$。定理 6 在 (H,S) 依赖关系方面有所改进，并达到近乎最优的 $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ 速率。为证明定理 6，我们将 NO-REGRET-ALG 实例化为在每个状态 $s\in\mathcal{S}$处分别运行一个 EXP3 算法，并将 OPTIMISTIC-REGRESS 直接设为逐状态乐观价值估计。我们证明，这些选择使条件 1 在以下奖励函数下成立：

$$G_{i,h}(s,\overline{\pi},K,\delta) := \widetilde{\Theta}\left(\eta_i^{-1}(J_h(s)+\iota)^{-1}+\eta_i H^2 A_i\right),$$

其中，η_i 是第 i名玩家的 NO-REGRET-ALG 的学习率，$J_h(s)$ 是使用滚入策略 $\overline{\pi}$ 运行 K 个回合时，在第 h 步访问状态 $\tilde{s}$ 的期望访问次数，且 $\iota=\mathcal{O}(1)$。

5 在受限策略类中学习粗相关均衡

本节将介绍一种在函数类的假设可能宽松得多的情况下，于受限策略类 Π 中学习粗相关均衡（以下简称 Π-粗相关均衡）的替代方法。

受限策略类 我们让每个玩家 $i\in[m]$ 都配备一个马尔可夫策略类 Π_i（以及其边际 Q 函数类 $\mathcal{F}_i$），并令 $\Pi := \prod_{i\in[m]}\Pi_i$ 表示在 $\{\Pi_i\}_{i\in[m]}$上的乘积策略集合。对于任意联合策略 Λ，如果 Λ 是一个 ε-近似 Π-粗相关均衡，则称其为

$$\text{CCEGap}^{\Pi}(\Lambda) := \max_{i\in[m]}\left(\max_{\pi_i^\dagger\in\Pi_i} V_{1,i}^{\pi_i^\dagger\times\Lambda_{-i}}(s_1)-V_{1,i}^{\Lambda}(s_1)\right)\le\varepsilon.$$

换言之，只要没有任何玩家通过偏离到 Π_i 中的其他策略而获得显著收益，Λ 就是一个近似的 Π-粗相关均衡。注意，我们始终有 $\text{CCEGap}^{\Pi}(\Lambda)\le\text{CCEGap}(\Lambda)$，并且即使 Π_i 是玩家 i 的所有可能马尔可夫策略的集合（此处允许的最大类），该不等式通常也是严格的[9]，因此，Π-粗相关均衡通常是一个限制性更强的概念。

[9]具体而言，存在一个马尔可夫博弈，其中存在一个 $\Lambda\in\Delta(\Pi^{\text{Mar}})$ 使得 $\text{CCEGap}^{\Pi^{\text{Mar}}}(\Lambda)=0$ 但 $\text{CCEGap}(\Lambda)\ge H/4$ 对于任意 $H\ge 2$；构造见附录 H 。

算法 4 DOPMD：去中心化乐观策略镜像下降

要求：学习率 $\{\eta_i\}_{i\in[m]}$、函数类 $\{\mathcal{F}_i\}_{i\in[m]}$、策略类 $\{\Pi_i\}_{i\in[m]}$、 $\{(K_i,\beta_i)\}_{i\in[m]}$。1：对所有 $i\in[m]$，初始化 $\Lambda_i^1 \leftarrow$均匀分布(Π_i)。2：对于轮次 $t=1$，...，T，执行 3：为每个 $i\in[m]$采样一个策略 $\pi_i^t\sim\Lambda_i^t$，并设置 $\pi^t=\pi_1^t\times\ldots\times\pi_m^t$。4：对于 $i\in[m]$，执行 5：获取第 i 个玩家的乐观估计 $\{\overline{V}^{(t}_{i\ \ \pi_i\times\pi_{-i}^t)}\}_{\pi_i\in\Pi_i}\leftarrow$APE$_i(\mathcal{F}_i,\Pi_i,\pi_{-i}^t,K_i,\beta_i)$。6：更新 $\Lambda_i^{t+1}(\pi_i)\propto_{\pi_i}\Lambda_i^t(\pi_i)\cdot\exp(\eta_i\cdot\overline{V}^{(t}_{i\ \ \pi_i\times\pi_{-i}^t)})$。输出：平均策略 $\overline{\Lambda}:=\frac{1}{T}\sum_{t\in[T]}\Lambda_1^t\times\cdots\times\Lambda_m^t$。

假设 我们的第一个假设要求每个函数类 $\mathcal{F}_i$ 相对于贝尔曼算子具有完备性$\{\mathcal{T}_{i,h}^{\pi}\}_{\pi,\ h}$这是通过平方损失回归确保准确估计价值的标准假设（Jin 等人，2021a）。由于该假设仅适用于 $f_{i,\ h+1}\in\mathcal{F}_{i,\ h+1}$（而假设4适用于任意 $f_{i,\ h+1}$），因此它放宽了假设 4 。

假设 7（Π-完备性）。对于每个 $i\in[m]$，函数类 $\mathcal{F}_i$ 相对于 Π满足完备性，也就是说，对于任意 $h\in[H]$ 和 $(f_{i,\ h+1},\pi)\in\mathcal{F}_{i,\ h+1}\times\Pi$，我们有 $\mathcal{T}_i^{\pi},\ _h f_{i,\ h+1}\in\mathcal{F}_{i,\ h}$。
我们还要求每个 $\mathcal{F}_i\subset((\mathcal{S}\times\mathcal{A}_i)\to[0,H])$ 具有有界的贝尔曼–埃卢德（BE）维数（Jin 等人，2021a），以确保样本高效强化学习。对于任意 $i\in[m]$，我们定义

$$d_i(\mathcal{F}_i,\Pi,\varepsilon):=\max_{\pi_{-i}\in\Pi_{-i}}d_{\pi_{-i}}^{\mathrm{BE}}(\mathcal{F}_i,\Pi_i,\varepsilon),\tag{4}$$

其中 $d_{\pi_{-i}}^{\mathrm{BE}}(\mathcal{F}_i,\ \Pi_i,\ \varepsilon)$ 表示 $\mathcal{F}_i$ 相对于贝尔曼算子的 Bellman-Eluder维数 $\{\mathcal{T}_{i,h}^{\pi_i\times\pi_{-i}}\}_{\pi_i\in\Pi_i}$（参见定义 16）。Bellman-Eluder维数是单智能体强化学习中的标准复杂度度量，用于控制探索的复杂度。我们假设所有参与者的边际价值函数的 Bellman-Eluder维数均有界 $i\in[m]$。

假设 8（有界 BE 维数）。存在标量 $\{d_i\}_{i\in[m]}$，使得对于所有 $i\in[m]$ 和 $\varepsilon\in$（0,1），均有 $d_i(\mathcal{F}_i,\ \Pi,\ \varepsilon)\le d_i\log(1/\varepsilon)$。
注意，假设 7 & 8 在本质上都是去中心化的，因为它们只要求关于固定（$\mathcal{F}_i,\ \Pi_i$）所诱导的单智能体马尔可夫决策过程中的 $\pi_{-i}\in\Pi_{-i}$性质。这些条件不同于以往使用一般函数逼近学习马尔可夫博弈的方法，后者要求其集中式函数类满足类似的结构条件（Jin 等人，2022年；Huang 等人，2021年；Chen 等人，2022年a）。

5.1 算法与保证

我们的算法去中心化乐观策略镜像下降（DOPMD，算法 4）是一种双循环算法。其外循环与（Liu 等人，2022年；Zhan 等人，2022年）中的策略镜像下降算法类似，其中每个玩家 $i\in[m]$ 维护 Λ_i^t——一个在 Π_i 上的策略分布。随后，玩家采样一个策略 $\pi_i^t\sim\Lambda_i^t$（第 3 行），获取乐观价值估计（第 5 行），并在策略空间中使用这些乐观价值估计执行镜像下降/Hedge（第 6 行），从而得到更新 $\Lambda_i^{t+1}\in\Delta(\Pi_i)$。

我们算法中的关键新要素是子程序 APE（探索式全策略评估；完整描述见算法 6），用于获取乐观价值估计。对于每个玩家 $i\in[m]$，子程序 APE$_i(\mathcal{F}_i,\ \Pi_i,\ \pi_{-i}^t,\ K_i,\ \beta_i)$执行 K_i 个回合，并在由（固定的）对手策略 π_{-i}^t 所诱导的马尔可夫决策过程中，获取对 所有 $\pi_i\in\Pi_i$ 的准确价值估计。概括而言，APE 对 Jin 等人 提出的 GOLF 算法进行了修改（2021a），通过执行使不确定性最大化的策略：

$$\pi_i^k:=\arg\max_{\pi_i\in\Pi_i}\left\{\max_{f:(f,\pi_i)\in\mathcal{B}^k}f_1(s_1,\pi_{i,1}(s_1))-\min_{f:(f,\pi_i)\in\mathcal{B}^k}f_1(s_1,\pi_{i,1}(s_1))\right\},$$

由平方损失置信集 $\mathcal{B}^k$ 指定，而不是像 GOLF 那样最大化乐观价值估计。

理论保证现在我们可以陈述 DOPMD 算法的保证。证明见附录I.2。

定理 9（DOPMD 的保证）。在假设 7 和 8 下，对于任意 $\varepsilon > 0$，算法 4 使用 $\eta_i = \sqrt{\log |\Pi_i|/(H^2T)}, K_i = \widetilde{\mathcal{O}}(H^4 d_i \log(\sum_{i\in[m]} |\Pi_i||\mathcal{F}_i|/\varepsilon^2)$，$\beta_i = \widetilde{\mathcal{O}}(H^2 \log(\sum_{i\in[m]} |\Pi_i||\mathcal{F}_i|))$，至多经过 $T \le \widetilde{\mathcal{O}}(H^2 \log(\sum_{i\in[m]} |\Pi_i|)/\varepsilon^2)$ 轮次，即可输出一个 ε-近似 Π-粗相关均衡。

所进行的回合总数至多为

$$T \times \left(\sum_{i\in[m]} K_i\right) = \widetilde{\mathcal{O}}\left(H^6 \left(\sum_{i\in[m]} d_i\right) \log^2(\textstyle\sum_i |\Pi_i||\mathcal{F}_i|)/\varepsilon^4\right).$$

其中，$\widetilde{\mathcal{O}}(\cdot)$ 隐去了关于 $H, d_i, \varepsilon, \delta, \log|\mathcal{F}_i|, \log|\Pi_i|, m$ 的多对数因子。

定理 9 关于学习 ε-近似 Π-粗相关均衡的样本复杂度，是关于 BE维度、函数类和策略类基数的对数，以及 $1/\varepsilon$ 的多项式。虽然 Π- 粗相关均衡保证弱于 VLPR 或 AVLPR 算法（定理2 & 18），但作为回报，定理 9 仅要求满足 BE维度和完备性假设，这些条件是一般函数逼近的标准假设，而且可能比第 1 节中要求的条件3更宽松得多。

去中心化执行注意，i号玩家的APE 仅使用自己的边际 Q 函数类 $\mathcal{F}_i$ 和局部观测来估计所有 $\pi_i \in \Pi_i$的值，因此算法4 可以通过让每位玩家在每轮按APE 的字典序执行，以去中心化方式运行。因此，玩家之间既不需要通信，也不需要共享随机性。这不同于 Liu 等人（Liu 等人(2022年）；Zhan 等人(2022年）采用联合 Q 类的集中式算法。

5.2 示例

我们首先说明，假设7 & 8 在学习 Π-粗相关均衡时成立，适用于线性二次博弈（Zhang 等人，2019年）——这是一类具有连续状态/动作和线性状态转移的特殊马尔可夫博弈，并采用线性策略类和线性值函数类。

示例 10（线性二次博弈（LQG））。我们考虑由状态空间$\mathcal{S} \subset \mathbb{R}^{d_S}$ 和动作空间 $\{\mathcal{A}_i \subset \mathbb{R}^{d_{A,i}}\}_{i\in[m]}$指定的 m 人有限时域线性二次博弈。初始状态 $s_1 \in \mathbb{R}^{d_s}$ 固定，且第 h 步的状态转移为

$$s_{h+1} = A_h s_h + \textstyle\sum_{i=1}^m B_{i,h} a_{i,h} + z_h, \tag{5}$$

其中，$A_h \in \mathbb{R}^{d_S \times d_S}$、$B_{i,h} \in \mathbb{R}^{d_S \times d_{A,i}}$ 是博弈参数，z_h 是相互独立的零均值噪声。对于所有（i、h）$\in [m] \times [H]$，奖励定义为 $r_{i,h}(s, a) = s_h^\top K_h^i s_h + \sum_{j=1}^m a_{j,h}^\top K_{j,h}^i a_{j,h}$，其中 $K_h^i \in \mathbb{R}^{d_S \times d_S}$、$K_{j,h}^i \in \mathbb{R}^{d_{A,j} \times d_{A,j}}$ 是博弈参数。

LQG 的一个重要策略类是线性策略类（记为 Π），其形式为 $\pi_{i,h}(s) = M_{i,h} s$；例如，在标准假设下，该类包含博弈的粗相关均衡（Basar and Bernhard，2008年）。在附录 I.5 中，我们证明，具有适当选择的线性策略类和线性价值类的此类 LQG 满足假设 7 和8，其中 $d_i = \mathcal{O}(d_s + d_{A,i})^2$；此外，DOPMD 可通过 $\widetilde{\mathcal{O}}$（poly（H，$\sum_{i\in[m]} d_i$）$/\varepsilon^4$）个样本实现对 Π-粗相关均衡的样本高效学习。

相比之下，VLPR/AVLPR 不太可能在例子 10—条件(1B) 通常需要Π-完备的乐观价值（即线性函数加奖励项）；一个充分条件是 Π-完备性，涵盖步 $h+1$ 的所有价值，如假设4所述。此处这类乐观价值不再是线性的，因此不太可能包含在步 h的线性函数类中。

其次，更一般地，由于假设 8 仅要求每位玩家在去中心化意义下具有有界的 Bellman-Eluder 维数（参见定义 16），因此它涵盖了丰富的子类，例如每位玩家所诱导的边缘 马尔可夫决策过程具有低 Eluder 维度或低贝尔曼秩；论证与（Jin 等人，2021a，命题 11 和 12）类似。

例 11 （低 Eluder 维度）。 假设对于所有 $i \in [m]$， $\mathcal{F}_i$ 具有 低 Eluder 维度 （Wang 等人 （2020年）； 参见定义 *14*），具体而言，$\max_{h\in[H]} d_{\mathrm{E}}(\mathcal{F}_{i,\ h},\ \varepsilon) \leq d_i \log(1/\varepsilon)$，并满足 Π-完备性（假设 *7*）。那么，假设 *8* 也成立，且具有相同的 $\{d_i\}_{i\in[m]}$。
特别是，低 Eluder 维度的函数类包含某些非线性函数类，例如广义线性模型（Russo 和 Van Roy，2013年），其形式为 $\mathcal{F}_{i,\ h} = \{Q_{i,h}(\cdot,\cdot) = \sigma(\phi_i(\cdot,\cdot)^\top \theta_{i,\ h}) : \theta_{i,\ h} \in \mathbb{R}^{d_i}\}$，其中 $\phi_i : \mathcal{S}\times\mathcal{A}_i \to \mathbb{R}^{d_i}$ 是特征映射，而 $\sigma : \mathbb{R} \to \mathbb{R}$ 是链接函数，且对于某个 $0 < c_1 < c_2$，有 $\sigma'(\cdot) \in [c_1, c_2]$。
例 12（低贝尔曼秩）。假设对于所有 $i \in [m]$，任意 $\pi_{-i} \in \Pi_{-i}$ 所诱导的单智能体马尔可夫决策过程都具有低贝尔曼秩（Jiang 等人，2017年），具体来说：对于任意固定的 $\pi_{-i} \in \Pi_{-i}$，存在映射 $\psi_{i\ h}^{\pi_{-i}}, : \Pi_i \to \mathbb{R}^{d_i}$，$\phi_{i\ h}^{\pi_{-i}}, : \mathcal{F}_i \times \Pi_i \to \mathbb{R}^{d_i}$，使得对于任意 $f \in \mathcal{F}_i, \pi_i, \pi_i' \in \Pi_i$

$$\mathbb{E}_{(s_h,a_{i,h})\sim\pi_i'\times\pi_{-i}}\left[(f_h - \mathcal{T}_h^{\pi_i\times\pi_{-i}} f_{h+1})(s_h, a_{i,h})\right] = \left\langle \phi_{i,h}^{\pi_{-i}}(f,\pi_i), \psi_{i,h}^{\pi_{-i}}(\pi_i')\right\rangle.$$

然后，假设*8* 在相同的 $\{d_i\}_{i\in[m]}$下成立。

6 结论

本文首次给出了在函数逼近下实现可证明高效的去中心化多智能体强化学习且避免多主体诅咒的一系列结果。我们提出了两种互补的方法：第一种基于策略重放和分阶段无遗憾学习，并具体应用于线性设置和表格设置，在这两种设置下，学习一个 ε-近似粗相关均衡的收敛速率均达到近乎最优的 $\widetilde{\mathcal{O}}(\varepsilon^{-2})$；第二种采用结合去中心化探索的策略镜像下降，能够学习粗相关均衡的一个受限版本，但适用于更广泛的问题类别。我们认为，本研究为未来工作开辟了许多有趣的方向，例如：(1) 推导这两种方法更精确的样本复杂度，尤其是改善线性设置下 AVLPR 对 $(d, \max_{i\in[m]} A_i)$ 的依赖，以及表格设置下对 S 的依赖；(2) 提高策略镜像下降方法的计算效率；(3) 确定适用于策略重放方法的新问题类别。

参考文献

Alekh Agarwal、Daniel Hsu、Satyen Kale、John Langford、Lihong Li 和 Robert Schapire。《驯服怪兽：一种快速而简单的上下文老虎机算法》。载于《国际机器学习会议》，第 1638–1646 页。PMLR，2014 年。

Mohammad Gheshlaghi Azar、Ian Osband 和 R´emi Munos。《强化学习的极小极大后悔界》。载于《国际机器学习会议》，第 263–272 页。PMLR，2017年。

Yu Bai 和 Chi Jin。《竞争性强化学习的可证明自我对弈算法》。载于《国际机器学习会议》，第 551–560 页。PMLR，2020年。

Yu Bai、Chi Jin 和 Tiancheng Yu。《通过自我博弈实现近乎最优的强化学习》。《神经信息处理系统进展》，33：2159–2170，2020年。

Yu Bai、Chi Jin、Huan Wang 和 Caiming Xiong。《一般和博弈中斯塔克尔伯格均衡的样本高效学习》。《神经信息处理系统进展》，34：25799–25811，2021年。

Yu Bai、Chi Jin、Song Mei、Ziang Song 和 Tiancheng Yu。通过在线镜像下降，在扩展式博弈中高效最小化 Φ-遗憾。arXiv 预印本 arXiv:2205.15294，2022年a。

Yu Bai、Chi Jin、Song Mei 和 Tiancheng Yu。对不完全信息扩展式博弈的近最优学习。arXiv 预印本 arXiv:2202.01752，2022b年。

Bowen Baker、Ingmar Kanitscheider、Todor Markov、Yi Wu、Glenn Powell、Bob McGrew 和 Igor Mordatch。多智能体自动课程中涌现出的工具使用。发表于 国际学习表征会议，2020年。网址：https://openreview.net/forum?id=SkxpxJBKwS。

Anton Bakhtin、Noam Brown、Emily Dinan、Gabriele Farina、Colin Flaherty、Daniel Fried、Andrew Goff、Jonathan Gray、Hengyuan Hu、Athul Paul Jacob 等人。结合语言模型与战略推理，在外交博弈中实现人类水平的对弈。《科学》，378(6624):1067–1074，2022年。Tamer Bas¸ar 和 Pierre Bernhard。《H∞最优控制与相关极小极大设计问题：一种动态博弈方法》。施普林格科学与商业媒体，2008年。Ronen I Brafman 和 Moshe Tennenholtz。R-max：一种用于近最优强化学习的通用多项式时间算法。《机器学习研究杂志》，3(10月):213–231，2002年。Manuele Brambilla、Eliseo Ferrante、Mauro Birattari 和 Marco Dorigo。群体机器人学：从群体工程视角出发的综述。《群体智能》，7(1):1–41，2013年。Noam Brown 和 Tuomas Sandholm。超人类人工智能玩单挑无限注扑克：Libratus 击败顶尖职业选手。《科学》，359(6374):418–424，2018年。Noam Brown 和 Tuomas Sandholm。超人类人工智能玩多人扑克。《科学》，365(6456):885–890，2019年。Nicolo Cesa-Bianchi 和 G´abor Lugosi。《预测、学习与博弈》。剑桥大学出版社，2006年。Fan Chen、Song Mei 和 Yu Bai。具有决策估计系数的强化学习统一算法：无遗憾、PAC 与无奖励学习。arXiv 预印本 arXiv:2209.11745，2022年a。Zixiang Chen、Dongruo Zhou 和 Quanquan Gu。双人零和线性混合马尔可夫博弈的近乎最优算法。发表于《国际算法学习理论会议》，第227–261页。PMLR，2022b年。Qiwen Cui 和 Simon S Du。通过策略级奖励实现可证明高效的离线多智能体强化学习。arXiv 预印本 arXiv:2206.00159，2022年。Qiwen Cui、Kaiqing Zhang 和 Simon S Du。打破大状态空间中多智能体的魔咒：采用独立线性函数逼近的马尔可夫博弈强化学习。arXiv 预印本 arXiv:2302.03673，2023年。Constantinos Daskalakis。关于近似纳什均衡的复杂度。《ACM算法汇刊》（TALG），9(3):1–35，2013年。Constantinos Daskalakis、Noah Golowich 和 Kaiqing Zhang。随机博弈中马尔可夫均衡的复杂度。arXiv 预印本 arXiv:2204.03991，2022年。Dongsheng Ding、Chen-Yu Wei、Kaiqing Zhang 和 Mihailo Jovanovic。大规模马尔可夫势博弈中的独立策略梯度：更优的收敛速率、函数逼近与博弈无关收敛。发表于《国际机器学习会议》，第5166–5220页。PMLR，2022年。Simon Du、Sham Kakade、Jason Lee、Shachar Lovett、Gaurav Mahajan、Wen Sun 和 Ruosong Wang。双线性类：用于强化学习中可证明泛化的结构框架。发表于《国际机器学习会议》，第2826–2836页。PMLR，2021年。Cˆome Fiegel、Pierre M´enard、Tadashi Kozuno、R´emi Munos、Vianney Perchet 和 Michal Valko。在零和不完全信息博弈中适应博弈树。arXiv 预印本 arXiv:2212.12567，2022年。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。Elad Hazan 和 Edgar Minasyan。更快的无投影在线学习。发表于《学习理论会议》，第1877–1893页。PMLR，2020年。Elad Hazan 等人。《在线凸优化导论》。《优化领域基础与趋势》，2(3-4):157–325，2016年。

Baihe Huang、Jason D Lee、Zhaoran Wang 和 Zhuoran Yang。迈向零和马尔可夫博弈中的一般函数逼近。arXiv 预印本 arXiv:2107.14702，2021年。Zeyu Jia、Lin F Yang 和 Mengdi Wang。双人随机博弈的基于特征的 Q 学习。arXiv 预印本 arXiv:1906.00423，2019年。Nan Jiang、Akshay Krishnamurthy、Alekh Agarwal、John Langford 和 Robert E Schapire。低贝尔曼秩的上下文决策过程具有 PAC 可学习性。载于《国际机器学习会议》，第 1704–1713 页。PMLR，2017年。Chi Jin、Zhuoran Yang、Zhaoran Wang 和 Michael I Jordan。通过线性函数逼近实现可证明高效的强化学习。载于《学习理论会议》，第 2137–2143 页。PMLR，2020年。Chi Jin、Qinghua Liu 和 Sobhan Miryoosefi。贝尔曼埃卢德维度：丰富的新型强化学习问题类别与样本高效算法。《神经信息处理系统进展》，34：13406–13418，2021a。Chi Jin、Qinghua Liu、Yuanhao Wang 和 Tiancheng Yu。V 学习——一种简单、高效的去中心化多智能体强化学习算法。arXiv 预印本 arXiv:2110.14555，2021b。Chi Jin、Qinghua Liu 和 Tiancheng Yu。利用者的力量：大规模状态空间中的可证明多智能体强化学习。载于《国际机器学习会议》，第 10251–10279 页。PMLR，2022年。Tadashi Kozuno、Pierre M´enard、R´emi Munos 和 Michal Valko。具有完美回忆的双人零和部分可观测马尔可夫博弈的无模型学习。arXiv 预印本 arXiv:2106.06279，2021年。Tor Lattimore 和 Csaba Szepesv´ari。《赌博机算法》。剑桥大学出版社，2020年。Stefanos Leonardos、Will Overman、Ioannis Panageas 和 Georgios Piliouras。多智能体策略梯度在马尔可夫势博弈中的全局收敛。arXiv 预印本 arXiv:2106.01969，2021年。Michael L Littman。作为多智能体强化学习框架的马尔可夫博弈。载于《机器学习论文集 1994》，第 157–163 页。爱思唯尔，1994年。Qinghua Liu、Tiancheng Yu、Yu Bai 和 Chi Jin。基于自我博弈的强化学习模型方法的精细分析。载于《国际机器学习会议》，第 7001–7010 页。PMLR，2021年。Qinghua Liu、Yuanhao Wang 和 Chi Jin。针对对抗性对手学习马尔可夫博弈：高效算法与基本极限。arXiv 预印本 arXiv:2203.06803，2022年。Weichao Mao 和 Tamer Bas¸ar。在去中心化一般和马尔可夫博弈中实现可证明高效的强化学习。《动态博弈与应用》，第 1–22 页，2022年。Weichao Mao、Lin Yang、Kaiqing Zhang 和 Tamer Basar。关于改进去中心化多智能体强化学习的无模型算法。载于《国际机器学习会议》，第 15007–15049 页。PMLR，2022年。Gergely Neu。无需再探索：改进非随机赌博机的高概率遗憾界。《神经信息处理系统进展》，28，2015年。Gergely Neu 和 Julia Olkhovskaya。用于对抗性线性上下文赌博机的高效且稳健的算法。载于《学习理论会议》，第 3049–3068 页。PMLR，2020年。Chengzhuo Ni、Yuda Song、Xuezhou Zhang、Chi Jin 和 Mengdi Wang。一般和低秩马尔可夫博弈的表征学习。arXiv 预印本 arXiv:2210.16976，2022年。OpenAI。OpenAI Five。https://blog.openai.com/openai-five/，2018年。

Francesco Orabona。《在线学习现代导论》。arXiv 预印本 arXiv:1912.13213，2019年。

Aviad Rubinstein。确定计算近似双人纳什均衡的复杂度。ACM SIGecom Exchanges，15(2):45–49，2017年。Daniel Russo 和 Benjamin Van Roy。Eluder 维度与乐观探索的样本复杂度。神经信息处理系统进展，26，2013年。Shai Shalev-Shwartz、Shaked Shammah 和 Amnon Shashua。面向自动驾驶的安全多智能体强化学习。arXiv 预印本 arXiv:1610.03295，2016年。Lloyd S Shapley。随机博弈。美国国家科学院院刊，39(10):1095–1100，1953年。Aaron Sidford、Mengdi Wang、Lin Yang 和 Yinyu Ye。以近乎最优的时间和样本复杂度求解折扣随机双人博弈。发表于国际人工智能与统计学会议，第2992–3002页。PMLR，2020年。David Silver、Aja Huang、Chris J Maddison、Arthur Guez、Laurent Sifre、George Van Den Driessche、Julian Schrittwieser、Ioannis Antonoglou、Veda Panneershelvam、Marc Lanctot 等。借助深度神经网络和树搜索精通围棋。《自然》，529(7587):484–489，2016年。David Silver、Julian Schrittwieser、Karen Simonyan、Ioannis Antonoglou、Aja Huang、Arthur Guez、Thomas Hubert、Lucas Baker、Matthew Lai、Adrian Bolton 等。不借助人类知识精通围棋。《自然》，550 (7676):354–359，2017年。Ziang Song、Song Mei 和 Yu Bai。何时能够以样本高效的方式学习参与者众多的一般和马尔可夫博弈？arXiv 预印本 arXiv:2110.04184，2021年。Ziang Song、Song Mei 和 Yu Bai。扩展式博弈中相关均衡的样本高效学习。arXiv 预印本 arXiv:2205.07223，2022年。Gilles Stoltz。个体序列预测中的不完全信息与内部遗憾。博士论文，Universit´e Paris Sud-Paris XI，2005年。Richard S Sutton 和 Andrew G Barto。强化学习：导论。麻省理工学院出版社，2018年。Yi Tian、Yuanhao Wang、Tiancheng Yu 和 Suvrit Sra。未知马尔可夫博弈中的在线学习。arXiv 预印本 arXiv:2010.15020，2021年。Roman Vershynin。高维概率：数据科学中的应用导论，第47卷。剑桥大学出版社，2018年。Oriol Vinyals、Igor Babuschkin、Wojciech M Czarnecki、Michael Mathieu、Andrew Dudzik、Junyoung Chung、David H Choi、Richard Powell、Timo Ewalds、Petko Georgiev 等。运用多智能体强化学习达到《星际争霸II》宗师级水平。《自然》，575(7782):350–354，2019年。Ruosong Wang、Russ R Salakhutdinov 和 Lin Yang。采用一般价值函数逼近的强化学习：通过有界 Eluder 维度实现可证明高效的方法。神经信息处理系统进展，33:6123–6135，2020年。Chen-Yu Wei、Yi-Te Hong 和 Chi-Jen Lu。随机博弈中的在线强化学习。发表于神经信息处理系统进展，第4987–4997页，2017年。Peter R Wurman、Samuel Barrett、Kenta Kawamoto、James MacGlashan、Kaushik Subramanian、Thomas J Walsh、Roberto Capobianco、Alisa Devlic、Franziska Eckert、Florian Fuchs 等。运用深度强化学习超越《跑车浪漫旅》冠军车手。《自然》，602(7896):223–228，2022年。Qiaomin Xie、Yudong Chen、Zhaoran Wang 和 Zhuoran Yang。利用函数逼近和相关均衡学习零和同时行动马尔可夫博弈。发表于学习理论会议，第3674–3682页。PMLR，2020年。

Wei Xiong、Han Zhong、Chengshuai Shi、Cong Shen 和 Tong Zhang。零和马尔可夫博弈的自博弈后验采样算法。载于《国际机器学习会议》，第 24496–24523 页。PMLR，2022年。Andrea Zanette 和 Martin Wainwright。利用线性架构稳定 Q 学习，实现可证明的高效学习。载于《国际机器学习会议》，第 25920–25954 页。PMLR，2022年。Andrea Zanette、Alessandro Lazaric、Mykel Kochenderfer 和 Emma Brunskill。学习固有贝尔曼误差较低的近最优策略。载于《国际机器学习会议》，第 10978–10989 页。PMLR，2020年。Wenhao Zhan、Jason D Lee 和 Zhuoran Yang。去中心化乐观超策略镜像下降：在马尔可夫博弈中实现可证明的无遗憾学习。arXiv 预印本 arXiv:2206.01588，2022年。Kaiqing Zhang、Zhuoran Yang 和 Tamer Bas¸ar。策略优化可证明地收敛到零和线性二次博弈中的纳什均衡。《神经信息处理系统进展》，第 32 卷，2019年。Kaiqing Zhang、Sham M Kakade、Tamer Bas¸ar 和 Lin F Yang。具有近最优样本复杂度的零和马尔可夫博弈中的基于模型的多智能体强化学习。arXiv 预印本 arXiv:2007.07461，2020 年。Kaiqing Zhang 、 Zhuoran Yang 和 Tamer Bas¸ar。多智能体强化学习：理论与算法概述。《强化学习与控制手册》，第 321–384 页，2021a。Runyu Zhang、Zhaolin Ren 和 Na Li。随机博弈中的梯度博弈：驻点、收敛与样本复杂度。arXiv 预印本 arXiv:2106.00198，2021b。Runyu Zhang、Qinghua Liu、Huan Wang、Caiming Xiong、Na Li 和 Yu Bai。马尔可夫博弈的策略优化：统一框架与更快收敛。arXiv 预印本 arXiv:2206.02640，2022年。Dongruo Zhou、Quanquan Gu 和 Csaba Szepesvari。线性混合马尔可夫决策过程的近乎极小极大最优强化学习。载于《学习理论会议》，第 4532–4576 页。PMLR，2021年。

A 技术工具

A.1 集中不等式

以下 Freedman 不等式可见于（Agarwal等人，2014年，引理 9）。

引理 13（Freedman 不等式）。设随机变量 $\{X_t\}_{t=1}^T$ 是一个鞅差序列，即 $X_t \in \mathcal{F}_t$，其中 $\{\mathcal{F}_t\}_{t\geq 1}$ 是一个滤过，且 $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$。假设几乎必然有 $X_t \leq R$，其中某个（非随机）$R > 0$。那么对于任意 $\lambda \in (0, 1/R]$，以至少 $1-\delta$ 的概率，有

$$\sum_{t=1}^{T} X_t \leq \lambda \cdot \sum_{t=1}^{T} \mathbb{E}\left[X_t^2|\mathcal{F}_{t-1}\right] + \frac{\log(1/\delta)}{\lambda}.$$

A.2 Eluder 与 Bellman-Eluder维数

我们首先给出函数类的标准 Eluder 维度定义（Russo 和 Van Roy，2013年；Wang 等人，2020年）。

定义 14（Eluder 维度）。对于任意函数类 $\mathcal{F} \subset (\mathcal{X} \to \mathbb{R})$，其 Eluder 维度 $d_{\mathrm{E}}(\mathcal{F}, \varepsilon)$ 定义为满足以下条件的最长序列 $\{x_1, x_2, \ldots, x_n\} \subset \mathcal{D}$ 的长度：存在 $\varepsilon' \geq \varepsilon$，使得对于所有 $i \in [n]$，x_i 相对于其前缀序列 $\{x_1, \ldots, x_{i-1}\}$ 是 ε'-独立的，即存在某个 $f_i, g_i \in \mathcal{F}$，使得

$$\sqrt{\sum_{j=1}^{i-1} [(f_i - g_i)(x_j)]^2} \leq \varepsilon' \quad \text{but} \quad |(f_i - g_i)(x_i)| \geq \varepsilon'.$$

定义 15（分布式Eluder维数）。对于任意函数类 $\mathcal{F} \subset (\mathcal{X} \to \mathbb{R})$，其相对于分布类 $\Pi \subset \Delta(\mathcal{X})$ 和 $\varepsilon > 0$ 的分布式Eluder维数 $d_{\mathrm{E}}(\mathcal{F}, \mathcal{D}, \varepsilon)$，定义为满足以下条件的最长序列 $\{\mu_1, \mu_2, \ldots, \mu_n\} \subset \mathcal{D}$ 的长度：存在 $\varepsilon' \geq \varepsilon$，使得对于所有 $i \in [n]$，μ_i 都与其前缀序列 $\{\mu_1, \ldots, \mu_{i-1}\}$ 具有 ε'-独立性，即存在某个 $f_i \in \mathcal{F}$，使得

$$\sqrt{\sum_{j=1}^{i-1} \left(\mathbb{E}_{X\sim\mu_j}[f_i(X)]\right)^2} \leq \varepsilon' \quad \text{but} \quad |\mathbb{E}_{X\sim\mu_i}[f_i(X)]| \geq \varepsilon'.$$

对于去中心化多智能体强化学习，我们采用以下贝尔曼-Eluder维数的定义。该定义类似于Jin 等人(2021a)针对玩家 i 面对固定马尔可夫对手时的单智能体马尔可夫决策过程 π_{-i}提出的原始定义，但这里我们考虑相对于所有策略 $\pi_i \in \Pi_i$ 的贝尔曼算子，而非贝尔曼最优算子。

定义 16（Bellman-Eluder维数）。对于任意玩家 $i \in [m]$、第 i 个玩家的任意马尔可夫策略类 Π_i、除第 i 个玩家外所有玩家的任意马尔可夫策略 π_{-i}，以及任意 $\varepsilon > 0$，定义

$$d_{\pi_{-i}}^{\mathrm{BE}}(\mathcal{F}_i, \Pi_i, \varepsilon) := \min_{\mathcal{D}\in\{\mathcal{D}_{\Pi_i\times\pi_{-i}}, \mathcal{D}_\Delta\}} \max_{h\in[H]} d_{\mathrm{E}}\Big(\Big\{f_h - \mathcal{T}_{i,h}^{\pi_i\times\pi_{-i}} f_{h+1} : (f, \pi_i) \in \mathcal{F} \times \Pi_i\Big\}, \mathcal{D}, \varepsilon\Big),$$

其中 $d_E(\cdot, \cdot, \varepsilon)$ 表示分布式Eluder维数（定义 *15*），并且

$$\mathcal{D}_{\Pi_i\times\pi_{-i}} := \left\{d_h^{\pi_i\times\pi_{-i}}(\cdot,\cdot) : \pi_i \in \Pi_i\right\} \subset \Delta(\mathcal{S}\times\mathcal{A}_i),$$
$$\mathcal{D}_\Delta := \left\{\delta_{(s,a_i)} : (s, a_i) \in \mathcal{S}\times\mathcal{A}_i\right\} \subset \Delta(\mathcal{S}\times\mathcal{A}_i),$$

其中，$d_h^{\pi_i\times\pi_{-i}}(\cdot,\cdot) \in \Delta(\mathcal{S}\times\mathcal{A}_i)$ 表示在博弈中采用策略 $\pi_i \times \pi_{-i}$ 时 $(s_h, a_{i,\ h})$ 的分布，而$\delta_{(s,a_i)} \in \Delta(\mathcal{S}\times\mathcal{A}_i)$ 表示位于 (s, a_i) 的点质量。

B 关于 V 型函数逼近的讨论

我们的元算法VLPR 和 AVLPR 及其保证可直接扩展到 V 型函数逼近。事实上，在元算法层面（算法1-3），VLPR 和 AVLPR 严格来说并不要求 $\mathcal{F}_i$是边际 Q 函数类——如果 $\{\mathcal{F}_i\}_{i\in[m]}$ 改为 V 类，它们也可直接照常应用，只要子程序 NO-REGRET-ALG 和 OPTIMISTIC-REGRESS 能够设计出奖励函数，使条件(1A)-(1C) （以及条件17）仍能得到满足 $\{G_{i,h}\}_{(i,h)\in[m]\times[H]}$。

然而，我们要指出的是，在具体实例化时，V 型函数逼近可能涵盖与 Q 型函数逼近具有相当不同结构的问题。例如，可以设想将第4.1 节中的线性函数逼近结果推广到线性 V 类。V 类的一种合理选择是 $\mathcal{F}_{i,\,h}\subset\{f_{i,h}(\cdot)=\phi_i(\cdot)^\top\theta_{i,\,h}:\,\theta_{i,\,h}\in\mathbb{R}^{d_i}\}$，其中 $\phi_i:\,\mathcal{S}\to\mathbb{R}^{d_i}$ 是针对状态的特征映射。在这种情况下，策略类可以合理地选择为如下形式的线性策略：$\pi_{i,\,h}(\cdot|s)=\arg\max_{a_i\in\mathcal{A}_i}\phi_{i,\,h}(s)^\top\theta_{i,\,h}^{a_i}$ 其中 $\{\theta_{i,h}^{a_i}\}_{a_i\in\mathcal{A}_i}\subset\mathbb{R}^{d_i}$是一个向量集合。然而，这样的策略类可以被理解为要求任意动作 $a_i\in\mathcal{A}_i$ 在所有状态下“具有相同的含义”，与 Q 型特征映射 $\phi_i(s,a_i)$ 相比，这可能相当不自然；后者允许 $a_i\in\mathcal{A}_i$ 作为一个通用的动作索引，而该索引在不同状态下可能表示不同的含义。

C 第 3 节的证明和补充细节3

C.1 定理 2的证明

根据贝尔曼最优性方程，对于所有 (t,i,h,s)，我们有 $\in[T]\times[m]\times[H]\times\mathcal{S}$

$$\max_{\mu_{i,h}\in\Delta(\mathcal{A}_i)}\mathbb{D}_{\mu_{i,h}\times\pi^t_{-i,h}}\left[r_{i,h}+\mathbb{P}_{h+1}V^{\dagger,\pi^t_{-i}}_{i,h+1}\right](s)=V^{\dagger,\pi^t_{-i}}_{i,h}(s). \tag{6}$$

另一方面，利用条件 (1A) 和条件 (1B)中的第一个不等式，我们有，以概率$1-2TH\delta$，对于所有 $(t,\ i,\ h,\ s)\in[T]\times[m]\times[H]\times\mathcal{S}$

$$\begin{aligned}&\max_{\mu_{i,h}\in\Delta(\mathcal{A}_i)}\mathbb{D}_{\mu_{i,h}\times\pi^t_{-i,h}}\left[r_{i,h}+\mathbb{P}_{h+1}\overline{V}^t_{i,h+1}\right](s)\\&\le\mathbb{D}_{\pi^t_h}\left[r_{i,h}+\mathbb{P}_{h+1}\overline{V}^t_{i,h+1}\right](s)+G_{i,h}(s,\overline{\pi}^{t-1},t-1,\delta)\le\overline{V}^t_{i,h}(s).\end{aligned} \tag{7}$$

因此，利用上述两个关系进行反向归纳，对于所有 (t,i,h,s)，我们有 $\in[T]\times[m]\times[H]\times\mathcal{S}$

$$\overline{V}^t_{i,h}(s)\ge V^{\dagger,\pi^t_{-i}}_{i,h}(s). \tag{8}$$

同样，利用条件 (1B)中的第二个不等式，通过反向归纳，我们可以证明，对于所有 $(t,\ i,\ h,\ s)\in[T]\times[m]\times[H]\times\mathcal{S}$

$$\overline{V}^t_{i,h}(s)\le V^{\pi^t}_{i,h}(s)+2\sum_{h'=h}^{H}\mathbb{E}_{\pi^t}\left[G_{i,h'}(s_{h'},\overline{\pi}^{t-1},t-1,\delta)\right]. \tag{9}$$

因此，我们可以将 CCE 遗憾上界估计为

$$\sum_{t=1}^{T}[V^{\dagger,\pi^t_{-i}}_{i,1}-V^{\pi^t}_{i,1}]\le 2\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[G_{i,h}(s_h,\overline{\pi}^{t-1},t-1,\delta)\right]\le\widetilde{\mathcal{O}}\left(\sqrt{H^2LT}\right),$$

其中，最后一个不等式由条件 (1C)得出。

最后，利用马尔可夫不等式以及对 . 的选择，可以界定输出策略 π^{out} $T=$ 的 CCEGap。$\widetilde{\mathcal{O}}\left(H^2L/\varepsilon^2\right)$ ■

C.2 加速算法

算法 5 带策略重放的加速 V 学习（AVLPR）

1：将 π^1 初始化为均匀策略：$\pi^1_{i,h}(\cdot|s) \leftarrow \mathrm{Unif}(\mathcal{A}_i)$，对所有 (i,s,h)，$\mathcal{B}_h^0 \leftarrow \emptyset, I_1 \leftarrow 0$。
2：对于迭代 $t = 1, \ldots, T$，执行 3：执行 π^t 以采样一个回合，并更新 $\mathcal{B}_h^t = \mathcal{B}_h^{t-1} \bigcup \{s_h\}$。
4：如果 $\exists (i,h) \in [m] \times [H]$ 满足 $\Psi_{i,h}(\mathcal{B}_h^t) \geq \Psi_{i,h}(\mathcal{B}_h^{I_t}) + 1$ 或 $t = 1$，则 5：将回放策略设为 $\overline{\pi}^t \leftarrow \mathrm{Unif}(\{\pi^\tau\}_{\tau \in [t]})$，并令 $\overline{V}_i^{t+1}{}_{,H+1} \leftarrow 0$。 6：对于 $h = H, \ldots, 1$，执行 7：计算 $\pi_h^{t+1} \leftarrow \text{CCE-APPROX}(\overline{\pi}^t, \{\overline{V}_{i,h+1}^{t+1}\}_{i \in [m]}, t)$。 8：计算 $\{\overline{V}_{i,h}^{t+1}\}_{i \in [m]} \leftarrow \text{V-APPROX}(\overline{\pi}^t, \{\overline{V}_{i,h+1}^{t+1}\}_{i \in [m]}, \pi_h^{t+1}, \{\mathcal{F}_i\}_{i \in [m]}, t)$。 9：设置 $I_{t+1} \leftarrow t$ 10：否则 11：设置 $I_{t+1} \leftarrow I_t$ 和 $\pi^{t+1} \leftarrow \pi^t$。输出：从 $\{\pi^t\}_{t \in [T]}$ 中均匀随机采样得到的 π^{out}。

算法 我们在 算法 中提出了加速算法 AVLPR。 AVLPR 的主要新要素是低频更新机制：该算法仅在满足某个 π^{t+1} 触发条件时执行策略重放并学习新策略（第 4 行），此时学习过程与 VLPR 中的相同。否则，它只执行当前策略 π^t 一个回合，将状态 s_h 添加到数据集 $\mathcal{B}_h^t$中，并设置 $\pi^{t+1} \leftarrow \pi^t$ （第 3 行）。

直观地说，触发条件要求自上次重放迭代 $I_t < t$ 以来，数据集已积累了足够多的新数据。这一设计受到一种倍增技巧式观察的启发：由 π^t 诱导的状态访问（因而样本复杂度）不会因 π^t 是否更新而产生显著差异，直到发现某个汇总统计量（例如表格情形下任一状态的访问次数）自上次重放以来增加到至少两倍（或任意常数倍 > 1）。我们用 $\Psi_{i,\ h}(\cdot)$ 表示此类汇总统计量的对数，因此只有当 $\Psi_{i,\ h}(\mathcal{B}_h^t) \geq \Psi_{i,\ h}(\mathcal{B}_h^{I_t}) + 1$ 时才会触发新的重放。

条件与保证 具体而言，AVLPR 要求满足以下附加条件，以确保低频更新机制有效。直观来说，该条件要求奖励函数在连续两次策略更新之间的增幅至多为一个常数倍。

条件 17 （低频策略更新的有效性）。 算法 5 中的触发准则 $\{\Psi_{i,h}\}_{(i,h) \in [m] \times [H]}$ 满足以下条件：

（a）以至少 $1-\delta$ 的概率，对于所有 (t,i,h)，若 $\Psi_{i,h}(\mathcal{B}_h^t) \leq \Psi_{i,h}(\mathcal{B}_h^{I_t}) + 1$，则必有 $G_{i,h}(s, \overline{\pi}^{I_t}, I_t, \delta) \leq 8 \times G_{i,h}(s, \overline{\pi}^t, t, \delta)$，对所有 $s \in \mathcal{S}$ 均成立；（b）以概率 1，在 T 次迭代内，算法 5 中触发的重放次数（即第 4 行）至多为 $d_{\mathrm{replay}} \log T$ 次迭代，其中 $d_{\mathrm{replay}} > 0$ 为某个常数。

现在我们陈述关于 AVLPR 的元保证；证明见附录 C.3。

定理 18 （AVLPR 的元保证）。 假设算法 *5* 中的子程序可以进行实例化，使得条件 *(1A)-(1C)& 17* 在使用相同的奖励函数 $\{G_{i,h}\}_{(i,h) \in [m] \times [H]}$ 和部署的触发函数$\{\Psi_{i,h}\}_{(i,h) \in [m] \times [H]}$时以至少 $1-\delta$ 的概率成立，则有

$$\mathrm{CCEReg}(T) := \max_{i \in [m]} \sum_{t=1}^{T} \left[V_{i,1}^{\dagger, \pi_{-i}^t}(s_1) - V_{i,1}^{\pi^t}(s_1) \right]_+ \leq \widetilde{\mathcal{O}}(\sqrt{H^2 LT}).$$

作为推论，选择 $T = \widetilde{\mathcal{O}}(H^2 L/\varepsilon^2)$ 可确保算法 *1* 的输出策略 π^{out} 满足 $\mathrm{CCEGap}(\pi^{\mathrm{out}}) \leq \varepsilon$，并且以至少[10] 0.99 的概率，所进行的回合总数至多为（其中 $\overline{\Gamma} := \max_{\overline{\pi},\ \pi'} |\Gamma_{\mathrm{explore}}(\overline{\pi}, \pi')|$）

$$\mathcal{O}\big(T + HT \times d_{\mathrm{replay}} \log T \times \overline{\Gamma}\big) = \widetilde{\mathcal{O}}\big(H^3 L \overline{\Gamma} \cdot d_{\mathrm{replay}}/\varepsilon^2\big).$$

[10]通过与定理 $1-\delta$ 2.，可以通过类似论证将成功概率进一步提升至任意值

C.3 定理18 的证明

令 $\mathcal{I}$ 表示 $[T]$ 中触发第 4 行的子集。根据条件 17， $|\mathcal{I}| \leq d_{\text{replay}} \log \qquad T$.

根据贝尔曼最优性方程，对于所有 (t, i, h, s)，我们有 $\in [T] \times [m] \times [H] \times \mathcal{S}$

$$\max_{\mu_{i,h} \in \Delta(\mathcal{A}_i)} \mathbb{D}_{\mu_{i,h} \times \pi^t_{-i,h}} \left[r_{i,h} + \mathbb{P}_{h+1} V^{\dagger, \pi^t_{-i}}_{i,h+1} \right](s) = V^{\dagger, \pi^t_{-i}}_{i,h}(s). \tag{10}$$

另一方面，根据条件(1A) 和条件(1B)中的第一个不等式，我们以概率$1 - 2TH\delta$得到，对于所有 $(t, \ i, \ h, \ s) \in \mathcal{I} \times [m] \times [H] \times \mathcal{S}$

$$\begin{aligned} &\max_{\mu_{i,h} \in \Delta(\mathcal{A}_i)} \mathbb{D}_{\mu_{i,h} \times \pi^{t+1}_{-i,h}} \left[r_{i,h} + \mathbb{P}_{h+1} \overline{V}^{t+1}_{i,h+1} \right](s) \\ &\leq \mathbb{D}_{\pi^{t+1}_h} \left[r_{i,h} + \mathbb{P}_{h+1} \overline{V}^{t+1}_{i,h+1} \right](s) + G_{i,h}(s, \overline{\pi}^t, t, \delta) \leq \overline{V}^{t+1}_{i,h}(s). \end{aligned} \tag{11}$$

因此，根据上述两个关系进行反向归纳，对于所有 (t, i, h, s)，我们有 $\in \mathcal{I} \times [m] \times [H] \times \mathcal{S}$

$$\overline{V}^{t+1}_{i,h}(s) \geq V^{\dagger, \pi^{t+1}_{-i}}_{i,h}(s), \tag{12}$$

这意味着，对于所有所有 (t, i, h, s)，我们有 $\in [T] \times [m] \times [H] \times \mathcal{S}$:

$$\overline{V}^{I_t+1}_{i,h}(s) \geq V^{\dagger, \pi^{I_t+1}_{-i}}_{i,h}(s).$$

类似地，根据条件(1B)中的第二个不等式进行反向归纳，我们可以证明，对于所有 $(t, \ i, \ h, \ s) \in \mathcal{I} \times [m] \times [H] \times \mathcal{S}$

$$\overline{V}^{t+1}_{i,h}(s) \leq V^{\pi^{t+1}}_{i,h}(s) + 2 \sum_{h'=h}^{H} \mathbb{E}_{\pi^{t+1}} \left[G_{i,h'}(s_{h'}, \overline{\pi}^t, t, \delta) \right],$$

这意味着，对于所有 $(t, i, h, s) \ \in [T] \times [m] \times [H] \times \mathcal{S}$:

$$\overline{V}^{I_t+1}_{i,h}(s) \leq V^{\pi^{I_t+1}}_{i,h}(s) + 2 \sum_{h'=h}^{H} \mathbb{E}_{\pi^{I_t+1}} \left[G_{i,h'}(s_{h'}, \overline{\pi}^{I_t}, I_t, \delta) \right].$$

因此，我们可以将 CCE 遗憾上界为

$$\begin{aligned} \sum_{t=1}^{T} [V^{\dagger, \pi^t_{-i}}_{i,1} - V^{\pi^t}_{i,1}] &\overset{(i)}{=} \sum_{t=1}^{T} [V^{\dagger, \pi^{I_t+1}_{-i}}_{i,1} - V^{\pi^{I_t+1}}_{i,1}] \\ &\leq 2 \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{\pi^{I_t+1}} \left[G_{i,h}(s_h, \overline{\pi}^{I_t}, I_t, \delta) \right] \\ &\overset{(ii)}{\leq} 16 \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{\pi^{I_t+1}} \left[G_{i,h}(s_h, \overline{\pi}^{t-1}, t-1, \delta) \right] \\ &\overset{(iii)}{=} 16 \sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{\pi^t} \left[G_{i,h}(s_h, \overline{\pi}^{t-1}, t-1, \delta) \right] \\ &\overset{(iv)}{\leq} \widetilde{\mathcal{O}} \left(\sqrt{H^2 LT} \right), \end{aligned}$$

其中，(i) 和 (iii) 利用了以下事实： $\pi^t = \pi^{I_t+1}$，(iv) 则由条件(1C)得出，而 (ii) 由引理19得出。

最后，利用马尔可夫不等式以及 的选取，可以界定输出策略 π^{out} 的 CCEGap。总样本复杂度可界定为 $T =$ $\widetilde{\mathcal{O}}\left(H^2L/\varepsilon^2\right)$

$$T + |\mathcal{I}| \times H \times \left(\underbrace{\mathcal{O}\left(\overline{\Gamma} \cdot T\right)}_{\text{cost of CCE-APPROX}} + \underbrace{\mathcal{O}\left(\overline{\Gamma} \cdot T\right)}_{\text{cost of V-APPROX}} \right) = \widetilde{\mathcal{O}}\left(H^3 L \overline{\Gamma} d_{\text{replay}}/\varepsilon^2\right).$$

引理 19。 假设条件 *17*成立，则以至少 $1-\delta$ 的概率，对所有 (t,i,h)，$G_{i,\ h}(s,\overline{\pi}^{I_t},I_t,\delta)\ \leq 8\times G_{i,\ h}(s_h,\overline{\pi}^{t-1},t-1,\delta)$ 对所有 $s\in\mathcal{S}$ 均成立。

证明。 如果第 4 行在第 $(t-1)^{\text{th}}$ 次迭代中被触发，则 $I_t = t-1$，结论成立。否则， $I_t = I_{t-1}$ 且对所有 (i,h)，$\Psi_{i,\ h}(\mathcal{B}_h^{t-1}) \leq \Psi_{i,\ h}(\mathcal{B}_h^{I_{t-1}})\ +1$，由条件 17可知

$$G_{i,h}(s,\overline{\pi}^{I_t},I_t,\delta) = G_{i,h}(s,\overline{\pi}^{I_{t-1}},I_{t-1},\delta) \leq 8\times G_{i,h}(s_h,\overline{\pi}^{t-1},t-1,\delta)$$

对所有 $s\in\mathcal{S}$均成立。 ■

C.4 VLPR 和 AVLPR 的去中心化执行协议

本节首先介绍我们的协议，然后说明在这些协议下，VLPR 和 AVLPR 均可实现去中心化（AVLPR 只需最小通信）。

我们考虑以下协议：在博弈开始之前，参与者对一个随机比特序列进行采样，其长度与所进行的回合数呈多项式关系，并且所有参与者都能观察到这个（共享的）随机比特序列。利用该序列，参与者便可以通过去中心化方式实现共享随机性。例如，执行 Unif($\{\pi^\tau\}_{\tau\in[T]}$)（其中每个 π^τ 都是乘积策略）可以通过使用共享随机比特（按照相同的预先确定协议）对共享的 $\tau\sim$Unif(T) 进行采样，然后执行 $\pi^\tau = \pi_1^\tau\times\cdots\times\pi_m^\tau$ 来完成；这可以通过完全去中心化的方式实现。

我们进一步假设，探索策略映射 $\Gamma_{\text{explore}}(\overline{\pi},\pi')$（回顾一下，它是由元组 $(\widetilde{\pi},P)$ 构成的有序集合）在以下意义上是边缘可执行的：所有参与者都知道其排序；对于有序排列中的每个 $(\widetilde{\pi},P)\ \in\Gamma_{\text{explore}}(\overline{\pi},\pi')$，所有参与者都知道 P；只要 $\overline{\pi}_i$ 和 π_i' 这两个边缘策略（以共享随机比特为条件）为第 i 个玩家所知，那么边缘策略 $\widetilde{\pi}_i$（以共享随机比特为条件）也为第 i 个玩家所知。

我们指出，对于常见的 Γ_{explore}取值，该假设均成立，例如我们在表格情形和线性情形中的具体设定。特别地，我们的表格设置选择 $\Gamma_{\text{explore}}(\overline{\pi},\ \pi')\ =\ [(\overline{\pi}_{1:h-1}\circ\pi_h',\ [m])]$，这直接满足边缘可执行性。对于线性设置，回顾式（15），我们选择了

$$\Gamma_{\text{explore}}(\overline{\pi},\pi') = \left[(\widetilde{\pi},P) = \left(\overline{\pi}_{1:h-1}\times(\text{Unif}(\mathcal{A}_i)\times\pi'_{-i,h}),\{i\}\right)\right]_{i=1}^m.$$

让所有参与者知晓并遵守 P的日程很简单（只需按字典序对 $\{i\}$ 中的 $i\in[m]$进行轮转）。此外，该列表中每个 $\widetilde{\pi}$ 的边缘策略 $\widetilde{\pi}_i$ 完全由 $\overline{\pi}_i$ 和{中的一个Unif($\mathcal{A}_i$) π_i'} 决定（取决于 $i\in P$是否成立），这验证了边缘可执行性假设。

VLPR 注意到，对于算法1-3中描述的 VLPR 算法，大多数步骤（如 NO-REGRET-ALG 和 OPTIMISTIC-REGRESS）本质上都是去中心化的，每个玩家都可以独立执行。唯一涉及的协调是执行回放策略 $\overline{\pi}^t$ =Unif($\{\pi^\tau\}_{\tau\in[T]}$)（算法2中的第1 行），或执行算法2 中的探索策略 $(\widetilde{\pi},P)\ \in\ \Gamma_{\text{explore}}(\overline{\pi}^t,\mu_h^k)$，以及算法3中的探索策略 $(\widetilde{\pi},P)\ \in\ \Gamma_{\text{explore}}(\overline{\pi}^t,\pi_h^{t+1})$。使用上述共享随机性即可执行 $\overline{\pi}^t$。此外，由于 $\mu_{i,h}^k$ 和 $\pi_{i\ h}^{t+1}$，对第 i 个玩家而言是已知的，并且满足边缘可执行性假设，因此所有探索策略都可以通过去中心化方式执行。这就验证了关于 VLPR 的论断。

AVLPR 与 VLPR 相比，AVLPR 唯一的区别是在算法 5 的第 4 行中检查触发条件；在每次迭代中，$t \in [T]$ 这需要通信 m 比特，每位玩家各占一比特（用于指示玩家 $i \in [m]$是否满足该条件）。只要至少有一位玩家满足触发条件，参与者就会进入重放部分；否则就开始下一回合。由于所有参与者都知道自己在每次迭代中是否进入了重放部分，因此重放索引 I_t 是所有参与者的共同知识，可以由所有参与者同时维护。此外，我们可以只在满足触发条件时才触发这次通信；根据条件 17，在 $\mathcal{O}(d_{\text{replay}} \log T)$ 次游戏迭代中，这种情况最多发生 T 次。

D 全策略完备性意味着“本质上是表格型”的博弈

在这里，我们通过说明不受限制的 全策略完备性 假设对博弈施加了强烈的隐含要求，论证假设 4 中对线性 argmax 策略的限制（或其他类型的限制）是必要的。

考虑以下针对去中心化线性函数逼近的全策略完备性假设。该假设通过移除 Π^{lin} 限制，强化了假设 4 。

假设 20（全策略完备性）。对于任意 $(i,h) \in [m] \times [H]$、任意函数 $\overline{V} = \overline{V}_{i,\,h+1} \colon \mathcal{S} \to [0,H]$ 以及任意策略 π，存在 $\theta^{h,\,\pi_{-i},\,\overline{V}} \in \mathbb{R}^d$，满足 $\|\theta^{h,\,\pi_{-i},\,\overline{V}}\|_2 \le B_\theta$，使得

$$\mathbb{D}_{\delta_{a_i} \times \pi_{-i}}\left[r_{i,h} + \mathbb{P}_h \overline{V}_{i,h+1}\right](s) = \phi_i(s,a_i)^\top \theta^{h,\pi_{-i},\overline{V}} \quad \textit{for all } (s,a_i) \in \mathcal{S} \times \mathcal{A}_i. \tag{13}$$

固定任意 $(h,s^\star) \in [H] \times \mathcal{S}$，再固定任意玩家 $i \in [m]$ 和 $s' \in \mathcal{S}$。令 π^1 和 π^2 为任意两个仅在 $(h,s^\star)$ 处不同的联合策略。将奖励设为零（即 $r_{i,\,h} = 0$），并取函数 $\overline{V}(\cdot) = 1\{s' = \cdot\}$，应用假设 20，可得存在 $\theta^{\pi^{\{1,2\}}}$，使得对于所有 $(s,a_i) \in \mathcal{S} \times \mathcal{A}_i$,

$$\begin{aligned}
\phi_i(s,a_i)^\top \theta^{\pi^1} &= \mathbb{E}_{a_{-i} \sim \pi^1_{-i}(\cdot|a)} \Pr[s'|s,a_i,a_{-i}], \\
\phi_i(s,a_i)^\top \theta^{\pi^2} &= \mathbb{E}_{a_{-i} \sim \pi^2_{-i}(\cdot|a)} \Pr[s'|s,a_i,a_{-i}].
\end{aligned}$$

对于任意 $s \ne s^\star$，在 (h,s) 处的 $\pi^1 = \pi^2$，我们对每个 $s \ne s^\star \in \mathcal{S}$ 和 $a_i \in \mathcal{A}_i$ 都有

$$\phi_i(s,a_i)^\top \left(\theta^{\pi^1} - \theta^{\pi^2}\right) = 0,$$

并且对于 $s = s^\star$，有

$$\phi_i(s^\star,a_i)^\top \left(\theta^{\pi^1} - \theta^{\pi^2}\right) = \mathbb{E}_{a_{-i} \sim \pi^1_{-i}(\cdot|a)} \Pr[s'|s^\star,a_i,a_{-i}] - \mathbb{E}_{a_{-i} \sim \pi^2_{-i}(\cdot|a)} \Pr[s'|s^\star,a_i,a_{-i}].$$

我们称状态 $(s,h) \in \mathcal{S} \times [H]$ 无关紧要，如果某些参与者的动作能够影响该状态的状态转移。根据定义，如果一个状态是 $s^\star$ 相关的，则存在 $\exists i \in [m]$、$s' \in \mathcal{S}$、$a_i \in \mathcal{A}_i$ 和 π^1、π^2，使得

$$\phi_i(s^\star,a_i)^\top \left(\theta^{\pi^1} - \theta^{\pi^2}\right) = \mathbb{E}_{a_{-i} \sim \pi^1_{-i}(\cdot|a)} \Pr[s'|s^\star,a_i,a_{-i}] - \mathbb{E}_{a_{-i} \sim \pi^2_{-i}(\cdot|a)} \Pr[s'|s^\star,a_i,a_{-i}] \ne 0.$$

由此可知：(1) $v \coloneqq \theta^{\pi^1} - \theta^{\pi^2} \ne 0$；(2) 对所有其他 $s \ne s^\star$ 和 $a_i' \in \mathcal{A}_i$，v 与 $\phi_i(s,a_i')$ 正交；(3) $\phi_i(s^\star,a_i)$ 与 v 不正交，因此与 $\{\phi_i(s,a_i')$ 线性无关，$a_i' \in \mathcal{A}_i\}$。由于特征 $\phi_i(s,a_i) \in \mathbb{R}^d$，至多有 d 个此类特征向量与其他所有特征向量线性无关，因此玩家 i 至多有 d 个相关状态。

由此可知，除至多 dm 个状态外，其他所有状态都无关紧要：这些状态下的转移概率不取决于参与者的联合动作。如果我们直接从轨迹中省略这些状态（访问这些状态时采取任意策略），那么由此产生的动态过程将构成一个状态数较少（至多 dm 个状态）的马尔可夫博弈动态过程。从这个意义上说，这种马尔可夫博弈可称为“本质上是表格型的”。

D.1 Cui 等人 Cui 等人 (2023年)

如果将Cui 等人（2023年）中的无悔学习预言机选为指数权重算法，那么它将导出如下形式的策略类: $\Pi^{\text{estimate}} = \Pi_1^{\text{estimate}} \times \cdots \times \Pi_m^{\text{estimate}}$其中

$$\Pi_i^{\text{estimate}} := \left\{ \pi_i(\cdot \mid s) \propto \exp\left(\eta \sum_{i=1}^{K} \left[\phi_i(s,\cdot)^\top \theta^k + \beta \|\phi_i(s,\cdot)\|_{\Sigma^{-1}} \right]_{[0,H]} \right) : \theta^k \in \mathbb{R}^d,\ \Sigma \in \mathbb{R}^{d\times d},\ \Sigma \succeq \lambda I \right\},$$

其中， $[\cdot]_{[0,H]}$ 表示一个截断算子，使得 $[x]_{[0,H]} = \min\{\max\{x,0\}, H\}$，而 η、λ、β 是其算法中的一些可调参数。请注意，线性 argmax 策略可以用单个 d维向量参数化，而上述策略类中的策略则由数量多得多的参数指定（K 个不同的 d维向量、一个 $d \times d$矩阵以及若干额外的标量），并且涉及 K 次截断，使得指数可能高度非线性。从这个意义上说，上述策略类比本文考虑的线性 argmax 策略类 Π^{lin} 更复杂。我们进一步指出，若移除截断算子，选取 $\beta = 0$，并令上述定义中的 η 趋于无穷，则 Π_i^{estimate}会退化为 Π_i^{lin}。

如果选择期望扰动领袖跟随算法作为无悔学习预言机，那么我们将得到 $\Pi^{\text{estimate}} = \Pi_1^{\text{estimate}} \times \cdots \times \Pi_m^{\text{estimate}}$，其中

$$\Pi_i^{\text{estimate}} := \left\{ \pi_i(a_i \mid s) = \mathbb{P}_{v\sim D_i}\left[a_i \in \arg\max_{\widehat{a}_i} \left(\sum_{i=1}^{K} \left[\phi_i(s,\widehat{a}_i)^\top \theta^k + \beta \|\phi_i(s,\widehat{a}_i)\|_{\Sigma^{-1}} \right]_{[0,H]} + \eta^{-1} v_{\widehat{a}_i} \right) \right] \right.$$
$$\left. : \theta^k \in \mathbb{R}^d,\ \Sigma \in \mathbb{R}^{d\times d},\ \Sigma \succeq \lambda I \right\},$$

其中，向量 $v \in \mathbb{R}^{A_i}$ 从 $\mathbb{R}^{A_i}$ 上的某个分布 D_i 中采样，$v_{\widehat{a}_i}$ 表示 v的第 $\widehat{a}_i$个坐标。与上文的论证类似，这个 Π_i^{estimate}仍然比 Π_i^{lin}更复杂。通过移除截断算子、选择 $\beta = 0$ 并取 $D = \delta_{\vec{0}}$（即点 0 处的狄拉克分布），可以再次将其简化为 $\Pi_{i\to}^{\text{lin}}$。

第 E 节的证明4.1

E.1 线性 AVLPR 算法的详细内容

理解假设4 假设4 对线性函数逼近情形的设计与分析有如下推论，并将在全文中使用。

注 1. 假设 4 蕴含以下结论。对于任意 $(i,h) \in [m]\times[H]$、任意函数 $\overline{V} = \overline{V}_{i,\ h+1}: \mathcal{S} \to [0,H]$ 以及任意策略 $\pi \in \Pi^{\text{lin}}$，存在 $\theta^{h,\ \pi_{-i},\ \overline{V}} \in \mathbb{R}^d$ 使得

$$\mathbb{D}_{\delta_{a_i}\times\pi_{-i,h}}\left[r_{i,h} + \mathbb{P}_h \overline{V}_{i,h+1} \right](s) = \phi_i(s,a_i)^\top \theta^{h,\pi_{-i},\overline{V}} \quad \textit{for all } (s,a_i) \in \mathcal{S}\times\mathcal{A}_i. \tag{14}$$

通过选取 $f_{i,\ h+1}(s,\ \ a_i) = \overline{V}(s)$，并应用假设 4。

选择 $\Psi_{i,\ h}$ 算法5 中的切换条件选为

$$\Psi_{i,h}(\mathcal{B}) := \log\det\left(I + \frac{1}{A_i} \sum_{s\in\mathcal{B}} \sum_{a_i\in\mathcal{A}_i} \phi_i(s,a_i)\phi_i(s,a_i)^\top \right).$$

正在处理 $\mathcal{D}_{\text{init}}$ 和 $\mathcal{D}_{\text{sample}}$ 对于线性函数逼近，随后将使用数据集 $\mathcal{D}_{\text{init}}^i$ 计算 H^特征协方差矩阵 $\{\Sigma_{i,h}^{\overline{\pi}}\}_{h\in[H]}$，用于衡量探索策略 $\overline{\pi}$的覆盖度，定义如下:

$$\widehat{\Sigma}_{i,h}^{\overline{\pi}} := \frac{1}{|\mathcal{D}_{\text{init}}^i| \cdot A_i} \sum_{s_h\in\mathcal{D}_{\text{init}}^i} \sum_{a_i\in\mathcal{A}_i} \phi_i(s_h,a_i)\phi_i(s_h,a_i)^\top.$$

此外，我们定义总体版本

$$\Sigma_{i,h}^{\overline{\pi}} = \mathbb{E}_{s_h \sim \overline{\pi}} \mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{A}_i)} \left[\phi_i(s_h, a_i) \phi_i(s_h, a_i)^\top \right].$$

对于线性函数逼近，我们在算法 $\Gamma_{\mathrm{explore}}(\overline{\pi}, \mu_h)$ 2 和 3 中选择的探索方案为有序集合

$$[(\overline{\pi}_{1:h-1} \circ (\mathrm{Unif}(\mathcal{A}_1) \times \mu_{-1,h}), \{1\}), \cdots, (\overline{\pi}_{1:h-1} \circ (\mathrm{Unif}(\mathcal{A}_m) \times \mu_{-m,h}), \{m\})]. \tag{15}$$

因此， $\mathcal{D}^{k\,i}$ 样本将恰好包含一个元素，记为 $(s_h^k, a_{i,h}^k, y_{i,h}^k)$。

No-Regret-Alg 条件 (1A) 可理解为关于损失函数 $\ell_{i,h}^k(s, a_i) := \mathbb{D}_{\delta_{a_i} \times \mu_{-i,h}^k}[r_{i,h} + \mathbb{P}_{h+1} \overline{V}_{i,h+1}](s)$ 的逐状态遗憾界。根据假设 4， $\ell_{i,h}^k$ 可写为线性函数 $\ell_{i,h}^k(s, a_i) = \langle \theta_{i,h}^k, \phi_i(s, a_i) \rangle$。为了保证逐状态遗憾，我们首先为所有 $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$ 构造 $\ell_{i,h}^k$ 的线性估计量：

$$\widehat{\ell}_{i,h}^k(s, a_i) = \langle \widehat{\theta}_{i,h}^k, \phi_i(s, a_i) \rangle,$$

其中

$$\widehat{\theta}_{i,h}^k := \left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1} \phi_i(s_{i,h}^k, a_{i,h}^k) y_{i,h}^k.$$

该估计量也用于对抗性线性老虎机（Neu 和 Olkhovskaya，2020年）。然而，在我们的设定中，直接使用该估计量运行指数权重算法行不通，因为假设 4 要求 μ_h^k 属于（.....的凸包） Π^{lin}；否则，在 μ_{-i}^k 下，所得的动作价值函数无法用线性函数逼近。为此，我们首先注意到，逐状态老虎机遗憾（比较对象位于 $\Delta(\mathcal{A}_i)$ 中）可以等价地视为一个在线线性优化问题的遗憾（比较对象位于动作特征向量的凸包中）

$$\begin{aligned}
&\max_{\mu_{i,h} \in \Delta_{\mathcal{A}_i}} \sum_{k=1}^K \left(\mathbb{D}_{\mu_{i,h} \times \mu_{-i,h}^k} - \mathbb{D}_{\mu_h^k} \right) (r_{i,h} + \mathbb{P}_h \overline{V}_{i,h+1})(s) \\
&= \max_{\mu_{i,h} \in \Delta_{\mathcal{A}_i}} \sum_{k=1}^K \langle \mu_{i,h} - \mu_{i,h}^k(\cdot \mid s), \ell_{i,h}^k(s, \cdot) \rangle = \max_{\phi \in CH(\Phi_i(s))} \sum_{k=1}^K \langle \phi - \Phi_i(s)^\top \mu_{i,h}^k, \theta_{i,h}^k \rangle.
\end{aligned}$$

这里， $\Phi_i(s) \in \mathbb{R}^{A_i \times d}$ 是一个矩阵，其中堆叠了所有特征向量 $\{\phi_i(s, \cdot)\}$；我们稍微滥用一下记号，用 $CH(\cdot)$ 表示该矩阵各行的凸包。

接下来，我们将把期望扰动跟随领先者算法（Hazan 和 Minasyan (2020年，算法3）；另见 Hazan 等人 (2016年））应用于在线线性优化问题，即选择

$$\Phi_i(s)^\top \mu_{i,h}^k = \mathbb{E}_{v \sim \mathcal{V}} \left[\operatorname*{arg\,max}_{\phi \in CH(\Phi_i(s))} \langle \phi, \sum_{k' \le k} \widehat{\theta}_{i,h}^{k'} + v/\eta \rangle \right],$$

其中， $\mathcal{V}$ 取为椭圆 $\{u \mid u^\top (\widehat{\Sigma}_{i,\ h}^{\overline{\pi}} + \lambda I) u \le 1\}$ 上的均匀分布，而 η 是一个作用类似于学习率的参数。这将导出如下策略

$$\mu_{i,h}^{k+1}(a_i | s) := \Pr_{v \sim \mathcal{V}} \left[a_i = \operatorname*{arg\,max}_{a_i' \in \mathcal{A}_i} \langle \phi_i(s, a_i'), \sum_{k' \le k} \widehat{\theta}_{i,h}^{k'} + v/\eta \rangle \right], \tag{16}$$

它位于 Π^{lin} 的凸包中，因此满足假设4的要求 .

O乐观-R回归 乐观回归通过在数据集 $\mathcal{D}^i_{\text{reg}}$ 上使用岭回归实现，该数据集包含 $(s_h, a_{i,h}, y_{i,h})$ 的样本，其中 $y_{i,h} = r_{i,h} + \overline{V}_{i,h+1}(s_{h+1})$。具体而言，

$$\widehat{\theta}_{i,h} \leftarrow \arg\min_{\theta} \frac{1}{K} \sum_{(s_h, a_{i,h}, y_{i,h}) \in \mathcal{D}^i_{\text{reg}}} \left[\phi_i(s_h, a_{i,h})^\top \theta - y_{i,h}\right]^2 + \lambda \|\theta\|_2^2,$$

$$\overline{Q}_{i,h}(s, a_i) \leftarrow \left(\phi_i(s, a_i)^\top \widehat{\theta}_{i,h} + \frac{3}{2} G_{i,h}(s, \overline{\pi}, K, \delta)\right) \wedge (H - h + 1),$$

$$\overline{V}_{i,h}(s) \leftarrow \langle \pi_{i,h}(\cdot \mid s), \overline{Q}_{i,h}(s, \cdot) \rangle.$$

计算效率我们在此指出，$\overline{V}_{i,h}(s)$ 无需针对每个 s 进行计算，只需针对数据集中的状态计算，这可以在多项式时间内完成。此外，也无需完整计算式（16）中的策略，因为执行算法只需从策略$\mu_i^{k+1}{}_h$，中高效采样，而这可以通过对 $v \sim \mathcal{V}$进行采样轻松实现。

E.2 条件的证明(1A)

如第4.1节所述，我们首先将条件(1A) 中的逐状态遗憾分解为基于损失估计器测量的逐状态遗憾项和统计误差项：

$$\begin{aligned}
&\max_{\mu_{i,h} \in \Delta_{\mathcal{A}_i}} \sum_{k=1}^{K} \left(\mathbb{D}_{\mu_{i,h} \times \mu^k_{-i,h}} - \mathbb{D}_{\mu^k_h}\right) (r_{i,h} + \mathbb{P}_h \overline{V}_{i,h+1})(s) \\
&= \underbrace{\max_{\phi \in CH(\Phi_i(s))} \sum_{k=1}^{K} \langle \phi - \Phi_i(s) \mu^k_{i,h}(\cdot \mid s), \widehat{\theta}^k_{i,h} \rangle}_{(A)} + \underbrace{\max_{\phi \in CH(\Phi_i(s))} \sum_{k=1}^{K} \langle \phi, \theta^k_{i,h} - \widehat{\theta}^k_{i,h} \rangle}_{(B)} \\
&\quad + \underbrace{\sum_{k=1}^{K} \langle \Phi_i(s) \mu^k_{i,h}(\cdot \mid s), \widehat{\theta}^k_{i,h} - \theta^k_{i,h} \rangle}_{(C)}.
\end{aligned}$$

在附录F中，我们证明，在选择 $\eta = 1/(dH\sqrt{K \log \delta^{-1}})$ 和 $\lambda = \overline{\Theta(d \max_i A_i / K)}$时，上述三项可分别控制如下：以至少 $1 - \delta$的概率，对所有 $s \in \mathcal{S}$

$$\begin{aligned}
\text{Term (A)} &\le \sup_{a_i \in \mathcal{A}_i} \|\phi_i(s, a_i)\|_{(\Sigma^{\overline{\pi}}_{i,h} + \lambda I)^{-1}} \cdot \widetilde{\mathcal{O}}\left(dH\sqrt{K(\max_i A_i)}\right), \\
\text{Term (B)} &\le \sup_{a_i \in \mathcal{A}_i} \|\phi_i(s, a_i)\|_{(\Sigma^{\overline{\pi}}_{i,h} + \lambda I)^{-1}} \times \widetilde{\mathcal{O}}\left(dH\sqrt{K(\max_i A_i)}\right), \\
\text{Term (C)} &\le \sup_{a_i \in \mathcal{A}_i} \|\phi_i(s, a_i)\|_{(\Sigma^{\overline{\pi}}_{i,h} + \lambda I)^{-1}} \times \widetilde{\mathcal{O}}\left(dH\sqrt{K(\max_i A_i)^3}\right) + \mathcal{O}(1).
\end{aligned}$$

因此，我们可以选取

$$G_{i,h}(s, \overline{\pi}, K, \delta) = \sup_{a_i \in \mathcal{A}_i} \|\phi_i(s, a_i)\|_{(\Sigma^{\overline{\pi}}_{i,h} + \lambda I)^{-1}} \times \widetilde{\Theta}\left(\frac{dH A_i^{1.5}}{\sqrt{K}}\right) + \Theta\left(\frac{1}{K}\right).$$

E.3 条件的证明(1B)

考虑一个固定的 $(i, h) \in [m] \times [H]$。记 $\mathcal{D}^i_{\text{reg}}$ 为

$$\left\{(s^j_h, a^j_{i,h}, y^j_{i,h})\right\}_{j \in [K]}.$$

根据假设 4，存在 $\theta_{i,h}^*$，使得对所有 j，

$$\mathbb{E}[y_{i,h}^j|s_h^j,a_{i,h}^j]=\mathbb{D}_{\delta_{a_{i,h}^j}\times\pi_{-i,h}}[r_{i,h}+\mathbb{P}_{h+1}\overline{V}_{i,h+1}](s_h^j)=\langle\phi_i(s_h^j,a_{i,h}^j),\theta_{i,h}^*\rangle.$$

定义 $\widehat{\Sigma}_{\mathrm{reg},\ h}=\frac{1}{K}\sum_{j=1}^K\phi_i(s_h^j,a_{i,\ h}^j)\phi_i(s_h^j,a_{i,\ h}^j)^\top+\lambda I$ 和 $\zeta_j=y_{i,\ h}^j-\mathbb{P}_h\left[(\overline{V}_{i,\ h+1}+r_{i,\ h})\right]\left(s_h^j,a_{i,h}^j\right)$。这里 ζ_j 均值为零且有界于 H。由此可得 $\forall(s,a_i)\in\mathcal{S}\times\mathcal{A}_i$

$$\begin{aligned}
&\left|\phi_i(s,a_i)^\top\widehat{\theta}_{i,h}-\mathbb{D}_{\delta_{a_i}\times\pi_{-i,h}}[r_{i,h}+\mathbb{P}_{h+1}\overline{V}_{i,h+1}](s)\right|\\
=&\left|\phi_i(s,a_i)^\top\widehat{\theta}_{i,h}-\phi_i(s,a_i)^\top\theta_{i,h}^\star\right|\\
=&\left|\phi_i(s,a_i)^\top\widehat{\Sigma}_{\mathrm{reg},h}^{-1}\frac{1}{K}\sum_{j=1}^K\phi_i(s_h^j,a_{i,h}^j)\left(\phi_i(s_h^j,a_{i,h}^j)^\top\theta_{i,h}^\star+\zeta_j\right)-\phi_i(s,a_{i,h})^\top\theta_{i,h}^\star\right|\\
\leq&\|\phi_i(s,a_{i,h})\|_{\widehat{\Sigma}_{\mathrm{reg},h}^{-1}}\times\left(\left\|\frac{1}{K}\sum_{j=1}^K\phi_i(s_h^j,a_{i,h}^j)\zeta_j\right\|_{\widehat{\Sigma}_{\mathrm{reg},h}^{-1}}+\sqrt{\lambda}B_\theta\right).
\end{aligned}$$

引理21. 假设我们选取 $\lambda=\Theta(d\log(dK/\delta)/K)$，则以概率 $1-\delta$

$$\left\|\sum_{j=1}^K\phi_i(s_h^j,a_{i,h}^j)\zeta_j\right\|_{\Sigma_{\mathrm{reg},h}^{-1}}\leq\mathcal{O}\left(\sqrt{KdH^2}\log(KdH/\delta)\right).$$

该引理的证明与引理 24的证明完全相同。最后注意，根据引理 22，以概率 $1-\delta$，

$$\widehat{\Sigma}_{\mathrm{reg},h}\succcurlyeq\frac{1}{2}\Sigma_{i,h}^{\overline{\pi}}+\lambda I-\mathcal{O}\left(\frac{d\log(dK/\delta)}{K}\right)I\succcurlyeq\frac{1}{2}\left(\Sigma_{i,h}^{\overline{\pi}}+\lambda I\right).$$

因此

$$\begin{aligned}
&\left|\phi_i(s,a_i)^\top\widehat{\theta}_{i,h}-\left[\mathbb{P}_h^{\pi^t}(\overline{V}_{i,h+1}+r_{i,h})\right](s,a_i)\right|\\
&\quad\leq\|\phi_i(s,a_i)\|_{\widehat{\Sigma}_{\mathrm{reg},h}^{-1}}\cdot\mathcal{O}\left(dH\sqrt{1/K}\log(dK/\delta)\right)\\
&\quad\leq\|\phi_i(s,a_i)\|_{(\Sigma_{i,h}^{\overline{\pi}}+\lambda I)^{-1}}\cdot\mathcal{O}\left(dH\sqrt{1/K}\log(dK/\delta)\right)\\
&\quad\leq\frac{1}{2}G_{i,h}(s,\overline{\pi},K,\delta).
\end{aligned}$$

我们得出 $\forall(s,a_i)$

$$\begin{aligned}
&\left(\left[\mathbb{P}_h^{\pi^t}(\overline{V}_{i,h+1}+r_{i,h})\right](s,a)+G_{i,h}(s,\overline{\pi},K,\delta)\right)\wedge(H-h+1)\leq\overline{Q}_{i,h}(s,a)\\
\leq&\left(\mathbb{P}_h^{\pi^t}(\overline{V}_{i,h+1}+r_{i,h})(s,a)+2G_{i,h}(s,\overline{\pi},K,\delta)\right)\wedge(H-h+1).
\end{aligned}$$

由此可得

$$\begin{aligned}
\min\left\{\mathbb{D}_{\pi_h}\left[r_{i,h}+\mathbb{P}_{h+1}\overline{V}_{i,h+1}\right](s)+G_{i,h}(s,\overline{\pi},K,\delta),H-h+1\right\}\leq\overline{V}_{i,h}(s),\\
\overline{V}_{i,h}(s)\leq\mathbb{D}_{\pi_h}\left[r_{i,h}+\mathbb{P}_{h+1}\overline{V}_{i,h+1}\right](s)+2G_{i,h}(s,\overline{\pi},K,\delta).
\end{aligned}$$

E.4 条件的证明(1C)

记

$$X_t := \mathbb{E}\left[\phi_i(s_h, a_{i,h})\phi_i(s_h, a_{i,h})^\top \mid s_h \sim \pi^t_{1:h-1},\ a_{i,h} \sim \mathrm{Unif}(\mathcal{A}_i)\right], \qquad S_t := \sum_{\tau=1}^{t} X_\tau + \lambda_0 I_{d\times d},$$

其中 $\lambda_0 = \mathcal{O}(d)$。然后根据 $G_{i,h}$ 方程 (3) 中的定义，

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{\pi^{t+1}}\left[G_{i,h}(s, \overline{\pi}^t, t, \delta)\right] \\
\le& \widetilde{\mathcal{O}}\left(\frac{d(\max_i A_i)^{1.5}H}{\sqrt{t}} \cdot \sum_{t=1}^{T} \mathbb{E}\left[\sqrt{t} \max_{a_{i,h}\in\mathcal{A}_i} \|\phi(s_h, a_{i,h})\|_{S_t^{-1}} \mid s_h \sim \pi^{t+1}_{1:h-1}\right]\right) + \widetilde{\mathcal{O}}(1) \\
\le& \widetilde{\mathcal{O}}\left(d(\max_i A_i)^{2.5}H \cdot \sum_{t=1}^{T} \mathbb{E}\left[\|\phi(s_h, a_{i,h})\|_{S_t^{-1}} \mid s_h \sim \pi^{t+1}_{1:h-1}, a_{i,h} \sim \mathrm{Unif}(\mathcal{A}_i)\right]\right) + \widetilde{\mathcal{O}}(1) \\
\le& \widetilde{\mathcal{O}}\left(d(\max_i A_i)^{2.5}H \cdot \sqrt{T \cdot \sum_{t=1}^{T} \mathbb{E}\left[\|\phi(s_h, a_{i,h})\|^2_{S_t^{-1}} \mid s_h \sim \pi^{t+1}_{1:h-1}, a_{i,h} \sim \mathrm{Unif}(\mathcal{A}_i)\right]}\right) + \widetilde{\mathcal{O}}(1) \\
\le& \widetilde{\mathcal{O}}\left(d(\max_i A_i)^{2.5}H \cdot \sqrt{T \cdot \sum_{t=1}^{T} \mathbb{E}\left[\mathrm{tr}(X_{t+1}S_t^{-1})\right]}\right) + \widetilde{\mathcal{O}}(1) = \widetilde{\mathcal{O}}\left(\sqrt{d^3(\max_i A_i)^5 H^2 T}\right).
\end{aligned}$$

E.5 条件的证明17

固定 $(i,h,t) \in [m]\times[H]\times[T]$。定义 $\widehat{S}_t := I + \frac{1}{A_i}\sum_{s\in\mathcal{B}_h^t}\sum_{a_i\in\mathcal{A}_i} \phi_i(s,a_i)\phi_i(s,a_i)^\top$。于是

$$\Psi_{i,h}(\mathcal{B}_h^t) - \Psi_{i,h}(\mathcal{B}_h^{I_t}) = \log\det\left(\widehat{S}_t\widehat{S}_{I_t}^{-1}\right).$$

因此，$\Psi_{i,h}(\mathcal{B}_h^t) - \Psi_{i,h}(\mathcal{B}_h^{I_t}) \le 1$ 意味着

$$\|\widehat{S}_t^{\frac{1}{2}}\widehat{S}_{I_t}^{-1}\widehat{S}_t^{\frac{1}{2}}\|_2 \le 2,$$

这进一步意味着

$$\widehat{S}_t \preccurlyeq 2\widehat{S}_{I_t}.$$

换言之，要证明条件17，只需证明 $S_t \preccurlyeq 2S_{I_t}$ 蕴含

$$t(\Sigma^{\overline{\pi}^t}_{i,h} + \lambda_t I) \le 8I_t\left(\Sigma^{\overline{\pi}^{I_t}}_{i,h} + \lambda_{I_t} I\right),$$

其中 $\lambda_t = \Theta(d\max_i A_i/t)$。这等价于证明

$$t(\Sigma^{\overline{\pi}^t}_{i,h} + \lambda_t I) \ge 8I_t\left(\Sigma^{\overline{\pi}^{I_t}}_{i,h} + \lambda_{I_t} I\right),$$

蕴含 $S_t \succcurlyeq 2S_{I_t}$。根据引理 22，以概率 $1-\delta$，

$$\widehat{S}_t \succcurlyeq \frac{t}{2}\Sigma^{\overline{\pi}^t}_{i,h} - \widetilde{\Theta}(d)I \succcurlyeq 4I_t\Sigma^{\overline{\pi}^{I_t}}_{i,h} + \widetilde{\Theta}(d)I \succcurlyeq 2\widehat{S}_{I_t}.$$

最后，针对 i、h 和 t 应用并集界，即可证明该条件的部分（a）。

至于第二部分，我们注意到 $\Psi_{i,h}(\emptyset) = \log\det I = 1$，并且

$$\Psi_{i,h}(\mathcal{D}_h^t) \le \log\det\widehat{S}_T \le d\log\left(\|S_T\|_2\right) \le d\log T.$$

因此，切换次数总计至多为 $dmH\log T$，即部分（b）以 $d_{\mathrm{replay}} = dmH$ 得到满足。

E.6 线性函数逼近的样本复杂度

第 E.2 节至 E.3 节表明，条件 (1A) 至(1C) 均以以下数值满足：

$$G_{i,h}(s,\overline{\pi},K,\delta)=\sup_{a_i\in\mathcal{A}_i}\|\phi_i(s,a_i)\|_{(\Sigma_{i,h}^{\overline{\pi}}+\lambda I)^{-1}}\times\widetilde{\Theta}\left(\frac{dH(\max_i A_i)^{1.5}}{\sqrt{K}}\right)+\Theta\left(\frac{1}{K}\right).$$

and

$$L=\widetilde{\mathcal{O}}\left(\max_{i\in[m]}d^3(\max_i A_i)^5H^2\right).$$

最后，在 E.5 中验证了条件 17 以 $d_{\text{replay}}=dmH$ 的形式成立。由（15），$\overline{\Gamma}=m$。因此，应用定理 18，我们得到以下寻找 ε-粗相关均衡的样本复杂度上界

$$\widetilde{\mathcal{O}}\left(\frac{H^3L\overline{\Gamma}d_{\text{replay}}}{\varepsilon^2}\right)=\widetilde{\mathcal{O}}\left(\frac{d^4m^2H^6\max_{i\in[m]}A_i^5}{\varepsilon^2}\right).$$

F 附录 E.2 的证明

F.1 相对集中性

考虑以下随机过程：在时间步 t，我们根据 $\{x_\tau\}_{\tau\in[t-1]}$（随机地）从 d 维单位球上选取一个分布 D_t，然后抽取样本 $x_t\sim D_t$。记 Σ_t 为 D_t 的协方差矩阵。关于经验时间平均协方差与总体协方差在乘法意义上的接近程度，我们有如下相对集中引理。

引理 22. 以至少 $1-\delta$ 的概率，对所有 $t\in[T]$

$$\frac{1}{2}\sum_{\tau\in[t]}\Sigma_\tau-\beta I\preceq\sum_{\tau\in[t]}x_\tau x_\tau^\top\preceq 2\sum_{\tau\in[t]}\Sigma_\tau+\beta I$$

其中 $\beta=\Theta(d\log(dT/\delta))$。

证明。首先固定 $t\in[T]$。任取满足 $\|w\|_2=1$ 的 $w\in\mathbb{R}^d$。定义 $W_\tau=\langle x_\tau,w\rangle^2$。由此可得 $\mathbb{E}[W_\tau]=w^\top\Sigma_\tau w$，且 $0\le W_\tau\le 1$。根据伯恩斯坦不等式，以概率 $1-\delta'$

$$\left|\sum_{\tau\in[t]}(W_\tau-\mathbb{E}[W_\tau])\right|\le\sqrt{4\log(1/\delta')\sum_{\tau\in[t]}\mathbb{E}[W_\tau^2]}+\mathcal{O}(\log 1/\delta') \tag{17}$$

$$\le\sqrt{4\log(1/\delta')\sum_{\tau\in[t]}\mathbb{E}[W_\tau]}+\mathcal{O}(\log 1/\delta') \tag{18}$$

$$\le\frac{1}{2}\sum_{\tau\in[t]}\mathbb{E}[W_\tau]+\mathcal{O}(\log 1/\delta'). \tag{19}$$

因此，概率为 $1-\delta'$

$$\sum_{\tau\le t}W_\tau\le 2\mathbb{E}\left[\sum_{\tau\le t}W_\tau\right]+O\left(\log\left(\frac{1}{\delta'}\right)\right),$$

$$\sum_{\tau\le t}W_\tau\ge\frac{1}{2}\mathbb{E}\left[\sum_{\tau\le t}W_\tau\right]-O\left(\log\left(\frac{1}{\delta'}\right)\right)$$

接下来需要构造一个 ε'-覆盖 $W_{\varepsilon'}$，覆盖 d-球面，其中我们选择 $\varepsilon' = 0.01/T$。由此可得，概率为$1-\delta$时，对于单位球面上的所有 w，

$$\sum_{\tau\le t}\langle w, x_\tau\rangle^2 \le 2\mathbb{E}[\sum_{\tau\le t}\langle w, x_\tau\rangle^2] + \mathcal{O}(\log(|W_{\varepsilon'}|/\delta),$$
$$\sum_{\tau\le t}\langle w, x_\tau\rangle^2 \ge \frac{1}{2}\mathbb{E}[\sum_{\tau\le t}\langle w, x_\tau\rangle^2] - \mathcal{O}(\log(|W_{\varepsilon'}|/\delta).$$

这意味着

$$\frac{1}{2}\sum_{\tau\in[t]}\Sigma_\tau - \mathcal{O}(\log(|W_{\varepsilon'}|/\delta)I \preceq \sum_{\tau\in[t]} x_\tau x_\tau^\top \preceq 2\sum_{\tau\in[t]}\Sigma_\tau + \mathcal{O}(\log(|W_{\varepsilon'}|/\delta))I$$

将 δ替换为 δ/T并代入 $|W_{\varepsilon'}| \le \left(\frac{3}{\varepsilon'}\right)^d$(Vershynin，2018年，推论 4.2.13) 即可证明该引理。 ■

F.2 控制条件 (A) 中的项(1A)

为了引出对期望 FPL 的分析，我们观察到

$$\begin{aligned}\text{Term (A)} &= \max_{\mu_i\in\Delta_{\mathcal{A}_i}}\sum_{k=1}^K\langle \mu_i - \mu_{i,h}^k(\cdot\mid s), \widehat{\ell}_{i,h}^k(s,\cdot)\rangle \\ &= \max_{\mu_i\in\Delta_{\mathcal{A}_i}}\sum_{k=1}^K\langle \mu_i - \mu_{i,h}^k(\cdot\mid s), \Phi_i(s,\cdot)^\top\widehat{\theta}_{i,h}^k\rangle \\ &= \max_{x\in CH(\{\phi_i(s,\cdot)\})}\sum_{k=1}^K\langle x - \Phi_i(s,\cdot)\mu_{i,h}^k(\cdot|s), \widehat{\theta}_{i,h}^k\rangle.\end{aligned}$$

注意，在我们的算法中，

$$\mu_{i,h}^k(a_i|s) := \Pr_{v\sim\mathcal{V}}\left[a_i = \arg\max\left\langle \phi_i(s,\cdot), \sum_{k'<k}\widehat{\theta}_{i,h}^{k'} + \frac{1}{\eta}(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I)^{-1/2}v\right\rangle\right],$$

这意味着

$$\Phi_i(s,\cdot)\mu_{i,h}^k(\cdot|s) = \mathbb{E}_{v\sim\mathcal{V}}\operatorname*{arg\,max}_{x\in CH(\{\phi_i(s,\cdot)\})}\left\langle (\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I)^{-1/2}x, (\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I)^{1/2}\sum_{k'<k}\widehat{\theta}_{i,h}^{k'} + \frac{1}{\eta}v\right\rangle.$$

这与期望 Follow-the-Perturbed-Leader 算法完全相同（参见例如(Hazan 等人，2016年，算法 17）用于一系列线性损失向量

$$(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I)^{1/2}\widehat{\theta}_{i,h}^1, \cdots, (\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I)^{1/2}\widehat{\theta}_{i,h}^K.$$

因此，根据期望 FPL 的遗憾界（Hazan 和 Minasyan，2020年，定理 10），选择 $\mathcal{V}$为 d维单位球上的均匀分布，

$$\text{Term(A)} \le \sup_{a_i\in\mathcal{A}_i}\|\phi_i(s,a_i)\|_{(\widehat{\Sigma}_{i,h}^{\overline{\pi}}+\lambda I)^{-1}}\cdot\left[\frac{1}{\eta} + \eta d\sum_{k=1}^K\|\widehat{\theta}_{i,h}^k\|^2_{(\widehat{\Sigma}_{i,h}^{\overline{\pi}}+\lambda I)}\right].$$

根据引理 23，其概率至少为 $1-\delta$

$$\sum_{k=1}^K\|\widehat{\theta}_{i,h}^k\|^2_{\widehat{\Sigma}_{i,h}^{\overline{\pi}}+\lambda I} = \mathcal{O}\left(dK + \frac{\log\delta^{-1}}{\lambda}\right).$$

因此　此外，将 $\eta = 1/(dH\sqrt{(\max_i A_i)K\log\delta^{-1}})$ 和 $\lambda = \widetilde{\Theta}(d(\max_i A_i)/K)$ 代入，可得　e

$$\begin{aligned}\mathrm{Term(A)} &\le \sup_{a\in\mathcal{A}_i} \|\phi_i(s,a_i)\|_{(\widehat{\Sigma}^{\overline{\pi}}_{i,h}+\lambda I)^{-1}} \cdot \mathcal{O}\left(dH\sqrt{K(\max_i A_i)\log\delta^{-1}}\right)\\ &= \sup_{a\in\mathcal{A}_i} \|\phi_i(s,a_i)\|_{(\Sigma^{\overline{\pi}}_{i,h}+\lambda I)^{-1}} \cdot \mathcal{O}\left(dH\sqrt{K(\max_i A_i)\log\delta^{-1}}\right),\end{aligned}$$

其中等式由引理22得出。

引理 23。在概率 $1-\delta$ 下，

$$\sum_{k=1}^{K} \|\widehat{\theta}^k_{i,h}\|^2_{(\widehat{\Sigma}^{\overline{\pi}}_{i,h}+\lambda I)^{-1}} = \mathcal{O}\left(dKH^2 + \frac{H^2\log\delta^{-1}}{\lambda}\right).$$

证明。定义

$$x_k := \phi_i(s^k_{i,h}, a^k_{i,h}), \quad z_k := x_k^\top(\widehat{\Sigma}^{\overline{\pi}}_{i,h}+\lambda I)^{-1}x_k.$$

根据定义

$$\sum_{k=1}^{K} \|\widehat{\theta}^k_{i,h}\|^2_{(\widehat{\Sigma}^{\overline{\pi}}_{i,h}+\lambda I)^{-1}} \le H^2\sum_{k=1}^{K} z_k.$$

此外，$\{z_k\}_{k\in[K]}$ 是满足以下条件的独立同分布样本

$$\mathbb{E}[z_k] = \mathrm{Tr}\left(\Sigma^{\overline{\pi}}_{i,h}\left(\widehat{\Sigma}^{\overline{\pi}}_{i,h}+\lambda I\right)^{-1}\right) \le \mathcal{O}(d),$$

其中不等式由引理 22 以及 λ的选取可得，并且

$$\mathbb{E}[z_k^2] \le \frac{1}{\lambda}\cdot\mathbb{E}\left[z_k\right] \le \mathcal{O}\left(\frac{d}{\lambda}\right),$$

and

$$|z_k| \le \frac{1}{\lambda}.$$

因此，根据伯恩斯坦不等式，以高概率有

$$\sum_k z_k \le \mathcal{O}\left(dK + \sqrt{\frac{dK\log\delta^{-1}}{\lambda}} + \frac{\log\delta^{-1}}{\lambda}\right) = \mathcal{O}\left(dK + \frac{\log\delta^{-1}}{\lambda}\right).$$

■

F.3 控制条件 (1A) 中的项（B）

考虑一个固定玩家 $i\in[m]$ 和第 $h\in[H]$ 步。为简化记号，记

$$x_k := \phi_i(s^k_{i,h}, a^k_{i,h}), \quad y_k := r^k_{i,h} + V_{i,h+1}(s^k_{h+1}), \quad \zeta_k := y_k - x_k^\top\theta^k_{i,h}.$$

对于任意 $(s,a_i)\in\mathcal{S}\times\mathcal{A}_i$:

$$\begin{aligned}&\sum_{k=1}^{K}\langle\phi_i(s,a_i), \widehat{\theta}^k_{i,h} - \theta^k_{i,h}\rangle\\ =&\left\langle\phi_i(s,a_i), \sum_{k=1}^{K}\left(\widehat{\theta}^k_{i,h} - \theta^k_{i,h}\right)\right\rangle\end{aligned}$$

$$
\begin{aligned}
&= \left\langle \phi_i(s,a_i), \sum_{k=1}^{K} \left(\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1} x_k y_k - \theta_{i,h}^{k} \right) \right\rangle \\
&= \phi_i(s,a)^\top \left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1} \left[\sum_{k=1}^{K} x_k \left(x_k^\top \theta_{i,h}^{k} + \zeta_k \right) - \left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right) \sum_{k=1}^{K} \theta_{i,h}^{k} \right] \\
&\le \|\phi_i(s,a_i)\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1}} \Bigg[\sqrt{\lambda} B_\theta K + \underbrace{\left\| \sum_{k=1}^{K} x_k \zeta_k \right\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1}}}_{\text{Term (B1)}} \\
&\quad + \underbrace{\left\| \sum_{k=1}^{K} \left(x_k x_k^\top - \widehat{\Sigma}_{i,h}^{\overline{\pi}} \right) \theta_{i,h}^{k} \right\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1}}}_{\text{Term (B2)}} \Bigg].
\end{aligned}
$$

~根据引理 24，25，选择 $\lambda = \mathcal{O}(d(\max_i A_i)/K)$ 以及相对集中性（引理 22），

$$
\text{Term (B)} \le \widetilde{\mathcal{O}} \left(\|\phi_i(s,a_i)\|_{\left(\Sigma_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1}} \times Hd \sqrt{K(\max_i A_i)} \right).
$$

引理24（项（B1））。以至少 $1-\delta$ 的概率，我们有

$$
\left\| \sum_{k=1}^{K} x_k \zeta_k \right\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1}} = \mathcal{O} \left(\sqrt{KdH^2 \log(KdH/\delta)} + \frac{dH \log(KdH/\delta)}{\sqrt{\lambda}} \right).
$$

证明。考虑一个固定的 $v \in \mathbb{R}^d$，满足 $\|v\|_2 = 1$。定义

$$
z_k := v^\top \left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1/2} x_k \zeta_k.
$$

注意，$\{z_k\}_{k=1}^K$ 是一个鞅，其条件方差和值域有界于

$$
|z_k| \le H \lambda^{-1/2},
$$

and

$$
\begin{aligned}
\text{Var}(z_k \mid z_{1:k-1}) &= \mathbb{E} \left[\zeta_k^2 v^\top \left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1/2} x_k x_k^\top \left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1/2} v \right] \\
&\le H^2 \left\| \left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1/2} \Sigma_{i,h}^{\overline{\pi}} \left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1/2} \right\|_2 \le \mathcal{O}(H^2),
\end{aligned}
$$

^，其中第二个不等式用到了引理22、$\Sigma_i^{\overline{\pi}}$ 的定义，h 以及 λ 的取值。

根据 Freedman 不等式，

$$
\left| \sum_{k=1}^{K} z_k \right| \le \mathcal{O} \left(\sqrt{KH^2 \log \delta^{-1}} + \frac{H \log \delta^{-1}}{\sqrt{\lambda}} \right).
$$

最后，通过取一个 对于来自 v 的所有离子界，取自 d维单位球的($\sqrt{\lambda}/(HK)$)-覆盖，我们得出 t

$$
\left\| \sum_{k=1}^{K} x_k \zeta_k \right\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1}} = \max_{v:\ \|v\|_2 = 1} \left| v^\top \left(\widehat{\Sigma}_{i,h}^{\overline{\pi}} + \lambda I \right)^{-1/2} \sum_{k=1}^{K} x_k \zeta_k \right|
$$

$$\leq \mathcal{O}\left(\sqrt{KdH^2\log(KdH/\delta)} + \frac{dH\log(KdH/\delta)}{\sqrt{\lambda}}\right).$$

■

引理25（项（B2））。以至少 $1-\delta$的概率，有

$$\left\|\sum_{k=1}^{K}\left(x_k x_k^\top - \widehat{\Sigma}_{i,h}^{\overline{\pi}}\right)\theta_{i,h}^k\right\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1}} = \mathcal{O}\left(\sqrt{KdB_\theta^2\log(KdB_\theta/\delta)} + \frac{dB_\theta\log(KdB_\theta/\delta)}{\sqrt{\lambda}}\right).$$

证明。 根据三角不等式和相对集中性（引理 22），我们有

$$\begin{aligned}
&\left\|\sum_{k=1}^{K}\left(x_k x_k^\top - \widehat{\Sigma}_{i,h}^{\overline{\pi}}\right)\theta_{i,h}^k\right\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1}} \\
\leq &\left\|\sum_{k=1}^{K}\left(x_k x_k^\top - \Sigma_{i,h}^{\overline{\pi}}\right)\theta_{i,h}^k\right\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1}} + \left\|\left(\Sigma_{i,h}^{\overline{\pi}} - \widehat{\Sigma}_{i,h}^{\overline{\pi}}\right)\sum_{k=1}^{K}\theta_{i,h}^k\right\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1}} \\
\leq &\mathcal{O}\left(\left\|\sum_{k=1}^{K}\left(x_k x_k^\top - \Sigma_{i,h}^{\overline{\pi}}\right)\theta_{i,h}^k\right\|_{\left(\Sigma_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1}} + \left\|\left(\Sigma_{i,h}^{\overline{\pi}} - \widehat{\Sigma}_{i,h}^{\overline{\pi}}\right)\sum_{k=1}^{K}\theta_{i,h}^k\right\|_{\left(\Sigma_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1}}\right).
\end{aligned}$$

考虑任意的 $v\in\mathbb{R}^d$，满足 $\|v\|_2=1$。定义

$$z_k := v^\top\left(\Sigma_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1/2}\left(x_k x_k^\top - \Sigma_{i,h}^{\overline{\pi}}\right)\theta_{i,h}^k.$$

注意，$\{z_k\}_{k=1}^K$ 是一个鞅，其条件方差和值域有界于

$$|z_k| \leq B_\theta\lambda^{-1/2},$$

and

$$\begin{aligned}
\text{Var}(z_k \mid z_{1:k-1}) &\leq \mathbb{E}\left[\left(v^\top\left(\Sigma_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1/2}x_k x_k^\top\theta_{i,h}^k\right)^2\right] \\
&\leq B_\theta^2\mathbb{E}\left[\left(v^\top\left(\Sigma_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1/2}x_k\right)^2\right] \\
&= B_\theta^2\mathbb{E}\left[v^\top\left(\Sigma_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1/2}x_k x_k^\top\left(\Sigma_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1/2}v\right] \\
&= B_\theta^2 v^\top\left(\Sigma_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1/2}(\Sigma_{i,h}^{\overline{\pi}})\left(\Sigma_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1/2}v \leq \mathcal{O}(B_\theta^2),
\end{aligned}$$

其中，第二个等式利用了这一事实：$\mathbb{E}[x_k x_k^\top] = \Sigma_{i,\ h}^{\overline{\pi}}$ 最后一个不等式利用了引理 22。根据 Freedman 不等式，

$$\left|\sum_{k=1}^{K} z_k\right| \leq \mathcal{O}\left(\sqrt{KB_\theta^2\log\delta^{-1}} + \frac{B_\theta\log\delta^{-1}}{\sqrt{\lambda}}\right).$$

最后，对 d维单位球的（$\sqrt{\lambda}/(B_\theta K)$）-覆盖中的所有 v 取并集界，我们可以得出

$$\left\|\sum_{k=1}^{K}\left(x_k x_k^\top - \Sigma_{i,h}^{\overline{\pi}}\right)\theta_{i,h}^k\right\|_{\left(\Sigma_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1}}$$

$$\leq \mathcal{O}\left(\sqrt{KdB_\theta^2 \log(KdB_\theta/\delta)} + \frac{dB_\theta \log(KdB_\theta/\delta)}{\sqrt{\lambda}}\right).$$

现在回顾一下，$\widehat{\Sigma}^{\overline{\pi}}_{i,\ h}$是通过使用从$\overline{\pi}^t$中独立同分布采样的 K 个样本进行估计的，因此我们可以简单地重复上述集中不等式论证，以控制 $\left\|\sum_{k=1}^K \left(x_k x_k^\top - \Sigma^{\overline{\pi}}_{i,h}\right)\theta^k_{i,\ h}\right\|_{(\Sigma^{\overline{\pi}}_{i,\ h}+\lambda I)^{-1}}$，从而得到 $\left\|\left(\Sigma^{\overline{\pi}}_{i,h} - \widehat{\Sigma}^{\overline{\pi}}_{i,h}\right)\sum_{k=1}^K \theta^k_{i,\ h}\right\|_{(\Sigma^{\overline{\pi}}_{i,\ h}+\lambda I)^{-1}}$的上界，其结果与上面的界相同。■

F.4 条件 (1A)中的项（C）控制

考虑固定玩家 $i \in [m]$ 和第 $h \in [H]$ 步。为简化记号，记作

$$x_k := \phi_i(s^k_{i,h}, a^k_{i,h}), \quad y_k := r^k_{i,h} + V_{i,h+1}(s^k_{h+1}), \quad \zeta_k := y_k - x_k^\top \theta^k_{i,h}.$$

我们有如下误差分解，与控制项（B）时的分解类似：对于任意 $s \in \mathcal{S}$,

$$\begin{aligned}
\text{Term (C)} &= \sum_{k=1}^K \langle \Phi_s \mu^k_{i,h}(\cdot \mid s), \theta^k_{i,h} - \widehat{\theta}^k_{i,h}\rangle \\
&= \sum_{a_i \in \mathcal{A}_i} \phi_i(s,a_i)^\top \left(\sum_{k=1}^K \mu^k_{i,h}(a_i \mid s)\theta^k_{i,h} - \sum_{k=1}^K \mu^k_{i,h}(a_i \mid s)\widehat{\theta}^k_{i,h}\right) \\
&= \sum_{a_i \in \mathcal{A}_i} \phi_i(s,a_i)^\top \left(\widehat{\Sigma}^{\overline{\pi}}_{i,h} + \lambda I\right)^{-1} \Bigg[\left(\widehat{\Sigma}^{\overline{\pi}}_{i,h} + \lambda I\right)\sum_{k=1}^K \mu^k_{i,h}(a_i \mid s)\theta^k_{i,h} \\
&\qquad - \sum_{k=1}^K \mu^k_{i,h}(a_i \mid s) x_k \left(x_k^\top \theta^k_{i,h} + \zeta_k\right)\Bigg] \qquad (20) \\
&\leq \sum_{a_i \in \mathcal{A}_i} \|\phi_i(s,a_i)\|_{\left(\widehat{\Sigma}^{\overline{\pi}}_{i,h}+\lambda I\right)^{-1}} \Bigg[\sqrt{\lambda} B_\theta K + \underbrace{\left\|\sum_{k=1}^K \mu^k_{i,h}(a_i \mid s) x_k \zeta_k\right\|_{\left(\widehat{\Sigma}^{\overline{\pi}}_{i,h}+\lambda I\right)^{-1}}}_{(C1)} \\
&\quad + \underbrace{\left\|\sum_{k=1}^K \mu^k_{i,h}(a_i \mid s)\left(x_k x_k^\top - \widehat{\Sigma}^{\overline{\pi}}_{i,h}\right)\theta^k_{i,h}\right\|_{\left(\widehat{\Sigma}^{\overline{\pi}}_{i,h}+\lambda I\right)^{-1}}}_{(C2)}\Bigg].
\end{aligned}$$

不难验证，用于界定项（B1）和项（B2）的相同论证，分别也可用于界定项（C1）和项（C2）。形式化地，我们有以下引理 24 和 25 的对应结论，用于界定项（C1）和项（C2）。

引理 26（项（C1））。考虑固定的状态与动作对（$s,\ a_i$）$\in \mathcal{S} \times \mathcal{A}_i$ 以及一个单位向量 $v \in \mathbb{R}^d$。以至少 $1-\delta$ 的概率，有

$$\begin{aligned}
&\left|v^\top \left(\widehat{\Sigma}^{\overline{\pi}}_{i,h} + \lambda I\right)^{-1/2} \sum_{k=1}^K \mu^k_{i,h}(a_i \mid s) x_k \zeta_k\right| \\
=&\mathcal{O}\left(\sqrt{KH^2 \log(1/\delta)} + \frac{H\log(1/\delta)}{\sqrt{\lambda}}\right).
\end{aligned}$$

引理 27（项（C2））。考虑一组固定的状态与动作 $(s,a_i) \in \mathcal{S} \times \mathcal{A}_i$ 以及一个单位向量 $v \in \mathbb{R}^d$。以至少 $1-\delta$ 的概率，我们有

$$\left|v^\top \left(\widehat{\Sigma}^{\overline{\pi}}_{i,h} + \lambda I\right)^{-1/2} \sum_{k=1}^K \mu^k_{i,h}(a_i \mid s)\left(x_k x_k^\top - \widehat{\Sigma}^{\overline{\pi}}_{i,h}\right)\theta^k_{i,h}\right|$$

$$=\mathcal{O}\left(\sqrt{KB_\theta^2\log(1/\delta)}+\frac{B_\theta\log(1/\delta)}{\sqrt{\lambda}}\right).$$

引理 26 和 27 的证明与引理 24 和 25 的前半部分（并集界之前）几乎完全相同，因此这里省略。

利用引理 26 和 27控制项（C）时，我们需要对所有状态与动作 $(s,a_i)\in\mathcal{S}\times\mathcal{A}_i$以及单位向量 $v\in\mathbb{R}^d$应用并集界。以下引理基本上说明，这样的并集界只会使上界额外增加一个 $\widetilde{}$因子 $\mathcal{O}(dA_i)$。

引理 28. 考虑一个策略 π，定义如下

$$\pi_{i,h}(a_i|s):=\Pr_{v\sim\mathcal{V}}\left[a_i=\arg\max\langle\phi_i(s,\cdot),w+Wv\rangle\right],$$

其中 $w\in\mathbb{R}^d$ 使得 $\|w\|_2\le\gamma$，$\alpha I_{d\times d}\preceq W\preceq\beta I_{d\times d}$，$\mathcal{V}$ 表示 d 维单位球上的均匀分布。那么，对于满足 $\max_{a_i\in\mathcal{A}_i}\|\phi_i(s,a_i)-\phi_i(s',a_i)\|_2\le\varepsilon$ 的任意状态 $s,s'\in\mathcal{S}$，我们有

$$\left\|\mathbb{E}_{a\sim\pi_{i,h}(\cdot|s)}[\phi_i(s,a_i)]-\mathbb{E}_{a\sim\pi_{i,h}(\cdot|s')}[\phi_i(s',a_i)]\right\|_2=\widetilde{\mathcal{O}}\left(\frac{d\beta\gamma\sqrt{\varepsilon}}{\alpha^2}\right).$$

引理 28 的证明推迟到本小节末尾。

根据标准的离散化论证，存在 $\mathcal{S}$ 的一个子集 $\mathcal{S}_\varepsilon$（即相对于度量 $d(s,s')=\max_{a_i\in\mathcal{A}_i}\|\phi_i(s,a_i)-\phi_i(s',a_i)\|_2$，对 $\mathcal{S}$ 的离散覆盖），使得

- 对于任意 $s\in\mathcal{S}$，都存在满足条件的 $s'\in\mathcal{S}_\varepsilon$

$$\max_{a_i\in\mathcal{A}_i}\|\phi_i(s,a_i)-\phi_i(s',a_i)\|_2\le\frac{1}{\mathrm{poly}(B_\theta,K,d,H,A_i,\lambda^{-1},\eta^{-1},\delta^{-1})},$$

- 且

$$\log|\mathcal{S}_\varepsilon|\le\widetilde{\mathcal{O}}(dA_i).$$

对于所有 $s\in\mathcal{S}$：记 s' 为 s 在 $\mathcal{S}_\varepsilon$ 中相对于度量 $d(s,s')=\max_{a_i\in\mathcal{A}_i}\|\phi_i(s,a_i)-\phi_i(s',a_i)\|_2$ 的最近邻，

$$\begin{aligned}
&\sum_{k=1}^{K}\langle\Phi_s\mu_{i,h}^k(\cdot\mid s),\widehat{\theta}_{i,h}^k-\theta_{i,h}^k\rangle\\
&\overset{(i)}{\le}\sum_{k=1}^{K}\langle\Phi_{s'}\mu_{i,h}^k(\cdot\mid s'),\widehat{\theta}_{i,h}^k-\theta_{i,h}^k\rangle+1\\
&\overset{(ii)}{\le}\sum_{a_i\in\mathcal{A}_i}\|\phi_i(s',a_i)\|_{\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1}}\Bigg[\sqrt{\lambda}B_\theta K\\
&\quad+\max_{v:\,\|v\|_2=1}\left|v^\top\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1/2}\sum_{k=1}^{K}\mu_{i,h}^k(a_i\mid s')\left(x_kx_k^\top-\widehat{\Sigma}_{i,h}^{\overline{\pi}}\right)\theta_{i,h}^k\right|\\
&\quad+\max_{v:\,\|v\|_2=1}\left|v^\top\left(\widehat{\Sigma}_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1/2}\sum_{k=1}^{K}\mu_{i,h}^k(a_i\mid s')\left(x_kx_k^\top-\widehat{\Sigma}_{i,h}^{\overline{\pi}}\right)\theta_{i,h}^k\right|\Bigg]+1\\
&\overset{(iii)}{\le}\max_{a_i\in\mathcal{A}_i}\|\phi_i(s',a_i)\|_{\left(\Sigma_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1}}\times\widetilde{\mathcal{O}}\left(dH\sqrt{K(\max_i A_i)^3}\right)+1\\
&\overset{(iv)}{\le}\max_{a_i\in\mathcal{A}_i}\|\phi_i(s,a_i)\|_{\left(\Sigma_{i,h}^{\overline{\pi}}+\lambda I\right)^{-1}}\times\widetilde{\mathcal{O}}\left(dH\sqrt{K(\max_i A_i)^3}\right)+1,
\end{aligned}$$

其中 (i) 使用 μ_i^k 的定义，h，$\mathcal{S}_\varepsilon$ 以及引理 28；(ii) 使用方程（20）；(iii) 使用引理 26 和 27，并对所有 $s' \in \mathcal{S}_\varepsilon$ 以及来自 v 使用并集界

$$1/\mathrm{poly}(B_\theta, K, d, H, A_i, \lambda^{-1}, \eta^{-1}, \delta^{-1})\text{-cover}$$

属于 d 维单位球，且 (iv) 利用了 s' 是 $\mathcal{S}_\varepsilon$ 中 s 的最近邻这一事实。

因此，

$$\text{Term (C)} \le \max_{a_i \in \mathcal{A}_i} \|\phi_i(s,a_i)\|_{(\Sigma_{i,h}^{\overline{\pi}}+\lambda I)^{-1}} \times \widetilde{\mathcal{O}}\left(dH\sqrt{K(\max_i A_i)^3}\right) + 1.$$

引理 28 的证明。为简化记号，记 $x_{a_i} = \phi_i(s,a_i)$ 和 $\overline{x}_{a_i} = \phi_i(s',a_i)$，$a_i \in \mathcal{A}_i$。我们按照以下规则将 $\mathcal{A}_i$ 中的动作划分为 $\{\mathcal{C}_v\}_{v=1}^n$ 个簇：当且仅当 $\|x_{a_i} - x_{a_i'}\|_2 \le \Delta$ 时，动作 a_i 和 a_i' 属于同一个簇，其中 $\Delta > 10\varepsilon$ 是稍后指定的参数。记 $y_v := \frac{1}{|\mathcal{C}_v|}\sum_{a_i \in \mathcal{C}_v} x_{a_i}$。进一步地，用 $c(a_i)$ 表示 $a_i \in \mathcal{A}_i$ 所属的簇。

不难验证，

$$\left\|\mathbb{E}_{a\sim\pi_{i,h}(\cdot|s)}[\phi_i(s,a_i)] - \mathbb{E}_{a\sim\pi_{i,h}(\cdot|s)}[y_{c(a_i)}]\right\|_2 \le \Delta,$$

and

$$\left\|\mathbb{E}_{a\sim\pi_{i,h}(\cdot|s')}[\phi_i(s',a_i)] - \mathbb{E}_{a\sim\pi_{i,h}(\cdot|s')}[y_{c(a_i)}]\right\|_2 \le 2\Delta.$$

因此，要证明引理 28，只需对其进行上界估计

$$\max_{v\in[n]} \left|\pi_{i,h}(\mathcal{C}_v \mid s) - \pi_{i,h}(\mathcal{C}_v \mid s')\right|.$$

对于任意 $(a_i, a_i', \theta) \in \mathcal{A}_i^2 \times \mathbb{R}^d$，定义事件

$$E_{a_i,a_i'}(v) = \{(x_{a_i} - x_{a_i'})^\top (w + Wv) > 0\}, \qquad E_{a_i}(v) = \bigcap_{a_i' \in (\mathcal{A}_i/\mathcal{C}_{c(a_i)})} E_{a_i,a_i'}(v).$$

类似地，我们将上述定义中的 x 替换为 $\overline{x}$，定义 $\overline{E}_{a_i,a_i'}(v)$ 和 $\overline{E}_{a_i}(v)$。我们有

$$\begin{aligned}
&\left|\pi_{i,h}(\mathcal{C}_v \mid s) - \pi_{i,h}(\mathcal{C}_v \mid s')\right| \\
&= \left|\mathbb{P}\left(\bigcup_{a_i\in\mathcal{C}_v} E_{a_i}(v)\right) - \mathbb{P}\left(\bigcup_{a_i\in\mathcal{C}_v} \overline{E}_{a_i}(v)\right)\right| \\
&= \left|\mathbb{P}\left(\bigcup_{a_i\in\mathcal{C}_v}\bigcap_{a_i'\in(\mathcal{A}_i/\mathcal{C}_v)} E_{a_i,a_i'}(v)\right) - \mathbb{P}\left(\bigcup_{a_i\in\mathcal{C}_v}\bigcap_{a_i'\in(\mathcal{A}_i/\mathcal{C}_v)} \overline{E}_{a_i,a_i'}(v)\right)\right| \\
&\le \sum_{a_i\in\mathcal{C}_v}\sum_{a_i'\in(\mathcal{A}_i/\mathcal{C}_v)} \left|\mathbb{P}(E_{a_i,a_i'}(v)) - \mathbb{P}(\overline{E}_{a_i,a_i'}(v))\right|.
\end{aligned}$$

根据 $E_{a_i,a_i'}(v)$ 的定义，

$$\begin{aligned}
\mathbb{P}(E_{a_i,a_i'}(v)) &= \mathbb{P}\big((x_{a_i} - x_{a_i'})^\top (w + Wv) > 0\big) \\
&= \mathbb{P}\big((x_{a_i} - x_{a_i'})^\top Wv > (x_{a_i'} - x_{a_i})^\top w\big) \\
&= \mathbb{P}\left(v_1 > \frac{(x_{a_i'} - x_{a_i})^\top w}{\|(x_{a_i} - x_{a_i'})^\top W\|_2}\right),
\end{aligned}$$

其中最后一个等式利用了分布的对称性 $\mathcal{V}$。通过简单的代数运算，可以证明 $_\sim v_1$ 的密度函数上界为 $\mathcal{O}(d)$。因此，我们有

$$\left|\mathbb{P}(E_{a_i,a_i'}(v)) - \mathbb{P}(\overline{E}_{a_i,a_i'}(v))\right|$$

$$
\begin{aligned}
&\le \widetilde{\mathcal{O}}(d) \times \left| \frac{(x_{a_i'} - x_{a_i})^\top w}{\|(x_{a_i} - x_{a_i'})^\top W\|_2} - \frac{(\overline{x}_{a_i'} - \overline{x}_{a_i})^\top w}{\|(\overline{x}_{a_i} - \overline{x}_{a_i'})^\top W\|_2} \right| \\
&\le \widetilde{\mathcal{O}}(d) \times \left| \frac{(x_{a_i'} - x_{a_i})^\top w \times \|(\overline{x}_{a_i} - \overline{x}_{a_i'})^\top W\|_2 - (\overline{x}_{a_i'} - \overline{x}_{a_i})^\top w \times \|(x_{a_i} - x_{a_i'})^\top W\|_2}{\|(x_{a_i} - x_{a_i'})^\top W\|_2 \times \|(\overline{x}_{a_i} - \overline{x}_{a_i'})^\top W\|_2} \right| \\
&\le \widetilde{\mathcal{O}}\left(\frac{d}{\alpha^2\Delta^2}\right) \times \left| (x_{a_i'} - x_{a_i})^\top w \times \|(\overline{x}_{a_i} - \overline{x}_{a_i'})^\top W\|_2 - (\overline{x}_{a_i'} - \overline{x}_{a_i})^\top w \times \|(x_{a_i} - x_{a_i'})^\top W\|_2 \right| \\
&\le \widetilde{\mathcal{O}}\left(\frac{d}{\alpha^2\Delta^2}\right) \times \mathcal{O}(\beta\gamma\varepsilon) = \widetilde{\mathcal{O}}\left(\frac{d\beta\gamma\varepsilon}{\alpha^2\Delta^2}\right)
\end{aligned}
$$

其中：(i) 第三个不等式利用了 a_i 和 $a_{i'}$ 来自不同的簇这一事实， $W \succeq \alpha I$，以及 $\|\overline{x}_{a_i} - x_{a_i}\|_2 \le \varepsilon \le 0.1\Delta$；(ii) 最后一个不等式利用了三角不等式、 $\|w\|_2 \le \gamma$、 $W \preceq \beta I$ 和 $\|\overline{x}_{a_i} - x_{a_i}\|_2 \le \varepsilon$。我们通过选择 $\Delta = 20\varepsilon^{1/4}$完成证明。■

第 G 节：第 4.2 节的证明

G.1 表格型 **AVLPR** 算法的详细内容

表格型 MG 情形是状态数量有限的线性函数逼近设置的特例，即$|\mathcal{S}| \le S$。对于表格设置，我们将切换判据函数 Ψ_h 设为

$$
\Psi_h(\mathcal{B}_h) := \ln \prod_{s \in \mathcal{S}} \max \left\{ \sum_{s_h \in \mathcal{B}_h} \mathbb{I}(s_h = s), 1 \right\},
$$

而探索方案选择为 $\Gamma_{\text{explore}}(\overline{\pi},\ \mu_h) :=\ \{(\overline{\pi}_{1:h-1} \odot \mu_h,\ [m])\}$。换言之，在算法 2 的第 4 行和算法3的第 3 行，所有参与者共同执行 μ_h^k （或 π_h）一次。

无后悔算法 **O-REGRET-ALG** 注意到 $\mathcal{D}^{k,\ i\text{ 样本}} = \{(s_h^k,\ a_{i,\ h}^k,\ r_{i,\ h}^k + V_{i,\ h+1}(s_{h+1}^k))\}$ 始终只包含一个样本。我们将用它执行一种 EXP3-IX 风格的更新（Neu，2015），即

$$
\widehat{\ell}_{i,h}^k(s, a_i) = \frac{H - r_{i,h}^k - V_{i,h+1}(s_{h+1}^k)}{\mu_{i,h}^k(a_i \mid s) + \gamma_i} \times \mathbb{I}((s, a_i) = (s_h^k, a_{i,h}^k)),
$$

$$
\mu_{i,h}^{k+1}(\cdot \mid s) \propto \exp\left(-\eta_i \sum_{k' \le k} \widehat{\ell}_{i,h}^{k'}(s, \cdot) \right),
$$

其中 $\eta_i = \sqrt{\frac{S \log A_i T}{H^2 A_i T}}$且

$\gamma_i = \frac{\eta}{2}$。

OPTIMISTIC-REGRESS 将 $\mathcal{D}_{\text{reg}}^i$ 中的数据元组记为 $\{(s_h^k, a_{i,h}^k, r_{i,h}^k + V_{i,h+1}(s_{h+1}^k))\}_{k \in [K]}$。定义 $N_h(s) :=\ \sum_{k=1}^K \mathbb{I}(s_h^k = s)$，并且

$$
\beta_i(n) := \Theta\left(\frac{\iota}{\eta_i(n + \iota)} + \eta_i H^2 A_i \right),
$$

其中 $\iota = \log(KSA_iHm/\delta)$。乐观回归通过带奖励项的经验平均步骤执行：如果$N_h(s) > 0$，则设置 $V_{i,h}(s) = H - h + 1$，否则，

$$
V_{i,h}(s) = \min \left\{ \frac{1}{N_h(s)} \sum_{k=1}^K (r_{i,h}^k + V_{i,h+1}(s_{h+1}^k)) \times \mathbb{I}(s_h^k = s) + \beta_i(N_h(s)), H - h + 1 \right\}.
$$

计算效率 很容易看出，由于我们的实例化仅涉及标准 EXP3 算法、指数权重更新、奖励项计算和简单平均，因此整个算法可在 $(T, H, S, \{A_i\}_{i\in[m]})$ 的多项式时间内运行。

本节剩余部分将通过验证条件 6至(1A) 17 ，然后应用定理 18，来证明定理 。

G.2 条件的证明(1A)

记 $N_h(s)$ 为在 CCE- APPROX 中执行 $\overline{\pi}$ 的 K 个回合里，状态 s 在第 h 步被访问的次数。令 $\iota = \log(mSK \max_i A_i/\delta)$ 且 $p_s := \mathbb{P}^{\overline{\pi}}(s_h = s)$。利用 EXP3-IX 的理论保证（例如 Lattimore 和 Szepesv´ari（2020年）中的定理 12.1），并对所有 $(i, s) \in [m] \times \mathcal{S}$ 应用并集界，可得以至少 $1-\delta$ 的概率，对所有 $(i, s) \in [m] \times \mathcal{S}$:

$$\max_{\mu_{i,h}\in\Delta_{A_i}} \sum_{k=1}^{K} \left(\mathbb{D}_{\mu_{i,h}\times\mu^k_{-i,h}} - \mathbb{D}_{\mu^k_h}\right)\left[r_{i,h} + \mathbb{P}_{h+1}V_{i,h+1}\right](s) \times \mathbb{I}(s^k_h = s) \le \mathcal{O}\left(\frac{\iota}{\eta_i} + \eta H^2 A_i N_h(s)\right).$$

根据 Freedman 不等式，并对所有 $(i, s,\mu_{i,\ h}) \in [m] \times \mathcal{S} \times \{e_i\}_{i\in[A_i]}$ 应用并集界，可得以至少 $1-\delta$ 的概率，对所有 $(i, s) \in [m] \times \mathcal{S}$:

$$\begin{aligned}&\max_{\mu_{i,h}\in\Delta_{A_i}} \sum_{k=1}^{K} \left(\mathbb{D}_{\mu_{i,h}\times\mu^k_{-i,h}} - \mathbb{D}_{\mu^k_h}\right)\left[r_{i,h} + \mathbb{P}_{h+1}V_{i,h+1}\right](s) \times \mathbb{I}(s^k_h = s)\\ \ge& p_s \times \max_{\mu_{i,h}\in\Delta_{A_i}} \sum_{k=1}^{K} \left(\mathbb{D}_{\mu_{i,h}\times\mu^k_{-i,h}} - \mathbb{D}_{\mu^k_h}\right)\left[r_{i,h} + \mathbb{P}_{h+1}V_{i,h+1}\right](s) - \mathcal{O}(H\sqrt{p_s K\iota} + H\iota),\end{aligned}$$

and

$$N_h(s) \le \mathcal{O}(p_s K + \iota).$$

结合以上所有关系可得

$$\begin{aligned}&\frac{1}{K}\max_{\mu_{i,h}\in\Delta_{A_i}} \sum_{k=1}^{K} \left(\mathbb{D}_{\mu_{i,h}\times\mu^k_{-i,h}} - \mathbb{D}_{\mu^k_h}\right)\left[r_{i,h} + \mathbb{P}_{h+1}V_{i,h+1}\right](s)\\ \le& \min\left\{\mathcal{O}\left(\frac{\iota}{\eta_i p_s K} + \eta_i H^2 A_i\left(1 + \frac{\iota}{p_s K}\right)\right), H\right\}\\ \le& \mathcal{O}\left(\frac{\iota}{\eta_i(p_s K + \iota)} + \eta_i H^2 A_i\right),\end{aligned}$$

其中，最后一个不等式利用了 $\eta_i^{-2} \ge A_i$ 这一事实。因此，我们可以选取

$$G_{i,h}(s, \overline{\pi}, K, \delta) = \mathcal{O}\left(\frac{\iota}{\eta_i(K\mathbb{P}^{\overline{\pi}}(s_h = s) + \iota)} + \eta_i H^2 A_i\right).$$

G.3 条件的证明(1B)

将 s 状态在执行 V-APPROX 的 K 个回合中第 h 步被访问的次数记为 $N_h(s)$。令 $\iota = \log(mSK \max_i A_i/\delta)$，且 $p_s := \mathbb{P}^{\overline{\pi}}(s_h = s)$。由于 $N_h(s) = 0$ 的情形显然成立，下面我们只考虑满足 $N_h(s) > 0$的状态 s。

根据 Azuma-Hoeffding不等式，并对所有 $(i,s,\mu_{i,h}) \in [m]\times\mathcal{S}\times\{e_i\}_{i\in[A_i]}$ 应用并集界，可得以至少 $1-\delta$ 的概率，对所有 $(i,s)\in[m]\times\mathcal{S}$ 均有：

$$\left|\frac{1}{N_h(s)}\sum_{k=1}^{K}\left(r_{i,h}^k+V_{i,h+1}(s_h^{k+1})\right)\times\mathbb{I}(s_h^k=s)-\mathbb{D}_{\pi_h}\left[r_{i,h}+\mathbb{P}_{h+1}V_{i,h+1}\right](s)\right|$$
$$\leq\mathcal{O}\left(H\sqrt{\frac{\iota}{N_h(s)}}\right)\leq\mathcal{O}\left(\frac{\iota}{\eta_i(N_h(s)+\iota)}+\eta_i H^2A_i\right),$$

其中第二个不等式利用了以下事实：$\eta_i^{-1}\geq\iota$。因此，要证明条件(1B)中的两个关系，只需证明对所有$(i,s)\in[m]\times\mathcal{S}$:

$$G_{i,h}(s,\overline{\pi},K,\delta)=\Theta\left(\frac{\iota}{\eta_i(N_h(s)+\iota)}+\eta_i H^2A_i\right).$$

根据Bernstein不等式，并对所有 $s\in\mathcal{S}$应用并集界，可知以至少 $1-\delta$的概率，对所有$s\in\mathcal{S}$:

$$\frac{1}{2}p_sK-\frac{1}{2}\iota\leq N_h(s)\leq 2p_sK+\frac{1}{2}\iota.$$

将上述夹逼关系代回$G_{i,h}(s,\overline{\pi},K,\delta)$的定义，即完成证明。

G.4 条件(1C)的证明

令 $\iota=\log(mSK\max_i A_i/\delta)$，$w_s^t:=\mathbb{P}^{\pi^t}(s_h=s)$和 $W_s^t:=\sum_{\tau\leq t}\mathbb{P}^{\pi^\tau}(s_h=s)$。代入$G_{i,h}$的定义可得

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}_{\pi^{t+1}}\left[G_{i,h}(s,\overline{\pi}^t,t,\delta)\right]\\=&\mathcal{O}\left(\sum_{t=1}^{T}\mathbb{E}_{\pi^{t+1}}\left[\frac{\iota}{\eta_i(W_s^t+\iota)}+\eta_iH^2A_i\right]\right)\\=&\mathcal{O}\left(\sum_{s\in\mathcal{S}}\sum_{t=1}^{T}w_s^{t+1}\left[\frac{\iota}{\eta_i(W_s^t+\iota)}+\eta_iH^2A_i\right]\right)=\mathcal{O}\left(\frac{S\iota\log(T)}{\eta_i}+\eta_iH^2A_iT\right).\end{aligned}$$

G.5 条件的证明17

记 $N_h(s,\mathcal{B}_h)$ 为状态 s 在数据集 $\mathcal{B}_h$ 中出现的次数。由 Bernstein不等式，并对所有 $(t,s)\in[T]\times\mathcal{S}$ 取并集界，可知以至少 $1-\delta$ 的概率，对于所有 $(t,s)\in[T]\times\mathcal{S}$:

$$\frac{1}{2}\sum_{\tau=1}^{t}\mathbb{P}^{\pi^\tau}(s_h=s)-\frac{1}{2}\iota\leq N_h(s,\mathcal{B}_h^t)\leq 2\sum_{\tau=1}^{t}\mathbb{P}^{\pi^\tau}(s_h=s)+\frac{1}{2}\iota.$$

由于 $\Psi_{i,h}(\mathcal{B}_h^t)\leq\Psi_{i,h}(\mathcal{B}_h^{I_t})+1$，因此有 $N_h(s,\mathcal{B}_h^t)\leq 2N_h(s,\mathcal{B}_h^{I_t})$。利用上述相对集中性结果，我们得到

$$\sum_{\tau=1}^{t}\mathbb{P}^{\pi^\tau}(s_h=s)+\iota\leq 2(N_h(s,\mathcal{B}_h^t)+\iota)\leq 4N_h(s,\mathcal{B}_h^{I_t})+2\iota\leq 8\sum_{\tau=1}^{I_t}\mathbb{P}^{\pi^\tau}(s_h=s)+4\iota.$$

最后，回顾17，我们完成条件 (a) 的证明

$$G_{i,h}(s,\overline{\pi}^t,t,\delta)=\frac{\iota}{\eta_i(\sum_{\tau=1}^{t}\mathbb{P}^{\pi^\tau}(s_h=s)+\iota)}+\eta_iH^2A_i.$$

至于条件 17(b)，只需注意：(1) $\Psi_{i,h}$ 不依赖于 i；(2) $\Psi_{i,h}(\mathcal{D}_h^t)\leq S\log t$。因此，截至迭代 T的切换次数总数至多为 $SH\log T$。换言之，条件 17(b) 得到满足，且$d_{\text{replay}}=SH$。

G.6 表格型马尔可夫博弈的样本复杂度

第 G.2 节和 G.3 节表明，条件 (1A) 和(1B) 均得到满足，且

$$G_{i,h}(s,\overline{\pi},t,\delta)=\mathcal{O}\left(\frac{\iota}{\eta_i(K\mathbb{P}^{\overline{\pi}}(s_h=s)+\iota)}+\eta_i H^2 A_i\right).$$

与此同时，第 G.4 节表明，这一 G 满足条件 (1C)，且

$$L=\widetilde{\mathcal{O}}\left(SH^2\max_i A_i\right).$$

最后，第 G.5 节表明，条件 17 在 $d_{\text{replay}}=SH$下成立。接下来只需应用定理18，即可得到样本复杂度上界：

$$\widetilde{\mathcal{O}}\left(\frac{H^3Ld_{\text{replay}}}{\varepsilon^2}\right)=\widetilde{\mathcal{O}}\left(\frac{S^2H^6\max_i A_i}{\varepsilon^2}\right).$$

请注意，在表格型算法中，Γ_{explore} 只包含一个元素，因此 $\overline{\Gamma}=1$。

H Π^{Mar}-粗相关均衡与粗相关均衡的区别

这里给出一个玩具马尔可夫博弈的例子，其中存在一个相关策略 $\Lambda\in\Delta(\Pi^{\mathsf{Mar}})$，其中$\Pi^{\mathsf{Mar}}$ 是所有马尔可夫乘积策略的集合，使得 $\mathsf{CCEGap}^{\Pi^{\mathsf{Mar}}}(\Lambda)=0$，但对于任意$H\geq 2$，$\mathsf{CCEGap}(\Lambda)\geq H/4$。

考虑以下时域为 $H\geq 2$ 的“序贯石头剪刀布”博弈。这是一个双人零和博弈（其中 $m=2$ 和 $r_2\equiv 1-r_1$）。状态空间只有一个状态（$\mathcal{S}=\{s_0\}$ 和 $S=1$），每位玩家有三个动作，分别对应石头、布和剪刀（$A_1=A_2=3$）。玩家 1 的即时奖励 $r_1(a_1,a_2)\in\{0,1/2,1\}$由标准石头剪刀布规则决定（例如，$r_1$(石头, 剪刀) $=1$ 和r_1(石头, 石头) $=1/2$）。令 $\Pi_1^{\mathsf{Mar}},\Pi_2^{\mathsf{Mar}}$表示每位玩家所有马尔可夫策略的集合，并令 $\Pi^{\mathsf{Mar}}=\Pi_1^{\mathsf{Mar}}\times\Pi_2^{\mathsf{Mar}}$。在此博弈中，马尔可夫策略对应于在每个阶段执行一种无记忆（不依赖历史）的策略$h\in[H]$。

令 $\Lambda=\mathsf{Unif}(\{\pi^{\text{rock}},\pi^{\text{paper}},\pi^{\text{scissors}}\})$，其中对每个 $\mathsf{a}\in\{$石头、剪刀、布$\}$，

$$\pi^{\mathsf{a}}:=\pi_1^{\mathsf{a}}\times\pi_2^{\mathsf{a}},\quad\text{where}\quad\pi_{i,h}^{\mathsf{a}}(\cdot|s_0)=\delta_{\mathsf{a}}\ \text{ for all }(i,h)\in[2]\times[H]$$

表示一种策略：双方在所有 H 步中都确定性地执行动作 a。注意，$\pi^{\mathsf{a}}\in\Pi^{\mathsf{Mar}}$，因此 $\Lambda\in\Delta(\Pi^{\mathsf{Mar}})$。

根据 Λ 的定义，我们有 $V_1^{\Lambda}=H/2$。此外，显然 $\max_{\pi_1^\dagger\in\Pi_1^{\mathsf{Mar}}}V^{\pi_1^\dagger,\ \Lambda_{-1}}_1=H/2$，因为选择 $\pi_1^\dagger=\pi_1^{\text{rock}}$ 即可达到这一结果；而面对每一步都在 $\{$石头、剪刀、布$\}$ 中均匀出拳的 Λ_{-1}，其他任何马尔可夫策略 $\pi_1^\dagger\in\Pi_1^{\mathsf{Mar}}$（它是无记忆的）都无法在任何一步获得超过 $1/2$ 的奖励。这表明 $\mathsf{CCEGap}^{\Pi^{\mathsf{Mar}}}(\Lambda)=0$。

然而，考虑非马尔可夫策略 $\widetilde{\pi}_1$：在 $h=1$ 随机均匀地出拳，观察对手执行的动作（或根据收到的奖励推断对手执行的动作），此后在第 $h\in\{2,\dots,H\}$ 步执行能够战胜该动作的获胜动作。根据 Λ 的定义，这种非马尔可夫策略将在所有步骤 $h\geq 2$ 中确定性地获得奖励 1，因此

$$V_1^{\widetilde{\pi}_1,\Lambda_{-1}}=\frac{1}{2}+H-1=H-\frac{1}{2},$$

这给出了

$$\mathrm{CCEGap}(\Lambda)\geq V_1^{\widetilde{\pi}_1,\Lambda_{-1}}-V_1^{\Lambda}=H-\frac{1}{2}-\frac{H}{2}\geq\frac{H}{4}$$

对于任意 $H\geq 2$。

I 第5节的证明

I.1 探索式全策略评估 (APE)

我们在算法APE 中完整描述了6。

算法 6 APE$_i(\mathcal{F}_i, \Pi_i, \pi_{-i}, K, \beta)$：探索式全策略评估（第 i 位玩家）

1：初始化置信集 $\mathcal{B}^1 \leftarrow \mathcal{F}_i \times \Pi_i, \mathcal{D}_h \leftarrow \{\}$。2：对于 $k = 1, ..$
., K，执行 3：计算所有 $\pi_i \in \Pi_i$ 的价值估计上界和下界：

$$(\overline{V}^{k,\pi_i\times\pi_{-i}}, \underline{V}^{k,\pi_i\times\pi_{-i}}) \leftarrow \left(\max_{f:(f,\pi_i)\in\mathcal{B}^k} f_1(s_1, \pi_{i,1}(s_1)), \min_{f:(f,\pi_i)\in\mathcal{B}^k} f_1(s_1, \pi_{i,1}(s_1))\right).$$

4：选择 $\pi_i^k \leftarrow \arg\max_{\pi_i\in\Pi_i}(\overline{V}^{k,\pi_i\times\pi_{-i}} - \underline{V}^{k,\pi_i\times\pi_{-i}})$。5：执行 $\pi_i^k \times \pi_{-i}$，并为第 i 位玩家收集轨迹 $(s_1^k, a_{i,1}^k, r_{i,1}^k, \ldots, s_H^k, a_{i,H}^k, r_{i,H}^k)$。6：对所有 $h \in [H]$ 更新 $\mathcal{D}_h \leftarrow \mathcal{D}_h \cup \{(s_h^k, a_{i,h}^k, r_{i,h}^k, s_{h+1}^k)\}$。7：更新置信集

$$\mathcal{B}^{k+1} = \left\{(f, \pi_i) \in \mathcal{F}_i \times \Pi_i : \mathcal{L}_h^{\mathcal{D}_h}(f_h, f_{h+1}, \pi_i) \le \min_{f_h'\in\mathcal{F}_{i,h}} \mathcal{L}_h^{\mathcal{D}_h}(f_h', f_{h+1}, \pi_i) + \beta, \forall h \in [H]\right\} \bigcap \mathcal{B}^k,$$

$$\text{where } \mathcal{L}_h^{\mathcal{D}_h}(f_h, f_{h+1}, \pi_i) := \sum_{(s,a_i,r,s')\in\mathcal{D}_h} \left[f_h(s, a_i) - r - f_{h+1}(s', \pi_{i,h}(s'))\right]^2.$$

输出：乐观价值估计 $\{\overline{V}^{K,\pi_i\times\pi_{-i}}\}_{\pi_i\in\Pi_i}$。

I.2 定理 9 的证明

本节证明定理 9。我们首先在以下命题中给出 APE 子程序的保证，其证明见附录 I.4。

命题 29 （通过 APE 学习所有策略的准确 Q 函数）。 在假设 *7* & *8*下，存在一个绝对常数 $c > 0$，使得对于任意玩家 $i \in [m]$，若在算法 $\beta = cH^2\log(|\Pi_i||\mathcal{F}_i|KH/\delta)$ 中选择 *6*，则以至少 $1-\delta$ 的概率有

(a) $\underline{V}^{K,\pi_i\times\pi_{-i}} \le V_{i,1}^{\pi_i,\pi_{-i}}(s_1) \le \overline{V}^{K,\pi_i\times\pi_{-i}}$ *for all* $\pi_i \in \Pi_i$.

(b) $\max_{\pi_i\in\Pi_i}\left(\overline{V}^{K,\pi_i\times\pi_{-i}} - \underline{V}^{K,\pi_i\times\pi_{-i}}\right) \le \mathcal{O}\left(H\sqrt{\frac{d_i \log K\cdot\beta}{K}}\right)$.

由于算法 4 调用 APE 子程序 T 轮次，每轮有 m 个参与者，参数为 $(\beta, K) \leftarrow (\beta_i, K_i)$，应用命题 29并使用并集界可得，以至少 $1-\delta/2$的概率，乐观价值估计$\{\overline{V}_i^{(t),\pi_i\times\pi_{-i}^t}\}_{\pi_i\in\Pi_i}$ 满足

$$\begin{aligned} V_{i,1}^{\pi_i,\pi_{-i}}(s_1) &\overset{(i)}{\le} V_{i,1}^{\pi_i,\pi_{-i}}(s_1) + \mathcal{O}\left(H^2\sqrt{\frac{d_i\log(K_i)\cdot\log(\sum_i|\Pi_i||\mathcal{F}_i|TK_iH/\delta)}{K_i}}\right) \\ &\overset{(ii)}{\le} V_{i,1}^{\pi_i,\pi_{-i}}(s_1) + \varepsilon/2 \end{aligned} \tag{21}$$

对于所有 $i \in [m]$、$\pi_i \in \Pi_i$，以及 $t \in [T]$ 同时。上文中，（i）用到了我们对 β_i 的选择，而（ii）可以通过选择 b 来满足 y

选择

$$K_i = \widetilde{\mathcal{O}}\left(\frac{H^4 d_i \cdot \log(\sum_i |\Pi_i||\mathcal{F}_i|)}{\varepsilon^2}\right). \tag{22}$$

接下来我们证明，DOPMD 对任何满足（21）的乐观价值估计，都能实现较小的遗憾。证明见附录I.3。

命题 30 （DOPMD的遗憾保证）。假设算法*4* 中的乐观价值估计具有有效乐观性和一致的小误差，即

$$V_{i,1}^{\pi_i \times \pi_{-i}^t}(s_1) \le \overline{V}_i^{(t),\pi_i \times \pi_{-i}^t} \le V_{i,1}^{\pi_i \times \pi_{-i}^t}(s_1) + \varepsilon \tag{23}$$

对所有 $t \in [T]$、 $i \in [m]$和 $\pi_i \in \Pi_i$均成立。那么，算法*4* 使用 $\eta_i = \sqrt{\log |\Pi_i|/(H^2 T)}$ 以至少 $1-\delta$ 的概率达到以下结果

$$\max_{i\in[m]} \max_{\pi_i \in \Pi_i} \sum_{t=1}^{T} \left[V_{i,1}^{\pi_i \times \Lambda_{-i}^t}(s_1) - V_{i,1}^{\Lambda^t}(s_1)\right] \le \varepsilon T + \mathcal{O}\left(H\sqrt{T \log\Big(\sum_{i\in[m]} |\Pi_i|/\delta\Big)}\right). \tag{24}$$

由（21）和命题30可知，以至少 $1-\delta$的概率，算法 $\overline{\Lambda}$ 4 的输出策略 达到

$$\mathrm{CCEGap}^{\Pi}(\overline{\Lambda}) = \max_{i\in[m]} \max_{\pi_i \in \Pi_i} \left(V_{i,1}^{\pi_i \times \overline{\Lambda}_{-i}} - V_{i,1}^{\overline{\Lambda}}\right) = \frac{1}{T} \max_{i\in[m]} \max_{\pi_i \in \Pi_i} \sum_{t=1}^{T} \left[V_{i,1}^{\pi_i \times \Lambda_{-i}^t}(s_1) - V_{i,1}^{\Lambda^t}(s_1)\right]$$

$$\le \varepsilon/2 + \mathcal{O}\left(H\sqrt{\log\Big(\sum_{i\in[m]} |\Pi_i|/\delta\Big)/T}\right) \le \varepsilon,$$

其中，最后一个不等式要求选择

$$T = \widetilde{\mathcal{O}}\left(\frac{H^2 \log(\sum_{i\in[m]} |\Pi_i|)}{\varepsilon^2}\right). \tag{25}$$

结合（22）和（25），已进行的回合总数至多为

$$T \times \left(\sum_{i\in[m]} K_i\right) = \widetilde{\mathcal{O}}\left(\frac{H^6\left(\sum_{i\in[m]} d_i\right) \cdot \log^2(\sum_i |\Pi_i||\mathcal{F}_i|)}{\varepsilon^4}\right).$$

定理的证明至此完成 9。 ■

I.3 命题证明30

任取一名玩家 $i \in [m]$。我们有

$$\begin{aligned}
\mathrm{Reg}_T^i &:= \max_{\pi_i \in \Pi_i} \sum_{t=1}^{T} \left[V_{i,1}^{\pi_i \times \Lambda_{-i}^t}(s_1) - V_{i,1}^{\Lambda^t}(s_1)\right] \\
&\le \underbrace{\max_{\pi_i \in \Pi_i} \sum_{t=1}^{T} \left[V_{i,1}^{\pi_i \times \pi_{-i}^t}(s_1) - V_{i,1}^{\Lambda_i^t \times \pi_{-i}^t}(s_1)\right]}_{\mathrm{I}} + \mathcal{O}\left(H\sqrt{T \log\Big(\sum_{i\in[m]} |\Pi_i|/\delta\Big)}\right)
\end{aligned}$$

以至少 $1-\delta$ 的概率成立，其中该不等式利用了以下事实：

$$\begin{aligned}
&\max_{\Lambda_i\in\Delta(\Pi_i)}\left|\sum_{t=1}^{T}\left[V_{i,1}^{\Lambda_i\times\pi_{-i}^t}(s_1)-V_{i,1}^{\Lambda_i\times\Lambda_{-i}^t}(s_1)\right]\right|\\
&=\max_{\Lambda_i\in\Delta(\Pi_i)}\left|\sum_{\pi_i\in\Pi_i}\Lambda_i(\pi_i)\sum_{t=1}^{T}\left[V_{i,1}^{\pi_i\times\pi_{-i}^t}(s_1)-V_{i,1}^{\pi_i\times\Lambda_{-i}^t}(s_1)\right]\right|\\
&=\max_{\pi_i\in\Pi_i}\left|\sum_{t=1}^{T}\left[V_{i,1}^{\pi_i\times\pi_{-i}^t}(s_1)-V_{i,1}^{\pi_i\times\Lambda_{-i}^t}(s_1)\right]\right|\le\mathcal{O}(H\sqrt{T\log\Big(\sum_{i\in[m]}|\Pi_i|/\delta\Big)}),
\end{aligned}$$

接下来，通过对所有 $i\in[m]$ 和所有 $\pi_i\in\Pi_i$ 同时应用 Azuma-Hoeffding 不等式。然后，为了界定项 I，我们有

$$\begin{aligned}
\mathrm{I}&=\max_{\pi_i\in\Pi_i}\left(\sum_{t=1}^{T}\left[V_{i,1}^{\pi_i\times\pi_{-i}^t}(s_1)-V_{i,1}^{\Lambda_i^t\times\pi_{-i}^t}(s_1)\right]\right)\\
&=\underbrace{\max_{\pi_i\in\Pi_i}\sum_{t=1}^{T}\left[\overline{V}_i^{(t),\pi_i\times\pi_{-i}^t}-\overline{V}_i^{(t),\Lambda_i^t\times\pi_{-i}^t}\right]}_{(a)}+\underbrace{\max_{\pi_i\in\Pi_i}\sum_{t=1}^{T}\left[V_{i,1}^{\pi_i\times\pi_{-i}^t}(s_1)-\overline{V}_i^{(t),\pi_i\times\pi_{-i}^t}\right]}_{(b)}\\
&\quad+\underbrace{\sum_{t=1}^{T}\left[\overline{V}_i^{(t),\Lambda_i^t\times\pi_{-i}^t}-V_{i,1}^{\Lambda_i^t\times\pi_{-i}^t}(s_1)\right]}_{(c)}.
\end{aligned}$$

由 (23)，我们有 $(b)\le 0$ 和 $(c)\le\varepsilon_1\cdot T$。为了界定 (a)，注意到根据算法 4，Λ_i^t 具有以下等价的跟随正则化领导者（FTRL）形式：$\Lambda_i^t(\pi_i)\propto_{\pi_i}\exp\Big(\eta_i\sum_{\tau=1}^{t-1}\overline{V}_i^{(t,\ \pi_i\times\pi_{-i}^t}\Big)$，其中每个 $\overline{V}_i^{(t,\ \pi_i\times\pi_{-i}^t}\in[0,H]$。因此，根据标准的 FTRL 分析（Orabona，2019 年，第 6.6 节），

$$\max_{\pi_i\in\Pi_i}\sum_{t=1}^{T}\left[\overline{V}_i^{(t),\pi_i\times\pi_{-i}^t}-\overline{V}_i^{(t),\Lambda_i^t\times\pi_{-i}^t}\right]\le\frac{\log|\Pi_i|}{\eta_i}+\frac{\eta_i}{2}H^2T\le\mathcal{O}\Big(H\sqrt{\log|\Pi_i|\cdot T}\Big),$$

其中，在最后一个不等式中，我们取了 $\eta_i=\sqrt{\log|\Pi_i|/(H^2T)}$。这表明 $\mathrm{I}\le\varepsilon T+\mathcal{O}\Big(H\sqrt{\log|\Pi_i|\cdot T}\Big)$。将其代回遗憾界可得，以至少 $1-\delta$ 的概率，对所有 $i\in[m]$ 同时成立：

$$\begin{aligned}
\mathrm{Reg}_T^i&\le\varepsilon T+\mathcal{O}\Big(H\sqrt{\log|\Pi_i|\cdot T}\Big)+\mathcal{O}\left(H\sqrt{T\log\Big(\sum_{i\in[m]}|\Pi_i|/\delta\Big)}\right)\\
&\le\varepsilon T+\mathcal{O}\left(H\sqrt{T\log\Big(\sum_{i\in[m]}|\Pi_i|/\delta\Big)}\right).
\end{aligned}$$

这就证明了所需的结果。 ■

I.4 命题的证明29

我们首先给出以下引理，它表明在每次迭代中，置信集都包含任意策略 π_i 的真实价值函数，并且针对已访问的状态-动作对，估计误差较小。证明依赖于 II-完备性假设（假设 7）以及针对平方损失的标准快速率集中性论证，可参见附录 I.4.1。

引理 31 （ $\mathcal{B}^k$ 的性质）。 在假设 7 下，存在一个绝对常数 $c > 0$，使得如果我们在算法 $\beta = cH^2\log(|\Pi_i||\mathcal{F}_i|KH/\delta)$ 中选择 6，那么以至少 $1-\delta$ 的概率，

(a) $(Q^{\pi_i}, \pi_{-i}, \pi_i) \in \mathcal{B}^k$ 对所有 $(\pi_i, k) \in \Pi_i \times [K]$，(b) $\sum_{t=1}^{k-1}\left[(f_h - \mathcal{T}_i^{\pi_i\times\pi_{-i}} f_{h+1})(s_h^t, a_{i,h}^t)\right]^2 \le \mathcal{O}(\beta)$ 对所有 $(k,h) \in [K]\times[H]$ 以及 $(f,\pi_i) \in \mathcal{B}^k$，(c) $\sum_{t=1}^{k-1}\mathbb{E}_{(s_h,a_{i,h})\sim\pi_i^t\times\pi_{-i}}\left[(f_h - \mathcal{T}_i^{\pi_i\times\pi_{-i}} f_{h+1})(s_h, a_{i,h})^2\right] \le \mathcal{O}(\beta)$ 对所有 $(k,h) \in [K]\times[H]$ 以及 $(f,\pi_i) \in \mathcal{B}^k$。

根据引理 31(a)，在良好事件上，它确保（概率至少为 $1-\delta/2$）并且根据 $\overline{V}^{K,\pi_i\times\pi_{-i}}$ 和 $\underline{V}^{K,\pi_i\times\pi_{-i}}$ 在算法 6 中的定义，我们立即得到 $\underline{V}^{K,\pi_i\times\pi_{-i}} \le V^{\pi_i,\pi_{-i}}_{i,1}(s_1) \le \overline{V}^{K,\pi_i\times\pi_{-i}}$ 对所有 $(\pi_i,k) \in \Pi_i\times[K]$ 均成立，由此证明部分（a）。

为证明部分（b），对于任意 $k \in [K]$，将“探索策略” π_i^k 的乐观和悲观 Q 估计记为

$$\overline{f}^k = \arg\max_{f:(f,\pi_i^k)\in\mathcal{B}^k} f_1(s_1, \pi_{i,1}^k(s_1)) \quad \text{and} \quad \underline{f}^k = \arg\min_{f:(f,\pi_i^k)\in\mathcal{B}^k} f_1(s_1, \pi_{i,1}^k(s_1)),$$

其中回顾一下， π_i^k 的选取旨在使上述两个值之差在所有 $\pi_i \in \Pi_i$ 中最大。结合 $\mathcal{B}^k$ 的单调性可知，对于任意固定的 $\pi_i \in \Pi_i$,

$$\begin{aligned}
&K \times \left(\overline{V}^{K,\pi_i\times\pi_{-i}} - \underline{V}^{K,\pi_i\times\pi_{-i}}\right)\\
\le&\sum_{k=1}^K\left(\max_{f:(f,\pi_i)\in\mathcal{B}^k} f_1(s_1,\pi_{i,1}(s_1)) - \min_{f:(f,\pi_i)\in\mathcal{B}^k} f_1(s_1,\pi_{i,1}(s_1))\right)\\
\le&\sum_{k=1}^K\left(\overline{f}_1^k(s_1,\pi_{i,1}^k(s_1)) - \underline{f}_1^k(s_1,\pi_{i,1}^k(s_1))\right)\\
=&\sum_{k=1}^K\left(\overline{f}_1^k(s_1,\pi_{i,1}^k(s_1)) - V_{i,1}^{\pi_i^k\times\pi_{-i}}(s_1,\pi_{i,1}^k(s_1))\right) + \sum_{k=1}^K\left(V_{i,1}^{\pi_i^k\times\pi_{-i}}(s_1,\pi_{i,1}^k(s_1)) - \underline{f}_1^k(s_1,\pi_{i,1}^k(s_1))\right).
\end{aligned}$$

上述两项可以用相同的方法进行界定。不失一般性，下文仅讨论第一项。

回顾 BE 维度假设（假设8）指出， $\mathcal{D}_{\Pi_i\times\pi_{-i}}$ 型或 $\mathcal{D}_\Delta$ 型分布式 Eluder 维数有界（参见定义 16）。我们先考虑 $\mathcal{D}_\Delta$ 型分布式 Eluder 维数的情形，此时对于任意 $\varepsilon > 0$,

$$d_i(\varepsilon) := \max_{h\in[H]} d_{\mathrm{E}}\left(\left\{f_h - \mathcal{T}_{i,h}^{\pi_i\times\pi_{-i}} f_{h+1} : (f,\pi_i) \in \mathcal{F}\times\Pi_i\right\}, \mathcal{D}, \varepsilon\right) \le d_i\log(1/\varepsilon).$$

在这种情况下，我们有

$$\begin{aligned}
&\sum_{k=1}^K\left(\overline{f}_1^k(s_1,\pi_{i,1}^k(s_1)) - V_1^{\pi_i^k\times\pi_{-i}}(s_1,\pi_{i,1}^k(s_1))\right)\\
=&\sum_{h=1}^H\sum_{k=1}^K\mathbb{E}_{\pi_i^k\times\pi_{-i}}\left[\overline{f}_h^k(s_h,\pi_{i,h}^k(s_h)) - r_h - \overline{f}_{h+1}^k(s_{h+1},\pi_{i,h+1}^k(s_{h+1}))\right]\\
\overset{(i)}{\le}&\sum_{h=1}^H\sum_{k=1}^K\left[\left(\overline{f}_h^k - \mathcal{T}_{i,h}^{\pi_i^k\times\pi_{-i}}\overline{f}_{h+1}^k\right)(s_h^k,a_{i,h}^k)\right] + \mathcal{O}\left(H\sqrt{K\log(H/\delta)}\right)\\
\overset{(ii)}{\le}&\mathcal{O}\left(H\sqrt{d_i(K^{-1/2})K\beta}\right) + \mathcal{O}\left(H\sqrt{K\log(H/\delta)}\right) \le \mathcal{O}\left(H\sqrt{d_iK\log K\cdot\beta}\right).
\end{aligned} \tag{26}$$

上文中，(i) 由 Azuma-Hoeffding 不等式推出；(ii) 结合引理 31(b) 在 $(\overline{f}^k,\ \pi_i^k)$上应用，并利用 Eluder 维度论证（Jin 等人，2021a，引理 41），可得对所有 $h\in[H]$,

$$\sum_{t=1}^{k-1}\left[(\overline{f}_h^k-\mathcal{T}_{i,h}^{\pi_i^k\times\pi_{-i}}\overline{f}_{h+1}^k)(s_h^t,a_{i,h}^t)\right]^2\le\mathcal{O}(\beta)\quad\text{for all } k\in[K]$$

$$\Longrightarrow\sum_{k=1}^{K}\left[\left(\overline{f}_h^k-\mathcal{T}_{i,h}^{\pi_i^k\times\pi_{-i}}\overline{f}_{h+1}^k)(s_h^k,a_{i,h}^k)\right)\right]\le\mathcal{O}\left(\sqrt{d_i(K^{-1/2})K\beta}\right)\le\mathcal{O}\left(\sqrt{d_i\log K\cdot K\beta}\right).$$

对于 $\mathcal{D}_{\Pi_i\times\pi_{-i}}$型分布型 Eluder 维度的另一种情况，我们采用相同的论证，直到 (26) 中不等式 (i) 之前的步骤，并结合针对滚入分布$\{d_h^{\pi_i^k\times\pi_{-i}}\}_{k\ge1}$的相同 Eluder 维度论证与引理 31(c)，得到与 $\mathcal{D}_\Delta$ 情形相同的界。

再结合第二项的相同界，可得

$$K\times\left(\overline{V}^{K,\pi_i\times\pi_{-i}}-\underline{V}^{K,\pi_i\times\pi_{-i}}\right)\le\mathcal{O}\left(H\sqrt{d_iK\log K\cdot\beta}\right).$$

将两边同时除以 K，即可证明所需结论。 ■

I.4.1 引理 31 的证明

证明与 Jin 等人 (2021a，引理 39(b) 和 40) 类似。回顾一下，我们考虑一个固定的 π_{-i}，并用 $\pi=\pi_i\times\pi_{-i}$ 作简写。定义随机变量

$$X_h^t(f,\pi_i):=2(f_h-\mathcal{T}_{i,h}^{\pi}f_{h+1})(s_h^t,a_{i,h}^t)\times\left[r_{i,h}^t+f_{h+1}(s_{h+1}^t,\pi_{i,h+1}(s_{h+1}^t))-(\mathcal{T}_{i,h}^{\pi}f_{h+1})(s_h^t,a_{i,h}^t)\right]$$

对所有 $(f,\pi_i,t,h)\in\mathcal{F}_i\times\Pi_i\times[K]\times[H]$。

考虑滤过 $\{\mathcal{G}_h^t\}_{t\ge1}$，它包含迭代 t 中截至 $(s_h^t,a_{i,h}^t)$ 的所有历史观测，但不包含 $(r_{i,h}^t,s_{h+1}^t)$。注意，$X_h^t(f,\pi_i)$ 关于 $\{\mathcal{G}_h^t\}_{t\in[K]}$ 构成鞅差序列（因为第二项在 $\mathcal{G}_h^t$ 上的均值为零）。此外，由于对所有 $h\in[H]$，$f_h(\cdot,\cdot)\in[0,H-h+1]$，我们几乎必然有 $X_h^t(f,\pi_i)\le2H^2$。因此，根据 Freedman 不等式（引理 13）和并集界，对于任意固定的 $\lambda\le1/(2H^2)$，以至少 $1-\delta$ 的概率，我们有

$$\begin{aligned}\sum_{t=1}^{k}X_h^t(f,\pi_i)&\le4\lambda H^2\sum_{t=1}^{k}\left[(f_h-\mathcal{T}_{i,h}^{\pi}f_{h+1})(s_h^t,a_{i,h}^t)\right]^2+\frac{\log(|\mathcal{F}_i||\Pi_i|KH/\delta)}{\lambda}\\&=\frac{1}{2}\sum_{t=1}^{k}\left[(f_h-\mathcal{T}_{i,h}^{\pi}f_{h+1})(s_h^t,a_{i,h}^t)\right]^2+8H^2\log(|\mathcal{F}_i||\Pi_i|KH/\delta).\end{aligned}\tag{27}$$

对所有 (f,π_i,k,h) 同时成立，其中在第二行中我们取了 $\lambda=1/(8H^2)$。

令 $\mathcal{D}_h^k$ 表示算法 $\mathcal{D}_h$ 6 中维护的数据集，且该数据集在第 k 次迭代开始前（即用于构造 $\mathcal{B}^k$）。为证明部分（b），任取 $(k,h)\in[K]\times[H]$ 和 $(f,\pi_i)\in\mathcal{B}^k$。根据 $\mathcal{B}^k$ 的定义，我们有

$$\begin{aligned}\beta&\ge\mathcal{L}_h^{\mathcal{D}_h^k}(f_h,f_{h+1},\pi_i)-\min_{f_h'\in\mathcal{F}_{i,h}}\mathcal{L}_h^{\mathcal{D}_h^k}(f_h',f_{h+1},\pi_i)\\&\overset{(i)}{\ge}\mathcal{L}_h^{\mathcal{D}_h^k}(f_h,f_{h+1},\pi_i)-\mathcal{L}_h^{\mathcal{D}_h^k}(\mathcal{T}_{i,h}^{\pi}f_{h+1},f_{h+1},\pi_i)\\&=\sum_{t=1}^{k-1}\left[f_h(s_h^t,a_{i,h}^t)-r_{i,h}^t-f_{h+1}(s_{h+1}^t,\pi_{i,h+1}(s_{h+1}^t))\right]^2\\&\quad-\sum_{t=1}^{k-1}\left[(\mathcal{T}_{i,h}^{\pi}f_{h+1})(s_h^t,a_{i,h}^t)-r_{i,h}^t-f_{h+1}(s_{h+1}^t,\pi_{i,h+1}(s_{h+1}^t))\right]^2\end{aligned}$$

$$= -\sum_{t=1}^{k-1} X_h^t(f,\pi_i) + \sum_{t=1}^{k-1}\left[(f_h - \mathcal{T}_{i,h}^{\pi} f_{h+1})(s_h^t, a_{i,h}^t)\right]^2$$
$$\overset{(ii)}{\geq} -8H^2\log(|\mathcal{F}_i||\Pi_i|KH/\delta) + \frac{1}{2}\sum_{t=1}^{k-1}\left[(f_h - \mathcal{T}_{i,h}^{\pi} f_{h+1})(s_h^t, a_{i,h}^t)\right]^2.$$

上文中，(i) 由 Π-完备性（假设 7）得出，而 (ii) 由（27）得出。因此，取 $\beta = 8H^2\log(|\mathcal{F}_i||\Pi_i|KH/\delta)$即可确保

$$\sum_{t=1}^{k-1}\left[(f_h - \mathcal{T}_{i,h}^{\pi} f_{h+1})(s_h^t, a_{i,h}^t)\right]^2 \leq 4\beta,$$

这就证明了部分（b）。

为证明部分（a），首先注意到 $Q_i^{\pi} = Q_i^{\pi_i\times\pi_{-i}} \in \mathcal{F}_i$，因为我们有 $Q_{i,\,h}^{\pi} \in \mathcal{F}_{i,\,h}$ 根据假设 7 $\sim$ $\sim$反复得到对 $h = H$，.....，1 成立。因此，固定任意的 $(k,h) \in [K]\times[H]$和 $f_h' \in \mathcal{F}_{i,\,h}$，并令 $Q \in \mathcal{F}$定义为 $Q_h = f_h'$and$\sim Q_{h'} = Q_{i,\,h'}^{\pi}$对所有 $h' \neq h$均成立。类似地，我们有

$$\begin{aligned}
&\mathcal{L}_h^{\mathcal{D}_h^k}(Q_{i,h}^{\pi}, Q_{i,h+1}^{\pi}, \pi_i) - \mathcal{L}_h^{\mathcal{D}_h^k}(f_h', Q_{i,h+1}^{\pi}, \pi_i)\\
&= \mathcal{L}_h^{\mathcal{D}_h^k}(\mathcal{T}_{i,h}^{\pi}\widetilde{Q}_h, \widetilde{Q}_{h+1}, \pi_i) - \mathcal{L}_h^{\mathcal{D}_h^k}(\widetilde{Q}_h, \widetilde{Q}_{h+1}, \pi_i)\\
&= \sum_{t=1}^{k-1} X_h^t(\widetilde{Q}, \pi_i) - \sum_{t=1}^{k-1}\left[(f_h' - Q_{i,h}^{\pi})(s_h^t, a_{i,h}^t)\right]^2\\
&\overset{(i)}{\leq} 8H^2\log(|\mathcal{F}_i||\Pi_i|KH/\delta) - \frac{1}{2}\sum_{t=1}^{k-1}\left[(f_h' - Q_{i,h}^{\pi})(s_h^t, a_{i,h}^t)\right]^2 \overset{(ii)}{\leq} \beta,
\end{aligned}$$

其中，(i) 由 (27) 得出，而 (ii) 则由我们对 $\beta = 8H^2$log$(|\mathcal{F}_i||\Pi_i|KH/\delta)$ 的选择得出。由于这对任意的$f_h' \in \mathcal{F}_{i,\,h}$都成立，对上述左侧取上确界可得

$$\mathcal{L}_h^{\mathcal{D}_h^k}(Q_{i,h}^{\pi}, Q_{i,h+1}^{\pi}, \pi_i) - \inf_{f_h'\in\mathcal{F}_{i,h}} \mathcal{L}_h^{\mathcal{D}_h^k}(f_h', Q_{i,h+1}^{\pi}, \pi_i) \leq \beta.$$

由于这对所有 $h \in [H]$都成立，根据定义，我们有 $(Q_{i,\,h}^{\pi}, \pi) \in \mathcal{B}^k$ 对所有 $k \in [K]$均成立。这就证明了部分（a）。

最后，部分（c）可以通过与部分（b）完全相同的论证来证明，只需重新定义滤过 $\{\mathcal{G}_h^t\}_{t\geq 1}$，使其包含 在回合 t 开始之前的所有历史观测，从而使 $(s_h^t, a_{i,h}^t) \sim d_h^{\pi_i^k\times\pi_{-i}}$ 在给定 $\mathcal{G}_h^t$的条件下，并对尾概率 $\delta \to \delta/2$ 在（27）以及相对于此处新滤过的对应式中重新缩放。■

I.5 线性二次博弈的细节

这里我们给出线性二次高斯示例（示例 10）的细节。对所有 $i \in [m]$定义如下特征映射（其中 $d_{\phi,i} := d_S + d_{A,\,i} + 1$）：

$$\phi_i(s, a_i) = \begin{bmatrix} s \\ a_i \\ 1 \end{bmatrix}\begin{bmatrix} s^\top & a_i^\top & 1 \end{bmatrix} \in \mathbb{R}^{d_{\phi,i}\times d_{\phi,i}}.$$

对于所有 $i \in [m]$，我们考虑以下线性价值类和线性策略类：

- $\mathcal{F}_{i,h} := \{f_{i,h}(s,a_i) = \langle\phi_i(s,a_i), \theta_h\rangle : \theta_h \in \mathbb{R}^{d_{\phi,i}\times d_{\phi,i}}, \|\theta_h\|_{\mathsf{Fr}} \leq B_{\theta,h}\}$.
- $\Pi_i := \{\pi_i = \{\pi_{i,h}(s) = M_{i,h}s\}_{h\in[H]} : M_{i,h} \in \mathbb{R}^{d_{A,i}\times d_S}, \|M_{i,h}\|_{\mathsf{Fr}} \leq B_{M,h}\}$.

固定对手的任意线性策略 $\pi_{-i} \in \Pi_{-i}$ 后，根据状态转移的结构（5）和奖励，玩家 i 所面对的 MDP 可简化为线性二次调节器（LQR）。为简化记号，我们将其记为

$$\begin{cases} s_{h+1} = C_h s_h + D_h a_{i,h} + z_h, \\ r_{i,h}(s, a_i) = \langle J_{i,h}, \phi_i(s, a_i) \rangle. \end{cases}$$

上述 C_h、D_h、J_i、$_h$ 可由 A_h、$\{B_{i,h}\}_i$、$\{K_h^i\}_i$、$\{K_{j,h}^i\}_i$、$_j$ 和 π_{-i} 计算得到。不难看出，只要适当选择 B_θ、$_h = \mathcal{O}$（poly(d_ϕ、$_i$、B_M)$^{H-h+1}$）（根据覆盖论证，最终的样本复杂度只取决于其对数），对于任意 $f_{h+1} \in \mathcal{F}_i$、$_{h+1}$，都有 $\mathcal{T}_i^{\pi_i \times \pi_{-i}}$、$_h$ $f_{h+1} \in \mathcal{F}_i$、$_h$。这验证了假设 7。

进一步地，注意到函数类

$$\left\{ f_h - \mathcal{T}_h^{\pi_i \times \pi^{-i}} f_{h+1} \mid (f, \pi_i) \in \mathcal{F}_i \times \Pi_i \right\}$$

是一个具有 d_ϕ^2、$_i$ 维特征映射 ϕ_i(·、·) 的线性函数类。根据线性函数类的标准 Eluder 维数界，$\mathcal{D}_\Delta$ 型 BE 维度（定义 16）由 $\widetilde{\mathcal{O}}(d_\phi^2$、$_i)$ 上界，因此结合 d_i : $= \mathcal{O}(d_\phi^2$、$_i)$ $= \mathcal{O}((d_S + d_A$、$_i)^2)$，验证了假设 8。进一步根据标准覆盖论证，我们可以构造 $\mathcal{F}_i$、Π_i 的有限覆盖，其对数基数均为 $\widetilde{\mathcal{O}}$(poly(H)$\cdot d_\phi^2$、$_i)$。将这些结果代入定理 9，我们得到 DOPMD 能够在以下范围内学习 LQGs 的 Π-粗相关均衡

$$\widetilde{\mathcal{O}}\left(\mathrm{poly}(H, \textstyle\sum_{i \in [m]} d_{\phi,i}) / \varepsilon^4 \right)$$

博弈回合。