arXiv:2305.00684v1 [cs.LG] 1 May 2023

多智能体决策的复杂度：从博弈学习到部分监控

DylanJ. Foster DeanP.Foster NoahGolowich*
AlexanderRakhlindylanfoster@microsoft.com dean@foster.net nzg@mit.edu rakhlin@mit.edu

2023年5月1日

摘要

多智能体强化学习（MARL）理论中的一个核心问题是，理解哪些结构条件和算法原理能够带来具有样本效率的学习保证，以及随着智能体数量从少量增加到大量，这些考量会发生怎样的变化。我们在一个用于多个智能体交互式决策的一般框架下研究这一问题，该框架涵盖具有函数逼近的马尔可夫博弈，以及具有老虎机反馈的标准形博弈。我们重点关注均衡计算，其中集中式学习算法旨在通过控制与（未知）环境交互的多个智能体来计算一个均衡。我们的主要贡献如下：

- 我们基于决策估计系数的多智能体推广形式，为多智能体决策的最优样本复杂度给出了上界和下界。该复杂度度量由Foster 等人(2021)在与我们设定相对应的单智能体设定中提出。与单智能体设定的最佳结果相比，我们的上界和下界之间还存在额外差距。我们证明，任何“合理”的复杂度度量都无法弥合这些差距，凸显了单智能体与多个智能体之间显著的差异。

- 我们证明，刻画多智能体决策的统计复杂度，等价于刻画单智能体决策的统计复杂度，但奖励是隐藏的（不可观测的）。这一框架涵盖了部分监控问题的各种变体。基于这一联系，我们尽可能充分地刻画了隐奖励交互式决策的统计复杂度。

在此基础上，我们给出了若干新的结构性结果，包括：1）多智能体决策的统计复杂度可降至单智能体情形的条件；2）避免所谓多智能体诅咒的条件。

目录

1 引言 **3**

1.1 多智能体交互式决策（MA-DMSO）.4 1.2 MA-DMSO：结果概述.8 1.3 隐奖励交互式决策（HR-DMSO）.10 1.4 HR-DMSO：结果概述.12 1.5 MA-DMSO：其他结果.13 1.6 预备知识.14 1.7 组织结构.16

2 MA-DMSO 与 HR-DMSO 框架的等价性 **16**

3 极小极大速率的上界和下界 **17**

*部分工作是在微软研究院实习期间完成的。

3.1 HR-DMSO：极小极大速率的上界和下界.18 3.2 HR-DMSO：界之间的差距及紧确刻画的不可能性.19 3.3 对 MA-DMSO 框架的启示.24

4 MA-DMSO：从多智能体到单智能体 **26**
4.1 凸决策空间下 MA-DEC 的界.26 4.2 马尔可夫博弈下 MA-DEC 的界.27

5 MA-DMSO：论多个智能体的诅咒 **28**

I 示例 **38**

A MA-DMSO：实例示例 **38**
A.1 均衡的补充示例.38 A.2 实例的补充示例.39 A.3 多智能体实例的 DEC 和极小极大风险的界的计算.41

II 证明 **51**

B 技术工具 **51**
B.1 信息论.51 B.2 集中不等式.52 B.3 拓扑引理.53 B.4 极小极大定理.53

C 第 2 节的证明 53

D 第 3 节的证明 59
D.1 第 3.1 节的证明.59 D.2 第 3.2 节的证明.61 D.3 第 3.3 节的证明.67

E 第4节的证明 **69**
E.1 凸决策空间一般博弈的界.70 E.2 马尔可夫博弈的界.71

F 第 5 节上界的证明 界 7 **7**
F.1 多智能体优化式探索目标.77 F.2 算法 1 性能的界定.79 F.3 多智能体参数化信息比率.81 F.4 多智能体信息比率与探索即优化目标之间的关系 . . . . . 84 F.5 汇总所有结果：定理 5.1 的证明.86

G 第 5 节下界的证明 **87**
G.1 命题 5.1 的证明.87 G.2 定理 5.2 的证明.87 G.3 补充引理.95

1 引言

人工智能最令人振奋的前沿领域中，有许多本质上属于博弈论范畴，涉及激励各异的多个智能体在动态环境中相互作用并做出决策，彼此合作或竞争。近期，不少采用多智能体强化学习（MARL）框架的方法，已在多智能体博弈领域达到人类水平的性能（Silver 等人，2016；Brown and Sandholm，2018年；Perolat 等人，2022年；Kramár 等人，2022年；Bakhtin 等人，2022年）。尽管多智能体强化学习在网络安全（Malialis andKudenko，2015年）、自动驾驶（Shalev-Shwartz 等人，2016）和经济政策（Zheng 等人，2022年）等领域具有巨大的进一步应用潜力，但样本效率和可靠性仍是其现实世界部署的障碍。因此，一个核心问题是，哪些建模假设与算法设计原则能够带来稳健且高样本效率的学习保证。这一问题在具有高维反馈和决策空间的领域尤为突出，因为在这些领域中，使用神经网络等灵活模型至关重要。

针对单智能体设定下的强化学习，大量研究探讨了在何种建模假设（或结构条件）下能够实现样本高效学习（Russo 和 Van Roy，2013年；Jiang 等人，2017年；Sun 等人，2019年；Wang 等人，2020年；Du 等人，2021年；Jin 等人，2021a；Foster 等人，2021年）。值得注意的是，Foster 等人（2021年、2022b、2023）提出了一种统计复杂度概念——决策估计系数（DEC），它是实现低样本复杂度的充分必要条件，并为算法设计带来了统一原则。对于多智能体强化学习，实现样本高效学习所需的结构条件也受到了广泛研究（Chen 等人，2022b；Li 等人，2022年；Xie 等人，2020年；Jin 等人，2022年；Huang 等人，2021年；Zhan 等人，2022年；Liu 等人，2022年），这些研究的灵感源自单智能体场景。然而，单智能体研究的见解并不总能以直观的方式推广到多个智能体（Daskalakis 等人，2022年），相关研究的发展在很大程度上仍是针对具体问题逐案推进。因此，如何对样本高效的多智能体强化学习形成统一理解，或确定其必要条件，仍是一个悬而未决的问题。

贡献。 我们考虑一个通用框架，结构化观测下的多智能体决策（MA-DMSO），它将 Foster 等人(2021）提出的单智能体 DMSO 框架加以推广，并涵盖了采用一般函数逼近的多智能体强化学习，以及具有老虎机反馈和结构化动作空间的标准形博弈。我们聚焦于集中式均衡计算：集中式学习算法控制所有智能体，并通过与（未知）环境交互来计算均衡。我们的主要结果如下：

- **多智能体决策的复杂度**。 我们提出一种新的复杂度度量，即多智能体决策估计系数，它推广了 Foster 等人(2021、2023）提出的决策估计系数，并证明它能够给出多智能体决策最优样本复杂度的上界和下界。与单智能体设定下的最佳结果（Foster 等人，2023）相比，我们的上界和下界之间还存在额外差距，并且我们证明没有任何（合理的）复杂度度量能够弥合这一差距。

- **隐藏奖励决策的复杂度**。我们证明，刻画多智能体决策的统计复杂度，等价于刻画单智能体决策在奖励隐藏（未被观测）时的统计复杂度；我们将这一框架称为结构化观测下的隐奖励决策（HR-DMSO）。利用这一联系，我们尽可能充分地刻画了 HR-DMSO 框架的统计复杂度，该框架涵盖了随机部分监测问题（Bartók 等人，2014年）的 PAC 变体（对任何合理的复杂度度量而言）。

- **关于多个智能体的进一步见解**。在上述结果的基础上，我们为多智能体决策和强化学习提供了一系列新的结构性结果和算法见解，包括：1）多智能体决策的复杂度可降低至单智能体决策复杂度的一般条件；2）移除所谓多智能体诅咒（Jin 等人，2021b）的一般条件。

我们的研究结果为进一步建立多智能体强化学习的统一理解奠定了基础。

学习与决策，并着重介绍若干令人振奋的开放问题。

1.1 多智能体交互式决策（MA-DMSO）

我们介绍了一个 结构化观测决策 框架的多智能体推广版本，该框架由 Foster 等人 (2021提出，我们称之为多智能体结构化观测决策(MA-DMSO)。该框架由以下部分构成： T 轮智能体与环境之间的交互。对于每个轮次 K ，智能体和环境。在每个轮次 $t = 1, 2, \ldots\ldots, T$:

1. The 智能体共同选择联合决策 $\pi^t \in \Pi$，其中 Π 是联合决策空间 .
2. 每个智能体 $k \in [K]$ 都会收到一个奖励 $r_k^t \in \mathcal{R} \subseteq \mathbb{R}$ 和一个纯观测 $o_\circ^t \in \mathcal{O}_\circ$ ，它们根据 $(r_1^t, \ldots, r_K^t, o_\circ^t) \sim M^\star(\pi^t)$ 采样，其中 $M^\star: \Pi \to \Delta(\mathcal{R}^K \times \mathcal{O}_\circ)$ 是底层模型。我们将 $\mathcal{R}$ 称为奖励空间，将 $\mathcal{O}_\circ$ 称为纯观测空间。我们将包含第 t 轮中向智能体揭示的所有信息的元组 $(r_1^t, \ldots, r_K^t, o_\circ^t)$ 称为完整观测。

在经过 T 轮交互后，智能体共同输出联合决策 $\widehat{\pi} \in \Pi$，该决策可以任意地根据 T 轮中观测到的数据来选择，也可以按照分布 $p \in \Delta(\Pi)$进行随机化。其目标（我们将在下文形式化说明）是选择 $\widehat{\pi}$ ，使其成为由 $M^\star$所诱导的平均奖励函数的均衡（例如纳什均衡或粗相关均衡）。模型 $M^\star$被形式化为从决策到完整观测的概率核（第 1.6 节），智能体对此一无所知，应将其理解为底层环境。

DMSO 框架涵盖了大多数由单个智能体与未知环境交互的在线决策问题，而 MA-DMSO 框架则进一步将其推广，以涵盖多智能体强化学习中的各种问题。例如，具有老虎机反馈的标准形博弈学习（Rakhlin 和 Sridharan，2013年；Foster 等人，2016；Heliou 等人，2017年；Wei 和 Luo，2018；Giannou 等人，2021），其中 $M^\star$ 表示博弈中每个条目对应的奖励分布；以及采用函数逼近的马尔可夫博弈学习（Chen 等人，2022b；Li 等人，2022；Xie 等人，2020年；Jin 等人，2022年；Huang 等人，2021；Zhan 等人，2022年；Liu 等人，2022），其中 $M^\star$ 表示底层的马尔可夫博弈。其他例子包括具有结构化（例如凸凹）奖励和高维动作空间的标准形博弈（Bravo 等人，2018；Maheshwari 等人，2022；Lin 等人，2021）。

可实现性性。 虽然模型 $M^\star$ 是未知的，但我们做出标准的可实现性 假设 .

假设 1.1 （MA-DMSO 的可实现性）。 智能体可以访问一个模型类 $\mathcal{M}$ ，该模型类由概率核 $M: \Pi \to \Delta(\mathcal{R}^K \times \mathcal{O}_\circ)$ 组成，并且包含真实模型 $M^\star$。

对于标准形博弈，类 $\mathcal{M}$ 在奖励（例如线性或凸性）或决策空间中编码结构；对于马尔可夫博弈，它在转移概率或价值函数中编码结构。示例请参见附录的 第一部分，以及 Foster 等人 （2021）关于单智能体情形的研究，其中 $K = 1$。

1.1.1 均衡

MA-DMSO 框架中智能体的目标是为底层博弈/模型$M^\star$生成一个均衡。我们以一种涵盖若干标准博弈论均衡的普遍方式，将均衡的概念形式化。为使记号简洁，定义 $\mathcal{O} := \mathcal{R}^K \times \mathcal{O}_\circ$ 为完整观察空间，并记作 $o^t := (r_1^t、\ldots\ldots、r_K^t、o_\circ^t)$来表示（完整）观测。对于 $M \in \mathcal{M}$和 $\pi \in \Pi$，令 $\mathbb{E}^{M,\pi}[\cdot]$表示过程（$r_1$、……、$r_K$、$o_\circ$）$\sim M(\pi)$ 下的期望；根据我们的记号 $\mathcal{O} = \mathcal{R}^K \times \mathcal{O}_\circ$，有时也将该过程记作 $o \sim M(\pi)$ 。对于每个 $k \in [K]$和 $M \in \mathcal{M}$，定义映射 $f_k^M: \Pi \to \mathbb{R}$为 $f_k^M(\pi) = \mathbb{E}^{M,\pi}[r_k]$，其表示在采取联合决策 π 时，智能体 k在 M 下的期望奖励。

对于每个智能体 k，我们假设其给定一个偏离空间 Π_k'以及一个切换函数,$U_k: \Pi_k' \times \Pi \to \Pi$。给定联合决策 $\pi \in \Pi$，每个智能体 k 都可以选择一个偏差 $\pi_k' \in \Pi_k'$，其效果是，智能体执行的联合策略变为 $U_k(\pi_k', \pi)$，而不是 π。我们的目标是使 MA-DMSO 设置所产生的输出策略$\widehat{\pi} \sim p$ 具备这样的性质：没有任何智能体能够通过偏差显著提高其

取值。我们通过以下方式对此进行量化：

$$\mathbf{Risk}(T) := \mathbb{E}_{\widehat{\pi}\sim p}\left[\sum_{k=1}^{K} \sup_{\pi'_k \in \Pi'_k} f_k^{M^\star}(U_k(\pi'_k, \pi)) - f_k^{M^\star}(\pi)\right]. \tag{1}$$

对于 $M \in \mathcal{M}$和 $\pi \in \Pi$，我们将 $h^M(\pi) = \sum_{k=1}^{K} \sup_{\pi'_k \in \Pi'_k} f_k^M(U_k(\pi'_k, \pi)) - f_k^M(\pi)$ 简记为，因此 **风险**(T) :$=\mathbb{E}_{\widehat{\pi}\sim p}[h^{M^\star}(\widehat{\pi})]$。量 $h^M(\pi)$ 衡量玩家偏离联合决策π在 M下的激励之和；如果 π对于 M满足 $h^M(\pi) = 0$，我们称其为一个均衡 。

概念 (1) 涵盖了均衡的标准概念，包括纳什均衡、相关均衡（CE）和粗相关均衡（CCE）。由于我们力求使本节中的设定尽可能一般化，因此我们作出两个正则性假设，以排除其他可能存在病态情形的均衡概念。第一个假设认为均衡存在，第二个断言每个智能体始终可以选择一个不会降低其取值的偏差。

假设 1.2 (均衡的存在性)。 对于任意模型 $M \in \mathcal{M}$，都存在 $\pi \in \Pi$，使得 $h^M(\pi) = 0$.

假设 1.3 (最优偏离的单调性)。 对于任意模型 $M \in \mathcal{M}$、智能体 $k \in [K]$和联合决策 $\pi \in \Pi$，都存在某个偏差 $\pi'_k \in \Pi'_k$，使得 $f_k^M(U_k(\pi'_k, \pi)) \geq f_k^M(\pi)$。

假设 1.3 意味着，风险的概念 K，就 (1) 而言，等价于任何智能体通过偏离所能获得的最大收益。这两个假设均适用于纳什均衡、CE 和 CCE（参见定义 1.1、1.2和 A.1）。

总之，MA-DMSO 框架刻画了均衡计算问题：智能体旨在找到一个（ε-近似）均衡 $\widehat{\pi}$，使得 **风险**$(T) \leq \varepsilon$，但底层博弈未知，因此它们必须通过与之交互并进行探索来收集信息。我们将元组 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 称为 MA-DMSO 框架的一个实例 。实例 $\mathscr{M}$ 规定了学习过程开始前智能体先验已知的所有信息。

备注 1.1。 如前所述，MA-DMSO 框架允许采用集中式学习协议，即单个学习算法可以集中控制所有智能体（等价地，在整个学习过程中，智能体之间可以进行不受限制的通信与协调）。针对集中式学习算法的下界当然也适用于去中心化算法，因为后者是前者的特例。不过，一般而言，集中式算法与去中心化算法的极小极大样本复杂度之间可能存在差距；我们将对分布式多智能体交互式决策的详细研究留待未来工作。

备注 1.2。 我们对 MA-DMSO 框架的介绍涵盖了这样一些设定：在这些设定中，（多智能体）学习算法仅根据输出决策 $\widehat{\pi}$ 与均衡的接近程度进行评估，而不是例如根据学习 T 轮次中所采取决策与均衡的平均接近程度进行评估。在单智能体设定中，这类保证通常称为 PAC（概率近似正确）保证，以区别于遗憾保证（Foster 等人，2023）。将我们的许多结果扩展到遗憾设定相当直接。

1.1.2 MA-DMSO 实例示例

现在，我们重点介绍 MA-DMSO 框架所涵盖的基本多智能体老虎机和多智能体强化学习问题。我们将介绍决策空间、偏离空间和切换函数的结构，这些结构使我们能够刻画具体的均衡概念，然后给出一些实例示例 $\mathscr{M}$。

均衡示例。在定义 1.1 和1.2 中，我们将具体说明可用于刻画纳什均衡和粗相关均衡（*CCE*）的决策空间、偏离空间和切换函数；更多示例请参见附录 A.1，其中包括相关均衡（*CE*）以及在马尔可夫博弈背景下研究过的 CCE 变体。

定义 1.1 （纳什均衡实例）。 一个 *MA-DMSO* 实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 是一个纳什均衡（NE）实例，如果满足以下条件：

1. 对于集合 $\Pi_1, \ldots, \Pi_K$，有 $\Pi = \Pi_1 \times \cdots \times \Pi_K$。*2.* 对于每个 $k \in [K]$，有 $\Pi'_k = \Pi_k$。*3.* 对于每个 $k \in [K]$，$\pi \in \Pi$ 和 $\pi'_k \in \Pi'_k$，都有 $U_k(\pi'_k, \pi) = (\pi'_k, \pi_{-k})$。[1]

我们称 *NE* 实例 $\mathscr{M}$ 为双人零和纳什均衡实例，如果 $K = 2$，并且对所有 $M \in \mathcal{M}$、$\pi \in \Pi$，均有 $f_1^M(\pi) + f_2^M(\pi) = 0$。

定义 1.1 中的纳什均衡概念包含了标准形博弈中混合纳什均衡的标准概念（例如（Nisan 等人，2007）），但其范围远不止于此。特别是，定义 1.1 并不假设决策空间 Π_k 是玩家 k的纯行动空间上的分布。因此，它涵盖了更精细的解概念，包括标准形博弈中的纯纳什均衡（Daskalakis 和 Papadimitriou，2006）以及马尔可夫博弈中的马尔可夫纳什均衡（例1.2）。由于这种一般性，根据定义 1.1 的 NE 实例并不保证满足假设 1.2，即不保证存在均衡；不过，我们将确保所考虑的所有 NE 实例均以满足假设 1.2 的方式构造。

定义 1.2 给出了定义 1.1 的一个类比，可以刻画（标准形式）粗相关均衡的概念。

定义 1.2（粗相关均衡实例）。如果 *MA-DMSO* 的一个实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 满足以下条件，则称其为粗相关均衡（CCE）实例：

1. 对于某些集合 $\Sigma_1, \ldots, \Sigma_K$（称为纯决策），有 $\Pi = \Delta(\Sigma_1 \times \cdots \times \Sigma_K)$。我们将写作 $\Sigma := \Sigma_1 \times \cdots \times \Sigma_K$。*2.* 对于每个 $\pi \in \Pi$ 和 $M \in \mathcal{M}$，都有 $M(\pi) = \mathbb{E}_{\sigma \sim \pi}[M(\sigma)]$。此外，存在一个可测函数 $\varphi : \mathcal{O} \to \Sigma$，使得对于每个 $M \in \mathcal{M}$ 和 $\sigma \in \Sigma$，都有 $\mathbb{P}_{o \sim M(\sigma)}(\varphi(o) = \sigma) = 1$（即，$M(\sigma)$ 揭示了 σ）。*3.* 对于每个 $k \in [K]$，都有 $\Pi'_k = \Sigma_k \cup \{\perp\}$。*4.* 对于每个 $k \in [K]$、$\pi \in \Pi$ 和 $\pi'_k \in \Pi'_k$，都有

$$U_k(\pi'_k, \pi) = \begin{cases} \mathbb{I}_{\pi'_k} \times \pi_{-k} & : \pi'_k \neq \perp \\ \pi & : \pi'_k = \perp \end{cases},$$

其中 $\mathbb{I}_{\pi'_k} \times \pi_{-k} \in \Pi$ 表示这样的乘积分布：智能体 k 采取 π'_k，而其他智能体则根据 其在 $\pi \in \Pi$ 下的联合边缘分布采取动作。

在定义 1.2中，加入 $\perp \in \Pi'_k$ 对应于玩家 k 选择不偏离。这对于满足假设 1.3是必要的，因为可能存在分布 $\pi \in \Pi$，使得如果玩家 k 偏离到 Σ_k中的任意一个固定选项，其取值都会下降。[2] 我们还指出，定义 1.2刻画了标准形博弈中的粗相关均衡概念（纯动作集合为 Σ_k）；在附录 A.1中，我们给出一个实例，刻画马尔可夫博弈中略有不同的粗相关均衡概念。

注释 1.3。我们在全文中（包括上述定义的条目 2）采用以下约定：为方便起见，我们将任意单点分布与该分布赋予其质量的元素对应起来。例如，对于 $\sigma = (\sigma_1, \ldots, \sigma_K) \in \Sigma_1 \times \cdots \times \Sigma_K$ 这一定义 1.2中的纯决策，我们将其对应的单点分布 $\mathbb{I}_\sigma \in \Delta(\Sigma) = \Pi$ 记作 $\sigma \in \Pi$。此外，在可能的情况下，我们约定 Σ 表示纯决策集，而 Π 表示可能为纯决策集或混合决策集的决策集（具体含义将由上下文明确）。

[1] 我们采用如下约定：$\pi_{-k} = (\pi_1, \ldots, \pi_{k-1}, \pi_{k+1}, \ldots)$ 和 $(\pi_k, \pi_{-k}) = (\pi_1, \ldots, \pi_k, \ldots, \pi_K)$。[2] 在某些语境中，粗相关均衡的定义不包含此类选项 $\perp \in \Pi'_k$；在目标仅为建立上界的情形下，加入 $\perp$ 不会产生实质影响（因为它唯一的作用是确保决策的次优性非负）。但由于我们还要证明下界，因此保留选项 $\perp \in \Pi'_k$至关重要。

均衡示例。 现在我们将为定义 1.1和1.2中的纳什均衡和粗相关均衡实例提供具体示例；有关更多示例（包括相关均衡）及讨论，请参见附录 A 。

例 1.1 （在标准形博弈中学习纳什均衡和粗相关均衡）。 我们首先描述具有老虎机反馈的标准形博弈中的学习问题。假设每个玩家 $k \in [K]$ 都有一个有限的行动集 $\mathcal{A}_k$，联合行动集记为 $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$。在采取联合动作配置 $a \in \mathcal{A}$后，真实模型（未知） $M^\star$对 $(r_1, \ldots\ldots, r_K) \sim M^\star(a)$进行采样，其中 r_k表示玩家 k获得的奖励。目标是计算一个关于联合动作配置的分布，使其成为该博弈的某种均衡，其中博弈的收益由 $M^\star$ 下的期望奖励给出。下面我们正式描述与纳什均衡和粗相关均衡计算问题相对应的 MA-DMSO 实例：

- 为表达纳什均衡计算问题，对于每个 k，令 $\Pi_k := \Delta(\mathcal{A}_k)$，令 $\Pi = \Pi_1 \times \cdots \Pi_K$表示 $\mathcal{A}$上的乘积分布空间，并按照定义 1.1 定义 Π'_k、U_k。此外，令 $\mathcal{R} = [0,1]$和 $\mathcal{O}_\circ = \mathcal{A}$、$\mathcal{O} = \mathcal{R}^K \times \mathcal{O}_\circ$。令 $\mathcal{M}$ 为满足以下条件的模型类：（a）对于所有单点分布$\mathbb{I}_a = \mathbb{I}_{a_1} \times \cdots \times \mathbb{I}_{a_K} \in \Pi$， $M(\mathbb{I}_a) \in \Delta(\mathcal{R}^K) \times \{\mathbb{I}_a\}$；（b）对于所有 $\pi \in \Pi$， $M(\pi) = \mathbb{E}_{a\sim\pi}[M(\mathbb{I}_a)]$。换言之，$M(\pi)$对动作配置 $a \sim \pi$ 进行采样（特别地，对于每个 k都有 $a_k \sim \pi_k$），揭示所采样的动作配置 a、[3] 以及从任意分布中抽取的 K $[0,1]$值奖励。于是，根据定义 1.1，实例$\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$是一个纳什均衡实例。对于 $\widehat{\pi} \in \Pi$， $h^{M^\star}(\widehat{\pi})$度量玩家在真实模型 $M^\star$下偏离 $\widehat{\pi}$ 的激励之和；特别地，当且仅当 $h^{M^\star}(\widehat{\pi}) = 0$ 时，$\widehat{\pi}$ 是一个博弈的纳什均衡，该博弈的收益函数由 $a \mapsto f_k^{M^\star}(a) := \mathbb{E}^{M^\star, a}[r_k]$给出。

- 为表述粗相关均衡计算问题，令 $\Pi = \Delta(\mathcal{A}_1 \times \cdots \times \mathcal{A}_K)$，并按照定义 1.2 定义 Π'_k, U_k，其中 $\Sigma_k = \mathcal{A}_k$ 对于每个 k。此外，令 $\mathcal{R} = [0,1]$，且 $\mathcal{O}_\circ = \mathcal{A}$, $\mathcal{O} = \mathcal{R}^K \times \mathcal{O}_\circ$。令 $\mathcal{M}$ 为满足以下条件的模型类：(a) 对于所有单点分布 $\mathbb{I}_a \in \Pi$, $M(\mathbb{I}_a) \in \Delta(\mathcal{R}^K) \times \{\mathbb{I}_a\}$，以及 (b)，对于 $\pi \in \Pi$, $M(\pi) = \mathbb{E}_{a\sim\pi}[M(\mathbb{I}_a)]$。那么，实例 $\mathscr{M} := (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 根据定义 1.2 是一个 CCE 实例。对于 $\widehat{\pi} \in \Pi$, $h^{M^\star}(\widehat{\pi})$ 衡量在真实模型 $M^\star$ 下，玩家偏离 $\widehat{\pi}$ 的非负偏离激励之和；特别地， $h^{M^\star}(\widehat{\pi}) = 0$ 当且仅当 $\widehat{\pi}$ 是一个博弈的粗相关均衡，该博弈的收益函数由 $a \mapsto f_k^{M^\star}(a) := \mathbb{E}^{M^\star, a}[r_k]$给出。

- MA-DMSO 框架还可以描述相关均衡计算问题。我们使用与上一点所述相同的 CCE 实例，但对 Π'_k、U_k 的定义略有不同；参见定义 A.1（见附录 A）。

在标准形博弈设定最基本的（“有限动作”）版本中，我们允许 $M^\star(a)$ 任意取值，但需满足约束 $r_k \in [0,1]$；同时假设 $A_k : = |\mathcal{A}_k| < \infty$ 对于所有 k。除了有限动作标准形博弈之外，MA-DMSO 框架还涵盖带有老虎机反馈的结构化标准形博弈（等价地说，即结构化老虎机问题的多智能体变体）。在这类博弈中，玩家的动作空间很大或无限，但奖励具有额外结构。示例包括线性、凸或凹收益（对老虎机凸优化的推广）（Bravo 等人，2018年；Maheshwari 等人，2022年；Lin 等人，2021），以及许多其他情形（Cui 等人，2022）。◁

例1.2 （在马尔可夫博弈中学习纳什均衡）。 接下来，我们考虑一种分幕式多智能体有限时域强化学习设定，其中未知的真实模型 $M^\star$ 是一个马尔可夫博弈。我们专注于计算马尔可夫纳什均衡的问题；关于计算粗相关均衡（CCE）和相关均衡（CE）的变体问题，将在附录 A.1中讨论。

形式上，每个模型 $M \in \mathcal{M}$ 都定义了如下形式的马尔可夫博弈： $M = (H, \{\mathcal{S}_h\}_{h\in[H]}, \{\mathcal{A}_k\}_{k\in[K]}, \{P_h^M\}_{h\in[H]}, \{R_{k,h}^M\}_{k\in[K],\ h\in[H]}, d_1)$，其中 $H \in \mathbb{N}$ 表示时域， $\mathcal{S}_h$ 表示第 h 层的状态空间， $\mathcal{A}_k$ 表示玩家 k 的动作空间， $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$ 表示联合动作空间， P_h^M: $\mathcal{S}_h \times \mathcal{A} \to \Delta(\mathcal{S}_{h+1})$表示第 h 层的概率转移核， $R_{k,h}^M$: $\mathcal{S}_h \times \mathcal{A} \to \Delta(\mathbb{R})$表示玩家 k 在第 h 层的奖励分布， $d_1 \in \Delta(\mathcal{S}_1)$表示初始状态分布。允许模型集合 $\mathcal{M}$ 中不同模型的转移核和奖励分布有所不同，但我们假设所有模型 $\mathcal{M}$ 的状态和动作空间、时域以及初始状态分布均相同。

每个智能体的决策空间 Π_k 是其 随机马尔可夫策略 $\pi_k = (\pi_{k,1}, \ldots, \pi_{k,H})$ 的空间，其中$\pi_{k,h}$： $\mathcal{S}_h \to \Delta(\mathcal{A}_k)$，联合决策空间为 $\Pi = \Pi_1 \times \cdots \times \Pi_K$。给定一个联合决策 $\pi \in \Pi$，一个

[3]出于技术原因，我们假设模型会揭示所采取的动作配置（参见假设 4.1）；这是一个非常宽松的假设，在几乎所有（集中式）场景中都成立，因为智能体知道自己采取了哪个动作。

观测根据以下过程从 $M(\pi)$ 中抽取，该过程称为一个 回合。首先，根据 $s_1 \sim d_1$ 抽取一个初始状态。然后，对于 $h \in [H]$，依次对以下随机变量进行采样：

- 对所有 $k \in [K]$, $a_{k,\,h} \sim \pi_{k,\,h}(s_h)$，以及 $r_{k,\,h} \sim R_{k,\,h}^{M}(s_h,\,(a_{1,\,h},\,\ldots,\,a_{K,\,h}))$。
- $s_{h+1} \sim P_h^M(\cdot|s_h,\,(a_{1,\,h},\,\ldots,\,a_{K,\,h}))$。

序列 $\tau = \{(s_h,\,(a_{1,\,h},\,\ldots,\,a_{K,\,h}),\,(r_{1,\,h},\,\ldots,\,r_{K,\,h})\}_{h\in[H]}$ 包含所有状态、动作和奖励，称为轨迹。$(r_1,\,\ldots,\,r_K,\,o_\circ) \sim M(\pi)$ 的分布由 $o_\circ = \tau$ 和 $r_k = \sum_{h=1}^{H} r_{k,\,h}$ 给出。特别地，纯观测空间 $\mathcal{O}_\circ$ 就是轨迹空间。我们假设 $\sum_{h=1}^{H} r_{k,\,h} \in [0,1]$ 以概率 1 成立，这意味着 $\mathcal{R} = [0,1]$，并记作 $\mathcal{O} = \mathcal{R}^K \times \mathcal{O}_\circ$。

令 Π'_k, U_k 如定义 1.1 中所定义。则实例 $\mathscr{M} := (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 是 MA-DMSO 的纳什均衡实例。对于 $\widehat{\pi} \in \Pi$，取值 $h^{M^\star}(\widehat{\pi})$ 衡量了在真实模型 $M^\star$ 下，玩家偏离 $\widehat{\pi}$ 的偏离激励之和，其中每个智能体都可以选择任意非平稳马尔可夫策略作为其偏差。特别地，$h^{M^\star}(\widehat{\pi}) = 0$ 当且仅当 $\widehat{\pi}$ 是 $M^\star$ 的马尔可夫纳什均衡（例如，Daskalakis 等人（2022））。

在（多智能体）在线强化学习中，一个关键问题是理解模型类 $\mathcal{M}$ 的哪些结构属性能够实现高效可学习性。在最简单的情形（称为表格情形）下，状态和动作空间 $\mathcal{S}_h$、$\mathcal{A}$ 都是有限的，而 $\mathcal{M}$ 由所有具有任意转移 P_h^M 和收益分布 $R_{k,\,h}^M$ 的模型组成，且这些分布的支撑集均有界。通过限制 $\mathcal{M}$，我们的表述还涵盖了包含函数逼近的更复杂设置（Chen 等人，2022b；Li 等人，2022；Xie 等人，2020；Jin 等人，2022；Huang 等人，2021；Zhan 等人，2022；Liu 等人，2022）；参见附录 A。

◁

我们参见附录 A，其中提供了更多示例和说明。

1.2 MA-DMSO：结果概述

我们利用一种新的复杂度度量——多智能体决策估计系数，给出了 MA-DMSO 框架的极小极大样本复杂度的上界和下界。该度量推广了约束决策估计系数，后者由Foster 等人（2023）在单智能体场景下提出。

多智能体决策估计系数。对于具有共同支配测度 ν 的概率测度 $\mathbb{P}$ 和 $\mathbb{Q}$，平方海林格距离定义为

$$D^2_{\mathsf{H}}(\mathbb{P},\mathbb{Q}) = \int \left(\sqrt{\frac{d\mathbb{P}}{d\nu}} - \sqrt{\frac{d\mathbb{Q}}{d\nu}}\right)^2 d\nu.$$

考虑 MA-DMSO 框架中的一个实例 $\mathscr{M} = (\mathcal{M},\ \Pi,\ \mathcal{O},\ \{\Pi'_k\}_k,\ \{U_k\}_k)$，以及一个参考模型 $M : \Pi \to \Delta(\mathcal{O})$。[4] 对于尺度参数 $\varepsilon > 0$，该实例 $\mathscr{M}$ 在参考模型 M 和尺度 ε 下的多智能体决策估计系数定义为

$$\mathsf{dec}_\varepsilon(\mathscr{M}, \overline{M}) := \inf_{p,q\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \left\{\mathbb{E}_{\pi\sim p}[h^M(\pi)] \mid \mathbb{E}_{\pi\sim q}[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)] \le \varepsilon^2\right\}; \tag{2}$$

当集合为空时，我们约定$\mathsf{dec}_\varepsilon(\mathscr{M},\ M) = 0$。

$$\mathcal{H}_{q,\varepsilon}(\overline{M}) := \{M \in \mathcal{M} \mid \mathbb{E}_{\pi\sim q}[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)] \le \varepsilon^2\} \tag{3}$$

此外，我们定义

$$\mathsf{dec}_\varepsilon(\mathscr{M}) := \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_\varepsilon(\mathscr{M}, \overline{M}), \tag{4}$$

其中，$\mathrm{co}(\mathcal{M})$ 表示类 $\mathcal{M}$ 的凸包。

[4] 参考模型M 可以是任意的，不要求属于 $\mathcal{M}$。

定义 (2) 是一个极小极大博弈，其含义如下。极大化玩家选取的模型 $M \in \mathcal{M}$ 代表对底层模型的最坏情况选择。极小化玩家选取的联合分布 $p, q \in \Delta(\Pi)$ 代表控制所有智能体的集中式学习算法所采用的策略。分布 $q \in \Delta(\Pi)$ 是探索分布，作为获取信息的策略；数量 $\mathbb{E}_{\pi\sim q}[D^2_{\mathrm{H}}(M(\pi)\ M(\pi))]$ 则代表其平均"信息增益"（即能够区分 $M \in \mathcal{M}$ 与参考模型 M 的信息量）。分布 $p \in \Delta(\Pi)$ 是利用分布，旨在使模型 $M \in \mathcal{M}$ 的策略接近均衡；其中，$\mathbb{E}_{\pi\sim p}[h^M(\pi)]$ 表示与均衡的距离。总而言之，对于 p 的最佳情况选择，取值 (2) 捕捉了 p 在模型 $M \in \mathcal{M}$ 下距离均衡的最坏情况；这些模型在 M 的意义上与之"接近"，即它们在 q 下的信息增益很小。

—

对于熟悉相关工作的读者，我们回顾一下：（单智能体）受约束 DEC 推广了更早的 偏移 DEC（由Foster 等人 (2021) 提出，它相当于一种拉格朗日松弛），并且总能给出更紧的保证（Foster 等人，2023）。我们的定义 (2) 推广了 Foster 等人(2023) 中受约束 DEC 所谓的 PAC 变体，而非遗憾 变体；后者仅限于 $p = q$。

主要结果。 我们的第一个结果基于多智能体决策估计系数，为 MA-DMSO 框架的极小极大样本复杂度给出了上界和下界。为以最简单的形式陈述该结果，我们假设 $|\mathcal{M}| < \infty$；更一般的结果见 第 3 节。

定理 1.1 （推论 3.1 和3.2的非正式表述）。对于 *MA-DMSO* 框架的任意实例 $\mathscr{M} = (\mathcal{M},\ \Pi,\ \mathcal{O},\ \{\Pi'_k\}_k,\ \{U_k\}_k)$ 以及 $T \in \mathbb{N}$:

- 上界：在 假设 *1.1* 下，存在一种算法能够达到

$$\mathbb{E}[\mathbf{Risk}(T)] \le \widetilde{O}(1) \cdot \mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{M}), \quad \text{where} \quad \bar{\varepsilon}(T) \le \widetilde{\Theta}\big(\sqrt{\log|\mathcal{M}|/T}\big). \tag{5}$$

- 下界：对于最坏情况模型 $M \in \mathcal{M}$，任何算法都必须满足

$$\mathbb{E}[\mathbf{Risk}(T)] \ge \widetilde{\Omega}(1) \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{M}), \quad \text{where} \quad \underline{\varepsilon}(T) \ \text{ solves } \ \mathsf{dec}_{\varepsilon}(\mathscr{M}) \ge \widetilde{\Omega}\big(\varepsilon^2 KT\big). \tag{6}$$

该结果表明，MA-DEC 是 MA-DMSO 框架中均衡计算的一项基本限制，并且只要 $\log|\mathcal{M}| < \infty$，就足以保证较低的样本复杂度。上界直接推论自 Foster 等人 (2023 在单智能体设定下给出的一个上界，而下界则需要采用一种新方法；这是因为单个智能体与多个智能体之间存在根本差异，我们将在下文重点说明。

为了建立直观认识，我们先来看一个基本例子。假设 $\mathscr{M}$ 是一个 CCE 实例，由两个玩家参与的 $A_1 \times A_2$ 标准形博弈组成（即 $|\mathcal{A}_1| = A_1$ 和 $|\mathcal{A}_2| = A_2$），具有老虎机反馈（例 1.1）和伯努利噪声。在这种情况下，可以证明 $\mathsf{dec}_\varepsilon(\mathscr{M}) \propto \varepsilon \cdot \sqrt{A_1 + A_2}$，因此上界 (5)给出

$$\mathbb{E}[\mathbf{Risk}(T)] \lesssim \sqrt{\frac{(A_1 + A_2)\log|\mathcal{M}|}{T}},$$

或者等价地，$\frac{(A_1+A_2)\log|\mathcal{M}|}{\varepsilon^2}$ 交互轮次足以找到一个 ε-粗相关均衡。对于此类，可以~取$\log|\mathcal{M}| \lesssim O(A_1 \cdot A_2)$。我们给出更精细的结果（第 5 节），允许将$\log|\mathcal{M}|$ 替换为$\max_k \log|\Pi'_k| \lesssim \log(A_1 + A_2)$，从而达到样本复杂度 $\widetilde{O}\big(\frac{A_1+A_2}{\varepsilon^2}\big)$，这是最优的。

转向下界，对于相同的标准形博弈实例 $\mathscr{M}$，可以选择 $\varepsilon(T) \gtrsim \sqrt{\frac{A_1+A_2}{T}}$，使得 (6)给出

$$\mathbb{E}[\mathbf{Risk}(T)] \gtrsim \frac{A_1 + A_2}{T},$$

或者等价地，找到一个 ε-CCE 需要 $\widetilde{\Omega}\big(\frac{A_1+A_2}{\varepsilon}\big)$ 个交互轮次。比较上界和下界，可以发现两处差距。第一处是上界中出现的 $\log|\mathcal{M}|$ 项，它表示

使用类别 $\mathcal{M}$进行统计估计所需的样本复杂度；一般而言，它会随着智能体数量的增加而迅速增大。这一点可以加以改进（参见 第 5 节），但一般无法将其完全消除，即使在单智能体设定下也是如此；进一步讨论请参见 Foster 等人(2021, 2023) 以及 第 3 节 。

第二个差距是上界和下界中出现的取值 $\varepsilon(T)$ 与 $\varepsilon(T)$ 之间的差异；我们令 $\varepsilon(T) \propto 1/\sqrt{T}$，而 $\varepsilon(T)$ 则选取为满足不动点方程 $\overline{\mathrm{dec}}_\varepsilon(\mathscr{M})_- \geq \widetilde{\Omega}\left(\varepsilon^2 T\right)$ 的解（本讨论聚焦于常数 K 的情形）。对于标准形博弈，这会使下界随 $\frac{1}{\varepsilon}$ 而非 $\frac{1}{\varepsilon^2}$ 缩放。在单智能体设定中则不存在这种差距（Foster 等，2023）：基于受约束 DEC 得到的最佳上界和下界满足 $\varepsilon(T) \approx \varepsilon(T)$（至多相差 $\log|\mathcal{M}|$ 的依赖）。我们证明（命题 3.1），对于大多数参数范围，

$$\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{M}) \lesssim \left(K^2 \log|\mathcal{M}| \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{M})\right)^{1/2},$$

即，上界与下界之间的差距通常至多为二次。这种差距是根本性的：我们证明（命题 3.2 和 3.3）存在一些实例，使得每个界（上界和下界）都是紧的；更令人意外的是，以下结果表明，没有任何复杂度度量能够在满足相当一般条件的同时，完全刻画多智能体决策的样本复杂度 $\widetilde{}$并将差距缩小到二次以内，即使 $\log|\mathcal{M}| = O(1)$。

定理 1.2 (定理 3.4的非正式版本。对于任意 $\varepsilon \in \mathbb{N}$，都存在两个双人零和纳什均衡 *MA-DMSO* 实例 $\mathscr{M}_1 = (\mathcal{M}_1, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 和 $\mathscr{M}_2 = (\mathcal{M}_2, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 以及一一映射 $\mathscr{E}: \mathcal{M}_1 \to \mathcal{M}_2$，满足：

1. 对所有 $M \in \mathcal{M}_1$、$f_k^M \equiv f_k^{\mathscr{E}(M)}$，以及所有 $k \in [2]$。
2. 对于所有 M、$M' \in \mathcal{M}_1$，以及所有 $\pi \in \Pi$，$D^2_{\mathsf{H}}(M(\pi), M'(\pi)) = D^2_{\mathsf{H}}(\mathscr{E}(M)(\pi), \mathscr{E}(M')(\pi))$。
3. 存在一种算法，可以使用 $O\left(\frac{1}{\varepsilon}\right)$ 轮次，在任何模型中求得一个 ε-纳什均衡；然而，对于 $\mathscr{M}_2$ 中的最坏情况模型，任何算法都需要 $\widetilde{\Omega}\left(\frac{1}{\varepsilon^2}\right)$ 轮次才能求得一个 ε-纳什均衡。

$\widetilde{}$此外，对数$|\mathcal{M}_1| =$ 对数$|\mathcal{M}_2| = O(1)$。

非正式地说，该结果表明，如果某个复杂度度量仅通过以下两者依赖于实例：1）奖励函数；2）模型之间的成对海灵格距离，且适用于 $\mathscr{M}$，那么除了前文中的差距之外，它无法刻画每个实例的最优样本复杂度。此外，完整结果并不局限于海林格距离，而是适用于一般的 f-散度，包括 KL-和 χ^2-散度。这排除了基于 DEC 的各种变体所能给出的更紧保证，也排除了大多数用于交互式决策的一般用途复杂度度量；详见第 3.2.2节。[5]

定理 1.2 （以及 命题 3.2 和 3.3）凸显了单智能体框架与多智能体框架之间的根本区别。在单智能体设定中，只要 $\widetilde{}$，受约束 DEC 就能在对数 ε因子范围内刻画学习一个 ε-最优决策所需的最优样本数（$\log|\mathcal{M}| = O(1)$(Foster 等人，2023）。对于两个或更多智能体，定理 1.2 和命题 3.2 和3.3 排除了这种刻画。

1.3 隐奖励交互式决策（HR-DMSO）

为了证明前文中的结果，我们确立了 MA-DMSO 框架与另一种单智能体设定之间的某种等价关系，我们将其称为结构化观测下的隐奖励决策(HR-DMSO)。该设定通过允许奖励对智能体隐藏，对单智能体DMSO框架（MA-DMSO，取 $K = 1$）进行了推广。这个设定本身也很有研究价值，可以将其视为部分监控问题的一种随机 PAC 变体（Bartók 等人，2014年）。下文将先介绍该框架，再说明：1）通过简单的归约，可以将 MA-DMSO 视为 HR-DMSO 框架的一个特例；2）反过来也成立，从而表明二者在某种意义上等价。随后，我们将讨论这两个框架的极小极大速率所受的影响。

[5]直接应用　　应用 定理 1.2 到受约束 DEC 会因 $M \in \mathrm{co}(\mathcal{M})$而产生复杂性；参见附录 3.2.2。

形式化地, HR-DMSO 框架分为 T轮次，其中对于每个轮次 $t = 1$，2，.....，T :

1. 学习者选择一个动作 $\pi^t \in \Pi$，其中 Π 是决策空间</style}，并获得（但无法观测）奖励 ()。2. 学习者接收一个通过 ()采样得到的观测 ，其中 : () 是底层模型。我们将 $\mathcal{O}$ 称为观测空间。

该过程结束后，学习者利用整个 T 轮交互过程中收集的数据生成输出决策 $\hat{\pi} \in \Pi$，该决策可能根据分布 $p \in \Delta(\Pi)$进行随机化。学习者的目标是选择决策 $\hat{\pi}$ ，以最大化其（不可观测的）奖励 $f^{M^\star}(\hat{\pi})$。形式化地，记$\pi_M := \arg\max_{\pi\in\Pi} f^M(\pi)$，我们将算法的风险 定义为:

$$\mathbf{Risk}(T) := \mathbb{E}_{\hat{\pi}\sim p}[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\hat{\pi})].$$

我们假设每个模型 M都关联一个（已知）函数 f^M: $\Pi \to \mathbb{R}$，其中 $f^M(\pi)$指定底层模型为 M时，学习者在决策 $\pi \in \Pi$ 下的取值。我们作出以下可实现性假设，它与 假设 1.1类似。

假设 1.4 (HR-DMSO 的可实现性)。 学习者可以使用一个模型类 $\mathcal{M}$ ，该模型类由概率核 M: $\Pi \to \Delta(\mathcal{O})$ 组成，并且包含真实模型 $M^\star$。

我们将元组 $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_{M\in\mathcal{M}})$ 称为 HR-DMSO 框架的一个实例 。它规定了学习者在与模型 $M^\star \in \mathcal{M}$ 交互之前先验已知的所有信息。

注释 1.4。 HR-DMSO 框架的一种等价表述是考虑模型 $M : \Pi \to \Delta(\mathcal{O} \times \mathcal{R})$，这些模型规定观测和奖励上的联合分布，并定义 $f^M(\pi) = \mathbb{E}^{M,\pi}[r]$，但只允许学习者在 $(o, r) \sim M(\pi)$ 下观测到 o。

我们将元组 $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_{M\in\mathcal{M}})$ 称为 HR-DMSO 框架的一个实例。我们将 Foster 等人（2023）提出的受约束决策估计系数扩展到 HR-DMSO，如下所示。对于实例 $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$、参考模型 M: $\Pi \to \Delta(\mathcal{O})$ 和尺度参数 $\varepsilon > 0$，受约束决策估计系数定义为[6]

$$\mathsf{dec}_\varepsilon(\mathscr{H}, \overline{M}) = \inf_{p,q\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \left\{\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi) \mid \mathbb{E}_{\pi\sim q}[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)] \le \varepsilon^2\right\}. \tag{7}$$

We 将实例 $\mathscr{H}$ 在尺度 ε 下的决策估计系数（DEC）定义为

$$\mathsf{dec}_\varepsilon(\mathscr{H}) = \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{dec}_\varepsilon(\mathscr{H}, \overline{M}). \tag{8}$$

此定义与受约束 PAC DEC 完全相同（Foster 等人，2023）；这是自然的，因为 HR-DMSO 框架与 DMSO 框架（Foster 等人，2023）之间唯一的区别，是我们放宽了智能体必须观测到其奖励这一约束。

备注 1.5。HR-DMSO 框架与部分监控问题有关（Bartók 等人，2014年）。部分监控领域的大多数研究关注遗憾保证（即，对于 π^1，.....，π^T 的累积次优性），而我们关注 PAC 保证（即，对于 $\hat{\pi}$ 的最终次优性）。这两种设定的另一项区别在于，部分监控通常考虑有限的决策空间和观测空间，而我们允许空间规模较大且具有结构（通过模型类 $\mathcal{M}$ 形式化），并致力于获得能够反映这些空间内在复杂度的样本复杂度保证。

备注 1.6 （与无奖励 DMSO 相比）。 尽管名称相似，HR-DMSO 框架与 Chen 等人最近的工作中考虑的“无奖励”DMSO 框架并不相同 （2022a）；后一框架专门针对马尔可夫决策过程，其中会明确向学习者提供奖励函数，但要等到学习过程结束之后才提供。

[6]请注意，在 HR-DMSO 和 MA-DMSO 两种设定中，我们对 DEC 使用相同的记号；通常，我们将使用字母 $\mathscr{H}$ 表示 HR-DMSO 实例，并使用 $\mathscr{M}$ 表示 MA-DMSO 实例，以避免歧义。

1.4 HR-DMSO：结果概述

不难看出，HR-DMSO 框架是 MA-DMSO 框架的推广。对于任意满足假设 1.3 和假设 1.2 的 MA-DMSO 实例 $\mathscr{M}=(\mathcal{M},\Pi,\mathcal{O},\{\Pi'_k\}_k,\{U_k\}_k)$，通过选择价值函数 $f^M(\cdot)=-h^M(\cdot)$，由元组 $\mathscr{H}=(\mathcal{M},\Pi,\mathcal{O},\{f^M\}_M)$ 指定的 HR-DMSO 框架实例（回顾 $\mathcal{O}=\mathcal{O}_\circ\times\mathcal{R}^K$）与 $\mathscr{M}$ 在统计上等价。[7]特别地，令 $\mathfrak{M}(\mathscr{M},T)$ 表示 MA-DMSO 框架中实例 $\mathscr{M}$ 的极小极大风险，并令 $\mathfrak{M}(\mathscr{H},T)$ 表示相应 HR-DMSO 实例 $\mathscr{H}$ 的极小极大风险（正式定义见第 1.6 节），则有：

1. For all models $\overline{M}$ and $\varepsilon>0$, $\mathsf{dec}_\varepsilon(\mathscr{H},\overline{M})=\mathsf{dec}_\varepsilon(\mathscr{M},\overline{M})$.
2. 对于所有 $T\in\mathbb{N}$, $\mathfrak{M}(\mathscr{H},T)=\mathfrak{M}(\mathscr{M},T)$。

自然会问，HR-DMSO 框架是否严格比 MA-DMSO 框架更一般。确实，由于允许奖励被隐藏，人们可能会认为 HR-DMSO 能够涵盖 MA-DMSO 之外的问题，因为 MA-DMSO 要求奖励必须可观测。接下来的结果表明事实并非如此：任何 HR-DMSO 实例都可以嵌入 MA-DMSO 的一个双人零和纳什均衡实例中，而统计复杂度仅会略有增加。

定理 1.3（定理 2.2的非正式版本）。考虑任意一个由元组$\mathscr{H}=(\mathcal{M},\ \Pi,\ \mathcal{O},\ \{f^M(\cdot)\}_M)$指定的 *HR-DMSO* 实例。对于任意 $\delta>0$，都存在一个双人零和纳什均衡 *MA-DMSO* 实例$\widetilde{\mathscr{M}}=(\mathcal{M},\ \Pi,\ \mathcal{O},\ \Pi'_k,\ U_k)$（定义 *1.1*），满足：*1.* 对所有 $\varepsilon>0$，$dec_\varepsilon(\mathscr{H})\le dec_\varepsilon(\mathscr{M})\le\delta+dec_{\varepsilon+\delta}(\mathscr{H})$。*2.* 对所有 $T\in\mathbb{N}$，有 $\mathfrak{M}(\mathscr{H},T)\le\mathfrak{M}(\mathscr{M},T)\le\mathfrak{M}(\mathscr{H},T)+\delta$. *3.* 如果 $\mathcal{M}$ 是有限的，那么 $\log|\mathcal{M}|\le\log|\mathcal{M}|+polylog(T,\delta^{-1})$。

这一结果表明，MA-DMSO 和 HR-DMSO 框架满足某种等价关系，并说明刻画 MA-DMSO 的极小极大样本复杂度并不比刻画 HR-DMSO 框架的极小极大样本复杂度更容易。证明通过将给定的 HR-DMSO 框架实例嵌入一个双人博弈来进行：博弈中的第一个智能体扮演 HR-DMSO 智能体的角色，第二个智能体则选择动作，以确保原始 HR-DMSO 实例的最优行动是新实例的纳什均衡，反之亦然。关键思想是，尽管博弈中的奖励是可观测的，但通过使博弈达到多项式规模，我们可以确保发现这些奖励需要付出极其巨大的探索成本，从而使它们实际上处于隐藏状态。

HR-DMSO：极小极大速率。为证明定理 1.1和1.2中的多智能体极小极大速率，我们首先证明 HR-DMSO 框架的类似界，然后利用上述等价关系将其推广到 MA-DMSO。具体而言，以下结果给出了 HR-DMSO 的主要样本复杂度界，并推广了定理 1.1。

定理 1.4（定理 3.1 至3.3的非正式版本）。对于任意实例 $\mathscr{H}=(\mathcal{M},\ \Pi,\ \mathcal{O},\ \{f^M(\cdot)\}_M)$，适用于 HR-DMSO 框架且满足 $T\in\mathbb{N}$：

- 上界：在假设 *1.1*下，存在一种算法可以达到

$$\mathbb{E}[\mathbf{Risk}(T)]\le\widetilde{O}(1)\cdot\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{H}) \tag{9}$$

  对于所有 $M\in\mathcal{M}$，其中 $\bar{\varepsilon}(T)\le\widetilde{\Theta}(\sqrt{\log|\mathcal{M}|/T})$。

- 下界：对于最坏情况模型 $M\in\mathcal{M}$，任何算法都必须满足

$$\mathbb{E}[\mathbf{Risk}(T)]\ge\widetilde{\Omega}(1)\cdot\mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{H}), \tag{10}$$

  其中，$\underline{\varepsilon}(T)$是满足以下条件的最大取值：$\varepsilon>0\,\mathsf{dec}_\varepsilon(\mathscr{H})\ge\widetilde{\Omega}(\varepsilon^2T)$。

[7]对于这一归约，$\mathscr{H}$中的奖励必须是隐藏的，因为一般而言，不可能使用来自 $M(\pi)$ 的样本来模拟均值为 $-h^M(\pi)$ 的奖励。

此外，任何仅通过奖励函数 $\mathscr{H}(\cdot)\}_{M\in\mathcal{M}}$ 以及模型 M、$M'\in\mathcal{M}$ 的成对海灵格距离来依赖实例的复杂度度量，都无法在超过二次差距的意义下刻画每个实例的最优样本复杂度。

1.5 MA-DMSO：其他结果

除了极小极大速率之外，我们还为 MA-DMSO 框架提供了一系列我们认为本身也值得关注的结构性结果，包括：(1) 多智能体 DEC 可由单智能体 DEC 控制的条件；以及 (2) 可避免所谓多智能体诅咒的条件。下面重点介绍这些结果。

从多智能体到单智能体。我们证明，在一般情况下，可以用每个玩家的单智能体 DEC 对 MA-DEC 给出上界 k。对此，最简洁的表述方式是采用由偏移版 DEC 的多智能体类比形式，该版本由Foster 等人（2021）提出。具体而言，我们考虑偏移 DEC 的一个遗憾变体，它对 $p=q$施加限制，从而将探索与利用耦合起来：对于一个实例 $\mathscr{M}$、参考模型 M和尺度参数 $\gamma>0$，我们定义

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M},\overline{M}) := \inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\left\{\mathbb{E}_{\pi\sim p}[h^{M}(\pi)]-\gamma\cdot\mathbb{E}_{\pi\sim p}[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)]\right\}. \tag{11}$$

由Foster 等人（2023）的结果（见命题 4.1）可立即推出：$\textsf{dec}_{\varepsilon}(\mathscr{M},M)\le\inf_{\gamma>0}\{\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M},M)\vee 0+\gamma\varepsilon^2\}$，因此，对$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M})$ 的上界可推出对$\textsf{dec}_{\varepsilon}(\mathscr{M})$ 的上界，进而可将其代入定理 1.1，得到极小极大风险的上界。虽然也可以不通过$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M},M)$，直接对$\textsf{dec}_{\varepsilon}(\mathscr{M},M)$ 给出上界，但使用$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\cdot)$更为方便，而且不会使最终得到的上界在定量上显著变差。

我们证明了实例 $\mathscr{M}$ 的多智能体 DEC 上界，该上界以 $\mathscr{M}$ 所定义的 K 个不同模型类 $\mathcal{M}_k$ 的（单智能体）DEC 表示。为定义这些模型类，对于 $M\in\mathcal{M}$ 和 $k\in[K]$，我们首先定义如下诱导的单智能体模型 $M|_k$：从 $M|_k(\pi)$ 中抽取的纯观测，其分布与 $o_{\circ}\sim M(\pi)$ 时的纯观测 $o_{\circ}$ 相同；从 $M|_k(\pi)$ 中抽取的奖励，其分布与 $(r_1,\dots,r_K)\sim M(\pi)$ 时的 r_k 相同。简言之，模型 $M|_k$ 与 M 相同，只是忽略了除 k 之外所有智能体的奖励。接下来，将模型类 $\mathcal{M}_k$ 定义为具有策略空间 Π_k，因此 $\mathcal{M}_k$ 中的 个模型是从 M 到 $\Pi_k\to\Delta(\mathcal{R}\times\mathcal{O}_{\circ})$ 的映射。最后，我们定义由 $\Pi_{-k}\times\mathcal{M}$ 索引的类 $\mathcal{M}_k$，如下所示：

$$\widetilde{\mathcal{M}}_k=\{\pi_k\mapsto M|_k(\pi_k,\pi_{-k})\ :\ \pi_{-k}\in\Pi_{-k},\ M\in\mathcal{M}\}. \tag{12}$$

定理 1.5。 设 $\mathscr{M}=(\mathcal{M},\Pi,\mathcal{O},\{\Pi'_k\}_k,\{U_k\}_k)$ 是满足 假设 4.1 的 *NE MA-DMSO* 实例。那么对于任意 $\gamma>0$，有[8]

$$\sup_{\overline{M}\in\mathrm{co}(\mathcal{M})}\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M},\overline{M})\le\sum_{k=1}^{K}\sup_{\overline{M}_k\in\mathrm{co}(\widetilde{\mathcal{M}}_k)}\textsf{r-dec}^{\mathrm{o}}_{\gamma/K}(\widetilde{\mathcal{M}}_k,\overline{M}_k).$$

<escape>这一结果使我们能够利用单智能体 DEC 的标准界来界定 MA-DEC（Foster 等人，2021）。例如，对于具有老虎机反馈的标准形博弈，其中每个玩家拥有 A_k 个动作，可得$\sup^{M\in\mathrm{co}(\mathcal{M})}\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M},M)\lesssim K\cdot\sum_{k=1}^{K}\frac{A_k}{\gamma}$。关于马尔可夫博弈的改进，请参见 第 4 节 。</escape>
定理 4.1 的证明采用了一种新颖的不动点论证：对于每个智能体 k，如果所有其他智能体都承诺采用某个联合分布，这就会诱导出一个单智能体 DMSO 实例；因此，智能体 k 自然会选择使该实例的单智能体 DEC 最小化的策略。利用 Kakutani 不动点定理，我们证明所有 K 智能体可以同时采用这一策略。

[8]此处，符号 $\textsf{r-dec}^{\mathrm{o}}{}_{\gamma/K}(\widetilde{\mathcal{M}}_k,M_k)$ 表示模型类 $\widetilde{\mathcal{M}}_k$对应的单智能体 DEC；参见 第 1.6 节。

关于多智能体诅咒。在多智能体强化学习中，多智能体诅咒指的是学习均衡所需的样本复杂度随玩家数量呈指数级增长的情形（Jin 等人，2021b）。一般而言，我们针对 MA-DMSO 框架（定理 1.1）给出的样本复杂度上界会受到多智能体诅咒的影响，这是因为其中包含估计复杂度项$\log|\mathcal{M}|$。例如，在一个有 K个玩家、每位玩家有 A 个动作的标准形博弈中，有$\log|\mathcal{M}| \approx A^K$ （对 $\mathcal{M}$进行适当离散化后）。我们的最终结果表明，可以用每位玩家偏离集合的最大规模$\log|\mathcal{M}|$取代估计复杂度$\max_k \log|\Pi'_k|$ ，从而避免多智能体诅咒；该规模通常关于智能体数量呈多项式增长。其代价是，结果会依赖于 MA-DMSO 实例的 MA-DEC，其中模型类 $\mathcal{M}$ 通过凸化，借助 $\mathcal{M} \leftarrow \mathrm{co}(\mathcal{M})$。

定理 1.6 （定理 5.1的非正式版本）。设 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 是 *MA-DMSO* 框架的一个粗相关均衡实例（定义 *1.2*）或相关均衡实例（定义 *A.1*）。那么，对于任意 $T \in \mathbb{N}$，算法 *1* 输出 $\widehat{\pi} \in \Pi$，使得其概率至少为 $1-\delta$，

$$\mathbf{Risk}(T) = h^{M^\star}(\widehat{\pi}) \le \widetilde{O}(K) \cdot \inf_{\gamma > 0} \left\{ \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathscr{M})) + \frac{\gamma}{T} \cdot \log\left(\frac{\max_k |\Pi'_k|}{\delta}\right) \right\},$$

其中我们采用如下约定：$\mathrm{co}(\mathscr{M}) \equiv (\mathrm{co}(\mathcal{M})\ \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$。

在拥有每位玩家 K 个动作的标准形博弈中，我们有 $\textsf{dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathscr{M})) \lesssim \frac{A}{\gamma}$ 以及 $\max_k \log|\Pi'_k| = \log(A)$，因此该结果给出

$$\mathbf{Risk}(T) \lesssim \sqrt{\frac{\mathrm{poly}(K) \cdot A}{T}}.$$

更广泛地说，定理 1.6 表明，对于凸类，通常可以避免多智能体诅咒，其中包括具有老虎机反馈的标准形博弈结构化类，例如具有线性或凸收益的博弈。不过，一般而言，它并不能为非凸类（如马尔可夫博弈）提供紧确保证，因此最好将其视为对此类情形相关结果的补充（Jin 等人，2021b；Song 等人，2021；Mao 和 Basar，2022年）。该结果通过改造单智能体设定中的强大探索即优化算法（Lattimore，2022年；Foster 等人，2022b），使其能够利用多智能体情形独特的反馈结构，从而证明。详情请参见第 5 节，其中还介绍了其他结果，指出在某些情形下，无法以定理 1.6所述的意义避免多智能体诅咒。

1.6 预备知识

下面我们补充一些技术性预备知识，供后续证明使用。

概率核。对于概率空间 $(\mathcal{X}, \mathscr{X})$ 和 $(\mathcal{Y}, \mathscr{Y})$，从 $(\mathcal{X}, \mathscr{X})$ 到 $(\mathcal{Y}, \mathscr{Y})$ 的概率核 $P(\cdot|\cdot)$ 是一个映射 P: $\mathscr{Y} \times \mathcal{X} \to [0,1]$，满足 (1) 对所有 $x \in \mathcal{X}$，$P(\cdot|x)$ 是 $(\mathcal{Y}, \mathscr{Y})$ 上的概率测度；并且 (2) 对所有 $Y \in \mathscr{Y}$，映射 $x \mapsto P(Y|x)$ 关于 $\mathscr{X}$ 可测。为简化记号，我们通常将概率核记作 $P : \mathcal{X} \to \Delta(\mathcal{Y})$。

MA-DMSO 框架。我们采用与 Foster 等人（2021，2023）相同的概率空间形式体系。决策与测度空间 $(\Pi, \mathscr{P})$ 相关联，观测与测度空间 $(\mathcal{O}, \mathscr{O})$ 相关联。在 MA-DMSO 框架中，纯观测与测度空间 $(\mathcal{O}_{\circ}, \mathscr{O}_{\circ})$ 相关联，奖励与测度空间 $(\mathcal{R}, \mathscr{R})$ 相关联；此外，我们有 $\mathcal{O} = \mathcal{O}_{\circ} \times \mathcal{R}^K$和 $\mathscr{O} = \mathscr{O}_{\circ} \otimes \mathscr{R}^{\otimes K}$。形式上，模型 $M(\cdot \mid \cdot)$ 是从 $(\Pi, \mathscr{P})$ 到 $(\mathcal{O}, \mathscr{O})$ 的概率核。我们将所有模型的集合记为 $\mathcal{M}^+$。注意，$\mathcal{M}^+$ 取决于测度空间 $(\Pi, \mathscr{P})$、$(\mathcal{O}, \mathscr{O})$；当我们希望明确表示这种依赖关系时，将记作 $\mathcal{M}^+_{\Pi,\mathcal{O}}$。截至时间 t 的历史为 $\mathfrak{H}^t = (\pi^1, o^1) \ldots, (\pi^t, o^t)$。我们定义

$$\Omega^t = \prod_{i=1}^{t} (\Pi \times \mathcal{O}), \qquad \mathscr{F}^t = \bigotimes_{i=1}^{T} (\mathscr{P} \otimes \mathscr{O}),$$

使得 $\mathfrak{H}^t$ 与空间 $(\Omega^t, \mathscr{F}^t)$ 相关联。

除非另有说明，本文始终假设 $\mathcal{R} = [0,1]$（这尤其意味着，对于所有 M, π，$h^M(\pi) \in [0, K]$）。为简化记号，对于每个 $\pi \in \Pi$ 和 $M \in \mathcal{M}$，我们记 $h_k^M(\pi) := \sup_{\pi'_k \in \Pi_k} f_k^M(U_k(\pi'_k, \pi)) - f_k^M(\pi)$，因此 $h^M(\pi) = \sum_{k=1}^K h_k^M(\pi)$。

规范的单智能体实例。给定一个决策空间 Π、一个观测空间 $\mathcal{O} = \mathcal{O}_\circ \times \mathcal{R}$ 和一个模型类 $\mathcal{M} \subset (\Pi \to \Delta(\mathcal{O}))$，存在一个与模型类 $\mathcal{M}$ 对应的规范单智能体实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$：我们取 $\Pi'_1 = \Pi$ 和 $U_1(\pi'_1, \pi) = \pi'_1$，这确保对于所有 $\pi \in \Pi, M \in \mathcal{M}$，$h^M(\pi) = \max_{\pi' \in \Pi} f^M(\pi') - f^M(\pi)$。1-player MA-DMSO 框架的单智能体实例 $\mathscr{M}$ 精确地刻画了 Foster 等人（2021，2023）针对模型类 $\mathcal{M}$ 提出的 DMSO 框架。此外，对于任意模型 M，我们将记 $\mathrm{dec}_\varepsilon(\mathcal{M}, M) = \mathrm{dec}_\varepsilon(\mathscr{M}, \overline{M})$（类似地，对于第 4 节中引入的偏移 DEC 的遗憾变体，我们将记 $\underline{\text{r-dec}}^\circ_\gamma(\mathcal{M}, M) = \underline{\text{r-dec}}^\circ_\gamma(\mathscr{M}, M)$）；量 $\mathrm{dec}_\varepsilon(\mathcal{M}, M) = \mathrm{dec}_\varepsilon(\mathscr{M}, M)$ 与 Foster 等人（2023）定义的模型类 $\mathcal{M}$ 的受约束（PAC）决策估计系数完全相同，而量 $\text{r-dec}^\circ_\gamma(\mathcal{M}, M) = \text{r-dec}^\circ_\gamma(\mathscr{M}, M)$ 与 Foster 等人（2021）定义的模型类 $\mathcal{M}$ 的偏移（遗憾）DEC 完全相同。

HR-DMSO 框架。与 MA-DMSO 框架一样，决策与测度空间 $(\Pi, \mathscr{P})$ 相关联，观测与测度空间 $(\mathcal{O}, \mathscr{O})$ 相关联，模型 $M(\cdot \mid \cdot)$ 是从 $(\Pi, \mathscr{P})$ 到 $(\mathcal{O}, \mathscr{O})$ 的概率核。时间 t 之前的历史由 $\mathfrak{H}^t = (\pi^1, o^1) \dots, (\pi^t, o^t)$ 给出，并与空间 $(\Omega^t, \mathscr{F}^t)$ 相关联，该空间定义为

$$\Omega^t = \prod_{i=1}^t (\Pi \times \mathcal{O}), \qquad \mathscr{F}^t = \bigotimes_{i=1}^T (\mathscr{P} \otimes \mathscr{O}).$$

我们将所有模型的集合记为 $\mathcal{M}^+$。除非另有说明，全文均假设对于所有 $M \in \mathcal{M}^+$ 和 $\pi \in \Pi$，$f^M(\pi) \in [0,1]$。

对于模型 M 和决策 $\pi \in \Pi$，$\mathbb{E}^{M,\pi}[\cdot]$ 表示过程 $o \sim M(\pi)$ 下的期望。为简化记号，我们常将 $g^M(\pi) := f^M(\pi_M) - f^M(\pi)$ 简记为，因此 **风险**$(T) = \mathbb{E}_{\widehat{\pi} \sim p}[g^{M^\star}(\widehat{\pi})]$。

密度比。对于 MA-DMSO 和 HR-DMSO，我们定义

$$V(\mathcal{M}) := \sup_{M, M' \in \mathcal{M}} \sup_{\pi \in \Pi} \sup_{A \in \mathscr{O}} \left\{ \frac{M(A \mid \pi)}{M'(A \mid \pi)} \right\} \vee e. \tag{13}$$

$V(\mathcal{M})$ 的有限性并非我们的结果成立所必需，但可使若干界改善一个 $\log(T)$ 因子。

散度。全变差距离定义为

$$D_{\mathsf{TV}}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathscr{F}} |\mathbb{P}(A) - \mathbb{Q}(A)| = \frac{1}{2} \int |d\mathbb{P} - d\mathbb{Q}|,$$

Kullback–Leibler散度定义为

$$D_{\mathsf{KL}}(\mathbb{P} \,\|\, \mathbb{Q}) = \begin{cases} \int \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P}, & \mathbb{P} \ll \mathbb{Q}, \\ +\infty, & \text{otherwise.} \end{cases}$$

极小极大样本复杂度。形式上，对于 $T \in \mathbb{N}$，算法（适用于 HR-DMSO 或 MA-DMSO 框架）是由概率核 $(p, q) = \left(p(\cdot \mid \cdot)\ \{q^t(\cdot \mid \cdot)\}_{t=1}^T\right)$ 构成的集合，其中每个 $q^t : \Omega^{t-1} \to \Delta(\Pi)$ 都是从 $(\Omega^{t-1}, \mathscr{F}^{t-1})$ 到 $(\Pi, \mathscr{P})$ 的概率核，而 $p : \Omega^T \to \Delta(\Pi)$ 是从 $(\Omega^T, \mathscr{F}^T)$ 到 $(\Pi, \mathscr{P})$ 的概率核。我们将 $\mathbb{P}^{M,(p,q)}$ 记为过程下 $(\mathfrak{H}^T, \widehat{\pi})$ 的概率律：

$$\pi^t \sim q^t(\cdot \mid \mathfrak{H}^{t-1}),\ o^t \sim M(\cdot \mid \pi^t),\ \forall t \in [T], \qquad \widehat{\pi} \sim p(\cdot \mid \mathfrak{H}^T),$$

并用 $\mathbb{E}^{M,(p,q)}$ 表示相应的期望。我们的主要目标是刻画 MA-DMSO 框架实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 或 HR-DMSO 框架实例 $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$ 的极小极大 PAC 样本复杂度。两种情况下的极小极大样本复杂度均以相同方式定义，具体如下：

$$\mathfrak{M}(\mathscr{M}, T) := \inf_{(p,q)} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star,(p,q)} \mathbb{E}_{\widehat{\pi} \sim p(\cdot|\mathfrak{H}^T)} \left[\sum_{k=1}^{K} \sup_{\pi'_k \in \Pi'_k} f_k^M(U_k(\pi'_k, \widehat{\pi})) - f_k^M(\widehat{\pi}) \right],$$
$$\mathfrak{M}(\mathscr{H}, T) := \inf_{(p,q)} \sup_{M^\star \in \mathcal{M}} \mathbb{E}^{M^\star,(p,q)} \mathbb{E}_{\widehat{\pi} \sim p(\cdot|\mathfrak{H}^T)} [f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\widehat{\pi})].$$

1.7 组织结构

Thi本文的组织结构如下。首先，第 2 节 和 第 3 节 介绍我们的主要结果 :

- 第 2 节 阐明了 MA-DMSO 与 HR-DMSO 之间的一种等价关系。
- 第 3 节 基于决策估计系数，给出了两种框架下极小极大速率的上界和下界，并指出了获得更精确保证所面临的障碍；这类保证类似于基本 DMSO 框架中的结果（Foster 等人，2023）。

节 4 和第 5 节随后介绍有关 MA-DMSO 框架的其他结果 rk:

- 第 4 节给出了可以用单智能体 DEC 来界定 MA-DEC 的一般条件。
- 第 5 节给出了在 MA-DMSO 框架中获得样本复杂度保证且避免所谓多智能体诅咒的条件，并举例说明在某些情况下这是不可能的。

所有证明均见 附录。两个框架的更多示例见附录 A .

补充记号。对于整数 $n \in \mathbb{N}$，我们用 $[n]$ 表示集合 $\{1, \dots, n\}$。对于集合 $\mathcal{X}$，我们用 $\Delta(\mathcal{X})$ 表示 $\mathcal{X}$ 上所有概率分布构成的集合。对于 $x \in \mathcal{X}$，我们用 $\mathbb{I}_x \in \Delta(\mathcal{X})$ 表示分布~，该分布将概率质量 1 赋予 x。我们采用标准的大 O 记号，并用 $f = O(g)$ 表示 $f = O(g \cdot \max\{1, \text{polylog}(g)\})$。我们仅在非正式表述中使用 $\lesssim$，以强调不等式中最相关的部分。对于集合 $\mathcal{X}$，令 $\mathcal{P}(\mathcal{X})$ 表示其幂集，即 $\mathcal{X}$ 的所有子集构成的集合。

2 MA-DMSO 与 HR-DMSO 框架的等价性

在本节中，作为主要结果的出发点，我们证明 MA-DMSO 和 HR-DMSO 框架满足某种统计等价性。首先，在 定理 2.1 中，我们形式化了这种等价性的平凡方向：也就是说，MA-DMSO 框架的任意实例都可以视为 HR-DMSO 框架的一个实例。为陈述这一结果，回顾一下，按照我们的约定，MA-DMSO 实例中的完整观察空间记作 $\mathcal{O} = \mathcal{O}_\circ \times \mathcal{R}^K$。

定理 2.1（将 MA-DMSO 归约为 HR-DMSO）。 考虑任意一个满足 假设 *1.3* 和 假设 *1.2* 、由元组 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 指定的 *MA-DMSO* 框架实例。~那么，对于某种价值函数 f^M，由元组 $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M\}_M)$ 指定的 *HR-DMSO* 框架实例满足：

1. 对于所有模型 M和 $\varepsilon > 0$，$\mathsf{dec}_\varepsilon(\mathscr{H}, M) = \mathsf{dec}_\varepsilon(\mathscr{M}, M)$。
2. 对于所有 $T \in \mathbb{N}$, $\mathfrak{M}(\mathscr{H}, T) = \mathfrak{M}(\mathscr{M}, T)$。

~该结果通过选择价值函数 $f^M(\pi) = K - h^M(\pi)$ 来实现。请注意，为使该归约可接受，奖励处于隐藏状态至关重要：在 MA-DMSO 框架中，函数 $h^M(\pi)$ 不会被直接观测到，正如我们将看到的，这正是根本性困难的来源。

定理 2.1 是一个相当直接的结果，自然会让人认为 HR-DMSO 框架可能确实比 MA-DMSO 框架更具一般性，尤其是因为在后者中可以观察到奖励。 以下结果是定理 1.3的正式版本，它表明，如果允许存在小幅近似，那么 HR-DMSO 框架中的任何实例都可以嵌入 MA-DMSO 的一个双人零和纳什均衡实例中，且复杂度仅有极小幅度的增加。

定理 2.2（将 HR-DMSO 约化为 MA-DMSO）。 考虑由元组 $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$ 指定的任意 *HR-DMSO* 框架实例。则对于任意 $V \in \mathbb{N}$，存在一个针对 *MA-DMSO* 框架（定义 *1.1*）的双人零和纳什均衡实例 $\widetilde{\mathscr{M}} = (\widetilde{\mathcal{M}}, \widetilde{\Pi}, \widetilde{\mathcal{O}}, \widetilde{\Pi}'_k, \widetilde{U}_k)$，满足：

1. 对所有 $\varepsilon > 0$，$\widetilde{dec}_\varepsilon(\widetilde{\mathscr{M}}) \le dec_\varepsilon(\mathscr{H}) \le 6/\sqrt{V} + \overline{dec}_{\varepsilon+(6/V)^{-1/2}}(\mathscr{M})$。 *2.* 对所有 $T \in \mathbb{N}$，有 $\mathfrak{M}(\widetilde{\mathscr{M}}, T) \le \mathfrak{M}(\mathscr{H}, T) \le \mathfrak{M}(\mathscr{M}, T) + O((T\log(T)/V)^{1/4})$。

3. $\widetilde{\mathcal{M}}$ 由元组 $(M, i) \in \mathcal{M} \times [V]$索引。特别地，如果 $\mathcal{M}$ 是有限的，那么 $\log|\widetilde{\mathcal{M}}| = \log|\mathcal{M}| + \log V$。

这一结果的主要推论是，刻画 MA-DMSO 的极小极大样本复杂度并不比刻画 HR-DMSO 框架的极小极大样本复杂度容易；因此，在接下来的结果中，我们可以将注意力限制在后者上。下面再补充几点说明。

- 随着参数 V 增大，定理 2.2 中对极小极大速率的近似会有所改善。对于所有我们感兴趣的情形，取 $V = \mathrm{poly}(T)$ 即可，唯一的权衡是模型类的规模 $\mathcal{M}$ 会从 $\log|\mathcal{M}|$ 增长到 $\log|\mathcal{M}| + \log V$。对于后续各节所考虑的结果，这一增长并无实质影响（除了 $\log(T)$ 因子外）。

- 除了保持极小极大风险外，这两种归约还保持决策估计系数的取值；这是因为它们保持了该类中模型的奖励和海林格距离。这一点与我们后续的结果（第 3 节）有关：我们将证明，DEC 与极小极大风险密切相关，但二者并不完全等价。

- 这两种归约在本质上都是算法性的。例如，假设我们从一个 HR-DMSO 实例 $\mathscr{H}$ 出发，并通过 中的归约构造一个 MA-DMSO 实例 $\widetilde{\mathscr{M}}$，该归约见 定理 2.2。那么，对于 $\widetilde{\mathscr{M}}$ 中的每个模型都能实现低风险的任何算法，都可以高效地提升为适用于原始类别 $\mathscr{H}$的算法。

定理 2.2 通过将给定的 HR-DMSO 框架实例 $\mathscr{H}$嵌入一个双人零和博弈实例 $\widetilde{\mathscr{M}}$来证明，其中两名智能体中的第一名扮演 HR-DMSO 智能体的角色。该嵌入的关键性质如下：

1. 第二名智能体选择动作，确保原始 HR-DMSO 实例中的近似最优决策构成新实例的纳什均衡，反之亦然。2. 尽管博弈实例 $\widetilde{\mathscr{M}}$ 中的奖励可被观测到，但通过增大博弈规模（使其成为参数 V的函数），我们可以确保发现一个非零奖励动作需要进行多到难以承受的探索，从而使这些奖励实际上不可见（误差至多为较小的近似误差）。

3 极小极大速率的上界和下界

本节介绍 MA-DMSO 和 HR-DMSO 框架下极小极大速率的相关结果。本节大部分内容采用 HR-DMSO 框架，并在最后利用第 2 节中的等价关系，说明这些结果对 MA-DMSO 的启示。具体而言：

- 在第 3.1 节中，我们给出了 HR-DMSO 框架下交互式决策的极小极大速率上界和下界，这些界随受约束 DEC 而变化。
- 接下来，我们将在第 3.2 节证明，在受约束 DEC 满足温和的正则性假设时，极小极大速率的上界和下界至多相差一个多项式因子（忽略估计误差项）；对于大多数参数范围，这些界之间的差距至多为二次。随后，我们将证明——或许令人惊讶的是——上界和下界都无法

有所改进，因为存在这样的实例：其中每一种都几乎是紧的。换言之，与 DMSO 框架（Foster 等人，2023年）相比，在 HR-DMSO 框架中，受约束 DEC 无法给出一个除二次因子外仍然紧的极小极大样本复杂度刻画。我们进一步表明，这一差距并不局限于受约束 DEC，事实上，它适用于一整类基于模型之间成对 f-散度的复杂度度量。因此，任何对 HR-DMSO 的极小极大速率进行刻画、且在多对数因子范围内紧的结果，都必须采用一种与近期工作（Foster 等人，2021、2022b、2023年）中所考虑的复杂度度量实质不同的度量。

- 最后，利用上一节中给出的等价关系，我们证明（第 3.3 节）上述所有结果在 MA-DMSO 框架中同样成立。

本节中的所有结果均以一般形式给出。有关其在特定实例中的应用，请参阅附录的第一部分。

3.1 HR-DMSO：极小极大速率的上界和下界

现在，我们给出 HR-DMSO 框架下极小极大风险的上界和下界。作为近期 从估计到决策$^+$ (E2D$^+$ forPAC算法遗憾界的直接推论，我们得到上界。该算法由 Foster 等人 (2023提出。E2D$^+$ forPAC 算法最初是在（单智能体／非隐藏奖励）DMSO 框架中提出的；在该框架下，它基于受约束 DEC 得到了紧的极小极大风险上界（Foster 等人，2023）。我们发现，无需任何修改，该算法在更一般的 HR-DMSO 框架中也能提供完全相同的保证；只要检查 E2D$^+$ forPAC 算法在 Foster 等人 (2023) 中的正确性证明，并注意到证明并未利用学习智能体能够观测奖励 r^1，……，r^T这一事实即可看出这一点。有关该算法的更多背景信息，请参阅附录 D.1。

我们的主要上界适用于 $\mathcal{M}$ 有限的情形（$|\mathcal{M}| < \infty$）；关于无限类的更一般保证见附录 D.1。

定理 3.1（HR-DMSO 的极小极大上界（Foster 等人，2023））。固定 $\delta \in \left(0, \frac{1}{10}\right)$ 和 $T \in \mathbb{N}$，并考虑任意实例 $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$。假设假设 *1.4* 成立。令$\varepsilon(T) := 16\sqrt{\lceil \log_T 2/\delta \rceil \cdot \log \frac{|\mathcal{M}|}{\delta}}$，则经过适当配置的~~*E2D$^+$ PAC* 算法~~保证以至少 $1-\delta$ 的概率，

$$\mathbf{Risk}(T) \le \mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{H}).$$

此外，若 $f^M(\cdot) \in [0, R]$ 对所有 $M \in \mathcal{M}$ 及某个 $R > 0$均成立，则期望风险有如下界：$\mathbb{E}[$**风险**$(T)] \le \mathsf{dec}^{\varepsilon(T)}(\mathscr{H}) + \delta R$。

在解读这一结果之前，我们先补充主要的下界，定理 3.2，该定理表明，对于适当选择的尺度参数，任何算法的极小极大风险都以下界受约束 DEC 为下界 $\varepsilon > 0$。该结果的陈述使用了定义 $C(T) := \log(T \wedge V(\mathcal{M}))$。此外，我们回顾一下，$g^M(\pi) := f^M(\pi_M) - f^M(\pi)$。

定理 3.2 (HR-DMSO 的极小极大下界**)**。考虑任意实例 $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$并记 $R := \sup_{\pi \in \Pi,\ M \in \mathcal{M}} g^M(\pi)$。给定 $T \in \mathbb{N}$，令 $\varepsilon(T) > 0$尽可能大，使得

$$\underline{\varepsilon}(T)^2 \cdot C(T) \cdot R \cdot T \le \frac{1}{8} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{H}). \tag{14}$$

那么，对于任意算法，都存在一个 $\mathcal{M}$ 中的模型，使得

$$\mathbb{E}[\mathbf{Risk}(T)] \ge \frac{1}{6} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{H}).$$

理解这些界。现在我们通过几个例子来了解定理 3.2的下界和定理 3.1的上界的表现。为简单起见，我们考虑以下情形：$R = \sup_{\pi \in \Pi,\ M \in \mathcal{M}} g^M(\pi) = 1$ **(**在定理 3.2的语境中）。

- $\sqrt{T}$ 阶。老虎机和强化学习文献中研究的大多数模型类都具有这样的性质：最优速率为 $O(\sqrt{T})$。其中许多问题具有可观测奖励的性质（即，它们属于 DMSO 框架），但在奖励不可观测的 HR-DMSO 问题中也会出现这样的速率；一个显著的例子是局部可观测的有限部分监控问题（Bartók 等人，2014年）。对于这类模型，有 $\mathrm{dec}_\varepsilon(\mathscr{H}) \propto \varepsilon \cdot \sqrt{C\mathrm{prob}}$，其中 $C\mathrm{prob}$ 是反映模型类复杂度的问题相关常数 > 0（示例参见 Foster 等人（2021，2023））。在这种情况下，通过选择失败概率 $\delta = 1/T$，我们有 $\varepsilon(T) \lesssim \sqrt{\log(T)\log(T|\mathcal{M}|)/T}$，因此定理 3.1 给出的上界为

$$\mathbb{E}[\mathbf{Risk}(T)] \le \widetilde{O}\left(\sqrt{\frac{C_{\mathrm{prob}} \log|\mathcal{M}|}{T}}\right)$$

关于极小极大风险。对于下界，若 $\mathrm{DEC}_\varepsilon(\mathscr{H}) \propto \varepsilon \cdot \sqrt{C_{\mathrm{prob}}}$，则不动点方程 (14) 的解为 $\varepsilon(T) \gtrsim \sqrt{C_{\mathrm{prob}}/(T \cdot C(T))}$。根据 定理 3.2，这意味着下界为

$$\mathbb{E}[\mathbf{Risk}(T)] \ge \widetilde{\Omega}\left(\frac{C_{\mathrm{prob}}}{T}\right)$$

关于极小极大风险，该下界与上界相差一个二次因子（忽略 $\log|\mathcal{M}|$ 因子）。根据Foster 等人 (2023) 的结果，在可观测到奖励的特殊情形下（即 DMSO 框架），$\widetilde{O}\left(\sqrt{\frac{C_{\mathrm{prob}} \log|\mathcal{M}|}{T}}\right)$ 是正确的速率（相差 $\log|\mathcal{M}|$ 因子和 $\log T$ 因子）。我们将在下文说明，在奖励不可观测的一般情形下，情况未必如此，而且下界可能是紧的。

- 非参数速率。对于非参数模型类，其最优遗憾为 $\omega(\sqrt{T})$，通常有 $\mathrm{dec}_\varepsilon(\mathscr{H}) \propto \varepsilon^{1-\rho}$，其中某个 $\rho \in (0, 1)$。对于这类问题，定理 3.1 给出的极小极大风险上界为 $\mathbb{E}[\text{风险}(T)] \le \widetilde{O}\left((\log|\mathcal{M}|/T)^{(1-\rho)/2}\right)$。相比之下，方程 (14) 的不动点方程所能得到的最佳解为 $\varepsilon(T) \gtrsim 1/(T \cdot C(T))^{\frac{1}{1+\rho}}$，根据定理 3.2，这对应于极小极大风险下界 $\mathbb{E}[\text{风险}(T)] \ge \widetilde{\Omega}\left(1/T^{\frac{1-\rho}{1+\rho}}\right)$。这里的下界与上界相差 $T^{\frac{2}{1+\rho}} \le 2$ 的一个幂次（忽略 $\log|\mathcal{M}|$ 因子）。根据 Foster 等人（2023）的结果，在观测到奖励这一特殊情形下，$\widetilde{O}\left((\log|\mathcal{M}|/T)^{(1-\rho)/2}\right)$ 的上界是正确的速率（忽略 $\log|\mathcal{M}|$ 因子和 $\log T$ 因子）。

我们参考Foster 等人（2023）以了解上文针对奖励可观测这一特殊情形（DMSO）所概述增长率的具体示例，并参考附录F中由 MA-DMSO 导出的示例。

3.2 HR-DMSO：界之间的差距以及不可能实现紧致刻画

现在，我们研究定理 3.1和3.2中上界和下界之间差距的性质。我们首先给出差距的一个一般界限，然后证明进一步缩小这一差距是不可能的——而且是在相当强的意义下不可能。

3.2.1 上界和下界之间的差距

忽略常数因子后，定理 3.1 和3.2 的上界与下界之间唯一的区别在于计算 DEC 时所采用的尺度 ε。定理 3.1 的上界使用尺度 $\varepsilon(T) = 8\sqrt{\lceil \log_T 2/\delta \rceil \cdot \log|\mathcal{M}|}$，而定理 3.2 的下界（其中 $R = 1$）使用尺度 $\varepsilon(T)$；该尺度通过满足约束 $\varepsilon(T)^2 \cdot C(T) \cdot T \le \frac{1}{8} \cdot \mathrm{dec}_{\varepsilon(T)}(\mathscr{H})$，隐式定义为尽可能大的值。因此，$\varepsilon(T)$ 与 $\varepsilon(T)$ 之间的差距大小决定了这些上界和下界的紧致程度。下文将给出一个界：只要受约束 DEC 满足以下正则性假设，该界便成立。

假设 3.1 (正则性)。 如果一个实例 $\mathscr{H}$ （属于 *HR-DMSO* 或 *MA-DMSO*）在尺度 $\varepsilon \in$ 上满足正则性条件，且常数为 C_{reg}、 $c_{\mathrm{reg}} > 1$，其中 $\varepsilon \in (0,2)$，则称其满足

$$\mathsf{dec}_{\varepsilon}(\mathscr{H}) \le c_{\mathrm{reg}}^2 \cdot \mathsf{dec}_{\varepsilon/C_{\mathrm{reg}}}(\mathscr{H}).$$

大多数自然类别都满足假设 3.1，其中常数为 c_{reg}， C_{reg} （特别地，当$\mathsf{dec}_{\varepsilon}(\mathscr{H}) \propto \varepsilon^p$ 对 $p < 2$成立时，该条件也成立）。我们注意到，Foster 等人(2023也使用了类似的假设，以给出 DMSO 框架中可达到的最优速率上界。

在假设 3.1下，以下结果表明，我们关于极小极大风险的上界（其随$\mathsf{dec}^{\varepsilon(T)}(\mathscr{H})$变化）被一个关于下界的多项式量所上界，即$\mathsf{dec}^{\varepsilon(T)}(\mathscr{H})$。

命题 3.1。 假设某个实例 $\mathscr{H}$ （适用于 *HR-DMSO* 或 *MA-DMSO*）满足假设 *3.1* ，其中常数取值为 $C_{\mathrm{reg}} > c_{\mathrm{reg}} > 1$，且对所有 $\varepsilon \in (\varepsilon(T) \cdot \frac{c_{\mathrm{reg}}}{C_{\mathrm{reg}}}, \varepsilon(T))$均成立。任取$\beta \ge \log c_{\mathrm{reg}} \log(C_{\mathrm{reg}}/c_{\mathrm{reg}})$。那么，对任意 $T \in \mathbb{N}$，

$$\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{H}) \le (C \log 1/\delta \cdot \log|\mathcal{M}| \cdot C(T) \cdot C_{\mathrm{reg}}/c_{\mathrm{reg}})^{\frac{\beta}{1+\beta}} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{H})^{\frac{1}{1+\beta}}.$$

我们指出，命题 3.1 是一个纯粹的代数事实，无需利用相关均衡的结构，特别地，对 HR-DMSO 和 MA-DMSO 框架的实例均成立。为使该结果具体化，我们重新考察第 3.1 节中讨论的每种情形，并说明应用命题 3.1如何使我们得出上界和下界之间相差一个多项式因子。

- $\sqrt{T}$-速率。假设 $\mathsf{dec}_{\varepsilon}(\mathscr{H}) \propto \varepsilon \cdot \sqrt{C\mathsf{prob}}$，其中 $C\mathsf{prob}$ 是某个问题相关常数，且 > 0。那么，对于任意常数 $\beta > 1$，都存在一个足够大的绝对常数 $C\mathsf{reg} > 1$，使得对所有 $\varepsilon > 0$，$\mathsf{dec}_{\varepsilon}(\mathscr{H}) \le C^{\beta}\mathsf{reg} \cdot \mathsf{dec}_{\varepsilon/C\mathsf{reg}}(\mathscr{H})$。由此可知，假设 3.1 对常数 $C\mathsf{reg}$ 和$c\mathsf{reg} := C^{\beta/2}\mathsf{reg}$ 成立（它们满足 $\beta \ge \log c\mathsf{reg} \log(C\mathsf{reg}/c\mathsf{reg})$），而命题 3.1 给出:

  $$\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{H}) \le \widetilde{O}(\log|\mathcal{M}|)^{\frac{\beta}{1+\beta}} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{H})^{\frac{1}{1+\beta}}.$$

  忽略估计误差并取 $\beta \to 1$，我们得出 $\mathsf{dec}^{\varepsilon(T)}(\mathscr{H}) \lesssim \mathsf{dec}^{\varepsilon(T)}(\mathscr{H})^{1/2-o(1)}$，即我们的上界和下界之间存在（近似）二次差距。
- 非参数速率。假设 $\mathsf{dec}_{\varepsilon}(\mathscr{H}) \propto \varepsilon^{1-\rho}$ 对于某个 $\rho \in (0, 1)$ 成立。那么对于任意常数 $\beta > \frac{1-\rho}{1+\rho}$，都存在一个足够大的常数 $C\mathsf{reg} > 1$，使得对于所有 $\varepsilon > 0$，$\mathsf{dec}_{\varepsilon}(\mathscr{H}) \le C^{\beta(1+\rho)}\mathsf{reg} \cdot \mathsf{dec}_{\varepsilon/C\mathsf{reg}}(\mathscr{H})$。因此，假设 3.1 满足常数 $C\mathsf{reg}$ 和 $c\mathsf{reg} := C^{\beta(1+\rho)/2}\mathsf{reg}$，它们满足 $\beta \ge \log c\mathsf{reg} \log(C\mathsf{reg}/c\mathsf{reg})$，而命题 3.1 给出

  $$\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{H}) \le \widetilde{O}(\log|\mathcal{M}|)^{\frac{\beta}{1+\beta}} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{H})^{\frac{1}{1+\beta}}.$$

  忽略估计误差并取 $\beta \to \frac{1-\rho}{1+\rho}$ （使得 $\frac{1}{1+\beta} \to \frac{1+\rho}{2}$ ），我们得出$\mathsf{dec}^{\varepsilon(T)}(\mathscr{H}) \lesssim \mathsf{dec}^{\varepsilon(T)}(\mathscr{H})^{\frac{1+\rho}{2}-o(1)}$，即上界和下界之间的差距小于二次差距。

当然，第 3.1 节 中的论证已经使我们能够直接得出这些结论；这里的目的是说明，如何将这一结论作为更一般的命题 3.1的特殊情形得到。

3.2.2 关于极小极大风险的紧确刻画

自然会让人想知道，能否缩小上界和下界之间的多项式间隙，从而将极小极大风险刻画到仅相差多对数因子的程度。本节将从几个方面说明这是不可能的。

上界和下界的紧致性。在命题 3.2 和 3.3 中，我们给出了两个实例 $\mathscr{H}_1\mathscr{H}_2$，因此，除对数 $\frac{1}{\varepsilon}$ 因子外，有 $\mathsf{dec}_\varepsilon(\mathscr{H}_1) \asymp \varepsilon$ 以及 $\mathsf{dec}_\varepsilon(\mathscr{H}_2) \asymp \varepsilon$。尽管这两个实例的 DEC 表现相同，但它们的极小极大速率不同：对于实例 $\mathscr{H}_1$，定理 3.1 给出的上界是紧的（$\mathfrak{M}(\mathscr{H}_1, T) \gtrsim 1/\sqrt{T}$）；而对于 $\mathscr{H}_2$，定理 3.2 给出的下界是紧的（$\mathfrak{M}(\mathscr{H}_2, T) \lesssim \log(T)/T$）。

命题 3.2（一个上界是紧的实例）。对于任意充分大的 $L, A \in \mathbb{N}$，存在一个实例 $\mathscr{H}_1 = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$，满足 $\log|\mathcal{M}| \le \log(LA)$，并且具有以下性质：

1. 对于所有 $T \le 2^{L/2}$，$\mathscr{H}_1$ 的极小极大速率由 $\mathfrak{M}(\mathscr{H}_1, T) = \Theta(\sqrt{A/T})$ 给出。
2. 对于所有 $\varepsilon \in (2^{-L}, 1/\sqrt{A})$，有 $c \cdot \varepsilon\sqrt{A} \le \mathsf{dec}_\varepsilon(\mathscr{H}_1) \le C \cdot \varepsilon\sqrt{A}$，其中常数为 $c, C > 0$。

该实例 $\mathscr{H}_1$ 在 命题 3.2 中的模型类由一类具有A 个臂和伯努利奖励的多臂老虎机问题的子类给出，且该命题中的界可直接由先前的工作推出。为完整起见，我们在 第 3.2 节 中给出证明。

命题 3.3（一个下界是紧的实例）。对于任意充分大的 $L \in \mathbb{N}$ 以及任意$C_{\mathrm{prob}} \ge 1$，存在一个实例 $\mathscr{H}_2 =$（$\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M$），满足 $\log|\mathcal{M}| \le L^2$，并具有以下性质：

1. 对所有 $T \le 2^L$，该实例的极小极大速率 $\mathscr{H}_2$满足如下界 $\mathfrak{M}(\mathscr{H}_2, T) \le \sqrt{\frac{8C^2_{\text{概率}}\log T}{T}}$.
2. 对所有 $\varepsilon \ge \frac{\sqrt{2}}{C_{\mathrm{prob}}\cdot 2^L}$，有 $\frac{C_{\mathrm{prob}}}{\sqrt{8\cdot L}}\cdot\varepsilon \le \mathrm{DEC}\varepsilon(\mathscr{H}_2) \le 2C_{\mathrm{prob}}\cdot\varepsilon$. 特别地，$\varepsilon(T) \ge \Omega\left(\sqrt{\frac{C_{\mathrm{prob}}}{T}\log(\cdot)\cdot L}\right)$ 只要 $T \le 2^L/L^3$。

特别地，对于任意 $T \in \mathbb{N}$，通过选择 $L = 100\log T$，可得对于所有 $\varepsilon \ge \Omega\left(\frac{1}{C_{\mathrm{prob}}\cdot T^{100}}\right)$，该实例 $\mathscr{H}_2$ 满足 $\Omega(\varepsilon\cdot C_{\mathrm{prob}}/\log\frac{1}{\varepsilon}) \le \mathrm{DEC}\varepsilon(\mathscr{H}_2) \le O(\varepsilon\cdot C_{\mathrm{prob}})$，而极小极大风险满足如下界：$\mathfrak{M}(\mathscr{H}_2, T) \le O(C^2_{\mathrm{prob}}\log(T)/T)$。

我们来比较 命题 3.2 和 命题 3.3中的实例。首先注意，对于这两个实例，估计复杂度对数$|\mathcal{M}|$ 按 $O(1)$的尺度变化。因此：

- 定理 3.1使用半径 $\varepsilon(T)$，给出了极小极大风险的上界 $\widetilde{O}(1/\sqrt{T})$，且该界对 $\mathscr{H}_1$是紧的。
- 定理 3.2使用半径 $\varepsilon(T)$，给出了极小极大风险的下界 $\Omega(1/T)$，且该界对 $\mathscr{H}_2$是紧的。

也就是说，实例 $\mathscr{H}_1$ 表明，我们的上界无法通过使用半径 $\varepsilon(T)$ 来改进；实例 $\mathscr{H}_2$ 则表明，我们的下界无法通过使用半径 $\varepsilon(T)$ 来改进。更一般地，由于 $\mathsf{dec}_\varepsilon(\mathscr{H}_1)$ 和 $\mathsf{dec}_\varepsilon(\mathscr{H}_2)$ 的表现相同，但 $\mathscr{H}_1$ 和 $\mathscr{H}_2$ 的极小极大速率不同，因此，受约束 DEC 无法对 HR-DMSO 框架下的极小极大风险给出紧确刻画。这与（观测到奖励的）DMSO 框架中的情况形成对比；在 Foster 等人（2023）的研究中，只要 $\log|\mathcal{M}| = O(1)$，受约束 DEC 就能在对数因子范围内刻画极小极大速率。

顺便指出，命题 3.2 和 3.3 中构造的实例满足 假设 3.1 的正则性条件，适用于 C_{reg}，$c_{\mathrm{reg}} \le O(\log T)$ 以及所有 $\varepsilon \ge \varepsilon(T)$。因此，正则性条件不足以弥合上界与下界之间的差距。

排除更一般的刻画。 我们现在展示，上述强调的差距并不局限于 DEC，事实上，它们是广泛一类复杂度度量所固有的。我们的主要结果——定理 3.3 表明，对于任意满足温和假设的 f-散度 $D(\cdot\|\cdot)$，都可以构造两个 HR-DMSO 实例 $\mathscr{H}_1$ 和 $\mathscr{H}_2$，使得它们的极小极大风险相差一个多项式因子，但 1) 与 $\mathscr{H}_1$ 和 $\mathscr{H}_2$ 相关联的价值函数完全相同，并且 2) $D(\cdot\|\cdot)$-散度 $\mathscr{H}_1$ 和 $\mathscr{H}_2$ 中所有模型之间的成对散度完全相同。换句话说：

不可能仅根据价值函数和成对的 f-散度，得到对极小极大风险的紧确刻画。

定义 3.1 给出了我们关于 f-散度的主要技术假设：粗略地说，该假设指出，定义 f-散度的函数在 0 和 ∞ 附近至多呈多项式增长。

定义 3.1 (有界 f-散度)。考虑一个凸函数 ϕ： $[0,\infty) \to [0,\infty]$ 使得 $\phi(1) = 0$且 $\phi(x)$ 对所有 $x > 0$ 都是有限的，并令

$$D_\phi(\mathbb{P} \,\|\, \mathbb{Q}) := \mathbb{E}_{\mathbb{Q}}\left[\phi\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right] \tag{15}$$

表示概率测度 $\mathbb{P}$ 和 $\mathbb{Q}$ 之间的相关均衡散度，其中 $\mathbb{P} \ll \mathbb{Q}$。对于常数 α 和 $\beta \geq 0$，如果对所有 $x \geq 1$ 均有 ϕ 满足(α, β)-有界，则称 ϕ 为(α, β)-有界。

$$\phi(1/x) + \frac{\phi(x)}{x} \leq \beta \cdot x^\alpha.$$

在这种情况下，我们称 f-散度 D_ϕ 是(α, β)-有界的。

几乎所有常用的 f-散度在 α 和 β较小时都满足定义 3.1。对于海林格散度，我们有 $\phi(x) = (\sqrt{x-1})^2,$，因此 $D^2_{\mathsf{H}}(\cdot,\cdot)$ 是 (0, 2)-有界的；对于 KL散度，我们有 $\phi(x) = x\ln(x) + 1 - x$，因此 $D_{\mathsf{KL}}(\cdot \,\|\, \cdot)$ 是 (0, 2)-有界的；对于 χ^2-散度，我们有$\phi(x) = (x-1)^2,$，因此 $D_{\chi^2}(\cdot,\cdot)$ 是 (1, 1)-有界的。

注 3.1（ ϕ的非负性）。需要指出的是，在介绍 f-散度时，通常假设函数 ϕ映射到 $[-\infty,\infty]$（而非 $[0,\infty]$）。假设 ϕ映射到 $[0,\infty]$并不失一般性，理由如下。众所周知，对于任意 $c \in \mathbb{R}$，以及任意满足 $\sim\phi(1) = 0$的凸函数 ϕ，令 $\phi(x) = \phi(x) + c\cdot(x-1)$，则有 $D_\phi = D_{\tilde{\phi}}$。因此，给定任意 ϕ：$[0,\infty) \to [-\infty,\infty]$，我们可以任取 $c \in -\partial\phi(1)$，使得 $0 \in \partial(\phi(x) + c\cdot(x-1))$，这尤其意味着对于所有 x，均有 $\phi(x) + c\cdot(x-1) \geq 0$，并且由 $\phi(x) + c\cdot(x-1)$ 诱导的 f-散度等价于 D_ϕ。

定理 *3.3*。对于某些常数 $\alpha, \beta \geq 0$，假设 D_ϕ 是一个 *(α, β)*-有界的 f-散度。那么对于任意 $T \in \mathbb{N}$, $\epsilon \in$ *(0, 1)*，以及 $C_{\mathrm{prob}} \geq 1$，*HR-DMSO* 框架中存在实例 $\mathscr{H}_1 =$ *$(\mathcal{M}_1, \Pi_1, \mathcal{O}_1, \{f_1^M(\cdot)\}_{M\in\mathcal{M}_1})$*、 $\mathscr{H}_2 =$*$(\mathcal{M}_2, \Pi_2, \mathcal{O}_2, \{f_2^M(\cdot)\}_{M\in\mathcal{M}_2})$*，使得 $\Pi_1 = \Pi_2$，$\mathcal{O}_1 = \mathcal{O}_2$，并且存在一一映射 $\mathscr{E}$*:* $\mathcal{M}_1 \to \mathcal{M}_2$ 满足：

1. 对所有 $M \in \mathcal{M}_1$, $f_1^M \equiv f_2^{\mathscr{E}(M)}$。
2. 对所有 M， $M' \in \mathcal{M}_1$，以及 $\pi \in \Pi_1$， $D_\phi(M(\pi) \,\|\, M'(\pi)) = D_\phi(\mathscr{E}(M)(\pi) \,\|\, \mathscr{E}(M')(\pi))$.
3. 存在一个常数 C_ϕ，它仅取决于 ϕ，使得对于所有满足 $T \leq T' \leq T^{3/2-2\epsilon}\cdot(C_\phi C^{1/2+\epsilon}{}_{\text{概率}} \ln T)^{-1}$ 的 T'，都有

$$\mathfrak{M}(\mathscr{H}_1, T') \leq \frac{1}{T} + 2\cdot\left(\frac{C_{\mathrm{prob}}}{T}\right)^{1/2+\epsilon/(2\alpha)}, \qquad \mathfrak{M}(\mathscr{H}_2, T') \geq 2^{-2-2/\epsilon}\cdot\left(\frac{C_{\mathrm{prob}}}{T}\right)^{1/2}.$$

当 $\alpha = 0$ 时，定理 3.3 陈述中的量 $(C_{\mathrm{prob}}/T)^{1/2+\epsilon/(2\alpha)}$ 应解释为 0。特别地，如果 $D(\cdot \,\|\, \cdot)$ 是海林格散度或 KL 散度，那么在条目 3 中有$\mathfrak{M}(\mathscr{H}_1, T') \leq 1/T$，从而得到二次分离。如果 $D(\cdot \,\|\, \cdot)$ 是 χ^2-散度，那么有$\mathfrak{M}(\mathscr{H}_1, T') \leq O(1/T^{1/2+\epsilon/2})$，这会导致较小但仍为多项式级的分离，对常数 $\epsilon > 0$ 的任意取值均成立。

DEC 的若干变体及相关复杂度度量仅取决于价值函数 $f^M(\cdot)$（针对 $M \in \mathcal{M}$）以及类 $\mathcal{M}$ 中模型之间的成对 f-散度，因此无法在 HR-DMSO 框架中给出一个精确到多对数因子的极小极大风险刻画。下面，我们重点介绍几个值得注意的例子。

- 分布偏移 DEC（Foster 等，2021；Chen 等人，2022a；Foster 等，2023）定义在$\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$上：[9]$\mathrm{deco}_{\mathrm{rnd}}{}^\gamma(\mathscr{H}) = \sup_{\nu\in\Delta(\mathcal{M})} \inf_{p,q\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \Big[\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - \gamma\cdot\mathbb{E}_{\pi\sim q}\big[\mathbb{E}_{M\sim\nu}\big[D^2_{\mathrm{H}}\big(M(\pi)\,\|\,M(\pi)\big)\big]\big]$.

[9]我们在此考虑偏移 DEC 的 PAC 变体（Foster 等人，2023），但显然，我们的论证同样适用于 DEC 的遗憾版本（Foster 等人，2021）。

显然，该定义仅取决于价值函数 $\{f^M\}_{M\in\mathcal{M}}$和集合 $\mathcal{M}$ 中模型的成对海灵格距离，因此只能将极小极大风险刻画到二次因子以内。

- 偏移 DEC（Foster 等人，2021、2023）针对 $\mathscr{H}=(\mathcal{M},\ \Pi,\ \mathcal{O},\ \{f^M(\cdot)\}_M)$ 定义如下：

$$\mathsf{dec}^{\mathrm{o}}_{\gamma}(\mathscr{H})=\sup_{\overline{M}\in\mathrm{co}(\mathcal{M})}\inf_{p,q\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\mathbb{E}_{\pi\sim p}[f^M(\pi_M)-f^M(\pi)]-\gamma\cdot\mathbb{E}_{\pi\sim q}[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)].$$

注意，$\mathsf{dec}^{\mathrm{o}}_{\gamma}(\mathscr{H})$ 取决于 $\mathcal{M}$ 中的模型与 $\mathrm{co}(\mathcal{M})$ 中的模型之间的散度，而 定理 3.3 并未涵盖这种情况。不过，Foster 等人 （2023，命题 D.2）证明了 $\mathsf{dec}^{\mathrm{o}}_{\gamma}(\mathscr{H})\le\mathsf{dec}^{\mathrm{o}\ \mathrm{rnd}}_{\gamma/4}(\mathscr{H})\le\mathsf{dec}^{\mathrm{o}}_{\gamma/4}(\mathscr{H})$（也就是说，$\mathsf{dec}^{\mathrm{o}}_{\gamma}(\mathscr{H})$ 与 $\mathsf{dec}^{\mathrm{o}\ \mathrm{rnd}}_{\gamma}(\mathscr{H})$ 在常数因子意义下等价），因此，根据前一个要点可知，这种复杂度度量只能在二次因子的范围内刻画极小极大风险。

- Foster 等人 （2021、2023）考虑了适用于局部化的模型类子集 $\mathcal{M}$的 DEC 变体。具体而言，Foster 等人（2021）考虑了以下两种局部化概念：对于某个局部化半径 $\alpha>0$、一个模型类 $\mathcal{M}$和一个参考模型 M，

$$\begin{aligned}\mathcal{M}_\alpha(\overline{M}):=&\{M\in\mathcal{M}\ :\ f^M(\pi_M)\le f^{\overline{M}}(\pi_{\overline{M}})+\alpha\},\quad\text{and}\\ \mathcal{M}^\infty_\alpha(\overline{M}):=&\left\{M\in\mathcal{M}\ :\ \left|(f^M(\pi_M)-f^M(\pi))-(f^{\overline{M}}(\pi_{\overline{M}})-f^{\overline{M}}(\pi))\right|\le\alpha\ \forall\pi\in\Pi\right\}.\end{aligned}$$

由于这些定义仅依赖于价值函数 $\{f^M\}_{M\in\mathcal{M}}$，定理 3.3 意味着，将局部化纳入上述考察的 DEC 变体中，无法帮助刻画极小极大风险。

- 信息比（Russo 和 Van Roy，2014年，2018年；Lattimore 和 György，2021）最初用于界定后验采样以及一种称为信息导向采样的更一般算法的贝叶斯遗憾。模型类 $\mathcal{M}$ 的信息比与其凸包 $\mathcal{M}$的 DEC 密切相关；具体而言，Foster 等人（2022b）证明， $\mathcal{M}$ 的信息比的一个参数化版本与 $\mathcal{M}$的凸包的 DEC 等价，二者相差常数因子。由于 $\mathrm{co}(\mathcal{M})$ 的 DEC 涉及凸包 $\mathcal{M}$中模型之间的成对海灵格距离，定理 3.3 并未彻底排除它能够刻画极小极大风险的可能性。然而，在许多自然的例子中（例如表格型强化学习（Foster 等人（2022b））），已知 $\mathrm{co}(\mathcal{M})$ 的 DEC 远大于极小极大风险，因此它似乎不太可能给出紧确刻画。

信息比也有一些变体，定理 3.3确实排除了以下情况：给定一个参考模型 $M\in\mathcal{M}$ 和一个分布 $\mu\in\Delta(\mathcal{M})$，可以定义（Foster 等人，2021）

$$\mathcal{I}(\mathscr{H},\overline{M},\mu):=\arg\min_{p,q\in\Delta(\Pi)}\frac{(\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}[f^M(\pi_M)-f^M(\pi)])^2}{\mathbb{E}_{\pi\sim q}\mathbb{E}_{M\sim\mu}[D_{\mathsf{KL}}\big(M(\pi)\,\|\,\overline{M}(\pi)\big)]}.$$

由于这个定义仅取决于 $\mathcal{M}$ 中模型的价值函数和成对 KL 散度，定理 3.3，因此 $\mathcal{I}(\mathscr{H},M,\mu)$ 的任何函数（例如由 $\max_{M\in\mathcal{M}}\max_{\mu\in\Delta(\mathcal{M})}\mathcal{I}(\mathscr{H},\overline{M},\mu)$ 定义的信息比的最坏情况版本）都无法刻画极小极大风险。

- 请注意，一般而言，受约束 DEC $\mathsf{dec}_\varepsilon(\mathscr{H})=\sup_{M\in\mathrm{co}(\mathcal{M})}\mathsf{dec}_\varepsilon(\mathscr{H},M)$ 取决于 $\mathcal{M}$ 中模型与 $\mathrm{co}(\mathcal{M})$ 中模型之间的海林格散度，因此定理 3.3 并未直接排除用 $\mathsf{dec}_\varepsilon(\mathscr{H})$ 来刻画的可能性。不过，我们已经在命题 3.2 和 3.3 中分别排除了这种刻画。当然，变体 $\sup_{M\in\mathcal{M}}\mathsf{dec}_\varepsilon(\mathscr{H},M)$ 仅限于$M\in\mathcal{M}$，只取决于 $\mathcal{M}$中模型的价值函数及两两之间的海林格散度，因此已由定理 3.3 涵盖。

我们要指出的是，目前我们的结果尚未排除一种复杂度度量，即 Lattimore（2022年）在对抗性部分监测工作中所考虑的广义信息比率。该比率使用基于对数障碍函数的未归一化类 KL 散度，无法用 f-散度表示。由Lattimore （2022年）给出的遗憾上界和下界相差 $\mathrm{poly}(|\Pi|)$ 倍，因此我们认为，在 $T\ll\min\{|\mathcal{M}|,|\Pi|\}$ 的“大决策空间/模型类”情形下，这种复杂度度量不太可能给出紧的保证，而这正是我们工作的重点。

备注 3.2。 虽然这超出了本文的范围，但我们指出，对于遗憾（而非 PAC）框架，同样可以建立类似的不可能性结果。

3.3 MA-DMSO 框架的启示

到目前为止，本节的所有结果都针对 HR-DMSO 框架。利用 定理 2.1和 2.2，我们可以立即推导出 MA-DMSO 框架的类似结果。下文将陈述这些类似结果（特别是极小极大风险的上界和下界，以及更紧结果的不可能性），它们都是前文结果的推论。有关这些结果的应用，请参见附录的第一部分。

极小极大风险的上界和下界。我们首先根据多智能体 DEC，给出 MA-DMSO 实例的极小极大风险上界和下界；这些结果是定理 3.1 和3.2的推论。

推论 3.1 （MA-DMSO 的极小极大上界）。固定 $\delta \in \left(0, \frac{1}{10}\right)$ 和 $T \in \mathbb{N}$，并考虑任意一个K-玩家 MA-DMSO 实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$。假设对于所有 $k \in [K]$ 和$M \in \mathcal{M}$， $f_k^M(\cdot) \in [0,1]$，并令 $\varepsilon(T) := 16\sqrt{\frac{\lceil \log 2/\delta \rceil \cdot \log \frac{|\mathcal{M}|}{\delta}}{T}}$。那么我们有

$$\mathfrak{M}(\mathscr{M}, T) \leq \mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{M}) + K\delta. \tag{16}$$

推论 3.1 的证明。给定一个 MA-DMSO 实例 $\mathscr{M}$，根据定理 2.1，考虑实例 $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M\}_M)$。我们有，对于所有 $M \in \mathcal{M}$， $h^M(\cdot) \in [0, K]$；这意味着，在定理 2.1 证明中的构造下，对于所有 $M \in \mathcal{M}$， $f^M(\cdot) \in [-K+1, 1]$。对 $f^M(\cdot)$ 进行缩放后，定理 3.1 的保证确保 $\mathfrak{M}(\mathscr{H}, T) \leq \mathsf{dec}_{\varepsilon(T)}(\mathscr{H}) + K\delta$，结合定理 2.1 即可推出式 (16)。这里~还用到了以下事实：对价值函数 $f^M(\cdot)$ 进行缩放时， $\mathfrak{M}(\mathscr{H}, T)$ 和 $\mathsf{dec}_\varepsilon(\mathscr{H})$ 都会线性缩放。 □

正如我们将在注 D.1中进一步讨论的那样，定理 D.1 给出的高概率保证也适用于 MA-DMSO 场景，即推论 3.1的语境中。

推论 3.2 （MA-DMSO 的极小极大下界）。 考虑 MA-DMSO 框架的任意实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$，满足 $\mathcal{R} = [0,1]$。给定 $T \in \mathbb{N}$，令 $\varepsilon(T) > 0$ 尽可能大，并满足 $\varepsilon(T)^2 \cdot C(T) \cdot K \cdot T \leq \frac{1}{8} \cdot \mathsf{dec}^{\varepsilon(T)}(\mathscr{M})$。那么

$$\mathfrak{M}(\mathscr{M}, T) \geq \frac{1}{6} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{M}).$$

推论 3.2. 给定一个 MA-DMSO 的实例 $\mathscr{M}$，根据定理 2.1，考虑实例 $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M\}_M)$。根据 f^M 的定义，我们有 $\sup_{\pi \in \Pi, M \in \mathcal{M}} \sup_{\pi' \in \Pi} f^M(\pi') - f^M(\pi) \leq K$。于是我们有$\mathfrak{M}(\mathscr{M}, T) = \mathfrak{M}(\mathscr{H}, T) \geq \frac{1}{6} \cdot \mathsf{dec}^{\varepsilon(T)}(\mathscr{H}) = \frac{1}{6} \cdot \mathsf{dec}_{\varepsilon(T)}(\mathscr{M})$，其中两个等式利用了定理 2.1，而不等式利用了定理 3.2。 □

正如我们此前所指出的，命题 3.1给出了基于 DEC 的上界和下界之间差距的界；只要假设 3.1 成立，该命题就适用于 MA-DMSO 的实例。特别是，这意味着只要 $\mathsf{dec}_\varepsilon(\mathscr{M}) \propto \varepsilon^{1-\rho}$ for $\rho \in [0,1)$，我们就有

$$\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathscr{M}) \leq \widetilde{O}(K^{\frac{1-\rho}{1+\rho}} \log^{\frac{1-\rho}{2}} |\mathcal{M}|) \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathscr{M})^{\frac{1+\rho}{2}}.$$

间隔的紧性。接下来，我们为 MA-DMSO 给出命题 3.2 和 3.3 的类似结论。这些结果构造了 MA-DMSO 实例 $\mathscr{M}_1$ （命题 3.4）和 $\mathscr{M}_2$ （命题 3.5），它们表现出相同的 DEC 行为，即 $\mathsf{dec}_\varepsilon(\mathscr{M}_1) \asymp \varepsilon$ and $\mathsf{dec}_\varepsilon(\mathscr{M}_2) \asymp \varepsilon,$，但其极小极大速率为： $\mathfrak{M}(\mathscr{M}_1, T) \gtrsim 1/\sqrt{T}and\mathfrak{M}(\mathscr{M}_2, T) \lesssim \log(T)/T$。特别地，下面的命题 3.4 表明，在上界推论 3.1 中，尺度 $\varepsilon(T)$ 不能减小；下面的命题 3.5 表明，在下界推论 3.2 中，尺度 $\varepsilon(T)$ 不能增大。

命题 3.4。 对于任意充分大的 L、$A \in \mathbb{N}$，存在一个实例 $\mathscr{M}_1 = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$，满足 $\log|\mathcal{M}| \le \log(LA)$，并且具有以下性质：

1. 对所有 $T \le 2^{L/2}$，实例 $\mathscr{M}_1$ 的极小极大速率为 $\mathfrak{M}(\mathscr{M}_1, T) = \Theta(\sqrt{A/T})$。*2.* 对所有 $\varepsilon \in (2^{-L}, 1/\sqrt{A})$, *it holds that* $c \cdot \varepsilon\sqrt{A} \le \mathsf{dec}_\varepsilon(\mathscr{M}_1) \le C \cdot \varepsilon\sqrt{A}$, *for some constants* $c, C > 0$.

命题 3.4 的证明。我们注意到，用于证明命题 3.2 的实例 $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$，可立即导出由 $\mathscr{M}_1 = (\mathcal{M}, \Pi, \mathcal{O}, \Pi'_1, U_1)$ 给出的 MA-DMSO 单玩家实例，其中有 $\Pi'_1 = \Pi$ 和 $U_1(\pi'_1, \pi) = \pi'_1$，因为在 $\mathcal{M}$中的所有模型下都能观测到奖励。由此立即得到命题 3.2 的结论。 □

命题 3.5。 对于任意充分大的 $L \in \mathbb{N}$ 以及任意 $C_{\mathrm{prob}} \ge 1$，存在一个实例 $\mathscr{M}_2 = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$，其满足 对数 $|\mathcal{M}| \le O(L^2 + \text{对数 } C_{\mathrm{prob}})$，并具有以下性质：

1. For all $T \le 2^L$，实例 $\mathscr{M}_2$ 的极小极大速率有界于 $\mathfrak{M}(\mathscr{M}_2, T) \le 8C^2 \text{概率} \log T \sqrt{\frac{T}{\log(\cdot)\cdot L}}$.

2. 对所有 $\varepsilon \ge \frac{\sqrt{2}}{C_{\mathrm{prob}} \cdot 2^L}$，我们有 $\frac{C_{\mathrm{prob}}}{\sqrt{8}\cdot L} \cdot \varepsilon \le \mathrm{DEC}\varepsilon(\mathscr{M}_2) \le 2C_{\mathrm{prob}} \cdot \varepsilon$. 特别地，$\varepsilon(T) \ge \Omega\left(T^{-\frac{C_{\mathrm{prob}}}{T \log(\cdot)\cdot L}}\right)$ 只要 $T \le 2^L/L^3$。

命题 3.5 的证明。给定 L和 C_{prob}，令 $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$ 为按照~~~命题 3.3 给出的实例。接下来，令 $\mathscr{M}_2 = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 为按照定理 2.2 针对实例 $\mathscr{H}$构造的实例，其中 $V = 100 \cdot C_{\mathrm{prob}}^2 \cdot 2^{2L}$。我们有 $\log|\widetilde{\mathcal{M}}| = \log|\mathcal{M}| + \log V \le O(L^2 + \log C_{\mathrm{prob}})$。利用命题 3.3 和定理 2.2 的保证，我们有 $\mathfrak{M}(\mathscr{M}_2, T) \le \mathfrak{M}(\mathscr{H}, T) \le 8C^2 \text{prob} \log T \sqrt{\frac{T}{}}$对于 $T \le 2^L$，并且对于~~所有 $\varepsilon \ge \frac{\sqrt{2}}{C_{\mathrm{prob}} \cdot 2^L}$~~（这确保了 $\varepsilon - \sqrt{6/V} \ge \frac{\sqrt{2}}{C_{\mathrm{prob}} \cdot 2^L}$），

$$\frac{C_{\mathrm{prob}}}{\sqrt{8}\cdot L} \cdot \left(\varepsilon - \sqrt{6/V}\right) \le \mathsf{dec}_{\varepsilon - (6/V)^{-1/2}}(\mathscr{H}) - 6/\sqrt{V} \le \mathsf{dec}_\varepsilon(\mathscr{M}_2) \le \mathsf{dec}_\varepsilon(\mathscr{H}) \le 2C_{\mathrm{prob}} \cdot \varepsilon.$$

由于 $\varepsilon \ge \frac{1}{C_{\mathrm{prob}} \cdot 2^L}$意味着 $\sqrt{6/V} \le \varepsilon/2$，因此可得 $\frac{C_{\mathrm{prob}}}{2\sqrt{8}L} \cdot \varepsilon \le \mathsf{dec}_\varepsilon(\mathscr{M}_2) \le 2C_{\mathrm{prob}} \cdot \varepsilon$. □

排除更一般的刻画。 最后，我们给出 定理 3.3 在 MA-DMSO 框架中的一个类似结果，该结果表明，任何仅通过价值函数和成对 $\mathscr{M}$-散度依赖于实例 f的复杂度度量，都只能在多项式因子的范围内刻画极小极大风险。

定理 *3.4*。对于某些常数 $\alpha, \beta \ge 0$，假设 D_ϕ 是一个 (α, β)-有界的 f-散度（定义 *3.1*）。那么对于任意 $T \in \mathbb{N}$, $\epsilon > 0$ 和 $C_{\mathrm{prob}} \ge 1$，*MA-DMSO* 框架中存在实例 $\mathscr{M}_1 = (\mathcal{M}_1, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$、$\mathscr{M}_2 = (\mathcal{M}_2, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$，使得存在一个一一映射$\mathscr{E}$: $\mathcal{M}_1 \to \mathcal{M}_2$，满足：

1. 对所有 $M \in \mathcal{M}_1$, $f_1^M \equiv f_2^{\mathscr{E}(M)}$。*2.* 对所有 $M, M' \in \mathcal{M}_1$，以及 $\pi \in \Pi$, $D_\phi(M(\pi) \,\|\, M'(\pi)) = D_\phi(\mathscr{E}(M)(\pi) \,\|\, \mathscr{E}(M')(\pi))$。

3. 存在一个仅取决于 ϕ 的常数 C_ϕ，使得对于所有满足 $T \le T' \le T^{3/2-2\epsilon} \cdot (C_\phi C^{1/2+\epsilon} \text{概率} \ln T)^{-1}$ 的 T'，都有

$$\mathfrak{M}(\mathscr{M}_1, T') \le \frac{1}{T} + 2 \cdot \left(\frac{C_{\mathrm{prob}}}{T}\right)^{1/2+\epsilon/(2\alpha)}, \quad \textit{yet} \quad \mathfrak{M}(\mathscr{M}_2, T') \ge 2^{-3-2/\epsilon} \cdot \left(\frac{C_{\mathrm{prob}}}{T}\right)^{1/2}.$$

该证明利用 定理 2.2 的等价性，将定理 3.3 中的 HR-DMSO 实例构造转换到 MA-DMSO 框架。由于 定理 3.3 讨论的是成对的 f-散度，而非该实例的受约束 DEC，$\mathsf{dec}_\varepsilon(\mathscr{H})$，因此我们无法完全以黑箱方式应用 定理 2.2，但其证明中的大部分推理仍然适用。

4 MA-DMSO：从多智能体到单智能体

在基于多智能体决策估计系数为 MA-DMSO 框架确立极小极大风险的上界和下界后，本文余下部分将给出结构性结果，以便将主要风险界应用于我们感兴趣的具体设定。为此，在本节中，我们给出一般性结果，说明在何种条件下多智能体 DEC 可由单智能体 DEC 控制，从而将单智能体设定中大量现有结果（Foster 等人，2021，2023）推广至多个智能体。

诱导的单智能体模型类。考虑 MA-DMSO 框架（定义 1.1）中的纳什均衡实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$，回顾一下 $\Pi = \Pi_1 \times \cdots \times \Pi_K$。我们将证明，该实例 $\mathscr{M}$ 的多智能体 DEC 上界可由一组基于 $\mathscr{M}$ 定义的 $\widetilde{\ }$ $\widetilde{\ }$ 诱导的单智能体模型类 $\mathcal{M}_k$ 的单智能体 DEC 来表示。为定义模型类 $\mathcal{M}_k$，对于 $M \in \mathcal{M}$ 和 $k \in [K]$，我们首先定义单智能体模型 $M|_k$ 如下：从 $M|_k(\pi)$ 抽取的纯观测，其分布与当 $o_\circ \sim M(\pi)$ 时的纯观测 $o_\circ$ 的分布相同；从 $M|_k(\pi)$ 抽取的奖励，其分布与当 $(r_1, \ldots\ldots, r_K) \sim M(\pi)$ 时的 r_k 的分布相同。换句话说，模型 $M|_k$ 与 M 完全相同，只是不考虑除 k 之外所有智能体的奖励。

$\widetilde{\ }$ $\widetilde{\ }$ 单智能体模型类 $\mathcal{M}_k$ 被定义为具有策略空间 Π_k，因此，$\mathcal{M}_k$ 中的模型是映射 $\sim M : \Pi_k \to \Delta(\mathcal{O}_\circ \times \mathcal{R})$。此外，$\mathcal{M}_k$ 由 $\Pi_{-k} \times \mathcal{M}$ 索引，其模型定义如下：

$$\widetilde{\mathcal{M}}_k = \{\pi_k \mapsto M|_k(\pi_k, \pi_{-k}) \ : \ \pi_{-k} \in \Pi_{-k}, \ M \in \mathcal{M}\}. \tag{17}$$

该定义背后的直觉是，对于每个智能体 k，如果其他智能体承诺采取 π_{-k}，这就为 k诱导出一个“单智能体”环境 $\widetilde{\ }$。如果 $M \in \mathcal{M}$ 是原始环境，那么模型 $M|_k(\cdot, \pi_{-k}) \in \mathcal{M}_k$正是为 k 诱导出的单智能体环境（在一种每个智能体都能观察自身奖励、但无法观察其他智能体奖励的去中心化协议中）。

偏移决策—估计系数。 本节中的结果用 偏移 版本的 DEC 来表述最为自然，该版本由 Foster 等人 (2021) 引入——具体而言，是 遗憾 版本，该版本限制于 $p = q$（即探索与利用相耦合）。对于一个实例 $\mathscr{M}$、参考模型 M和尺度参数 $\gamma > 0$，我们定义

$$\text{r-dec}^{\text{o}}_\gamma(\mathscr{M}, \overline{M}) := \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{\mathbb{E}_{\pi \sim p}[h^M(\pi)] - \gamma \cdot \mathbb{E}_{\pi \sim p}[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)]\right\}. \tag{18}$$

我们通过 Foster 等人 (2023) 指出，这一概念可以与受约束（PAC）决策估计系数建立如下联系。

命题 4.1 (Foster 等人 (2023))**.** 对于所有 $M \in \mathcal{M}^{\mp}$ 和 $\varepsilon > 0$,

$$\text{dec}_\varepsilon(\mathscr{M}, \overline{M}) \le \inf_{\gamma > 0} \left\{\text{r-dec}^{\text{o}}_\gamma(\mathscr{M}, \overline{M}) \vee 0 + \gamma\varepsilon^2\right\}. \tag{19}$$

命题 4.1 足以为我们将要考虑的所有示例推导出受约束 DEC 的紧界。也可以反向关联这两个复杂度度量，但这可能会导致松散的结果（Foster 等人，2023）；这对我们的目的并非必要。

4.1 凸决策空间的 MA-DEC 界

我们的第一个结果考察了一类一般实例，其中智能体的决策空间 Π_k 满足一项 凸性性质，具体表述如下：假设 4.1。

假设 4.1 (决策空间的凸性)。对于每个 $k \in [K]$，存在一个有限集 $\mathcal{A}_k$（称为纯决策集），使得 $\Pi_k = \Delta(\mathcal{A}_k)$。此外，以下条件成立：

1. 每个 $M \in \mathcal{M}$ 关于 π 都是线性的，即，对于 $\pi \in \Pi$, $M(\pi) = \mathbb{E}_{a_k \sim \pi_k \forall k}[M(a)]$,，其中我们记 $a = (a_1, \ldots, a_K)$。
2. 存在一个可测函数 $\varphi : \mathcal{O} \to \mathcal{A}$，使得对于所有 $a \in \mathcal{A}$ 和 $M \in \mathcal{M}$，$\mathbb{P}_{o \sim M(a)}(\varphi(o) = a) = 1$，也就是说，$M(a)$ 揭示了 a。

这一假设相当宽松，只要玩家 1）可以随机选择动作，并且 2）能够观察每一轮采样得到的动作，就能满足该假设。特别地，这涵盖了带有老虎机反馈的（结构化）标准形博弈（参见附录 A.3）。为简化记号，我们将记$\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$和 $\mathcal{A}_{-k} = \prod_{k' \neq k} \mathcal{A}_{k'}$。

本小节的主要结果，定理 4.1，表明，对于任意 $M \in \Delta(\mathcal{M})$，我们都可以用模型类 $K \sim \mathcal{M}_k$ 和参考模型 M_k的单智能体决策估计系数$\overline{\text{r-dec}}^{\text{o}}_{\gamma/K}(\mathcal{M}_k, M_k)$来界定$\widetilde{\ }$多智能体 DEC $\text{r-dec}^{\text{o}}_{\gamma}(\overline{\mathcal{M}}, \overline{M})$。

定理 4.1（重述定理 1.5）。设 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 是 *MA-DMSO* 框架中的一个纳什均衡实例，且满足假设 *4.1*。那么，对于任意 $\gamma > 0$，都有

$$\sup_{\overline{M} \in \text{co}(\mathcal{M})} \text{r-dec}^{\text{o}}_{\gamma}(\mathscr{M}, \overline{M}) \leq \sum_{k=1}^{K} \sup_{\overline{M}_k \in \text{co}(\widetilde{\mathcal{M}}_k)} \text{r-dec}^{\text{o}}_{\gamma/K}(\widetilde{\mathcal{M}}_k, \overline{M}_k).$$

这一结果颇为直观：它表明，集中式均衡计算的复杂度，不会高于各智能体在其他玩家所诱导的最坏情形环境中优化自身奖励所需的复杂度。证明采用以下不动点论证：给定$\widetilde{\ }$智能体 k，如果所有其他智能体都选定一个联合分布，就会诱导出一个单智能体 DMSO 类 $\mathcal{M}_k$，因此，智能体 k 采用能使该类的单智能体 DEC 最小化的策略是很自然的。但仅凭这一点还不足以界定 MA-DEC，因为我们需要为所有智能体指定策略；不过，通过应用卡库塔尼不动点定理，我们证明了所有 K 智能体可以同时针对其他智能体的策略，使各自的单智能体 DEC 最小化。此外，我们指出，定理 4.1 的一个直接推论是，$\text{r-dec}^{\text{o}}_{\gamma}$（$\mathscr{M}$）的同一上界在 $\mathscr{M}$ 是 CCE 或 CE 实例时也成立，因为纳什均衡总是（粗）相关均衡（见附录 A.3）。

举一个具体例子，对于具有 A 个动作的多臂老虎机问题，我们有 $\text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M}) \leq O(\frac{A}{\gamma})$(Foster 等人，2021)。根据 定理 4.1，可知如果 $\mathscr{M}$ 是具有老虎机反馈的 K人标准形博弈的类，且每位玩家有 A_k 个动作，那么

$$\sup_{\overline{M} \in \text{co}(\mathcal{M})} \text{r-dec}^{\text{o}}_{\gamma}(\mathscr{M}, \overline{M}) \leq O(K) \cdot \frac{\sum_{k=1}^{K} A_k}{\gamma}.$$

根据命题 4.1，我们得出$\text{dec}_{\varepsilon}(\mathscr{M}) \leq O\left(\varepsilon \cdot \sqrt{K \sum_{k=1}^{K} A_k}\right)$。有关细节以及更多示例，请参阅附录 A.3，其中包括收益为线性或凹函数的结构化正规形博弈。对于其中许多示例，应用定理 4.1 可得出$\text{r-dec}^{\text{o}}_{\gamma}(\mathscr{M})$的界，且该界几乎是紧的。但情况并非总是如此：在命题 A.11（附录 A.3.5）中，我们证明存在一些实例 $\mathscr{M}\widetilde{\ }$，使得$\text{r-dec}^{\text{o}}_{\gamma}(\mathcal{M}_k)$远大于$\text{r-dec}^{\text{o}}_{\gamma}(\mathscr{M})$。

4.2 马尔可夫博弈的 MA-DEC 界

虽然 假设 4.1 相当一般，适用于大多数标准形博弈设置，但它无法涵盖的一种重要情形是马尔可夫博弈，其中联合决策空间 Π由随机非平稳策略构成（形式化定义见下文的假设 4.2）。[10] 本节给出一个类似的结果，专门适用于这种一般的非凸设置。

假设 4.2 (马尔可夫博弈实例)。该实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 满足：对于某个 $H \in \mathbb{N}$、有限状态空间 $\mathcal{S}$以及有限联合动作空间 $\mathcal{A} = \prod_{k=1}^{K} \mathcal{A}_k$，每个模型 $M \in \mathcal{M}$ 都是一个K玩家、时域为H 的马尔可夫博弈，其状态空间为 $\mathcal{S}$，联合动作空间为 $\mathcal{A}$ *(*参见 *例1.2*）。此外，对于每个 k，类 Π_k 由非平稳随机马尔可夫策略构成，即：

$$\Pi_k = \{(\pi_{k,1}, \ldots, \pi_{k,H}) \mid \pi_{k,h} : \mathcal{S} \to \Delta(\mathcal{A}_k) \ \ \forall h \in [H]\}.$$

[10]一种可能的做法是，通过对每个智能体的决策空间进行凸化来满足假设 41 Π_k；然而，在马尔可夫$\widetilde{\ }$博弈设定中，这会导致 $\mathcal{M}_k$ 在(17) 中定义的模型类 π_{-k} 变得庞大，因为策略 $\widetilde{\mathcal{M}}_k$ 将变为非平稳马尔可夫策略的混合分布。具体而言，由此得到的模型类的 DEC 通常会随着 MDP 混合分布类的 DEC 一同增长，即使在表格设定下也呈指数级（Foster 等人，2022b）。

有限性 $\mathcal{S}$ 和 $\mathcal{A}$ 在 假设 4.2中的设定出于技术原因，旨在能够应用不动点定理；本节中的界在数量上不会依赖于 $|\mathcal{S}|$ 或 $|\mathcal{A}|$，我们预计这一假设可以放宽。

在假设 4.2下，我们给出定理 4.1的以下类似结论。

定理 4.2。 存在一个常数 $C>0$ 使得以下结论成立。设 $\mathscr{M}=(\mathcal{M},\Pi,\mathcal{O},\{\Pi'_k\}_k,\{U_k\}_k)$是满足假设 *4.2*的 *DMSO* 框架的一个纳什均衡实例。则对于任意 $\gamma>0$，有

$$\sup_{\overline{M}\in\mathcal{M}} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M},\overline{M}) \le \frac{CKH\log H}{\gamma} + \sum_{k=1}^{K} \sup_{\overline{M}_k\in\widetilde{\mathcal{M}}_k} \textsf{r-dec}^{\mathrm{o}}_{\gamma/(CKH\log H)}(\widetilde{\mathcal{M}}_k,\overline{M}_k).$$

例如，当 $\mathcal{M}$ 是具有 $|\mathcal{S}|=S$ 和 $|\mathcal{A}|=A$的表格型马尔可夫决策过程类时，我们有 $\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}) \le \mathrm{poly}(S,A,H)\cdot\gamma$（Foster 等，2021）。定理 4.2 随即表明，对于具有 $|\mathcal{S}|\le S$ 和$|\mathcal{A}_k|\le A$的表格型马尔可夫博弈，有 $\sup_{M\in\mathcal{M}}\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M},M) \le \mathrm{poly}(S,A,H,K)\cdot\gamma$，并且通过式 (19),

$$\sup_{\overline{M}\in\mathcal{M}} \textsf{dec}_{\varepsilon}(\mathscr{M},\overline{M}) \le \varepsilon\cdot\sqrt{\mathrm{poly}(S,A,H,K)}.$$

我们指出，尽管 定理 4.1 允许使用不适当的参考模型 $M\in\mathrm{co}(\mathcal{M})$，定理 4.2 则仅适用于适当的参考模型 $M\in\mathcal{M}$，因此主要适用于可使用适当估计量的情形（例如表格型 MG）。示例及更多细节请参见 附录 A.3 。

5 MA-DMSO：关于多智能体诅咒

在多智能体强化学习研究中经常遇到的一个难题，是样本复杂度随智能体数量增加而扩展不佳 K。具体而言，直接估计模型 $M^\star$ 或智能体的 Q-价值函数的算法，通常会导致样本复杂度关于 K呈指数增长，因为模型和智能体的 Q-价值函数至少需要 $\exp(K)$ 个参数才能表示；这一现象被称为多智能体诅咒 (Jin 等人，2021b）。本节将通过多智能体决策估计系数的视角，探究 MA-DMSO 框架中的多智能体诅咒。

我们首先指出，我们的上界定理 3.1（以及更一般的版本定理 D.1）中关于 DEC 的极小极大风险上界，确实受到多智能体诅咒的影响：即使对于非常简单的模型类，例如 K 玩家标准形博弈，定理 3.1 中的估计误差 $\log|\mathcal{M}|$ 也会随 K 指数增长（详细示例和讨论见附录 A.3），因此定理 3.1 的上界也会随 K 指数增长，尽管 MA-DEC 本身并非指数级。请注意，我们的下界（定理 3.2）对 K 的依赖并非指数级，因为 (a) DEC 通常按 $\textsf{dec}_{\varepsilon}(\mathscr{M}) \asymp C_{\mathrm{prob}}\cdot\varepsilon$, 的形式增长，其中问题相关常数 C_{prob} 仅取决于智能体各自行动集的大小，因此避免了随 K 指数增长；(b) 定理 3.2 的界不包含任何涉及模型估计误差的项（尤其不会乘以评估 DEC 时所用的尺度 $\varepsilon(T)$）。[11]

–

规避多智能体诅咒。多智能体（老虎机）学习领域的经典结果表明，至少对于具有老虎机反馈的多人标准型博弈，多智能体诅咒并非不可避免：如果每位玩家都运行一个对抗性老虎机无悔算法，那么经过 T 个时间步后，其联合动作配置的经验平均会以$\mathrm{poly}(K,\max_k A_k)/\sqrt{T}$的速率趋近于该博弈的（粗）相关均衡（例如, Rakhlin 和 Sridharan（2013年）），其中 A_k 是玩家k的动作数。此外，近期一系列研究已将这些结果推广到马尔可夫博弈设定（Jin 等人，2021b；Song 等人，2021；Mao 和 Basar，2022年）。

[11]我们回顾一下，即使在单智能体设定中，上界中出现估计误差项而下界中没有，也会导致二者之间存在差距。 Foster 等人（2023）强调，缩小这一差距是一个重要的开放问题。

自然会让人想知道，我们能否通过本框架和多智能体决策估计系数来刻画这些避免随 K 指数级增长的结果。鉴于上述讨论，这个问题可以转化为：定理 3.1 中的 $\log|\mathcal{M}|$ 项（更一般地说，定理 D.1 中的 $\mathrm{Est}_{\mathsf{H}}(T)$ 项，该项可通过覆盖数加以控制）是否能够减小；该项源自估计误差。请注意，一般而言，正如 Foster 等人（2021）所观察到的，即使在单智能体设定中，定理 3.1 中出现的估计误差项 $\log|\mathcal{M}|$也无法完全消除，但或许可以用一个更弱的量来替代它。一种可能的途径是（如果可行的话）用 $\log|\mathcal{F}_{\mathcal{M}}|$ 替换 $\log|\mathcal{M}|$，其中 $\mathcal{F}_{\mathcal{M}}$ 表示诱导出的价值函数类；Foster 等人（2022a）在单智能体设定中探索了这一方法，并由此为无模型强化学习设定带来了更紧的保证。然而，这种方法不足以规避多智能体诅咒，因为即使对于具有老虎机反馈的标准形博弈（例 1.1），价值函数类 $\mathcal{F}$ 的（ε-覆盖）规模通常也会随 K 指数级增长。

鉴于上述讨论，规避多智能体诅咒最有希望的方法或许是寻求类似于定理 3.1 的界，但将因子 $\log|\mathcal{M}|$ 替换为智能体决策集合大小的对数。事实上，联合（纯）决策集合的大小通常不会随 K 呈指数级增长。例如，对于每个玩家有 A 个动作的 K 人标准形博弈，纯动作配置的数量为 A^K，因此其对数仅随 K 线性增长；等价地，我们可以寻求随各智能体单独决策集合大小的对数之和增长的界。在单智能体 DMSO 场景中，Foster 等人（2021, 2022b）确实得到了随 $\log|\Pi|$ 增长的界，而不是随 $\log|\mathcal{M}|$ 增长。不过，这种改进也有代价：Foster 等人（2021, 2022b）以 $\log|\Pi|$ 替换 $\log|\mathcal{M}|$ 的上界，依赖于 $\mathcal{M}$的凸包的 DEC，而不是 $\mathcal{M}$ 本身的 DEC。

我们的上界。 在下文的 定理 5.1 中，我们给出一个上界，以 $\log|\mathcal{M}|$替代 定理 3.1 中出现的因子，并改为 $\max_k \log|\Pi'_k|$，代价是其随该实例凸化版本的 MA-DEC 缩放 $\mathscr{M}$。在 CCE 实例（$\max_k \log|\Pi'_k|$）的特殊情况下，该量等于 $\max_k \log(|\Sigma_k|+1)$，但对于 CE 实例（定义 1.2）也很小（定义 A.1）；对于下面这种更一般的相关均衡概念，它同样很小，我们称之为“广义相关均衡”。

假设 5.1 (广义相关均衡)。 若 *MA-DMSO* 实例 $\mathscr{M}=(\mathcal{M},\Pi,\mathcal{O},\{\Pi'_k\}_k,\{U_k\}_k)$满足以下条件，则称其满足广义相关均衡假设：存在 $\Pi=\Delta(\Sigma_1\times\cdots\times\Sigma_K)$，其中 $\Sigma_1,\ldots,\Sigma_K$ 为有限集，称为纯决策集。此外，记 $\Sigma:=\Sigma_1\times\cdots\times\Sigma_K$，该实例 $\mathscr{M}$ 满足：

1. 每个 $M\in\mathcal{M}$ 关于 π 都是线性的，即对于 $\pi\in\Pi$, $M(\pi)=\mathbb{E}_{\sigma\sim\pi}[M(\sigma)]$。
2. 偏离函数 U_k 满足线性性质，具体而言，对于所有 $k\in[K]$, $M\in\mathcal{M}$，以及 $\pi\in\Pi,\pi'_k\in\Pi'_k$，都有 $f_k^M(U_k(\pi'_k,\pi))=\mathbb{E}_{\sigma\sim\pi}[f_k^M(U_k(\pi'_k,\sigma)]$。

只要纯决策集 Σ_k 均为有限集，就很容易验证，粗相关均衡实例（定义 1.2）和相关均衡实例（定义 A.1）都满足 假设 5.1 。

为陈述我们的结果，对于 MA-DMSO 框架中的一个实例 $\mathscr{M}=(\mathcal{M},\ \Pi,\ \mathcal{O},\ \{\Pi'_k\}_k,\ \{U_k\}_k)$，我们将实例 $\mathscr{M}$的凸包 定义为实例 $\mathrm{co}(\mathscr{M})$：$=(\mathrm{co}(\mathcal{M})\ \Pi,\ \mathcal{O},\ \{\Pi'_k\}_k,\ \{U_k\}_k)$。结合上一节的结果，我们的保证用 MA-DEC 的遗憾变体（$\mathsf{r}\text{-}\mathsf{dec}^{\mathrm{o}}_{\gamma}$；参见 (18)）来表述最为自然。

定理 5.1。设 $\mathscr{M}=(\mathcal{M},\ \Pi,\ \mathcal{O},\ \{\Pi'_k\}_k,\ \{U_k\}_k)$ 是满足假设 *5.1*的 *MA-DMSO* 实例。那么，对于任意 $T\in\mathbb{N}$和 $\delta\in(0,\ 1)$，存在一种算法（MAEx0；附录*F*中的算法 *1*），该算法生成 $\widehat{\pi}\in\Pi$，且其成立的概率至少为 $1-\delta$，

$$\mathbf{Risk}(T)=h^{M^\star}(\widehat{\pi})\le O(K)\cdot\inf_{\gamma>0}\left\{\mathsf{r}\text{-}\mathsf{dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathscr{M}))+\frac{\gamma}{T}\cdot\log\left(\frac{K\cdot\max_k|\Pi'_k|}{\delta}\right)\right\}.$$

我们将这一结果视为对单智能体设定中以 $\log|\mathcal{M}|$ 替代 $\log|\Pi|$ 的保证的扩展（Foster 等人，2021，2022b）；与此前的结果一样，其代价在于，DEC 被应用于凸包

实例。对于计算具有 K 个玩家、每个玩家有 A 个动作的标准式博弈中的粗相关均衡（CCE）问题，我们有 $\mathsf{dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathscr{M})) \lesssim \frac{A}{\gamma}$ and $\max_k \log|\Pi'_k| = \log(A)$，因此这一结果给出了

$$\mathbf{Risk}(T) \lesssim \sqrt{\frac{\mathrm{poly}(K)\cdot A}{T}};$$

见 附录 A.3了解详情和更多示例。定理 5.1 表明，对于凸类，可以避免多智能体诅咒，并为具有老虎机反馈的标准形博弈中的结构化类提供紧确保证，例如具有线性或凸收益的博弈。不过，一般而言，对于马尔可夫博弈等非凸类，它无法给出紧确保证。我们通过改造单智能体设定中的强大探索-优化算法（Lattimore, 2022年；Foster 等人，2022b）来证明这一结果；这种改造利用了多智能体情形特有的反馈结构。或许有人会想知道，定理 5.1 的保证与以下方法所得的结果相比如何：让每个智能体 k 分别运行Foster 等人（2022b）提出的（单智能体）探索-优化算法（应用于(17)中定义的模型类 $\mathcal{M}_k$），再利用Foster 等人（2022b）针对每个智能体给出的遗憾界来获得近似粗相关均衡。正如命题 A.11所示，定理 5.1 的保证可能远优于这一替代方法，因为前者涉及多智能体 DEC，即$\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathscr{M}))$，它可能远小于单智能体类对应的 DEC，即$\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathcal{M}_k))$。

将结果扩展到无限决策集。接下来，我们将说明如何把定理 5.1 的保证扩展到纯决策集 Σ_k 和偏离集合 Π'_k 并非有限的情形。我们将重点讨论粗相关均衡实例：考虑一个满足假设 5.1 的 MA-DMSO 实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$。对于每个 k，考虑子集 $\Sigma_k \subseteq \Sigma_k$ 和 $\Pi'_k \subseteq \Pi'_k$，并记 $\Pi = \Delta(\Sigma_1 \times \cdots \times \Sigma_K) \subset \Pi$。（例如，如果 $\mathscr{M}$ 是粗相关均衡实例，我们通常取 $\Pi'_k = \Sigma_k \cup \{\perp\}$。）不难看出，实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 满足假设 5.1（纯决策集为 Σ_k）。现在，我们定义一种意义，在这种意义下，实例 $\mathscr{M}$ 是 $\mathscr{M}$ 的良好覆盖。

定义 5.1。令 $\mathscr{M}$、$\mathscr{M}$如上定义。对于 $\varepsilon \geq 0$，如果 $\mathscr{M}$是 $\mathscr{M}$ 的一个 ε-决策空间覆盖，则称其为 ε 马会?

$$\forall M \in \mathcal{M}, \quad \forall k \in [K], \quad \forall \widetilde{\pi} \in \widetilde{\Pi}, \quad \exists \widetilde{\pi}'_k \in \widetilde{\Pi}'_k \quad s.t. \quad \max_{\pi'_k \in \Pi'_k} f^M_k(U_k(\pi'_k, \widetilde{\pi})) - f^M_k(U_k(\widetilde{\pi}'_k, \widetilde{\pi})) \leq \frac{\varepsilon}{K}.$$

我们令 $\mathcal{N}_{\Pi}(\mathscr{M}, \varepsilon) := \max_{k\in[K]} |\Pi'_k|$ 表示最小此类覆盖中最大偏离集合的大小，并对 $T \in \mathbb{N}$ 定义，

$$\mathsf{est}_{\Pi}(\mathscr{M}, T) = \inf_{\varepsilon \geq 0} \left\{ \log \mathcal{N}_{\Pi}(\mathscr{M}, \varepsilon) + \varepsilon T \right\}.$$

令 $\mathscr{M} = (\mathcal{M}$、$\Pi$、$\mathcal{O}$、$\{\Pi'_k\}_k$、$\{U_k\}_k)$ 为 ε-决策空间覆盖，用于 $\mathscr{M}$。注意，对于任意 $\widehat{\pi} \in \Pi$，根据定义 5.1可知

$$h^M(\widehat{\pi}) = \sum_{k=1}^{K} \max_{\pi'_k \in \Pi'_k} f^M_k(U_k(\pi'_k, \widehat{\pi})) - f^M_k(\widehat{\pi}) \leq \sum_{k=1}^{K} \max_{\widetilde{\pi}'_k \in \widetilde{\Pi}_k} f^M_k(U_k(\widetilde{\pi}'_k, \widehat{\pi})) - f^M_k(\widehat{\pi}) + \varepsilon.$$

因此，将定理 5.1 的算法应用于实例 $\mathscr{M}$的适当决策空间覆盖（选择适当的 ε），即可立即得到以下推论：

推论 5.1。假设 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 是一个满足 假设 5.1 的 MA-DMSO 实例。那么，对于任意 $T \in \mathbb{N}$ 和 $\delta \in (0,1)$，存在一种算法，可生成 $\widehat{\pi} \in \Pi$，使得其概率至少为 $1-\delta$。

$$\mathbf{Risk}(T) = h^{M^\star}(\widehat{\pi}) \leq O(K) \cdot \inf_{\gamma > 0} \left\{ \mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathscr{M})) + \frac{\gamma}{T} \cdot (\mathsf{est}_{\Pi}(\mathscr{M}, T) + \log(K/\delta)) \right\}.$$

纳什均衡实例的下界。 定理 5.1 依赖于以下假设： $\mathscr{M}$ 是一个广义相关均衡实例（假设 5.1）。为结束本节，我们通过证明无法为纳什均衡实现类似的保证来补充这一结果。首先，在 命题 5.1中，我们针对 K人玩家的纳什均衡实例证明了这一不可能性结果：我们给出一个实例，其中 定理 5.1 中的上界关于 K是多项式级的，但极小极大风险关于 K是指数级的。

命题 5.1。 存在一个常数 $c_0 > 0$，使得以下结论成立。对于任意 $K \in \mathbb{N}$，存在一个 K-玩家纳什均衡实例 $\mathscr{M} =$（$\mathcal{M}$、Π、$\mathcal{O}$、$\{\Pi'_k\}_k$、$\{U_k\}_k$），使得：

1. 最大值$_k |\Pi'_k| = 2$。*2.* 对于所有 $\gamma > 0$，
r-dec$^{\mathrm{o}}_\gamma$（co($\mathscr{M}$)） $\le O(K/\gamma)$。

3. 不存在任何算法能够自适应地抽取 $2^{o(K)}$ 个样本，并输出一个期望风险至多为 $c_0 \cdot K$ 的策略。

对于命题 5.1 中的实例 $\mathscr{M}$，我们有 $|\Pi'_k| = O(1)$，因此，定理 5.1 中形式的界将~意味着只需 $O(\mathrm{poly}(K)/\epsilon^2)$ 个样本即可学习一个 ϵ-近似纳什均衡；而命题 5.1 给出的 $2^{\Omega(K)}$ 的样本复杂度下界排除了这种可能性。命题 5.1 的证明直接源于关于 K-玩家纳什均衡查询复杂度的众所周知的下界（Rubinstein，2016；Babichenko，2016；Chen 等人，2017）。

作为我们的最后一个结果 定理 5.2，我们更进一步，证明对于纳什均衡实例，即使在 $K = 2$ 且博弈为零和的情况下，证明任何形式的定理 5.1 都是不可能的，这一结论依然成立。

定理 5.2。 存在一个常数 $C_0 > 0$，使得以下结论成立。任取满足 $N \ge C_0$ 且$\epsilon \in (1/N, 1)$的 $N \in \mathbb{N}$。存在一个双人零和纳什均衡实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$，使得以下结论成立：

1. 最大值$\{|\Pi'_1|, |\Pi'_2|\} \le |\Pi| \le C_0 \cdot N^2/\epsilon^2$。
2. 对所有 $\gamma \ge C_0$，r-dec$^{\mathrm{o}}_\gamma(\mathrm{co}(\mathscr{M})) \le \epsilon$。

3. 不存在一种算法，能够自适应地抽取 $\sqrt{N/C_0}$个样本，并输出一个期望风险至多为 $1/C_0$ 的策略。

注意，对于 $\mathscr{M}$ 中定理 5.2的实例，我们有$\log|\Pi| \lesssim \log(N/\epsilon)$，因此，定理 5.1 中的形式的界意味着大约需要 $\log(N/\epsilon)$ ϵ 个样本，就足以学得一个 ϵ-近似均衡。定理 5.2 中的样本复杂度下界表明，至少需要$\Omega(\sqrt{N})$ 个样本，因此以相当强的方式排除了这种类型的保证。

我们指出， $\mathscr{M}$ 中构造的定理 5.2实例虽然按照定义 1.1是一个纳什均衡实例，但并不对应于标准形博弈中混合纳什均衡的标准概念（参见定义 1.1之后的讨论）。由于双人零和博弈中粗相关均衡的边缘分布构成混合纳什均衡，定理 5.1 排除了对定理 5.2 作出如下加强：构造一个对应于混合纳什均衡标准概念的纳什均衡实例。

定理5.2 的证明（考虑到已有研究成果）比命题 5.1的证明难得多。它利用经典的支撑集估计问题（例如，Paninski (2008年)；Canonne (2020年)）构造一个实例，使其 DEC 很小，但极小极大风险很大。这一思路很自然，因为支撑集估计问题具有较大的模型估计误差，而命题D.1的上界涉及模型估计误差，因此该实例必须满足这一上界 $\mathscr{M}$。以支撑集估计问题为构建模块，我们构造了一类双人零和博弈，这与定理 2.2证明中的构造有些相似。然而，后一个结果中的构造无法确保r-dec$^{\mathrm{o}}_\gamma(\mathrm{co}(\mathscr{M}))$ 仍然很小，因此需要采用更精巧的方法。为确保r-dec$^{\mathrm{o}}_\gamma(\mathrm{co}(\mathscr{M}))$保持较小，同时维持极小极大风险的下界，我们需要在构造中嵌入一些额外组件，即里德–所罗门码与随机性提取器的复合。详情请参见附录 G 。

致谢

感谢 Rob Schapire、Yunzong Xu 和 Yanjun Han 提供有益的评论并参与讨论。NG 在麻省理工学院获得 Fannie & John Hertz Foundation Fellowship 和美国国家科学基金会研究生奖学金的资助。AR 感谢美国海军研究办公室通过资助项目 N00014-20-1-2336 以及美国陆军研究办公室通过奖项 W911NF-21-1-0328 提供的支持。

参考文献

Jacob Abernethy、Elad Hazan 和 Alexander Rakhlin。暗中竞争：老虎机线性优化的高效算法。发表于第*21*届学习理论年会（*COLT*）论文集，2008年。Ioannis Anagnostides、Gabriele Farina 和 Tuomas Sandholm。扩展式博弈中的近最优 phi-遗憾学习，2022年。Jean-Yves Audibert 和 Sébastien Bubeck。对抗性与随机老虎机的极小极大策略。发表于$COLT$，第7卷，第1–122页，2009年。Yakov Babichenko。近似纳什均衡的查询复杂度。《*ACM*学报》，63(4)，2016年10月。ISSN 0004-5411。Yu Bai、Chi Jin 和 Tiancheng Yu。通过自我博弈实现近最优强化学习。发表于第*34*届神经信息处理系统国际会议论文集，NIPS'20，美国纽约州雷德胡克，2020年。Curran Associates Inc. ISBN 9781713829546。Anton Bakhtin、Noam Brown、Emily Dinan、Gabriele Farina、Colin Flaherty、Daniel Fried、Andrew Goff、Jonathan Gray、Hengyuan Hu、Athul Paul Jacob、Mojtaba Komeili、Karthik Konath、Minae Kwon、Adam Lerer、Mike Lewis、Alexander H. Miller、Sasha Mitts、Adithya Renduchintala、Stephen Roller、Dirk Rowe、Weiyan Shi、Joe Spisak、Alexander Wei、David Wu、Hugh Zhang 和 Markus Zijlstra。结合语言模型与战略推理，在 外交博弈中达到人类水平。《科学》，378 (6624)：1067–1074，2022年。doi：10.1126/science.ade9097。URL https://www.science.org/doi/abs/10.1126/science.ade9097。Gábor Bartók、Dean P Foster、Dávid Pál、Alexander Rakhlin 和 Csaba Szepesvári。部分监测——分类、遗憾界与算法。《运筹学数学》，39(4)：967–997，2014年。Gerald Beer。闭集与闭凸集上的拓扑。Kluwer Academic Publishers，1993年。Mario Bravo、David Leslie 和 Panayotis Mertikopoulos。凹 n 人博弈中的老虎机学习。神经信息处理系统进展，31，2018年。Noam Brown 和 Tuomas Sandholm。用于单挑无限注扑克的超人类人工智能：Libratus击败顶尖职业选手。《科学》，359(6374)：418–424，2018年。doi：10.1126/science.aao1733。URL https://www.science.org/doi/abs/10.1126/science.aao1733。Sébastien Bubeck。凸优化：算法与复杂度。基础与趋势® 机器学习，8(3-4)：231–357，2015年。Sébastien Bubeck、Nicolo Cesa-Bianchi 和 Sham M Kakade。迈向带老虎机反馈的在线线性优化极小极大策略。发表于学习理论会议，第41–1页。《JMLR研讨会与会议论文集》，2012年。Sébastien Bubeck、Yin Tat Lee 和 Ronen Eldan。基于核的老虎机凸优化方法。发表于第*49*届*ACM SIGACT*计算理论年会论文集，第72–85页，2017年。Clément L Canonne。分布检验综述：你的数据很庞大，但它是否"蓝"？《计算理论》，第1–100页，2020年。Fan Chen、Song Mei 和 Yu Bai。具有决策估计系数的统一强化学习算法：无遗憾、PAC与无奖励学习。*arXiv*预印本 *arXiv:2209.11745*，2022a。Xi Chen、Yu Cheng 和 Bo Tang。良好支撑的纳什均衡与近似纳什均衡：大型博弈的查询复杂度。发表于 Christos H. Papadimitriou 编，第*8*届理论计算机科学创新会议（*ITCS 2017*），信息学第67卷莱布尼茨国际信息学论文集（*LIPIcs*），第57:1–57:9页，德国达格施图尔，2017年。Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik。ISBN 978-3-95977-029-3。doi：10.4230/LIPIcs.ITCS.2017.57。URL http://drops.dagstuhl.de/opus/volltexte/2017/8163。

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。发表于算法学习理论国际会议，第227–261页。PMLR，2022b。Qiwen Cui、Zhihan Xiong、Maryam Fazel 和 Simon S Du。带老虎机反馈的拥塞博弈学习。*arXiv*预印本 *arXiv:2206.01880*，2022年。Varsha Dani、Thomas P Hayes 和 Sham Kakade。老虎机信息对在线优化的代价。2007年。Constantinos Daskalakis 和 Christos H Papadimitriou。通过马尔可夫随机场计算图博弈中的纯策略纳什均衡。发表于第七届*ACM*电子商务会议论文集，第91–99页，2006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。Simon S Du、Sham M Kakade、Jason D Lee、Shachar Lovett、Gaurav Mahajan、Wen Sun 和 Ruosong Wang。双线性类：强化学习中可证明泛化的结构框架。国际机器学习大会，2021。Eyal Even-Dar、Yishay Mansour 和 Uri Nadav。凹博弈中遗憾最小化动态的收敛性。发表于第四十一届*ACM*计算理论年度研讨会论文集，第523–532页，2009年。Abraham D Flaxman、Adam Tauman Kalai 和 H Brendan McMahan。老虎机设定下的在线凸优化：无梯度的梯度下降。发表于第十六届*ACM-SIAM*离散算法年度研讨会论文集，第385–394页，2005年。Dylan J Foster、Zhiyuan Li、Thodoris Lykouris、Karthik Sridharan 和 Eva Tardos。博弈学习：快速收敛的鲁棒性。神经信息处理系统进展，29，2016。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。Dylan J Foster、Noah Golowich、Jian Qian、Alexander Rakhlin 和 Ayush Sekhari。关于采用决策估计系数的无模型强化学习的说明。*arXiv*预印本 *arXiv:2211.14250*，2022a。Dylan J Foster、Alexander Rakhlin、Ayush Sekhari 和 Karthik Sridharan。对抗性决策的复杂度。*arXiv*预印本 *arXiv:2206.13063*，2022b。Dylan J. Foster、Noah Golowich 和 Yanjun Han。采用决策估计系数进行交互式决策的紧致保证。*arXiv*预印本 *arXiv:2301.08215*，2023。Angeliki Giannou、Emmanouil-Vasileios Vlatakis-Gkaragkounis 和 Panayotis Mertikopoulos。关于博弈中正则化学习的收敛速率：从老虎机与不确定性到乐观主义及更进一步。神经信息处理系统进展，34：22655–22666，2021。Geoffrey J. Gordon、Amy Greenwald 和 Casey Marks。凸博弈中的无遗憾学习。发表于第二十五届机器学习国际会议论文集，ICML '08，第360–367页，美国纽约州纽约市，2008年。计算机协会。ISBN 9781605582054。Venkatesan Guruswami、Atri Ruda 和 Madhu Sudan。基础编码理论。2022年。Amélie Heliou、Johanne Cohen 和 Panayotis Mertikopoulos。势博弈中的老虎机反馈学习。神经信息处理系统进展，30，2017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。Nan Jiang、Akshay Krishnamurthy、Alekh Agarwal、John Langford 和 Robert E Schapire。具有低贝尔曼秩的上下文决策过程是PAC可学习的。发表于国际机器学习大会，第1704–1713页，2017年。

Chi Jin、Qinghua Liu 和 Sobhan Miryoosefi。Bellman eluder 维度：强化学习问题中的新颖丰富类别及样本高效算法。神经信息处理系统，2021a。Chi Jin、Qinghua Liu、Yuanhao Wang 和 Tiancheng Yu。V-learning——一种简单、高效的多智能体强化学习去中心化算法。*arXiv* 预印本 *arXiv:2110.14555*，2021b。Chi Jin、Qinghua Liu 和 Tiancheng Yu。利用者的力量：大状态空间中的可证明多智能体强化学习。载于国际机器学习大会，第 10251–10279 页。PMLR，2022。Robert Kleinberg。连续臂老虎机问题的近乎紧确界。神经信息处理系统进展，17：697–704，2004。János Kramár、Tom Eccles、Ian Gemp、Andrea Tacchetti、Kevin R. McKee、Mateusz Malinowski、Thore Graepel 和 Yoram Bachrach。人工智能方法在棋盘游戏《外交》中的谈判与诚实。《自然·通讯》，13(1)：7214，2022年12月。ISSN 2041-1723。doi：10. 1038/s41467-022-34473-5。URL https://www.nature.com/articles/s41467-022-34473-5。编号：1；出版商：自然出版集团。Tor Lattimore。对抗性老虎机凸优化的零阶方法改进遗憾界。《数理统计与学习》，2(3)：311–334，2020年。Tor Lattimore。部分监控的极小极大遗憾：无限结果与 Rustichini 的遗憾。*arXiv* 预印本 *arXiv:2202.10997*，2022年。Tor Lattimore 和 Andras György。镜像下降与信息比。载于学习理论会议，第 2965–2992 页。PMLR，2021。Chris Junchi Li、Dongruo Zhou、Quanquan Gu 和 Michael I Jordan。双人混合马尔可夫博弈学习：核函数近似与相关均衡。*arXiv* 预印本 *arXiv:2208.05363*，2022年。Tianyi Lin、Zhengyuan Zhou、Wenjia Ba 和 Jiawei Zhang。强单调博弈中具有老虎机反馈的最优无遗憾学习。*arXiv* 预印本 *arXiv:2112.02856*，2021。Qinghua Liu、Tiancheng Yu、Yu Bai 和 Chi Jin。对基于模型且采用自我博弈的强化学习的精细分析。载于 Marina Meila 和 Tong Zhang（编），第*38*届国际机器学习大会论文集，第 139 卷机器学习研究论文集，第 7001–7010 页。PMLR，2021年7月18–24日。Qinghua Liu、Csaba Szepesvári 和 Chi Jin。部分可观测马尔可夫博弈的样本高效强化学习。*arXiv* 预印本 *arXiv:2206.01315*，2022年。Chinmay Maheshwari、Chih-Yuan Chiu、Eric Mazumdar、Shankar Sastry 和 Lillian Ratliff。凸凹极小极大问题的零阶方法：在决策依赖风险最小化中的应用。载于国际人工智能与统计大会，第 6702–6734 页。PMLR，2022年。Kleanthis Malialis 和 Daniel Kudenko。运用多智能体强化学习分布式应对网络入侵。《人工智能工程应用》，41：270–284，2015年。ISSN 0952-1976。doi：https://doi.org/10.1016/j.engappai.2015.01.013。URL https://www.sciencedirect.com/ science/article/pii/S095219761500024X。Yishay Mansour、Mehryar Mohri、Jon Schneider 和 Balasubramanian Sivan。在贝叶斯博弈中针对学习者制定策略。载于 Po-Ling Loh 和 Maxim Raginsky（编），第三十五届学习理论大会论文集，第 178 卷机器学习研究论文集，第 5221–5252 页。PMLR，2022年7月2–5日。Weichao Mao 和 Tamer Basar。去中心化一般和马尔可夫博弈中的可证明高效强化学习。《动态博弈与应用》，第 1–22 页，2022年。Noam Nisan、Tim Roughgarden、Eva Tardos 和 Vijay V Vazirani。算法博弈论，第 1 卷。剑桥大学出版社，剑桥，2007。Martin J Osborne 和 Ariel Rubinstein。博弈论教程。麻省理工学院出版社，1994。

Liam Paninski。基于极稀疏采样离散数据的均匀性巧合检验。*IEEE* 信息论汇刊，54(10)：4750–4755，2008年。Julien Perolat、Bart De Vylder、Daniel Hennes、Eugene Tarassov、Florian Strub、Vincent de Boer、Paul Muller、Jerome T. Connor、Neil Burch、Thomas Anthony、Stephen McAleer、Romuald Elie、Sarah H. Cen、Zhe Wang、Audrunas Gruslys、Aleksandra Malysheva、Mina Khan、Sherjil Ozair、Finbarr Timbers、Toby Pohlen、Tom Eccles、Mark Rowland、Marc Lanctot、Jean-Baptiste Lespiau、Bilal Piot、Shayegan Omidshafiei、Edward Lockhart、Laurent Sifre、Nathalie Beauguerlange、Remi Munos、David Silver、Satinder Singh、Demis Hassabis 和 Karl Tuyls。运用无模型多智能体强化学习精通战略棋。《科学》，378(6623)：990–996，2022年。doi：10.1126/science.add4679。URL https: //www.science.org/doi/abs/10.1126/science.add4679。Yury Polyanskiy 和 Yihong Wu。信息论讲义。2014。Alexander Rakhlin 和 Karthik Sridharan。具有可预测序列的优化、学习与博弈。载于神经信息处理系统进展（*NIPS*），第 3066–3074 页，2013。J. B. Rosen。凹 n 人博弈均衡点的存在性与唯一性。《计量经济学》，33 (3)：520–534，1965年。Aviad Rubinstein。厘清计算近似双人纳什均衡的复杂度。载于计算机科学基础年会（*FOCS*），第 258–265 页。IEEE，2016。Daniel Russo 和 Benjamin Van Roy。Eluder 维数与乐观探索的样本复杂度。载于神经信息处理系统进展，第 2256–2264 页，2013。Daniel Russo 和 Benjamin Van Roy。通过后验采样学习优化。《运筹学数学》，39(4)：1221–1243，2014。Daniel Russo 和 Benjamin Van Roy。通过信息导向采样学习优化。《运筹学》，66(1)：230–252，2018。Shai Shalev-Shwartz、Shaked Shammah 和 Amnon Shashua。用于自动驾驶的安全多智能体强化学习。*CoRR*，abs/1610.03295，2016。URL http://arxiv.org/abs/1610.03295。David Silver、Aja Huang、Chris J Maddison、Arthur Guez、Laurent Sifre、George Van Den Driessche、Julian Schrittwieser、Ioannis Antonoglou、Veda Panneershelvam、Marc Lanctot 等。运用深度神经网络和树搜索精通围棋。《自然》，529(7587)：484，2016。Ziang Song、Song Mei 和 Yu Bai。何时能够以样本高效的方式学习玩家众多的一般和马尔可夫博弈？*arXiv* 预印本 *arXiv:2110.04184*，2021。Wen Sun、Nan Jiang、Akshay Krishnamurthy、Alekh Agarwal 和 John Langford。上下文决策过程中的基于模型的强化学习：PAC 界及相较于无模型方法的指数级改进。载于学习理论会议，第 2898–2933 页。PMLR，2019。Salil Vadhan。伪随机性。《理论计算机科学基础与趋势》，7：1–336，2012。Ruosong Wang、Russ R Salakhutdinov 和 Lin Yang。具有一般价值函数近似的强化学习：通过有界 Eluder 维数实现可证明高效的方法。神经信息处理系统进展，33，2020。Chen-Yu Wei 和 Haipeng Luo。更具自适应性的对抗性老虎机算法。载于学习理论会议，第 1263–1291 页。PMLR，2018。Qiaomin Xie、Yudong Chen、Zhaoran Wang 和 Zhuoran Yang。运用函数逼近和相关均衡学习零和同时行动马尔可夫博弈。载于学习理论会议，第 3674–3682 页。PMLR，2020。Wenhao Zhan、Jason D Lee 和 Zhuoran Yang。去中心化乐观超策略镜像下降：马尔可夫博弈中可证明的无遗憾学习。*arXiv* 预印本 *arXiv:2206.01588*，2022年。

Stephan Zheng、Alexander Trott、Sunil Srinivasa、David C. Parkes 和 Richard Socher。《人工智能经济学家：通过双层深度多智能体强化学习进行税收政策设计》。《科学进展》，8(18)：eabk2607，2022年。doi：10.1126/sciadv.abk2607。URL：https://www.science.org/doi/abs/10.1126/sciadv. abk2607。

第一部分

示例

MA-DMSO：实例示例

在本附录的这一节中，我们给出 MA-DMSO 框架的实例，并应用我们的结果推导极小极大风险的上界和下界。

- 在 附录 A.1 中，我们给出 MA-DMSO 框架可涵盖的其他均衡示例，重点讨论相关均衡及其变体。

- 在附录 A.2 我们给出了 MA-DMSO 实例的详细示例，包括具有线性或凹收益的标准形博弈（附录 A.2.1）和马尔可夫博弈（附录 A.2.2）。
- 最后，在 附录 A.3 中，我们给出了方差实例的多智能体决策估计系数和极小极大风险的界，其中包括有限动作标准式博弈（附录 A.3.1）、结构化正规形博弈（附录 A.3.2、附录 A.3.3）以及表格型马尔可夫博弈（附录 A.3.4）。此外，在 附录 A.3.5 中，我们给出了一个实例，表明定理 4.1 中的多智能体到单智能体归约一般而言可能并不紧。

A.1 均衡的其他示例

定义 A.1 展示了如何利用 MA-DMSO 框架来刻画博弈中的（标准形）相关均衡计算问题。该定义与 CCE 实例（定义 1.2）类似，不同之处在于，玩家的偏离集合由从其纯决策集到自身的映射构成；这些映射描述了玩家如何根据其纯决策进行偏离。

定义 A.1（相关均衡实例）。 我们称 *MA-DMSO* 实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k$) 是一个相关均衡（CE）实例，如果满足以下条件：

1. 对于某些有限集 $\Sigma_1, \ldots, \Sigma_K$（称为纯决策），有 $\Pi = \Delta(\Sigma_1 \times \cdots \times \Sigma_K)$。我们记作 $\Sigma = \Sigma_1 \times \cdots \times \Sigma_K$。 *2.* 对于每个 $\pi \in \Pi$ 和 $M \in \mathcal{M}$，有 $M(\pi) = \mathbb{E}_{\sigma\sim\pi}[M(\sigma)]$。 *3.* 对于 $k \in [K]$，有 $\Pi'_k = \Sigma_k^{\Sigma_k}$，即 Π'_k 是函数 $\phi : \Sigma_k \to \Sigma_k$ 的集合。 *4.* 对于每个 $k \in [K]$、 $\pi \in \Pi$ 和 $\phi \in \Pi'_k$， $U_k(\phi, \pi) \in \Delta(\Sigma)$ 是一个分布，其概率质量函数如下所示：

$$\forall \sigma \in \Sigma, \quad U_k(\phi, \pi)(\sigma) = \pi\left(\{(\sigma'_k, \sigma_{-k}) \in \Sigma \ : \ \phi(\sigma'_k) = \sigma_k\}\right).$$

换言之， $U_k(\phi, \pi)$ 是$(\phi(\sigma_k)\ \sigma_{-k})$的分布，其中 $\sigma \sim \pi$。

下一个例子将考察专用于马尔可夫博弈的均衡概念。回想一下，定义 1.2和 A.1 描述了能够刻画标准形博弈中（粗）相关均衡概念的实例，其中纯动作属于 $\Sigma = \Sigma_1 \times \cdots \times \Sigma_K$。在马尔可夫博弈的情形下，通常采用略有不同的（粗）相关均衡概念，我们将在下面的例 A.1 中展示如何刻画这一概念。

例A.1（马尔可夫博弈中的马尔可夫（粗）相关均衡）。在例1.2中，我们将展示如何在 MA-DMSO 框架中刻画计算马尔可夫粗相关均衡（粗相关均衡）和马尔可夫相关均衡（相关均衡）的问题（例如，Bai 等人（2020）；Liu 等人（2021）；Daskalakis 等人（2022）），从而推广例1.2中的马尔可夫 Has 均衡概念。与例1.2中一样，我们假设模型类 $\mathcal{M}$ 由时域为 $H \in \mathbb{N}$、状态空间为 $\mathcal{S}_h$（对应于 $h \in [H]$）、动作空间为 $\mathcal{A}_k$（对应于 $k \in [K]$）且分布为 $d_1 \in \Delta$（$\mathcal{S}_1$）的有限时域马尔可夫博弈组成，并且这些要素在模型类中的所有模型间均相同。纯观测空间 $\mathcal{O}$。由轨迹构成，奖励空间为 $\mathcal{R} = [0, 1]$。对于马尔可夫相关均衡和马尔可夫粗相关均衡，联合决策空间是马尔可夫相关策略的集合 Π，

即策略 $\pi = (\pi_1, \ldots, \pi_H)$，其中每个 $\pi_h: \mathcal{S}_h \to \Delta(\mathcal{A}_1 \times \cdots \times \mathcal{A}_K)$ 都指定了一个从状态到动作联合分布的映射。对于模型 M 和联合决策 $\pi \in \Pi$，观测（轨迹）$o = \{(s_h, (a_{1,h}, \ldots, a_{K,h}), (r_{1,h}, \ldots, r_{K,h}))\}_{h\in[H]}$ 按如下方式抽取：首先，$s_1 \sim d_1$，然后对于 $h \in [H]$:

- $(a_{1,h}, \ldots, a_{K,h}) \sim \pi_h(s_h)$ 和 $r_{k,h} \sim R_k^M(s_h, (a_{1,h}, \ldots, a_{K,h}))$。
- $s_{h+1} \sim P_h^M(\cdot|s_h, (a_{1,h}, \ldots, a_{K,h}))$.

还需要指定偏离集合 Π'_k 和切换函数 U_k:

- 对于马尔可夫粗相关均衡的情形，对于每个 $k \in [K]$，偏离集合 Π'_k 是玩家 k 的确定性马尔可夫策略集合，其形式为 $\pi'_k = (\pi'_{k,1}, \ldots, \pi'_{k,H})$，其中 $\pi'_{k,h}: \mathcal{S}_h \to \mathcal{A}_k$。对于联合策略 $\pi \in \Pi$, $U_k(\pi'_k, \pi) \in \Pi$ 是马尔可夫相关策略，其中玩家 k 在每个状态下按照 $\pi'_{k,h}$ 行动，所有其他玩家则按照 π 行动。特别地，记 $\widetilde{\pi} := U_k(\pi'_k, \pi)$，我们有 $\widetilde{\pi}_h(s_h) = \pi'_{k,h}(s_h) \times \pi_{-k,h}(s_h)$，其中 $\pi_{-k,h}(s_h)$ 表示 $\pi_h(s_h)$ 在除 k 外所有玩家动作上的边缘分布。总而言之，对于 MA-DMSO 实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$，我们有 $\widehat{\pi} \in \Pi$, $h^{M^\star}(\widehat{\pi}) = 0$ 当且仅当 $\widehat{\pi}$ 是 $M^\star$ 的马尔可夫粗相关均衡。
- 对于马尔可夫相关均衡的情况，对于每个 $k \in [K]$，偏离集合 Π'_k 仅由元组集合 $\phi = (\phi_{k,h,s})_{h\in[H], s\in\mathcal{S}_h}$ 构成，其中每个 $\phi_{k,h,s}: \mathcal{A}_k \to \mathcal{A}_k$ 都是一个从 $\mathcal{A}_k$ 到其自身的函数。对于联合策略 $\pi \in \Pi$，$U_k(\phi, \pi)$ 是如下定义的马尔可夫相关策略 $\widetilde{\pi}$：在步 h、状态 $s_h \in \mathcal{S}_h$ 下，$\widetilde{\pi}$ 的联合动作分布由以下分布给出：

$$\widetilde{\pi}_h(s)(a) = \pi_h(s)(\{(a'_k, a_{-k}) \in \mathcal{A} \ : \ \phi(a'_k) = a_k\}),$$

对于联合动作 $a \in \mathcal{A}$。换言之，$\widetilde{\pi}_h(s)$ 是 $(\phi(a'_k)\ a_{-k})$ 的分布，其中 $a \sim \pi_h(s)$。总而言之，对于 MA-DMSO 实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$，当且仅当 $\widehat{\pi}$ 是 $M^\star$ 的一个马尔可夫相关均衡时，$\widehat{\pi} \in \Pi$, $h^{M^\star}(\widehat{\pi}) = 0$。

请注意，上文构造的实例并非我们针对标准形博弈所考虑的粗相关均衡或相关均衡实例（定义 1.2和 A.1）的特例。这是因为，上文讨论的马尔可夫（粗）相关均衡概念限制更严格，要求联合决策 $\widehat{\pi}$ 必须是联合马尔可夫策略，而不能是联合策略上的任意分布。尽管如此，正如 例A.1 所示，MA-DMSO 框架足够通用，能够涵盖所有这些均衡概念。◁

A.2 实例的其他示例

本节将介绍文献中常见的标准均衡学习问题所对应的其他实例。我们先在 附录 A.2.1 中介绍结构化正规形博弈的示例，然后在附录 A.2.2中讨论多智能体强化学习问题。

A.2.1 用于老虎机的实例

本节将介绍若干结构化正规形博弈的实例，可将其视为单智能体设定中结构化老虎机问题的多智能体推广。对于每个示例，模型都具有以下共同结构（与 例 1.1 相似）。

- 每个智能体 $k \in [K]$ 都有一个集合 $\mathcal{A}_k$，称为其纯动作集合，联合策略空间 Π是 $\Delta(\mathcal{A}) = \Delta(\mathcal{A}_1 \times \cdots \times \mathcal{A}_K)$ 的一个子集，其中包含所有单点分布 $\mathbb{I}_a$。
- 我们将 $\mathcal{R} := [-1, 1]$ 作为奖励空间，将 $\mathcal{O}_\circ := \mathcal{A}$ 作为纯观测空间。
- 给定一类平均奖励函数 $\mathcal{F} \subseteq (\mathcal{A} \to \mathcal{R}^K)$。我们将模型类 $\mathcal{M}_\mathcal{F}$ 定义为模型 M 的集合：$\Pi \to \Delta(\mathcal{R}^K \times \mathcal{O}_\circ)$，其中存在某个 $(f_1, \ldots, f_K) \in \mathcal{F}$，使得：(a) 对所有单点分布 $\mathbb{I}_a \in \Pi$，$(r_1, \ldots, r_K, o_\circ) \sim M(\mathbb{I}_a)$ 的分布几乎处处满足 $o_\circ = a$，且 $\mathbb{E}^{M, \mathbb{I}_a}[r_k] = f^M(\mathbb{I}_a) = f_k(a)$；(b) 对所有 $\pi \in \Pi$，$M(\pi) = \mathbb{E}_{a\sim\pi}[M(\mathbb{I}_a)]$。

用文字来说，$\mathcal{M}_\mathcal{F}$ 由模型 M 构成，其中（i）价值函数 $f_k^M(\cdot)$ 由 $\mathcal{F}$ 中的某个元素给出；（ii）观测会揭示所执行的动作（通过纯观测）。

首先，在示例 A.2中，我们考虑一个具有线性结构奖励的标准形博弈，将单智能体线性老虎机问题推广开来（Dani 等人，2007；Abernethy 等人，2008年；Bubeck 等人，2012年）。这个示例推广了例 1.1，后者可以被视为每个玩家的行动集都是标准单位向量集合的特殊情形。

示例 A.2（具有线性奖励的标准形博弈）。固定 $K \in \mathbb{N}$；对于每个玩家 $k \in [K]$，$\mathcal{A}_k \subset \mathbb{R}^{d_k}$ 对于某个 $d_k \in \mathbb{N}$。记作 $d = d_1 d_2 \cdots d_K$。假设 Θ_1，……，$\Theta_K \subset \mathbb{R}^d$ 是凸集，使得对于所有 $a_1 \in \mathcal{A}_1$，……，$a_K \in \mathcal{A}_K$、$k \in [K]$和 $\theta_k \in \Theta_k$均成立。定义 $\mathcal{F} \subset (\mathcal{A} \to \mathbb{R}^K)$ 为 $\mathcal{F} = \{(a_1, \ldots\ldots, a_K) \mapsto (\langle a_1 \otimes \cdots \otimes a_K, \theta_k \rangle)_{k \in [K]} : \theta_1 \in \Theta_1, \ldots\ldots, \theta_K \in \Theta_K\}$。现在，我们可以考虑以下实例：对于收益由 $\mathcal{F}$ 中的函数给出的博弈类，求解纳什均衡、相关均衡和粗相关均衡：

- 我们首先讨论纳什均衡：假设我们设定 $\Pi_k = \Delta(\mathcal{A}_k)$ 对于每个 $k \in [K]$ 和 $\Pi = \Pi_1 \times \cdots \times \Pi_k$，并按照 Π'_k，U_k 中的定义 定义 1.1。我们定义 $\mathcal{M} = \mathcal{M}_{\mathcal{F}}$。那么，实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$刻画了在未知线性老虎机博弈中寻找纳什均衡的问题。
- 接下来我们讨论（粗）相关均衡：我们设定 $\Pi = \Delta(\mathcal{A}_1 \times \cdots \times \mathcal{A}_K)$，并按照 Π'_k，U_k 中的定义 定义 A.1（分别对应 定义 1.2）。我们定义 $\mathcal{M} = \mathcal{M}_{\mathcal{F}}$，使得实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 刻画了在未知线性老虎机博弈中寻找（粗）相关均衡的问题。

◁

接下来，示例 A.3 讨论了凹博弈（带有老虎机反馈）的情形，该情形在博弈论文献（Rosen，1965年；Even-Dar 等人，2009）以及机器学习领域（Bravo 等人，2018年；Maheshwari 等人，2022年；Lin 等人，2021）中都受到了广泛关注。它也可以被视为单人凹老虎机问题的推广（Kleinberg，2004；Flaxman 等人，2005；Bubeck 等人，2017年；Lattimore，2020）。[12]

示例 A.3（凹博弈）。给定 $K \in \mathbb{N}$，对于每个 $k \in [K]$，令 $d_k \in \mathbb{N}$ 和 $\mathcal{A}_k \subset \mathbb{R}^{d_k}$ 为具有非空内部的凸紧子集。设 $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_K \subset \mathbb{R}^d$，其中 $d = d_1 + \cdots + d_K$。定义$\mathcal{F} \subset (\mathcal{A} \to \mathbb{R}^K)$为

$$\mathcal{F} = \left\{ f : \mathcal{A} \to [0,1]^K \mid \forall k \in [K],\ \forall a_{-k} \in \mathcal{A}_{-k},\ \ \mathcal{A}_k \ni a_k \mapsto f_k(a_k, a_{-k}) \ \text{ is concave and 1-Lipschitz} \right\}.$$

上述，1-Lipschitzn ess 是相对于 ℓ_2 范数而言的。我们考虑以下纳什均衡和粗相关均衡实例 ：

- 我们首先考虑纳什均衡：按照定义 1.1 定义 Π'_k, U_k，并设 $\Pi = \mathcal{A}$, $\mathcal{M} = \mathcal{M}_{\mathcal{F}}$。那么，实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$刻画了在凹博弈中寻找纳什均衡的问题，这是一个经典问题（Rosen，1965年）。在双人零和情形（即当 $K = 2$ 和$f_1(a) + f_2(a) = 0$ 对所有 $a \in \Pi$均成立时），我们所涵盖的老虎机反馈问题已受到广泛关注（Bravo 等，2018；Maheshwari等，2022年；Lin等，2021）。
- 接下来我们考虑粗相关均衡。定义 $\Pi := \Delta(\mathcal{A}_1 \times \cdots \times \mathcal{A}_K)$，即紧致集合 $\mathcal{A}_1 \times \cdots \times \mathcal{A}_K \subset \mathbb{R}^d$上的博雷尔测度空间，并令 $\mathcal{M} = \mathcal{M}_{\mathcal{F}}$。进一步定义 Π'_k，U_k如 定义 1.2所述。于是，实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 刻画了在凹博弈中寻找粗相关均衡的问题；与凹博弈中的纳什均衡相比，这一问题受到的关注较少，但近年来已有相关研究

由于在此设定中动作集合 $\mathcal{A}_k$ 是无限的，因此按照 定义 A.1中的方式定义 CE 实例并不特别自然。◁

A.2.2 多智能体强化学习实例

现在给出马尔可夫博弈类的具体示例 $\mathcal{M}$。第一个示例考虑了计算马尔可夫纳什均衡和马尔可夫（粗）相关均衡的实例的特殊情形，这些实例在示例 1.2 和 A.1 中有所描述，其中所考虑的马尔可夫博弈是表格型（即具有有限状态和行动）。

[12]通常称为凸 *bandit*或零阶凸优化，因为它通常采用损失最小化的表述形式，而我们考虑的是奖励最大化。

示例 A.4（表格型马尔可夫博弈中的均衡）。固定参数 K，$H \in \mathbb{N}$ 分别表示玩家数量和时域；对于每个玩家 $k \in [K]$，给定有限动作空间 $\mathcal{A}_k$（大小为 $A_k \in \mathbb{N}$），以及每个步 $h \in [H]$对应的有限状态空间 $\mathcal{S}_h$（大小为 $S \in \mathbb{N}$）。三种均衡类型（纳什、CE、CCE）的实例共享同一个观测空间 $\mathcal{O}$：具体而言，它们的纯观测空间为 $\mathcal{O}$。，即状态和动作空间上所有可能的 H-步轨迹 $\mathcal{S}_1$、……、$\mathcal{S}_H$和 $\mathcal{A}$所构成的空间，而奖励空间为 $\mathcal{R} = [0,1]$。

我们将表格设定称为模型类 $\mathcal{M}$，它由所有可能的 K 人马尔可夫博弈构成，这些博弈的时域为 H、状态空间为 $\mathcal{S}_h$、动作空间为 $\mathcal{A}_k$，并且在任意正概率轨迹上，每位玩家的奖励总和均有界于 $[0,1]$。[13] 接下来，对于例1.2中描述的偏差函数和切换函数 Π'_k、U_k，实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$刻画了在未知表格型马尔可夫博弈中计算马尔可夫纳什均衡的问题；对于例A.1中分别对应马尔可夫粗相关均衡或马尔可夫相关均衡概念的 Π'_k、U_k，实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$刻画了在未知表格型马尔可夫博弈中计算马尔可夫粗相关均衡或马尔可夫相关均衡的问题。◁

示例 A.5（线性混合马尔可夫博弈中的均衡）。固定参数 K，$H \in \mathbb{N}$ 分别表示玩家数量和时域；对于每个 $k \in [K]$，给定有限动作空间 $\mathcal{A}_k$；对于每个$h \in [H]$，给定有限状态空间 $\mathcal{S}_h$。[14]对于一个维度参数 $d \in \mathbb{N}$，给定映射 $\phi_h \colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$, $\psi_{k,\,h} \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$，使得对于所有 $h \in [H]$ 和 $k \in [K]$

$$\sum_{s_{h+1} \in \mathcal{S}_{h+1}} \phi_h(s_{h+1} | s_h, a) = \mathbf{1} \in \mathbb{R}^d, \quad \text{and} \quad \|\psi_{k,h}(s_h, a)\|_2 \le 1$$

对于所有 $s_h \in \mathcal{S}_h$，$a \in \mathcal{A}$，$s_{h+1} \in \mathcal{S}_{h+1}$。[15] 我们构造的实例具有纯观测空间 $\mathcal{O}$。，它由动作空间和状态空间 $\mathcal{A}$ 与 $\mathcal{S}_h$ 上所有可能的 H 步轨迹构成，并具有奖励空间$\mathcal{R} = [0,1]$。

对于某个 $B \in \mathbb{N}$，线性混合马尔可夫博弈的集合是模型类 $\mathcal{M}$，由所有满足以下条件的 K 人马尔可夫博弈 M 构成：时域为 H，状态空间为 $\mathcal{S}_h$，动作空间为 $\mathcal{A}_k$，且存在向量 $\theta_h^M \in \mathbb{R}^d$ 满足 $\|\theta_h^M\|_2 \le B$ 和

$$P_h^M(s_{h+1} | s_h, a) = \langle \theta_h^M, \phi_h(s_{h+1} | s_h, a) \rangle, \qquad R_{k,h}^M(s_h, a) = \langle \theta_h^M, \psi_{k,h}(s_h, a) \rangle$$

对于所有 $h \in [H], k \in [K]$, $s_h \in \mathcal{S}_h, a \in \mathcal{A}, s_{h+1} \in \mathcal{S}_{h+1}$，且对于任意正概率轨迹，$\sum_{h=1}^H r_{k,h} \in [0,1]$。（为简单起见，我们假设奖励是确定性的，且等于上文定义的量$R_{k,h}^M(s_h, a)$。）

对于偏差函数和切换函数 Π'_k，U_k如 例1.2所述，实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$刻画了在未知线性混合马尔可夫博弈中计算马尔可夫纳什均衡的问题；对于 Π'_k，U_k 如 例A.1 所述，分别对应于马尔可夫粗相关均衡或马尔可夫相关均衡的概念，实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 刻画了在未知线性混合马尔可夫博弈中分别计算马尔可夫粗相关均衡或马尔可夫相关均衡的问题。

◁

A.3 计算 DEC 和多智能体实例的极小极大风险界 s

在本节中，我们将运用第 3节至第 5节的研究结果，(a) 给出各种 MA-DMSO 实例的 DEC 界；(b) 利用这些 DEC 界，推导多智能体交互式决策中学习均衡的极小极大风险界。

[13]该假设允许我们取 $\mathcal{R} = [0,1]$。[14]出于技术原因，我们要求状态空间是有限的，但我们的界不会取决于状态空间的大小。[15] $\phi_H(s_{h+1} | s_h, a)$ 的取值并不重要，因此我们可以令 $\mathcal{S}_{H+1}$ 为例如由单个状态组成的集合。

A.3.1 有限动作空间的标准式博弈

我们从一个或许最简单的例子开始：具有老虎机反馈的有限动作标准式博弈。我们考虑例 1.1中所述的纳什、CE 和 CCE 实例。固定 $K \in \mathbb{N}$ 及行动集 $\mathcal{A}_1$、……、$\mathcal{A}_K$，对应 K 个玩家，其联合行动集为 $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$。记 $A_k := |\mathcal{A}_k|$表示 $k \in [K]$。令 $\mathscr{M}_{\mathrm{NE}}$ $\mathscr{M}^{\mathrm{CE}}$、$\mathscr{M}_{\mathrm{CCE}}$ 分别表示在例 1.1中构造的 NE、CE 和 CCE 实例。本节将为这些实例的 DEC 给出界；我们先给出偏移 DEC 的上界，由此可通过命题 4.1立即得到受约束 DEC 的上界。

Pr命题 A.1. 对于任意 $\gamma > 0$，上述定义的实例 $\mathscr{M}_{\text{纳什均衡}}$ $\mathscr{M}^{\text{相关均衡}}$ $\mathscr{M}^{\mathrm{CCE}}$ 满足

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{CCE}}) \le \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{CE}}) \le \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{NE}}) \le \frac{K \cdot \sum_{k=1}^{K} A_k}{\gamma}.$$

命题 A.1 的证明。注意，实例 $\mathscr{M}^{\mathrm{NE}}$、$\mathscr{M}^{\mathrm{CE}}$、$\mathscr{M}^{\mathrm{CCE}}$ 具有相同的观测空间 $\mathcal{O}$，也就是说，它们具有纯观测空间 $\mathcal{O}_{\circ} = \mathcal{A}$ 和奖励空间 $\mathcal{R} = [0,1]$。[16] 因此，我们记 $\mathscr{M}$纳什均衡 $= (\mathcal{M}, \Pi^{\mathrm{NE}}, \mathcal{O}, \{(\Pi'_k)^{\mathrm{NE}}\}_k, \{U_k^{\mathrm{NE}}\}_k)$、$\mathscr{M}$相关均衡 $= (\mathcal{M}, \Pi^{\mathrm{CE}}, \mathcal{O}, \{(\Pi'_k)^{\mathrm{CE}}\}_k, \{U_k^{\mathrm{CE}}\}_k)$，以及 $\mathscr{M}$粗相关均衡 $= (\mathcal{M}, \Pi^{\mathrm{CCE}}, \mathcal{O}, \{(\Pi'_k)^{\mathrm{CCE}}\}_k, \{U_k^{\mathrm{CCE}}\}_k)$。为区分这三种不同的情形，我们在函数 $f^M(\cdot)$ 和 $h^M(\cdot)$ 上添加上标 NE/CE/CCE。例如，对于实例 $\mathscr{M}^{\mathrm{NE}}$，当 $M \in \mathcal{M}, \pi \in \Pi^{\mathrm{NE}}$ 时，

$$f_k^{M,\mathrm{NE}}(\pi) := \mathbb{E}^{M,\pi}[r_k], \quad \text{and} \quad h^{M,\mathrm{NE}}(\pi) = \sum_{k=1}^{K} \max_{\pi'_k \in (\Pi'_k)^{\mathrm{NE}}} f_k^{M,\mathrm{NE}}(U_k^{\mathrm{NE}}(\pi'_k, \pi)) - f_k^{M,\mathrm{NE}}(\pi).$$

函数 $h^{M,\mathrm{CE}} : \Pi^{\mathrm{CE}} \to \mathbb{R}$ 和 $h^{M,\mathrm{CCE}} : \Pi^{\mathrm{CCE}} \to \mathbb{R}$ 以类似方式定义。

有 $\Pi^{\mathrm{CE}} = \Pi^{\mathrm{CCE}}$；此外，对于任意 $M \in \mathcal{M}$ 和 $\pi \in \Pi^{\mathrm{CE}} = \Pi^{\mathrm{CCE}}$，有 $h^{M,\mathrm{CCE}}(\pi) \le h^{M,\mathrm{CE}}(\pi)$。由此立即可得 $\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{CCE}}) \le \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{CE}})$。接下来注意到 $\Pi^{\mathrm{NE}} \subset \Pi^{\mathrm{CE}}$，并且对于任意 $\pi \in \Pi^{\mathrm{NE}}$ 和 $M \in \mathcal{M}$，有 $h^{M,\mathrm{NE}}(\pi) = h^{M,\mathrm{CE}}(\pi)$。因此，对于 $M \in \mathrm{co}(\mathcal{M})$，

$$\begin{aligned}
\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{NE}}, \overline{M}) &= \inf_{p \in \Delta(\Pi^{\mathrm{NE}})} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[h^{M,\mathrm{NE}}(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \\
&\ge \inf_{p \in \Delta(\Pi^{\mathrm{CE}})} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[h^{M,\mathrm{CE}}(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] = \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{CE}}, \overline{M}).
\end{aligned}$$

这就证明了

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{CCE}}) \le \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{CE}}) \le \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{NE}}).$$

还需给 $\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{NE}})$取一个上界。对于 $k \in [K]$，记 $\Pi_k := \Delta(\mathcal{A}_k)$ 和 $\Pi_{-k} := \prod_{k' \neq k} \Pi_{k'}$。对于每个 $k \in [K]$，定义模型类 $\mathcal{M}_k \subset (\Pi_k \to \Delta(\mathcal{R} \times \mathcal{O}_{\circ}))$，如 式 (17)所示；特别地：

$$\widetilde{\mathcal{M}}_k = \{\pi_k \mapsto M|_k(\pi_k, \pi_{-k}) \ : \ \pi_{-k} \in \Pi_{-k}, M \in \mathcal{M}\}.$$

接下来定义模型类 $\mathcal{M}'_k \subset (\mathcal{A}_k \to \Delta(\mathcal{R} \times \{\perp\}))$：

$$\mathcal{M}'_k = \{M \ : \ M(a_k) \in \Delta(\mathcal{R} \times \{\perp\}) \ \forall a_k \in \mathcal{A}_k\},$$

也就是说，$M(a_k)$ 可以是对于每个 a_k 在 $\mathcal{R} \times \{\perp\}$ 上的任意分布。Foster 等人（2021）的命题 5.2 表明 $\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}'_k) \le \frac{A_k}{\gamma}$。接下来，固定 $\overline{M} \in \mathrm{co}(\widetilde{\mathcal{M}}_k)$，并令 $M' \in \mathrm{co}(\mathcal{M}'_k)$ 为唯一模型，使得

[16] 严格来说，实例 $\mathscr{M}^{\mathrm{NE}}$ 的模型类只作用于 $\Pi^{\mathrm{NE}} = \Delta(\mathcal{A}_1) \times \cdots \times \Delta(\mathcal{A}_K)$ 中的乘积分布，而非 $\Pi^{\mathrm{CCE}} = \Pi^{\mathrm{CE}} = \Delta(\mathcal{A}) \supset \Pi^{\mathrm{NE}}$；为避免符号过于繁杂，我们将实例 $\mathscr{M}^{\mathrm{NE}}$ 的 $\mathcal{M}$ 的定义域形式化地解释为 Π^{NE}。

奖励 $r\sim$ 对于所有 $a_k\in\mathcal{A}_k$，$M'(a_k)$ 与奖励 $r\sim M(a_k)$ 同分布。那么我们有 e

$$
\begin{aligned}
&\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\widetilde{\mathcal{M}}_k,\overline{M})\\
&=\inf_{p\in\Delta(\Pi_k)}\sup_{M\in\mathcal{M},\pi_{-k}\in\Pi_{-k}}\mathbb{E}_{\pi_k\sim p}\left[\max_{\pi'_k\in\Pi_k} f_k^{M,\mathrm{NE}}(\pi'_k,\pi_{-k})-f_k^{M,\mathrm{NE}}(\pi_k,\pi_{-k})-\gamma\cdot D^2_{\mathsf{H}}\big(M(\pi_k,\pi_{-k}),\overline{M}(\pi_k)\big)\right]\\
&\le\inf_{p\in\Delta(\mathcal{A}_k)}\sup_{M\in\mathcal{M},\pi_{-k}\in\Pi_{-k}}\mathbb{E}_{a_k\sim p}\left[\max_{a'_k\in\mathcal{A}_k} f_k^{M,\mathrm{NE}}(a'_k,\pi_{-k})-f_k^{M,\mathrm{NE}}(a_k,\pi_{-k})-\gamma\cdot D^2_{\mathsf{H}}\big(M(a_k,\pi_{-k}),\overline{M}(a_k)\big)\right]\\
&\le\inf_{p\in\Delta(\mathcal{A}_k)}\sup_{M'\in\mathcal{M}'_k}\mathbb{E}_{a_k\sim p}\left[\max_{a'_k\in\mathcal{A}_k} f_k^{M'}(a'_k)-f_k^{M'}(a_k)-\gamma\cdot D^2_{\mathsf{H}}\big(M'(a_k),\overline{M}'(a_k)\big)\right]=\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}'_k,\overline{M}'),
\end{aligned}
$$

其中，第一个不等式成立，是因为 $\mathcal{A}_k\subset\Pi_k$（将每个动作 $a_k\in\mathcal{A}_k$ 视为其指示分布 $\mathbb{I}_{a_k}\in\Pi_k$）；第二个不等式成立，是因为对于任意 $M\in\mathcal{M},\pi_{-k}\in\Pi_{-k}$，都存在一个模型 $M'\in\mathcal{M}'_k$，使得对于所有 $a_k\in\mathcal{A}_k$，奖励 $r\sim M(a_k,\pi_{-k})$ 的分布与 $r\sim M'(a_k)$ 的分布相同。请注意，在上面的公式中，按照我们的约定，我们将动作 $a_k\in\mathcal{A}_k$ 与其单点分布 $\mathbb{I}_{a_k}\in\Pi_k$相关联。由此可得，对于所有 $\gamma>0$，$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}_k)\le\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}'_k)$。最后，将定理 4.1 应用于实例 $\mathscr{M}^{\mathrm{NE}}$，可得

$$
\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{NE}})\le\sum_{k=1}^{K}\textsf{r-dec}^{\mathrm{o}}_{\gamma/K}(\widetilde{\mathcal{M}}_k)\le\sum_{k=1}^{K}\textsf{r-dec}^{\mathrm{o}}_{\gamma/K}(\mathcal{M}'_k)\le\frac{K\cdot\sum_{k=1}^{K}A_k}{\gamma}.
$$

请注意，我们对定理 4.1 的应用是有效的，因为根据假设 4.1 的定义，$\mathscr{M}^{\mathrm{NE}}$在例 1.1 中满足该假设（特别是，我们假设 $M(\mathbb{I}_a)\in\Delta(\mathcal{R}^K)\times\{\mathbb{I}_a\}$，即 M 揭示 a，满足假设 4.1的第二点）。

□

利用命题 A.1，我们现在对实例 $\mathscr{M}_{\text{纳什均衡}}\mathscr{M}^{\mathrm{CE}}$、$\mathscr{M}_{\text{粗相关均衡}}$的极小极大速率进行界定。为稍微简化问题，我们考虑这些实例的简化特例，其中模型类被限制为输出服从伯努利分布的奖励的模型（即奖励取$\{0,1\}$值）。[17] 此外，为简单起见，我们假设对所有 k，$A_k\ge 2$。我们将对应的具有伯努利奖励的 MA-DMSO 实例记为 $\mathscr{M}_0^{\mathrm{NE}}$、$\mathscr{M}_0^{\mathrm{CE}}$、$\mathscr{M}_0^{\mathrm{CCE}}$。首先，我们对$\mathscr{M}_0^{\mathrm{NE}}$的极小极大速率进行界定：

命题 A.2. 对于实例 $\mathscr{M}_{\text{纳什均衡}0}$，存在一种算法，保证以至少 $1-\delta$ 的概率，**风险**$(T)\le\sqrt{(\max_k A_k)}\cdot A\cdot T^{-1}\cdot\mathrm{polylog}(T,A,\delta^{-1})$，其中 $A=A_1A_2\cdots A_K$。

显然，命题 A.2 中 $\mathscr{M}$纳什均衡0 的风险上界同样适用于 $\mathscr{M}_0^{\mathrm{CE}},\mathscr{M}$粗相关均衡0，因为对于任意决策 $\widehat{\pi}\in\Pi^{\mathrm{NE}}\subset\Pi^{\mathrm{CE}}=\Pi^{\mathrm{CCE}}$，都有 $h^{M,\mathrm{CCE}}(\widehat{\pi})\le h^{M,\mathrm{CE}}(\widehat{\pi})\le h^{M,\mathrm{NE}}(\widehat{\pi})$（回顾命题 A.1 证明中 $h^{M,\mathrm{NE}},h^{M,\mathrm{CE}},h^{M,\mathrm{CCE}}$ 的定义）。

命题 A.2 的证明。命题 A.1 与命题 4.1 结合可得 $\mathrm{dec}_{\varepsilon}(\mathscr{M}_0^{\mathrm{NE}})\le\varepsilon\cdot 2\sqrt{K\cdot\sum_{k=1}^{K}A_k}$。由于 $|\mathcal{O}|\le 2^K A$（假设奖励服从伯努利分布），类 $\mathcal{M}$ 满足假设 D.2，且 $B=2^K A$，因此命题 D.1 给出 Est $_{\mathsf{H}}=O\big(\mathcal{M}\ T+\delta^{-1}\cdot K^2\cdot{}^{2}{}_{k}A_k\big)$ ε^2 (T,δ) (est(,) log) log (max)。最后，将奖励均值离散化为 的倍数，可知 $\mathcal{N}(\mathcal{M},\varepsilon)\le(1/\varepsilon^2)^{AK}$，这意味着 est$(\mathcal{M},T)\le O(AK\cdot\log(T))$。因此，定理 D.1 与定理 2.1 结合可得，存在一种算法，其.....

$$
\mathbf{Risk}(T)\le\sqrt{K(A_1+\cdots+A_K)}\cdot\sqrt{\frac{\mathbf{Est}_{\mathsf{H}}(T,\delta)}{T}}\cdot\mathrm{polylog}(T,1/\delta)\le\sqrt{\frac{\max_k A_k\cdot A}{T}}\cdot\mathrm{polylog}(T,1/\delta,A).
$$

□

请注意，命题 A.2 的上界受到多智能体诅咒的影响：联合行动组合的数量 A 相对于智能体数量呈指数级增长 K。众所周知，这种指数级

[17]对模型类施加这一限制，实质上并不失一般性：给定 $[0,1]$, 中任意一个奖励分布一般的模型类，我们都可以模拟出一个具有相同价值函数 f_k^M(·) 和伯努利奖励分布的模型类的样本。具体而言，收到奖励 $(r_1,\ldots\ldots,r_K)\sim M(\pi)$ 后，将每个 r_k 替换为一个来自 $r'_k\sim$伯努利分布(r_k) 的样本。

对 K 的依赖对于学习是必要的（例如，Rubinstein (2016)；参见命题 5.1），但对于学习（粗）相关均衡则并非必要。接下来，我们将说明，第 5 节中的结果使我们能够为（粗）相关均衡恢复这一改进后的（多项式）上界：

命题 A.3. 任取 $T \in \mathbb{N}, \delta \in (0,1)$。对于实例 $\mathscr{M}_0^{\mathrm{CE}}$，存在一个算法可生成$\widehat{\pi} \in \Pi^{\mathrm{CE}}$，使得其概率至少为 $1-\delta$，

$$\mathbf{Risk}(T) \le \sqrt{\frac{K^4 \max_k A_k^2}{T}} \cdot \mathrm{polylog}\left(K, \max_k A_k, \delta^{-1}\right).$$

此外，对于实例 $\mathscr{M}_0^{\mathrm{CCE}}$，存在一个算法可生成 $\widehat{\pi} \in \Pi^{\mathrm{CCE}}$，使得其概率至少为 $1-\delta$，

$$\mathbf{Risk}(T) \le \sqrt{\frac{K^3 \sum_{k=1}^K A_k}{T}} \cdot \mathrm{polylog}\left(K, \max_k A_k, \delta^{-1}\right).$$

命题 A.3 的证明。 该命题的结论可由定理 5.1直接推出。对于实例 $\mathscr{M}_0^{\mathrm{CCE}}$，我们有 $\textsf{r-dec}_\gamma^{\mathrm{o}}(\mathrm{co}(\mathscr{M}_0^{\mathrm{CCE}})) \le \textsf{r-dec}_\gamma^{\mathrm{o}}(\mathrm{co}(\mathscr{M}^{\mathrm{CCE}})) = \textsf{r-dec}_\gamma^{\mathrm{o}}(\mathscr{M}^{\mathrm{CCE}}) \le \frac{K \sum_{k=1}^K A_k}{\gamma}$，其中用到了模型类 $\mathcal{M}$ 是凸的这一点以及 命题 A.1。因此，定理 5.1 表明存在一个算法可达到

$$\begin{aligned}\mathbf{Risk}(T) \le & O\left(K \cdot \inf_{\gamma>0}\left\{\frac{K\sum_{k=1}^K A_k}{\gamma} + \frac{\gamma}{T} \cdot \log\left(\frac{K \cdot \max_k A_k}{\delta}\right)\right\}\right) \\ \le & \sqrt{\frac{K^3 \sum_{k=1}^K A_k}{T}} \cdot \mathrm{polylog}\left(K, \max_k A_k, \delta^{-1}\right).\end{aligned}$$

接下来，对于实例 $\mathscr{M}_0^{\mathrm{CE}}$，$\mathrm{co}(\mathscr{M}_0^{\mathrm{CE}})$ 的 DEC 上界同样成立，但偏离集合更大：我们有 $\max_k |\Pi_k'| = \max_k |A_k^{A_k}|$，因此 定理 5.1 给出

$$\begin{aligned}\mathbf{Risk}(T) \le & O\left(K \cdot \inf_{\gamma>0}\left\{\frac{K\sum_{k=1}^K A_k}{\gamma} + \frac{\gamma}{T} \cdot \log\left(\frac{K \cdot \max_k A_k^{A_k}}{\delta}\right)\right\}\right) \\ \le & \sqrt{\frac{K^4 \max_k A_k^2}{T}} \cdot \mathrm{polylog}\left(K, \max_k A_k, \delta^{-1}\right).\end{aligned}$$

□

下界。接下来我们讨论实例 $\mathscr{M}_0^{\mathrm{CCE}}, \mathscr{M}\mathrm{CE0}, \mathscr{M}\mathrm{NE0}$ 的下界。显然，通过将模型类 $\mathcal{M}$ 限制为奖励分布仅取决于任意单个玩家所采取动作的模型，就能在每个实例中嵌入一个单玩家的 $\max_k A_k$ 臂老虎机实例。由此，根据 Foster 等人（2021）的命题 5.3 的证明可知，$\mathrm{dec}_\varepsilon(\mathscr{M}\mathrm{NE0}) \ge \mathrm{dec}_\varepsilon(\mathscr{M}_0^{\mathrm{CE}}) \ge \mathrm{dec}_\varepsilon(\mathscr{M}_0^{\mathrm{CCE}}) \ge \Omega(\varepsilon\sqrt{\max_k A_k})$ 对于 $\varepsilon > 0$ 成立；事实上，这些下界是由 $\mathcal{M}$ 的子类得到的，其中 $C(T) = \log(T \wedge V(\mathcal{M})) = O(1)$。因此，定理 3.2（当 T 足够小时，取 $\varepsilon(T) = \sqrt[c]{\max_{KT}{}^{A_k}}$）与定理 2.1 一起表明，对于上述任一实例 $\mathscr{M}_0^{\mathrm{CCE}}, \mathscr{M}_0^{\mathrm{CE}}, \mathscr{M}_0^{\mathrm{NE}}$ 以及任意算法，在这三种实例中的任意一种下，都存在一个模型，使得 $\mathbb{E}[\widehat{\mathbf{Risk}}(T)] \ge \Omega(\max_k A_k/(KT))$。

对于实例 $\mathscr{M}\mathrm{CCE0}$，在可学习区域 $T > \max_k A_k$ 中，该下界与命题 A.3 的上界相差一个 $\sqrt{T/\max_k A_k} \cdot \mathrm{poly}(K, \max_k \log A_k, \log T)$ 的因子；对于 $\mathscr{M}\mathrm{CE0}$，差距增至 $\sqrt{T/\max_k A_k} \cdot \max_k \sqrt{A_k} \cdot \mathrm{poly}(K, \max_k \log A_k, \log T)$；而对于 $\mathscr{M}\mathrm{NE0}$，差距进一步增至 $\sqrt{T/\max_k A_k} \cdot \sqrt{A} \cdot \mathrm{polylog}(T, A)$。在所有这些情况下，差距中的 $\sqrt{T}$ 因子源于附录 3.2.2 中讨论的不可能性结果，其余项则源于上界中出现、但下界中没有的模型估计误差。特别是，对于这些实例，我们计算出的 MA-DEC 上界和下界之间没有差距（相差一个 $O(K)$ 因子）。

A.3.2 具有线性收益的标准式博弈

本节给出 示例 A.2 中定义的线性结构标准形博弈实例的 DEC 和极小极大遗憾的界。具体而言，固定行动集 $\mathcal{A}_k \subset \mathbb{R}^{d_k}$ 对于每个 $k \in [K]$，以及凸集$\Theta_1, \ldots, \Theta_K \subset \mathbb{R}^d$(其中 $d = d_1 \cdots d_K$)，使得 $|\langle a_1 \otimes \cdots \otimes a_K, \theta_k \rangle| \le 1$对于所有 $(a_1, \ldots, a_K) \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$，$k \in [K]$，$\theta_k \in \Theta_k$。令 $\mathscr{M}^{\mathrm{NE}}$，$\mathscr{M}^{\mathrm{CE}}$，$\mathscr{M}^{\mathrm{CCE}}$ 表示给定集合$\mathcal{A}_k$，Θ_k 后按照示例 A.2构造的纳什均衡、相关均衡和粗相关均衡实例。下面的命题给出了这些实例的（遗憾）偏移 DEC 的界：

命题 A.4. 对于任意 $\gamma > 0$，上述实例 $\mathscr{M}_{\text{纳什均衡}}$ $\mathscr{M}_{\text{相关均衡}}$ $\mathscr{M}^{\mathrm{CCE}}$ 满足

$$\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{CCE}}) \le \text{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{CE}}) \le \text{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{NE}}) \le \frac{K \cdot \sum_{k=1}^{K} d_k}{\gamma}.$$

命题 A.4 的证明。 证明与 命题 A.1 的证明基本相同，唯一不同之处在于，每个 $\widetilde{\ }$诱导模型类 $\mathcal{M}_k$ 都可以视为一个 d 维的单智能体线性老虎机问题，因此我们可以使用 Foster 等人 (2021) 的命题 6.1 来界定（单玩家）线性老虎机的 DEC。 □

利用命题 A.4，我们现在界定实例 $\mathscr{M}^{\mathrm{NE}}, \mathscr{M}^{\mathrm{CE}}, \mathscr{M}^{\mathrm{CCE}}$ 的极小极大速率。与上一小节一样，为简化讨论，我们将实例限制为模型类仅包含输出（随机）奖励且其取值属于 $\{-1, 1\}$ 的模型（回想一下，对于示例 A.2 中定义的线性老虎机实例，$f_k^M(a) \in [-1, 1]$ 对所有 $M \in \mathcal{M}, a \in \mathcal{A}$ 都成立）。此外，我们假设对于每个 k，$d_k \ge 2$ 以及所有 $\theta_k \in \Theta_k$ 都满足 $\|\theta_k\|_2 \le D$，且所有 $a_k \in \mathcal{A}_k$ 都满足 $\|a_k\|_2 \le D$，其中某个 $D > 0$。由此可得，对于所有 $a_1 \in \mathcal{A}_1, \ldots, a_k \in \mathcal{A}_k$，$\|a_1 \otimes \cdots \otimes a_K\|_2 \le D^K$；我们的界仅以对数方式依赖于 D。我们将相应的、奖励取值属于 $\{-1, 1\}$ 的 MA-DMSO 实例记为 $\mathscr{M}_0^{\mathrm{NE}}, \mathscr{M}_0^{\mathrm{CE}}, \mathscr{M}_0^{\mathrm{CCE}}$。首先，我们界定 $\mathscr{M}_0^{\mathrm{NE}}$ 的极小极大速率：

命题 A.5。 对于任意 $T \in \mathbb{N}$、$\delta \in (0, 1)$，存在一个针对实例 $\mathscr{M}_0^{\mathrm{NE}}$的算法，并保证以至少 $1 - \delta$ 的概率，**风险**$(T) \le \sqrt{}$(最大$_k$ d_k) $\cdot d \cdot T^{-1} \cdot \mathrm{polylog}(T$、$d$、$\delta^{-1})$，其中 $d = d_1 d_2 \cdots d_K$。

命题 A.5 的证明。类似于我们对有限动作标准形博弈所采用的记号，记 $\mathscr{M}$纳什均衡$_0$ $=$($\mathcal{M}$、Π^{NE}、$\mathcal{O}$、$\{(\Pi'_k)^{\mathrm{NE}}\}_k$、$\{U_k^{\mathrm{NE}}\}_k$)。命题 A.1 与命题 4.1 的结合表明，dec$_\varepsilon$($\mathscr{M}_0^{\mathrm{NE}}$) $\le 2\varepsilon \cdot \sqrt{K \cdot \sum_{k=1}^{K} d_k}$。对于任意 $\pi \in \Pi^{\mathrm{NE}} \subset \Delta(\mathcal{A})$，在 $\mathcal{O} = \mathcal{R}^K \times \mathcal{O}_{\mathrm{o}} = \mathcal{R}^K \times \mathcal{A}$上由 $\nu(\cdot|\pi)$:= 均匀分布($\{-1, 1\}^K$) $\times\, \pi$ 定义的分布满足：$\mathcal{M}$ 满足带有 $B = 2^K$ 的假设 D.2，因此命题 D.1 给出 Est$_{\mathsf{H}}$(T、δ) $=$ O(est($\mathcal{M}$、T) $+\log\, \delta^{-1})) \cdot K^2$。最后注意，对于 $k \in [K]$，关于欧几里得范数的 ε^2/D^K-覆盖的乘积产生了一个 ε-模型类覆盖$\mathcal{M}$，其含义见定义 D.1。由于每个 Θ_k 都有一个大小为 $O(D^{K+1}/\varepsilon^2)^d$的 ε^2/D^K-覆盖，因此可得$\mathcal{N}(\mathcal{M}$、$\varepsilon) \le (D^{K+1}/\varepsilon^2)^{Kd}$，这意味着 est($\mathcal{M}$、$T$) $\le O($ $K^2 d \cdot \log(TD))$。因此，定理 D.1 与定理 2.1 相结合，表明存在一个算法，其

$$\mathbf{Risk}(T) \le \sqrt{K(d_1 + \cdots + d_K)} \cdot \sqrt{\frac{\mathbf{Est}_{\mathsf{H}}(T, \delta)}{T}} \cdot \mathrm{polylog}(T, 1/\delta) \le \sqrt{\frac{\max_k d_k \cdot d}{T}} \cdot \mathrm{polylog}(T, 1/\delta, d).$$

□

与有限动作标准式博弈的情形一样，命题 A.5 中的上界（也适用于实例 $\mathscr{M}_0^{\mathrm{CE}}$、$\mathscr{M}_0^{\mathrm{CCE}}$）受到多智能体诅咒的影响。对于实例 $\mathscr{M}_0^{\mathrm{CCE}}$，借助我们在 第 5 节中的结果，我们得到了改进的界，其极小极大风险随 K 仅呈多项式增长。

命题 A.6。 对于任意 $T \in \mathbb{N}, \delta \in (0, 1)$，存在一个针对实例 $\mathscr{M}_0^{\mathrm{CCE}}$的算法，它保证以至少 $1 - \delta$ 的概率，

$$\mathbf{Risk}(T) \le \sqrt{\frac{K^5 \cdot \max_k\{d_k\}}{T}} \cdot \log(KDT/\delta).$$

人们可能会想，能否为实例 $\mathscr{M}_0^{\mathrm{CE}}$建立类似的上界。根据我们对 $\mathscr{M}_0^{\mathrm{CE}}$的定义（根据 定义 A.1，这是一个 CE 实例），我们有 $|\Pi'_k| = |\mathcal{A}_k|^{|\mathcal{A}_k|}$ 对于每个 k，这意味着 定理 5.1 的上界将给出一个对 $|\mathcal{A}_k|$具有多项式依赖关系的风险上界，这

在线性老虎机设定中是不可接受的，因为 $\mathcal{A}_k$ 通常被认为是指数级大，甚至是无穷大。即使我们试图利用推论 5.1 来缩小偏离集合的规模，普遍适用的偏离集合也只有一种选择，即 $\widetilde{\Pi}'_k := \widetilde{\Sigma}_k^{\Sigma_k}$其对数随维度 d_k 呈指数级增长。更有希望的方向是考虑介于粗相关均衡和相关均衡之间的均衡概念（有时称为 Φ-均衡），例如 Gordon 等人 （2008年）； Anagnostides 等人 （2022年）； Mansour 等人 （2022年）； 我们将这一方向留待未来工作。

命题 A.6 的证明。该命题可由推论 5.1 推出。仿照标准形博弈的记号，我们记作 $\mathscr{M}\mathrm{CCE}_0 = (\mathcal{M}, \Pi^{\mathrm{CCE}}, \mathcal{O}, \{(\Pi'_k)^{\mathrm{CCE}}\}_k, \{U_k^{\mathrm{CCE}}\}_k)$。记$\mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_K := \{a_1 \otimes \cdots \otimes a_K : a_1 \in \mathcal{A}_1, \ldots, a_K \in \mathcal{A}_K\}$。对于每个 $k \in [K]$，存在一个相对于 ℓ_2-范数的 $\mathcal{A}_k$ 的 $\varepsilon/(KD^K)$-覆盖，其大小至多为 $O(KD^{K+1}/\varepsilon)^{d_k}$。将这样的覆盖记为 $\widetilde{\mathcal{A}}_k \subseteq \mathcal{A}_k$。

记 $\widetilde{\mathcal{A}} = \widetilde{\mathcal{A}}_1 \times \cdots \times \widetilde{\mathcal{A}}_K$, $\Pi^{\mathrm{CCE}} = \Delta(\widetilde{\mathcal{A}})$，以及对于每个 $k \in [K]$, $(\Pi'_k)^{\mathrm{CCE}} := \widetilde{\mathcal{A}}_k \cup \{\perp\}$。考虑任意模型 $M \in \mathcal{M}$。注意，对于任意 $k \in [K]$和 $a_k \in \mathcal{A}_k$，都存在某个 $\widetilde{a}'_k \in \widetilde{\mathcal{A}}_k$，使得对于所有 $\widetilde{a} \in \widetilde{\mathcal{A}}$,

$$\begin{aligned}|f_k^M(U_k(a_k, \widetilde{a})) - f_k^M(U_k(\widetilde{a}'_k, \widetilde{a}))| =& |\langle \widetilde{a}_1 \otimes \cdots \otimes a_k \otimes \cdots \otimes \widetilde{a}_K, \theta_k^M \rangle - \langle \widetilde{a}_1 \otimes \cdots \otimes \widetilde{a}'_k \otimes \cdots \otimes \widetilde{a}_K, \theta_k^M \rangle| \\ \le& \|\theta_k^M\|_2 \cdot D^{K-1} \cdot \|\widetilde{a}_k - \widetilde{a}'_k\|_2 \le \varepsilon/K,\end{aligned}$$

这尤其意味着，实例 $\widetilde{\mathscr{M}}\mathrm{CCE0} := (\mathcal{M}, \Pi^{\mathrm{CCE}}, \mathcal{O}, \{(\Pi'_k)^{\mathrm{CCE}}\}_k, \{U_k^{\mathrm{CCE}}\}_k)$ 是 ε-决策空间覆盖，可覆盖 $\mathscr{M}\mathrm{CCE}$（根据定义 5.1）。因此可得，$\mathrm{est}_\Pi(\widetilde{\mathscr{M}}\mathrm{CCE0}, T) \le K \cdot \max_k\{d_k\} \cdot \log(KDT)$。根据命题 A.4 以及类 $\mathcal{M}$ 的凸性（这是因为集合 Θ_k 是凸的），~~有 r-dec~~$^{\mathrm{o}}_\gamma(\mathrm{co}(\widetilde{\mathscr{M}}_0^{\mathrm{CCE}})) \le \text{r-dec}^{\mathrm{o}}_\gamma(\mathrm{co}(\mathscr{M}\mathrm{CCE})) = \text{r-dec}^{\mathrm{o}}_\gamma(\mathscr{M}\mathrm{CCE}) \le \frac{K \cdot \sum_{k=1}^K d_k}{\gamma}$。由推论 5.1 可知，存在一种算法，其

$$\begin{aligned}\mathbf{Risk}(T) \le& O(K) \cdot \inf_{\gamma > 0} \left\{ \frac{K \cdot \sum_{k=1}^K d_k}{\gamma} + \frac{\gamma}{T} \cdot K \cdot \max_k\{d_k\} \cdot \log(KDT/\delta) \right\} \\ \le& \sqrt{\frac{K^5 \cdot \max_k\{d_k\}}{T}} \cdot \log(KDT/\delta).\end{aligned}$$

□

下界。现在，在假设 $\mathcal{A}_k, \Theta_k$ 分别包含其所在空间的单位 ℓ_2 球的前提下，我们推导实例 $\mathscr{M}_0^{\mathrm{CCE}}, \mathscr{M}\mathrm{CE0}, \mathscr{M}\mathrm{NE0}$ 的下界。[18] 由 Foster 等人（2021）命题 6.2 的证明可知，$\mathrm{dec}_\varepsilon(\mathscr{M}_0^{\mathrm{NE}}) \ge \mathrm{dec}_\varepsilon(\mathscr{M}_0^{\mathrm{CE}}) \ge \mathrm{dec}_\varepsilon(\mathscr{M}_0^{\mathrm{CCE}}) \ge \Omega(\varepsilon\sqrt{\max_k d_k})$，这是利用了这些实例都嵌入了一个维度为 $\max_k d_k$ 的单玩家线性老虎机实例这一事实。因此，定理 3.2（当 $\varepsilon(T) = \sqrt[c]{\max_{KT}}^{k\ d_k}$ ~~且 $c > 0$~~ 足够小时）结合定理 2.1 表明，对于任意一个实例 $\mathscr{M}_0^{\mathrm{CCE}}, \mathscr{M}_0^{\mathrm{CE}}, \mathscr{M}_0^{\mathrm{NE}}$ 以及任意算法，都存在一个模型，使得在这三种实例中的任意一种下，$\mathbb{E}[\mathbf{Risk}(T)] \ge \Omega(\max_k d_k/(KT))$。关于上界与下界之间差距的类似讨论见附录 A.3.1。

A.3.3 凹（老虎机）博弈

现在我们界定示例 A.3 中给出的凹奖励标准形博弈的 DEC 和极小极大遗憾。固定示例 A.3 中描述的集合 $\mathcal{A}_k \subset \mathbb{R}^{d_k}$ 和类 $\mathcal{F} \subset (\mathcal{A} \to \mathbb{R}^K)$。我们假设对于所有 $a \in \mathcal{A}$, $\|a\|_2 \le D$ 均成立，其中 $D > 0$ 为某个值；我们的界仅以对数方式依赖于 D。令 $\mathscr{M}\mathrm{NE} = (\mathcal{M}, \Pi^{\mathrm{NE}}, \mathcal{O}, \{(\Pi'_k)^{\mathrm{NE}}\}_k, \{U_k^{\mathrm{NE}}\}_k)$ 表示示例 A.3 中构造的纳什均衡实例，并令$_{\mathscr{M}}\mathrm{CCE} = (\mathcal{M}, \Pi^{\mathrm{CCE}}, \mathcal{O}, \{(\Pi'_k)^{\mathrm{CCE}}\}_k, \{U_k^{\mathrm{CCE}}\}_k)$ 表示示例 A.3 中构造的粗相关均衡实例。[19] 下面的命题给出了这些实例的（遗憾）偏移 DEC 的界：

[18]在其他动作集和参数集下，也可以得到类似的下界；例如，如果 $\mathcal{A}_k$ 每个都包含 ℓ_1 球，且 Θ_k 每个都包含 ℓ_∞ 球，那么我们就可以嵌入上一小节中的标准形博弈设定。[19]与前文一样，对于两个实例 $\mathscr{M}^{\mathrm{NE}}$, $\mathscr{M}^{\mathrm{CCE}}$，我们都使用模型类 $\mathcal{M}$，其中默认模型的定义域适用于各自的实例。

命题 A.7. 对于任意 $\gamma > 0$，上述定义的实例 $\mathscr{M}^{\mathrm{NE}}, \mathscr{M}^{\mathrm{CCE}}$ 满足

$$\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{CCE}}) \le \text{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{NE}}) \le \frac{K \cdot \sum_{k=1}^{K} d_k^4}{\gamma} \cdot \mathrm{polylog}\left(\max_k\{d_k\}, D, \gamma\right).$$

命题 A.7 的证明。由命题 A.1 的证明可知，$\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}_{\text{粗相关均衡}}) \le \text{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{NE}})$，这是因为 Π^{NE}可视为 Π^{CCE} 的子集（也就是说，Π^{NE} 由 Π^{CCE}中的单点分布组成）。为证明第二个上界，我们将定理 4.1 应用于实例 $\mathscr{M}^{\mathrm{NE}}$，由此得到 $\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{NE}}) \le \sum_{k=1}^{K} \text{r-dec}^{\mathrm{o}}_{\gamma/K}(\widetilde{\mathcal{M}}_k)$，其中 $\widetilde{\mathcal{M}}_k$ 定义为$\widetilde{\ }$（见式 (17)）。接下来，为了界定 $\mathcal{M}_k$的 DEC，我们定义模型类 $\mathcal{M}'_k \subset (\mathcal{A}_k \to \Delta(\mathcal{R} \times \{\perp\}))$，其定义为：$\mathcal{M}'_k = \{M\colon\ f^M(\cdot)$ 是凹的$\}$。由于对于任意 $k \in [K]$、$M \in \mathcal{M}^{\mathrm{NE}}$、$a_{-k} \in \mathcal{A}_{-k}$，都存在一个模型 $M'_k \in \mathcal{M}'_k$，使得对所有 $a_k \in \mathcal{A}_k$，$r \sim M'_k(a_k)$的分布与 $r_k \sim M(a_k, a_{-k})$的分布相同，因此$\widetilde{\ }$有 $\text{r-dec}^{\mathrm{o}}_{\gamma/K}(\mathcal{M}_k) \le \text{r-dec}^{\mathrm{o}}_{\gamma/K}(\mathcal{M}'_k)$。最后，Foster 等人（2021）的命题 6.3（这是 Lattimore（2020）定理 3 的重述）给出，对于所有 $\gamma' > 0$，$\text{r-dec}^{\mathrm{o}}_{\gamma'}(\mathcal{M}'_k) \le \frac{d_k^4}{\gamma'} \cdot \mathrm{polylog}(d_k, \mathrm{diam}(\mathcal{A}_k), \gamma')$，从而得到 $\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{NE}}) \le \frac{K \sum_{k=1}^{K} d_k^4}{\gamma} \cdot \mathrm{polylog}\left(\max_k\{d_k\}, D, \gamma\right)$。

□

现在，我们将注意力转向极小极大风险的界的推导。对于我们的凹博弈实例，其模型类 $\mathcal{M}^{\mathrm{NE}}$、$\mathcal{M}^{\mathrm{CCE}}$ 规模极其庞大：根据 $\mathcal{M}^{\mathrm{NE}}$ 或 $\mathcal{M}^{\mathrm{CCE}}$ 的 定义 D.1，其覆盖的对数必须随维度 d_k 呈指数增长，因此，来自定理 D.1 的基于模型的保证并不特别有意义，即使在 K 较小的情况下也是如此。因此，我们直接转向 第 5 节 中给出的基于策略的保证，并将证明实例 $\mathscr{M}^{\mathrm{CCE}}$ 的极小极大风险上界。事实证明，在以下假设下，这样的上界将立即推出实例 $\mathscr{M}^{\mathrm{NE}}$ 的上界，该假设是对 Even-Dar 等人 (2009) 的特化。

假设 A.1 (零和社会凹性)。我们称模型类 $\mathcal{M}$ 为零和社会凹性，如果对于所有 $M \in \mathcal{M}$、$k \in [K]$ 和 $a_k \in \mathcal{A}_k$，映射 $\mathcal{A}_{-k} \ni a_{-k} \mapsto f_k^M(a_k, a_{-k})$ 是一个凸函数，并且对于所有 $a \in \mathcal{A}$、$\sum_{k=1}^{K} f_k^M(a) = 0$。

在模型 M 是双人零和凹博弈（即 $f_1^M + f_2^M \equiv 0$）这一特殊情形下，零和社会凹性必然成立。

命题 A.8. 那么，对于任意 $T \in \mathbb{N}, \delta \in (0, 1)$，存在一个针对实例 $\mathscr{M}^{\mathrm{CCE}}$ 的算法，保证以至少 $1-\delta$ 的概率，

$$\mathbf{Risk}(T) \le \frac{K^2 \cdot \max_k\{d_k^{2.5}\}}{\sqrt{T}} \cdot \mathrm{polylog}\left(D, T, \gamma, \max_k\{d_k\}, K, 1/\delta\right). \tag{20}$$

进一步假设模型类 $\mathcal{M}$ 具有零和社会凹性（即满足假设 *A.1*）。那么，存在一个针对实例 $\mathscr{M}^{\mathrm{NE}}$ 的算法，以至少 $1-\delta$ 的概率保证式 *(20)* 中的风险上界不变。

命题 A.8 的证明。对于每个 $k \in [K]$，都存在一个关于 ℓ_2-范数的 $\mathcal{A}_k$ 的 ε-覆盖，其大小至多为 $O(D/\varepsilon)^{d_k}$。将这样的覆盖记为 $\widetilde{\mathcal{A}}_k \subset \mathcal{A}_k$。记 $\widetilde{\mathcal{A}} := \widetilde{\mathcal{A}}_1 \times \cdots \times \widetilde{\mathcal{A}}_K$。令 $\widetilde{\Pi}^{\mathrm{CCE}} := \Delta(\widetilde{\mathcal{A}})$，$\widetilde{\ }$和 $(\Pi'_k)^{\mathrm{CCE}} := \mathcal{A}_k \cup \{\perp\}$。注意，对于任意 $M \in \mathcal{M}$、$k \in [K]$和 $a_k \in \mathcal{A}_k$，都存在某个 $\widetilde{a}'_k \in \mathcal{A}_k$，使得对于所有 $\widetilde{a} \in \mathcal{A}$，$\widetilde{\ }$

$$|f_k^M(U_k(a_k, \widetilde{a})) - f_k^M(U_k(\widetilde{a}'_k, \widetilde{a}))| = |f_k^M(a_k, \widetilde{a}_{-k}) - f_k^M(\widetilde{a}'_k, \widetilde{a}_{-k})| \le \|a_k - \widetilde{a}'_k\|_2 \le \varepsilon,$$

$\widetilde{\ }$ $\widetilde{\ }$ $\widetilde{\ }$其中，第一个不等式使用了 $f_k^M(\cdot)$ 的 1-利普希茨性。因此，CCE 实例 $\mathscr{M}_{\text{粗相关均衡}} := (\mathcal{M},\ \Pi^{\mathrm{CCE}},\ \mathcal{O},\ \{(\Pi'_k)^{\mathrm{CCE}}\}_k,\ \{U^{\text{粗相关均衡}}_k\}_k)$ 是 ε-决策空间覆盖，覆盖 $\mathscr{M}_{\text{粗相关均衡}}$（根据 定义 5.1）。由此可得 $\mathrm{est}_{\Pi}(\mathscr{M}_{\text{粗相关均衡}}, T) \le \max_k\{d_k\} \cdot \log(DT)$。

接下来，我们有 $\text{r-dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathscr{M}^{\mathrm{CCE}})) = \text{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}_{\text{粗相关均衡}}) \le \frac{K \cdot \sum_{k=1}^{K} d_k^4}{\gamma} \cdot \mathrm{polylog}(\max_k\{d_k\},\ D,\ \gamma)$ 根据 命题 A.7以及该类的凸性 $\mathcal{M}$ (这是因为凹函数与 1-Lipschitz 函数的凸组合

是凹且 1-Lipschitz 的）。根据 推论 5.1，对于任意 $T \in \mathbb{N}$ 和 $\delta > 0$，存在一个算法，其输出$\widehat{\pi} \in \Pi^{\mathrm{CCE}}$，使得其以至少 $1-\delta$的概率，

$$\begin{aligned}\mathbf{Risk}(T) \leq & K \cdot \inf_{\gamma>0}\left\{\frac{K \cdot \sum_{k=1}^{K} d_k^4}{\gamma}+\frac{\gamma}{T} \cdot \max_k\{d_k\}\right\} \cdot \operatorname{polylog}\left(D, T, \gamma, \max_k\{d_k\}, K, 1/\delta\right) \\ \leq & \frac{K^2 \cdot \max_k\{d_k^{2.5}\}}{\sqrt{T}} \cdot \operatorname{polylog}\left(D, T, \gamma, \max_k\{d_k\}, K, 1/\delta\right).\end{aligned}$$

接下来我们证明 $\mathscr{M}^{\mathrm{NE}}$ 的上界。如本节前文所述，对于 $\widehat{\pi} \in \Pi^{\mathrm{CCE}}$ 和$M \in \mathcal{M}$，我们用 $h^{M,\mathrm{CCE}}(\widehat{\pi})$ 表示 $\widehat{\pi}$ 相对于实例 $\mathscr{M}^{\mathrm{CCE}}$ 的次优性；对于$\widehat{\pi} \in \Pi^{\mathrm{NE}}$，我们用 $h^{M,\mathrm{NE}}(\widehat{\pi})$ 表示 $\widehat{\pi}$ 相对于实例 $\mathscr{M}^{\mathrm{NE}}$ 的次优性。给定 $\widehat{\pi} \in \Pi^{\mathrm{CCE}}$，定义 $\widehat{a} := \mathbb{E}_{a\sim\widehat{\pi}}[a] \in \mathcal{A}$。对于每个 $k \in [K]$ 和 $M \in \mathcal{M}$，有

$$\begin{aligned}h^{M,\mathrm{NE}}(\widehat{a}) &= \sum_{k=1}^{K} \max_{a_k' \in \mathcal{A}_k} f_k^M(a_k', \widehat{a}_{-k}) - f_k^M(\widehat{a}) \\ &\leq \sum_{k=1}^{K} \max_{a_k' \in \mathcal{A}_k} \mathbb{E}_{a_{-k}\sim\widehat{\pi}}[f_k^M(a_k', a_{-k})] - f_k^M(\widehat{a}) \\ &= \sum_{k=1}^{K} \max_{a_k' \in \mathcal{A}_k} \mathbb{E}_{a_{-k}\sim\widehat{\pi}}[f_k^M(a_k', a_{-k})] - \mathbb{E}_{a\sim\widehat{\pi}}[f_k^M(a)] = h^{M,\mathrm{CCE}}(\widehat{\pi}),\end{aligned}$$

其中，第一个不等式由社会凹性推出，第二个等式则由以下事实推出：$\sum_{k=1}^{K} f_k^M(\widehat{a}) = 0 = \sum_{k=1}^{K} \mathbb{E}_{a\sim\widehat{\pi}}[f_k^M(a)]$。因此，给定算法针对实例$\mathscr{M}^{\mathrm{CCE}}$输出的决策 $\widehat{\pi}$，我们只需输出 $\widehat{a} = \mathbb{E}_{a\sim\widehat{\pi}}[a]$，即可得到相同的风险上界。□

下界。假设 $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_K$包含单位 ℓ_2球。那么，实例 $\mathscr{M}$CCE $\mathscr{M}$NE 各自都嵌入了一个维度为 $\max_k d_k$ 的单玩家线性赌博机实例（即将 $\mathcal{F}$ 的子类取为仅包含 a_k 中的线性函数），因此附录 A.3.2 中的下界给出 $\mathrm{dec}_\varepsilon(\mathscr{M}^{\mathrm{NE}}) \geq \mathrm{dec}_\varepsilon(\mathscr{M}\mathrm{CCE}) \geq \Omega(\varepsilon\sqrt{\max_k d_k})$，以及 $\Omega(\max_k d_k/(KT))$ 的极小极大风险下界。在这种设定下，即使是 DEC 下界（在单智能体设定中）也与命题 A.7 和命题 4.1（Foster 等，2021）所蕴含的上界相差甚远。

A.3.4 表格型马尔可夫博弈

接下来，我们给出表格型马尔可夫博弈中与马尔可夫纳什均衡、马尔可夫 CE 和马尔可夫 CCE 对应的实例的极小极大风险界，如示例 A.4 所述。给定 $H \in \mathbb{N}$，状态空间$\mathcal{S}_h$ 的大小均为 S，动作空间 $\mathcal{A}_k$ 的大小为 $A_k := \mathcal{A}_k$，以及初始分布 $d^1 \in \Delta(\mathcal{S}_1)$，令 $\mathscr{M}^{\mathrm{NE}} = (\mathcal{M}, \Pi^{\mathrm{NE}}, \mathcal{O}, \{(\Pi_k')^{\mathrm{NE}}\}_k, \{U^{\mathrm{NE}}_k\}_k)$、$\mathscr{M}\mathrm{CE} = (\mathcal{M}, \Pi^{\mathrm{CE}}, \mathcal{O}, \{(\Pi_k')^{\mathrm{CE}}\}_k, \{U^{\mathrm{CE}}_k\}_k)$ 和 $\mathscr{M}\mathrm{CCE} = (\mathcal{M}, \Pi^{\mathrm{CCE}}, \mathcal{O}, \{(\Pi_k')^{\mathrm{CCE}}\}_k, \{U_k^{\mathrm{CCE}}\}_k)$ 分别表示示例 A.4 中定义的、对应于马尔可夫纳什均衡、马尔可夫 CE 和马尔可夫 CCE 的 MA-DMSO 实例。严格来说，$\mathscr{M}^{\mathrm{NE}}$ 对应的模型类 $\mathcal{M}$ 作用于 Π^{NE} 中的策略，而 $\mathscr{M}^{\mathrm{CE}}$ 和 $\mathscr{M}^{\mathrm{CCE}}$ 对应的模型类 $\mathcal{M}$ 作用于 $\Pi^{\mathrm{CE}} = \Pi^{\mathrm{CCE}} \neq \Pi^{\mathrm{NE}}$ 中的策略；为避免符号过于繁杂，我们将每个实例对应的模型类记为 $\mathcal{M}$，并将其定义域形式化地解释为相应的决策空间。

在命题 A.9 中，我们首先给出其偏移 DEC 的一个上界，由此可通过命题 4.1立即得到受约束 DEC 的一个上界。

命题 A.9。 对于任意 $\gamma > 0$ 和任意 $\overline{M} \in \mathcal{M}$，上文定义的实例 $\mathscr{M}^{\mathrm{NE}}$、$\mathscr{M}^{\mathrm{CE}}$、$\mathscr{M}^{\mathrm{CCE}}$ 满足 y

$$\text{r-dec}^{\mathrm{o}}_\gamma(\mathscr{M}^{\mathrm{CCE}}, \overline{M}) \leq \text{r-dec}^{\mathrm{o}}_\gamma(\mathscr{M}^{\mathrm{CE}}, \overline{M}) \leq \text{r-dec}^{\mathrm{o}}_\gamma(\mathscr{M}^{\mathrm{NE}}, \overline{M}) \leq \frac{27KH^3\log(H)S\sum_{k=1}^{K} A_k}{\gamma}.$$

证明命题 A.9。与命题 A.1的证明类似，我们给函数 $f^M(\cdot)$和 $h^M(\cdot)$添加上标 NE/CE/CCE，以区分三种不同模型的价值函数

实例。举例来说，对于实例 $\mathscr{M}^{\mathrm{NE}}$，对于 $M \in \mathcal{M}$，$\pi \in \Pi^{\mathrm{NE}}$，

$$f_k^{M,\mathrm{NE}}(\pi) := \mathbb{E}^{M,\pi}\left[\sum_{h=1}^{H} r_{k,h}\right], \qquad h^{M,\mathrm{NE}}(\pi) = \sum_{k=1}^{K} \max_{\pi'_k \in (\Pi'_k)^{\mathrm{NE}}} f_k^{M,\mathrm{NE}}(U_k^{\mathrm{NE}}(\pi'_k, \pi)) - f_k^{M,\mathrm{NE}}(\pi).$$

函数 $h^{M,\mathrm{CE}}: \Pi^{\mathrm{CE}} \to \mathbb{R}$ 和 $h^{M,\mathrm{CCE}}: \Pi^{\mathrm{CCE}} \to \mathbb{R}$ 的定义类似。

我们有 $\Pi^{\mathrm{CE}} = \Pi^{\mathrm{CCE}}$；此外，对于任意 $M \in \mathcal{M}$ 和 $\pi \in \Pi^{\mathrm{CE}} = \Pi^{\mathrm{CCE}}$，有 $h^{M,\mathrm{CCE}}(\pi) \le h^{M,\mathrm{CE}}(\pi)$。因此，$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}$粗相关均衡$M) \le \overline{\textsf{r-dec}}^{\mathrm{o}}_{\gamma}(\mathscr{M}$相关均衡$\overline{M})$。接下来，注意到我们可以如下将 Π^{NE}视为 Π^{CE}的子集：对于 $\pi = (\pi_1, \ldots, \pi_K) \in \Pi^{\mathrm{NE}}$，我们将其对应到联合马尔可夫策略 $\widetilde{\pi} = (\widetilde{\pi}_1, \ldots, \widetilde{\pi}_H) \in \Pi^{\mathrm{CE}}$，其中 $\widetilde{\pi}_h(s_h)$ 是乘积分布 $\widetilde{\pi}_h(s_h) := \pi_{1,h}(s_h) \times \cdots \times \pi_{K,h}(s_h)$。显然，对于这样的 π以及任意模型 $M \in \mathcal{M}$，$M(\pi)$ 和 $M(\widetilde{\pi})$ 的分布相同。因此，稍微滥用一下记号，对于 $\pi \in \Pi^{\mathrm{NE}}$，我们也将其在 Π^{CE} 中的对应策略记作 π。于是我们有$h^{M,\mathrm{NE}}(\pi) = h^{M,\mathrm{CE}}(\pi)$，并且对于任意 $M \in \mathcal{M}$，有

$$\begin{aligned}\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{NE}}, \overline{M}) &= \inf_{p \in \Delta(\Pi^{\mathrm{NE}})} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[h^{M,\mathrm{NE}}(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] \\ &\ge \inf_{p \in \Delta(\Pi^{\mathrm{CE}})} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[h^{M,\mathrm{CE}}(\pi) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)\right] = \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{CE}}, \overline{M}).\end{aligned}$$

仍需给出$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{NE}})$的上界。对于 $k \in [K]$，令 Π_k 为玩家 k的随机马尔可夫策略类（使得 $\Pi^{\mathrm{NE}} = \Pi_1 \times \cdots \times \Pi_K$）。对于每个 $k \in [K]$，定义模型类 $\mathcal{M}_k \subset (\Pi_k \to \Delta(\mathcal{R} \times \mathcal{O}_{\circ}))$如(17)所示：

$$\widetilde{\mathcal{M}}_k = \{\pi_k \mapsto M|_k(\pi_k, \pi_{-k}) \ : \ \pi_{-k} \in \Pi_{-k}, \ M \in \mathcal{M}\}.$$

定义 $\mathcal{M}'_k$ 为由所有时域为H、行动集为 $\mathcal{A}_k$且状态空间为 $\mathcal{S}_1$、…、$\mathcal{S}_H$的马尔可夫决策过程组成的模型类，并且任何以正概率发生的轨迹所对应的奖励之和均有界于 $[0,1]$。形式上，$\mathcal{M}'_k$ 的纯观测空间是轨迹$\{(s_h, a_{k,h}, r_{k,h})\}_{h \in [H]}$构成的空间 $\mathcal{O}'_{\circ}$，其中 $s_h \in \mathcal{S}_h$、a_k、$_h \in \mathcal{A}_k$、r_k、$_h \in \mathbb{R}$；其奖励空间为 $\mathcal{R} = [0,1]$，决策空间为 Π_k。因此，$\mathcal{M}'_k \subset (\Pi_k \to \Delta(\mathcal{R} \times \mathcal{O}'_{\circ}))$。Foster 等人（2021）的命题 5.4 表明，对所有 $M \in \mathcal{M}'_k$，$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}'_k, M) \le 26\frac{H^2 S A_k}{\gamma}$。

接下来，固定$M \in \mathcal{M}_k$。根据 $\mathcal{M}_k$的定义，我们可以找到 $M \in \mathcal{M}$ 和 $\pi_{-k} \in \Pi_{-k}$，使得对于所有 $\pi_k \in \Pi_k$，$M(\pi_k) = M|_k(\pi_k, \pi_{-k})$。令 $M' \in \mathcal{M}'_k$ 为唯一模型，使得对于所有 $\pi_k \in \Pi_k$，从 $M(\pi_k)$中抽取的轨迹的 $\{(s_h, a_{k,h}, r_{k,h})$的边缘分布$\}_{h \in [H]}$与从 $M'(\pi_k)$中抽取的纯观测的分布相同。这样的模型存在，因为对于每个状态 $s_h \in \mathcal{S}_h$ 和动作a_k、$_h \in \mathcal{A}_k$，转移分布 $P_h^{M'}(\cdot|s_h, a_{k,h}) \in \Delta(\mathcal{S}_{h+1})$定义为$\mathbb{E}_{a_{k'}、\ _h \sim \pi_{k'}、\ _h(s_h)\ \forall k' \ne k}[P_h^M(\cdot|s_h, (a_{k,h}, a_{-k,h}))]$，而收益分布$R_k^{M'}$、$(s_h, a_{k,h}) \in \Delta(\mathbb{R})$定义为$\mathbb{E}_{a_{k'}、\ _h \sim \pi_{k'}、\ _h(s_h)\ \forall k' \ne k}[R_h^M(s_h, (a_{k,h}, a_{-k,h}))]$。现在我们计算

$$\begin{aligned}&\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\widetilde{\mathcal{M}}_k, \widetilde{M}) \qquad (21)\\ &= \inf_{p \in \Delta(\Pi_k)} \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} \mathbb{E}_{\pi_k \sim p}\left[\max_{\pi'_k \in \Pi_k} f_k^{M,\mathrm{NE}}(\pi'_k, \pi_{-k}) - f_k^{M,\mathrm{NE}}(\pi) - \gamma \cdot D^2_{\mathsf{H}}\Big(M(\pi_k, \pi_{-k}), \widetilde{M}(\pi_k)\Big)\right] \\ &\le \inf_{p \in \Delta(\Pi_k)} \sup_{M' \in \mathcal{M}'_k} \mathbb{E}_{\pi_k \sim p}\left[\max_{\pi'_k \in \Pi_k} f_k^{M'}(\pi'_k) - f_k^{M'}(\pi_k) - \gamma \cdot D^2_{\mathsf{H}}\big(M'(\pi_k), \overline{M}'(\pi_k)\big)\right] \\ &= \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}'_k, \overline{M}'),\end{aligned}$$

其中，该不等式成立是因为：通过与构造 M' 时相同的论证，对于任意 $M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}$，都存在某个 $M' \in \mathcal{M}'_k$，使得对于任意 $\pi_k \in \Pi_k$，从 $M(\pi_k, \pi_{-k})$ 中抽取一条轨迹时，$\{(s_h, a_{k,h}, r_{k,h})\}_{h \in [H]}$ 的边缘分布与从 $M'(\pi_k)$ 中抽取一条轨迹的分布相同。此外，我们应用了海林格距离的数据处理不等式，从而得出 $D^2_{\mathsf{H}}\big(M(\pi_k, \pi_{-k}) \ \widetilde{M}(\pi_k)\big)$ 是 $M(\pi_k, \pi_{-k})$ 和 $M(\pi_k)$ 下 $\{(s_h, a_{k,h}, r_{k,h})\}_{h \in [H]}$ 的边缘分布之间平方海林格距离的一个上界。最后，将定理 4.2 应用于该实例

$\mathscr{M}^{\mathrm{NE}}$，我们得到

$$\begin{aligned}\sup_{\overline{M}\in\mathcal{M}} \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}^{\mathrm{NE}},\overline{M}) \leq& \frac{CKH\log H}{\gamma}+\sum_{k=1}^{K}\sup_{\widetilde{M}_k\in\widetilde{\mathcal{M}}_k}\textsf{r-dec}^{\mathrm{o}}_{\gamma/(CKH\log H)}(\widetilde{\mathcal{M}}_k,\widetilde{M}_k)\\ \leq& \frac{CKH\log H}{\gamma}+\sum_{k=1}^{K}\sup_{\overline{M}'_k\in\mathcal{M}'_k}\textsf{r-dec}^{\mathrm{o}}_{\gamma/(CKH\log H)}(\mathcal{M}'_k,\overline{M}'_k)\\ \leq& \frac{CKH\log H}{\gamma}+\sum_{k=1}^{K}26H^2SA_k\cdot\frac{CKH\log H}{\gamma}\\ \leq& \frac{27KH^3\log(H)S\sum_{k=1}^{K}A_k}{\gamma}.\end{aligned}$$

□

利用命题 A.9，我们现在对实例 $\mathscr{M}^{\mathrm{NE}}$、$\mathscr{M}^{\mathrm{CE}}$、$\mathscr{M}$粗相关均衡的极小极大速率进行界定。为简化问题，我们假设奖励分布已知。形式化地说，我们固定一些函数 $R^{\star}_{k,h}:\mathcal{S}_h\times\mathcal{A}\to\Delta(\mathcal{R})$（对于 $k\in[K]$、$h\in[H]$），并将模型类 $\mathcal{M}$ 限制为满足以下条件的模型：对于所有 $M\in\mathcal{M}$，$R^{\star}_{k,h}(s_h,a)\equiv R^{M}_{k,h}(s_h,a)\in\Delta([0,1/H])$。我们还假设对于所有 k，$A_k\geq 2$ 成立。固定函数 $R^{\star}_{k,h}$ 后，将所得实例记为 $\mathscr{M}$NE0、$\mathscr{M}^{\mathrm{CE}}_0$、$\mathscr{M}$粗相关均衡0。[20]

命题 A.10。 对于每个实例 $\mathscr{M}^{\mathrm{NE}}_0$、$\mathscr{M}_{\text{相关均衡}0}$、$\mathscr{M}_{\text{粗相关均衡}0}$，都存在一个算法，使得以至少 $1-\delta$ 的概率，**风险**$(T)\leq\sqrt{\frac{\text{最大值}_k\, A_k\cdot AS^3H^4}{T}}\cdot\mathrm{polylog}(T,\delta^{-1},A,S,H)$。

命题 A.10 的证明。注意，只需对实例 $\mathscr{M}^{\mathrm{NE}}_0$ 的极小极大风险进行上界估计，因为对于任意 $\widehat{\pi}\in\Pi^{\mathrm{NE}}\subset\Pi^{\mathrm{CE}}=\Pi^{\mathrm{CCE}}$，我们有 $h^{M,\mathrm{CCE}}(\widehat{\pi})\leq h^{M,\mathrm{CE}}(\widehat{\pi})\leq h^{M,\mathrm{NE}}(\widehat{\pi})$。（回顾命题 A.9 的证明中 $h^{M,\mathrm{CCE}}, h^{M,\mathrm{CE}}, h^{M,\mathrm{NE}}$ 的定义。）结合命题 A.9 和命题 4.1 可得，对于任意 $M\in\mathcal{M}$，

$$\mathsf{dec}_{\varepsilon}(\mathscr{M}^{\mathrm{NE}}_0,\overline{M})\leq O\left(\varepsilon\cdot\sqrt{KH^3\log(H)S\sum_{k=1}^{K}A_k}\right).$$

由于 $M\in\mathcal{M}$ 对 DEC 上界的约束，我们需要一个适当的估计算法，即一个满足 $\widehat{\mathcal{M}}=\mathcal{M}$ 的算法（在假设 D.1 的语境下）。为此，我们采用 Foster 等人（2021）的逐层估计器方法。注意，模型类 $\mathcal{M}$ 具有乘积结构 $\mathcal{M}=\mathcal{M}_1\times\cdots\times\mathcal{M}_H$，其中每个 $\mathcal{M}_h$ 都是转移核 $\mathcal{S}_h\times\mathcal{A}\to\Delta(\mathcal{S}_{h+1})$ 的集合；这是一个凸集，因此满足 Foster 等人（2021）的假设 7.2。此外，将转移密度网格化为 ε^2的倍数后，我们有 $\mathcal{N}(\mathcal{M}_h,\varepsilon)\leq(1/\varepsilon^2)^{S^2A}$。因此，根据 Foster 等人（2021）的命题 7.1 和引理 A.16，存在一个估计算法 $\mathrm{Alg}_{\mathsf{Est}}$，其满足 $\widehat{\mathcal{M}}=\mathcal{M}$，且估计误差为 $\mathsf{Est}_{\mathsf{H}}(T,\delta)\leq O(S^2AH)\cdot\mathrm{polylog}(S,H,\delta^{-1},T)$。因此，定理 D.1 与定理 2.1 结合可得，存在一个算法，其

$$\begin{aligned}\mathbf{Risk}(T)\leq&\sqrt{KH^3\log(H)S\sum_{k=1}^{K}A_k}\cdot\sqrt{\frac{\mathbf{Est}_{\mathsf{H}}(T,\delta)}{T}}\cdot\mathrm{polylog}(T,1/\delta)\\ \leq&\sqrt{\frac{\max_k A_k\cdot AS^3H^4}{T}}\cdot\mathrm{polylog}(T,\delta^{-1},A,S,H).\end{aligned}$$

□

[20]基本上，命题 A.10 中的相同论证允许我们对原始实例$\mathscr{M}_{\text{纳什均衡}}$、$\mathscr{M}^{\mathrm{CE}}$、$\mathscr{M}_{\text{粗相关均衡}}$的极小极大风险给出上界；这些实例的奖励未知，但要做到这一点，需要进行重新归一化论证（并损失一个 H 因子），以确保奖励始终有界于 $[0,1]$，，为简洁起见，我们省略该论证。

下界。显然，对于任意 k，每个 $\mathscr{M}^{\mathrm{NE}}, \mathscr{M}^{\mathrm{CE}}, \mathscr{M}$粗相关均衡实例都嵌入了一个对应于状态空间为 $\mathcal{S}_h$、动作空间为 $\mathcal{A}_k$、时域为 H的单玩家马尔可夫决策过程类的实例：具体而言，取 $\mathcal{M}$ 的子类，其中转移和奖励在每一步都仅依赖于玩家 k的动作。由 Foster 等人（2021）命题 5.8 的证明可知，$\mathrm{dec}_\varepsilon(\mathscr{M}$纳什均衡$) \geq \mathrm{dec}_\varepsilon(\mathscr{M}$相关均衡$) \geq \mathrm{dec}_\varepsilon(\mathscr{M}$粗相关均衡$) \geq \Omega(\varepsilon\sqrt{SH \cdot \max_k A_k})$。因此，定理 3.2（其中 $\varepsilon(T) = c\sqrt{SH \cdot \max_{KT} \frac{\max_k A_k}{T} \log}$，对充分小的 $c > 0$）与定理 2.1 结合可得，对于任意 $\sim\mathscr{M}$粗相关均衡 $\mathscr{M}$相关均衡 $\mathscr{M}$纳什均衡实例以及任意算法，都存在一个模型，使得 $\mathbb{E}[\mathrm{Risk}(T)] \geq \Omega(SH \cdot \max_k A_k/(KT))$。

A.3.5 多智能体 DEC 与单智能体 DEC 的一种分离

在前面的小节中，我们对多智能体 DEC 进行了界定，从而也得到了若干赌博机问题的极小极大风险上界（通过应用 定理 D.1 和 定理 2.1）。在所有情况下，我们对多智能体 DEC（针对粗相关均衡、相关均衡和纳什实例）的上界，都是通过应用 定理 4.1，将多智能体 DEC 上界为$\widetilde{\ }$由模型类 $\mathcal{M}_k$ 定义的单智能体 DEC；这些模型类见式 (17)。接下来的命题（证明直接）表明，这种方法一般而言并不紧，说明多智能体 DEC 是一种不同于现有复杂度度量的基本复杂度度量。

命题 *A.11.* 对任意 K、$A \in \mathbb{N}$，存在一个 K-玩家 *MA-DMSO* 纳什均衡实例 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$$\widetilde{\ }$，使得 $\textit{r-dec}^{\mathrm{o}}_\gamma(\mathscr{M}) = 0$，但对所有 $\gamma > 0$，$\textit{r-dec}^{\mathrm{o}}_\gamma(\mathcal{M}_k) \geq \Omega(A/\gamma)$，其中 $\mathcal{M}_k$ 按式 *(17)* 中的定义。

命题 A.11 的证明。固定 K、$A \in \mathbb{N}$，并为每个 k 设置 $\Pi_k = \{0, 1, \ldots, A\}$。令 $\mathcal{R} := [-1, 1]$ 和 $\mathcal{O}_{\mathrm{o}} := \mathcal{A}$。定义 $\mathcal{F} \subseteq (\Pi \to \mathcal{R}^K)$ 为所有满足以下条件的元组 $(f_1, \ldots, f_K)$ 构成的类：$f_k : \Pi \to \mathcal{R}^K$，且对于所有 $\pi \in \Pi$，如果存在某个 k 使得 $\pi_k = 0$，那么对于所有 $k' \in [K]$，$f_{k'}(\pi) = 0$。令 $\mathcal{M} := \mathcal{M}_{\mathcal{F}}$，并按照定义 1.1 定义 Π'_k、U_k。

定义 $\pi_0 = (0, \ldots\ldots, 0)$。由于对于所有 $M \in \mathcal{M}$，$h^M(\pi_0) = 0$，因此可得 $\textsf{r-dec}^{\mathrm{o}}_\gamma(\mathscr{M}) = 0$。另一方面，$\widetilde{\ }$不难看出，每个类 $\mathcal{M}_k$ 都包含一个具有 A 个臂的标准多臂老虎机实例，$\widetilde{\ }$这意味着根据 Foster 等人（2021）的命题 5.3，对于所有 $\gamma > 0$，$\textsf{r-dec}^{\mathrm{o}}_\gamma(\mathcal{M}_k) \geq \Omega(A/\gamma)$。 □

第二部分

证明

B 技术工具

B.1 信息论

本节汇总了若干技术引理，这些引理将在我们的证明中使用。

引理 *B.1*。设 $(\mathcal{X})$、$(\mathscr{X})$ 为测度空间。假设对于每个 $i \in \mathcal{I}$，存在分布P_i、$P'_i \in \Delta(\mathcal{X})$ 以及 $Q \in \Delta(\mathcal{I})$。进一步假设存在可测函数 $\varphi : \mathcal{X} \to \mathcal{I}$，使得，对于每个 $i \in \mathcal{I}$，$\mathbb{P}_{x \sim P_i}(\varphi(x) = i) = \mathbb{P}_{x \sim P'_i}(\varphi(x) = i) = 1$。那么对于任意 f-散度 $D(\cdot \,\|\, \cdot)$，有

$$D(\mathbb{E}_{i \sim Q}[P_i] \,\|\, \mathbb{E}_{i \sim Q}[P'_i]) = \mathbb{E}_{i \sim Q}[D(P_i \,\|\, P'_i)].$$

证明 引理 B.1。由 $D(\mathbb{E}_{i \sim Q}[P_i] \,\|\, \mathbb{E}_{i \sim Q}[P'_i]) \leq \mathbb{E}_{i \sim Q}[D(P_i \,\|\, P'_i)]$ 是由 $D(\cdot \,\|\, \cdot)$的凸性得出的。为证明反方向，我们关于函数 φ 的假设结合数据处理不等式可得

$$D(\mathbb{E}_{i \sim Q}[P_i] \,\|\, \mathbb{E}_{i \sim Q}[P'_i]) \geq D(\mathbb{E}_{i \sim Q}[\mathbb{I}_i \times P_i] \,\|\, \mathbb{E}_{i \sim Q}[\mathbb{I}_i \times P'_i]) = \mathbb{E}_{i \sim Q}[D(P_i \,\|\, P'_i)],$$

其中，最后一个不等式可由例如 Polyanskiy 和 Wu （2014年）推出。 □

引理 B.2（例如，Polyanskiy 和 Wu （2014年））。令（$\mathcal{X}$、$\mathscr{X}$）和（$\mathcal{Y}$、$\mathscr{Y}$）为测度空间，并令 $\mathcal{X}\times\mathcal{Y}$ 配备乘积 σ-代数 $\mathscr{X}\otimes\mathscr{Y}$。令（$x$、$y$）为 $\mathcal{X}\times\mathcal{Y}$ 上的一对随机变量，其服从某个分布 $\mathbb{P}_{x,\,y}$。对于任意 f-散度 $D(\cdot\,\|\,\cdot)$，均有

$$\mathbb{E}_{x\sim\mathbb{P}_x}\left[D\big(\mathbb{P}_{y|x}\,\|\,\mathbb{P}_y\big)\right]=\mathbb{E}_{y\sim\mathbb{P}_y}\left[D\big(\mathbb{P}_{x|y}\,\|\,\mathbb{P}_x\big)\right].$$

引理 B.3（Foster 等人 （2022b）中的引理 B.5）。设 $\mathbb{P}$、$\mathbb{Q}$ 是测度空间上的概率分布（$\mathcal{X}$，$\mathscr{X}$）。对于任意 $\alpha\geq 1$，设 $\mathcal{G}_\alpha:=\{g\colon\mathcal{X}\to\mathbb{R}\colon\ \|g\|_\infty\leq\alpha\}$。则

$$\frac{1}{2}D_{\mathsf{H}}^2(\mathbb{P},\mathbb{Q})\leq\sup_{g\in\mathcal{G}_\alpha}\left\{1-\mathbb{E}_{\mathbb{P}}[e^g]\cdot\mathbb{E}_{\mathbb{Q}}[e^{-g}]\right\}+4e^{-\alpha}.$$

引理 B.4（例如，Foster 等人 （2022b））。考虑测度空间（$\mathcal{X},\mathscr{X}$）、（$\mathcal{Y},\mathscr{Y}$），并设（x,y）是一对随机变量，其服从定义在（$\mathcal{X}\times\mathcal{Y},\mathscr{X}\otimes\mathscr{Y}$）上的某个分布 $\mathbb{P}_{x,\,y}$。则

$$\mathbb{E}_{x\sim\mathbb{P}_x}\left[D_{\mathsf{H}}^2\big(\mathbb{P}_{y|x},\mathbb{P}_y\big)\right]\leq 4\cdot\inf_{\mathbb{Q}\in\Delta(\mathcal{Y})}\mathbb{E}_{x\sim\mathbb{P}_x}\left[D_{\mathsf{H}}^2\big(\mathbb{P}_{y|x},\mathbb{Q}\big)\right].$$

引理 B.4 的证明。任取 $\mathbb{Q}\in\Delta(\mathcal{Y})$。利用海林格距离满足三角不等式这一事实，可得

$$\begin{aligned}\mathbb{E}_{x\sim\mathbb{P}_x}\left[D_{\mathsf{H}}^2\big(\mathbb{P}_{y|x},\mathbb{P}_y\big)\right]\leq&\mathbb{E}_{x\sim\mathbb{P}_x}\left[2\cdot D_{\mathsf{H}}^2\big(\mathbb{P}_{y|x},\mathbb{Q}\big)+2\cdot D_{\mathsf{H}}^2(\mathbb{Q},\mathbb{P}_y)\right]\\ \leq&2\cdot\mathbb{E}_{x\sim\mathbb{P}_x}\left[D_{\mathsf{H}}^2\big(\mathbb{P}_{y|x},\mathbb{Q}\big)\right]+2\cdot D_{\mathsf{H}}^2\big(\mathbb{Q},\mathbb{E}_{x\sim\mathbb{P}_x}[\mathbb{P}_{y|x}]\big)\\ \leq&4\cdot\mathbb{E}_{x\sim\mathbb{P}_x}\left[D_{\mathsf{H}}^2\big(\mathbb{P}_{y|x},\mathbb{Q}\big)\right],\end{aligned}$$

其中，最后一个不等式由平方海林格距离的凸性得出。 □

引理 B.5（Donsker-Varadhan；参见 Polyanskiy 和 Wu （2014年））。设（$\mathcal{X}$，$\mathscr{X}$）是一个测度空间，并设$\mathbb{P}$，$\mathbb{Q}$ 是定义在（$\mathcal{X}$，$\mathscr{X}$）上的概率测度。则

$$D_{\mathsf{KL}}(\mathbb{P}\,\|\,\mathbb{Q})=\sup_{h:\mathcal{X}\to\mathbb{R}}\left\{\mathbb{E}_{X\sim\mathbb{P}}[h(X)]-\log\mathbb{E}_{X\sim\mathbb{Q}}[\exp(h(X))]\right\},$$

其中上确界　取遍所有满足 $h\colon\mathcal{X}\to\mathbb{R}$ $\mathbb{E}_{X\sim\mathbb{Q}}[\exp(h(X))]<\infty$ 的（可测）函数。

引理 B.6。设 $\mathbb{P},\mathbb{Q}$ 是某个概率空间 $(\Omega\ \mathscr{F})$ 上的概率测度。考虑某个事件 $\mathcal{E}\in\mathscr{F}$，使得 $\mathbb{P}(\mathcal{E})\geq 1-\delta$，其中某个 $\delta\in(0,\ 1)$。此外，假设对于所有事件 $\mathcal{F}\in\mathscr{F}$，都有 $\mathbb{P}(\mathcal{E}\cap\mathcal{F})=\mathbb{Q}(\mathcal{E}\cap\mathcal{F})$。那么 $D_{\mathsf{TV}}(\mathbb{P},\mathbb{Q})\leq\delta$。

引理 B.6 的证明。取 $\mathcal{E}'=\Omega$ 可得 $\mathbb{Q}(\mathcal{E})=\mathbb{P}(\mathcal{E})\geq 1-\delta$。于是对于任意事件 $\mathcal{F}\in\mathscr{F}$，有

$$\begin{aligned}|\mathbb{P}(\mathcal{F})-\mathbb{Q}(\mathcal{F})|\leq&|\mathbb{P}(\mathcal{F}\cap\mathcal{E})-\mathbb{Q}(\mathcal{F}\cap\mathcal{E})|+|\mathbb{P}(\mathcal{F}\cap\mathcal{E}^{\mathrm{c}})-\mathbb{Q}(\mathcal{F}\cap\mathcal{E}^{\mathrm{c}})|\\ =&|\mathbb{P}(\mathcal{F}\cap\mathcal{E}^{\mathrm{c}})-\mathbb{Q}(\mathcal{F}\cap\mathcal{E}^{\mathrm{c}})|\leq\delta.\end{aligned}$$

□

B.2 集中不等式

引理 B.7（Foster 等人（2021）的引理 A.4）。设 $(X_t)^{t\in[T]}$ 是适应于滤子 $\mathscr{F}^t$的任意实值随机变量序列。那么，其概率至少为 $1-\delta$，

$$\sum_{t=1}^{T}X_t\leq\sum_{t=1}^{T}\log\left(\mathbb{E}\left[e^{X_t}|\mathscr{F}^{t-1}\right]\right)+\log(1/\delta).$$

B.3 拓扑引理

下面的引理是贝尔热最大值定理的一个特例。

引理 B.8. 设 $\mathcal{U}$、$\mathcal{V}$ 是欧几里得空间中的紧子集，并考虑任意连续函数 $G:\mathcal{U}\times\mathcal{V}\to\mathbb{R}$。定义 $\mathcal{C}\colon \mathcal{V}\to\mathcal{P}(\mathcal{U})$ 为 $\mathcal{C}(v):=\arg\min_{u\in\mathcal{U}}\{G(u,v)\}$。则 $\mathcal{C}$ 是上半连续的。

引理 B.8 的证明。考虑任意序列 $u_n\to u\in\mathcal{U}$、$v_n\to v\in\mathcal{V}$，使得对所有 n，$u_n\in\mathcal{C}(v_n)$。我们希望证明 $u\in\mathcal{C}(v)$，即对所有 $u'\in\mathcal{U}$，$G(u,v)\le G(u',v)$（这足以由 $\mathcal{V}$ 的紧致性证明上半连续性；参见（Beer，1993年，引理 6.2.6））。为此，固定任意 $u'\in\mathcal{U}$ 和 $\epsilon>0$。由 G 的连续性，存在 N，使得对于 $n\ge N$，有 $|G(u_n,v_n)-G(u,v)|\le\epsilon$，且 $|G(u',v_n)-G(u',v)|\le\epsilon$。于是

$$G(u,v)\le G(u_n,v_n)+\epsilon\le G(u',v_n)+\epsilon\le G(u',v)+2\epsilon,$$

并取 $\epsilon\to 0$，我们得到 $G(u,v)\le G(u',v)$。 □

下一个引理是卡库塔尼不动点定理的直接推论。在其表述中，我们记 $\mathcal{X}_{-k}:=\prod_{k'\neq k}\mathcal{X}_{k'}$ 以及 $\mathcal{X}=\prod_{k\in[K]}\mathcal{X}_k$。

引理 B.9。 假设 $\mathcal{X}_1$、……、$\mathcal{X}_K$ 是欧几里得空间中的非空、紧致且凸的子集。假设对于每个 $k\in[K]$，给定一个上半连续函数 $F_k:\mathcal{X}\to\mathcal{P}(\mathcal{X}_k)$，使得对于所有 $x\in\mathcal{X}$，$F_k(x)$ 都是非空、闭且凸的。则存在某个 $x\in\mathcal{X}$，使得

$$x\in F_1(x)\times\cdots\times F_K(x).$$

引理 B.9 的证明。定义 $F\colon \mathcal{X}\to\mathcal{P}(\mathcal{X})$，其中 $F(x):=F_1(x)\times\cdots\times F_K(x)$。显然，对于每个 $x\in\mathcal{X}$，$F(x)$ 都是非空、闭且凸的。此外，我们断言 F 是上半连续的。为证明这一点，考虑任意序列 $x_n\to x$ 和 $y_n\to y$，使得对于每个 $n\in\mathbb{N}$，$y_n\in F(x_n)$。记 $y_n=(y_{n,1},\dots,y_{n,K})$ 和 $y=(y_1,\dots,y_K)$，根据 $F(x_n)$ 的乘积结构，对于每个 $k\in[K]$，有 $y_{n,k}\in F_k(x_n)$。由 F_k 的上半连续性以及 $y_{n,k}\to y_k$ 这一事实，可得 $y_k\in F_k(x)$。因此，$y\in F(x)$。根据卡库塔尼不动点定理（Osborne and Rubinstein，1994 年，引理 20.1），F 有一个不动点，即存在某个 $x\in\mathcal{X}$，使得 $x\in F(x)$。 □

B.4 极小极大定理

定理B.1 (Sion 极小极大定理)。 设 $\mathcal{X}$、$\mathcal{Y}$ 是拓扑向量空间的凸子集，其中 $\mathcal{X}$ 是紧致的。设 $F\colon \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ 是一个函数，满足：*(a)* 对所有 $x\in\mathcal{X}$，映射 $y\mapsto F(x,y)$ 都是凹且上半连续的；*(b)* 对所有 $y\in\mathcal{Y}$，映射 $x\mapsto F(x,y)$ 都是凸且下半连续的。那么

$$\inf_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}}F(x,y)=\sup_{y\in\mathcal{Y}}\inf_{x\in\mathcal{X}}F(x,y).$$

C 第 2 节的证明第 2 节

定理的证明 定理 2.1。 考虑 MA-DMSO 框架的一个实例 $\mathscr{M}=(\mathcal{M},\ \Pi,\ \mathcal{O}_\circ,\ \mathcal{R},\ \{\Pi'_k\}_k,\ \{U_k\}_k)$。

对于所有模型 M 和决策 $\pi\in\Pi$，定义[21]

$$\widetilde{f}^M(\pi):=K-h^M(\pi)=K-\sum_{k=1}^{K}\sup_{\pi'_k\in\Pi'_k}\left\{f_k^M(U_k(\pi'_k,\pi))-f_k^M(\pi)\right\}.$$

[21]在 $\widetilde{f}^M(\pi)$ 的定义中加入 K 是为了方便，从而确保如果对于所有 M，π，$h^M(\pi)\in[0,K]$，那么对于 $\widetilde{f}^M(\pi)$ 也同样成立。

现在固定任意 $M \in \mathcal{M}_\circ$。根据 假设 1.2，存在某个 $\pi^\star \in \Pi$ 使得 $h^M(\pi^\star) = 0_\circ$ 根据 假设 1.3，有 $h^M(\pi) \geq 0$ 对所有 $\pi \in \Pi$均成立。于是

$$\sup_{\pi' \in \Pi} \widetilde{f}^M(\pi') - \widetilde{f}^M(\pi) = K - (K - h^M(\pi)) = h^M(\pi). \tag{22}$$

~注意，实例 $\mathscr{H} = (\mathcal{M}、\ \Pi、\ \mathcal{O}、\ \{f^M\}_M)$ 是定义良好的，因为模型 M 是概率核 $M: \ \Pi \to \mathcal{O} = \mathcal{O}_\circ \times \mathcal{R}^K$，且实例 $\mathscr{H}$ 中的观测空间根据定义为 $\mathcal{O}_\circ$ 因此，第一个所述结论由 (22) 得出，因为对于任意 M，我们有：

$$\begin{aligned} \mathsf{dec}_\varepsilon(\mathscr{M}, \overline{M}) &= \inf_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{H}_{q,\varepsilon}(\overline{M})} \mathbb{E}_{\pi \sim p}[h^M(\pi)] \\ &= \inf_{p,q \in \Delta(\Pi)} \sup_{M \in \mathcal{H}_{q,\varepsilon}(\overline{M})} \mathbb{E}_{\pi \sim p}\left[\sup_{\pi' \in \Pi} \widetilde{f}^M(\pi') - \widetilde{f}^M(\pi)\right] = \mathsf{dec}_\varepsilon(\mathscr{H}, \overline{M}). \end{aligned} \tag{23}$$

最后，我们注意到，由于 $\mathscr{H}$、$\mathscr{M}$ 的决策空间和（完整）观测空间相同，因此两个框架中的算法空间（p、q）和分布 $\mathbb{P}^{M^\star}$、$(p、\ q)$ 也相同。根据 $\mathfrak{M}$（$\mathscr{H}$、T）和 $\mathfrak{M}$（$\mathscr{M}$、T）的定义可知，它们相等。

□

定理 2.2的证明。 考虑一个实例 $\mathscr{H} = (\mathcal{M}、\ \Pi、\ \mathcal{O}、\ \{f^M(\cdot)\}_M)$ 以及某个 $V \in \mathbb{N}_\circ$ 我们首先通过定义实例 $\mathscr{M}$ 的各个组成部分来对其进行具体说明：

- 定义 $\Sigma_1 = \Pi$ 和 $\Sigma_2 = \{0, 1, \ldots, V\}$，$\widetilde{\Pi}_k = \Delta(\Sigma_k)$ 对于 $k \in \{1, 2\}$，以及 $\widetilde{\Pi} := \widetilde{\Pi}_1 \times \widetilde{\Pi}_2$.
- 我们按照纳什均衡实例的标准方式，根据 Π'_k，U_k 对于 $k \in [2]$ 中的 定义 1.1；特别地，~将 $\Pi'_k = \Pi_k = \Delta(\Sigma_k)$ 对于每个 k 以及 $U_k(\pi'_k, \ \pi) = (\pi'_k, \ \pi_{-k})$。
- 定义 $\mathcal{O}_\circ := \mathcal{O} \cup \{\perp\}$，$\mathcal{R} = [-1, 1]$，并设定 $\widetilde{\mathcal{O}} = \mathcal{O}_\circ \times \mathcal{R}^2$。
- 模型类 $\mathcal{M}$ 由元组 $(\widetilde{M}, \ v) \in \mathcal{M} \times \{1, 2, \ldots, V\} = \mathcal{M} \times [V]$索引。特别地，对于每个~ ~这样的元组 (M, v)，我们有一个模型 $M_{M,\ v} \in \mathcal{M}$，其定义如下所述。由于我们正在构造的实例 $\mathscr{M}$ 对应于在一个博弈中计算混合纳什均衡的实例，而该博弈的纯动作集合为 $\Sigma_1, \ \Sigma_2$，因此我们将 $\Sigma_1 \times \Sigma_2$ 纯决策 的元素称为“纯决策”。

~–对于形式为 $(\sigma_1, 0) \in \Sigma_1 \times \Sigma_2$ 的纯决策，$(o_\circ, r_1, r_2) \sim M_{M,v}((\sigma_1, 0))$ 的分布由下式给出：

$$o_\circ \sim M(\sigma_1) \in \mathcal{O} \subset \mathcal{O}_\circ, \qquad r_1 = r_2 = 0.$$

–对于形式为 $(\sigma_1, i) \in \Sigma_2 \times \Sigma_2$ 且满足 $i > 0$ 的纯决策，$(o_\circ, r_1, r_2) \sim M_{M,v}((\sigma_1, i))$ 的分布为：

$$o_\circ = \perp, \qquad r_2 = -r_1 = \begin{cases} -1 & : i \neq v \\ g^M(\sigma_1) & : i = v. \end{cases}$$

~– 对于一般决策 $\pi \in \Pi$，我们可以将 $\pi_k \in \Delta(\Sigma_k)$ 写作 $\pi = \pi_1 \times \pi_2$，其中 $k \in [2]_\circ$ 那么，~ ~分布 $M_{M,v}(\pi)$ 就是 $M_{M,v}(\sigma)$ 的分布，其中 $\sigma = (\sigma_1, \sigma_2)$ 的分布如下：$\sigma_k \sim \pi_k$，其中$k \in [2]_\circ$

为便于在后续证明中引用，我们陈述一个基本的技术引理，它直接推论自 ~constructionof $\mathcal{M}_\circ$

~ ~ ~**引理 C.1.** 对于任意 $\pi = \pi_1 \times \pi_2 \in \Pi$，以及任意 $M' = M_{M,\ v} \in \mathcal{M}$，均有

$$h^{M'}(\pi) \geq \pi_2(\Sigma_2 \backslash \{v\}) \cdot \mathbb{E}_{\sigma_1 \sim \pi_1}[g^M(\sigma_1)] + \pi_2(\Sigma_2 \backslash \{0, v\}).$$

引理 C.1 的证明。 考虑偏差 $\pi_2' = v$，我们有

$$\begin{aligned} h^{M'}(\pi) \geq & \sup_{\pi_2' \in \Pi_2'} \{f_2^{M'}(U_2(\pi_2', \pi)) - f_2^{M'}(\pi)\} \\ \geq & f_2^{M'}(\pi_1 \times \mathbb{I}_v) - f_2^{M'}(\pi_1 \times \pi_2) \\ = & \mathbb{E}_{\sigma_1 \sim \pi_1}[g^M(\sigma_1)] + \pi_2(\Sigma_2 \backslash \{0, v\}) - \pi_2(v) \cdot \mathbb{E}_{\sigma_1 \sim \pi_1}[g^M(\sigma_1)] \\ = & \pi_2(\Sigma_2 \backslash \{v\}) \cdot \mathbb{E}_{\sigma_1 \sim \pi_1}[g^M(\sigma_1)] + \pi_2(\Sigma_2 \backslash \{0, v\}). \end{aligned}$$

□

用 $\mathfrak{M}(\mathscr{M}, T)$ 界定 $\mathfrak{M}(\mathscr{H}, T)$。考虑任意一个能实现 $\mathfrak{M}(\mathscr{M}, T)$ 的算法 $(\widetilde{p}, \widetilde{q})$。我们有$\widetilde{p}$：$\prod_{t=1}^{T}(\widetilde{\Pi} \times \widetilde{\mathcal{O}}) \to \Delta(\widetilde{\Pi})$ 以及 $\widetilde{q} = (\widetilde{q}^1, \ldots, \widetilde{q}^T)$，其中对于每个 $\widetilde{q}^t$：$\prod_{i=1}^{t-1}(\widetilde{\Pi} \times \widetilde{\mathcal{O}}) \to \Delta(\widetilde{\Pi})$（关于 MA-DMSO 框架和 HR-DMSO 框架中的算法如何形式化，请参阅第 1.6 节以了解背景）。

给定 $(\widetilde{p}, \widetilde{q})$，我们为实例 $\mathscr{H}$定义如下算法 (p, q)。对于任意模型 $M \in \mathcal{M}$，该$\widetilde{\ }$算法仅通过与 $M \in \mathcal{M}$ 交互，尝试模拟 $(\widetilde{p}, \widetilde{q})$ 与 $M_{M,v}$ 的交互。对于每个 $i \in [T]$，该算法都会存储内部状态，记为 $(\widetilde{\pi}^i, (\widetilde{o}^i, r_1^i, r_2^i))$，其中存储了相对于 $M_{M,v}$ 所采取的"模拟"$\widetilde{\ }$决策和观测。由于这一内部状态，下面的描述不会明确标出概率核 $p(\cdot|\cdot)$ $q^t(\cdot|\cdot)$。由于这些核将完整历史作为输入，因此存在核 $p(\cdot|\cdot)$ $q^t(\cdot|\cdot)$，它们在轨迹上的分布与下面的算法完全相同，但将其明确写出会比较繁琐。

I特别地，分布 q^t（对于 $t \in [T]$）和 p 按如下方式（隐式地）定义 ：

1. 对于 $t = 1, 2, \ldots, T$:

$\widetilde{\ }$(a) 抽取 $\widetilde{\pi}^t \sim \widetilde{q}^t(\cdot|(\widetilde{\pi}^1, (\widetilde{o}^1, r_1^1, r_2^1)), \ldots, (\widetilde{\pi}^{t-1}, (\widetilde{o}^{t-1}, r_1^{t-1}, r_2^{t-1})))$，使得 $\pi^t \in \Pi$。(b) 抽取 $(\sigma_1^t, \sigma_2^t) \sim \widetilde{\pi}^t$。(c) 分布 q^t 通过取决策 $\sigma_1^t \in \Sigma_1 = \Pi$（隐式地）定义。

(d) 用于选择未来的决策：作为在 σ_1^t 被执行后收到的观测 o^t 的函数，定义

$$(\widetilde{o}^t, r_1^t, r_2^t) = \begin{cases} (o^t, 0, 0) & : \sigma_2^t = 0 \\ (\perp, 1, -1) & : \sigma_2^t > 0. \end{cases}$$

2. 最后，分布 p定义为 $\widehat{\sigma}_1$ 的分布，其中 $\widehat{\pi} \sim \widetilde{p}(\cdot|(\widetilde{\pi}^1, (\widetilde{o}^1, r_1^1, r_2^1)), \ldots, (\widetilde{\pi}^T, (\widetilde{o}^T, r_1^T, r_2^T)))$，且 $\widehat{\sigma}_1 \sim \widehat{\pi}_1$。

$\widetilde{\ }$ $\widetilde{\ }$为了分析该算法，对于每个 $M \in \mathcal{M}$，我们引入一个模型 $M_{M,0}$，其定义与 $M_{M,v}$$\widetilde{\ }$完全相同，只是对于任意 $v \in [V]$，$M_{M,0}((\sigma_1, i))$ 都会以概率 1 输出 $(\perp, 1, -1)$，对于任意 $\sigma_1 \in \Sigma_1, i \in [V]$ 均如此。显然，如果存在某个底层模型 $M \in \mathcal{M}$，使得在使用上述定义的算法 (p, q) 时，$o^t \sim M(\sigma_1^t)$，那么上述定义的 $\{(\widetilde{\pi}^t, (\widetilde{o}^t, r_1^t, r_2^t))\}_{t=1}^T$ 的分布恰好就是在 $\mathbb{P}^{M_M,0,(\widetilde{p},\widetilde{q})}$ 下的$\widetilde{\ }$历史分布。接下来，我们将借助以下论断，它指出，对于某个 $v^\star \in [V]$，我们可以将这个$\widetilde{\ }$分布转换为分布 $\mathbb{P}^{M_{M,v^\star},(\widetilde{p},\widetilde{q})}$：

引理 C.2。 存在一个绝对常数 $C > 0$，使得对于任意算法 $(\widetilde{p}, \widetilde{q})$ 和模型$M \in \mathcal{M}$，都存在 $v^\star \in [V]$，使得：

1. $D_{\mathsf{TV}}\left(\mathbb{P}^{\widetilde{M}_{M,0},(\widetilde{p},\widetilde{q})}, \mathbb{P}^{\widetilde{M}_{M,v^\star},(\widetilde{p},\widetilde{q})}\right) \leq C\sqrt{T\log(T)/V}$.
2. $\mathbb{E}^{\widetilde{M}_{M,v^\star},(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}[\widehat{\pi}_2(v^\star)] \leq C\sqrt{T\log(T)/V}$。

引理 C.2 的证明如下所述。设 $\delta := C\sqrt{T\log(T)/V}$，其中 C 是引理 C.2 中的常数。如果 $\delta > 1$，则显然有 $\mathfrak{M}(\mathscr{H}, T) \le \mathfrak{M}(\mathscr{M}, T) + O(\delta)$，因此下文可假设 $\delta \le 1$。于是我们有：

$$\begin{aligned}
\mathbb{E}^{M,(p,q)}\mathbb{E}_{\widehat{\sigma}_1\sim p}[g^M(\widehat{\sigma}_1)] &= \mathbb{E}^{\widetilde{M}_{M,0},(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}\mathbb{E}_{\widehat{\sigma}_1\sim\widehat{\pi}_1}[g^M(\widehat{\sigma}_1)] \\
&\le \mathbb{E}^{\widetilde{M}_{M,v^\star},(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}\mathbb{E}_{\widehat{\sigma}_1\sim\widehat{\pi}_1}[g^M(\widehat{\sigma}_1)] + \delta \\
&\le \mathbb{E}^{\widetilde{M}_{M,v^\star},(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}\left[\min\left\{\frac{h^{\widetilde{M}_{M,v^\star}}(\widehat{\pi})}{1-\widehat{\pi}_2(v^\star)}, 1\right\}\right] + \delta \\
&\le (1+\sqrt{\delta})\cdot\mathbb{E}^{\widetilde{M}_{M,v^\star},(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}[h^{\widetilde{M}_{M,v^\star}}(\widehat{\pi})] + 2\sqrt{\delta} \\
&\le \mathbb{E}^{\widetilde{M}_{M,v^\star},(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}[h^{\widetilde{M}_{M,v^\star}}(\widehat{\pi})] + 6\sqrt{\delta},
\end{aligned}$$

其中，第一个不等式由引理 C.2的第一点推出，第二个不等式由引理 C.1推出；倒数第二个不等式结合引理 C.2 的第二点与马尔可夫不等式，得出 $\mathbb{P}^{\widetilde{M}_{M,v^\star},(\widetilde{p},\widetilde{q})}(\widehat{\pi}_2(v^\star) \ge \sqrt{\delta}) < \sqrt{\delta}$；最后一个不等式利用了 $h^{M'}(\pi) \le 4$ 对所有 $\widetilde{M}$ 成立，其中 $M' \in \mathcal{M}$，$\pi \in \Pi$。

对所有模型 $M \in \mathcal{M}$ 取上确界，可得

$$\sup_{M\in\mathcal{M}} \mathbb{E}^{M,(p,q)}\mathbb{E}_{\widehat{\sigma}_1\sim p}[g^M(\widehat{\sigma}_1)] \le \sup_{\widetilde{M}\in\widetilde{\mathcal{M}}} \mathbb{E}^{\widetilde{M},(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}[h^{\widetilde{M}}(\widehat{\pi})] + 6\sqrt{\delta}.$$

将 $\mathfrak{M}(\mathscr{M}, T)$ 界定为 $\mathfrak{M}(\mathscr{H}, T)$。 考虑任意一个实现 $\mathfrak{M}(\mathscr{H}, \ldots, T)$ 的算法 (p, q)。我们有 $p : \prod_{t=1}^{T}(\Pi\times\mathcal{O}) \to \Delta(\Pi)$，以及 $q = (q^1, \ldots, q^T)$，其中每个 $q^t : \prod_{i=1}^{t-1}(\Pi\times\mathcal{O}) \to \Delta(\Pi)$。

我们定义实例 $\mathscr{M}$ 上的算法 $(\widetilde{p}, \widetilde{q})$ 如下。给定 $\pi^t \in \mathcal{O}, o^t \in \mathcal{O}_\circ, r_1^t, r_2^t \in \mathcal{R}$，对于每个 $t \in [T]$，定义

$$\widetilde{p}(\cdot|(\pi^1,(o^1,r_1^1,r_2^1)),\ldots,(\pi^T,(o^T,r_1^T,r_2^T))) \in \Delta(\widetilde{\Pi})$$

为通过对每个 t 抽取 $(\sigma_1^t, \sigma_2^t) \sim \pi^t$，并取纯决策 $(\widehat{\sigma}, 0)$ 所得到的分布，其中 $\widehat{\sigma}$ 按照 $p(\cdot|(\sigma_1^1, o^1)\ldots,(\sigma_1^T, o^T))$ 分布。类似地，定义

$$\widetilde{q}^t(\cdot|(\pi^1,(o^1,r_1^1,r_2^1)),\ldots,(\pi^{t-1},(o^{t-1},r_1^{t-1},r_2^{t-1}))) \in \Delta(\widetilde{\Pi})$$

为通过对每个 $i < t$ 抽取 $(\sigma_1^i, \sigma_2^i) \sim \pi^i$，并取纯决策 $(\sigma_1^t, 0)$ 所得到的分布，其中 σ_1^t 按照 $q^t(\cdot|(\sigma_1^1, o^1)\ldots,(\sigma_1^{t-1}, o^{t-1}))$ 分布。由于每个 $\widetilde{q}^t$ 仅支撑于 $\Sigma_1 \times \{0\}$ 中的（纯）$\widetilde{}$ 决策，对于任意模型 $M \in \mathcal{M}$ 和任意 $v \in [V]$，令 $M' = M_{M,v}$，则在 $\mathbb{P}^{M',(\widetilde{p},\widetilde{q})}$ 条件下，$\{(\sigma_1^t, o^t)\}_{t=1}^T$ 的分布与在 $\mathbb{P}^{M,(p,q)}$ 条件下 $\{(\pi^t, o^t)\}_{t=1}^T$ 的分布相同。因此，我们有

$$\begin{aligned}
\mathbb{E}^{M,(p,q)}\mathbb{E}_{\widehat{\pi}\sim p}\left[g^M(\widehat{\pi})\right] =& \mathbb{E}^{M',(\widetilde{p},\widetilde{q})}\mathbb{E}_{(\widehat{\sigma}_1,0)\sim\widetilde{p}}[g^M(\widehat{\sigma}_1)] \\
=& \mathbb{E}^{M',(\widetilde{p},\widetilde{q})}\mathbb{E}_{(\widehat{\sigma}_1,0)\sim\widetilde{p}}\left[\sup_{\pi_2'\in\Pi_2'}\{f_2^{M'}(U_2(\pi_2',(\widehat{\sigma}_1,0))) - f_2^{M'}((\widehat{\sigma}_1,0))\}\right] \\
=& \mathbb{E}^{M',(\widetilde{p},\widetilde{q})}\mathbb{E}_{(\widehat{\sigma}_1,0)\sim\widetilde{p}}\left[h^{M'}((\widehat{\sigma}_1,0))\right],
\end{aligned}$$

其中，我们将 $p = p(\cdot|(\pi^1, o^1)\ldots,(\pi^T, o^T))$ 简写为表示在 $\mathbb{P}^{M,(p,q)}$ 下的随机变量，并将 $\widetilde{p} = \widetilde{p}(\cdot|(\pi^1,(o^1,r_1^1,r_2^1))\ldots,(\pi^T,(o^T,r_1^T,r_2^T)))$ 简写为表示在 $\mathbb{P}^{M',(\widetilde{p},\widetilde{q})}$ 下的随机变量。由此可得 $\mathfrak{M}(\mathscr{H}, T) \le \mathfrak{M}(\mathscr{M}, T)$ 】 【。

通过 $\mathsf{dec}_\varepsilon(\mathscr{H})$ 对 $\mathsf{dec}_{\varepsilon'}(\mathscr{M})$ 进行上界估计。任取参考模型 $M \in \mathsf{co}(\mathcal{M})$。给定 $\varepsilon > 0$，令 $\varepsilon' := \varepsilon + \sqrt{6/V}$。对于某个 $\widetilde{M} \in \mathsf{co}(\widetilde{\mathcal{M}})$，我们将通过 $\mathsf{dec}_{\varepsilon'}(\mathscr{M}, \widetilde{M})$ 对 $\mathsf{dec}_\varepsilon(\mathscr{H}, M)$ 进行上界估计。对于某个分布 $\nu \in \Delta(\mathcal{M})$，我们可以写作 $M(\pi) = \mathbb{E}_{M\sim\nu}[M(\pi)]$ 对所有 $\pi \in \Pi$ 均成立。定义 $\mu := \nu \times \mathsf{Unif}([V]) \in \Delta(\mathcal{M}\times[V])$，并令 $\sim$ $\sim$ $\sim M(\pi) := \mathbb{E}_{(M,v)\sim\mu}[M_{M,v}(\pi)]$ 对所有 $\pi \in \Pi$ 均成立。选择某个 $\widetilde{p}, \widetilde{q} \in \Delta(\Pi)$，使得

$$\mathsf{dec}_{\varepsilon'}(\mathscr{M}, \widetilde{M}) = \sup_{M'\in\widetilde{\mathcal{M}}}\left\{\mathbb{E}_{\pi\sim\widetilde{p}}[h^{M'}(\pi)] \mid \mathbb{E}_{\pi\sim\widetilde{q}}\left[D^2_{\mathsf{H}}\left(M'(\pi), \widetilde{M}(\pi)\right)\right] \le (\varepsilon')^2\right\}.$$

将 $p \in \Delta(\Pi)$ 定义为 σ_1 的分布，其中 $\pi = \pi_1 \times \pi_2 \sim \widetilde{p}$ 且 $\sigma_1 \sim \pi_{1\circ}$ 类似地，将 $q \in \Delta(\Pi)$ 定义为 σ_1 的分布，其中 $\pi = \pi_1 \times \pi_2 \sim \widetilde{q}$ 且 $\sigma_1 \sim \pi_{1\circ}$ 现在按如下方式选择 $v^\star \in [V]$：

$$v^\star := \mathop{\arg\min}_{v \in [V]} \left\{ \mathbb{E}_{\pi \sim \widetilde{p}}[\pi_2(v)] + \mathbb{E}_{\pi \sim \widetilde{q}}[\pi_2(v)] \right\},$$

其中我们使用了上述 $\pi = \pi_1 \times \pi_2$ 的约定。于是我们有

$$\mathbb{E}_{\pi \sim \widetilde{p}}[\pi_2(v^\star)] + \mathbb{E}_{\pi \sim \widetilde{q}}[\pi_2(v^\star)] \le \frac{2}{V}.$$

考虑任意模型 $M \in \mathcal{M}$，并令 $M' := M_{M,\, v^\star\circ}$ 现在我们计算

$$\begin{aligned}
\mathbb{E}_{\sigma_1 \sim p}[g^M(\sigma_1)] =& \mathbb{E}_{\pi \sim \widetilde{p}} \mathbb{E}_{\sigma_1 \sim \pi_1}[g^M(\sigma_1)] \\
\le& \mathbb{E}_{\pi \sim \widetilde{p}} \left[\min \left\{ \frac{h^{M'}(\pi) - \pi_2(\Sigma_2 \backslash \{0, v^\star\})}{\pi_2(\Sigma_2 \backslash \{v^\star\})},\ 1 \right\} \right] \\
\le& 2V^{-1/2} + \mathbb{E}_{\pi \sim \widetilde{p}} \left[\frac{h^{M'}(\pi)}{1 - V^{-1/2}} \right] \\
\le& 2V^{-1/2} + (1 + 2V^{-1/2}) \cdot \mathbb{E}_{\pi \sim \widetilde{p}}[h^{M'}(\pi)],
\end{aligned}$$

其中，第一个不等式使用了 引理 C.1，倒数第二个不等式则利用马尔可夫不等式得出 $\mathbb{P}_{\pi \sim \widetilde{p}}[\pi_2(v^\star) > V^{-1/2}] \le 2V^{-1/2}$。此外，我们有

$$\begin{aligned}
\mathbb{E}_{\pi \sim \widetilde{q}} \left[D^2_{\mathsf{H}} \Big(M'(\pi), \widetilde{M}(\pi) \Big) \right] \le& \mathbb{E}_{\pi \sim \widetilde{q}} \left[\mathbb{E}_{(\sigma_1, \sigma_2) \sim \pi} \left[D^2_{\mathsf{H}} \Big(M'((\sigma_1, \sigma_2)), \widetilde{M}((\sigma_1, \sigma_2)) \Big) \right] \right] \\
\le& \mathbb{E}_{\pi \sim \widetilde{q}} \left[\pi_2(0) \cdot \mathbb{E}_{\sigma_1 \sim \pi_1} \left[D^2_{\mathsf{H}} \Big(M'((\sigma_1, 0)), \widetilde{M}((\sigma_1, 0)) \Big) \right] \right] \\
&+ \mathbb{E}_{\pi \sim \widetilde{q}} \left[\pi_2(v^\star) \cdot \mathbb{E}_{\sigma_1 \sim \pi_1} \left[D^2_{\mathsf{H}} \Big(M'((\sigma_1, v^\star)), \widetilde{M}((\sigma_1, v^\star)) \Big) \right] \right] \\
&+ \mathbb{E}_{\pi \sim \widetilde{q}} \left[\pi_2(\Sigma \backslash \{0, v^\star\}) \cdot D^2_{\mathsf{H}}(\mathrm{Ber}(0), \mathrm{Ber}(1/V)) \right] \\
\le& \mathbb{E}_{\sigma_1 \sim q} \left[D^2_{\mathsf{H}} \Big(M'((\sigma_1, 0)), \widetilde{M}((\sigma_1, 0)) \Big) \right] + 2 \cdot \mathbb{E}_{\pi \sim \widetilde{q}}[\pi_2(v^\star)] + 2/V \\
\le& \mathbb{E}_{\sigma_1 \sim q} \left[D^2_{\mathsf{H}} \big(M(\pi), \overline{M}(\pi) \big) \right] + 6/V,
\end{aligned}$$

其中，第一个等式利用了平方海林格距离的凸性，第二个不等式利用了 $\widetilde{M}$ 是 $\widetilde{M}_{M,\, v}$ 与 $v \sim \mathrm{Unif}([V])$ 的混合分布这一事实，而第三个不等式利用了 $D^2_{\mathsf{H}}(\mathrm{Ber}(0),\ \mathrm{Ber}(1/V)) \le 2 \cdot D_{\mathsf{TV}}(\mathrm{Ber}(0),\ \mathrm{Ber}(1/V)) = 2/V$ 这一事实。因此可得

$$\begin{aligned}
\mathsf{dec}_\varepsilon(\mathscr{H}, \overline{M}) \le& \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\sigma_1 \sim p}[g^M(\pi)] \mid \mathbb{E}_{\sigma_1 \sim q} \left[D^2_{\mathsf{H}} \big(M(\sigma_1), \overline{M}(\sigma_1) \big) \right] \le \varepsilon^2 \right\} \\
\le& \sup_{M' \in \widetilde{\mathcal{M}}} \left\{ 2V^{-1/2} + (1 + 2V^{-1/2}) \cdot \mathbb{E}_{\pi \sim \widetilde{p}}[h^{M'}(\pi)] \mid \mathbb{E}_{\pi \sim \widetilde{q}} \left[D^2_{\mathsf{H}} \Big(M'(\pi), \widetilde{M}(\pi) \Big) \right] \le \varepsilon^2 + 6/V \right\} \\
\le& 2V^{-1/2} + (1 + 2V^{-1/2}) \cdot \mathsf{dec}_{\varepsilon + (6/V)^{-1/2}}(\mathscr{M}, \widetilde{M}) \\
\le& 6V^{-1/2} + \mathsf{dec}_{\varepsilon + (6/V)^{-1/2}}(\mathscr{M}, \widetilde{M}).
\end{aligned}$$

用 $\mathsf{dec}_\varepsilon(\mathscr{M})$ 对 $\mathsf{dec}_\varepsilon(\mathscr{H})$ 给出上界。接下来，考虑任意参考模型 $M \in \mathsf{co}(\mathcal{M})$。对于某个 $M \in \mathsf{co}(\mathcal{M})$，我们将用 $\mathsf{dec}_\varepsilon(\mathscr{H}, M)$ 对 $\mathsf{dec}_\varepsilon(\mathscr{M}, M)$ 给出上界。对于某个分布 $\mu \in \Delta(\mathcal{M} \times [V])$，有 $M(\pi) = \mathbb{E}_{(M,v) \sim \mu}[M_{M,v}(\pi)]$ 对所有 $\pi \in \Pi$ 均成立。定义 M，令 $\nu \in \Delta(\mathcal{M})$ 为 μ 在 $\mathcal{M}$ 上的边缘分布，接着：$M(\pi) := \mathbb{E}_{M \sim \nu}[M(\pi)]$ 对于每个 $\pi \in \Pi$。选择某个 $p, q \in \Delta(\Pi)$，使得

$$\mathsf{dec}_\varepsilon(\mathscr{H}, \overline{M}) = \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}[g^M(\pi)] \mid \mathbb{E}_{\pi \sim q}[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)] \le \varepsilon^2 \right\}.$$

令 $\widetilde{p} \in \Delta(\Pi)$ 为 $(\sigma_1, 0)$ 的分布，其中 $\sigma_1 \sim p$；令 $\widetilde{q} \in \Delta(\Pi)$ 为 $(\sigma_1, 0)$ 的分布，其中 $\sigma_1 \sim q$。根据模型 $M_{M,v}$ 的定义，对于任意 $M' = M_{M,v} \in \mathcal{M}$，我们有：

$$\begin{aligned}
\mathbb{E}_{\pi\sim\widetilde{p}}[h^{M'}(\pi)] =& \mathbb{E}_{\pi\sim\widetilde{p}}\left[\sup_{\pi_1'\in\Pi_1'}\{f_1^{M'}(U_1(\pi_1',\pi)) - f_1^{M'}(\pi)\} + \sup_{\pi_2'\in\Pi_2'}\{f_2^{M'}(U_2(\pi_2',\pi)) - f_2^{M'}(\pi)\}\right] \\
=& \mathbb{E}_{\pi\sim\widetilde{p}}\left[\sup_{\pi_2'\in\Pi_2'}\{f_2^{M'}(U_2(\pi_2',\pi))\}\right] \\
=& \mathbb{E}_{\sigma_1\sim p}[g^M(\sigma_1)].
\end{aligned}$$

此外，由于 $\widetilde{q}$ 完全支撑于 $\Sigma_1 \times \{0\}$，且在此类策略下，$\mathcal{M}$ 中的所有模型几乎必然满足 $r_1 = r_2 = 0$，因此有

$$\mathbb{E}_{\pi\sim\widetilde{q}}\left[D_{\mathsf{H}}^2\left(M'(\pi), \widetilde{M}(\pi)\right)\right] = \mathbb{E}_{\sigma_1\sim q}\left[D_{\mathsf{H}}^2\left(M(\pi), \overline{M}(\pi)\right)\right],$$

这证明了

$$\mathsf{dec}_\varepsilon(\mathscr{M}, \widetilde{M}) \le \sup_{M'\in\widetilde{\mathcal{M}}}\left\{\mathbb{E}_{\pi\sim\widetilde{p}}[h^{M'}(\pi)] \mid \mathbb{E}_{\pi\sim\widetilde{q}}\left[D_{\mathsf{H}}^2\left(M'(\pi), \widetilde{M}(\pi)\right)\right] \le \varepsilon^2\right\} \le \mathsf{dec}_\varepsilon(\mathscr{H}, \overline{M}).$$

对于每个 $\overline{\pi} \in \Pi$，令 $M(\pi)$ 表示当 $(o_\circ, r_1, r_2) \sim M_{M,v}((\pi, 0))$ 且 $(M, v) \sim \mu$ 时 $o_\circ$ 的分布。 □

引理 C.2 的证明。我们将从任意分布 $M_{M,v}$（对于 $v \ge 0$）中抽取的历史记为 $\{(\widetilde{\pi}^t, (\widetilde{o}^t, r_1^t, r_2^t))\}_{t=1}^T$。此外，我们将 $\widetilde{q}^t = \widetilde{q}^t(\cdot|(\widetilde{\pi}^1, (\widetilde{o}^1, r_1^t, r_2^t)) \ldots, (\widetilde{\pi}^{t-1}, (\widetilde{o}^{t-1}, r_1^{t-1}, r_2^{t-1})))$ 和 $\widetilde{p} = \widetilde{p}(\cdot|(\widetilde{\pi}^1, (\widetilde{o}^1, r_1^t, r_2^t)) \ldots, (\widetilde{\pi}^{T-1}, (\widetilde{o}^{T-1}, r_1^{T-1}, r_2^{T-1})))$ 简记为。定义 $M' := M_M$,0 并选取

$$v^\star := \arg\min_{v\in[V]}\left\{\sum_{t=1}^T \mathbb{E}^{M',(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widetilde{\pi}^t\sim\widetilde{q}^t}[\widetilde{\pi}_2^t(v)] + T\cdot\mathbb{E}^{M',(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}[\widehat{\pi}_2(v)]\right\}.$$

然后则 对 $v^\star$ 的选择，加上所有 $\pi \in \widetilde{\Pi}$ 均满足 $\sum_{v=1}^V \pi_2(v) \le 1$ 这一事实，确保了

$$\sum_{t=1}^T \mathbb{E}^{M',(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widetilde{\pi}^t\sim\widetilde{q}^t}[\widetilde{\pi}_2^t(v^\star)] + T\cdot\mathbb{E}^{M',(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}[\widehat{\pi}_2(v^\star)] \le \frac{2T}{V}. \tag{24}$$

写出 $M'' = \widetilde{M}_{M,v^\star}$。接下来，利用（Foster 等人，2021，引理 A.13），[22] 我们有：

$$\begin{aligned}
D_{\mathsf{H}}^2\left(\mathbb{P}^{M',(\widetilde{p},\widetilde{q})}, \mathbb{P}^{M'',(\widetilde{p},\widetilde{q})}\right) =& O(\log T)\cdot\mathbb{E}^{M',(\widetilde{p},\widetilde{q})}\left[\sum_{t=1}^T D_{\mathsf{H}}^2(M'(\widetilde{\pi}^t), M''(\widetilde{\pi}^t))\right] \\
\le& O(\log T)\cdot\mathbb{E}^{M',(\widetilde{p},\widetilde{q})}\left[\sum_{t=1}^T \mathbb{E}_{(\sigma_1^t,\sigma_2^t)\sim\widetilde{\pi}^t}[D_{\mathsf{H}}^2(M'((\sigma_1^t,\sigma_2^t)), M''((\sigma_1^t,\sigma_2^t)))]\right] \\
\le& O(\log T)\cdot\mathbb{E}^{M',(\widetilde{p},\widetilde{q})}\left[\sum_{t=1}^T \mathbb{E}_{(\sigma_1^t,\sigma_2^t)\sim\widetilde{\pi}^t}[2\cdot\mathbb{I}\{\sigma_2^t = v^\star\}]\right] \\
\le& O\left(\frac{T\log T}{V}\right),
\end{aligned}$$

其中，最后一个不等式用到了 (24)。由于全变差距离有海林格距离作为上界，因此可得 $D_{\mathsf{TV}}\left(\mathbb{P}^{M',(\widetilde{p},\widetilde{q})}, \mathbb{P}^{M'',(\widetilde{p},\widetilde{q})}\right) \le C\sqrt{T\log(T)/V}$ 其中某个常数 $C > 0$。结合这一事实与(24)，我们得到

$$\mathbb{E}^{M'',(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}[\widehat{\pi}_2(v^\star)] \le \mathbb{E}^{M',(\widetilde{p},\widetilde{q})}\mathbb{E}_{\widehat{\pi}\sim\widetilde{p}}[\widehat{\pi}_2(v^\star)] + C\sqrt{T\log(T)/V} \le 2T/V + C\sqrt{T\log(T)/V},$$

[22] 特别地，我们将该引理应用于序列 $X_1, \ldots, X_{2T}$。对于 t 的奇数值，有 $X_t = \widetilde{\pi}^t$，$X_{t+1} = (\widetilde{o}^t, r_1^t, r_2^t)$。并利用这样一个事实：在分布 $\mathbb{P}^{M',(\widetilde{p},\widetilde{q})}$ 和 $\mathbb{P}^{M'',(\widetilde{p},\widetilde{q})}$ 下，给定截至第 $t-1$ 步的历史，$\widetilde{\pi}^t$ 的条件分布相同，因为算法 $(\widetilde{p}, \widetilde{q})$ 相同。

其中，上述第二个不等式用到了 (24)。 □

D 第3 节的证明

D.1 第3.1 节中的证明

D.1.1 上界的更多细节

定理 3.1 中的上界是通过借助 Foster 等人（2023）的 E2D$^+$ PAC 算法推导出来的。下文将介绍该算法的一些背景，并给出一个更一般的上界。简而言之，E2D$^+$ PAC 算法的步骤如下：该算法使用一个在线估计预言机，记为 $\mathrm{Alg}_{\mathrm{Est}}$（在假设 D.1 中给出正式定义），它以模型类 $\mathcal{M}$ 为输入，并尝试根据在 $M^\star$ 下执行各种决策所获得的数据来估计真实模型 $M^\star \in \mathcal{M}$。为了在第 t 次迭代生成随后输入估计预言机 $\mathrm{Alg}_{\mathrm{Est}}$ 的下一个数据点，E2D$^+$ PAC 算法求解式 (7) 中的极小极大问题，以计算分布 p^t, q^t，其中模型 M 设为估计预言机在上一次迭代中的输出。然后，从 q^t 中采样一个决策 π^t，并观测由此产生的观测 $o^t \sim M^\star(\pi^t)$。接着将元组 (π^t, o^t) 输入估计预言机 $\widehat{\ }$，由其生成下一个估计 M^{t+1}。算法经过 T 次迭代后的输出，是从某个分布 $p^{t^\star}$ 中采样得到的，其中 $t^\star \sim [T]$ 服从均匀分布。更多背景请参见 Foster 等人（2023）。

假设 D.1 (针对 $\mathcal{M}$ 的估计预言机)。对于每个时刻 $t \in [T]$，一个面向函数类 $\mathcal{M}$ 的在线估计预言机 $\mathbf{Alg}_{\mathrm{Est}}$ 以 $\mathfrak{H}^{t-1} = (\pi^1, o^1) \ldots, (\pi^{t-1}, o^{t-1})$ 作为输入，其中 $o^i \sim M^\star(\pi^i)$ 且 $\pi^i \sim q^i$，适用于任意 $\widehat{\ }$（自适应）选择的分布 $q^i \in \Delta(\Pi)$。于是，对于某个函数类 $\mathcal{M} \subseteq \mathrm{co}(\mathcal{M})$，预言机 $\mathbf{Alg}_{\mathrm{Est}}$ $\widehat{\ }$ $\widehat{\ }$会返回一个估计量 $M^t \in \mathcal{M}$。我们假设，若 $M^\star \in \mathcal{M}$，该算法生成的估计量满足

$$\mathbf{Est}_{\mathsf{H}}(T) := \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim q^t}\left[D^2_{\mathsf{H}}\left(M^\star(\pi^t), \widehat{M}^t(\pi^t)\right)\right] \le \mathbf{Est}_{\mathsf{H}}(T, \delta),$$

其成立的概率至少为 $1-\delta$，其中$\mathbf{Est}_{\mathsf{H}}(T, \delta)$ 是一个已知的上界。

$\widehat{\ }$对于大多数估计预言机，$\mathcal{M}$ 假设 D.1 中的$\mathrm{co}(\mathcal{M})$将采用该类，不过在某些情况下可以选取更小的类（示例见命题 A.10）。

定理 *D.1*（*Foster* 等人（*2023*），定理 *3.1*；*HR-DMSO* 的上界）。固定 $\delta \in \left(0, \frac{1}{10}\right)$ 和 $T \in \mathbb{N}$，并考虑任意实例 $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$。假设模型类 $\mathcal{M}$ 满足假设 *1.4* 和 *D.1*，且某个类 $\widehat{\mathcal{M}} \subseteq \mathrm{co}(\mathcal{M})$ 也满足这些条件，并令 $\mathit{Est}_{\mathsf{H}} := \mathit{Est}_{\mathsf{H}}\left(\lceil 2T \log 2/\delta \rceil, 4\lceil \delta \log 2/\delta \rceil\right)$。令$\varepsilon(T) := 8\sqrt{\lceil \log \frac{2/\delta}{T} \rceil} \cdot \mathit{Est}_{\mathsf{H}}$，对于 *PAC*，使用预言机 $\mathit{Alg}_{\mathrm{Est}}$，可保证其概率至少为 $1-\delta$，

$$\mathbf{Risk}(T) \le \sup_{\overline{M} \in \widehat{\mathcal{M}}} \mathsf{dec}_{\varepsilon(T)}(\mathscr{H}, \overline{M}) \le \mathsf{dec}_{\varepsilon(T)}(\mathscr{H}).$$

若进一步对所有 $M \in \mathcal{M}$ 和某个 $R > 0$ 都有 $f^M(\cdot) \in [0, R]$，则期望风险有如下界：$\mathbb{E}[$**风险**$(T)] \le$解$^{\varepsilon(T)}(\mathscr{H})^- + R\delta$。

$\widehat{\ }$我们指出，定理 D.1 在Foster 等人(2023) 中仅针对 $\mathcal{M} = \mathrm{co}(\mathcal{M})$陈述，但检查$\widehat{\ }$证明可知，同样的保证也适用于任意一个 $\mathcal{M} \subseteq \mathrm{co}(\mathcal{M})$ 子类，其中**算法**$_{\mathrm{Est}}$会生成其预测（无需对证明作任何修改）。

定理 3.1 可由定理 D.1 推出，只需注意到，对于有限类，存在一个估计预言机，其满足$\mathbf{Est}_{\mathsf{H}}(T, \delta) \le 2\log(|\mathcal{M}|/\delta)$（Foster 等人，2023）。

注 D.1（MA-DMSO 的类似结论）。使用 定理 2.1 中的变换（该变换不会改变实例的模型类别，因此能够保留估计误差保证），也存在一个类似的结论

关于定理 D.1 在多智能体情形下的表述。具体而言，对于 MA-DMSO 的任意实例 $\mathscr{M}$，在假设 1.1 和D.1 下，存在一种算法，能够以概率 $1-\delta$ 确保**风险**$(T) \le \mathsf{dec}^{\varepsilon(T)}(\mathscr{M})$。

无限模型类。 正如附录 A 中的一些应用涉及无限模型类 $\mathcal{M}$，接下来我们将介绍一种简单的方法来界定此类模型的估计误差$\mathbf{Est}_{\mathsf{H}}(T,\ \delta)$，该方法遵循Foster 等人(2021)的思路。

定义 *D.1*（模型类覆盖；*Foster* 等人（*2021*），定义 *3.2*）。如果对于所有 $M \in \mathcal{M}$，都存在 $M' \in \mathcal{M}'$，使得 $\sup_{\pi\in\Pi} D^2_{\mathsf{H}}(M(\pi)\ M'(\pi)) \le \varepsilon^2$，则模型类 $\mathcal{M}' \subseteq \mathcal{M}$ 是一个ε-覆盖。令 $\mathcal{N}(\mathcal{M},\varepsilon)$ 表示此类最小覆盖 $\mathcal{M}'$的大小，并定义

$$\mathsf{est}(\mathcal{M},T) := \inf_{\varepsilon\ge 0}\left\{\log \mathcal{N}(\mathcal{M},\varepsilon) + \varepsilon^2 T\right\}.$$

我们将根据量 $\mathsf{est}(\mathcal{M},\ T)$ 来界定某个模型类的估计误差 $\mathbf{Est}_{\mathsf{H}}(T,\ \delta)$；为此，我们需要以下温和的假设。

假设 D.2。 假设存在一个核 ν，其从 $(\Pi,\mathscr{P})$ 映射到 $(\mathcal{O},\mathscr{O})$，使得 $M(\pi) \ll \nu(\pi)$ 对所有$M \in \mathcal{M}$、$\pi \in \Pi$ 均成立，并令 $m^M(\cdot|\pi)$ 表示 $M(\cdot|\pi)$ 关于 $\nu(\cdot|\pi)$的密度。此外，假设存在一个常数 $B \ge e$，使得

1. $\nu(\mathcal{O}|\pi) \le B$ 对所有 $\pi \in \Pi$ 均成立。*2.* $\sup_{\pi\in\Pi}$ $\sup_{o\in\mathcal{O}} m^M(o|\pi) \le B$ 对所有 $M \in \mathcal{M}$ 均成立。

命题 D.1 表明，估计误差 $\mathbf{Est}_{\mathsf{H}}(T,\delta)$ 随 对数 B增长。这个量通常较小：例如，对于标准的多臂老虎机问题（如伯努利老虎机和高斯老虎机），它是一个常数；而对于状态和动作空间有限的强化学习问题，它关于状态和动作空间的大小呈多重对数级增长。

命题 D.1 (Foster 等人 (2021）中的引理 A.16）。 假设 假设 *D.2* 成立。固定 $T \in \mathbb{N}, \delta \in (0, e^{-1})$，并记 $b_T = \log(2B^2T)$。那么存在一个算法 $\mathbf{Alg}_{\mathsf{Est}}$，它保证以概率$1-\delta$满足

$$\mathbf{Est}_{\mathsf{H}}(T) \le O(b_T \cdot \mathsf{est}(\mathcal{M},T) + b_T^2 \log(\delta^{-1})),$$

即，对于某个普适常数 C，我们可以取 $\mathbf{Est}_{\mathsf{H}}(T,\delta) = C\cdot(b_T\cdot\mathsf{est}(\mathcal{M},T) + b_T^2\log(\delta^{-1}))$。

D.1.2 定理 3.2 的证明

证明 定理 3.2。 固定 $T \in \mathbb{N}$ 以及一个算法 $(p,q) = \{q^t(\cdot|\cdot)\ p(\cdot|\cdot)\}_{t=1}^T$。对于每个模型 $M \in \mathcal{M}^+$，我们使用缩写 $\mathbb{P}^M \equiv \mathbb{P}^{M,(p,q)}$，并将相应的期望记为 $\mathbb{E}^M$。我们还定义

$$p_M = \mathbb{E}^M[p(\cdot|\mathfrak{H}^T)], \qquad q_M = \mathbb{E}^M\left[\frac{1}{T}\sum_{t=1}^T q^t(\cdot|\mathfrak{H}^{t-1})\right].$$

按照定理陈述选择 $\varepsilon(T)$，并记 $\varepsilon = \varepsilon(T)$。选择 $M \in \mathsf{co}(\mathcal{M})$，使得 $\mathsf{dec}_\varepsilon(\mathcal{M}) = \overline{\mathsf{dec}}_\varepsilon(\mathcal{M},M)$。[23] 我们将证明期望风险的下界与 $\mathsf{dec}_\varepsilon(\mathcal{M},M)$ 有关。定义

$$M := \arg\max_{M\in\mathcal{M}}\left\{\mathbb{E}_{\pi\sim p_{\overline{M}}}[g^M(\pi)] \mid \mathbb{E}_{\pi\sim q_{\overline{M}}}[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)] \le \varepsilon^2\right\},$$

其中我们回顾，$C(T) := \log(T \wedge V(\mathcal{M}))$。注意，若 $\mathcal{H}_{q_M,\varepsilon}(M) = \varnothing$，则根据定义$\mathsf{dec}_\varepsilon(\mathcal{M},M) = 0$，结论即得。因此，我们可以假设 $\mathcal{H}_{q_M,\varepsilon}(M) \ne \varnothing$，故上文对 M 的选取是良定义的。此外，M 的选取确保了

$$\mathbb{E}_{\pi\sim p_{\overline{M}}}[g^M(\pi)] \ge \mathsf{dec}_\varepsilon(\mathcal{M},\overline{M}) = \mathsf{dec}_\varepsilon(\mathcal{M}). \tag{25}$$

[23]如果 M 上的上确界无法达到，那么我们可以对一个达到该上确界的序列应用接下来的论证。

根据 Foster 等人 (2021) 的引理 A.13，我们有[24]

$$D^2_{\mathsf{H}}\big(\mathbb{P}^M, \mathbb{P}^{\overline{M}}\big) \le C(T)\cdot T\cdot \mathbb{E}_{\pi\sim q_{\overline{M}}}[D^2_{\mathsf{H}}\big(M(\pi), \overline{M}(\pi)\big)] \le C(T)\cdot T\cdot \varepsilon^2.$$

根据数据处理不等式，可得

$$D^2_{\mathsf{H}}(p_M, p_{\overline{M}}) \le C(T)\cdot T\cdot \varepsilon^2 \le \frac{1}{8R}\cdot \mathsf{dec}_\varepsilon(\mathcal{M}), \tag{26}$$

其中，第二个不等式由 $\varepsilon = \varepsilon(T)$ 的选择得出。

接下来，利用Foster 等人（2021）中的引理 A.11，以及 $g^M(\pi) \in [0, R]$ 对所有 π成立这一事实，可得

$$\mathbb{E}_{\pi\sim p_{\overline{M}}}[g^M(\pi)] \le 3\cdot \mathbb{E}_{\pi\sim p_M}[g^M(\pi)] + 4R\cdot D^2_{\mathsf{H}}(p_M, p_{\overline{M}}).$$

结合上述表达式与（25）和（26），并重新整理，可得

$$\frac{1}{6}\cdot \mathsf{dec}_\varepsilon(\mathcal{M}) \le \mathbb{E}_{\pi\sim p_M}[g^M(\pi)] = \mathbb{E}^M\mathbb{E}_{\pi\sim p(\cdot|\mathfrak{H}^T)}[g^M(\pi)] = \mathbb{E}^M[\mathbf{Risk}(T)],$$

由此得到期望风险的所需下界。

□

D.2 第 3.2 节中的证明

D.2.1 命题 3.1 的证明

证明命题 3.1。定义 $\Delta := \mathsf{dec}_{\varepsilon(T)}(\mathcal{M})\, 8\cdot\varepsilon(T)^2\cdot C(T)\cdot T$。若 $\Delta \ge 1$，则有 $\varepsilon(T) \ge \varepsilon(T)$，因此从现在起可以假设 $\Delta < 1$。取

$$\alpha = \left\lceil \frac{\log 1/\Delta}{2\log(C_{\mathrm{reg}}/c_{\mathrm{reg}})} \right\rceil \ge 1,$$

其中，特别地，它是使得$(C^2_{\mathrm{reg}}/c^2_{\mathrm{reg}})^\alpha \ge 1/\Delta$成立的最小正整数。$\alpha$ 根据我们关于 $C_{\mathrm{reg}} > c_{\mathrm{reg}}$ 的假设以及 $1/\Delta > 1$，是良定义的。应用假设 3.1 于 $\varepsilon = \varepsilon(T)\cdot(c_{\mathrm{reg}}/C_{\mathrm{reg}})^j$ 当$0 \le j < \alpha$时，可得

$$\begin{aligned}
\mathsf{dec}_{\bar\varepsilon(T)}(\mathcal{M}) &\le c^{2\alpha}_{\mathrm{reg}}\cdot \mathsf{dec}_{\bar\varepsilon(T)/C^\alpha_{\mathrm{reg}}}(\mathcal{M}) \\
&\le \Delta\cdot C^{2\alpha}_{\mathrm{reg}}\cdot \mathsf{dec}_{\bar\varepsilon(T)/C^\alpha_{\mathrm{reg}}}(\mathcal{M}) \\
&\le \mathsf{dec}_{\bar\varepsilon(T)}(\mathcal{M})\cdot \frac{\mathsf{dec}_{\bar\varepsilon(T)/C^\alpha_{\mathrm{reg}}}(\mathcal{M})}{8\cdot(\bar\varepsilon(T)/C^\alpha_{\mathrm{reg}})^2\cdot C(T)\cdot T}.
\end{aligned}$$

因此 $\varepsilon(T) \ge \varepsilon(T)/C^\alpha_{\mathrm{reg}}$，于是

$$\mathsf{dec}_{\bar\varepsilon(T)}(\mathcal{M}) \ge \mathsf{dec}_{\bar\varepsilon(T)/C^\alpha_{\mathrm{reg}}}(\mathcal{M}) \ge \frac{1}{c^{2\alpha}_{\mathrm{reg}}}\cdot \mathsf{dec}_{\bar\varepsilon(T)}(\mathcal{M}).$$

α 的定义和 $\varepsilon(T)$可得

$$\left(\frac{C_{\mathrm{reg}}}{c_{\mathrm{reg}}}\right)^{2\alpha} \le \frac{C_{\mathrm{reg}}}{c_{\mathrm{reg}}}\cdot\frac{1}{\Delta} \le \frac{C_{\mathrm{reg}}}{c_{\mathrm{reg}}}\cdot\frac{8^3\cdot\lceil\log 2/\delta\rceil\cdot\overline{\mathbf{Est}}_{\mathsf{H}}\cdot C(T)}{\mathsf{dec}_{\bar\varepsilon(T)}(\mathcal{M})}.$$

我们的 β 定义保证了 $c^{2\alpha}_{\mathrm{reg}} \le ((C_{\mathrm{reg}}/c_{\mathrm{reg}})^{2\alpha})^\beta$，这意味着，对于某个常数 C，

$$\mathrm{DEC}_{\varepsilon(T)}(\mathcal{M}) \le c^{2\alpha}_{\mathrm{reg}}\cdot \mathrm{DEC}_{\bar\varepsilon(T)}(\mathcal{M}) \le (C\cdot C_{\mathrm{reg}}/c_{\mathrm{reg}})^\beta\cdot\log^\beta 1/\delta\cdot\overline{\mathbf{Est}}^\beta_{\mathsf{H}}\cdot C(T)^\beta\cdot \mathrm{DEC}_{\varepsilon(T)}(\mathcal{M})^{-\beta}\cdot \mathrm{DEC}_{\bar\varepsilon(T)}(\mathcal{M}),$$

[24]为了应用这一结果，我们需要确保对于所有可测集 $\mathcal{E}\subseteq\mathcal{O}$ 以及所有 $\pi\in\Pi$，都有 $M(\mathcal{E}|\pi)\, M(\mathcal{E}|\pi) \le V(\mathcal{M})$。根据 $V(\mathcal{M})$ 在 (13) 中的定义以及 $M\in\mathrm{co}(\mathcal{M})$这一事实，即可得出该结论。

重新整理可得：

$$\mathsf{dec}_{\bar{\varepsilon}(T)}(\mathcal{M}) \le \left(C \log 1/\delta \cdot \overline{\mathbf{Est}}_{\mathsf{H}} \cdot C(T) \cdot C_{\mathrm{reg}}/c_{\mathrm{reg}}\right)^{\frac{\beta}{1+\beta}} \cdot \mathsf{dec}_{\underline{\varepsilon}(T)}(\mathcal{M})^{\frac{1}{1+\beta}}.$$

□

D.2.2 命题 3.2 和命题 3.3 的证明

命题 3.2 的证明。令 $\Pi = [A]$ 和 $\mathcal{O} = \{0,1\}$。对于每个 $\delta > 0$，我们定义模型类$\mathcal{M}_\delta \subset (\Pi \to \Delta(\mathcal{O}))$如下：$\mathcal{M}_\delta = \{M_{\delta,a} : a \in [A]\}$，并定义 $M_{\delta,a}(\pi) =$ 伯努利分布$(1/2 + \delta \mathbb{I}\{\pi = a\})$。现在令 $\mathcal{M} = \bigcup_{i=2}^{L} \mathcal{M}_{2^{-i}}$，由此可得 $|\mathcal{M}| \le LA$。定义 $f^M(\pi) = \mathbb{E}^{M,\pi}[r]$，其中 $r \sim M(\pi)$。最后令 $\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$。注意，实例 $\mathscr{H}$实际上是 Foster 等人（2021）意义下的标准（非隐藏奖励）DMSO 实例。

由于模型类 $\mathcal{M}$ 是所有 A 臂赌博机问题类的子类，根据 Foster 等人(2021) 的命题 5.1 以及 命题 4.1 (除了 MA-DMSO 实例外，该命题同样适用于 HR-DMSO 实例)，我们有 $\mathsf{dec}^{\varepsilon}(\mathscr{H}) \le O(\varepsilon\sqrt{A})$。此外，我们有 $\mathfrak{M}(\mathscr{H}, T) \le O(\sqrt{TA})$ (Audibert 和 Bubeck, 2009)（忽略对数因子后，该界也可由例如 定理 3.1 得出）。

对于每个 $\delta > 0$，写作 $\mathscr{H}_\delta := (\mathcal{M}_\delta, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$。另记 $M_\delta = \frac{1}{A}\sum_{a=1}^{A} M_{\delta,a} \in \mathrm{co}(\mathcal{M}_\delta)$。由于对于所有 $\pi \in \Pi$ 和 $a \in [A]$，$D^2_{\mathsf{H}}\left(M_\delta(\pi) \| M_{\delta,a}(\pi)\right) \le 4\delta^2$，显然 $\mathsf{dec}^{4\delta/\sqrt{A}}(\mathscr{H}_\delta) \ge \mathsf{dec}^{4\delta/\sqrt{A}}(\mathscr{H}_\delta, M_\delta) \ge \Omega(\delta)$。由于增大模型类的规模不会降低 DEC，因此对于所有满足 $1/\sqrt{A} > \varepsilon > 2^{-L}$ 的 ε，都有 $\mathsf{dec}^{\varepsilon}(\mathscr{H}) \ge \Omega(\varepsilon\sqrt{A})$。最后，由于奖励是在实例 $\mathscr{H}$ 中观测到的，我们可以使用 Foster 等人（2023）的定理 2.1 得出，对于至少为某个充分大的常数的 A，以及 $T \le 2^{L/2}$, $\mathfrak{M}(\mathscr{H}, T) \ge \Omega(\sqrt{A/T})$。

□

证明命题 3.3。令 L 大于某个普适常数（其取值将在下文给出），并考虑常数 $C_{\mathrm{prob}} \ge 1$的任意取值。我们定义如下实例$\mathscr{H} = (\mathcal{M}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_{M \in \mathcal{M}})$，其各个组成部分定义如下：

- 对于 $1 \le \ell \le L$，定义 $\alpha_\ell := 1/L$, $N_\ell = 2^\ell$，以及 $\delta_\ell = \frac{1}{(C_{\text{概率}} N_\ell)^2}$。
- 令 $\mathcal{V} := \prod_{\ell=1}^{L} [N_\ell]$，并集合 $\Pi := \mathcal{V}$。
- 令 $\mathcal{O} = \prod_{\ell=1}^{L} ([N_\ell] \cup \{\perp_\ell\})$。为便于记号，我们记作 $\mathcal{O}_\ell := [N_\ell] \cup \{\perp_\ell\}$。
- 对于 $o_\ell \in \mathcal{O}_\ell$ 和 $v_\ell \in [N_\ell]$，定义

$$P_{v_\ell}(o_\ell) := \begin{cases} 1 - \delta(N_\ell - 1) & : o_\ell = \perp_\ell \\ \delta_\ell & : o_\ell \in [N_\ell] \backslash \{v_\ell\} \\ 0 & : o_\ell = v_\ell. \end{cases}$$

  那么 $P_{v_\ell} \in \Delta(\mathcal{O}_\ell)$。

- 类 $\mathcal{M}$ 由元组 $v \in \mathcal{V}$索引；特别地，对于每个 $v = (v_1, \ldots, v_\ell) \in \mathcal{V}$，都有一个模型$M_v$，定义如下。对于 $\pi \in \Pi$, $M_v(\pi) \in \Delta(\mathcal{O})$ 是如下不依赖于 π的分布：对于 $o = (o_1, \ldots, o_L) \in \mathcal{O}$,

$$M_v(\pi)(o) = \prod_{\ell=1}^{L} P_{v_\ell}(o_\ell).$$

  由于分布 $M_v(\pi)$ 不依赖于 π，我们通常会省略参数 π，直接写作 $M_v \in \Delta(\mathcal{O})$。相应地，两个模型 $M, M' \in \mathrm{co}(\mathcal{M})$ 的观测分布之间的海林格距离记为 $D^2_{\mathsf{H}}(M, M')$。

- 对于所有 $v \in \mathcal{V}$ 和 $\pi = (\pi_1, \dots, \pi_L) \in \Pi$，价值函数 $f\colon \Pi \to [0,1]$ 定义如下：

$$f^{M_v}(\pi) := \sum_{\ell=1}^{L} \alpha_\ell \cdot (1 - \mathbb{I}\{\pi_\ell = v_\ell\}).$$

为方便起见，我们记 $f_\ell^{M_v}(\pi) := (1 - \mathbb{I}\{\pi_\ell = v_\ell\})$，于是 $f^{M_v}(\pi) = \sum_{\ell=1}^{L} \alpha_\ell \cdot f_\ell^{M_v}(\pi)$。显然，对于所有 $v \in \mathcal{V}$，都存在某个 π（即任意一个满足对所有 ℓ 均有 $\pi_\ell \neq v_\ell$ 的 π），使得 $f^{M_v}(\pi) = 1$，也就是说，$g^{M_v}(\pi) = 1 - f^{M_v}(\pi)$。

极小极大样本复杂度的上界。固定某个 $T \in \mathbb{N}$；接下来我们给出 $\mathfrak{M}(\mathscr{H}, T)$ 的上界。由于该类中所有模型 $\mathcal{M}$ 对观测的分布不依赖于决策，因此，要指定一个算法（p，q），我们只需指定分布 p；它是从由 T 个观测组成的元组到决策分布的映射。为定义 p，我们先定义映射 $p_\ell\colon \mathcal{O}_\ell^T \to [N_\ell]$，如下：

$$p_\ell(o_{\ell,1}, \dots, o_{\ell,T}) := \begin{cases} \mathrm{Unif}(\Pi) & : o_{\ell,1} = \cdots = o_{\ell,T} = \perp_\ell, \\ \mathbb{I}_{o_{\ell,t}} & : t := \arg\min\{s \in [T] \mid o_{\ell,s} \neq \perp_\ell\} \text{ exists.} \end{cases}$$

具体而言，p_ℓ 输出第一个不为 $\perp_\ell$的观测索引；如果不存在这样的索引，则 p_ℓ输出 $[N_\ell]$上的均匀分布。现在我们定义

$$p(o_1, \dots, o_T) := (p_\ell(o_{\ell,1}, \dots, o_{\ell,T}))_{\ell=1}^{L},$$

其中，对于每个 $T \in [T]$，我们将 $o_t = (o_{1,\,t}, \dots\dots, o_{L,\,t})$记作。

现在我们给算法 p的风险取一个上界。对于给定模型 $M_v \in \mathcal{M}$ 下历史记录的分布，我们简记为 $\mathbb{P}^{M_v}(\cdot)$，并将相应的期望记为 $\mathbb{E}^{M_v}[\cdot]$。对于每个 $\ell \in [L]$，对所有 $M_v \in \mathcal{M}$，都有：

$$\mathbb{E}^{M_v}\mathbb{E}_{\pi \sim p(o_1,\dots,o_T)}\left[1 - f_\ell^{M_v}(\pi)\right] \leq (1 - \delta_\ell(N_\ell - 1))^T \cdot \frac{1}{N_\ell},$$

因为不存在满足$o_{\ell,\,t} \neq \perp_\ell$的 $t \in [T]$ 的概率为$(1 - \delta_\ell(N_\ell - 1))^T$，而在该事件的补集上（即存在这样的 t 时），$p_\ell(o_1, \dots\dots, o_T)$将其全部概率质量赋予这样的$o_{\ell,\,t} \neq v_\ell$，因此 $f_\ell^{M_v}(\pi) = 1$。因此

$$\mathbb{E}^{M_v}\mathbb{E}_{\pi \sim p(o_1,\dots,o_T)}[g^{M_v}(\pi)] \leq \sum_{\ell=1}^{L} \frac{\alpha_\ell}{N_\ell} \cdot (1 - \delta_\ell(N_\ell - 1))^T \leq \sum_{\ell=1}^{L} \frac{\alpha_\ell}{N_\ell} \cdot \left(1 - \frac{1}{2C_{\mathrm{prob}}^2 N_\ell}\right)^T.$$

给定 $T \leq N_L$，选择尽可能大的 $\ell_\star = \ell_\star(T) \in [L]$，使得 $T \geq 2\log(N_{\ell_\star}) \cdot 2C_{\mathrm{prob}}^2 N_{\ell_\star}$，从而得到

$$\begin{aligned} \mathbb{E}^{M_v}\mathbb{E}_{\pi \sim p(o_1,\dots,o_T)}[g^{M_v}(\pi)] &\leq \sum_{\ell=1}^{\ell_\star} \frac{\alpha_\ell}{N_\ell} \cdot \exp\left(-\frac{2\log N_{\ell_\star}}{T}\right)^T + \sum_{\ell=\ell_\star+1}^{L} \frac{\alpha_\ell}{N_\ell} \\ &\leq \sum_{\ell=1}^{\ell_\star} \frac{\alpha_\ell}{N_\ell} \cdot \frac{1}{N_{\ell_\star}^2} + \sum_{\ell=\ell_\star+1}^{L} \frac{\alpha_\ell}{N_\ell} \leq \frac{1}{N_{\ell_\star}} \leq \frac{8C_{\mathrm{prob}}^2 \log T}{T}, \end{aligned}$$

其中，最后一个不等式利用了我们对 $\ell_\star$ 的选择可得 $N_{\ell_\star} \geq \frac{T}{8C_{\mathrm{prob}}^2 \text{对数} T}$，

DEC 的下界。根据平方海林格距离的张量化性质，对于任意两个模型 M_v，$M_u \in \mathcal{M}$，我们有

$$1 - \frac{1}{2}D_{\mathsf{H}}^2(M_u, M_v) = \prod_{\ell=1}^{L}\left(1 - \frac{1}{2}D_{\mathsf{H}}^2(P_{v_\ell}, P_{u_\ell})\right) = \prod_{\ell=1}^{L}\left(1 - \frac{1}{2} \cdot \mathbb{I}\{u_\ell \neq v_\ell\} \cdot 2\delta_\ell\right) \geq 1 - \sum_{\ell=1}^{L} \delta_\ell \cdot \mathbb{I}\{u_\ell \neq v_\ell\},$$

这意味着 该 $D_{\mathsf{H}}^2(M_u, M_v) \leq 2\sum_{\ell=1}^{L} \delta_\ell \cdot \mathbb{I}\{u_\ell \neq v_\ell\}$。令 $v := (1, 1, \dots, 1) \in \mathcal{V}$，并集合 $\bar{\mathcal{M}} := M_{\bar{v}}$.

现在考虑任意 $2 \geq \varepsilon \geq \sqrt{2\delta_L}$，将 $\ell^\star = \ell^\star(\varepsilon)$ 选择为 $\ell \in [L]$ 的最小可能取值，使得 $\varepsilon^2 \geq 2\delta_\ell$。对于每个 $i \in [N_{\ell^\star}]$，定义 $v^i \in \mathcal{V}$ 如下：

$$\bar{v}^i_\ell = \begin{cases} \bar{v}_\ell & : \ell \neq \ell^\star \\ i & : \ell = \ell^\star, \end{cases}$$

并记为 $M^i := M_{v^i}$。那么对于所有 $i \in [N_{\ell^\star}]$，我们有 $D^2_{\mathsf{H}}\big(M, M^i\big) \leq 2\delta_{\ell^\star} \leq \varepsilon^2$。对于任意分布$p \in \Delta(\Pi)$，必定存在某个 $i^\star \in [N_{\ell^\star}]$，使得

$$\mathbb{E}_{\pi\sim p}[1 - f^{M^{i^\star}}_{\ell^\star}(\pi)] = \mathbb{P}_{\pi\sim p}(\pi_{\ell^\star} = i^\star) \geq 1/N_{\ell^\star}.$$

因此，

$$\mathsf{dec}_\varepsilon(\mathcal{M}, \overline{M}) \geq \frac{\alpha_{\ell^\star}}{N_{\ell^\star}} = \alpha_{\ell^\star} C_{\mathrm{prob}} \sqrt{\delta_{\ell^\star}} \geq \frac{\alpha_{\ell^\star} C_{\mathrm{prob}}}{\sqrt{8}} \cdot \varepsilon = \frac{C_{\mathrm{prob}}}{\sqrt{8} \cdot L} \cdot \varepsilon, \tag{27}$$

其中，最后一个不等式利用了 $\varepsilon^2 \leq 8\delta_\ell$，因为对于所有 $\ell < L$ 和 $\varepsilon \leq 2$，都有 $\delta_{\ell+1} = \delta_\ell/4$。

DEC 上界。 接下来我们给出 $\mathrm{DEC}_\varepsilon(\mathcal{M})$ 在 $\varepsilon \in (0,2)$上的上界；虽然求 $\varepsilon(T)$的下界并不需要这一点，但给出 $\mathrm{DEC}_\varepsilon(\mathcal{M})$ 的上界，可以确保类 $\mathcal{M}$ 满足 假设 3.1 中的正则性条件。这样可以证明我们构造的实例 $\mathscr{H}$ 满足我们在根据 DEC 给出极小极大风险的上界和下界时所用的假设。

考虑任意的 $\overline{M} \in \mathrm{co}(\mathcal{M})$。我们可以将其写为 $M = \mathbb{E}_{v\sim\mu}[M_v]$，其中分布为 $\mu \in \Delta(\mathcal{V})$。对于每个 $\ell \in [L]$，令$\mu_\ell \in \Delta([N_\ell])$ 表示 μ 在 $[N_\ell]$ 上的边缘分布（回忆一下 $\mathcal{V} = \prod_{\ell=1}^L [N_\ell]$）。由于 $D^2_{\mathsf{H}}(P_{u_\ell}, P_{v_\ell}) = 2\delta_\ell$ 对于 $u_\ell \neq v_\ell$，任何两个不同的取值 $v_\ell, v'_\ell \in [N_\ell]$，只要满足 $D^2_{\mathsf{H}}(\mathbb{E}_{u_\ell\sim\mu_\ell}[P_{u_\ell}], P_{v_\ell}) \leq \varepsilon^2$ 和 $D^2_{\mathsf{H}}\Big(\mathbb{E}_{u_\ell\sim\mu_\ell}[P_{u_\ell}], P_{v'_\ell}\Big) \leq \varepsilon^2$，就必须进一步满足

$$2\delta_\ell = D^2_{\mathsf{H}}\Big(P_{v_\ell}, P_{v'_\ell}\Big) \leq 2 \cdot D^2_{\mathsf{H}}(\mathbb{E}_{u_\ell\sim\mu_\ell}[P_{u_\ell}], P_{v_\ell}) + 2 \cdot D^2_{\mathsf{H}}\Big(\mathbb{E}_{u_\ell\sim\mu_\ell}[P_{u_\ell}], P_{v'_\ell}\Big) \leq 4\varepsilon^2. \tag{28}$$

现在考虑任意一个 $\varepsilon \in (\sqrt{\delta_L}, 2)$。定义 $\ell = \ell(\varepsilon)$ 为 $\ell \in [L]$ 的最大可能取值，使得 $\varepsilon^2 < \delta_\ell/2$。由(28) 可知，对于所有 $\ell \leq \ell$，至多存在一个 $v_\ell \in [N_\ell]$，使得 $D^2_{\mathsf{H}}(\mathbb{E}_{u_\ell\sim\mu_\ell}[P_{u_\ell}], P_{v_\ell}) \leq \varepsilon^2$。若存在这样的 v_ℓ，将其记为 v_ℓ；若不存在，则任意选取 $v_\ell \in [N_\ell]$。

根据数据处理不等式，对于任意 $v \in \mathcal{V}$ 以及每个 $\ell \in [L]$，都有 $D^2_{\mathsf{H}}(M, M_v) \geq D^2_{\mathsf{H}}(\mathbb{E}_{u_\ell\sim\mu_\ell}[P_{u_\ell}], P_{v_\ell})$。因此，对于每个满足 $D^2_{\mathsf{H}}(M, M_v) \leq \varepsilon^2$ 的 $M_v \in \mathcal{M}$，我们必须有：对于所有 $\ell \leq \ell$，$v_\ell = v_\ell$。

现在任取 $v^\star \in \mathcal{V}$，使得对所有 $\ell \leq \ell$都有 $v^\star_\ell \neq v_\ell$，并将 $p^\star \in \Delta(\Pi)$ 定义如下：

$$p^\star := \mathrm{Unif}\left(\left\{v \in \mathcal{V} \mid v_\ell = v^\star_\ell \ \forall \ell \leq \bar{\ell}\right\}\right).$$

现在我们可以计算：

$$\mathsf{dec}_\varepsilon(\mathcal{M}, \overline{M}) \leq \sup_{M\in\mathcal{M}} \left\{\mathbb{E}_{\pi\sim p^\star}[g^M(\pi)] \mid D^2_{\mathsf{H}}\big(\overline{M}, M\big) \leq \varepsilon^2\right\} \leq \sum_{\ell=\bar{\ell}+1}^{L} \frac{\alpha_\ell}{N_\ell} \leq \frac{1}{N_{\bar{\ell}+1}} \leq 2C_{\mathrm{prob}}\varepsilon,$$

wh这里，最后一个不等式利用了 ℓ 的定义以及 $\varepsilon \geq \sqrt{\delta_L}$ 这一事实，因此有 $\varepsilon^2 \geq \delta_\ell/8$。

Bo束 $\varepsilon(T)$。任取 $T \leq N_L/L^3$，这便确保（当 L 充分大时）

$$\frac{C_{\mathrm{prob}}}{8\sqrt{8} \cdot L \cdot C(T) \cdot T} \geq \frac{\sqrt{2}}{C_{\mathrm{prob}} \cdot N_L} = \sqrt{2\delta_L}, \tag{29}$$

这里我们利用了：对于某个普适常数 C_0，有 $C(T) \leq C_0 \cdot \log(T)$。回顾一下，$\varepsilon(T)$ 的定义是：尽可能取大，使得 $\varepsilon(T)^2 \cdot C(T) \cdot T \leq \frac{1}{8}\mathsf{dec}_{\varepsilon(T)}(\mathcal{M})$。令

$$\varepsilon_0 := \frac{C_{\mathrm{prob}}}{8\sqrt{8} \cdot L \cdot C(T) \cdot T} \geq \Omega\left(\frac{C_{\mathrm{prob}}}{T \log(T) \cdot L}\right),$$

利用(29)和(27)，可得 $\varepsilon_0^2 \cdot C(T) \cdot T \leq \frac{1}{8} \cdot$ 衰减$_{\varepsilon_0}(\mathcal{M})$，因此 $\varepsilon(T) \geq \varepsilon_0$。 □

D.2.3 定理3.3的证明

定理 3.3 的证明。给定任意 $C_{\mathrm{prob}} \geq 1$，固定 $N = \lceil \sqrt{T/C_{\mathrm{prob}}} \rceil$。对于实数 $\delta, \beta \in (0, 1)$，我们将定义 HR-DMSO 框架的实例 $\mathscr{H}^{\delta,\beta}$。稍后，我们将针对某些 δ, β 的取值，选择 $\mathscr{H}_1, \mathscr{H}_2$ 作为这样的实例。对于某些模型类 $\mathcal{M}^{\delta,\beta}$，每个模型类的大小为 N，我们将证明，对于所有 δ, β，$\mathscr{H}^{\delta,\beta} = (\mathcal{M}^{\delta,\beta}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$，也就是说，实例 $\mathscr{H}^{\delta,\beta}$ 具有相同的决策空间、观测空间和价值函数。接下来定义这些组成部分：

- $\Pi = [N]$ 和 $\mathcal{O} = [N] \cup \{\perp\}$。
- 对于所有 δ，β，我们有 $\mathcal{M}^{\delta,\beta} = \{M_1^{\delta,\beta}, \ldots, M_N^{\delta,\beta}\}$。对于 $i \in [N]$ 和 $\pi \in \Pi$，$M^{\delta,\beta}_i(\pi) \in \Delta(\mathcal{O})$ 是如下分布，它不依赖于 π：

$$M_i^{\delta,\beta}(\pi)(j) = \begin{cases} 1 - \delta(N-1) - \beta & : j = \perp \\ \delta & : j \in [N] \backslash \{i\} \\ \beta & : j = i. \end{cases}$$

  由于分布 $M^{\delta,\beta}_i(\pi)$ 不依赖于 π，我们通常省略参数 π，直接写作 $M^{\delta,\beta}_i \in \Delta(\mathcal{O})$。

- 对于所有 δ，β 以及所有 $\pi \in \Pi$、$i \in [N]$，价值函数 $f^{M^{\delta,\beta}_i}: \Pi \to [0, 1]$ 定义如下：

$$f^{M_i^{\delta,\beta}}(\pi) := 1 - \mathbb{I}\{i = \pi\}.$$

  由于上述价值函数不依赖于 δ，β，我们将直接写作 $f^i(\pi) := f^{M^{\delta,\beta}_i}(\pi)$ 和 $g^i(\pi) := g^{M^{\delta,\beta}_i}(\pi) = \mathbb{I}\{i = \pi\}$。

在模型 $M^{\delta,\beta}_i$ 中，除决策 i 外，所有决策都是最优的。此外，我们总有 $\beta < \delta$，这意味着，在 $M^{\delta,\beta}_i$ 下，观测到 $[N]\backslash\{i\}$ 中任意给定索引的可能性高于观测到 i 的可能性。

对极小极大风险进行上界估计。接下来，对于 $T \in \mathbb{N}$，我们给出 $\mathfrak{M}(\mathscr{H}^{\delta,\beta}, T)$ 的上界。由于类别 $\mathcal{M}^{\delta,\beta}$ 中所有模型的观测分布不依赖于决策，要指定一个算法 (p, q)，我们只需指定分布 p，它是从 T 元组的观测到决策分布的映射。此外，为了指定给定模型 $M^{\delta,\beta}_i$ 下的历史记录分布，我们记为 $\mathbb{E}^{M^{\delta,\beta}_i}[\cdot]$。现在考虑由 p 定义的算法：

$$p(o_1, \ldots, o_T) := \begin{cases} \mathrm{Unif}(\Pi) & : o_1 = \cdots = o_T = \perp \\ \mathbb{I}_{o_t} & : t := \arg\min\{s \in [T] \mid o_s \neq \perp\}, \ \text{exists}. \end{cases} \tag{30}$$

特别地，p 输出第一个不为 $\perp$ 的观测的索引；如果不存在这样的索引，则 p 输出决策上的均匀分布。为了给出 p 的期望风险上界，注意到，对于任意模型 $M^{\delta,\beta}_i$，我们有

$$\mathbb{E}^{M^{\delta,\beta}_i}\mathbb{E}_{\pi \sim p(o_1,\ldots,o_T)}[g^i(\pi)] \leq \frac{(1 - \delta(N-1) - \beta)^T}{N} + \frac{\beta}{\beta + \delta(N-1)} \leq \frac{(1 - \delta(N-1))^T}{N} + \frac{\beta}{\delta(N-1)}, \tag{31}$$

其中，右侧第一项对应于 $o_1 = \cdots = o_T = \perp$ 的情况，而第二项给出这样的概率：在存在满足 $o_s \neq \perp$ 的 s 的条件下，第一个满足条件的观测的索引 t 满足 $o_t = i$。

极小极大风险的下界。接下来，我们在以下引理中为实例 $\mathscr{H}^{\delta,\beta}$ 给出极小极大风险的下界；证明见本节末尾。

引理 D.1。任取实数 $C \geq 1$ 和 $\epsilon \in (0, 1)$，假设 $\delta \leq 1/N$，并记 $\beta = \delta/C$。实例 $\mathscr{H}^{\delta,\beta}$ 的极小极大风险具有如下下界：对于 $S \leq 1/\delta^{1-\epsilon}$，$\mathfrak{M}(\mathscr{H}^{\delta,\beta}, S) \geq \frac{1}{2NC^{2/\epsilon}}$。

计算 D_ϕ。现在可以直接计算 $\mathcal{M}^{\delta,\beta}$ 中任意两个模型之间的 D_ϕ-散度。特别地，对于 $i \in [N]$，我们有：

$$D_\phi(M_i^{\delta,\beta} \parallel M_j^{\delta,\beta}) = \begin{cases} 0 & : i = j \\ \beta \cdot \phi(\delta/\beta) + \delta \cdot \phi(\beta/\delta) & : i \neq j. \end{cases} \tag{32}$$

选取 δ，β。令 $\epsilon \in (0, 1)$ 和 $T \in \mathbb{N}$ 如定理陈述所述。由于假设 ϕ 是 (α, β)-有界的，因此

$$\frac{\phi(N^{\epsilon/\alpha}) \cdot N^{-\epsilon/\alpha} + \phi(N^{-\epsilon/\alpha})}{\phi(2)/2 + \phi(1/2)} \leq \frac{\beta \cdot N^\epsilon}{\phi(2)/2 + \phi(1/2)} \leq \beta' \cdot N^\epsilon, \tag{33}$$

其中我们记 $\beta' := \max\left\{1, \frac{\beta}{\phi(2)/2+\phi(1/2)}\right\}$。

对于下面将要指定的某个常数 $C_0 > 0$，我们选择

$$\delta_1 = \frac{C_0 \ln T}{(N-1)T}, \quad \beta_1 = \frac{\delta_1}{N^{\epsilon/\alpha}}; \qquad \delta_2 = \frac{\phi(N^{\epsilon/\alpha}) \cdot N^{-\epsilon/\alpha} + \phi(N^{-\epsilon/\alpha})}{\phi(2)/2 + \phi(1/2)} \cdot \delta_1, \quad \beta_2 = \frac{\delta_2}{2}. \tag{34}$$

δ_1，β_1，δ_2，β_2 的选取确保

$$\beta_1 \cdot \phi(\delta_1/\beta_1) + \delta_1 \cdot \phi(\beta_1/\delta_1) = \beta_2 \cdot \phi(\delta_2/\beta_2) + \delta_2 \cdot \phi(\beta_2/\delta_2), \tag{35}$$

这与 (32)(32) 一起确保对所有 i，$j \in [N]$，$D_\phi(M^{\delta_1, \beta_1}_{i} \parallel M^{\delta_1, \beta_1}_{j}) = D_\phi(M^{\delta_2, \beta_2}_{i} \parallel M^{\delta_2, \beta_2}_{j})$。

收尾。我们令 $\mathscr{H}_1 = \mathscr{H}^{\delta_1, \beta_1}$ 和 $\mathscr{H}_2 = \mathscr{H}^{\delta_2, \beta_2}$，相应地令 $\mathcal{M}_1 = \mathcal{M}^{\delta_1, \beta_1}$ 和 $\mathcal{M}_2 = \mathcal{M}^{\delta_2, \beta_2}$。我们定义一一映射 $\mathscr{E}$：$\mathcal{M}_1 \to \mathcal{M}_2$，该映射将 $M^{\delta_1, \beta_1}_{i} \mapsto M^{\delta_2, \beta_2}_{i}$ 对所有 $i \in [N]$。显然，这些定义满足命题陈述的 条目 1 和 条目 2 。

根据 （31），p 在 $\mathcal{M}^{\delta_1, \beta_1}$中针对最坏情况模型的期望风险的上界如下：

$$\begin{aligned} \sup_{M^\star \in \mathcal{M}^{\delta_1,\beta_1}} \mathbb{E}^{M^\star, p}[\mathbf{Risk}(T)] &\leq \left(1 - \frac{C_0 \ln T}{T}\right)^T + \frac{\beta_1}{\delta_1 \cdot (N-1)} \\ &\leq \exp\left(-\frac{C_0 \ln T}{T}\right)^T + \frac{2}{N^{1+\epsilon/\alpha}} \\ &\leq T^{-1} + 2 \cdot \left(\frac{C_{\text{prob}}}{T}\right)^{1/2+\epsilon/(2\alpha)}, \end{aligned}$$

其中，只要我们选择 $C_0 = 1$，最后一个不等式便成立；回顾 $N := \lceil \sqrt{T/C_{\text{prob}}} \rceil$。上述推导确立了条目 3的上界。

接下来，对于下界，回顾 （33）给出 $\delta_2 \leq \beta' \cdot N^\epsilon \cdot \delta_1$，因此

$$\frac{1}{\delta_2^{1-\epsilon}} \geq \frac{1}{(\beta' N^\epsilon \cdot \delta_1)^{1-\epsilon}} \geq \frac{(N^{1-\epsilon}T)^{1-\epsilon}}{2\beta' C_0 \ln T} \geq \frac{N^{1-4\epsilon} \cdot T}{2\beta' C_0 C_{\text{prob}}^\epsilon \ln T} \geq \frac{T^{3/2-2\epsilon}}{2\beta' C_0 C_{\text{prob}}^{1/2+\epsilon} \ln T},$$

其中，倒数第二个不等式利用了 $T^\epsilon \leq N^{2\epsilon} \cdot C_{\text{prob}}^\epsilon$。因此，根据引理 D.1 ，取 $(\delta, \beta) = (\delta_2, \beta_2)$ （从而 $C = 2$），对于所有 $T' \leq \frac{T^{3/2-2\epsilon}}{2\beta' C_0 C^{1/2+\epsilon} \ln T}$，我们有

$$\mathfrak{M}(\mathscr{H}^{\delta_2,\beta_2}) \geq \frac{1}{2^{1+2/\epsilon}} \cdot \frac{1}{N} \geq \frac{1}{2^{2+2/\epsilon}} \cdot \sqrt{\frac{T}{C_{\text{prob}}}}.$$

因此，取 $C_\phi = 2\beta' C_0$，上述不等式验证了 条目 3的下界。

□

引理 D.1 的证明。 考虑任意算法 $p\colon \mathcal{O}^S \to \Pi$。注意，$M^{\delta}_{1}$、$M^{\delta}_{2}$、……、$M^{\delta}_{N}$ 的分布都相同。因此，存在某个 $i \in [N]$，使得

$$\mathbb{E}^{M_i^{\delta,\delta}}\mathbb{E}_{\pi\sim p(o_1,\ldots,o_S)}[\mathbb{I}\{\pi = i\}] = \mathbb{E}^{M_i^{\delta,\delta}}\mathbb{E}_{\pi\sim p(o_1,\ldots,o_S)}[g^i(\pi)] \geq 1/N.$$

对于 $\lambda \geq 0$，定义

$$\mathcal{S}_\lambda := \left\{(o_1,\ldots,o_S) \in \mathcal{O}^S \ :\ |\{t \in [S] \ :\ o_t = i\}| > \lambda\right\}.$$

那么，$(o_1,\ldots,o_S) \in \mathcal{S}_\lambda$ 的概率上界如下：

$$\mathbb{P}^{M_i^{\delta,\delta}}\left((o_1,\ldots,o_S) \in \mathcal{S}_\lambda\right) \leq \binom{S}{\lambda}\cdot\delta^\lambda \leq (S\delta)^\lambda \leq \delta^{\epsilon\lambda}.$$

选择 $\lambda = 2/\epsilon$ 可得 $\delta^{\epsilon\lambda} = \delta^2 \leq 1/N^2$，这意味着

$$\mathbb{E}^{M_i^{\delta,\delta}}\mathbb{E}_{\pi\sim p(o_1,\ldots,o_S)}\left[\mathbb{I}\left\{(o_1,\ldots,o_S) \notin \mathcal{S}_{2/\epsilon}\right\}\cdot\mathbb{I}\{\pi = i\}\right] \geq 1/N - 1/N^2 \geq 1/(2N).$$

对于任意 $(o_1,\ldots,o_S) \notin \mathcal{S}_\lambda$，有

$$\frac{M_i^{\delta,\beta}((o_1,\ldots,o_S))}{M_i^{\delta,\delta}((o_1,\ldots,o_S))} \geq (\beta/\delta)^\lambda \leq 1/C^\lambda.$$

因此，

$$\begin{aligned}
\mathbb{E}^{M_i^{\delta,\beta}}\mathbb{E}_{\pi\sim p(o_1,\ldots,o_S)}[g^i(\pi)] &\geq \mathbb{E}^{M_i^{\delta,\beta}}\mathbb{E}_{\pi\sim p(o_1,\ldots,o_S)}[\mathbb{I}\{\pi = i\}\cdot\mathbb{I}\left\{(o_1,\ldots,o_S) \notin \mathcal{S}_{2/\epsilon}\right\}] \\
&\geq \mathbb{E}^{M_i^{\delta,\delta}}\mathbb{E}_{\pi\sim p(o_1,\ldots,o_S)}[\mathbb{I}\{\pi = i\}\cdot\mathbb{I}\left\{(o_1,\ldots,o_S) \notin \mathcal{S}_{2/\epsilon}\right\}]\cdot 1/C^{2/\epsilon} \\
&\geq \frac{1}{2NC^{2/\epsilon}}.
\end{aligned}$$

□

D.3 来自 第3.3节的证明

定理 3.4 的证明。给定 $C_{\text{prob}} \geq 1$，固定 $N = \lceil\sqrt{T/C_{\text{prob}}}\rceil$。回顾定理 3.3 的证明中定义的 HR-DMSO 框架的实例 $\mathscr{H}^{\delta,\beta} = (\mathcal{M}^{\delta,\beta}, \Pi, \mathcal{O}, \{f^M(\cdot)\}_M)$（对于 $\delta,\beta \in (0,1)$），其中有 $\Pi = [N]$ 和 $\mathcal{O} = [N]\cup\{\perp\}$。对于每个 δ,β，我们现在将 $\mathscr{M}^{\delta,\beta} = (\widetilde{\mathcal{M}}^{\delta,\beta}, \widetilde{\Pi}, \widetilde{\mathcal{O}}, \{\Pi'_k\}_k, \{U_k\}_k)$定义为根据定理 2.2 证明中的构造，由实例 $\mathscr{H}^{\delta,\beta}$ 构造出的（双人）MA-DMSO 框架的实例，其取值为 V，具体将在下文说明。特别地，$\Pi, \mathcal{O}, \Pi'_k, U_k$ 不依赖于 δ,β。为清楚起见，我们明确写出 $\mathscr{M}^{\delta,\beta}$ 的各组成部分如何用 $\mathscr{H}^{\delta,\beta}$的各组成部分表示：

- 为 $k \in \{1,2\}$ 和 $\Pi = \Pi_1 \times \Pi_2$ 定义 $\Sigma_1 = \Pi = \widetilde{[N]}$ 和 $\Sigma_2 = \{0,1,\ldots,V\}$、$\widetilde{\Pi_k} = \widetilde{\Delta(\Sigma_k)}$。
- 定义 Π'_k，U_k 对于 $k \in [2]$ 使得 $\mathscr{M}^{\delta,\beta}$ 是一个纳什均衡实例（定义 1.1）；特别地，$\Pi'_k = \widetilde{\Pi}_k$ 对于每个 k 和 $U_k(\pi'_k, \pi) = (\pi_k, \pi_{-k})$。~
- 定义纯观测空间为 $\mathcal{O}_\circ := \mathcal{O}\cup\{\perp\}$，奖励空间为 $\mathcal{R} = [-1,1]$，完整观察空间为 $\widetilde{\mathcal{O}} := \mathcal{O}_\circ \times \mathbb{R}^2$。
- 模型类 $\widetilde{\mathcal{M}}^{\delta,\beta}$ 由元组 $(M,v) \in \mathcal{M}^{\delta,\beta}\times\{1,2,\ldots,V\} = \mathcal{M}^{\delta,\beta}\times[V]$索引。（因此$|\widetilde{\mathcal{M}}^{\delta,\beta}| = NV$。）特别地，对于每个这样的元组 (M,v)，我们都有一个模型 $\widetilde{M}_{M,v}$，其定义如下：

– 对于形式为 $(\sigma_1,0) \in \Sigma_1 \times \Sigma_2$ 的纯决策，$(o_\circ, r_1, r_2) \sim M_{M,v}((\sigma_1,0))$ 的分布为：

$$o_\circ \sim M(\sigma_1) \in \mathcal{O} \subset \mathcal{O}_\circ, \qquad r_1 = r_2 = 0.$$

– 对于形式为 $(\sigma_1, i) \in \Sigma_1 \times \Sigma_2$ 且满足 $i > 0$ 的纯决策，$(o_\circ, r_1, r_2) \sim M_{M,v}((\sigma_1, i))$ 的分布为：

$$o_\circ = \perp, \qquad r_2 = -r_1 = \begin{cases} -1 & : i \neq v \\ g^M(\sigma_1) & : i = v, \end{cases} \tag{36}$$

这里回顾一下，$g^M(\sigma_1) = \max_{\sigma_1' \in \Sigma_1}\{f^M(\sigma_1')\} - f^M(\sigma_1)$。

– 对于一般决策 $\pi \in \Pi$，我们可以写成 $\pi = \pi_1 \times \pi_2$，其中 $\pi_k \in \Pi_k$，对于 $k \in [2]$。那么，分布 $M_{M,v}(\pi)$ 是 $M_{M,v}(\sigma)$ 的分布，其中 $\sigma = (\sigma_1, \sigma_2)$ 的分布为：$\sigma_k \sim \pi_k$，对于 $k \in [2]$。

接下来，在给定 $T, N, \epsilon, \phi, \alpha$ 的情况下，定义 δ_1, δ_2，与定理 3.3 的证明中相同（特别是，它们在 (34) 中给出）。我们记 $\mathscr{M}_1 = \mathscr{M}^{\delta_1,\beta_1}$ 和 $\mathscr{M}_2 = \mathscr{M}^{\delta_2,\beta_2}$，相应地记 $\mathcal{M}_1 = \widetilde{\mathcal{M}}^{\delta_1,\beta_1}$ 和 $\mathcal{M}_2 = \widetilde{\mathcal{M}}^{\delta_2,\beta_2}$。此外，我们以与定理 3.3 证明中的定义类似的方式定义映射 $\mathscr{E}$：$\mathcal{M}_1 \to \mathcal{M}_2$。特别地，对于每个 δ, β，都有 $\mathcal{M}^{\delta,\beta} = \{M_1^{\delta,\beta}, \ldots, M_N^{\delta,\eta}\}$。首先定义 $\mathscr{E}_0$：$\mathcal{M}^{\delta_1,\beta_1} \to \mathcal{M}^{\delta_2,\beta_2}$，使得 $\mathscr{E}_0(M^{\delta_1,\beta_1}_i) = M^{\delta_2,\beta_2}_i$，对于 $i \in [N]$（与定理 3.3 的证明中完全相同）。然后，对于每个形式为 $\widetilde{M}_{M,v} \in \widetilde{M}^{\delta_1,\beta_1} = \mathcal{M}_1$ 的模型（因此 $M \in \mathcal{M}^{\delta_1,\beta_1}, v \in [V]$），定义 $\mathscr{E}(M_{M,v}) := M_{\mathscr{E}_0(M),v_\circ}$。现在我们可以验证该定理的各项结论：

条目1的证明。 任取 $\widetilde{M}_{M,v} \in \widetilde{\mathcal{M}}^{\delta_1,\beta_1}$（因此 $M \in \mathcal{M}^{\delta_1,\beta_1}$，$v \in [V]$）。对于任意 $(\sigma_1, \sigma_2) \in \Sigma_1 \times \Sigma_2$，根据 $\mathscr{E}$ 的定义，我们有，

$$f_2^{\widetilde{M}_{M,v}}(\sigma_1,\sigma_2) = f_2^{\mathscr{E}(\widetilde{M}_{M,v})}(\sigma_1,\sigma_2) = \begin{cases} 0 & : \sigma_2 = 0 \\ -1 & : \sigma_2 \in [V]\backslash\{v\} \\ g^M(\sigma_1) & : \sigma_2 = v, \end{cases}$$

这就证明了条目1，因为所有实例都是双人零和实例。

条目 2 的证明。 考虑任意两个模型 $\widetilde{M}_{M,v}$，$\widetilde{M}_{M',v'} \in \widetilde{\mathcal{M}}^{\delta_1,\beta_1}$（因此 M，$M' \in \mathcal{M}^{\delta_1,\beta_1}$，且$v$，$v' \in [V]$）。对于任意 $\pi_1 \in \Pi_1$，我们有

$$\begin{aligned} &D_\phi(\widetilde{M}_{M,v}(\pi_1,0) \,\|\, \widetilde{M}_{M',v'}(\pi_1,0)) = D_\phi(M(\pi_1) \,\|\, M'(\pi_1)) \\ =&D_\phi(\mathscr{E}_0(M)(\pi_1) \,\|\, \mathscr{E}_0(M')(\pi_1)) = D_\phi(\widetilde{M}_{\mathscr{E}_0(M),v}(\pi_1,0) \,\|\, \widetilde{M}_{\mathscr{E}_0(M'),v'}(\pi_1,0)) \\ =&D_\phi(\mathscr{E}(\widetilde{M}_{M,v})(\pi_1,0) \,\|\, \mathscr{E}(\widetilde{M}_{M',v'})(\pi_1,0)), \end{aligned} \tag{37}$$

其中，第一个和第三个等式根据上文 $\widetilde{\mathcal{M}}^{\delta,\beta}$ 的定义得出，第二个等式由定理 3.3的 条目 2 以及以下事实得出：我们选择的 δ_1，β_1，δ_2，β_2 与定理 3.3的证明中所选取的完全相同（参见公式（34） 和 公式（35）），第四个等式则根据 $\mathscr{E}$ 的定义得出。

接下来，对于任意 $\sigma_1 \in \Sigma_1$ 和 $\sigma_2 \in \Sigma_2\backslash\{0\}$，注意到分布 $M_{M,v}(\sigma_1,\sigma_2)$ 和 $\mathscr{E}(M_{M,v})(\sigma_1,\sigma_2) = M_{\mathscr{E}_0(M),v}(\sigma_1,\sigma_2)$ 完全相同：这两个分布下的纯观测几乎必然都是 $\perp$，奖励则由式 (36) 给出；其中我们注意到，对于所有 $\sigma_1 \in \Sigma_1$，$g^M(\sigma_1) = g^{\mathscr{E}_0(M)}(\sigma_1)$。因此，对于任意 $\pi_1 \in \Pi_1$ 和 $\sigma_2 \in \Delta(\Sigma_2\backslash\{0\})$，分布 $M_{M,v}(\pi_1,\pi_2)$ 和 $\mathscr{E}(M_{M,v})(\pi_1,\pi_2)$ 完全相同。类似地，对于任意这样的 π_1, π_2，分布 $M_{M',v}(\pi_1,\pi_2)$ 和 $\mathscr{E}(M_{M',v'})(\pi_1,\pi_2)$ 完全相同。因此，

$$D_\phi(\widetilde{M}_{M,v}(\pi_1,\pi_2) \,\|\, \widetilde{M}_{M',v'}(\pi_1,\pi_2)) = D_\phi(\mathscr{E}(\widetilde{M}_{M,v})(\pi_1,\pi_2) \,\|\, \mathscr{E}(\widetilde{M}_{M',v'})(\pi_1,\pi_2)). \tag{38}$$

现在考虑任意联合决策 $(\pi_1,\pi_2)\in\Pi_1\times\Pi_2$。记 $\pi_2=\pi_2(0)\cdot\mathbb{I}_0+(1-\pi_2(0))\cdot\pi_2'$，其中 $\pi_2'\in\Delta(\Sigma_2\backslash\{0\})$。由于对于任意模型 $\widetilde{M}\in\widetilde{\mathcal{M}}^{\delta,\beta}$（对于任意 δ,β），分布 $\widetilde{M}(\pi_1,0)$ 和 $\widetilde{M}(\pi_1,\pi_2')$ 的支撑集互不相交（具体而言，在第二个分布下，纯观测始终为 $\perp$；而在第一个分布下，纯观测绝不会为 $\perp$），根据引理 B.1，对于任意两个模型 $\widetilde{M},\widetilde{M}'\in\mathcal{M}^{\delta,\beta}$，

$$\begin{aligned}D_\phi(\widetilde{M}(\pi_1,\pi_2)\parallel\widetilde{M}'(\pi_1,\pi_2))=&\pi_2(0)\cdot D_\phi(\widetilde{M}(\pi_1,0)\parallel\widetilde{M}'(\pi_1,0))\\&+(1-\pi_2(0))\cdot D_\phi(\widetilde{M}(\pi_1,\pi_2')\parallel\widetilde{M}'(\pi_1,\pi_2')).\end{aligned}\tag{39}$$

那么，对于决策 $(\pi_1,\ \pi_2)\in\Pi_1\times\Pi_2$，其中 π_2' 定义如上，我们有

$$\begin{aligned}&D_\phi(\widetilde{M}_{M,v}(\pi_1,\pi_2)\parallel\widetilde{M}_{M',v'}(\pi_1,\pi_2))\\=&\pi_2(0)\cdot D_\phi(\widetilde{M}_{M,v}(\pi_1,0)\parallel\widetilde{M}_{M',v'}(\pi_1,0))+(1-\pi_2(0))\cdot D_\phi(\widetilde{M}_{M,v}(\pi_1,\pi_2')\parallel\widetilde{M}_{M',v'}(\pi_1,\pi_2'))\\=&\pi_2(0)\cdot D_\phi(\mathscr{E}(\widetilde{M}_{M,v})(\pi_1,0)\parallel\mathscr{E}(\widetilde{M}_{M',v'})(\pi_1,0))+(1-\pi_2(0))\cdot D_\phi(\mathscr{E}(\widetilde{M}_{M,v})(\pi_1,\pi_2')\parallel\mathscr{E}(\widetilde{M}_{M',v'})(\pi_1,\pi_2'))\\=&D_\phi(\mathscr{E}(\widetilde{M}_{M,v})(\pi_1,\pi_2)\parallel\mathscr{E}(\widetilde{M}_{M',v'})(\pi_1,\pi_2)),\end{aligned}$$

其中，第一个和第三个等式用到了 (39)，第二个等式用到了 (37)和 (38)。上述推导验证了 条目 2。

条目 3 的证明。对于每个 $\delta,\ \beta\in(0,\ 1)$，根据 $\mathscr{H}^{\delta,\ \beta}$在 定理 2.2的证明中给出的构造，构造 $\mathscr{M}^{\delta,\ \beta}$，结合 定理 2.2的结论可得，对所有 $T'\in\mathbb{N}$，

$$\mathfrak{M}(\mathscr{M}^{\delta,\beta},T')\le\mathfrak{M}(\mathscr{H}^{\delta,\beta},T')\le\mathfrak{M}(\mathscr{M}^{\delta,\beta},T')+O((T'\log(T')/V)^{1/4}).\tag{40}$$

那么，定理 定理 3.3的条目 3，结合我们选择 $\delta_1,\ \beta_1,\ \delta_2,\ \beta_2$ 来仿照定理 3.3证明中的做法，可得对于所有满足 $T\le T'\le T^{3/2-2\epsilon}\cdot(C_\phi C^{1/2+\epsilon}\text{概率}\ln T)^{-1}$的 T'

$$\begin{aligned}\mathfrak{M}(\mathscr{M}_1,T')=\mathfrak{M}(\mathscr{M}^{\delta_1,\beta_1},T')\le&\mathfrak{M}(\mathscr{H}^{\delta_1,\beta_1},T')\le\frac{1}{T}+2\cdot\left(\frac{C_{\text{prob}}}{T}\right)^{1/2+\epsilon/(2\alpha)}\\\mathfrak{M}(\mathscr{M}_2,\ T')=\mathfrak{M}(\mathscr{M}^{\delta_2,\ \beta_2},\ T')\ge&\mathfrak{M}(\mathscr{H}^{\delta_2,\ \beta_2},\ T')-O((T'\log(T')/V)^{1/4})\\\ge&2^{-2-2/\epsilon}\cdot\left(\frac{C_{\text{prob}}}{T}\right)^{1/2}-O((T'\log(T')/V)^{1/4}).\end{aligned}\tag{41}$$

选择 $V=T^{100}\cdot2^{8+8/\epsilon}$ 可确保

$$\mathfrak{M}(\mathscr{M}_2,T')\ge2^{-3-2/\epsilon}\cdot\left(\frac{C_{\text{prob}}}{T}\right)^{1/2}.\tag{42}$$

一起（41）和（42）验证条目 3。

□

第 4 节的证明

在本节中，我们考虑 MA-DMSO 的一个实例 $\mathscr{M}=(\mathcal{M},\ \Pi,\ \mathcal{O},\ \{\Pi_k'\}_k,\ \{U_k\}_k)$，它是一个纳什均衡实例（定义 1.1）。特别地，对于任意 $M\in\mathcal{M},\ \pi\in\Pi$，我们有

$$h^M(\pi)=\sum_{k=1}^K h_k^M(\pi)=\sum_{k=1}^K\sup_{\pi_k'\in\Pi_k}f_k^M(\pi_k',\pi_{-k})-f_k^M(\pi).$$

E.1 一般博弈的界（凸决策空间）

定理 4.1 的证明。 对于每个 $k \in [K]$ 和 $\pi_{-k} \in \Pi_{-k}$，定义

$$\widetilde{\mathcal{M}}_k(\pi_{-k}) := \{\pi_k \mapsto M|_k(\pi_k, \pi_{-k}) \ : \ M \in \mathcal{M}\}.$$

根据 $\mathcal{M}_k$ 在 (17) 中的定义，对于每个 $k \in [K]$，$\mathcal{M}_k = \bigcup_{\pi_{-k} \in \Pi_{-k}} \mathcal{M}_k(\pi_{-k})$、$\widetilde{\quad}$，显然有 $\bigcup_{\pi_{-k} \in \Pi_{-k}} \mathsf{co}(\mathcal{M}_k(\pi_{-k})) \subseteq \mathsf{co}(\mathcal{M}_k)$。对于任意 $\pi_{-k} \in \Pi_{-k}$ 和 $M \in \mathsf{co}(\mathcal{M})$，我们将 $\mathsf{co}(\mathcal{M}_k(\pi_{-k}))$ 中的相应元素记为 (M, π_{-k})。（特别地，(M, π_{-k}) 是将 $\pi_k \mapsto M|_k(\pi_k, \pi_{-k})$ 的模型。）因此，只需证明以下更强的结果：对于每个 $M \in \mathsf{co}(\mathcal{M})$，

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathcal{M}, \overline{M}) \le \sum_{k=1}^{K} \sup_{\overline{\pi}_{-k} \in \Pi_{-k}} \textsf{r-dec}^{\mathrm{o}}_{\gamma/K}(\widetilde{\mathcal{M}}_k, (\overline{M}, \overline{\pi}_{-k})). \tag{43}$$

接下来，注意对于任意 $M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}$，模型 $\pi_k \mapsto M|_k(\pi_k, \pi_{-k})$ 的价值函数为 $f^{M|_k}(\pi_k) = f^M_k(\pi_k, \pi_{-k})$，对于 $\pi \in \Pi$（这是因为在 $M|_k(\pi_k, \pi_{-k})$ 下的奖励分布，正是智能体 k 在 $M(\pi_k, \pi_{-k})$ 下的奖励分布）。那么对于任意 $M \in \mathsf{co}(\mathcal{M})$，$\pi_{-k} \in \Pi_{-k}$，我们有

$$\begin{aligned}
&\textsf{r-dec}^{\mathrm{o}}_{\gamma/K}(\widetilde{\mathcal{M}}_k, (\overline{M}, \overline{\pi}_{-k})) \\
&= \inf_{p_k \in \Delta(\Pi_k)} \sup_{\substack{M \in \mathcal{M} \\ \pi_{-k} \in \Pi_{-k}}} \mathbb{E}_{\pi_k \sim p_k} \left[\max_{\pi'_k \in \Pi_k} f^M_k(\pi'_k, \pi_{-k}) - f^M_k(\pi_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D^2_{\mathsf{H}}\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big) \right] \\
&\ge \inf_{p_k \in \Delta(\Pi_k)} \sup_{\substack{M \in \mathcal{M} \\ \pi_{-k} \in \Pi_{-k}}} \mathbb{E}_{\substack{\pi_k \sim p_k \\ a_k \sim \pi_k}} \left[\max_{\pi'_k \in \Pi_k} f^M_k(\pi'_k, \pi_{-k}) - f^M_k(a_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D^2_{\mathsf{H}}\big(M(a_k, \pi_{-k}), \overline{M}(a_k, \overline{\pi}_{-k})\big) \right] \\
&= \inf_{\pi_k \in \Pi_k} \sup_{\substack{M \in \mathcal{M} \\ \pi_{-k} \in \Pi_{-k}}} \mathbb{E}_{a_k \sim \pi_k} \left[\max_{\pi'_k \in \Pi_k} f^M_k(\pi'_k, \pi_{-k}) - f^M_k(a_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D^2_{\mathsf{H}}\big(M(a_k, \pi_{-k}), \overline{M}(a_k, \overline{\pi}_{-k})\big) \right],
\end{aligned} \tag{44}$$

其中，不等式利用了平方海林格距离的联合凸性，最后一个不等式利用了这一事实：任意分布 $p_k \in \Delta(\Pi_k)$ 都可以替换为决策 $\widetilde{\pi}_k$ 的单点分布，而不改变表达式的取值 $= \mathbb{E}_{\pi_k \sim p_k}[\pi_k]$。

Thus

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma/K}(\widetilde{\mathcal{M}}_k, (\overline{M}, \overline{\pi}_{-k})) \ge \inf_{\pi_k \in \Pi_k} \sup_{\substack{M \in \mathcal{M} \\ \pi_{-k} \in \Pi_{-k}}} \mathbb{E}_{a_k \sim \pi_k} \left[h^M_k(a_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D^2_{\mathsf{H}}\big(M(a_k, \pi_{-k}), \overline{M}(a_k, \overline{\pi}_{-k})\big) \right].$$

不动点的存在性。 对于每个 $k \in [K]$，定义集值函数 $\mathcal{C}_k : \Pi \to \mathcal{P}(\Pi_k)$ 如下

$$\mathcal{C}_k(\overline{\pi}) := \arg\min_{\pi_k \in \Pi_k} \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} \mathbb{E}_{a_k \sim \pi_k} \left[h^M_k(a_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D^2_{\mathsf{H}}\big(M(a_k, \pi_{-k}), \overline{M}(a_k, \overline{\pi}_{-k})\big) \right].$$

此外，对于 $\pi_{-k} \in \Pi_{-k}$，$M \in \mathcal{M}$，定义函数 $G_{M, \pi_{-k}} : \Pi_k \times \Pi_{-k} \to \mathbb{R}$ 如下

$$G_{M, \pi_{-k}}(\pi_k, \overline{\pi}_{-k}) = \mathbb{E}_{a_k \sim \pi_k} \left[h^M_k(a_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D^2_{\mathsf{H}}\big(M(a_k, \pi_{-k}), \overline{M}(a_k, \overline{\pi}_{-k})\big) \right].$$

假设 4.1 表明，对于所有 a_k，映射 $\pi_{-k} \mapsto M(a_k, \pi_{-k})$ 是线性的。由控制收敛定理可知，对于所有 M，π_{-k}，a_k，函数 $\pi_{-k} \mapsto D^2_{\mathsf{H}}\big(M(a_k, \pi_{-k}) M(a_k, \pi_{-k})\big)$ 是连续的。因此，$G_{M, \pi_{-k}}(\pi_k, \pi_{-k})$ 关于 (π_k, π_{-k}) 连续，并且函数

$$\widetilde{G}_k(\pi_k, \overline{\pi}_{-k}) := \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} G_{M, \pi_{-k}}(\pi_k, \overline{\pi}_{-k})$$

在 (π_k, π_{-k}) 中也连续。此外，由于对于每个 π_{-k}，函数 $G_{M, \pi_{-k}}(\pi_k, \pi_{-k})$ 关于 π_k（假设 4.1）是线性的，因此 $G_k(\pi_k, \pi_{-k})$ 关于 π_k 是凸的。由此可知，对于所有 π，$\mathcal{C}_k(\pi) = \arg\min_{\pi_k \in \Pi_k} \{G_k(\pi_k, \pi_{-k})\}$ 是闭、非空且凸的 Π_k 子集。此外，根据 G_k 的连续性和引理 B.8，可知 $\mathcal{C}_k(\pi)$ 是上半连续的。根据引理 B.9，可知映射 $\pi \mapsto \mathcal{C}_1(\pi) \times \cdots \times \mathcal{C}_K(\pi)$ 有一个不动点，即存在某个 $\pi \in \Pi$ 使得 $\pi \in \prod_{k \in [K]} \mathcal{C}_k(\pi)$。

应用不动点策略。令 $\pi \in \prod_{k\in[K]} \mathcal{C}_k(\pi)$ 为 $\mathcal{C}_1 \times \cdots \times \mathcal{C}_K$的一个不动点。则

$$
\begin{aligned}
\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}, \overline{M}) &\le \sup_{M\in\mathcal{M}} \left\{h^M(\overline{\pi}) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\overline{\pi}), \overline{M}(\overline{\pi})\big)\right\} \\
&= \sup_{M\in\mathcal{M}} \left\{\sum_{k=1}^K h^M_k(\overline{\pi}) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\overline{\pi}), \overline{M}(\overline{\pi})\big)\right\} \\
&\le \sum_{k=1}^K \sup_{M\in\mathcal{M}} \left\{h^M_k(\overline{\pi}) - \frac{\gamma}{K} \cdot D^2_{\mathsf{H}}\big(M(\overline{\pi}), \overline{M}(\overline{\pi})\big)\right\} \\
&\le \sum_{k=1}^K \sup_{M\in\mathcal{M}, \pi_{-k}\in\Pi_{-k}} \left\{h^M_k(\overline{\pi}_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D^2_{\mathsf{H}}\big(M(\overline{\pi}_k, \pi_{-k}), \overline{M}(\overline{\pi}_k, \overline{\pi}_{-k})\big)\right\} \\
&= \sum_{k=1}^K \sup_{M\in\mathcal{M}, \pi_{-k}\in\Pi_{-k}} \mathbb{E}_{a_k\sim\overline{\pi}_k}\left[h^M_k(a_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D^2_{\mathsf{H}}\big(M(a_k, \pi_{-k}), \overline{M}(a_k, \overline{\pi}_{-k})\big)\right] \\
&= \sum_{k=1}^K \inf_{\pi_k\in\Pi_k} \sup_{M\in\mathcal{M}, \pi_{-k}\in\Pi_{-k}} \mathbb{E}_{a_k\sim\pi_k}\left[h^M_k(a_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D^2_{\mathsf{H}}\big(M(a_k, \pi_{-k}), \overline{M}(a_k, \overline{\pi}_{-k})\big)\right] \\
&\le \sum_{k=1}^K \textsf{dec}_{\gamma/K}(\widetilde{\mathcal{M}}_k, (\overline{M}, \overline{\pi}_{-k})).
\end{aligned}
$$

上文中，我们用到了以下事实：

1. 第二个等式利用假设 4.1 得出，对于所有 a_k，π_k，π_{-k}，M，$\overline{M}$，

$$\mathbb{P}_{o\sim M(a_k,\pi_{-k})}(\varphi(o) = a_k) = 1, \quad \mathbb{P}_{o\sim \overline{M}(a_k,\overline{\pi}_{-k})}(\varphi(o) = a_k) = 1,$$

   从而可以应用引理 B.1 得到

$$D^2_{\mathsf{H}}\big(M(\overline{\pi}_k, \pi_{-k}), \overline{M}(\overline{\pi}_k, \overline{\pi}_{-k})\big) = \mathbb{E}_{a_k\sim\overline{\pi}_k}\left[D^2_{\mathsf{H}}\big(M(a_k, \pi_{-k}), \overline{M}(a_k, \overline{\pi}_{-k})\big)\right].$$

2. 第三个等式源于以下事实：对于所有 $k \in [K]$，$\pi_k \in \mathcal{C}_k(\pi)$。
3. 最后一个不等式由(44)得出。

□

E.2 马尔可夫博弈的界

这里，我们证明定理 4.2。证明中用到了若干技术引理，这些引理将在下文陈述并证明。

证明定理 4.2。与定理 4.1的证明类似，对于每个 $k \in [K]$ 和 $\pi_{-k} \in \Pi_{-k}$，我们定义

$$\widetilde{\mathcal{M}}_k(\pi_{-k}) := \{\pi_k \mapsto M|_k(\pi_k, \pi_{-k}) \ : \ M \in \mathcal{M}\}.$$

对于任意 $\pi_{-k} \in \Pi_{-k}$ 和 $M \in \mathcal{M}$，我们将 $\mathcal{M}_k(\pi_{-k}) \subseteq \mathcal{M}_k$ 中对应的元素记为(M, π_{-k})。我们将证明以下更强的结果：存在某个常数 $C' > 0$，使得对于每个 $M \in \mathcal{M}$，

$$\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathscr{M}, \overline{M}) \le \frac{C'KH\log H}{\gamma} + \sum_{k=1}^K \sup_{\overline{\pi}_{-k}\in\Pi_{-k}} \textsf{r-dec}^{\mathrm{o}}_{\gamma/(C'KH\log H)}(\widetilde{\mathcal{M}}_k, (\overline{M}, \overline{\pi}_{-k})). \tag{45}$$

固定任意 $\epsilon > 0$。对于每个 $k \in [K]$，令 Π^ϵ_k 为 Π_k 的一个有限 ϵ-覆盖，意即对于所有 $\pi_k \in \Pi_k$，都存在某个元素 $\pi^\epsilon_k \in \Pi^\epsilon_k$，使得对于所有 $M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}$，

$$D^2_{\mathsf{H}}(M(\pi_k, \pi_{-k}), M(\pi^\epsilon_k, \pi_{-k})) \le \epsilon^2.$$

此外，我们要求映射 $\pi_k \mapsto \pi_k^\epsilon$ 关于 Borel σ-代数上的 Π_k是可测的。由 $\mathcal{S}$的有限性，$\mathcal{A}_k$，不难看出这样的有限覆盖 Π_k^ϵ 存在。覆盖的大小 Π_k^ϵ可能取决于 $|\mathcal{S}|$，$|\mathcal{A}|$但这并不重要，因为 $|\Pi_k^\epsilon|$ 不会出现在我们最终的界中。（我们在此引入离散化，只是为了确保在应用 引理 B.8时 Π_k^ϵ 是紧致的。）

我们在下面的引理中总结了 Π_k^ϵ 的一些基本性质，证明见本节末尾：

引理 E.1. 对于任意 $\pi_k \in \Pi_k$，存在某个 $\pi_k^\epsilon \in \Pi_k^\epsilon$，使得以下结论成立。对于任意 M、$M \in \mathcal{M}$、$\pi_{-k} \in \overline{\Pi}_{-k}$，

$$\begin{aligned} D_{\mathsf{H}}^2\big(M(\pi_k^\epsilon, \pi_{-k}), \overline{M}(\pi_k^\epsilon, \overline{\pi}_{-k})\big) \geq & \frac{1}{3} \cdot D_{\mathsf{H}}^2\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big) - 2\epsilon^2 \\ |h_k^M(\pi_k, \pi_{-k}) - h_k^M(\pi_k^\epsilon, \pi_{-k})| \leq & \epsilon. \end{aligned}$$

不动点的存在性。令 $C > 0$ 为引理 E.4 的常数引理 E.4，并记 $\gamma' = \gamma/(CKH$ 对数 $H)$。对于每个 $k \in [K]$，定义函数 $\mathcal{C}_k$： $\Pi \to \Delta(\Pi_k^\epsilon)$ 为

$$\mathcal{C}_k(\overline{\pi}) = \operatorname*{arg\,min}_{p_k \in \Delta(\Pi_k^\epsilon)} \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} \mathbb{E}_{\pi_k \sim p_k}\left[h_k^M(\pi_k, \pi_{-k}) - \gamma' \cdot D_{\mathsf{H}}^2\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big)\right] + \epsilon \cdot \|p_k\|_2^2,$$

其中，$\|p_k\|_2^2$ 表示 p_k的平方 ℓ_2 范数，并将其视为欧几里得空间 $\mathbb{R}^{|\Pi_k^\epsilon|}$中的向量。

此外，对于 $\pi_{-k} \in \Pi_{-k}$，$M \in \mathcal{M}$，定义函数 $G_{M,\pi_{-k}}$： $\Delta(\Pi_k) \times \Pi_{-k} \to \mathbb{R}$ 为

$$G_{M,\pi_{-k}}(p_k, \overline{\pi}_{-k}) = \mathbb{E}_{\pi_k \sim p_k}\left[h_k^M(\pi_k, \pi_{-k}) - \gamma' \cdot D_{\mathsf{H}}^2\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big)\right].$$

我们可以将 $\overline{\pi}_{-k}$ 视为 $\Delta(\mathcal{A}_k)^{\mathcal{S}\times[H]}$ 的一个元素，它是欧几里得空间的子集（因为假设 $\mathcal{A}_k$，$\mathcal{S}$ 均为有限集）。由于状态和动作的数量有限，根据控制收敛定理，对于所有 M，π_k，π_{-k}，函数 $\pi_{-k} \mapsto D_{\mathsf{H}}^2\big(M(\pi_k, \pi_{-k}), M(\pi_k, \pi_{-k})\big)$ 连续。因此，$G_{M,\pi_{-k}}(p_k, \pi_{-k})$ 关于 (p_k, π_{-k}) 连续。因此函数

$$\widetilde{G}_k(p_k, \overline{\pi}_{-k}) := \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} G_{M,\pi_{-k}}(p_k, \overline{\pi}_{-k}) + \epsilon \cdot \|p_k\|_2^2$$

$\widetilde{\ }$ 也连续。此外，对于固定的 π_{-k}，$G_{M,\pi_{-k}}(p_k, \pi_{-k})$ 关于 p_k 是线性的，因此对于固定的 $\overline{\pi}_{-k}$，$G_k(p_k, \pi_{-k})$ 关于 p_k 是强凸的。因此，对于所有 $\overline{\pi}$，$\mathcal{C}_k(\overline{\pi}) = \arg\min_{p_k \in \Delta(\Pi_k^\epsilon)}\{G_{\overline{\pi}}(p_k, \pi_{-k})\}$ 是一个单元素集。此外，根据 G_k 的连续性、$\Delta(\Pi_k^\epsilon)$ 和 Π_{-k} 的紧致性，以及引理 B.8，我们有 $\mathcal{C}_k(\overline{\pi})$ 是上半连续的；这意味着，根据其单值性，它实际上是连续的。

给定 $\overline{M} \in \mathcal{M}$、$\overline{\pi}_{-k} \in \Pi_{-k}$，注意到模型 $\pi_k \mapsto M(\pi_k, \pi_{-k})$ 的纯观测分布恰好就是一个 MDP的分布，我们将其记为 $M_{\pi_{-k}}$：它的时域为 H，状态空间为 $\mathcal{S}$，动作空间为 $\mathcal{A}_k$；当每个智能体 $k' \neq k$ 在每个状态 s 和步 h 均按照 $\pi_{k'}$、$_h(\cdot|s)$ 采取动作时，其奖励和转移分别由 M 的奖励和转移给出（严格来说，$M_{\pi_{-k}}$ 的奖励由 k 在 M 中的奖励给出）。注意，$M_{\overline{\pi}_{-k}}$ 的随机非平稳策略空间为 Π_k（使用假设 4.2）。

由于我们不假设 Π_k 的凸性，$p_k \in \Delta(\Pi_k^\epsilon)$ 中的元素可能不属于 Π_k。接下来，我们在 Π_k 中引入一个决策集合：相对于参考模型 $\overline{M}$ 和参考决策$\overline{\pi}_{-k}$，其中的决策与 p_k “等价”。具体而言，对于 $M \in \mathcal{M}$、$\pi_{-k} \in \Pi_{-k}$和 $p_k \in \Delta(\Pi_k^\epsilon)$，令 $\overline{\Pi}^\star_{M,\pi_{-k}}(p_k) \subset \Pi_k$ 为满足引理 E.2 中针对 p_k 和 $\pi_k \mapsto M(\pi_k, \pi_{-k})$ 的式 (50) 的所有策略$\pi_k^\star \in \Pi_k$所构成的集合。注意，$\overline{\Pi}^\star_{M,\pi_{-k}}(p_k)$ 是一个非空凸集：作为 $\Delta(\mathcal{A}_k)^{\mathcal{S}\times[H]}$的子集，它是若干集合的笛卡尔积（每个因子对应 $\Delta(\mathcal{A}_k)$ 的一个因子），其中每个集合要么是单元素集，要么是整个 $\Delta(\mathcal{A}_k)$。根据定义可直接得出，映射 $(p_k, \pi_{-k}) \mapsto \overline{\Pi}^\star_{M,\pi_{-k}}(p_k)$ 是上半连续的。于是，引理E.4 表明，对于任意 M、$\overline{\pi}_{-k}$ 和 p_k，若$\pi_k^\star \in \overline{\Pi}^\star_{M,\pi_{-k}}(p_k)$ 是式 (50) 中相应的策略，则对于 $\gamma > 0$，

$$\begin{aligned} & \sup_{M \in \mathcal{M},\ \pi_{-k} \in \Pi_{-k}} \left\{h_k^M(\pi_k^\star, \pi_{-k}) - \frac{\gamma}{K} \cdot D_{\mathsf{H}}^2\big(M(\pi_k^\star, \pi_{-k}), \overline{M}(\pi_k^\star, \overline{\pi}_{-k})\big)\right\} \\ \leq & \frac{1}{\gamma'} + \sup_{M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}} \mathbb{E}_{\pi_k \sim p_k}\left[h_k^M(\pi_k, \pi_{-k}) - \gamma' \cdot D_{\mathsf{H}}^2\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big)\right]. \end{aligned} \tag{46}$$

由于映射 $\pi \mapsto \mathcal{C}_k(\pi)^- \in \Delta(\Pi_k^\epsilon)$ 是连续的，复合映射 $\mathcal{C}_k^\star(\pi) := \Pi^\star_{M,\,\pi_{-k}}(\mathcal{C}_k(\pi))$ 是上半连续的。因此，根据卡库塔尼不动点定理（Osborne andRubinstein，1994，引理 20.1），取值映射 $C^\star(\pi) := \mathcal{C}_1^\star(\pi) \times \cdots \times \mathcal{C}_K^\star(\pi)$ 存在一个不动点。

应用不动点策略。 令 $\pi \in \Pi$ 为 $C^\star$的一个不动点，因此对于每个$k \in [K]$，$\pi_k \in \mathcal{C}_k^\star(\pi)$。于是

$$
\begin{aligned}
&\textsf{r-dec}^{\mathrm{o}}_\gamma(\mathscr{M}, \overline{M}) \\
\le& \sup_{M\in\mathcal{M}} \left\{ \sum_{k=1}^K h_k^M(\overline{\pi}) - \gamma \cdot D^2_{\mathsf{H}}\big(M(\overline{\pi}), \overline{M}(\overline{\pi})\big) \right\} \\
\le& \sum_{k=1}^K \sup_{M\in\mathcal{M}} \left\{ h_k^M(\overline{\pi}_k, \overline{\pi}_{-k}) - \frac{\gamma}{K} \cdot D^2_{\mathsf{H}}\big(M(\overline{\pi}_k, \overline{\pi}_{-k}), \overline{M}(\overline{\pi}_k, \overline{\pi}_{-k})\big) \right\} \\
\le& \sum_{k=1}^K \sup_{M\in\mathcal{M},\ \pi_{-k}\in\Pi_{-k}} \left\{ h_k^M(\overline{\pi}_k, \pi_{-k}) - \frac{\gamma}{K} \cdot D^2_{\mathsf{H}}\big(M(\overline{\pi}_k, \pi_{-k}), \overline{M}(\overline{\pi}_k, \overline{\pi}_{-k})\big) \right\} \\
\le& \frac{1}{\gamma'} + \sum_{k=1}^K \sup_{M\in\mathcal{M},\pi_{-k}\in\Pi_{-k}} \mathbb{E}_{\pi_k\sim\mathcal{C}_k(\overline{\pi})} \left[h_k^M(\pi_k, \pi_{-k}) - \gamma' \cdot D^2_{\mathsf{H}}\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big) \right] \qquad (47) \\
\le& \frac{1}{\gamma'} + \sum_{k=1}^K \epsilon + \inf_{p_k\in\Delta(\Pi_k^\epsilon)} \sup_{M\in\mathcal{M},\pi_{-k}\in\Pi_{-k}} \mathbb{E}_{\pi_k\sim p_k} \left[h_k^M(\pi_k, \pi_{-k}) - \gamma' \cdot D^2_{\mathsf{H}}\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big) \right] \qquad (48) \\
\le& \frac{1}{\gamma'} + \sum_{k=1}^K 2\epsilon + \gamma' \cdot 2\epsilon^2 + \inf_{p_k\in\Delta(\Pi_k)} \sup_{M\in\mathcal{M},\pi_{-k}\in\Pi_{-k}} \mathbb{E}_{\pi_k\sim p_k} \left[h_k^M(\pi_k, \pi_{-k}) - \frac{\gamma'}{3} \cdot D^2_{\mathsf{H}}\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big) \right]. \qquad (49)
\end{aligned}
$$

其中，(47) 使用了等式 (46) 以及如下事实：对于每个 k，$\pi_k \in \mathcal{C}_k^\star(\pi) = \Pi^\star_{M,\pi_{-k}}(\mathcal{C}_k(\pi))$；(48) 使用了 $\mathcal{C}_k(\pi)$ 的定义。最后，(49) 使用了引理 E.1，具体如下：给定任意分布 $p_k \in \Delta(\Pi_k)$，我们考虑分布 $p_k^\epsilon \in \Delta(\Pi_k^\epsilon)$，它是将 p_k 通过映射 $\pi_k \mapsto \pi_k^\epsilon$推前得到的（这里我们利用 $\pi_k \mapsto \pi_k^\epsilon$ 可测这一点来确保 $p_k^\epsilon p$ 定义良好）。于是根据引理 E.1，对于所有$M \in \mathcal{M}, \pi_{-k} \in \Pi_{-k}, \pi_{-k} \in \Pi_{-k}$，有

$$
\begin{aligned}
\mathbb{E}_{\pi_k\sim p_k^\epsilon}[h_k^M(\pi_k, \pi_{-k})] \le& \mathbb{E}_{\pi_k\sim p_k}[h_k^M(\pi_k, \pi_{-k})] + \epsilon \\
-\gamma' \cdot \mathbb{E}_{\pi_k\sim p_k^\epsilon}[D^2_{\mathsf{H}}\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big)] \le& -\frac{\gamma'}{3} \cdot \mathbb{E}_{\pi_k\sim p_k}[D^2_{\mathsf{H}}\big(M(\pi_k, \pi_{-k}), \overline{M}(\pi_k, \overline{\pi}_{-k})\big)] + \gamma' \cdot 2\epsilon^2.
\end{aligned}
$$

取 $\epsilon \to 0$，我们得到，对于某个常数 $C > 0$，

$$
\begin{aligned}
\textsf{r-dec}^{\mathrm{o}}_\gamma(\mathscr{M}, \overline{M}) \le& \frac{CKH\log H}{\gamma} + \sum_{k=1}^K \textsf{r-dec}^{\mathrm{o}}_{\gamma/(CKH\log H)}(\widetilde{\mathcal{M}}_k, (\overline{M}, \overline{\pi}_{-k})) \\
\le& \frac{CKH\log H}{\gamma} + \sum_{k=1}^K \sup_{\widetilde{\pi}_{-k}\in\Pi_{-k}} \textsf{r-dec}^{\mathrm{o}}_{\gamma/(CKH\log H)}(\widetilde{\mathcal{M}}_k, (\overline{M}, \widetilde{\pi}_{-k})),
\end{aligned}
$$

从而验证了(45)。 □

E.2.1 辅助引理

引理 引理 E.1 的证明。 为证明第一个性质，我们使用 Π_k^ϵ 的定义以及海林格距离的三角不等式，得出

$$\begin{aligned}
&D_{\mathsf{H}}^2\big(M(\pi_k,\pi_{-k}),\overline{M}(\pi_k,\overline{\pi}_{-k})\big)\\
\leq&3\cdot\Big(D_{\mathsf{H}}^2(M(\pi_k^\epsilon,\pi_{-k}),M(\pi_k,\pi_{-k}))+D_{\mathsf{H}}^2\big(\overline{M}(\pi_k^\epsilon,\overline{\pi}_{-k}),\overline{M}(\pi_k,\overline{\pi}_{-k})\big)+D_{\mathsf{H}}^2\big(M(\pi_k^\epsilon,\pi_{-k}),\overline{M}(\pi_k^\epsilon,\overline{\pi}_{-k})\big)\Big)\\
\leq&3\cdot\Big(D_{\mathsf{H}}^2(M(\pi_k^\epsilon,\pi_{-k}),M(\pi_k,\pi_{-k}))+2\epsilon^2\Big),
\end{aligned}$$

整理后即可得到引理中所述的第一个不等式。

为了证明第二个不等式，我们注意到，对于每个 $\pi_k\in\Pi_k$，覆盖元素 $\pi_k^\epsilon\in\Pi_k^\epsilon$ 满足以下条件：对于所有 $M\in\mathcal{M}$，$\pi_{-k}\in\Pi_{-k}$

$$|h_k^M(\pi_k,\pi_{-k})-h_k^M(\pi_k^\epsilon,\pi_{-k})|=|f_k^M(\pi_k,\pi_{-k})-f_k^M(\pi_k^\epsilon,\pi_{-k})|\leq D_{\mathsf{H}}(M(\pi_k,\pi_{-k}),M(\pi_k^\epsilon,\pi_{-k}))\leq\epsilon.$$

□

以下引理表明，对于任意马尔可夫决策过程 M 和分布 $p\in\Delta(\Pi_{\mathrm{RNS}})$，都存在一个相应的随机策略属于 Π_{RNS}，它在 M中诱导出相同的占用度。

引理 E.2. 考虑任意有限时域马尔可夫决策过程 $M=(\mathcal{S},H,\mathcal{A},P,R,\mu)$，其状态和动作空间$\mathcal{S},\mathcal{A}$均为有限集合。令 Π_{RNS} 表示 M的随机非平稳策略集合。假设 $p\in\Delta(\Pi_{\mathrm{RNS}})$ 是定义在 Π_{RNS} 上的有限支撑分布。考虑任意满足以下条件的策略 $\pi^\star\in\Pi_{\mathrm{RNS}}$：

$$\forall a\in\mathcal{A},\ s\in\mathcal{S}\ s.t.\ \sum_{\pi'\in\Pi_{\mathrm{RNS}}}p(\pi')\cdot d_h^{\overline{M},\pi'}(s)>0:\quad \pi_h^\star(a|s)=\sum_{\pi\in\Pi_{\mathrm{RNS}}:\ p(\pi)>0}\frac{p(\pi)\cdot d_h^{\overline{M},\pi}(s)}{\sum_{\pi'\in\Pi_{\mathrm{RNS}}}p(\pi')\cdot d_h^{\overline{M},\pi'}(s)}\cdot\pi_h(a|s). \tag{50}$$

那么，对于所有状态 $s\in\mathcal{S}$，$d_h^{M,\pi^\star}(s)=\sum_{\pi\in\Pi_{\mathrm{RNS}}}p(\pi)\cdot d_h^{M,\pi}(s)$，以及所有 $(s,a)\in\mathcal{S}\times\mathcal{A}$，$d_h^{M,\pi^\star}(s,a)=\sum_{\pi\in\Pi_{\mathrm{RNS}}}p(\pi)\cdot d_h^{M,\pi}(s,a)$。

因此可得 $V_1^{M,\pi^\star}=\sum_{\pi\in\Pi_{\mathrm{RNS}}}p(\pi)\cdot V_1^{M,\pi}$。

引理 E.2 的证明。在整个证明过程中，我们省略所有相关量中的上标 M。我们对 h使用归纳法，并注意到基础情形 $h=1$ 是显然的，因为 d_1^π 对于所有 $\pi\in\Pi_{\mathrm{RNS}}$都相同。固定$p\in\Delta(\Pi_{\mathrm{RNS}})$，并令 $\pi^\star$ 按照式 (50)进行选取。假设引理的陈述在第$h-1$步成立，我们计算

$$\begin{aligned}
d_h^{\pi^\star}(s)=&\sum_{\substack{s',a':\\ d_{h-1}^{\pi^\star}(s')>0}}d_{h-1}^{\pi^\star}(s')\cdot\pi_{h-1}^\star(a'|s')\cdot P_{h-1}(s|s',a')\\
=&\sum_{\substack{s',a':\\ d_{h-1}^{\pi^\star}(s')>0}}\left(\sum_{\pi'}p(\pi')\cdot d_{h-1}^{\pi'}(s')\right)\cdot\sum_{\pi}\frac{p(\pi)\cdot d_{h-1}^{\pi}(s')}{\sum_{\pi'}p(\pi')\cdot d_{h-1}^{\pi'}(s)}\cdot\pi_{h-1}(a'|s')\cdot P_{h-1}(s|s',a')\\
=&\sum_{\pi}p(\pi)\cdot\sum_{\substack{s',a':\\ d_{h-1}^{\pi^\star}(s')>0}}d_{h-1}^{\pi}(s')\cdot\pi_{h-1}(a'|s')\cdot P_{h-1}(s|s',a')\\
=&\sum_{\pi}p(\pi)\cdot\sum_{s',a'}d_{h-1}^{\pi}(s')\cdot\pi_{h-1}(a'|s')\cdot P_{h-1}(s|s',a')\\
=&\sum_{\pi}p(\pi)\cdot d_h^{\pi}(s),
\end{aligned}$$

其中，倒数第二个不等式成立是因为，如果 $d_{h-1}^{\pi^\star}(s')=0$，那么（利用归纳假设）对于所有 π, $p(\pi)\cdot d_{h-1}^{\pi}(s')=0$。上述等式链由此完成归纳步骤。根据 $\pi^\star$ 的定义，立即可得 $d_h^{\pi^\star}(s,\ a)=\sum_{\pi\in\Pi_{\mathrm{RNS}}}p(\pi)\cdot d_h^{\pi}(s,\ a)$。

关于价值函数的最后一个结论成立，这是因为对于所有策略 π,

$$V_1^{\pi}=\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\pi}(s,a)\cdot r_h(s,a).$$

□

其余引理证明了策略 $\pi^\star\in\Pi_{\mathrm{RNS}}$ 所具有的一些技术性质，该策略构造于引理 E.2中。

引理 *E.3.* 存在一个常数 $C>0$，使得以下结论成立。考虑任意具有有限状态和动作空间 $\mathcal{S},\mathcal{A}$ 的有限时域马尔可夫决策过程$M=(\mathcal{S},H,\mathcal{A},P^{M},R^{M},\mu^{M})$。令 Π_{RNS} 表示 M 的随机非平稳策略集合，并令 $p\in\Delta(\Pi_{\mathrm{RNS}})$ 为一个有限支撑分布。考虑满足关于 p 的式 *(50)* 的任意策略$\overline{\pi^\star}\in\Pi_{\mathrm{RNS}}$。那么，对于任意马尔可夫决策过程 $\mathcal{M}=(\mathcal{S},H,\mathcal{A},P^{\mathcal{M}},R^{\mathcal{M}},\mu^{\mathcal{M}})$,

$$\mathbb{E}_{\pi\sim p}\left[D_{\mathsf{H}}^2\big(M(\pi),\overline{M}(\pi)\big)\right]\le CH\log H\cdot D_{\mathsf{H}}^2\big(M(\pi^\star),\overline{M}(\pi^\star)\big).$$

引理 E.3 的证明。对于任意 $\pi\in\Pi_{\mathrm{RNS}}$，完整观测 $(r,o_\circ)\sim M(\pi)$ 由轨迹 $(s_1,a_1,r_1,\ldots,s_H,a_H,r_H)$ 组成，其中 $s_1\sim\mu^M$, $s_{h+1}\sim P_h^M(s_h,a_h)$ 对于 $h\in[H-1]$, $r_h\sim R_h^M(s_h,a_h)$ 对于 $h\in[H]$，且 $a_h\sim\pi_h(s_h)$ 对于 $h\in[H]$。我们用记号 $\tau_{1:h}$ 表示轨迹中由 $(s_1,a_1,r_1,\ldots,s_h,a_h,r_h)$ 组成的部分。

我们用 $\mathbb{P}^{M,\pi}$ 表示轨迹 $\tau_H\sim M(\pi)$的分布，并用 $\mathbb{P}^{\overline{M},\pi}$ 表示轨迹 $\tau_H\sim\overline{M}(\pi)$的分布。我们用 $\mathbb{E}^{M,\pi}[\cdot]$ 和 $\mathbb{E}^{\overline{M},\pi}[\cdot]$ 表示相应的期望。根据 Foster 等人 (2021）中的引理 A.13，对于某个常数 $C>0$,

$$\begin{aligned}
&\mathbb{E}_{\pi\sim p}\left[D_{\mathsf{H}}^2\big(M(\pi),\overline{M}(\pi)\big)\right]\\
\le& C\log(H)\cdot\mathbb{E}_{\pi\sim p}\mathbb{E}^{\overline{M},\pi}\left[\sum_{h=1}^{H}D_{\mathsf{H}}^2\big(\mathbb{P}^{M,\pi}(s_h|\tau_{1:h-1}),\mathbb{P}^{\overline{M},\pi}(s_h|\tau_{1:h-1})\big)\right]\\
&+C\log(H)\cdot\mathbb{E}_{\pi\sim p}\mathbb{E}^{\overline{M},\pi}\left[\sum_{h=1}^{H}D_{\mathsf{H}}^2\big(\mathbb{P}^{M,\pi}(r_h|\tau_{1:h-1},s_h,a_h),\mathbb{P}^{\overline{M},\pi}(r_h|\tau_{1:h-1},r_h,a_h)\big)\right]\\
=& C\log(H)\cdot D_{\mathsf{H}}^2\big(\mu^M,\mu^{\overline{M}}\big)+C\log(H)\cdot\mathbb{E}_{\pi\sim p}\mathbb{E}^{\overline{M},\pi}\left[\sum_{h=1}^{H-1}D_{\mathsf{H}}^2\big(P_h^M(s_h,a_h),P_h^{\overline{M}}(s_h,a_h)\big)\right]\\
&+C\log(H)\cdot\mathbb{E}_{\pi\sim p}\mathbb{E}^{\overline{M},\pi}\left[\sum_{h=1}^{H}D_{\mathsf{H}}^2\big(R_h^M(s_h,a_h),R_h^{\overline{M}}(s_h,a_h)\big)\right].
\end{aligned}\tag{51}$$

根据 引理 E.2 以及 $\pi^\star$，对于每个 $h\in[H]$, $s\in\mathcal{S}$, $a\in\mathcal{A}$，有

$$\mathbb{E}_{\pi\sim p}\left[d_h^{\overline{M},\pi}(s,a)\right]=d_h^{\overline{M},\pi^\star}(s,a).$$

因此，我们可以替换对 $\pi\sim p$ 的期望，$(s_h,a_h)\sim M(\pi)$在 (51)中，取 $(s_h,a_h)\sim M(\pi^\star)$，并得到

$$\begin{aligned}
&\mathbb{E}_{\pi\sim p}\left[D_{\mathsf{H}}^2\big(M(\pi),\overline{M}(\pi)\big)\right]\\
\le& C\log(H)\cdot\left(D_{\mathsf{H}}^2\big(\mu^M,\mu^{\overline{M}}\big)+\mathbb{E}^{\overline{M},\pi^\star}\left[\sum_{h=1}^{H-1}D_{\mathsf{H}}^2\big(P_h^M(s_h,a_h),P_h^{\overline{M}}(s_h,a_h)\big)\right]\right.\\
&\left.+\mathbb{E}^{\overline{M},\pi^\star}\left[\sum_{h=1}^{H}D_{\mathsf{H}}^2\big(R_h^M(s_h,a_h),R_h^{\overline{M}}(s_h,a_h)\big)\right]\right).
\end{aligned}$$

作者：Foster 等人 （2021，引理 A.9）以及数据处理不等式，我们有 t:

$$\begin{aligned}\mathbb{E}^{\overline{M},\pi^\star}\left[\sum_{h=1}^{H-1} D^2_{\mathsf{H}}\big(P^M_h(s_h,a_h), P^{\overline{M}}_h(s_h,a_h)\big)\right] \leq & 4H\cdot D^2_{\mathsf{H}}\big(M(\pi^\star),\overline{M}(\pi^\star)\big),\\ \mathbb{E}^{\overline{M},\pi^\star}\left[\sum_{h=1}^{H} D^2_{\mathsf{H}}\big(R^M_h(s_h,a_h), R^{\overline{M}}_h(s_h,a_h)\big)\right] \leq & 4H\cdot D^2_{\mathsf{H}}\big(M(\pi^\star),\overline{M}(\pi^\star)\big),\\ D^2_{\mathsf{H}}\big(\mu^M,\mu^{\overline{M}}\big) \leq & D^2_{\mathsf{H}}\big(M(\pi^\star),\overline{M}(\pi^\star)\big).\end{aligned}$$

由此可得，对于某个常数 $C>0$,

$$\mathbb{E}_{\pi\sim p}\left[D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right] \leq C\cdot H\log(H)\cdot D^2_{\mathsf{H}}\big(M(\pi^\star),\overline{M}(\pi^\star)\big),$$

正如所期望的那样。 □

引理 E.4. 存在一个常数 $C>0$，使得以下结论成立。考虑任意模型类 $\mathcal{M}$，其中包含时域固定为 H、有限状态空间 $\mathcal{S}$、有限动作空间 $\mathcal{A}$且累积奖励有界于$[0,1]$的马尔可夫决策过程。令 Π_{RNS} 为随机非平稳策略类。考虑任意 $M\in\mathcal{M}$ 和有限支撑集分布 $p\in\Delta(\Pi_{\mathrm{RNS}})$，并令 $\pi^\star\in\Pi_{\mathrm{RNS}}$ 表示任意满足–式 （*50*） 的策略，其中 M 且 p。则对任意 $\gamma>0$,

$$\begin{aligned}&\sup_{M\in\mathcal{M}}\left\{f^M(\pi_M)-f^M(\pi^\star)-\gamma\cdot D^2_{\mathsf{H}}\big(M(\pi^\star),\overline{M}(\pi^\star)\big)\right\}\\ \leq&\frac{CH\log H}{\gamma}+\sup_{M\in\mathcal{M}}\mathbb{E}_{\pi\sim p}\left[f^M(\pi_M)-f^M(\pi)-\frac{\gamma}{CH\log H}\cdot D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right].\end{aligned}\tag{52}$$

由此直接得出引理 E.4的结论是

$$\begin{aligned}&\inf_{\pi\in\Pi_{\mathrm{RNS}}}\sup_{M\in\mathcal{M}}\left\{f^M(\pi_M)-f^M(\pi)-\gamma\cdot D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right\}\\ \leq&\frac{CH\log H}{\gamma}+\inf_{p\in\Delta(\Pi_{\mathrm{RNS}})}\sup_{M\in\mathcal{M}}\mathbb{E}_{\pi\sim p}\left[f^M(\pi_M)-f^M(\pi)-\frac{\gamma}{CH\log H}\cdot D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right].\end{aligned}$$

证明 引理 E.4。 考虑任意 $M\in\mathcal{M}$、有限支撑集的 $p\in\Delta(\Pi_{\mathrm{RNS}})$，并令 $\pi^\star$ 按引理陈述中的定义给出。引理 E.2 给出 $f^M(\pi^\star)=\mathbb{E}_{\pi\sim p}\left[f^M(\pi)\right]$。令 C为引理 E.3中的常数，并令 $C'=C+\frac{1}{2}$。则对任意 $\gamma>0$,

$$\begin{aligned}&\sup_{M\in\mathcal{M}} f^M(\pi_M)-f^M(\pi^\star)-C'H\log H\cdot\gamma\cdot D^2_{\mathsf{H}}\big(M(\pi^\star),\overline{M}(\pi^\star)\big)\\ \leq&\sup_{M\in\mathcal{M}} f^M(\pi_M)-f^{\overline{M}}(\pi^\star)-C'H\log H\cdot\gamma\cdot D^2_{\mathsf{H}}\big(M(\pi^\star),\overline{M}(\pi^\star)\big)+\frac{1}{2\gamma}+\frac{\gamma}{2}\cdot D^2_{\mathsf{H}}\big(M(\pi^\star),\overline{M}(\pi^\star)\big)\\ =&\sup_{M\in\mathcal{M}} f^M(\pi_M)-f^{\overline{M}}(\pi^\star)-CH\log H\cdot\gamma\cdot D^2_{\mathsf{H}}\big(M(\pi^\star),\overline{M}(\pi^\star)\big)+\frac{1}{2\gamma}\\ \leq&\sup_{M\in\mathcal{M}}\mathbb{E}_{\pi\sim p}\left[f^M(\pi_M)-f^{\overline{M}}(\pi)-\gamma\cdot D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right]+\frac{1}{2\gamma}\qquad(53)\\ \leq&\sup_{M\in\mathcal{M}}\mathbb{E}_{\pi\sim p}\left[f^M(\pi_M)-f^M(\pi)-\gamma\cdot D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)+\frac{1}{2\gamma}+\frac{\gamma}{2}\cdot D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right]+\frac{1}{2\gamma}\\ =&\sup_{M\in\mathcal{M}}\mathbb{E}_{\pi\sim p}\left[f^M(\pi_M)-f^M(\pi)-\frac{\gamma}{2}\cdot D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right]+\frac{1}{\gamma}.\qquad(54)\end{aligned}$$

其中（53）使用引理 ma E.3。将 γ 替换为 $\gamma\cdot C'H\log H$，即可得到该命题的结论。 □

第 F 节 第 5 节上界证明

本节证明定理 5.1，该定理为 MA-DMSO 框架中的 CCE 和 CE 实例学习均衡给出了一个上界，同时避免了多智能体诅咒，即避免了随着玩家数量 K 呈指数级增长。在附录 F.1中，我们介绍用于证明算法 1（定理 5.1）所采用的算法。该算法基于探索即优化这一思想，此思想此前已见于Foster 等人 (2022b)；Lattimore 和 György (2021)。在附录 F.2 至F.4 中，我们证明定理 5.1；我们的证明大体遵循Foster 等人 (2022b) 的证明，但需要进行一些细微的修改，以适应我们问题的多智能体特性，以及我们所研究的更一般的偏离集合概念 Π'_k。

F.1 多智能体优化式探索目标

我们首先介绍用于证明多智能体探索即优化(MAExO；算法 1)，定理 5.1的算法。该算法是Foster 等人+提出的探索即优化（ExO+算法的多智能体对应版本。从整体上看，MAExO （以及其前身ExO+）是EXP3的一种变体，它将指数权重算法应用于一系列奖励估计器，这些估计器充当真实奖励函数的重要性加权估计。不过，与EXP3 和ExO+不同，MAExO 并不将指数权重应用于智能体自身的纯策略，而是应用于它们可能采取的偏离 Π'_k。

具体而言，MAExO 在与环境交互的 T 轮次中运行。在每一轮 $t \in [T]$，算法首先根据此前轮次中由该算法构造的一系列向量 f_k^1、……、f_k^{t-1}，通过指数权重更新计算每位玩家 k 在其偏离空间上 Π'_k上的参考分布 $q_k^t \in \Delta(\Pi'_k)$（第 4 行）。粗略地说，对于 $s \le t-1$，这些向量中各项 $f_k^s(\pi'_k)$、$\pi'_k \in \Pi'_k$ 可解释为：在其他智能体的决策遭遇对抗性选择的情况下，智能体 k 偏离至 π'_k 所能获得的潜在价值增益。因此，参考分布 q_k^t 会将更多概率质量分配给能带来更大价值增益的偏离。

接下来，在第 5 行，玩家共同求解一个优化问题。为了定义这个优化问题，我们先引入一些记号。对于每个 $k \in [K]$，令 $\mathcal{G}_k$ 表示所有函数 $g_k : \Pi'_k \times \Sigma \times \mathcal{O} \to \mathbb{R}$的集合，并令 $\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_K$。给定 $q \in \prod_{k=1}^K \Delta(\Pi'_k)$，$\eta > 0$，$\pi \in \Pi$，$g \in \mathcal{G}$，$\pi^\star = (\pi_1^\star, \ldots\ldots, \pi_K^\star) \in \prod_{k=1}^K \Pi'_k$，以及$M \in \mathcal{M}$，定义

$$\begin{aligned}\Gamma_{q,\eta}(\pi, g; \pi^\star, M) :=& \mathbb{E}_{\sigma\sim\pi}\left[\sum_{k=1}^K f_k^M(U_k(\pi_k^\star, \sigma)) - f_k^M(\sigma)\right] \\ &+ \frac{1}{\eta}\cdot\sum_{k=1}^K \mathbb{E}_{\sigma\sim\pi, o\sim M(\sigma)}\mathbb{E}_{\pi'_k\sim q_k}\left[\exp\left(\frac{\eta}{\pi(\sigma)}\cdot(g_k(\pi'_k;\sigma,o) - g_k(\pi_k^\star;\sigma,o))\right) - 1\right].\end{aligned} \tag{55}$$

根据这个定义，MAExO 第5 行中求解的优化问题如下 s:

$$(\pi^t, g^t) \leftarrow \operatorname*{arg\,min}_{\pi\in\Pi, g\in\mathcal{G}} \sup_{M\in\mathcal{M}, \pi^\star\in\prod_{k=1}^K \Pi'_k} \Gamma_{q^t,\eta}(\pi, g; \pi^\star, M). \tag{56}$$

目标函数 (55) 和优化问题 (56) 的含义如下。粗略地说，对于每个 $k \in [K], \pi'_k \in \Pi'_k, \sigma \in \Sigma, o \in \mathcal{O}$，对于 $g \in \mathcal{G}_k$，取值 $g_k(\pi'_k; \sigma, o)$ 可以解释为：在未知模型 M 下，玩家 k 偏离联合决策剖面 σ、改为选择 π'_k 所带来的价值收益估计量；该未知模型与决策—观测对 (σ, o)“一致”。随后，通过求解 (56)，算法希望找到一个联合决策 $\pi^t \in \Pi$ 和估计量 $g^t = (g_1^t, \ldots, g_K^t) \in \mathcal{G}$，使得对于每个玩家 $k \in [K]$，都满足以下两个性质：

- 首先，对应于(55)中的第一项，对于最坏情况下的未知模型 M 和未知偏差 $\pi_k^\star$，玩家 k 在给定策略 $\pi_k^\star$ 时，不应能通过偏离至 π^t而获得太多取值。这里 $\pi_k^\star$ 应解释为算法终止时事后最佳偏离。

- 其次，对应于 (55) 中的第二项：π^t 和 g_k^t 应当这样选择，使得在 $\sigma \sim \pi^t$ 下以高概率，g_k^t 不会低估偏离到 $\pi_k^\star$ 相较于选择一个

从参考分布 π'_k 中抽取的样本 q^t_k。式 (55) 中的第二项可以看作是对对抗性选择 $\pi^\star_k$的正则化项，类似于偏移 DEC 中减去平方海林格距离的那一项（见 (11)）：具体而言，如果 $\pi^\star_k$ 在估计值 g_k下的取值显著偏高，那么这一项就会非常负，从而抵消第一项（可能很大）的影响。

给定在公式 (56) 中计算的 (π^t, g^t)，算法 1 会采样一个决策 $\sigma^t \sim \pi^t$，并从真实模型中获得观测 o^t。最后，在第 7 行，玩家使用 $g^t_k(\cdot; \sigma^t, o^t)$ 构造奖励估计器 f^t_k（供未来迭代 $t' > t$ 使用，以构造 $q^{t'}_k$）。所有 T 轮次结束后，该算法输出联合决策 $\widehat{\pi}$，即对 T 个纯决策 $\sigma^1, \ldots, \sigma^T$ 求得的均匀平均。需要指出的是，算法 1 不同于让每位玩家分别运行 Foster 等人（2022b）的探索-优化算法：在后者中，每个智能体单独优化自己的目标；而式 (56) 中的优化问题则是为所有智能体同时求解的。MAExO 的这一特性使我们能够获得按 r-dec$^{\mathrm{o}}_\gamma$(co($\mathscr{M}$)) 缩放的保证，而该量可以任意小于采用 Foster 等人（2022b）的方法所得的量（见命题 A.11）。

在定义 F.1 中，我们将形式化定义极小极大目标 (56) 在算法 1执行过程中计算得到的取值。

算法 1 通过优化进行多智能体探索（MAExO）

1: **参数**：学习率 $\eta > 0$。2: 初始化 $f^0_k(\pi'_k) := 0$ 对于所有 $k \in [K]$、$\pi'_k \in \Pi'_k$。3: **对于** $t = 1, 2, \ldots T$**执行**4: 对于每个智能体 $k \in [K]$，通过指数权重更新定义 $q^t_k \in \Delta(\Pi'_k)$：对于 $\pi'_k \in \Pi'_k$,

$$q^t_k(\pi'_k) := \frac{\exp\left(\eta \sum_{i=1}^{t-1} \widehat{f}^i_k(\pi'_k)\right)}{\sum_{\pi''_k \in \Pi'_k} \exp\left(\eta \sum_{i=1}^{t-1} \widehat{f}^i_k(\pi''_k)\right)}.$$

5: 定义 $q^t = q^t_1 \times \cdots \times q^t_K$。玩家共同求解以下目标: // Eq. (55)

$$(\pi^t, g^t) \leftarrow \arg\min_{\pi \in \Pi, g \in \mathcal{G}} \sup_{M \in \mathcal{M}, \pi^\star \in \Pi'} \Gamma_{q^t, \eta}(\pi, g; \pi^\star, M).$$

6: 采样 $\sigma^t \sim \pi^t$，进行 σ^t，并观测 $o^t \sim M^\star(\sigma^t)$。

7: 每位玩家 $k \in [K]$ 按如下方式构建其奖励估计量 f^t_k：对于 $\pi'_k \in \Pi'_k$,

$$\widehat{f}^t_k(\pi'_k) = \frac{g^t_k(\pi'_k; \sigma^t, o^t)}{\pi^t(\sigma^t)}.$$

8: **返回** 联合决策 $\widehat{\pi} := \frac{1}{T} \sum_{t=1}^T \mathbb{I}_{\sigma^t}$。

定义 F.1（通过优化进行探索的目标）。考虑任意满足 $\mathscr{M} = (\mathcal{M}, \Pi, \mathcal{O}, \{\Pi'_k\}_k, \{U_k\}_k)$ 的实例 假设 5.1。对于任意尺度参数 $\eta > 0$ 和分布 $q \in \prod_{k=1}^K \Delta(\Pi'_k)$，定义

$$\mathsf{exo}_\eta(\mathscr{M}, q) = \inf_{\pi \in \Pi, g \in \mathcal{G}} \sup_{M \in \mathcal{M}, \pi^\star \in \prod_{k=1}^K \Pi'_k} \Gamma_{q, \eta}(\pi, g; \pi^\star, M),$$

并令 $\mathsf{exo}\eta(\mathscr{M}) := \sup_{q \in \prod_{k=1}^K \Delta(\Pi'_k)} \mathsf{exo}\eta(\mathscr{M}, q)$。

为证明 定理 5.1，我们首先（附录 F.2）以 算法 1 的 $\mathsf{exo}\eta(\mathscr{M})$来界定其性能。随后，在附录 F.3 和 附录 F.4中，我们将利用一个称为$\mathsf{exo}\eta(\mathscr{M})$ 的dec$^{\mathrm{o}}_\gamma(\mathscr{M})$ 上界，并适当选择 γ，以多智能体（参数化）信息比 作为中间量。最后，在附录 F.5中，我们将这些结果结合起来并证明定理 5.1。

F.2 界定算法1的性能

以下结果以算法 1 的性能（即数量 $h^{M^\star}(\widehat{\pi})$）来界定其$\mathsf{exo}_\eta(\mathscr{M})$。

引理 F. **1.** 对于任意 $\eta > 0$，*算法 1* 确保对于所有 $\delta > 0$，以至少 $1-\delta$的概率 ，

$$h^{M^\star}(\widehat{\pi}) = \sum_{k=1}^{K} \max_{\pi'_k \in \Pi'_k} f_k^{M^\star}(U_k(\pi'_k, \widehat{\pi})) - f_k^{M^\star}(\widehat{\pi}) \le \mathsf{exo}_\eta(\mathscr{M}) + \frac{2}{T\eta} \cdot \sum_{k=1}^{K} \log\left(\frac{K \cdot |\Pi'_k|}{\delta}\right).$$

引理引理 F.1的证明。对于任意 $\pi_k^\star \in \Pi'_k$ 和玩家 $k \in [K]$，我们将玩家 k的遗憾相对于偏差 $\pi_k^\star \in \Pi'_k$ 定义如下:

$$\mathrm{Reg}_k(\pi_k^\star) = \sum_{t=1}^{T} \mathbb{E}_{\sigma^t \sim \pi^t}[f_k^{M^\star}(U_k(\pi_k^\star, \sigma^t)) - f_k^{M^\star}(\sigma^t)] = T \cdot \left(f_k^{M^\star}(U_k(\pi_k^\star, \widehat{\pi})) - f_k^{M^\star}(\widehat{\pi})\right),$$

其中，上面的第二个等式使用了 $\widehat{\pi}$ 在第 8 行的算法 1 中的定义以及假设 5.1中的第二个性质。因此，只需界定 $\frac{1}{T} \cdot \sum_{k=1}^{K} \max^{\pi_k^\star \in \Pi'_k} \mathrm{Reg}_k(\pi_k^\star)$ 即可证明引理中的结论。

在整个证明中，我们采用以下约定：对于函数 $f_k:\ \Pi'_k \to \mathbb{R}$ （例如，奖励估计器 $\widehat{f}_k^t$ 定义于第 7 行的算法 1），我们将 f_k 视为 $\mathbb{R}^{|\Pi'_k|}$中的一个向量，其坐标为 $f_k(\pi'_k)$的取值，其中 $\pi'_k \in \Pi'_k$。此外，对于每个 $\pi'_k \in \Pi'_k$，我们记 $e_{\pi'_k} \in \mathbb{R}^{|\Pi'_k|}$ 为对应的单位向量，其 π'_k分量为 1，所有其他分量均为 0。

通过加上并减去 $\sum_{t=1}^{T} \langle e_{\pi_k^\star}, \widehat{f}_k^t \rangle$，我们得到

$$\begin{aligned}
\mathrm{Reg}_k(\pi_k^\star) &= \sum_{t=1}^{T} \mathbb{E}_{\sigma^t \sim \pi^t}[f_k^{M^\star}(U_k(\pi_k^\star, \sigma^t)) - f_k^{M^\star}(\sigma^t)] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\sigma^t \sim \pi^t}[f_k^{M^\star}(U_k(\pi_k^\star, \sigma^t)) - f_k^{M^\star}(\sigma^t)] + \sum_{t=1}^{T} \langle e_{\pi_k^\star}, \widehat{f}_k^t \rangle - \sum_{t=1}^{T} \langle e_{\pi_k^\star}, \widehat{f}_k^t \rangle. \qquad (57)
\end{aligned}$$

根据 引理 F.2 以及 q_k^t 在 第 4 行 的 算法 1中乘法权重更新的定义，可得

$$\begin{aligned}
&\sum_{t=1}^{T} \langle e_{\pi_k^\star}, \widehat{f}_k^t \rangle \\
&\le \sum_{t=1}^{T} \langle q_k^{t+1}, \widehat{f}_k^t \rangle - \frac{1}{\eta} \sum_{t=1}^{T} D_{\mathsf{KL}}(q_k^{t+1} \,\|\, q_k^t) + \frac{1}{\eta} D_{\mathsf{KL}}\left(e_{\pi_k^\star} \,\|\, q_k^1\right) \\
&\le \sum_{t=1}^{t} \langle q_k^{t+1}, \widehat{f}_k^t \rangle - \frac{1}{\eta} \sum_{t=1}^{T} D_{\mathsf{KL}}(q_k^{t+1} \,\|\, q_k^t) + \frac{\log |\Pi'_k|}{\eta}. \qquad (58)
\end{aligned}$$

根据 引理 B.5，对于每个 $t \in [T]$,

$$\langle q_k^{t+1}, \widehat{f}_k^t \rangle - \frac{1}{\eta} D_{\mathsf{KL}}(q_k^{t+1} \,\|\, q_k^t) \le \frac{1}{\eta} \log\left(\sum_{\pi'_k \in \Pi'_k} q_k^t(\pi'_k) \cdot \exp(\eta \cdot \widehat{f}_k^t(\pi'_k))\right).$$

结合上述内容以及（58）和（57），我们得到

$$\begin{aligned}
\mathrm{Reg}_k(\pi_k^\star) \le &\sum_{t=1}^{T} \mathbb{E}_{\sigma^t \sim \pi^t}[f_k^{M^\star}(U_k(\pi_k^\star, \sigma^t)) - f_k^{M^\star}(\sigma^t)] + \frac{1}{\eta} \log\left(\sum_{\pi'_k \in \Pi'_k} q_k^t(\pi'_k) \cdot \exp(\eta \cdot \widehat{f}_k^t(\pi'_k))\right) \\
&+ \frac{\log |\Pi'_k|}{\eta} - \sum_{t=1}^{T} \langle e_{\pi_k^\star}, \widehat{f}_k^t \rangle. \qquad (59)
\end{aligned}$$

令 $\mathscr{F}^t$ 表示由 σ-代数生成的（σ^1，o^1，……，σ^t，o^t）（其中随机变量 σ^s，o^s 按照算法 1中的方式抽取）。注意，$\mathscr{F}^t$ 是一个滤子，并记 $\mathbb{E}_t[\cdot]=\mathbb{E}[\cdot|\mathscr{F}^t]$。对于每个 $\pi_k^\star\in\Pi'_k$，我们定义一个随机变量序列，记为 $\{X_t(\pi_k^\star)\}_{t\in[T]}$，定义如下：

$$X_t(\pi_k^\star):=\log\left(\sum_{\pi'_k\in\Pi'_k}q_k^t(\pi'_k)\cdot\exp(\eta\cdot\widehat{f}_k^t(\pi'_k))\right)-\langle e_{\pi_k^\star},\eta\cdot\widehat{f}_k^t\rangle.$$

根据引理 B.7 以及并集界，以至少 $1-\delta/K$ 的概率，对所有 $\pi_k^\star\in\Pi'_k$，均有，

$$\sum_{t=1}^T X_t(\pi_k^\star)\le\sum_{t=1}^T\log\mathbb{E}_{t-1}[e^{X_t(\pi_k^\star)}]+\log\left(\frac{K\cdot|\Pi'_k|}{\delta}\right). \tag{60}$$

注意，π^t，q^t 均相对于 $\mathscr{F}^{t-1}$ 可测。于是，对于任意 $\pi_k^\star\in\Pi'_k$ 和任意 $t\in[T]$，我们可以计算

$$\begin{aligned}
&\log\mathbb{E}_{t-1}[e^{X_t(\pi_k^\star)}]\\
&=\log\mathbb{E}_{t-1}\left[\exp\left(\log\left(\sum_{\pi'_k\in\Pi'_k}q_k^t(\pi'_k)\cdot\exp(\eta\widehat{f}_k^t(\pi'_k))\right)-\eta\widehat{f}_k^t(\pi_k^\star)\right)\right]\\
&=\log\mathbb{E}_{\sigma^t\sim\pi^t}\mathbb{E}_{o^t\sim M(\sigma^t)}\left[\mathbb{E}_{\pi'_k\sim q_k^t}\left[\exp(\eta\widehat{f}_k^t(\pi'_k))\right]\cdot\exp(-\eta\widehat{f}_k^t(\pi_k^\star))\right]\\
&=\log\mathbb{E}_{\sigma^t\sim\pi^t}\mathbb{E}_{o^t\sim M(\sigma^t)}\left[\mathbb{E}_{\pi'_k\sim q_k^t}\left[\exp\left(\frac{\eta}{\pi^t(\sigma^t)}\cdot g_k^t(\pi'_k;\sigma^t,o^t)\right)\right]\cdot\exp\left(-\frac{\eta}{\pi^t(\sigma^t)}\cdot g_k^t(\pi_k^\star;\sigma^t,o^t)\right)\right]\\
&\le\mathbb{E}_{\sigma^t\sim\pi^t}\mathbb{E}_{o^t\sim M(\sigma^t)}\mathbb{E}_{\pi'_k\sim q_k^t}\left[\exp\left(\frac{\eta}{\pi^t(\sigma^t)}\cdot(g_k^t(\pi'_k;\sigma^t,o^t)-g_k^t(\pi_k^\star;\sigma^t,o^t))\right)\right]-1,
\end{aligned}\tag{61}$$

其中，最后一个不等式利用了对所有 $x>0$，均有 $\log(x)\le x-1$。

根据 (59)、(60)、以及 (61)，并对 $k\in[K]$ 使用并集界，可得以至少 $1-\delta$ 的概率，对所有$\pi_1^\star\in\Pi'_1$、……、$\pi_K^\star\in\Pi'_K$，令 $\pi^\star=(\pi_1^\star$、……、$\pi_K^\star)$，

$$\begin{aligned}
&\sum_{k=1}^K\mathrm{Reg}_k(\pi_k^\star)\\
&\le\sum_{k=1}^K\left(\frac{\log|\Pi'_k|}{\eta}+\frac{\log\left(\frac{K\cdot|\Pi'_k|}{\delta}\right)}{\eta}\right)+\sum_{t=1}^T\sum_{k=1}^K\mathbb{E}_{\sigma^t\sim\pi^t}\left[f_k^{M^\star}(U_k(\pi_k^\star,\sigma^t))-f_k^{M^\star}(\sigma^t)\right]\\
&\quad+\sum_{t=1}^T\sum_{k=1}^K\frac{1}{\eta}\left(\mathbb{E}_{\sigma^t\sim\pi^t}\mathbb{E}_{o^t\sim M(\sigma^t)}\mathbb{E}_{\pi'_k\sim q_k^t}\left[\exp\left(\frac{\eta}{\pi^t(\sigma^t)}\cdot(g_k^t(\pi'_k;\sigma^t,o^t)-g_k^t(\pi_k^\star;\sigma^t,o^t))\right)\right]-1\right)\\
&\le\frac{2}{\eta}\cdot\sum_{k=1}^K\log\left(\frac{K\cdot|\Pi'_k|}{\delta}\right)+\sum_{t=1}^T\Gamma_{q^t,\eta}(\pi^t,g^t;\pi^\star,M^\star)\\
&\le\frac{2}{\eta}\cdot\sum_{k=1}^K\log\left(\frac{K\cdot|\Pi'_k|}{\delta}\right)+\sum_{t=1}^T\sup_{\widetilde{\pi}^\star\in\Pi',\widetilde{M}\in\mathcal{M}}\Gamma_{q^t,\eta}(\pi^t,g^t;\widetilde{\pi}^\star,\widetilde{M})\\
&\le\frac{2}{\eta}\cdot\sum_{k=1}^K\log\left(\frac{K\cdot|\Pi'_k|}{\delta}\right)+T\cdot\mathsf{exo}_\eta(\mathscr{M}),
\end{aligned}$$

其中，第二个不等式利用了$\Gamma_{q^t\text{的定义},\,\eta}(\pi^t,\ g^t;\ \pi^\star,\ M^\star)$ 中的(55)，而最后一个等式成立，是因为 π^t，g^t 的选取旨在最小化多智能体优化式探索目标（算法 1 的第 5 行）。

□

引理 F.2. 考虑任意的 $d \in \mathbb{N}$，并令 $f^1, \ldots, f^T \in \mathbb{R}^d$ 为任意向量序列。对于 $\eta > 0$，令 $q^1, \ldots, q^T \in \Delta^d$ 表示步长为 η、奖励向量依次为 $f^1, \ldots, f^T$ 时的指数权重更新迭代；特别地，对于 $t \in [T]$:

$$q^t(i) = \frac{\exp(\eta \sum_{s \le t} f^s(i))}{\sum_{j=1}^d \exp(\eta \sum_{s \le t} f^s(j))}. \tag{62}$$

那么，对于任意 $q \in \Delta^d$,

$$\sum_{t=1}^T \langle q, f^t \rangle \le \sum_{t=1}^T \langle q^{t+1}, f^t \rangle - \frac{1}{\eta} \sum_{t=1}^T D_{\mathsf{KL}}(q^{t+1} \,\|\, q^t) + \frac{1}{\eta} D_{\mathsf{KL}}(q \,\|\, q^1).$$

证明 引理 F.2。通过重新整理并进行裂项，只需证明，对于每个 $t \in [T]$，

$$\langle q - q^{t+1}, f^t \rangle = \frac{1}{\eta} \cdot \left(D_{\mathsf{KL}}(q \,\|\, q^t) - D_{\mathsf{KL}}(q \,\|\, q^{t+1}) - D_{\mathsf{KL}}(q^{t+1} \,\|\, q^t) \right).$$

为证明这个不等式，我们注意到，乘法权重更新（62）等价于以下使用负熵正则项 $\Phi(q) = \sum_{i=1}^d q_i \cdot \log q_i$ 的镜像下降更新:

$$\nabla \Phi(p^{t+1}) = \nabla \Phi(q^t) + \eta \cdot f^t, \quad q^{t+1} = \frac{p^{t+1}}{\langle \mathbf{1}, p^{t+1} \rangle},$$

其中 $\mathbf{1} \in \mathbb{R}^d$ 表示全 1 向量。利用对于所有 $x, y, z \in \Delta^d$ 的 Bubeck 等式（4.1）（Bubeck（2015年））

$$\langle \nabla \Phi(y) - \nabla \Phi(x), x - z \rangle = D_{\mathsf{KL}}(z \,\|\, y) - D_{\mathsf{KL}}(z \,\|\, x) - D_{\mathsf{KL}}(x \,\|\, y)$$

对于 $z = q$, $y = q^t$, $x = q^{t+1}$，我们得到

$$\begin{aligned}
\frac{1}{\eta} \cdot \left(D_{\mathsf{KL}}(q \,\|\, q^t) - D_{\mathsf{KL}}(q \,\|\, q^{t+1}) - D_{\mathsf{KL}}(q^{t+1} \,\|\, q^t) \right) &= \frac{1}{\eta} \cdot \langle \nabla \Phi(q^t) - \nabla \Phi(q^{t+1}), q^{t+1} - q \rangle \\
&= \frac{1}{\eta} \cdot \langle \nabla \Phi(q^t) - \nabla \Phi(p^{t+1}), q^{t+1} - q \rangle \\
&= \langle f^t, q - q^{t+1} \rangle,
\end{aligned}$$

其中，在第二个等式中，我们用到了 $\nabla \Phi(q^{t+1}) = \nabla \Phi(p^{t+1}) - \log\left(\langle \mathbf{1}, p^{t+1} \rangle\right) \cdot \mathbf{1}$。 □

F.3 多智能体参数化信息比

本节中，我们介绍 (Foster 等人，2022b，定义 3.1) 中参数化信息比的多智能体版本，并以上界为凸包 $\mathscr{M}$ 的 DEC 来约束该信息比。下一节中，我们将以上界为该信息比来约束 $\mathsf{exo}^\eta(\mathscr{M})$。

我们先引入一些记号。我们需要讨论 $\mathcal{M} \times \Pi'_1 \times \cdots \times \Pi'_K$ 上概率测度的空间。由于 $|\mathcal{M}|$ 可能是无限的，为避免测度论问题，我们稍微滥用记号，用 $\Delta(\mathcal{M} \times \Pi'_1 \times \cdots \times \Pi'_K)$ 表示 $\mathcal{M} \times \Pi'_1 \times \cdots \times \Pi'_K$ 上有限支撑的概率测度集合。这一约定确保对于任意函数 $h: \mathcal{M} \times \Pi'_1 \times \cdots \times \Pi'_K \to \mathbb{R}$ 以及任意 $\mu \in \Delta(\mathcal{M} \times \Pi'_1 \times \cdots \times \Pi'_K)$，$\mathbb{E}_{(M, \pi'_1, \ldots, \pi'_K) \sim \mu}[h(M, \pi'_1, \ldots\ldots, \pi'_K)]$ 都是良定义的。

考虑任意 $k \in [K]$、分布 $\mu \in \Delta(\mathcal{M} \times \Pi'_1 \times \cdots \times \Pi'_K)$ 和分布 $\pi \in \Delta(\Sigma)$。令 $\mathbb{P}$ 表示过程 $(M, \pi^\star_1, \ldots\ldots, \pi^\star_K) \sim \mu$、$\sigma \sim \pi$ 和 $o \sim M(\sigma)$ 的概率律。我们引入以下分布，它们取决于 μ 和 k:

- 定义分布 $\mu^k{}_{\mathrm{pr}} \in \Delta(\Pi'_k)$，使得 $\mu^k{}_{\mathrm{pr}}(\pi'_k) = \mathbb{P}(\pi^\star_k = \pi'_k)$，其中 $\pi'_k \in \Pi'_k$。
- 对于每个 $\sigma \in \Sigma$ 和 $o \in \mathcal{O}$，定义分布 $\mu^k{}_{\mathrm{po}} \in \Delta(\Pi'_k)$ 为 $\mu^k{}_{\mathrm{po}}(\pi'_k; \sigma, o) = \mathbb{P}(\pi^\star_k = \pi'_k | (\sigma, o))$，其中 $\pi'_k \in \Pi'_k$。

应将分布 μ_{pr}^k 视为关于偏差 $\pi_k^\star$的先验分布，而分布$\mu_{\mathrm{po}}^k(\cdot;\sigma,o)$ 应视为在观测到纯决策 σ 以及观测结果 $o\sim M(\sigma)$ 后关于 $\pi_k^\star$的后验分布。

定义 F.2 （多智能体信息比率）。 给定一个实例 $\mathscr{M}=(\mathcal{M},\Pi,\mathcal{O},\{\Pi'_k\}_k,\{U_k\}_k)$，它是一个广义相关均衡实例，该实例$\mathscr{M}$ 的参数化多智能体信息比率 定义为

$$\begin{aligned}\mathsf{infr}_\gamma(\mathscr{M}):=\sup_{\mu\in\Delta(\mathcal{M}\times\Pi'_1\times\cdots\times\Pi'_K)}\inf_{\pi\in\Pi}\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{(M,\pi_1^\star,\ldots,\pi_K^\star)\sim\mu}\left[\sum_{k=1}^K f_k^M(U_k(\pi_k^\star,\sigma))-f_k^M(\sigma)\right]\\ -\gamma\cdot\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}\left[\sum_{k=1}^K D^2_{\mathsf{H}}\left(\mu_{\mathrm{po}}^k(\cdot;\sigma,o),\mu_{\mathrm{pr}}^k(\cdot)\right)\right].\end{aligned}$$

在上述表达式中，当我们写作 $U_k(\pi_k^\star,\sigma)$ 和 $f_k^M(\sigma)$ 时，我们将 $\sigma\in\Sigma$视为 Π 的一个元素，方法是将其与 σ 上的单点分布关联起来，回顾一下 $\Pi=\Delta(\Sigma)$。

引理 F.3 用给定实例凸包的多智能体偏移 DEC 给出了多智能体信息比率的上界。

引理 F.3。 考虑任意满足 $\mathscr{M}=(\mathcal{M},\Pi,\mathcal{O},\{\Pi'_k\}_k,\{U_k\}_k)$ 的实例 假设 *5.1*，且满足 $\mathrm{co}(\mathscr{M})$ 满足 假设 *1.3*。那么对所有 $\gamma>0$,

$$\mathsf{infr}_\gamma(\mathscr{M})\le K\cdot\mathsf{r}\text{-}\mathsf{dec}^{\mathrm{o}}_\gamma(\mathrm{co}(\mathscr{M})).$$

引理 F.3 的证明。我们将实例 $\mathscr{M}$ 的纯决策集记为 $\Sigma_1,\ldots,\Sigma_K$，并将联合决策集记为 $\Sigma=\Sigma_1\times\cdots\times\Sigma_K$。固定一个先验 $\mu\in\Delta(\mathcal{M}\times\Pi'_1\times\cdots\times\Pi'_K)$和一个分布 $\pi\in\Delta(\Sigma)$。回顾上文的记号：令 $\mathbb{P}$ 表示过程 $\sigma\sim\pi,(M,\pi_1^\star,\ldots,\pi_K^\star)\sim\mu,o\sim M(\sigma)$ 的概率律。对于每个 $k\in[K]$，令 $\mu_{\mathrm{pr}}^k(\pi'_k)=\mathbb{P}(\pi_k^\star=\pi'_k)$，并令 $\mu_{\mathrm{po}}^k(\pi'_k;\sigma,o)=\mathbb{P}(\pi_k^\star=\pi'_k|(\sigma,o))$。

考虑给定 μ，π 的选择时，多智能体信息比率的取值：

$$\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{(M,\pi_1^\star,\ldots,\pi_K^\star)\sim\mu}\left[\sum_{k=1}^K f_k^M(U_k(\pi_k^\star,\sigma))-f_k^M(\sigma)\right]-\gamma\cdot\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}\left[\sum_{k=1}^K D^2_{\mathsf{H}}\left(\mu_{\mathrm{po}}^k(\cdot;\sigma,o),\mu_{\mathrm{pr}}^k(\cdot)\right)\right].$$

对于每个 $k\in[K]$、$\pi'_k\in\Pi'_k$和 $\pi\in\Pi$，定义$\overline{M}^k_{\pi'_k}(\pi):=\mathbb{E}_\mu[M(\pi)|\pi_k^\star=\pi'_k]$。进一步定义 $\overline{M}(\pi)=\mathbb{E}_\mu[M(\pi)]$。注意，$M^k_{\pi'_k}(\sigma)=\mathbb{P}_{o|\sigma,\ \pi'_k}$且 $M(\sigma)=\mathbb{P}_{o|\sigma}$。

接下来，注意对于每个固定的 $\sigma\in\Sigma$,

$$\begin{aligned}\mathbb{E}_{o|\sigma}\left[D^2_{\mathsf{H}}\left(\mu_{\mathrm{po}}^k(\cdot;\sigma,o),\mu_{\mathrm{pr}}(\cdot)\right)\right]=&\mathbb{E}_{o|\sigma}\left[D^2_{\mathsf{H}}\left(\mathbb{P}_{\pi_k^\star|\sigma,o},\mathbb{P}_{\pi_k^\star}\right)\right]\\ =&\mathbb{E}_{o|\sigma}\left[D^2_{\mathsf{H}}\left(\mathbb{P}_{\pi_k^\star|\sigma,o},\mathbb{P}_{\pi_k^\star|\sigma}\right)\right]\\ =&\mathbb{E}_{\pi_k^\star\sim\mu}\left[D^2_{\mathsf{H}}\left(\mathbb{P}_{o|\sigma,\pi_k^\star},\mathbb{P}_{o|\sigma}\right)\right]\\ =&\mathbb{E}_{\pi_k^\star\sim\mu}\left[D^2_{\mathsf{H}}\left(\overline{M}^k_{\pi_k^\star}(\sigma),\overline{M}(\sigma)\right)\right],\end{aligned}\tag{63}$$

其中第二个等式成立，是因为 $\sigma\sim\pi$和 $(\pi_1^\star,\ldots,\pi_K^\star)\sim\mu$彼此（边缘分布）独立，而第三个等式根据 引理 B.2成立。此外，我们还有

$$\begin{aligned}\mathbb{E}_{(M,\pi_1^\star,\ldots,\pi_K^\star)\sim\mu}\mathbb{E}_{\sigma\sim\pi}\left[f_k^M(U_k(\pi_k^\star,\sigma))-f_k^M(\sigma)\right]=&\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{\pi_k^\star\sim\mu}\mathbb{E}_\mu\left[f_k^M(U_k(\pi_k^\star,\sigma))-f_k^M(\sigma)|\pi_k^\star\right]\\ =&\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{\pi_k^\star\sim\mu}\left[f_k^{\overline{M}^k_{\pi_k^\star}}(U_k(\pi_k^\star,\sigma))-f_k^{\overline{M}^k_{\pi_k^\star}}(\pi)\right]\\ \le&\mathbb{E}_{\pi_k^\star\sim\mu}\max_{\pi'_k\in\Pi'_k}\mathbb{E}_{\sigma\sim\pi}\left[f_k^{\overline{M}^k_{\pi_k^\star}}(U_k(\pi'_k,\sigma))-f_k^{\overline{M}^k_{\pi_k^\star}}(\sigma)\right].\end{aligned}\tag{64}$$

接下来，对于任意 $M_1 \in \mathsf{co}(\mathcal{M})$，我们有，对于 $\gamma > 0$,

$$
\begin{aligned}
&\mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathscr{M}), \overline{M}_1) \\
=& \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathrm{co}(\mathcal{M})} \mathbb{E}_{\pi\sim p}\left[\sum_{k=1}^{K}\left(\max_{\pi'_k\in\Pi'_k} f_k^M(U_k(\pi'_k,\pi)) - f_k^M(\pi)\right) - \gamma\cdot D^2_{\mathsf{H}}\big(M(\pi), \overline{M}_1(\pi)\big)\right] \\
=& \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathrm{co}(\mathcal{M})} \mathbb{E}_{\pi\sim p}\left[\sum_{k=1}^{K}\max_{\pi'_k\in\Pi'_k} \mathbb{E}_{\sigma\sim\pi}\left[f_k^M(U_k(\pi'_k,\sigma)) - f_k^M(\sigma)\right] - \gamma\cdot D^2_{\mathsf{H}}\big(\mathbb{E}_{\sigma\sim\pi}[M(\sigma)], \mathbb{E}_{\sigma\sim\pi}[\overline{M}_1(\sigma)]\big)\right] \\
\geq& \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathrm{co}(\mathcal{M})} \mathbb{E}_{\pi\sim p}\left[\sum_{k=1}^{K}\max_{\pi'_k\in\Pi'_k} \mathbb{E}_{\sigma\sim\pi}\left[f_k^M(U_k(\pi'_k,\sigma)) - f_k^M(\sigma)\right] - \gamma\cdot \mathbb{E}_{\sigma\sim\pi}[D^2_{\mathsf{H}}\big(M(\sigma), \overline{M}_1(\sigma)\big)]\right] \\
\geq& \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathrm{co}(\mathcal{M})} \sum_{k=1}^{K}\max_{\pi'_k\in\Pi'_k} \mathbb{E}_{\pi\sim p}\mathbb{E}_{\sigma\sim\pi}\left[f_k^M(U_k(\pi'_k,\sigma)) - f_k^M(\sigma)\right] - \gamma\cdot \mathbb{E}_{\pi\sim p}\mathbb{E}_{\sigma\sim\pi}\left[D^2_{\mathsf{H}}\big(M(\sigma), \overline{M}_1(\sigma)\big)\right] \\
\geq& \inf_{\pi\in\Delta(\Sigma)} \sup_{M\in\mathrm{co}(\mathcal{M})} \sum_{k=1}^{K}\max_{\pi'_k\in\Pi'_k} \mathbb{E}_{\sigma\sim\pi}\left[f_k^M(U_k(\pi'_k,\sigma)) - f_k^M(\sigma)\right] - \gamma\cdot \mathbb{E}_{\sigma\sim\pi}\left[D^2_{\mathsf{H}}\big(M(\sigma), \overline{M}_1(\sigma)\big)\right],
\end{aligned}
$$

其中，第二个等号由假设 5.1得出，第一个不等式源于平方海林格距离的凸性，第二个不等式由詹森不等式得出，最后一个不等式则是通过将任意 $p \in \Delta(\Pi)$ 替换为决策 $\pi := \mathbb{E}_{\pi\sim p}[\pi] \in \Delta(\Sigma) = \Pi$得出。

由上面的式子可知，对于任意 $M_1 \in \mathsf{co}(\mathcal{M})$，都存在某个 $\pi \in \Pi$，使得对于每个$M_2 \in \mathsf{co}(\mathcal{M})$,

$$
\sum_{k=1}^{K}\max_{\pi'_k\in\Pi'_k} \mathbb{E}_{\sigma\sim\pi}\left[f_k^{\overline{M}_2}(U_k(\pi'_k,\sigma)) - f_k^{\overline{M}_2}(\sigma)\right] - \gamma\cdot \mathbb{E}_{\sigma\sim\pi}\left[D^2_{\mathsf{H}}\big(\overline{M}_2(\sigma), \overline{M}_1(\sigma)\big)\right] \leq \mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathscr{M})). \tag{65}
$$

由于我们已假设$\mathrm{co}(\mathscr{M})$ 满足假设 1.3，因此，对于每个 $k \in [K]$，$\pi \in \Pi$以及 $M_1 \in \mathrm{co}(\mathcal{M})$，有（再次利用假设 5.1）

$$
\max_{\pi'_k\in\Pi'_k} \mathbb{E}_{\sigma\sim\pi}\left[f_k^{\overline{M}_1}(U_k(\pi'_k,\sigma)) - f_k^{\overline{M}_1}(\sigma)\right] = \max_{\pi'_k\in\Pi'_k} f_k^{\overline{M}_1}(U_k(\pi'_k,\pi)) - f_k^{\overline{M}_1}(\pi) \geq 0. \tag{66}
$$

于是，由(66) 和(65)可知，对于每个 $k \in [K]$，对于任意 $M_1 \in \overline{\mathrm{co}}(\mathcal{M})$，存在 $\pi \in \Pi$，使得对于每个$M_2 \in \overline{\mathrm{co}}(\mathcal{M})$,

$$
\max_{\pi'_k\in\Pi'_k} \mathbb{E}_{\sigma\sim\pi}\left[f_k^{\overline{M}_2}(U_k(\pi'_k,\sigma)) - f_k^{\overline{M}_2}(\sigma)\right] - \gamma\cdot \mathbb{E}_{\sigma\sim\pi}\left[D^2_{\mathsf{H}}\big(\overline{M}_2(\sigma), \overline{M}_1(\sigma)\big)\right] \leq \mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathscr{M})). \tag{67}
$$

接下来，选择$\pi \in \Pi$，给定$\overline{M}_1 = \overline{M}$ 以确保对于所有 $M_2 \in \mathrm{co}(\mathcal{M})$都成立。然后，对于每个 $k \in [K]$和每个 $\pi^\star_k \in \Pi'_k$，在 () 中选择 $M_2 = M^k_{\pi^\star_k}$ 67,

$$
\max_{\pi'_k\in\Pi'_k} \mathbb{E}_{\sigma\sim\overline{\pi}}\left[f_k^{\overline{M}^k_{\pi^\star_k}}(U_k(\pi'_k,\sigma)) - f_k^{\overline{M}^k_{\pi^\star_k}}(\sigma)\right] - \gamma\cdot \mathbb{E}_{\sigma\sim\overline{\pi}}\left[D^2_{\mathsf{H}}\Big(\overline{M}^k_{\pi^\star_k}(\sigma), \overline{M}(\sigma)\Big)\right] \leq \mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathscr{M})).
$$

对 $\pi^\star_k \sim \mu$ 取期望，并利用（63）和（64），我们得到

$$
\begin{aligned}
&\mathbb{E}_{(M,\pi^\star_k)\sim\mu}\mathbb{E}_{\sigma\sim\overline{\pi}}\left[f_k^M(U_k(\pi^\star_k,\sigma)) - f_k^M(\sigma)\right] - \gamma\cdot \mathbb{E}_{\sigma\sim\overline{\pi}}\mathbb{E}_{o|\sigma}\left[D^2_{\mathsf{H}}\big(\mu^k_{\mathrm{po}}(\cdot;\sigma,o), \mu_{\mathrm{pr}}(\cdot)\big)\right] \\
\leq& \mathbb{E}_{\pi^\star_k\sim\mu}\left[\max_{\pi'_k\in\Pi'_k} \mathbb{E}_{\sigma\sim\overline{\pi}}\left[f_k^{\overline{M}^k_{\pi^\star_k}}(U_k(\pi'_k,\sigma)) - f_k^{\overline{M}^k_{\pi^\star_k}}(\sigma)\right] - \gamma\cdot \mathbb{E}_{\sigma\sim\overline{\pi}}\left[D^2_{\mathsf{H}}\Big(\overline{M}^k_{\pi^\star_k}(\sigma), \overline{M}(\sigma)\Big)\right]\right] \\
\leq& \mathsf{r\text{-}dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathscr{M})).
\end{aligned}
$$

请注意， π 的选择仅取决于 M，尤其不取决于 k。因此，我们可以对上述表达式关于 $k \in [K]$求和，从而得到

$$\mathbb{E}_{(M,\pi_1^\star,\ldots,\pi_K^\star)\sim\mu}\mathbb{E}_{\sigma\sim\bar{\pi}}\left[\sum_{k=1}^{K} f_k^M(U_k(\pi_k^\star,\sigma)) - f_k^M(\sigma)\right] - \gamma\cdot\mathbb{E}_{\sigma\sim\bar{\pi}}\mathbb{E}_{o|\sigma}\left[\sum_{k=1}^{K} D^2_{\mathsf{H}}\big(\mu^k_{\mathrm{po}}(\cdot;\sigma,o),\mu_{\mathrm{pr}}(\cdot)\big)\right]$$
$$\leq K\cdot \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathscr{M})).$$

利用 $\mu \in \Delta(\mathcal{M}\times\Pi'_1\times\cdots\times\Pi'_K)$ 的选择是任意的这一点，我们得到 $\textsf{infr}_\gamma(\mathscr{M}) \leq \textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathscr{M}))$，如所需。 □

F.4 多智能体信息比率与通过优化进行探索的目标之间的关系

在本节中，我们证明以下结果：在 $1/(8\eta)$ 的尺度下，$\textsf{exo}_\eta(\mathscr{M})$ 的上界由 $\mathscr{M}$ 的多智能体信息比率给出。

引理 F.4. 考虑任意一个实例 $\mathscr{M} = (\mathcal{M},\ \Pi,\ \mathcal{O},\ \{\Pi'_k\}_k,\ \{U_k\}_k)$ 满足 假设 *5.1*。那么对于所有$\eta > 0$,

$$\textsf{exo}_\eta(\mathscr{M}) \leq \textsf{infr}_{1/(8\eta)}(\mathscr{M}).$$

引理 F.4 的证明引理 F.4。 在整个证明过程中，我们将由 假设 5.1 保证的（有限）纯决策集记为 Σ_1、.....、Σ_K，并将联合决策集记为 $\Sigma := \Sigma_1\times\cdots\times\Sigma_K$。此外，我们记 $\Pi' := \prod_{k=1}^K \Pi'_k$，表示偏离集合 Π'_k 的乘积。我们可以写成

$$\textsf{exo}_\eta(\mathscr{M}) = \sup_{q\in\prod_{k=1}^K\Delta(\Pi'_k)}\ \inf_{\pi\in\Delta(\Sigma),\ g\in\mathcal{G}}\ \sup_{\mu\in\Delta(\mathcal{M}\times\Pi')}\ \mathbb{E}_{(M,\pi^\star)\sim\mu}\left[\Gamma_{q,\eta}(\pi,g;\pi^\star,M)\right].$$

对于 $\alpha \geq \max\{1, 1/\eta\}$ 和 $\varepsilon \in (0,1)$，定义

$$\mathcal{G}_\alpha = \{(g_1,\ldots,g_k)\in\mathcal{G} \ :\ \|g_k\|_\infty \leq \alpha\ \forall k\in[K]\},\qquad \mathcal{P}_\varepsilon = \{\pi\in\Pi\ :\ \pi(\sigma)\geq\epsilon|\Sigma|^{-1}\ \forall\sigma\}.$$

现在，我们将使用 Sion 极小极大定理（定理B.1），取 $\mathcal{X} = \mathcal{P}_\varepsilon\times\mathcal{G}_\alpha$ 和 $\mathcal{Y} = \Delta(\mathcal{M}\times\Pi')$，在上文的定义 ot $\textsf{exo}_\eta(\mathscr{M})$ 中交换 $\inf_{\pi\in\Pi,g\in\mathcal{G}}$ 与 $\sup_{\mu\in\Delta(\mathcal{M}\times\Pi')}$。我们先检查其前提是否成立：

- 令集合 $\mathcal{P}_\varepsilon$ 取由 Π 诱导的标准拓扑，使得 $\mathcal{P}_\varepsilon$ 紧致，并令 $\mathcal{G}_\alpha$ 取乘积拓扑。Tychanoff 定理表明 $\mathcal{G}_\alpha$ 紧致，因此 $\mathcal{X} = \mathcal{P}_\varepsilon\times\mathcal{G}_\alpha$ 也紧致。它显然也是凸的。

- 令 $\mathcal{Y} = \Delta(\mathcal{M}\times\Pi')$，回顾一下，它是 $\mathcal{M}\times\Pi'$ 上有限支撑分布构成的空间）取弱拓扑。该拓扑是最粗的拓扑，使得对于所有有界函数 $\phi: \mathcal{M}\times\Pi'\to\mathbb{R}$，泛函 $\mu\mapsto\int\phi d\mu$ 都是连续的。

- 为验证其余前提条件，我们需要映射 $(\pi,\ g,\ \mu)\mapsto\mathbb{E}_{(M,\ \pi^\star)\sim\mu}[\Gamma_{q,\eta}(\pi,\ g;\ \pi^\star,\ M)$ 对于 $(\pi,\ g)\in\mathcal{P}_\varepsilon\times\mathcal{G}_\alpha$ 和 $\mu\in\Delta(\mathcal{M}\times\Pi')$，该映射是一致有界的。这直接由 $\Gamma_{q,\ \eta}(\pi,\ g;\ \pi^\star,\ M)$ 的定义以及 $\mathcal{P}_\varepsilon$ 和 $\mathcal{G}_\alpha$的定义域得出。]
- 显然，对于每个 $\pi,\ g$，映射 $\mu\mapsto\mathbb{E}_{(M,\ \pi^\star)}[\Gamma_{q,\eta}(\pi,\ g;\ \pi^\star,\ M)]$ 是线性的，因此也是凹的。此外，由于 $\Gamma_{q,\ \eta}(\pi,\ g;\ \pi^\star,\ M)$ 有界，且 $\Delta(\mathcal{M}\times\Pi')$ 具有弱拓扑，它还是连续的。
- 根据引理 F.5，对于任意固定的 μ，映射 $(\pi,\ g)\mapsto\mathbb{E}_{(M,\ \pi^\star)\sim\mu}[\Gamma_{q,\eta}(\pi,\ g;\ \pi^\star,\ M)]$ 关于 $(\pi,\ g)$ 是凸的。此外，根据定义，由于采用乘积拓扑，且 $\pi(\sigma)$ 对于 $\pi\in\mathcal{P}_\varepsilon$ 一致有下界，因此该映射连续。

验证了满足定理B.1适用的所有条件后，我们现在得到：

$$\begin{aligned}\mathsf{exo}_\eta(\mathscr{M}) &\le \sup_{q\in\prod_{k=1}^K\Delta(\Pi'_k)}\inf_{\pi\in\mathcal{P}_\varepsilon,g\in\mathcal{G}_\alpha}\sup_{\mu\in\Delta(\mathcal{M}\times\Pi')}\mathbb{E}_{(M,\pi^\star)\sim\mu}[\Gamma_{q,\eta}(\pi,g;\pi^\star,M)]\\ &= \sup_{q\in\prod_{k=1}^K\Delta(\Pi'_k)}\sup_{\mu\in\Delta(\mathcal{M}\times\Pi')}\inf_{\pi\in\mathcal{P}_\varepsilon,g\in\mathcal{G}_\alpha}\mathbb{E}_{(M,\pi^\star)\sim\mu}\left[\Gamma_{q,\eta}(\pi,g;\pi^\star,M)\right], \end{aligned} \tag{68}$$

因为我们将范围限制在更小的集合上，不等式成立 $\mathcal{G}_\alpha\subset\mathcal{G}$，并且 $\mathcal{P}_\varepsilon\subset\Pi$ 下确界中的等式利用了定理B.1。给定 $q\in\prod_{k=1}^K\Delta(\Pi'_k)$，$\mu\in\Delta(\mathcal{M}\times\Pi')$，$\pi\in\mathcal{P}_\varepsilon$考虑以下取值：

$$\begin{aligned}&\inf_{g\in\mathcal{G}_\alpha}\mathbb{E}_{(M,\pi^\star)\sim\mu}[\Gamma_{q,\eta}(\pi,g;\pi^\star,M)]\\ =&\mathbb{E}_{(M,\pi^\star)\sim\mu}\mathbb{E}_{\sigma\sim\pi}\left[\sum_{k=1}^K f_k^M(U(\pi_k^\star,\sigma))-f_k^M(\sigma)\right]\\ &+\frac{1}{\eta}\inf_{g\in\mathcal{G}_\alpha}\sum_{k=1}^K\mathbb{E}_{(M,\pi^\star)\sim\mu}\mathbb{E}_{\sigma\sim\pi,o\sim M(\sigma)}\mathbb{E}_{\pi'_k\sim q_k}\left[\exp\left(\frac{\eta}{\pi(\sigma)}\cdot(g_k(\pi'_k;\sigma,o)-g_k(\pi_k^\star;\sigma,o))\right)-1\right]\\ =&\mathbb{E}_{(M,\pi^\star)\sim\mu}\mathbb{E}_{\sigma\sim\pi}\left[\sum_{k=1}^K f_k^M(U(\pi_k^\star,\sigma))-f_k^M(\sigma)\right]\\ &+\frac{1}{\eta}\sum_{k=1}^K\inf_{g_k\in\mathcal{G}_{k,\alpha}}\mathbb{E}_{(M,\pi^\star)\sim\mu}\mathbb{E}_{\sigma\sim\pi,o\sim M(\sigma)}\mathbb{E}_{\pi'_k\sim q_k}\left[\exp\left(\frac{\eta}{\pi(\sigma)}\cdot(g_k(\pi'_k;\sigma,o)-g_k(\pi_k^\star;\sigma,o))\right)-1\right], \end{aligned} \tag{69}$$

其中 我们使用 $\mathcal{G}_{k,\ \alpha}$ 表示 $\{g_k\in\mathcal{G}_k:\ \|g_k\|_\infty\le\alpha\}$，从而有 $\mathcal{G}_\alpha=\mathcal{G}_{1,\ \alpha}\times\cdots\times\mathcal{G}_{K,\ \alpha}$。

令 $\mathbb{P}$ 表示过程（M、$\pi^\star$）$\sim\mu$、$\sigma\sim\pi$、$o\sim M(\sigma)$ 的分布律，并且对于 $k\in[K]$、$\mu^k_{\mathrm{pr}}(\pi'_k)=\mathbb{P}(\pi_k^\star=\pi'_k)$ 以及 $\mu^k_{\mathrm{po}}(\pi'_k;\ \sigma,\ o)=\mathbb{P}(\pi_k^\star=\pi'_k|(\sigma,\ o))$ 进行定义。对于每个 $k\in[K]$，上文（69）中第二项内对应于智能体 k 的项可利用后验分布 $\mu^k_{\mathrm{po}}(\pi'_k;\ \sigma,\ o)$ 的定义重写如下：

$$\begin{aligned}&\inf_{g_k\in\mathcal{G}_{k,\alpha}}\mathbb{E}_{(M,\pi^\star)\sim\mu}\mathbb{E}_{\sigma\sim\pi,o\sim M(\sigma)}\mathbb{E}_{\pi'_k\sim q_k}\left[\exp\left(\frac{\eta}{\pi(\sigma)}\cdot(g_k(\pi'_k;\sigma,o)-g_k(\pi_k^\star;\sigma,o))\right)-1\right]\\ =&\inf_{g_k\in\mathcal{G}_{k,\alpha}}\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}\left[\mathbb{E}_{\pi'_k\sim q_k}\left[\exp\left(\eta\cdot\frac{g_k(\pi'_k;\sigma,o)}{\pi(\sigma)}\right)\right]\cdot\mathbb{E}_{\pi_k^\star\sim\mu^k_{\mathrm{po}}(\cdot;\sigma,o)}\left[\exp\left(-\eta\cdot\frac{g_k(\pi_k^\star;\sigma,o)}{\pi(\sigma)}\right)\right]-1\right]. \end{aligned}$$

对于任意 $g_k\in\mathcal{G}_{k,\ \eta\alpha}$，我们有 $(\pi'_k,\ \sigma,\ o)\mapsto\frac{\pi}{(\sigma)\eta}\cdot g_k(\pi'_k;\ \sigma,\ o)$ 和 $(\pi_k^\star,\ \sigma,\ o)\mapsto\frac{\pi}{(\sigma)\eta}\cdot g_k(\pi_k^\star;\ \sigma,\ o)$ 均属于 $\mathcal{G}_{k,\ \alpha}$，这意味着上述量的上界为

$$\inf_{g_k\in\mathcal{G}_{k,\eta\alpha}}\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}\left[\mathbb{E}_{\pi'_k\sim q_k}\left[\exp(g_k(\pi'_k;\sigma,o))\right]\cdot\mathbb{E}_{\pi_k^\star\sim\mu^k_{\mathrm{po}}(\cdot;\sigma,o)}[\exp(-g_k(\pi_k^\star;\sigma,o))]-1\right].$$

该表达式等于

$$V_k(\pi,q,\mu):=\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}\inf_{g_k:\Pi'_k\to\mathbb{R},\|g_k\|_\infty\le\alpha\eta}\left\{\mathbb{E}_{\pi'_k\sim q_k}[\exp(g_k(\pi'_k))]\cdot\mathbb{E}_{\pi_k^\star\sim\mu^k_{\mathrm{po}}(\cdot;\sigma,o)}[\exp(-g_k(\pi_k^\star))]-1\right\}.$$

根据引理 B.3，对于所有 π，q，μ，我们有

$$\begin{aligned}V_k(\pi,q,\mu)&=-\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}\sup_{g_k:\Pi'_k\to\mathbb{R},\|g_k\|_\infty\le\alpha\eta}\left\{-\mathbb{E}_{\pi'_k\sim q_k}[\exp(g_k(\pi'_k))]\cdot\mathbb{E}_{\pi_k^\star\sim\mu^k_{\mathrm{po}}(\cdot;\sigma,o)}[\exp(-g_k(\pi_k^\star))]+1\right\}\\ &\le-\frac{1}{2}\cdot\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}\left[D^2_{\mathsf{H}}(\mu^k_{\mathrm{po}}(\cdot;\sigma,o),q_k)\right]+4e^{-\alpha\eta}. \end{aligned} \tag{70}$$

结合（68）、（69）和（70，我们得到以下上界：

$$\begin{aligned}\mathsf{exo}_\eta(\mathscr{M})\le\sup_{q\in\prod_{k=1}^K\Delta(\Pi'_k)}\sup_{\mu\in\Delta(\mathcal{M}\times\Pi')}\inf_{\pi\in\mathcal{P}_\varepsilon}\Bigg\{&\mathbb{E}_{(M,\pi^\star)\sim\mu}\mathbb{E}_{\sigma\sim\pi}\left[\sum_{k=1}^K f_k^M(U(\pi_k^\star,\sigma))-f_k^M(\sigma)\right]\\ &-\frac{1}{2\eta}\sum_{k=1}^K\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}\left[D^2_{\mathsf{H}}(\mu^k_{\mathrm{po}}(\cdot;\sigma,o),q_k)\right]+\frac{4K}{\eta}\cdot e^{-\alpha\eta}\Bigg\}. \end{aligned}$$

由于 $f_k^M \in [0,1]$ 对所有 k, M 以及 $D^2_{\mathsf{H}}(\cdot,\cdot) \in [0,2]$，因此我们可以将上述表达式中的 $\inf_{\pi\in\mathcal{P}_\varepsilon}$ 替换为 $\inf_{\pi\in\Pi}$，并支付加性代价 $K\varepsilon\cdot(1+1/\eta)$，因此

$$\begin{aligned}\mathsf{exo}_\eta(\mathscr{M}) \le \sup_{q\in\prod_{k=1}^K\Delta(\Pi'_k)}\sup_{\mu\in\Delta(\mathcal{M}\times\Pi')}\inf_{\pi\in\Pi}\Bigg\{&\mathbb{E}_{(M,\pi^\star)\sim\mu}\mathbb{E}_{\sigma\sim\pi}\left[\sum_{k=1}^K f_k^M(U(\pi_k^\star,\sigma))-f_k^M(\sigma)\right]\\&-\frac{1}{2\eta}\sum_{k=1}^K\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}\left[D^2_{\mathsf{H}}\big(\mu^k_{\mathrm{po}}(\cdot;\sigma,o),q_k\big)\right]+\frac{4}{\eta}\cdot e^{-\alpha\eta}+K\varepsilon\cdot(1+1/\eta)\Bigg\}.\end{aligned}$$

由于上述结论对任意 $\varepsilon\in(0,1)$ 和 $\alpha\ge\max\{1,1/\eta\}$ 都成立，我们可以取极限 $\varepsilon\to 0,\alpha\to\infty$，得到

$$\begin{aligned}\mathsf{exo}_\eta(\mathscr{M}) \le \sup_{q\in\prod_{k=1}^K\Delta(\Pi'_k)}\sup_{\mu\in\Delta(\mathcal{M}\times\Pi')}\inf_{\pi\in\Pi}\Bigg\{&\mathbb{E}_{(M,\pi^\star)\sim\mu}\mathbb{E}_{\sigma\sim\pi}\left[\sum_{k=1}^K f_k^M(U(\pi_k^\star,\sigma))-f_k^M(\sigma)\right]\\&-\frac{1}{2\eta}\sum_{k=1}^K\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}\left[D^2_{\mathsf{H}}\big(\mu^k_{\mathrm{po}}(\cdot;\sigma,o),q_k\big)\right]\Bigg\}.\end{aligned}$$

接下来，对于任意选择的 $q_k\in\Delta(\Pi'_k)$，有

$$\begin{aligned}\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}\left[D^2_{\mathsf{H}}\big(\mu^k_{\mathrm{po}}(\cdot;\sigma,o),\mu^k_{\mathrm{pr}}\big)\right]=&\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}\left[D^2_{\mathsf{H}}\big(\mu^k_{\mathrm{po}}(\cdot;\sigma,o),\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}[\mu_{\mathrm{po}}(\cdot;\sigma,o)]\big)\right]\\\le&4\cdot\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}\left[D^2_{\mathsf{H}}\big(\mu^k_{\mathrm{po}}(\cdot;\sigma,o),q_k\big)\right],\end{aligned}$$

其中，等式成立是因为对于 $\pi'_k\in\Pi'_k$，$\mu^k_{\mathrm{pr}}(\pi'_k)=\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}[\mu_{\mathrm{po}}(\pi'_k;\sigma,o)]$（根据贝叶斯定理），而不等式则由引理 B.4得出。

因此，我们有

$$\begin{aligned}\mathsf{exo}_\eta(\mathscr{M})\le&\sup_{\mu\in\Delta(\mathcal{M}\times\Pi')}\inf_{\pi\in\Pi}\left\{\mathbb{E}_{(M,\pi^\star)\sim\mu}\mathbb{E}_{\sigma\sim\pi}\left[\sum_{k=1}^K f_k^M(U(\pi_k^\star,\sigma))-f_k^M(\sigma)\right]-\frac{1}{8\eta}\sum_{k=1}^K\mathbb{E}_{\sigma\sim\pi}\mathbb{E}_{o|\sigma}\left[D^2_{\mathsf{H}}\big(\mu^k_{\mathrm{po}}(\cdot;\sigma,o),\mu^k_{\mathrm{pr}}\big)\right]\right\}\\=&\mathsf{infr}_{1/(8\eta)}(\mathscr{M}),\end{aligned}$$

正如所需。 □

引理 F.5。对于任意固定的 $\eta>0$, $q\in\prod_{k=1}^K\Delta(\Pi'_k)$、$M\in\mathcal{M}$ 和 $\pi^\star\in\Pi'$，映射 $(\pi,g)\mapsto\Gamma_{q,\eta}(\pi,g;\pi^\star,M)$ 关于 $(\pi,g)\in\Pi\times\mathcal{G}$ 联合凸。

引理 引理 F.5 的证明。任取 η，q，M，$\pi^\star$如引理陈述所示。回顾$\Gamma_{q,\eta}(\pi,g;\pi^\star,M)$在(55)中的定义。由于凸性在求和下保持不变，只需证明对于每个 k，从 $\Pi\times\mathcal{G}_k\to\mathbb{R}$映射而来的函数

$$\begin{aligned}(\pi,g_k)\mapsto&\mathbb{E}_{\sigma\sim\pi}\left[f_k^M(U_k(\pi_k^\star,\sigma))-f_k^M(\sigma)\right]\\&+\frac{1}{\eta}\cdot\mathbb{E}_{\sigma\sim\pi,o\sim M(\sigma)}\mathbb{E}_{\pi'_k\sim q_k}\left[\exp\left(\frac{\eta}{\pi(\sigma)}\cdot(g_k(\pi'_k;\sigma,o)-g_k(\pi_k^\star;\sigma,o))\right)-1\right]\end{aligned}$$

是凸的。这直接由Foster 等人（2022b）的引理 C.1 得出。 □

F.5 综合起来：定理 定理 5.1 的证明

定理 定理 5.1的证明直接由本节前面证明的引理得出。

定理 5.1 的证明： 考虑一个如下所述的实例： $\mathscr{M}$ 定理 5.1。根据引理 F.3 和引理 F.4，对于任意 $\eta>0$,

$$\mathsf{exo}_\eta(\mathscr{M})\le\mathsf{infr}_{1/(8\eta)}(\mathscr{M})\le K\cdot\mathsf{r\text{-}dec}^{\mathrm{o}}_{1/(8\eta)}(\mathrm{co}(\mathscr{M})).$$

另一方面，引理 F.1 表明，对于任意 η， $\delta > 0$，算法 1 以取值 η 运行时，以至少 $1-\delta$的概率，

$$\mathbf{Risk}(T) = h^{M^\star}(\widehat{\pi}) \le \mathsf{exo}_\eta(\mathscr{M}) + \frac{2K}{T\eta} \cdot \log\left(\frac{K \cdot \max_k |\Pi'_k|}{\delta}\right).$$

对 $\eta > 0$ 取最小值并代入 $\gamma = 1/(8\eta)$ 可得，存在一个 η 使得 算法 1给出的风险上界为

$$\mathbf{Risk}(T) = h^{M^\star}(\widehat{\pi}) \le K \cdot \inf_{\gamma>0}\left\{\textsf{r-dec}^{\mathrm{o}}_\gamma(\mathrm{co}(\mathscr{M})) + \frac{16\gamma}{T} \cdot \log\left(\frac{K \cdot \max_k |\Pi'_k|}{\delta}\right)\right\},$$

由此得到 定理 5.1所述结论。 □

G 第 5 节下界的证明

G.1 命题5.1的证明

命题 5.1 的证明。固定 $K \in \mathbb{N}$，考虑例 1.1 中的 K 玩家纳什均衡实例 $\mathscr{M}$，其中$\mathcal{A}_k = \{1,2\}$ 对于每个 $k \in [K]$。显然，对于所有 k，我们都有 $|\Pi'_k| = |\mathcal{A}_k| = 2$。根据命题 A.1，对于所有 $\gamma > 0$，我们有 $\textsf{r-dec}^{\mathrm{o}}_\gamma(\mathscr{M}^{\mathrm{NE}}) \le O(K/\gamma)$。最后，Rubinstein（2016）表明，不存在这样的算法：它抽取 $2^{o(K)}$ 个样本（每个样本都需要查询真实收益函数 $a \mapsto (f_1^{M^\star}(a) \dots, f_K^{M^\star}(a))$ 一次），并以至少 $2/3$ 的概率输出一个 c_0 近似纳什均衡，其中 $c_0 > 0$ 是一个足够小的普适常数；这便得到了命题 5.1 所述的第三个结论。 □

G.2 定理 5.2 的证明

本节中，我们证明定理 5.2。在证明该结果之前，我们先引入一些将在本节余下部分用到的记号。

- 对于整数 $N \ge N' \ge 0$，我们用 $\binom{[N]}{N'}$ 表示 $[N] = \{1, 2, \dots, N\}$ 中所有大小为 N' 的子集构成的集合。
- 对于正整数 $n \le n'$，令 $[n, n'] = \{n, n+1, \dots, n'\}$。
- 对于集合 $\mathcal{X}$，$\mathcal{Y}$，$\mathcal{X} \sqcup \mathcal{Y}$表示 $\mathcal{X}$ 和 $\mathcal{Y}$ 的不交并；其形式化定义为 $\{(x,\ 0) :\ x \in \mathcal{X}\} \cup \{(y,\ 1) :\ y \in \mathcal{Y}\}$。
- 对于有限集 $\mathcal{X}, \mathcal{Y}$，我们用 $\mathcal{X}^{\mathcal{Y}}$表示所有函数 $\phi \colon \mathcal{Y} \to \mathcal{X}$组成的集合。注意，在$\mathcal{Y} = [n]$对某个 $n \in \mathbb{N}$成立的情况下，集合 $\mathcal{X}^n$（即 $\mathcal{X}$ 的 n 重积）与 $\mathcal{X}^{[n]}$ 之间存在双射。我们有时会稍微滥用记号，将这两个集合视为相同。
- 对于有限集 $\mathcal{X}$，令 $\mathrm{Unif}(\mathcal{X})$ 表示 $\mathcal{X}$ 上的均匀分布。

定理 5.2 的证明。固定 $\epsilon > 0$ 和 $N \in \mathbb{N}$；通过增大定理陈述中的常数 C_0，不失一般性，可以假设 N 是 3 的倍数。令 $N_1 = N/3$ 且 $N_2 = 2N/3 = N - N_1$。定义

$$k = N, \qquad q = n = \frac{2k}{\epsilon} = \frac{2N}{\epsilon}, \tag{71}$$

这保证了对于足够大的 N，$q^k \ge \binom{N}{N_1}$ 成立。记 $\mathscr{T}_1 := \binom{[N]}{N_1}$ 和 $\mathscr{T}_2 := \binom{[N]}{N_2}$。现在我们将定义一个随机函数 $\widetilde{\Phi} \colon \mathscr{T}_1 \cup \mathscr{T}_2 \to [q]^{[n]} \sqcup [q+1, 2q]^{[n+1,2n]}$，使得 $\widetilde{\Phi}$ 将 $\mathscr{T}_1$ 映射到 $[q]^{[n]}$，并将 $\mathscr{T}_2$ 映射到 $[q+1, 2q]^{[n+1,2n]}$。我们将证明，$\widetilde{\Phi}$ 以正概率满足某些条件。

1. 首先，令 $\Gamma \colon \mathscr{T}_1 \cup \mathscr{T}_2 \to [q]^k \sqcup [q]^k$ 表示一个随机函数，定义如下：Γ 将 $\mathscr{T}_1$ 映射到 $[q]^k$ 的第一份副本（均匀随机地），并将 $\mathscr{T}_2$ 映射到 $[q]^k$ 的第二份副本（均匀随机地）。特别地，对于每个 $\mathcal{S} \in \mathscr{T}_1 \cup \mathscr{T}_2$，$\Gamma(\mathcal{S})$ 相互独立，并分别从 $[q]^k$的对应副本中均匀随机选取。

2. 接下来，我们定义一个映射 $\Sigma:\ [q]^k \sqcup [q]^k \to [q]^{[n]} \sqcup [q+1,2q]^{[n+1,2n]}$ 它根据 $[q]^k$的里德–所罗门码，将其第一份副本映射到 $[q]^{[n]}$，并将 $[q]^k$ 的第二份副本映射到 $[q+1,2q]^{[n+1,2n]}$。该码见引理 G.6。（这里我们以自然的方式将 $[q]^{[n]}$ 和 $[q+1,2q]^{[n+1,2n]}$ 分别认作 $[q]^n$。）

3. 然后我们令 $\Phi = \Sigma \circ \Gamma$。

接下来我们说明，存在某种选择 Γ，使得由此得到的 Φ 满足以下条件 G.1和 G.2。

条件 G.1。对于每个 $i \in \{1,2\}$，对于所有满足 $\mathcal{T} \neq \mathcal{T}'$ 的集合 $\mathcal{T}, \mathcal{T}' \in \mathscr{T}_i$，均有 $d_{\mathrm{Ham}}(\Phi(\mathcal{T}) \Phi(\mathcal{T}')) \geq q-k+1$。

条件 G.2。对于任意满足 $|\mathcal{Q}| \leq \sqrt{N}$ 的子集 $\mathcal{Q} \subset [N]$，

$$\forall a_1 \in [n], a_2 \in [q], \quad \mathbb{P}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_1)} \left(\widetilde{\Phi}(\mathcal{T})(a_1) = a_2 | \mathcal{Q} \subset \mathcal{T} \right) \leq 2/q \tag{72}$$

$$\forall a_1 \in [n+1,2n], a_2 \in [q+1,2q], \quad \mathbb{P}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_2)} \left(\widetilde{\Phi}(\mathcal{T})(a_1) = a_2 | \mathcal{Q} \subset \mathcal{T} \right) \leq 2/q. \tag{73}$$

为了说明存在这样的选择使得 Γ，我们作如下观测。

1. 由于 $q^k > 10 \cdot \binom{N}{N_1}^2$ 当 N 足够大时（根据 公式 (71)），以至少 $1 - \binom{N}{N_1}^2 / q^k > 9/10$ 的概率，函数 Γ 是单射。条件是其为单射，引理 G.6 给出条件 G.1成立，因为 Σ 在 $[q]^k$ 的每个副本上的动作都定义为里德–所罗门码的动作。因此，条件 G.1 以至少 $9/10$ 的概率成立，其中概率是针对 Γ的选择而言。

2. 考虑任意固定的 $\mathcal{T} \in \mathscr{T}_1$。注意，对于每个坐标 $a_1 \in [n]$，映射 $\mathcal{T} \mapsto \Phi(\mathcal{T})(a_1) = \Sigma(\Gamma(\mathcal{T}))(a_1)$，对于 $\mathcal{T} \in \mathscr{T}_1$，服从来自 $\mathscr{T}_1 \to [q]$ 的均匀随机函数分布（关于 Γ 中的随机性）。这一事实可由引理 G.6 的最后一句以及 Γ 是一个均匀随机函数这一事实得出。因此，根据引理 G.7，取 $N_0 = N_1$，并对 a_1 的所有 n 个可能取值应用并集界，可知以 Γ 的选取为条件，概率为 $1 - n \cdot N^{\sqrt{N}+1} \cdot 2^{-\binom{5N/6}{N/6}/(Cq^2)}$ 时，对于大小为 $|\mathcal{Q}| \leq \sqrt{N}$ 的任意子集 $\mathcal{Q} \subset [N]$，式 (72) 成立。类似地，应用引理 G.7，取$N_0 = N_2$，可得以 Γ 的选取为条件，概率为 $1 - n \cdot N^{\sqrt{N}+1} \cdot 2^{-\binom{5N/6}{N/6}/(Cq^2)}$ 时，对于大小为 $|\mathcal{Q}| \leq \sqrt{N}$ 的任意子集 $\mathcal{Q} \subset [N]$，式 (73) 成立。注意，我们对 q、N、ϵ 的选取确保，对于引理 G.7 中的常数C，只要 N 足够大，

$$3 \log q \leq 6 \log(N/\epsilon) \leq 12 \log(N) \leq N/6 - C \leq \log \binom{5N/6}{N/6} - C,$$

这意味着可以应用引理 G.7。最后需要指出，我们对 N，q 的选择确保，只要 N 足够大，

$$\binom{5N/6}{N/6} > Cq^2 \cdot \left(\log(2n) + \log(N^{\sqrt{N}+1}) + 5 \right),$$

因此，(72)和(73)对所有 $\mathcal{Q} \subset [N]$满足 $|\mathcal{Q}| \leq \sqrt{N}$的情形均成立，且其概率至少为 $1 - 2^{-5}$。特别地，条件 G.2 成立的概率至少为 $1 - 2^{-5}$，其中 Γ是随机选取的。

综上所述，在 $1 - 1/10 - 2^{-5} > 0$ 的选择下，Γ满足条件 G.1和G.2 的概率至少为 Γ。我们任取一个满足这两个条件的 $\Phi = \Sigma \circ \Gamma$，并令其为

现在可以定义双人实例 $\mathscr{M} = (\mathcal{M},\ \Pi,\ \mathcal{O},\ \{\Pi'_k\}_k,\ \{U_k\}_k)$。

策略空间。令 $\Pi_1 = \{1, 2, \ldots, 2n\}$ 和 $\Pi_2 = \{0, 1, \ldots, 2q\}$，并用 $\Pi = \Pi_1 \times \Pi_2$ 表示联合策略空间。

偏离集合与切换函数。偏离集合 Π'_k 和切换函数 U_k 按照定义 1.1 设定，使得 $\mathscr{M}$ 成为一个双人纳什均衡实例。具体而言，对于 $\Pi'_k = \Pi_k$ 每个 k，且$U_k(\pi'_k,\ \pi) = (\pi'_k,\ \pi_{-k})$。

模型类 $\mathcal{M}$。类 $\mathcal{M}$ 由 $\mathscr{T}_1 \cup \mathscr{T}_2$ 索引。给定一个集合 $\mathcal{T} \in \mathscr{T}_1 \cup \mathscr{T}_2$，我们将对应的模型记作 $M_{\mathcal{T}}$。我们通常考虑分解 $\mathcal{M} = \mathcal{M}_1 \sqcup \mathcal{M}_2$，其中 $\mathcal{M}_1 := \{M_{\mathcal{T}}: \ \mathcal{T} \in \mathscr{T}_1\}$ 和 $\mathcal{M}_2 := \{M_{\mathcal{T}}: \ \mathcal{T} \in \mathscr{T}_2\}$。对于每个 $M_{\mathcal{T}} \in \mathcal{M}$，我们需要指定分布 $o = (r_1, \ r_2, \ o_\circ) \sim M_{\mathcal{T}}(\pi)$，对于每个 $\pi \in \Pi$。为此，我们首先定义一个映射 $\mathcal{B}^\star$：$\mathscr{T}_1 \cup \mathscr{T}_2 \to \mathcal{P}([2n] \times [2q]) \subset \mathcal{P}(\Pi)$，如下所示：回想一下，$\Phi$ 将 $\mathscr{T}_1$ 映射到 $[q]^{[n]}$，并将 $\mathscr{T}_2$ 映射到 $[q+1, 2q]^{[n+1,2n]}$。然后对于 $\mathcal{T} \in \mathscr{T}_1 \cup \mathscr{T}_2$，定义 $\mathcal{B}^\star(\mathcal{T}) \subset [2n] \times [2q] \subset \Pi$，使得

$$\mathcal{B}^\star(\mathcal{T}) = \begin{cases} \{(i, \Phi(\mathcal{T})(i)) : i \in [n]\} & : M_{\mathcal{T}} \in \mathcal{M}_1 \\ \{(i, \Phi(\mathcal{T})(i)) : i \in [n+1, 2n]\} & : M_{\mathcal{T}} \in \mathcal{M}_2. \end{cases}$$

注意，这里我们将定义域中的每个集合 $\mathcal{T}$ in Φ 的 $\Phi(\mathcal{T})$ 视为一个函数，该函数映射到 $[n] \to [q]$（对于 $M_{\mathcal{T}} \in \mathcal{M}_1$）或 $[n+1, 2n] \to [q+1, 2q]$（对于 $M_{\mathcal{T}} \in \mathcal{M}_2$）。

我们将奖励空间设为 $\mathcal{R} = [0, 1]$，将纯观测空间设为 $\mathcal{O}_\circ = [N]$。现在，对于每个$M_{\mathcal{T}} \in \mathcal{M}$ 和 $\pi \in \Pi$，完整观测 $o = (r_1, r_2, o_\circ) \sim M_{\mathcal{T}}(\pi)$ 按如下方式抽取：

- 纯观测 $o_\circ \in \mathcal{O}_\circ$ 只是集合 $\mathcal{T}$ 中的一个均匀随机元素。
- 奖励是确定性的，即对于每个 $k \in [K]$，几乎必然有 $r_k = f_k^{M_{\mathcal{T}}}(\pi)$。此外，我们定义

$$f_1^{M_{\mathcal{T}}}(\pi) = -f_2^{M_{\mathcal{T}}}(\pi) = \begin{cases} 0 & : \pi \in \Pi_1 \times \{0\} \\ 1 & : \pi \in (\Pi_1 \times \{1, 2, \ldots, 2q\}) \backslash \mathcal{B}^\star(\mathcal{T}) \\ -\delta & : \pi \in \mathcal{B}^\star(\mathcal{T}), \end{cases} \tag{74}$$

其中我们设 $\delta := 10^{-3}$。

证明所声称的结论。根据 Π 的定义，显然有 $|\Pi| = 2n \cdot (2q+1) = O(N^2/\epsilon^2)$，因此定理的第一个所声称结论得证。接下来，下面的 引理 G.1 给出了$\textsf{r-dec}^{\circ}_{\gamma}(\textsf{co}(\mathscr{M}))$ 的界，从而证明了第二个所声称的结论。

引理 G.1。对于任意 $\gamma > 0$，有$\textsf{r-dec}^{\circ}_{\gamma}(\textsf{co}(\mathscr{M})) \leq \epsilon$。

引理 G.1 的证明利用了 Φ 满足 条件 G.1这一事实。最后，以下引理证明第三个所声称的结论。

引理 G.2。存在一个常数 $C > 0$，使得以下结论成立。对于任意最多进行$T \leq \sqrt{N}/C$轮交互的算法，存在某个模型 $M^\star \in \mathcal{M}$，使得

$$\mathbb{E}^{M^\star}[\mathbf{Risk}(T)] = \mathbb{E}^{M^\star}\left[h^{M^\star}(\widehat{\pi})\right] > \delta/100 = 10^{-5}.$$

回顾一下，上文中 $\widehat{\pi}$ 表示该算法的输出策略。

引理 G.2 的证明利用了 Φ满足 条件 G.2这一事实。还需证明引理 G.1 和 G.2；我们将在本节余下部分完成证明。 □

证明 引理 G.1。对于 $M \in \textsf{co}(\overline{\mathcal{M}})$ 和 $i \in [N]$，令 $M[i] \in [0, \overline{1}]$ 表示概率 $\mathbb{P}_{(r_1, r_2, o_\circ) \sim M(\pi)}[o_\circ = i]$，其中决策 $\pi \in \Pi$ 任意（注意，决策的选择不会影响纯观测 $o_\circ$上的分布）。对于任意 $M \in \textsf{co}(\mathcal{M})$，我们将集合 $\mathcal{T}(M) \subset [N]$定义如下：

$$\mathcal{T}(\overline{M}) := \left\{i \in [N] \ : \ \overline{M}[i] \geq 1/N\right\},$$

现在固定任意一个 $M \in \textsf{co}(\mathcal{M})$。将 $p^\star \in \Delta(\Pi)$ 定义如下，它是 M 的函数：

$$p^\star = \begin{cases} \text{Unif}(\{(1,0), \ldots, (n,0)\}) & : |\mathcal{T}(\overline{M})| \geq N/2 \\ \text{Unif}(\{(n+1,0), \ldots, (2n,0)\}) & : |\mathcal{T}(\overline{M})| < N/2. \end{cases}$$

我们有

$$\begin{aligned}\textsf{r-dec}^{\mathrm{o}}_{\gamma}(\mathrm{co}(\mathscr{M}),\overline{M}) &\le \sup_{M\in\mathrm{co}(\mathcal{M})}\mathbb{E}_{\pi\sim p^\star}\left[h^M(\pi)-\gamma\cdot D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right]\\ &= \sup_{M\in\mathrm{co}(\mathcal{M})}\mathbb{E}_{\pi\sim p^\star}\left[\sum_{k=1}^{2}\max_{\pi'_k\in\Pi_k} f_k^M(\pi'_k,\pi_{-k})-f_k^M(\pi)-\gamma\cdot D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right]\\ &= \sup_{M\in\mathrm{co}(\mathcal{M})}\mathbb{E}_{\pi\sim p^\star}\left[\max_{\pi'_2\in\Pi_2} f_2^M(\pi_1,\pi'_2)-f_2^M(\pi)-\gamma\cdot D^2_{\mathsf{H}}\big(M(\pi),\overline{M}(\pi)\big)\right], \qquad (75)\end{aligned}$$

因为对所有 $M\in\mathrm{co}(\mathcal{M})$ 以及 $p^\star$ 的支撑集中的所有 π，$\max_{\pi'_1\in\Pi_1} f_1^M(\pi'_1,\ \pi_2)=0=f_1^M(\pi)$，所以最后一个等号成立。

固定 $\nu\in\Delta(\mathcal{M})$，使得 $M=\overline{M}_\nu(\pi):=\mathbb{E}_{M\sim\nu}[M(\pi)]$ 在 式（75）中达到上确界。我们考虑以下可能性：

情况1。首先假设 $|\mathcal{T}(M)|\ge N/2$。我们考虑以下子情况：

1. 首先假设 $\nu(\mathcal{M}_1)\le\frac{1}{1+\delta}$，这里回顾一下 $\delta:=10^{-3}$（式（74））。那么对所有 $\pi_1\in[n]$ 和 $\pi_2\in[2q]$，有

$$f_2^{\overline{M}_\nu}((\pi_1,\pi_2))\le\frac{1}{1+\delta}\cdot\delta-\left(1-\frac{1}{1+\delta}\right)=0, \qquad (76)$$

因为 $f_2^{M_\mathcal{T}}((\pi_1,\ \pi_2))$ 仅当 $\pi\in\mathcal{B}^\star(\mathcal{T})$时为正，而在 $M_\mathcal{T}\sim\nu$ 下，这种情况发生的概率至多为 $\frac{1}{1+\delta}$，因为 $\pi_1\in[n]$；此外，当其为正时，它等于 δ，而当其不为正时，它等于 -1。利用 (76)，由于所有决策 π 都属于 $p^\star$ 的支撑集（在这种子情况下，它们具有 $\pi_1\in[n]$），所以 (75) 中的表达式有上界 0。

2. 接下来，假设存在某个模型 $M_\mathcal{T}\in\mathcal{M}_1$ 使得 $\nu(M_\mathcal{T})\ge 14/15$。因此，我们必有$\sum_{j\in\mathcal{T}(M)\setminus\mathcal{T}} M_\nu[j]\le\sum_{j\in[N]\setminus\mathcal{T}} M_\nu[j]\le\frac{1}{15}\cdot$另一方面，由于 $|\mathcal{T}(M)|\ge N/2$，我们有$\sum_{j\in\mathcal{T}(M)\setminus\mathcal{T}} M[j]\ge\frac{1}{N}\cdot\frac{N}{6}=1/6$。因此，对于任意决策 $\pi\in\Pi$，

$$D^2_{\mathsf{H}}\big(\overline{M}_\nu(\pi),\overline{M}(\pi)\big)\ge\big(D_{\mathsf{TV}}\big(\overline{M}_\nu(\pi),\overline{M}(\pi)\big)\big)^2\ge(1/6-1/15)^2=1/100,$$

因此，只要 $\gamma\ge 100$，由于 $f_2^M(\pi_1,\ \pi'_2)\le 1$对于所有 $\pi_1,\ \pi'_2$，如果我们记得 $M=M_\nu$被选为使下列表达式最大化 (75)，那么该表达式有上界 0。

3. 在剩余情形下，我们必有 $\nu(\mathcal{M}_1)\ge\frac{1}{1+\delta}$，但对于每个 $M_\mathcal{T}\in\mathcal{M}_1$，$\nu(M_\mathcal{T})<14/15$。假设为反证目的，

$$\mathbb{E}_{\pi\sim p^\star}\left[\max_{\pi'_2\in\Pi_2} f_2^{\overline{M}_\nu}(\pi_1,\pi'_2)-f_2^{\overline{M}_\nu}(\pi)\right]=\mathbb{E}_{\pi\sim p^\star}\left[\max_{\pi'_2\in\Pi_2} f_2^{\overline{M}_\nu}(\pi_1,\pi'_2)\right]>\epsilon. \qquad (77)$$

写出 $\mathcal{I}=\{\pi_1\in[n]:\ \max_{\pi'_2\in\Pi_2} f_2^{\overline{M}_\nu}(\pi_1,\pi'_2)\ge 0\}$；由于对于所有 π_1，$\max_{\pi'_2\in\Pi_2} f_2^{M_\nu}(\pi_1,\pi'_2)\le 1$，式 (77) 告诉我们 $|\mathcal{I}|\ge\epsilon n$。根据构造，对于每个 $\pi_1\in\mathcal{I}$，至多存在一个 $\pi_2\in[q]$ 的取值使得 $f_2^{M_\nu}(\pi_1,\ \pi_2)\ge 0$；如果给定 π_1时存在这样的 π_2，则将 π_2 的该取值记为 $\pi_2^\star(\pi_1)$，否则令 $\pi_2^\star(\pi_1)=-1$。

注意，如果对于任意 $\pi_1\in\Pi_1$，$f_2^{M_\nu}(\pi_1,\ \pi_2)\ge 0$，那么必有 $\nu(\{M_\mathcal{T}\in\mathcal{M}_1:\ \Phi(\mathcal{T})(\pi_1)=\pi_2\})\ge\frac{1}{1+\delta}>1-\delta.$。因此，对于所有 π_1，如果 $\pi_2^\star(\pi_1)>0$，那么

$$\nu(\{M_\mathcal{T}\in\mathcal{M}_1:\Phi(\mathcal{T})(\pi_1)=\pi_2^\star(\pi_1)\})>1-\delta. \qquad (78)$$

对于每个 $M_\mathcal{T}\in\mathcal{M}_1$，定义

$$\zeta(\mathcal{T}):=\left|\{\pi_1\in\mathcal{I}\ :\ \Phi(\mathcal{T})(\pi_1)\ne\pi_2^\star(\pi_1)\}\right|.$$

我们有

$$|\mathcal{I}| - \sum_{M_{\mathcal{T}} \in \mathcal{M}_1} \nu(M_{\mathcal{T}})\zeta(\mathcal{T}) = \sum_{\pi_1 \in \mathcal{I}} \sum_{M_{\mathcal{T}} \in \mathcal{M}_1} \nu(M_{\mathcal{T}}) \cdot \mathbb{I}\{\Phi(\mathcal{T})(\pi_1) = \pi_2^\star(\pi_1)\} \geq |\mathcal{I}| \cdot (1-\delta),$$

其中，不等式用到了 (78)。因此，根据马尔可夫不等式，对于某个子集 $\mathcal{M}_1' \subset \mathcal{M}_1$，有 $\nu(\mathcal{M}_1 \backslash \mathcal{M}_1') \leq \sqrt{\delta}$，且对所有 $M_{\mathcal{T}} \in \mathcal{M}_1'$，$\zeta(\mathcal{T}) \leq |\mathcal{I}| \cdot \sqrt{\delta}$。由于 $\nu(\mathcal{M}_1) \geq 1-\delta$，可得 $\nu(\mathcal{M}_1') \geq 1 - \delta - \sqrt{\delta} \geq 1 - 2\sqrt{\delta}$。根据我们对 $\delta = 10^{-3}$ 的选择，既然 $1 - 2\sqrt{\delta} > 14/15$，那么 $\mathcal{M}_1'$ 中至少有两个不同的元素，记为 $M_{\mathcal{T}_1}$ 和 $M_{\mathcal{T}_2}$。

接下来，根据 $\mathcal{M}_1'$ 的定义，有

$$|\{\pi_1 \in \mathcal{I} \ : \ \Phi(\mathcal{T}_1)(\pi_1) = \Phi(\mathcal{T}_2)(\pi_1) = \pi_2^\star(\pi_1)\}| \geq |\mathcal{I}| \cdot (1 - 2\sqrt{\delta}) \geq |\mathcal{I}|/2 \geq n\epsilon/2.$$

由此可得 $d_{\mathrm{Ham}}(\Phi(\mathcal{T}_1)\ \Phi(\mathcal{T}_2)) \leq n - n\epsilon/2 = n(1-\epsilon/2)$，这与条件 G.1 矛盾，因为 $n(1-\epsilon/2) = \frac{2N}{\epsilon} \cdot (1-\epsilon/2) < \frac{2N}{\epsilon} - N + 1 = q - k + 1$。因此，(77) 不成立，所以 (75) 中对应于选择 $M = M_\nu$ 的表达式至多为 ϵ。

情形 2。 现在假设 $|\mathcal{T}(M)| < N/2$。在这种情况下，可以采用与 $|\mathcal{T}(M)| \geq N/2$ 情形中对称的论证，来建立多智能体 DEC 的相同上界。（具体而言，$\mathcal{M}_1$、$\mathcal{M}_2$ 的角色互换；这种对称性源于 $\mathscr{T}_1$ 中集合的大小为 $N/3 = N/2 - N/6$，而 $\mathscr{T}_2$ 中集合的大小为 $2N/3 = N/2 + N/6$。）下面我们详细展开论证，以求完整。

1. 若 $\nu(\mathcal{M}_2) \leq \frac{1}{1+\delta}$，则对所有 $\pi_1 \in [n+1, 2n]$ 和 $\pi_2 \in [2q]$，$f_2^{M_\nu}((\pi_1, \pi_2)) \leq 0$，这意味着，由于对支撑集 π 中的所有决策 $p^\star$，$\pi_2 \in [n+1, 2n]$，式 (75) 均为非正。
2. 接下来，假设存在某个模型 $M_{\mathcal{T}} \in \mathcal{M}_2$，使得 $\nu(M_{\mathcal{T}}) \geq 14/15$。我们必须有 $\sum_{j \in \mathcal{T} \backslash \mathcal{T}(M)} M[j] \leq |\mathcal{T} \backslash \mathcal{T}(M)| \cdot \frac{1}{N}$。另一方面，由于对于每个 $i \in \mathcal{T}$，我们都有 $M_{\mathcal{T}}[i] = 3/(2N)$，因此有 $\sum_{j \in \mathcal{T} \backslash \mathcal{T}(M)} M_\nu[j] \geq \frac{14}{15} \cdot \frac{3}{2N} \cdot |\mathcal{T} \backslash \mathcal{T}(M)| \geq \frac{7}{5N} \cdot |\mathcal{T} \backslash \mathcal{T}(M)|$。因此，对于任意 $\pi \in \Pi$，由于 $|\mathcal{T}(M)| \leq N/2$ 且 $|\mathcal{T}| = 2N/3$（因为 $M_{\mathcal{T}} \in \mathcal{M}_2$），

$$D_{\mathsf{H}}^2\big(\overline{M}_\nu(\pi), \overline{M}(\pi)\big) \geq \big(D_{\mathsf{TV}}\big(\overline{M}_\nu(\pi), \overline{M}(\pi)\big)\big)^2 \geq \left(|\mathcal{T} \backslash \mathcal{T}(\overline{M})| \cdot \left(\frac{7}{5N} - \frac{1}{N}\right)\right)^2$$
$$\geq \left(\frac{N}{6} \cdot \left(\frac{7}{5N} - \frac{1}{N}\right)\right)^2 = \frac{1}{225}.$$

因此，只要 $\gamma \geq 225$，由于 $f_2^M(\pi_1, \pi_2') \leq 1$ 对于所有 π_1，π_2'，(75) 中的表达式关于 $M = M_\nu$ 有上界 0。

3. 在剩余情形下，必有 $\nu(\mathcal{M}_2) \geq \frac{1}{1+\delta}$，然而对于每个 $M_{\mathcal{T}} \in \mathcal{M}_2$，$\nu(M_{\mathcal{T}}) < 14/15$。在这种情况下，通过与上文 (75) 中针对 $M = M_\nu$ 的论证完全相同的推理（将所有 $\mathcal{M}_1$ 替换为 $\mathcal{M}_2$），可知该式有上界 ϵ。条目 3

总之，我们已经证明，对于任意选取的 M，(75) 都有上界 ϵ，这就完成了该引理的证明。

□

引理 G.2 的证明。任取 $T \leq \sqrt{N}/C$（其中常数 C 将在下文指定），并考虑任意算法 $(p, q) = \{(q^t(\cdot|\cdot)\ p(\cdot|\cdot)\}_{t=1}^T$。回忆一下，对于任意模型 M，$\mathfrak{H}^T$ 表示算法 (p, q) 与模型 M 之间的交互历史，其定义为 $\mathfrak{H}^T = (\pi^1, o^1) \ldots, (\pi^T, o^T)$。$\mathfrak{H}^T$ 与测度空间 $(\Omega^T, \mathscr{F}^T)$ 相关联。对于每个模型 $M \in \mathcal{M}$，我们用缩写 $\mathbb{P}^M \equiv \mathbb{P}^{M,(p,q)}$ 表示 $\mathfrak{H}^T$ 的概率律，并用 $\mathbb{E}^M$ 表示相应的期望。我们将证明一个更强的结论：对于均匀随机选取的模型 $M^\star \in \mathcal{M}$，算法 (p, q) 的风险较大；特别地，

$$\mathbb{E}_{M^\star \sim \mathrm{Unif}(\mathcal{M})} \mathbb{E}^{M^\star}\left[h^{M^\star}(\widehat{\pi})\right] > \delta/100. \tag{79}$$

显然 (79) 推出了引理 G.2。

为了证明引理 G.2，我们先证明几个中间结果。首先，我们定义一个额外的模型 M_0：$(r_1,\ r_2,\ o_\circ) \sim M_0(\pi)$ 的分布如下：

- 奖励 r_1 r_2 如 (74) 所示， $\mathcal{B}^\star = \varnothing$；特别地， $r_k = f_k^{M_0}(\pi)$ 是确定性的，其中 h

$$f_1^{M_0}(\pi) = -f_2^{M_0}(\pi) = \begin{cases} 0 & : \pi \in \Pi_1 \times \{0\} \\ 1 & : \pi \in \Pi_2 \times \{1, 2, \ldots, 2q\}. \end{cases}$$

- 纯观测 $o_\circ \in [N]$ 是从 $[N]$ 中均匀随机选取的元素。

接下来，回顾一下，对于 $\pi \in \Pi$， $k \in \{1, 2\}$， $M \in \mathcal{M}$, $h_k^M(\pi) = \max_{\pi'_k \in \Pi'_k} f_k^M(U_k(\pi'_k, \pi)) - f_k^M(\pi)$ 的记法。

引理 G.3 表明，对于每个 $i \in \{1, 2\}$，在模型 M_0下，以常数概率， $\mathcal{M}_1$ 中的所有模型或 $\mathcal{M}_2$ 中的所有模型，相对于算法的输出策略 $\widehat{\pi}$都具有较高风险。

引理 G.3. 存在某个 $i \in \{1, 2\}$ （取决于算法 (p, q)），使得

$$\mathbb{P}^{M_0}\left(\forall M \in \mathcal{M}_i : h^M(\widehat{\pi}) \geq \delta\right) \geq \frac{1}{2}.$$

引理 G.3 的证明将在本节末尾给出。由于 M_0 不属于 $\mathcal{M}$，引理 G.3 不足以证明引理 G.2；接下来，我们将使用一系列测度变换论证，分析真实模型是 M中的均匀随机模型时的交互历史。具体而言，对于每个模型 $M_\mathcal{T} \in \mathcal{M}$，我们定义一个中间模型 $M_{\mathcal{T},\,0}$：$(r_1,\ r_2,\ o_\circ) \sim M_{\mathcal{T},\,0}(\pi)$ 的分布如下：

- 奖励 $(r_1,\ r_2)$ 与 $M_0(\pi)$ 下的奖励完全相同（特别是，它们是确定性的）。

- 纯观测 $o_\circ$ 是 $\mathcal{T}$中的均匀随机元素。

引理 G.4 表明，对于从 $M_{\mathcal{T},\,0}$ 中抽取的历史，当 $\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_i)$ 为均匀随机变量时，算法以高概率不会查询属于 $\mathcal{B}^\star(\mathcal{T}) \subset \Pi$ 的任何决策；此外，对于均匀随机的 $\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_i)$，历史 $\mathfrak{H}^T$ 在 M_0 和 $M_{\mathcal{T},\,0}$ 下的分布也很接近：

引理 G.4。 对于每个 $i \in \{1, 2\}$，以下结论成立：

$$\mathbb{E}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_i)} \mathbb{E}^{M_\mathcal{T}, 0}\left[\mathbb{I}\left\{\{\pi^1, \ldots, \pi^T\} \cap \mathcal{B}^\star(\mathcal{T}) \neq \varnothing\right\}\right] \leq \frac{2T}{q} + \frac{1}{100}. \tag{80}$$

此外，对于任意历史记录的可测子集 $\mathcal{F} \in \mathscr{F}^T$，

$$\left|\mathbb{E}^{M_0}\left[\mathbb{I}\left\{\mathfrak{H}^T \in \mathcal{F}\right\}\right] - \mathbb{E}_{\mathcal{T} \sim \mathrm{Unif}(\mathscr{T}_i)} \mathbb{E}^{M_\mathcal{T}, 0}\left[\mathbb{I}\left\{\mathfrak{H}^T \in \mathcal{F}\right\}\right]\right| \leq \frac{1}{100}. \tag{81}$$

引理 G.4的证明见本节末尾。

接下来，引理 G.5 表明，如果对于某个模型 $M_\mathcal{T}$，算法以高概率不查询 $\mathcal{B}^\star(\mathcal{T})$ 中的任何决策，那么在 $\mathbb{P}^{M_\mathcal{T},\,0}$ 和 $\mathbb{P}^{M_\mathcal{T}}$ 下的历史分布彼此接近。

引理 G.5。 固定某个模型 $M_\mathcal{T} \in \mathcal{M}$，使得 $\mathbb{P}^{M_\mathcal{T},\,0}(\{\pi^1, \ldots, \pi^T\} \cap \mathcal{B}^\star(M_\mathcal{T}) \neq \varnothing) \leq \eta$对于某个 $\eta > 0$ 成立。于是 $D_{\mathsf{TV}}(\mathbb{P}^{M_\mathcal{T},\,0}, \mathbb{P}^{M_\mathcal{T}}) \leq \eta$。

引理 G.5的证明见本节末尾。利用上述引理，我们现在证明(79)。为达到反证目的，假设 $\mathbb{E}_{M \sim \mathrm{Unif}(\mathcal{M})} \mathbb{E}^M\left[h^M(\widehat{\pi})\right] \leq \delta/100$。那么根据马尔可夫不等式， $\mathbb{E}_{M \sim \mathrm{Unif}(\mathcal{M})} \mathbb{E}^M\left[\mathbb{I}\left\{h^M(\widehat{\pi}) \geq \delta\right\}\right] \leq 1/100$。由于 $\mathrm{Unif}(\mathcal{M})$ 是$\mathrm{Unif}(\mathcal{M}_1)$ 与$\mathrm{Unif}(\mathcal{M}_2)$的均匀平均，因此对于每个 $i \in \{1, 2\}$,

$$\mathbb{E}_{M \sim \mathrm{Unif}(\mathcal{M}_i)} \mathbb{E}^M\left[\mathbb{I}\left\{h^M(\widehat{\pi}) \geq \delta\right\}\right] \leq 1/50. \tag{82}$$

接下来注意到引理 G.3表明存在某个 $i^\star \in \{1, 2\}$,

$$\mathbb{P}^{M_0}\left(\forall M \in \mathcal{M}_{i^\star} : h^M(\widehat{\pi}) \geq \delta\right) \geq \frac{1}{2}.$$

由(81) 和引理 G.4可得

$$\mathbb{E}_{\mathcal{T}\sim\mathrm{Unif}(\mathscr{T}_{i^\star})}\mathbb{E}^{M_{\mathcal{T}},0}\left[\mathbb{I}\{\forall M\in\mathcal{M}_{i^\star}: h^M(\widehat{\pi})\geq\delta\}\right]\geq 1/2-1/100. \tag{83}$$

接下来，根据 (80) 中的 引理 G.4 ，并利用 $2T\leq\sqrt{N}$ 和 $\sqrt{N}/q\leq 1\overline{/}\sqrt{N}\leq 1\overline{/}100$ 对充分大的 N，

$$\mathbb{E}_{\mathcal{T}\sim\mathrm{Unif}(\mathscr{T}_{i^\star})}\mathbb{E}^{M_{\mathcal{T}},0}\left[\mathbb{I}\left\{\{\pi^1,\ldots,\pi^T\}\cap\mathcal{B}^\star(M_{\mathcal{T}})\neq\varnothing\right\}\right]\leq\frac{\sqrt{N}}{q}+\frac{1}{100}\leq\frac{1}{50}. \tag{84}$$

现在，对于 $\eta=1/7$，记 $\chi(\mathcal{T}):=\mathbb{I}\left\{\mathbb{E}^{M_{\mathcal{T}},0}\left[\mathbb{I}\{\{\pi^1,\ldots,\pi^T\}\cap\mathcal{B}^\star(M_{\mathcal{T}})\neq\varnothing\}\right]>\eta\right\}\in\{0,1\}$；式 (84)结合马尔可夫不等式可得 $\mathbb{E}_{\mathcal{T}\sim\mathrm{Unif}(\mathscr{T}_{i^\star})}[\chi(\mathcal{T})]\leq 1/7$。
接下来，引理 G.5 表明，对所有 $\mathcal{T}\in\mathscr{T}_1\cup\mathscr{T}_2$,

$$\mathbb{P}^{M_{\mathcal{T}}}\left(\forall M\in\mathcal{M}_{i^\star}:\ h^M(\widehat{\pi})\geq\delta\right)\geq\mathbb{P}^{M_{\mathcal{T}},0}\left(\forall M\in\mathcal{M}_{i^\star}:\ h^M(\widehat{\pi})\geq\delta\right)-\chi(\mathcal{T})-\eta,$$

并对 $\mathcal{T}\sim\mathrm{Unif}(\mathscr{T}_{i^\star})$ 取期望，利用 $\mathbb{E}_{\mathcal{T}\sim\mathrm{Unif}(\mathscr{T}_{i^\star})}[\chi(\mathcal{T})]\leq 1/7$ 以及$\eta=1/7$ 的选取，可得

$$\begin{aligned}&\mathbb{P}_{\mathcal{T}\sim\mathrm{Unif}(\mathscr{T}_{i^\star})}\mathbb{E}^{M_{\mathcal{T}}}\left[\mathbb{I}\left\{\forall M\in\mathcal{M}_{i^\star}:\ h^M(\widehat{\pi})\geq\delta\right\}\right]\\ \geq&\mathbb{P}_{\mathcal{T}\sim\mathrm{Unif}(\mathscr{T}_{i^\star})}\mathbb{E}^{M_{\mathcal{T}},0}\left[\mathbb{I}\left\{\forall M\in\mathcal{M}_{i^\star}:\ h^M(\widehat{\pi})\geq\delta\right\}\right]-2/7\geq 1/2-1/100-2/7,\end{aligned}$$

其中最后一个不等式由 式 (83)得出。特别地，利用 $M_{\mathcal{T}}\in\mathcal{M}_{i^\star}$ 若 $\mathcal{T}\in\mathscr{T}_{i^\star}$，我们有

$$\mathbb{E}_{\mathcal{T}\sim\mathrm{Unif}(\mathscr{T}_{i^\star})}\mathbb{E}^{M_{\mathcal{T}}}\left[\mathbb{I}\left\{h^{M_{\mathcal{T}}}(\widehat{\pi})\geq\delta\right\}\right]\geq 1/2-1/100-2/7>1/5,$$

这与式（82）矛盾，因此证明完毕。 □

引理 G.3 的证明引理 G.3。我们写作 $\Pi:=\Pi_1\times(\Pi_2\setminus\{0\})=\Pi_1\times\{1,2,\ldots,2q\}\subset\Pi$。首先，我们声称，对于所有$\pi\in\Pi$，以及所有 $M\in\mathcal{M}$，都有 $h^M(\pi)=h_1^M(\pi)+h_2^M(\pi)\geq\delta$。为说明这一点，任取 $M=M_{\mathcal{T}}\in\mathcal{M}$，并考虑以下两种情况：

- 如果 $\pi\notin\mathcal{B}^\star(\mathcal{T})$，那么 $f_2^M((\pi_1,0))-f_2^M(\pi)=0-(-1)=1$。
- 如果 $\pi\in\mathcal{B}^\star(\mathcal{T})$，那么必定存在某个 $\pi_1'\in\Pi_1$，使得 $(\pi_1',\pi_2)\notin\mathcal{B}^\star(\mathcal{T})$，因此 $f_1^M((\pi_1',\pi_2))-f_1^M(\pi)=1-(-\delta)=1+\delta$。

接下来，注意到

$$\max\left\{\mathbb{P}^{M_0}\left(\widehat{\pi}\in\widetilde{\Pi}\cup([n]\times\{0\})\right),\mathbb{P}^{M_0}\left(\widehat{\pi}\in\widetilde{\Pi}\cup([n+1,2n]\times\{0\})\right)\right\}\geq 1/2.$$

我们首先假设 $\mathbb{P}^{M_0}\left(\widehat{\pi}\in\widetilde{\Pi}\cup([n]\times\{0\})\right)\geq 1/2$。注意，如果 $M=M_{\mathcal{T}}\in\mathcal{M}_1$ 和 $\pi\in[n]\times\{0\}$，那么$h^M(\pi)\geq h_2^M(\pi)=f_2^M((\pi_1,\Phi(\mathcal{T})(\pi_1)))-f_2^M(\pi)=\delta-0=\delta$。此外，上述两个项目符号所列的内容表明，~如果 $\widehat{\pi}\in\Pi$，那么 $h^M(\widehat{\pi})\geq 1>\delta$。因此，在这种情况下，我们已经证明 $\mathbb{P}^{M_0}(\forall M\in\mathcal{M}_1,\ h^M(\widehat{\pi})\geq\delta)\geq 1/2$。

在另一种情况下，当 $\mathbb{P}^{M_0}\left(\widehat{\pi}\in\widetilde{\Pi}\cup([n+1,2n]\times\{0\})\right)\geq 1/2$ 时，以对称的方式可得，$\mathbb{P}^{M_0}(\forall M\in\mathcal{M}_2,\ h^M(\widehat{\pi})\geq\delta)\geq 1/2$。 □

引理 G.4 的证明。固定任意 $i\in\{1,2\}$。对于模型 $M\in\{M_0\}\cup\bigcup_{\mathcal{T}\in\mathscr{T}_i}\{M_{\mathcal{T},0}\}$，考虑一次抽样：$\mathfrak{H}^T=(\pi^1,(r_1^1,r_2^1,o_\circ^1)\ldots,\pi^T,(r_1^T,r_2^T,o_\circ^T))\sim\mathbb{P}^M$，其中我们已写出完整观测$o^t=(r_1^t,r_2^t,o_\circ^t)$。由于纯观测的分布 $o_\circ^t\sim M(\pi)$ 不依赖于策略π，因此 $\mathfrak{H}^T$的分布与以下情况相同：首先，从 $M(\pi_0)$ 中独立同分布地抽取 $o_\circ^1,\ldots,o_\circ^T$ （对于任意决策 π_0），然后自适应地选择决策 π^t，$\pi^t\sim q^t(\cdot|\mathfrak{H}^{t-1})$，奖励 r_1^t,r_2^t 则由 π^t 决定。

对于任意 $\mathcal{T}\in\mathscr{T}_i$，对于任意 $t,\ t'\in[T]$ 满足 $t\neq t'$ 的情况，我们有 $\mathbb{P}^{M_{\mathcal{T},0}}[o_\circ^t=o_\circ^{t'}]=1/|\mathcal{T}|\le 3/N$。Th us

$$\mathbb{P}^{M_{\mathcal{T},0}}\left(\exists t\neq t'\ :\ o_\circ^t=o_\circ^{t'}\right)\le T^2\cdot 3/N\le 1/100, \tag{85}$$

其中，最后一个不等式成立是因为 $T\le\sqrt{N/300}$ (只要 C 在引理 G.2的陈述中足够大)。令 $\mathcal{E}\in\mathscr{F}^T$ 表示对所有 $t\neq t',\ o_\circ^t\neq o_\circ^{t'}$ 均成立的事件。由不等式(85) 可得

$$\mathbb{E}_{\mathcal{T}\sim\mathrm{Unif}(\mathscr{T}_i)}\mathbb{E}^{M_{\mathcal{T},0}}\left[\mathbb{I}\left\{\mathcal{E}\right\}\right]\ge 1-1/100. \tag{86}$$

类似地，我们也有：

$$\mathbb{E}^{M_0}[\mathbb{I}\left\{\mathcal{E}\right\}]\ge 1-1/100. \tag{87}$$

现在，我们可以计算了。

$$\begin{aligned}
&\mathbb{E}_{\mathcal{T}\sim\mathrm{Unif}(\mathscr{T}_i)}\mathbb{E}^{M_{\mathcal{T},0}}\left[\mathbb{I}\left\{\pi^t\in\mathcal{B}^\star(\mathcal{T})\right\}\mid\mathcal{E}\right]\\
=&\sum_{\mathcal{T}\in\mathscr{T}_i}\frac{1}{|\mathscr{T}_i|}\sum_{\omega_1,\ldots,\omega_T\in[N]}\mathbb{P}^{M_{\mathcal{T},0}}\left(o_\circ^{1:T}=\omega_{1:T}\mid\mathcal{E}\right)\cdot\mathbb{E}^{M_{\mathcal{T},0}}\left[\mathbb{I}\left\{\pi^t\in\mathcal{B}^\star(\mathcal{T})\right\}\mid o_\circ^{1:T}=\omega_{1:T}\right]\\
=&\sum_{\mathcal{T}\in\mathscr{T}_i}\frac{1}{|\mathscr{T}_i|}\sum_{\substack{\omega_1,\ldots,\omega_T\in\mathcal{T}\\ \text{distinct}}}\frac{1}{N_i(N_i-1)\cdots(N_i-T+1)}\cdot\mathbb{E}^{M_0}\left[\mathbb{I}\left\{\pi^t\in\mathcal{B}^\star(\mathcal{T})\right\}\mid o_\circ^{1:T}=\omega_{1:T}\right]\\
=&\sum_{\substack{\omega_1,\ldots,\omega_T\in[N]\\ \text{distinct}}}\frac{1}{N(N-1)\cdots(N-T+1)}\mathbb{E}_{\mathcal{T}\sim\mathrm{Unif}(\mathscr{T}_i)}\mathbb{E}^{M_0}\left[\mathbb{I}\left\{\pi^t\in\mathcal{B}^\star(\mathcal{T})\right\}\mid o_\circ^{1:T}=\omega_{1:T},\ \{\omega_1,\ldots,\omega_T\}\subset\mathcal{T}\right]
\end{aligned} \tag{88}$$

$\le 2/q,$

其中：

- 第二个等式利用了这样一个事实：在给定 $o^{1:T}$ 的条件下，$\mathfrak{H}^T$ 的分布在 $\mathbb{P}^{M_0}$ 和$\mathbb{P}^{M_{\mathcal{T},0}}$下完全相同。
- 第三个等式交换了求和顺序，并利用了 $1/|\mathscr{T}_i|=\binom{N}{N_i}^{-1}=\frac{N_i!()(N-N_i+1)}{N\ N-1\ \cdots}$，以及包含任意由互不相同的整数 $\omega_1,\ldots,\omega_T\in[N]$组成的元组的集合 $\mathcal{T}$的数量为 $\frac{(N-T)(N-T-1)\cdots(N-N_i+1)\,(N_i-T)!}{}$这一事实。

- 最后一个不等式利用了这样一个事实：对于固定的 $\omega_1,\ldots,\omega_T$，$\mathcal{T}\sim\mathrm{Unif}(\mathscr{T}_i)|\{\omega_1,\ldots,\omega_T\}\subset\mathcal{T}$的分布与 $\mathfrak{H}^T\sim\mathbb{P}^{M_0}|o^{1:}{}_\circ{}^T=\omega_{1:T}$ 的分布相互独立。此外，根据 $\Phi(\mathcal{T})$ 定义 $\mathcal{B}^\star(\mathcal{T})$，并结合 Φ 满足条件 G.2 这一事实可知，对于任意固定的 $\pi=(\pi_1,\pi_2)\in\Pi$ with$\pi_1>0$,

$$\mathbb{P}_{\mathcal{T}\sim\mathrm{Unif}(\mathscr{T}_i)}(\pi\in\mathcal{B}^\star(\mathcal{T})|\{\omega_1,\ldots,\omega_T\}\subset\mathcal{T})=\mathbb{P}_{\mathcal{T}\sim\mathrm{Unif}(\mathscr{T}_i)}(\Phi(\mathcal{T})(\pi_1)=\pi_2|\{\omega_1,\ldots,\omega_T\}\subset\mathcal{T})\le 2/q,$$

其中，当 π_1 不在 $\Phi(\mathcal{T})$ 的定义域内时，我们取 $\Phi(\mathcal{T})(\pi_1)=-1$。（这里我们还用到了$T\le\sqrt{N}$。）特别地，将随机选择的 $\pi^t\sim\mathbb{P}^{M_0}|o_\circ^{1:T}=\omega_{1:T}$替换 π后，上述不等式仍然成立。

对所有项应用并集界 并 T 取值 $t\in[T]$，并应用(86)，引理的第一个结论(80) 即得 s.

为证明引理的第二个结论(81)，注意，对于任意固定的子集 $\mathcal{F}\in\mathscr{F}^T$（不依赖于$\mathcal{T}$），以(88) 结束的等式链蕴含

$$\mathbb{E}_{\mathcal{T}\sim\mathrm{Unif}(\mathscr{T}_i)}\mathbb{E}^{M_{\mathcal{T},0}}[\mathbb{I}\left\{\mathfrak{H}^T\in\mathcal{F}\right\}\mid\mathcal{E}]=\mathbb{E}^{M_0}[\mathbb{I}\left\{\mathfrak{H}^T\in\mathcal{F}\right\}\mid\mathcal{E}],$$

式（81）由上述等式结合（86）和（87）得出。 □

引理 G.5 的证明。令 $\mathcal{E}$ 表示以下事件：$\{\pi^1,\dots,\pi^T\}\cap\mathcal{B}^\star(M_{\mathcal{T}})=\varnothing$。任取历史记录的子集 $\mathcal{F}\subset\mathscr{F}^T$。则

$$\mathbb{P}^{M_{\mathcal{T},0}}(\mathcal{E}\cap\mathcal{F})=\mathbb{P}^{M_{\mathcal{T}}}(\mathcal{E}\cap\mathcal{F}),$$

这是因为对于任意决策 $\pi\notin\mathcal{B}^\star(M_{\mathcal{T}})$，关于完整观测 $o\sim M_{\mathcal{T}}(\pi)$ 和 $o\sim M_{\mathcal{T},0}(\pi)$ 的分布相同。由此，结合引理 B.6即可得出该引理的结论。 □

G.3 补充引理

下面的引理是编码理论中的一个基本事实，给出了里德–所罗门码的维度和距离性质。为陈述该引理，我们回顾汉明距离的定义：对于 $q,n\in\mathbb{N}$ 和 $w,w'\in[q]^n$，令 $d_{\mathrm{Ham}}(w,w')=|\{i\in[n]\colon w_i\neq w'_i\}|$ 表示 w,w' 不同的位置数。

引理 G.6 (里德–所罗门码；Guruswami 等人 (2022年)。 任取满足以下条件的整数 n， q， k。则存在一个映射 $\Phi\colon[q]^k\to[q]^n$ 使得对于任意两个向量 v， $v'\in[q]^k$，若满足 $v\neq v'$，则有 $d_{\mathrm{Ham}}(\Phi(v)\ \Phi(v'))\geq n-k+1$。

此外，可以选择 Φ，使得如果 $X\in[q]^k$ 是均匀随机的，那么对于每个 $i\in[n]$，取值 $\Phi(X)_i\in[q]$ 都是均匀随机的。

引理 G.7 表明，某种随机性提取器 是存在的。

引理 G.7。 存在一个足够大的常数 $C\geq1$，使得以下结论成立。考虑任意正整数 N、N_0、R、q，满足 $N_0\leq2N/3$、$R\leq N/6\leq N_0-N/6$，以及 $3\log q\leq\log\binom{5N/6}{N/6}-C$。令 $\Psi\colon\binom{[N]}{N_0}\to[q]$ 为一个均匀随机函数。那么，对于 Ψ 的选择，以至少 $1-N^{R+1}\cdot2^{-\binom{5N/6}{N/6}/(Cq^2)}$ 的概率，对所有大小为 $|\mathcal{Q}|\leq R$ 的子集 $\mathcal{Q}\subset[N]$ 以及所有 $j\in[q]$，

$$\mathbb{P}_{\mathcal{T}\sim\mathrm{Unif}\binom{[N]}{N_0}}(\Psi(\mathcal{T})=j|\mathcal{Q}\subset\mathcal{T})\leq\frac{2}{q}.$$

我们澄清一下，上述引理陈述中的均匀随机函数 $\Psi\colon\binom{[N]}{N_0}\to[q]$ 的分布定义如下：对于每个 $\mathcal{S}\in\binom{[N]}{N_0}$，$\Psi(\mathcal{S})$ 都是一个独立随机变量，在 $[q]$ 上服从均匀分布。

引理 G.7 的证明。由于 $R\leq N/6\leq N_0-N/6$ 和 $N_0\leq2N/3$，对于任意大小为 $|\mathcal{Q}|\leq R$ 的子集 $\mathcal{Q}\subset[N]$，分布 $\mathcal{T}\sim$均匀分布 $\binom{[N]}{N_0}|\mathcal{Q}\subset\mathcal{T}$ 在任意子集 $\mathcal{T}$ 上赋予的概率质量至多为 $1/\binom{5N/6}{N/6}$（这种分布称为某个 $k\geq\log\binom{5N/6}{N/6}$ 的平坦 k-源）。由 Vadhan（2012年，命题 6.12），取 $\varepsilon=1/q$，对于足够大的常数 C，只要 $3\log q\leq\log\binom{5N/6}{N/6}-C$，那么在选择 Ψ 时，以至少 $1-2^{-\binom{5N/6}{N/6}/(Cq^2)}$ 的概率，对任意大小至多为 R 的固定 $\mathcal{Q}$，分布 $\Psi(\mathcal{T})$（其中 $\mathcal{T}\sim$均匀分布 $\binom{[N]}{N_0}|\mathcal{Q}\subset\mathcal{T}$）与 $[q]$ 上的均匀分布之间的全变差距离至多为 $1/q$。这尤其意味着，对于任意 $j\in[q]$，$\Psi(\mathcal{T})=j$ 的概率至多为 $2/q$（同样地，在 $\mathcal{T}\sim$均匀分布 $\binom{[N]}{N_0}|\mathcal{Q}\subset\mathcal{T}$ 的条件下）。

对所有 $\sum_{r=0}^{R}\binom{N}{R}\leq N^{R+1}$ 可能的集合 $\mathcal{Q}$应用并集界，我们得到 Ψ 以至少 $1-N^{R+1}\cdot2^{-\binom{5N/6}{N/6}/(Cq^2)}$ 的概率满足所需性质。

□