arXiv:2302.02571v1 [cs.LG] 6 Feb 2023

一般函数逼近下马尔可夫博弈中的离线学习

Yuheng Zhang* Yu Bai† Nan Jiang*

2023年2月7日

摘要

我们研究马尔可夫博弈中的离线多智能体强化学习（RL），其目标是从预先收集的离线数据集中学习近似均衡，例如纳什均衡和（粗）相关均衡。现有研究主要考虑较为受限的表格型或线性模型，并分别处理各种均衡。本文提出了首个在一般函数逼近下实现样本高效的马尔可夫博弈离线学习框架，以统一的方式处理全部三种均衡。通过采用贝尔曼一致悲观性，我们对策略回报进行区间估计，并利用上界和下界对候选策略的间隙进行松弛，由此得到我们的优化目标。我们的结果推广了先前的研究，并提供了若干额外见解。值得注意的是，我们要求的数据覆盖条件优于近期提出的"单边集中性"。我们的条件允许选择性地覆盖偏离策略，使其在贪婪程度（作为近似最佳响应）与覆盖之间实现最优权衡；我们还展示了在某些情形下，这能带来显著更好的保证。作为一项新的联系，我们还说明，我们的算法框架如何涵盖专为双人零和博弈这一特殊情形设计、看似不同的解概念。

1 引言

离线强化学习旨在从预先收集的历史数据集中学习一个良好的策略。由于其非交互式特性，离线强化学习已成为将强化学习应用于现实场景的重要范式，尤其适用于在真实系统中部署自适应算法在经济上成本高昂和/或在伦理上存在问题的场景。 [Levine 等人， 2020]。尽管离线强化学习已在单智能体设定中得到广泛研究，但许多现实应用都涉及多个智能体之间的策略性交互。这使得引入博弈论推理成为必要，而在强化学习理论文献中，这类推理通常采用马尔可夫博弈 [Shapley， 1953] 进行建模。马尔可夫博弈可视为马尔可夫决策过程（MDP）的多智能体扩展，其中智能体共享相同的状态信息，系统动态由所有智能体的联合动作决定。

尽管近年来马尔可夫博弈中的在线强化学习取得了显著进展 [Bai and Jin, 2020, Liu et al., 2021, Song et al., 2021, Jin et al., 2021b], ， 马尔可夫博弈中的离线学习才刚刚开始受到学界关注。早期工作 [Cui 和 Du, 2022b, Zhong et al., 2022] 聚焦于表格情形或线性函数逼近，无法处理需要高级函数逼近技术的复杂环境。尽管已有大量关于采用一般函数逼近的单智能体强化学习的研究 [Jiang et al., 2017, Jin et al., 2021a, Wang et al., 2020, Huang et al., 2021a], ， 但这些方法能否以及如何扩展到离线马尔可夫博弈，仍不甚明了。此外，学习目标

*伊利诺伊大学厄巴纳-香槟分校计算机科学系。电子邮箱：yuhengz2@illinois.edu，nanjiang@illinois.edu。

†Salesforce 研究院。电子邮箱：yu.bai@salesforce.com

在马尔可夫博弈中不再是回报优化，而是寻找一个均衡。然而，均衡有多种流行概念，而先前关于离线设定的研究主要聚焦于其中一种（纳什均衡） [Cui 和 Du，2022a，b，Zhong et al.， 2022]。这些考量促使我们研究以下问题：

我们能否为离线马尔可夫博弈设计样本高效算法，在一般函数逼近下处理不同均衡，并将其纳入统一框架？

统一框架 本文给出了信息论结果，对这一问题给出了肯定回答。首先，我们以统一的方式表述了 3 种常见均衡概念——纳什均衡（NE）、相关均衡（CE）和粗相关均衡（CCE）——对应的均衡差距（即我们希望最小化的目标）（第 3 节）。接着，我们以单智能体离线强化学习中的贝尔曼一致悲观框架 [Xieetal.，2021a],为基础，该框架使我们能够构造用于策略评估的置信集，并获得策略回报的置信区间。一个重要的区别在于，Xie et al. [2021a]在单智能体情形下只需要 悲观 评估；相比之下，我们既需要乐观 评估，也需要悲观评估，以进一步计算每个候选策略的均衡差距替代上界，从而能够以可证明的方式获得强有力的离线学习保证（第 4 节）。

关于数据条件的新见解 我们的算法和分析也为马尔可夫博弈的离线可学习性带来了新的认识。在单智能体离线强化学习中，人们已经了解到，只要数据覆盖一个策略，就可以学习到一个良好的策略，而这一条件通常被称为“单策略集中性/覆盖” [Jin 等人、2021c、Zhan et al.、 2022]。相比之下，在马尔可夫博弈中，仅覆盖一个均衡在直觉上是不充分的，因为均衡的一个基本方面是推理其他智能体发生偏离时会发生什么。为解决这一差异，已有工作提出了“单边集中性”这一概念，将其作为离线马尔可夫博弈的充分数据条件 [Cui 和 Du、2022a] （另见Zhonget al. [2022]），该概念要求均衡及其所有单边偏离都得到覆盖。尽管这一条件是充分的，并且在最坏情况下也是必要的，但不那么严格的条件是否同样足够，仍不清楚。我们的工作放宽了这一假设，并提供了更灵活的保证。我们的误差界不再依赖于所有单边偏离策略的最坏情况估计误差，而是体现了策略覆盖误差项与策略次优性项之间的权衡。它会自动适应最优权衡；我们在附录B中展示了若干情形，在这些情形下，该界相较于单边覆盖结果有显著改进 [Cui 和 Du、2022b]。

V型变体 我们的主要算法估计策略的 Q 函数，这些函数将所有智能体的动作作为输入。在表格型设定下，这会导致样本复杂度随智能体数量呈指数增长。虽然可以通过使用强大的函数逼近在联合动作空间上进行泛化来避免这一问题 [Zhong et al.， 2022], 但这会导致我们无法复现并涵盖先前的研究 [Cui 和 Du，2022a、b]。为了解决这一问题，我们提出了算法的 V 型变体，该变体改为估计状态价值函数，并使用重要性采样来校正动作不匹配。它自然避免了指数级依赖，并复现了 [Cui 和 Du，2022b] 的收敛速率（仅有细微差异）；其分析针对表格型设定（第 5节）。

双人零和博弈的新联系 作为一项额外发现，我们揭示了本研究与以往算法思想之间有趣的联系 [Jin 等人，2022，Cui 和 Du，2022b] 这些思想专为双人零和博弈设计。乍看之下，它们似乎截然不同，但我们在附录 B 中证明，这些思想可以纳入我们的算法框架，而且我们的分析与保证也可以直接扩展。

1.1 相关工作

离线强化学习 离线强化学习旨在从预先收集的数据集中学习一个优良策略，而无需与环境直接交互。许多已有研究探讨了单智能体离线强化学习问题，涵盖表格设定 [Yin et al., 2021b,a, Yin and Wang, 2021, Rashidinejad et al., 2021, Xie et al., 2021b, Shi et al., 2022, Li et al., 2022] 以及函数逼近设定 [Antos et al., 2008, Precup, 2000, Chen and Jiang, 2019, Xie and Jiang, 2020, 2021, Xie等人，2021a, Jin et al., 2021c, Zanette et al., 2021, Uehara and Sun, 2021, Yin et al., 2022, Zhan et al., 2022]。值得注意的是，Xie et al. [2021a] 提出了贝尔曼一致悲观性的概念，而我们的技术正是建立在此基础之上。

马尔可夫博弈是多智能体强化学习中广泛使用的框架。马尔可夫博弈的在线学习均衡已得到广泛研究，包括双人零和马尔可夫博弈 [Wei 等人，2017，Bai 和 Jin，2020，Bai 等人，2020，Liu 等人，2021，Dou 等人， 2022], 以及多人一般和马尔可夫博弈 [Liu 等人，2021，Song 等人，2021，Jin 等人，2021b，Mao 和 Bas.ar， 2022]。通常将三种均衡作为学习目标——纳什均衡（NE）、相关均衡（CE）和粗相关均衡（CCE）。近期，一系列研究致力于通过函数逼近求解马尔可夫博弈，包括线性函数逼近 [Xie et al.，2020，Chen 等人， 2022]以及一般函数逼近 [Huang 等人，2021b，Jin 等人， 2022]。一项密切相关的工作是 Jin 等人 [2022], 提出了一种多智能体版本的贝尔曼-埃卢德维度，用于在一般函数逼近下求解零和马尔可夫博弈。然而，他们关注的是在线设定，这与我们的离线设定不同。

离线马尔可夫博弈自从Cui 和 Du [2022的]关于离线表格型零和马尔可夫博弈的开创性工作以来，已有多项后续研究探讨离线马尔可夫博弈，涉及表格型零和/一般和马尔可夫博弈 [Cui 和 Du、2022b、Yan 等人， 2022]以及线性函数逼近 [Zhonget al.、2022、Xiong et al.， 2022]。在本文中，我们研究多人一般和马尔可夫博弈中的一般函数逼近，这是一种更一般的框架。在技术上，我们与这些先前工作的区别在于处理策略评估中不确定性量化的方式；这是离线学习中的一个重要技术问题：我们在策略评估中采用初始状态乐观/悲观性，而先前的工作则依赖带有奖励项的前状态悲观性。此外，先前的工作要求满足所谓的数据覆盖“单边集中性”假设。[1] 尽管在最坏情况下这一假设不可避免，但我们的方法所需的条件从不比它更差（且在最坏情况下与之相同），并且在某些实例中可以显著更好。

2 预备知识

记号 我们用 $\Delta(\cdot)$ 表示概率单纯形。我们用粗体字母表示向量，例如 $\mathbf{a}$，而 $\mathbf{a}$ 的 j^{th} 个元素记作 $\mathbf{a}_j$。我们用 $-i$ 表示除玩家 i 以外的所有玩家。对于正整数 m，$[m]$ 表示集合 $\{1, 2, \cdots, m\}$。$\|f\|_{2,\,d}^2$ 表示 $\mathbb{E}_d[f^2]$，而 $f(s,\ \pi)$ 表示 $_\sim\mathbb{E}_{a\sim\pi(\cdot|s)}[f(s,\ a)]$。我们用 $\mathcal{O}(\cdot)$ 隐去绝对常数，用 $\mathcal{O}(\cdot)$ 隐去对数因子。

2.1 多人一般和马尔可夫博弈

我们考虑无限时域折扣设定下的多人一般和马尔可夫博弈。这种马尔可夫博弈由 $(\mathcal{S},\ \mathcal{A} = \prod_{i\in[m]}\mathcal{A}_i,\ P,\ r,\ \gamma,\ s_0)$ 指定，其中 $\mathcal{S}$ 是状态空间，$|\mathcal{S}| = S$，$\mathcal{A}_i$ i $|\mathcal{A}_i| = A_i$ $\in \mathcal{A}$ m 是玩家的动作空间，一个是所有玩家采取的联合动作，

[1] Zhong et al. [2022] 提出了“相对不确定性”的概念，它是“单边集中性”的线性版本 .

P： $\mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ 是转移函数， $P(\cdot|s,$ $\mathbf{a})$ 描述在状态 s 采取联合动作 $\mathbf{a}$ 时下一状态的概率分布；$r = \{r_i\}_{i\in[m]}$ 是奖励函数的集合，其中r_i： $\mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$ 是玩家 i 的确定性奖励函数；$\gamma \in [0, 1)$ 是折扣因子，而 s_0 是固定的初始状态，这不失一般性。

乘积策略与相关策略 马尔可夫联合策略 π： $\mathcal{S} \to \Delta(\mathcal{A})$ 指定了所有玩家的决策策略，并诱导出一条轨迹 s_0， $\mathsf{a}_0, \mathsf{r0}, s_1, \mathsf{a}_1, \mathsf{r1}, \cdots, s_t, \mathsf{a}_t, \mathsf{r}_t, \cdots$， 其中 $\mathsf{a}_t \sim \pi(\cdot|s_t)$、$\mathsf{r}_{t,i} = r_i(s_t, \mathsf{a}_t)$ 和 $s_{t+1} \sim P(\cdot|s_t, \mathsf{a}_t)$。对于联合策略 π， π_i 是玩家i 的边缘策略，而 π_{-i} 是其余玩家的边缘策略。若联合策略 π 满足$\pi = \pi_1 \times \pi_2 \times \cdots \pi_m$，且每个玩家 i 都根据 π_i 独立地采取动作，则称其为乘积策略。若 π 不是联合策略，有时我们称 π 是相关的，此时玩家需要依据公共随机性来决定自己的动作。

价值函数与占用度 对于玩家 i 和联合策略 π，我们定义价值函数 $V_i^\pi(s) := \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t r_i(s_t, \textbf{动作}_t)|s_0 = s]$ 以及 Q函数 $Q_i^\pi(s, \textbf{动作}) := \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t r_i(s_t, \textbf{动作}_t)|s_0 = s, a_0 = \textbf{动作}]$，它们在 $[0, V_{\max}]$ 中有界，其中 $V_{\max} = R_{\max}/(1-\gamma)$。对于每个联合策略 π，第 i^{th} 个玩家 $\mathcal{T}_i^\pi$ 的策略特定贝尔曼算子 $\mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ 定义为

$$(\mathcal{T}_i^\pi f)(s, \mathbf{a}) = r_i(s, \mathbf{a}) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,\mathbf{a})}[f(s', \pi)],$$

且 Q_i^π是 $\mathcal{T}_i^\pi$的唯一不动点。请注意，一旦策略固定，博弈论方面的考量便不再相关，价值函数的定义方式也与单智能体设定中的常见方式类似，唯一的区别在于，由于每位玩家的奖励函数各不相同，每位玩家 i 都有自己的价值函数 r_i。与单智能体情形类似，我们还考虑折扣状态-动作占用度$d^\pi(s,$ **动作**$) \in \Delta(\mathcal{S} \times \mathcal{A})$，其定义为 $d^\pi(s,$ **动作**$) = (1-\gamma)\mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t \mathbb{I}[s_t = s, \textbf{动作}_t = \textbf{动作}]]$。

2.2 马尔可夫博弈的离线学习

在离线学习设定中，我们假设可以访问预先收集的数据集，且无法再与环境进行交互。离线数据集 $\mathcal{D}$ 由 n 个独立元组 $(s,$ **动作**$,$ **奖励** $s')$ 组成，这些元组由某个数据分布 $d_D \in \Delta(\mathcal{S} \times \mathcal{A})$ 按照 $(s,$ **动作**$) \sim d_D,$ **奖励** $_i \sim r_i(s,$ **动作**$)$ 和 $s' \sim P(\cdot|s,$ **动作**$)$ 的方式生成。[2]

策略类在状态空间较大的实际问题中，所有可能的马尔可夫联合策略构成的空间过于庞大，难以处理。为此，我们假设给定了一个预先指定的策略类 $\Pi \subset (\mathcal{S} \to \Delta(\mathcal{A}))$，并希望从中找到一个在给定准则下近似达到均衡的策略。[3] 令 $\Pi_i = \{\pi_i : \pi \in \Pi\}$ 表示玩家 i所诱导的边缘化策略类，并类似地定义Π_{-i}。

扩展类 正如我们将在第 3 节中看到的，均衡的一个基本方面是反事实推理：其他智能体会如何偏离给定策略并作出回应。在考虑玩家 i 针对每个策略 $\pi \in \Pi$可能采取的偏离行为后，我们得到一个扩展类 $\Pi_i^{\text{ext}} \supseteq \Pi$ 用于

[2]对于非独立同分布自适应数据，我们可以在分析中使用鞅集中不等式。若不作进一步的混合型^假设，只需将定义（例如 d_D （一个静态对象）中的（1）和（2））替换为 d_D，即状态-动作对上的经验分布。由此得到的定义（2）例如，对应于 $\widehat{C}$（π）中的量Cui 和 Du [2022b针对表格型设定定义的定义 3]。本文仅考虑在一类平稳马尔可夫策略中最小化均衡差距。关于它们如何满足纳什均衡和粗相关均衡等标准均衡概念，见Daskalakis 等人[2022] 及其中的参考文献；关于相关均衡的情形，见Nowak 和 Raghavan [1992]。下文也仅考虑用于纳什均衡/粗相关均衡的平稳马尔可夫响应策略，这也有理据：一旦固定一个平稳马尔可夫 π_{-i} 策略，优化玩家 i的行为以求最佳响应，就转化为一个单智能体马尔可夫决策过程问题。

玩家 i。 Π_i^{ext}的具体形式将在第3 节中定义，并且可能取决于所考虑的均衡概念。目前只需指出，Π_i^{ext}是 Π 的超集，其中包含玩家 i需要进行推理的所有策略。

价值函数近似 我们使用 $\mathcal{F}_i \subset (\mathcal{S} \times \mathcal{A} \to [0, V_{\max}])$ 来近似每个玩家 Q_i^π的 Q函数 i。遵循 [Xie et al., 2021a], , 我们对 $\mathcal{F}_i$作出两个标准假设：

假设 A (近似可实现性)。 对于任意玩家 $i \in [m]$ 和任意 $\pi \in \Pi_i^{\text{ext}}$，我们有

$$\inf_{f \in \mathcal{F}_i} \sup_{\text{admissible } d} \|f - \mathcal{T}_i^\pi f\|_{2,d}^2 \le \varepsilon_{\mathcal{F}},$$

若 $d \in \{d^{\pi'} : \pi' \in \Pi_i^{\text{ext}}\} \cup d_D$，则数据分布 d 是可容许的。

对于每个玩家 i 和联合策略 π，假设 A 要求存在 $f \in \mathcal{F}_i$，使得 f 在由扩展策略类 Π_i^{ext}和数据分布诱导的所有可能分布下，贝尔曼误差都很小。当 $Q_i^\pi \in \mathcal{F}_i$， $\forall \pi \in \Pi$， $i \in [m]$时，我们有 $\varepsilon_{\mathcal{F}} = 0$。

假设 B (近似完备性)。 对于任意玩家 $i \in [m]$ 和任意 $\pi \in \Pi_i^{\text{ext}}$，我们有

$$\sup_{f \in \mathcal{F}_i} \inf_{f' \in \mathcal{F}_i} \left\|f' - \mathcal{T}_i^\pi f\right\|_{2,d_D}^2 \le \varepsilon_{\mathcal{F},\mathcal{F}}. \tag{1}$$

假设B 要求 $\mathcal{F}_i$ 在算子 $\mathcal{T}_i^\pi$下近似封闭。这两个假设都是对离线强化学习文献中广泛使用的对应假设的直接扩展。

分布失配与数据覆盖与Xie et al. [2021a], 类似，我们使用 π 下贝尔曼误差的差异度来衡量任意分布 d 与数据分布 d_D之间的分布失配：

$$\mathscr{C}(d; d_D, \mathcal{F}_i, \pi) := \max_{f \in \mathcal{F}_i} \frac{\|f - \mathcal{T}_i^\pi f\|_{2,d}^2}{\|f - \mathcal{T}_i^\pi f\|_{2,d_D}^2}. \tag{2}$$

我们指出， $\mathscr{C}(d; d_D, \mathcal{F}_i, \pi) \le \sup_{s,\mathbf{a}} \frac{d(s,\mathbf{a})}{d_D(s,\mathbf{a})}$，这意味着 $\mathscr{C}(d; d_D, \mathcal{F}_i, \pi)$ 是比原始密度比更精确的度量。

3 均衡

我们考虑博弈论中的三种常见均衡：纳什均衡（NE）、相关均衡（CE）和粗相关均衡（CCE）。我们使用响应类映射这一概念，以统一的方式定义这三种均衡，使得每种均衡都是相对于相应响应类中的相对最佳响应来定义的。

响应类映射 $\Pi^\dagger(\cdot)$ 将策略 π 映射到类， $\Pi^\dagger(\pi) := \bigcup_{i \in [m]} \Pi_i^\dagger(\pi)$。粗略地说， $\Pi_i^\dagger(\pi)$ 是通过以下方式得到的：先取一个候选策略 π，再考虑玩家 i 如何以各种方式将其行为从 π_i 偏离到 $\pi_i^\dagger$，然后将 $\pi_i^\dagger$ 和 π_{-i} 重新组合成联合策略。玩家 i 可选择的 $\pi_i^\dagger$ 空间决定了该映射；正如接下来将要说明的，在不同的均衡概念下，这个空间会有不同的形式。

[4]因此 ，策略类 $\Pi_i^\dagger(\pi)$ 始终满足以下条件：对于任意 $i \in [m]$ 以及任意 $\pi' \in \Pi_i^\dagger(\pi)$， $\pi'_{-i} = \pi_{-i}$.

1. 如果乘积策略满足任何玩家都无法通过偏离自己的策略来增加收益，那么它就是纳什均衡。因此，纳什均衡对应的响应类映射定义为 $\Pi^{\dagger,\mathrm{NE}}(\pi) := \{\Pi_i^{\dagger,\mathrm{NE}}(\pi)\}_{i\in[m]}$，其中 $\prod^{\dagger}{}_{,i}{}^{\mathrm{NE}}(\pi) := \{\pi_i^\dagger \times \pi_{-i} : \pi_i^\dagger \in \Pi_i\}$。请注意，此处 $\Pi^{\dagger}{}_{,i}{}^{\mathrm{NE}}$ 不依赖于输入 π_i，而玩家i 只需考虑采用某个 $\pi_i^\dagger \in \Pi_i$ 来替换 π_i。

2. 相关均衡（CE）由一类策略修改 $\Phi = (\Phi_i)_{i\in[m]}$定义，其中，$\Phi_i \subseteq (\mathcal{S} \times \mathcal{A}_i \to \mathcal{A}_i)$是玩家 i^{th} 的策略修改集合，且每个 $\phi_i \in \Phi_i$ 都是一个映射 $\phi_i : \mathcal{S} \times \mathcal{A}_i \to \mathcal{A}_i$。对于任意联合策略 π，修改后的策略 $\phi_i \diamond \pi$ 定义如下：在状态 $s \in \mathcal{S}$下，所有玩家采样一个 $\sim \pi(\cdot|s)$，玩家i^{th} 将动作$\mathbf{a}_i$ 改为 $\phi_i(s,\ a_i)$，而$\mathbf{a}_{-i}$ 保持不变。对于相关均衡（CE），每个联合策略 π的响应类映射定义为 $\Pi^{\dagger}(\pi) := \{\Pi_i^\dagger(\pi)\}_{i\in[m]}$，其中 $\Pi_i^\dagger(\pi) = \{(\phi_i \diamond \pi_i) \odot \pi_{-i} : \phi_i \in \Phi_i\}$。

3. 粗相关均衡（CCE）针对一般的（即可能相关的）联合策略定义，是纳什均衡（NE）的一种松弛。唯一的区别在于，粗相关均衡（CCE）不要求候选策略 π 是乘积策略。因此，粗相关均衡（CCE）的响应类映射与纳什均衡（NE）相同。

根据响应类映射的定义，对于均衡 $\in$ {纳什均衡、相关均衡、粗相关均衡}，我们将任意联合策略 π 相对于 $\Pi^{\dagger,\ \mathrm{EQ}}(\cdot)$ 的间隙定义为

$$\mathrm{Gap}^{\Pi^{\dagger,\mathrm{EQ}}}(\pi) := \max_{i\in[m]} \max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} V_i^{\pi^\dagger}(s_0) - V_i^{\pi}(s_0).$$

现在，我们可以给出三种均衡的定义了。

定义 1（均衡；纳什均衡、相关均衡和粗相关均衡）。对于均衡 $\in$ {纳什均衡、相关均衡和 粗相关均衡}，若某个联合策略（纳什均衡下为乘积策略）相对于 $\Pi^{\dagger,\ \mathrm{EQ}}(\cdot)$ 是一个 ε-均衡，则对于响应类 $\Pi^{\dagger,\ {}_i{}^{\mathrm{EQ}}}(\pi)$，

$$\mathrm{Gap}^{\Pi^{\dagger,\mathrm{EQ}}}(\pi) \le \varepsilon.$$

定义 1 是相对于策略类 Π 和策略修改类 Φ （对于相关均衡）定义的。全文中，我们关注此类“类内”间隙的理论保证；如果假设所有玩家的表示能力有限，必须使用受限策略类，那么这是一种合理的间隙定义。在附加假设下（我们称之为“策略完备性”；参见附录 A），此类“类内”间隙可以与一种更强的间隙概念联系起来，在该概念中，最佳响应考虑的是不受限制的偏离策略。

借助响应类映射，我们还定义了扩展策略类 $\Pi_i^{\mathrm{ext}} := (\bigcup_{\pi\in\Pi} \Pi_i^\dagger(\pi)) \bigcup \Pi$，它刻画了所有可能的、相对于 i^{th} 玩家发生偏离的策略。此外，我们定义 $\Pi^{\mathrm{ext}} := \bigcup_{i=1}^m \Pi_i^{\mathrm{ext}}$。

4 多玩家一般和马尔可夫博弈的信息论结果

4.1 算法

由于我们的学习目标是找到一项均衡差距较小的策略 $\pi \in \Pi$，即 $\mathrm{Gap}^{\Pi^{\dagger},\ \mathrm{EQ}}(\pi)$（对于 EQ∈{NE、CE、CCE}），一种自然的想法是直接估计该差距，并在 $\pi \in \Pi$ 上将其最小化。不幸的是，我们处于离线设定，只能访问从任意数据分布d_D中采样的数据，而这些数据可能不足以评估某些策略的差距。

由于间隙并非总是可以估计的，我们转而寻求一种替代目标，使其始终是每个候选策略 $\pi \in \Pi$ 的均衡差距上界 。该上界应当

算法1基于离线数据的Bellman一致均衡学习（BCEL） et

1: **输入：** 离线数据集 $\mathcal{D}$、参数 β_f、均衡 EQ $\in$ {纳什均衡、 相关均衡、粗相关均衡}2: 对每个玩家 $i \in [m]$ 和策略 $\pi \in \Pi_i^{\text{ext}}$，构造函数版本空间

$$\mathcal{F}^{\pi}{}_{,i}{}^{\beta_f} = \{f_i \in \mathcal{F}_i \colon \mathcal{E}_i(f_i, \pi; \mathcal{D}) \leq \beta_f\}. \tag{3}$$

3: 对每个玩家 $i \in [m]$，计算

$$\overline{V}_i^{\pi^\dagger}(s_0) = \max_{f \in \mathcal{F}_i^{\pi^\dagger, \beta_f}} f(s_0, \pi^\dagger), \quad \forall \pi^\dagger \in \Pi_i^{\text{ext}}. \tag{4}$$

$$\underline{V}_i^{\pi}(s_0) = \min_{f \in \mathcal{F}_i^{\pi, \beta_f}} f(s_0, \pi), \quad \forall \pi \in \Pi. \tag{5}$$

4: 对每个策略 $\pi \in \Pi$，计算估计间隔

$$\widehat{\text{Gap}}_{\text{EQ}}(\pi) \coloneqq \max_{i \in [m]} \max_{\pi^\dagger \in \Pi_i^{\dagger, \text{EQ}}(\pi)} \overline{V}_i^{\pi^\dagger}(s_0) - \underline{V}_i^{\pi}(s_0). \tag{6}$$

5: 输出 $\hat{\pi} \leftarrow \min_{\pi \in \Pi}$间隙EQ （？）

当策略被数据覆盖且我们拥有足够信息来准确确定其间隙时，该方法也具有紧确性。为实现这一目标，我们回顾间隙的定义：

$$\text{Gap}^{\Pi^{\dagger}, \text{EQ}}(\pi) \coloneqq \max_{i \in [m]} \max_{\pi^\dagger \in \Pi_i^{\dagger, \text{EQ}}(\pi)} V_i^{\pi^\dagger}(s_0) - V_i^{\pi}(s_0).$$

我们算法的核心思想是

$$V_i^{\pi^\dagger}(s_0) - V_i^{\pi}(s_0) \leq \overline{V}_i^{\pi^\dagger}(s_0) - \underline{V}_i^{\pi}(s_0),$$

其中

- $\overline{V}_i^{\pi^\dagger}(s_0) \geq V_i^{\pi^\dagger}(s_0)$ 是对 $\pi^\dagger$ 的乐观评估。
- $\underline{V}_i^{\pi}(s_0) \leq V_i^{\pi}(s_0)$ 是对 π 的悲观评估。

通过这种松弛，该问题可归结为乐观和悲观策略评估，我们可以借鉴单智能体强化学习中的现有技术。

贝尔曼一致悲观性与乐观性我们采用Xie et al. [2021提出的] 贝尔曼一致悲观性框架来构造乐观和悲观策略评估。对于每位玩家 i，我们首先使用数据集 $\mathcal{D}$ 计算所有函数在贝尔曼算子 $f_i \in \mathcal{F}_i$ 下的经验贝尔曼误差 $\mathcal{T}_i^{\pi}$，

$$\mathcal{E}_i(f_i, \pi; \mathcal{D}) \coloneqq \mathcal{L}_i(f_i, f_i, \pi; \mathcal{D}) - \min_{f_i' \in \mathcal{F}_i} \mathcal{L}_i(f_i', f_i, \pi; \mathcal{D}),$$

$$\mathcal{L}_i(f_i', f_i, \pi; \mathcal{D}) \coloneqq \frac{1}{n} \sum_{(s, \mathbf{a}, \mathbf{r}, s') \in \mathcal{D}} \left(f_i'(s, \mathbf{a}) - \mathbf{r}_i - \gamma f_i(s', \pi)\right)^2.$$

与单智能体设定类似， $\mathcal{E}_i(f_i, \pi; \mathcal{D})$ 是 f_i 关于 π 的真实贝尔曼误差的良好近似，即 $\mathcal{E}_i(f_i, \pi; \mathcal{D}) \approx \|f_i - \mathcal{T}_i^{\pi} f_i\|_{2, d_D}^2$，因此我们可以为每个玩家 i 构造一个版本空间 $\mathcal{F}^{\pi}{}_{, i}{}^{\beta_f}$

和策略 $\pi \in \Pi_i^{\text{ext}}$在 (3) 中。为确保 Q_i^π 的最佳近似包含在 $\mathcal{F}^{\pi,\beta_f}_{,i}$ 中，给定失败概率 $\delta > 0$，我们按如下方式选取阈值参数 β_f，

$$\beta_f = \frac{80V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} + 30\varepsilon_{\mathcal{F}},$$

其中 $\mathcal{F} = \bigcup_{i=1}^m \mathcal{F}_i$。随后，只需分别取版本空间中所有函数对初始状态 s_0 的最高和最低预测值，即可得到乐观评估和悲观评估（(4) 和（5）。

有了$\overline{V}_i^{\pi^\dagger}(s_0)$ 和 $\underline{V}_i^\pi(s_0)$ 之后，我们为每个 $\widehat{\ }$估计间隔$^{\text{EQ}}(\pi)$ 计算估计间隙 $\pi \in \Pi$ 在 (6)中。我们选择估计间隙最低的策略 $\widehat{\pi}$，算法总结见算法 1。

4.2 理论保证

在介绍理论保证之前，我们先引入区间宽度 $\Delta_i^\pi$of $\mathcal{F}^{\pi,\beta_f}_{,i}$，它将在主定理的表述中发挥关键作用：

$$\Delta_i^\pi := \max_{f_i \in \mathcal{F}_i^{\pi,\beta_f}} f_i(s_0,\pi) - \min_{f_i \in \mathcal{F}_i^{\pi,\beta_f}} f_i(s_0,\pi).$$

正如我们将看到的，Δ_i^π是衡量数据分布 d_D 对 d^π（即 π的状态-动作占用）的覆盖程度。覆盖越好，Δ_i^π越小。以下命题对此进行了形式化：

命题 2（区间宽度的界）。 以至少 $1-\delta$ 的概率，对于任意玩家 $i \in [m]$ 和任意$\pi \in \Pi_i^{\text{ext}}$，都有

$$\Delta_i^\pi \le \min_d \frac{1}{1-\gamma}\sqrt{\mathscr{C}(d; d_D, \mathcal{F}_i, \pi)}\varepsilon_{\text{apx}} + \frac{1}{1-\gamma}\sum_{s,\mathbf{a}}(d^\pi \setminus d)(s,\mathbf{a})\left[\Delta f_i^\pi(s,\mathbf{a}) - \gamma(P^\pi \Delta f_i^\pi)(s,\mathbf{a})\right], \quad (7)$$

where $\varepsilon_{\text{apx}} = \mathcal{O}\left(V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}\right)$, $(d^\pi \setminus d)(s,\mathbf{a}) := \max(d^\pi(s,\mathbf{a}) - d(s,\mathbf{a}), 0)$, $\Delta f_i^\pi(s,\mathbf{a}) := f_i^{\pi,\max}(s,\mathbf{a}) - f_i^{\pi,\min}(s,\mathbf{a})$, *and* $(P^\pi f)(s,\mathbf{a}) = \mathbb{E}_{s' \sim P(\cdot|s,\mathbf{a})}[f(s',\pi)]$.

这里，引入式（7）中的分布 $d \in \Delta(\mathcal{S} \times \mathcal{A})$，以处理 d_D 与 d^π之间的差异度。式（7）中的第一项刻画了 d 与 d_D之间的分布失配，第二项则表示在 π 下的支持集外贝尔曼误差。当数据分布 d_D 在 d^π上实现完全覆盖时，d 可取为 d^π，此时第二项为零。因此，为了便于建立直觉，我们总可以选择 $d = d^\pi$ 并将 $\Delta_i^\pi \propto \sqrt{\mathscr{C}(d^\pi; d_D, \mathcal{F}_i, \pi)}$视为如此；不过，一般而言，某些 $d \neq d^\pi$ 可能实现更好的权衡，并得到更紧的界。

对 Δ_i^π有了直观的理解后，我们可以给出以下定理，以说明所提出算法的性质。

定理3。 以至少 $1-\delta$的概率，对于任意 $\pi \in \Pi$和 均衡概念 $\in$ {纳什均衡、 相关均衡、粗相关均衡}，算法 $\widehat{\pi}$的输出策略$\widehat{\pi}$满足

$$\text{Gap}^{\Pi^\dagger,\text{EQ}}(\widehat{\pi}) \le \text{Gap}^{\Pi^\dagger,\text{EQ}}(\pi) + \frac{4\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} + \max_{i\in[m]} \min_{\widetilde{\pi}_i \in \Pi_i^{\dagger,\text{EQ}}(\pi)} \left(\Delta_i^{\widetilde{\pi}_i} + \Delta_i^\pi + \text{subopt}_i^\pi(\widetilde{\pi}_i)\right),$$

其中 $\text{subopt}_i^\pi(\widetilde{\pi}_i) := \max_{\pi^\dagger \in \Pi_i^\dagger(\pi)} \overline{V}_i^{\pi^\dagger}(s_0) - \underline{V}_i^{\widetilde{\pi}_i}(s_0)$。

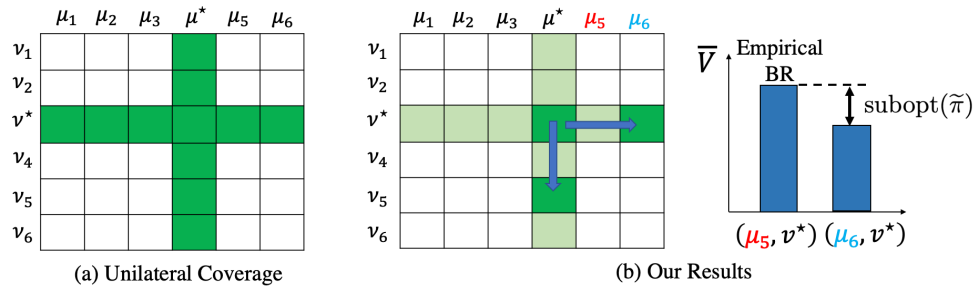

图 1：零和示例中的单边覆盖示意图及我们的结果。（**a**）单边覆盖要求数据集覆盖所有单边策略对（$\mu^\star$，ν'）和（μ'，$\nu^\star$），其中（$\mu^\star$，$\nu^\star$）是纳什均衡。（**b**）我们的方法具有自适应性，放宽了条件。首先，如果数据仅覆盖乐观最佳响应（在此示例中为（μ_5，$\nu^\star$）），即当$\Delta^{\mu_5,\,\nu^\star}$ 较小时，我们就已经能够实现较好的样本复杂度。即使数据集对（μ_5，$\nu^\star$）的覆盖较差，也可能存在某个其他的μ_6，使得 $\Delta^{\mu_5,\,\nu^\star} \gg \Delta^{\mu_6,\,\nu^\star}$。我们的方法不会承受 $\Delta^{\mu_5,\,\nu^\star}$带来的影响，而是会自动适应策略$\widetilde{\pi} =$ （μ_6，$\nu^\star$），在策略覆盖项 Δ^π 与次优性项 subopt($\widetilde{\pi}$)之间实现更好的权衡。

4.3 单边覆盖的改进

为了解读定理 3 并将其与现有保证进行比较，我们首先给出定理 3 + 命题 2的一个直接推论。该推论是我们结果的一种弱化形式，与 Cui 和 Du [2022a，b]。

推论 4。对于纳什均衡策略 $\pi^\star \in \Pi$，假设存在单边系数 $C(\pi^\star)$，使得下列不等式成立

$$\max_{i\in[m]} \max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{NE}}(\pi^\star)} \mathscr{C}(d^{\pi^\dagger}; d_D, \mathcal{F}_i, \pi^\star) \le C(\pi^\star). \tag{8}$$

以至少 $1-\delta$ 的概率，我们有

$$\mathrm{Gap}^{\Pi^\dagger,\mathrm{NE}}(\widehat{\pi}) \le \mathcal{O}\left(\frac{V_{\max}\sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\mathrm{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}}{1-\gamma} \sqrt{C(\pi^\star)} \right).$$

推论 4 中的差距上界形式简洁：第一部分包含一个 $O(1/\sqrt{n})$ 统计误差（其规模受函数类和策略类的复杂度影响，这完全在预期之中），以及一个依赖于 $\varepsilon_{\mathcal{F}}, \varepsilon_{\mathcal{F},\mathcal{F}}$ 的近似误差项；当我们的函数类完全可实现且贝尔曼完备时，该误差项趋于 0。

界中的关键项是 $\sqrt{C(\pi^*)}$ 因子，它衡量分布失配，并隐式决定数据覆盖条件。 $C(\pi^*)$ 在式 (8) 中定义。可以看出，要使 $C(\pi^*)$较小，数据不仅要覆盖 π^* 自身[5]，还必须覆盖 $\Pi^\dagger$ 中的所有策略，$_i^{\mathrm{NE}}(\pi^\star) = \{\pi_i \times \pi^\star_{-i} : \pi_i \in \Pi_i\}$。这就是 Cui 和 Du [2022a] 以及 Zhong et al. 提出的单边覆盖概念 [2022]。在图 1(a) 中，我们通过一个简化的双人矩阵博弈场景对此进行可视化；这样的条件意味着数据覆盖以 NE 为中心的整个“十字形”区域。

[5]注意，$\pi^* \in \Pi^\dagger_{\cdot,i}{}^{\mathrm{NE}}(\pi^\star)$。

尽管 [Cui 和 Du，2022a] 认为单边覆盖在最坏情况下“充分且必要”，但他们的论证并未排除一种改进版本：在某些条件下，该版本可以大幅放宽。我们的定理 3 正是这样一种版本。现在，我们来分解定理 3中的界：

1. 首先，界的右侧取决于 Δ_i^π，其中 π 是我们要与之竞争的策略，对应于推论 π^*中的 4。回想一下，$\Delta_i^\pi \propto \sqrt{\mathscr{C}(d^\pi; d_D, \mathcal{F}_i, \pi)}$这一项对应于 π^*上的数据覆盖；若希望与 π^*竞争，这始终是必需的。

2. 右侧还取决于 $\Delta_i^{\widetilde{\pi}_i} + \text{subopt}_i^\pi(\widetilde{\pi}_i)$，其中 $\widetilde{\pi}_i$ 在 $\Pi^\dagger_{\cdot,i}{}^{\text{NE}}(\pi^\star)$ 上取最小值，当 EQ=NE 时（而(8) 则在 $\pi^\dagger$ 上取最大值）。特别地，我们总可以选择 $\widetilde{\pi}_i$ 作为使 V_i 最大化的策略，也就是乐观最佳响应。这样一来，$\text{subopt}_i^\pi(\widetilde{\pi}_i) = 0$，这表明我们只需要覆盖乐观最佳响应策略，而不必像单边假设所要求的那样覆盖 $\Pi^\dagger_{\cdot,i}{}^{\text{NE}}(\pi^\star)$中的所有策略。

3. 最后，我们的界还提供了进一步的松弛：当乐观最佳响应覆盖不足时，我们可以改选其他覆盖良好的 $\widetilde{\pi}_i$，并额外付出一项 $\text{subopt}_i^\pi(\widetilde{\pi}_i)$，该项衡量 $\widetilde{\pi}_i$ 在多大程度上是基于 V_i 的近似最佳响应。

再次，我们在图 1(b) 中展示了我们的界所具有的灵活性。下面我们还给出一个具体示例，在该示例中，与单边条件所能提供的保证相比，我们的保证显著改善了样本速率。

示例 考虑一个简单的双人零和矩阵博弈，其收益矩阵为 ：

	b_1	b_2	b_3
a_1	0.5	0.75	0.75
a_2	0.25	0	0
a_3	0.25	0	0

其中列玩家旨在最大化收益，而行玩家旨在最小化收益。显然，(a_1, b_1) 是纳什均衡。离线数据集 $\mathcal{D}$ 是从以下分布中收集的，

	b_1	b_2	b_3
a_1	p_1	p_2	p_2
a_2	p_2	p_3	p_3
a_3	p_2	p_3	p_3

其中 $0 < p_2 \ll p_1$ 和 $p_3 = \frac{1-p_1-4p_2}{4}$。根据推论 4（即单边覆盖 [Cui 和 Du，2022a]），样本复杂度界为 $\widetilde{\mathcal{O}}(\frac{1}{p_2\epsilon^2})$。然而，当 $n > \widetilde{\mathcal{O}}(\frac{1}{p_2})$ 时，我们已经能以高概率识别出 (a_1, b_2)、(a_1, b_3)、(a_2, b_1) 和 (a_3, b_1) 是次优动作。在这种情况下，定理 3 表明，我们只需承担乐观最佳响应（其本身就是 (a_1, b_1)）上的覆盖系数，因此样本复杂度界变为 $\widetilde{\mathcal{O}}(\max\{\frac{1}{p_2}, \frac{1}{p_1\epsilon^2}\}) \ll \widetilde{\mathcal{O}}(\frac{1}{p_2\epsilon^2})$。

5 V 型函数逼近

我们在第4 节所采用方法的一个潜在缺陷是，我们对以联合动作为输入的 Q 函数进行建模。在表格型设定中，完整 Q 函数类的复杂度会随智能体数量呈指数级增长 m，而先前针对表格型设定的研究结果则不存在这种依赖。

尽管已知，对动作进行联合特征化可以避免这种指数级依赖 [Zhonget al.， 2022]（这与线性 MDP 结果在单智能体情形下不会产生 $|\mathcal{A}|$ 依赖的方式类似

设定中 [Jin 等人， 2020]），本节提供一种替代方法，可直接涵盖此前的表格结果，并得到相同的速率（除细微差异外，后文将对此进行讨论）。我们提出了一个BCEL的 V 型变体算法，借助函数类 V_i^π，直接对状态价值函数进行建模 $\mathcal{G}_i \subset (\mathcal{S} \to [0, V_{\max}])$，每位玩家 i各自使用一个。

与之前一样，我们假设元组 $(s, \mathbf{a}, \mathbf{r}\ s') \in \mathcal{D}$ 由 $(s, \mathbf{a}) \sim d_D$、$\mathbf{r}_i \sim r_i(s, \mathbf{a})$ 和$s' \sim P(\cdot|s, \mathbf{a})$ 生成。在本节中，我们记作 $d_D = d_S \times d_A$，即 $(s, \mathbf{a}) \sim d_D \Leftrightarrow s \sim d_S, a \sim d_A(\cdot|s)$。此外，我们假设 (1) $d_A(\mathbf{a}|s) > 0, \forall(s, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$,[6] ；以及 (2) 学习者已知行为策略 $d_A(\cdot|s)$。我们利用行为策略对动作进行重要性加权，以纠正 d_A 与 π 之间的不匹配，并按如下方式修改损失函数 $\mathcal{L}$：对于任意函数 $g_i \in \mathcal{G}_i$，定义

$$\mathcal{L}_i(g_i', g_i, \pi; \mathcal{D}) \coloneqq \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s') \in \mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(g_i'(s, \mathbf{a}) - \mathbf{r}_i - \gamma g_i(s', \pi)\right)^2.$$

与之前类似，我们计算经验贝尔曼误差 $\mathcal{E}_i(g_i, \pi; \mathcal{D}) \coloneqq \mathcal{L}_i(g_i, g_i, \pi; \mathcal{D}) - \min_{g_i' \in \mathcal{G}_i} \mathcal{L}_i(g_i', g_i, \pi; \mathcal{D})$，并构造版本空间 $\mathcal{G}^{\pi, \varepsilon}_i = \{g_i \in \mathcal{G}_i : \mathcal{E}_i(g_i, \pi; \mathcal{D}) \le \beta_g\}$。略有不同的是，我们将参数 β_g 设为

$$\beta_g \coloneqq \frac{80 C_A(\pi) V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} + 30\varepsilon_{\mathcal{F}},$$

其中 $C_A(\pi) \coloneqq \max_{s, \mathbf{a}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}$。与算法 β_f 中的算法 1 相比，额外的 $C_A(\pi)$ 项来自重要性加权。利用已知的 $\mathcal{G}^{\pi, \beta_g}_i$，我们定义

$$g_i^{\pi,\max} \coloneqq \operatorname*{argmax}_{g_i \in \mathcal{G}_i^{\pi,\beta_g}} g_i(s_0), \quad g_i^{\pi,\min} \coloneqq \operatorname*{argmin}_{g_i \in \mathcal{G}_i^{\pi,\beta_g}} g_i(s_0).$$

We 然后计算 $\text{Gap}^{\text{EQ}}(\pi)$，它是任意 $\pi \in \Pi$的均衡差距上界：

$$\widehat{\text{Gap}}_{\text{EQ}}(\pi) \coloneqq \max_{i \in [m]} \max_{\pi^\dagger \in \Pi_i^{\dagger,\text{EQ}}(\pi)} g_i^{\pi^\dagger,\max}(s_0) - g_i^{\pi,\min}(s_0).$$

我们通过最小化估计的均衡差距来选择策略：

$$\widehat{\pi} = \operatorname*{argmin}_{\pi \in \Pi} \widehat{\text{Gap}}_{\text{EQ}}(\pi), \tag{9}$$

其性能保证如下所示。

定理 5 (V 型保证)。 以至少 $1 - \delta$ 的概率，对于任意 $\pi \in \Pi$ 以及均衡 $\in$ {纳什均衡、相关均衡、粗相关均衡}，由式 *(9)* 得到的输出策略 $\widehat{\pi}$ 满足*(9)* 如下性质

$$\text{Gap}^{\Pi^{\dagger,\text{EQ}}}(\widehat{\pi}) \le \text{Gap}^{\Pi^{\dagger,\text{EQ}}}(\pi) + \frac{4\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} + \max_{i \in [m]} \min_{\widetilde{\pi}_i \in \Pi_i^{\dagger,\text{EQ}}(\pi)} \left(\Delta_i^{\widetilde{\pi}_i} + \Delta_i^\pi + \text{subopt}_i^\pi(\widetilde{\pi}_i)\right),$$

其中 $\Delta_i^\pi = g^{\pi, \max}_i(s_0) - g^{\pi, \min}_i(s_0)$ 以及次优$\widetilde{\pi}_i$ = 最大值$_{\pi^\dagger \in \Pi^{\dagger, \text{EQ}}_i(\pi)} g^{\pi^\dagger, \text{最大值} i}(s_0) - g^{\widetilde{\pi}_i, \text{最大值} i}(s_0)$。此外，以至少 $1 - \delta$ 的概率，对于任意玩家 $i \in [m]$ 和任意 $\pi \in \Pi^{\text{ext}}_i$，我们有

$$\Delta_i^\pi \le \min_{d \in \Delta(S)} \frac{1}{1-\gamma} \sqrt{\mathscr{C}(d; d_S, \mathcal{G}_i, \pi)\varepsilon_{\text{apx}}} + \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} (d^\pi \setminus d)(s) \left[\Delta g_i^\pi(s) - \gamma (P_i^\pi \Delta g_i^\pi)(s)\right],$$

[6]该假设不失一般性，且仅为技术上的便利，因此动作重要性权重始终有定义。否则，我们可以直接忽略任何满足 π where $\pi(\mathbf{a}|s)/d_A(\mathbf{a}|s)$ 无界的策略，并假设此类 π的最大值为 Δ_i^π。

where $\varepsilon_{\text{apx}} = \mathcal{O}\left(V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}}+\varepsilon_{\mathcal{F},\mathcal{F}}}\right)$, $(d^{\pi}\backslash d)(s) \coloneqq \max(d^{\pi}(s)-d(s),0)$, $\Delta g_i^{\pi}(s) \coloneqq g_i^{\pi,\max}(s) - g_i^{\pi,\min}(s)$, *and* $(P_i^{\pi}g)(s) = \mathbb{E}_{\mathbf{a}\sim\pi(\cdot|s),s'\sim P(\cdot|s,\mathbf{a})}[r_i(s,\mathbf{a})+g(s')]$.

与第 4 节中的结果类似，我们的界具有自适应性，并会自动选择最佳策略 $\widetilde{\pi}_i$，从而在次优误差 次优$_i^{\pi}$（$\widetilde{\pi}_i$）与数据集 $\widetilde{}$覆盖误差 $\Delta_i^{\pi_i}$之间实现权衡。此外，当数据集 $\mathcal{D}$ 满足单边覆盖假设时，我们有如下推论。

推论6。 对于纳什均衡策略 $\pi^\star \in \Pi$，若存在单边系数 $C_S(\pi^\star)$ 使得下列不等式成立

$$\max_{i\in[m]}\max_{\pi^\dagger\in\Pi_i^{\dagger,\text{NE}}(\pi^\star)} \mathscr{C}(d^{\pi^\dagger};d_S,\mathcal{G}_i,\pi^\star) \le C_S(\pi^\star). \tag{10}$$

以至少 $1-\delta$ 的概率，有

$$\text{Gap}^{\Pi^{\dagger,\text{NE}}}(\widehat{\pi}) \le \mathcal{O}\left(\frac{V_{\max}\sqrt{\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}}+\varepsilon_{\mathcal{F},\mathcal{F}}}}{1-\gamma}\sqrt{C_A(\pi^\star)C_S(\pi^\star)}\right).$$

与推论 4 相比，这里的界对 V函数类 $\mathcal{G}$ 呈对数依赖，而非对 Q 函数类 $\mathcal{F}$ 呈对数依赖。在表格型设定中，当我们使用完全表达（且平稳）的函数类时，$\log|\mathcal{G}| \approx \mathcal{O}(S)$（通过简单的覆盖论证）成立，因此我们的界避免了对 m 的指数依赖（即对 $\prod_{i=1}^m A_i$ 的依赖）。相比之下，Cui 和 Du [2022b] ~~为有限时域表格化马~~尔可夫博弈建立了 $\widetilde{\mathcal{O}}\left(\sqrt{H^4S^2\log(\mathcal{N}(\Pi))C(\pi^\star)/n}\right)$ 误差界，其中 H 是时域长度，而 $\mathcal{N}(\Pi)$ 大致对应于我们的 $|\Pi^{\text{ext}}|$。虽然有限时域结果与折扣结果通常不可比较，但在标准转换下，[7] 我们的界具有相同的速率 $\widetilde{\mathcal{O}}(n^{-1/2})$;$\sqrt{\log|\mathcal{G}|} \approx \sqrt{SH}$ [8]，这使得它对 S 的依赖更优（节省了一个 $\sqrt{S}$ 因子），但整体上对 H 的依赖更差（我们有 $\sqrt{H^5}$）。一个小缺点是，推论6分别衡量动作和状态上的分布失配（而不是像推论 4 中那样联合衡量 $C(\pi^*)$），因此界更松。

6 讨论与结论

双人零和博弈的算法 本文大部分内容考虑的是多人一般和马尔可夫博弈的一般情形。我们发现，当将我们的算法专门应用于双人零和（2p0s）这一特殊情形时，它似乎不同于另一种专门为 2p0s 设计、受 Jin 等人 [2022],Cui 和 Du [2022b]启发的样本高效算法。在附录B中，我们说明这种差异只是表面上的，这些专门化算法都可以作为我们算法的小型变体而被涵盖。

结论与开放问题 在本文中，我们研究了马尔可夫博弈中的离线学习。我们设计了一个框架，在一般函数近似下以统一的方式学习三种流行的均衡概念。我们框架的自适应性使我们能够在某些情况下放宽“单边集中性”条件，并在此基础上实现显著改进。

一个尚未解决的问题是，能否设计出一种计算高效的算法，在离线马尔可夫博弈中学习相关均衡/粗相关均衡，即使是在表格型设定下也是如此。一个潜在方向是改进计算效率，

[7] Cui 和 Du [2022b] 假设奖励取值于 $[0,1]$,，因此我们将 $V_{\max} = 1/(1-\gamma) = H$。当使用完全表达的表格类时，$\varepsilon_{\mathcal{F}} = \varepsilon_{\mathcal{F},\ \mathcal{F}} = 0$。[8]在有限时域问题中，我们需要使用非平稳函数类，因此会多出 H 因子。

V-学习算法 [Song 等人，2021，Jin 等人，2021b]——该算法在每个状态下运行无悔学习动态——并将其应用于离线设定，但这可能需要新的思路。

参考文献

Andr´as Antos、Csaba Szepesv´ari 和 R´emi Munos。基于贝尔曼残差最小化的拟合策略迭代与单条样本路径下的近最优策略学习。机器学习，71(1):89–129，2008。Robert J Aumann。随机化策略中的主观性与相关性。《数学经济学杂志》，1(1):67–96，1974年。Yu Bai 和 Chi Jin。用于竞争性强化学习的可证明自博弈算法。载于国际机器学习会议，第 551–560 页。PMLR，2020。Yu Bai、Chi Jin 和 Tiancheng Yu。通过自博弈实现近最优强化学习。神经信息处理系统进展，33:2159–2170，2020。Jinglin Chen 和 Nan Jiang。批量强化学习中的信息论考量。载于国际机器学习会议，第 1042–1051 页。PMLR，2019。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。载于国际算法学习理论会议，第 227–261 页。PMLR，2022。Qiwen Cui 和 Simon S Du。离线双人零和马尔可夫博弈何时可解？*arXiv* 预印本 *arXiv:2201.03522*，2022a。Qiwen Cui 和 Simon S Du。通过逐策略奖励实现可证明高效的离线多智能体强化学习。*arXiv* 预印本 *arXiv:2206.00159*，2022b。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。Zehao Dou、Zhuoran Yang、Zhaoran Wang 和 Simon Du。双人马尔可夫博弈中依赖间隙的界。载于国际人工智能与统计会议，第 432–455 页。PMLR，2022。Baihe Huang、Kaixuan Huang、Sham Kakade、Jason D Lee、Qi Lei、Runzhe Wang 和 Jiaqi Yang。超越线性强化学习：样本高效的神经函数逼近。神经信息处理系统进展，34:8968–8983，2021a。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*，2021b。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。载于国际机器学习会议，第 1704–1713 页。PMLR，2017。Chi Jin、Zhuoran Yang、Zhaoran Wang 和 Michael I Jordan。通过线性函数逼近实现可证明高效的强化学习。载于学习理论会议，第 2137–2143 页。PMLR，2020。

Chi Jin、Qinghua Liu 和 Sobhan Miryoosefi。贝尔曼 eluder 维度：强化学习问题的新丰富类别及样本高效算法。神经信息处理系统进展，34:13406–13418，2021a。Chi Jin、Qinghua Liu、Yuanhao Wang 和 Tiancheng Yu。V 学习——一种简单、高效的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：在大状态空间中实现可证明的多智能体强化学习。载于国际机器学习会议，第 10251–10279 页。PMLR，2022。Ying Jin、Zhuoran Yang 和 Zhaoran Wang。悲观主义是否能为离线强化学习带来可证明的效率？载于国际机器学习会议，第 5084–5096 页。PMLR，2021c。Sergey Levine、Aviral Kumar、George Tucker 和 Justin Fu。离线强化学习：教程、综述及开放问题展望。*arXiv* 预印本 *arXiv:2005.01643*，2020。Gen Li、Laixi Shi、Yuxin Chen、Yuejie Chi 和 Yuting Wei。确定基于模型的离线强化学习的样本复杂度。*arXiv* 预印本 *arXiv:2204.05275*，2022。Qinghua Liu、Tiancheng Yu、Yu Bai 和 Chi Jin。对通过自博弈进行的基于模型强化学习的精确分析。载于国际机器学习会议，第 7001–7010 页。PMLR，2021。Weichao Mao 和 Tamer Bas.ar。在去中心化一般和马尔可夫博弈中实现可证明高效的强化学习。动态博弈与应用，第 1–22 页，2022。John Nash Jr.。非合作博弈。载于博弈论论文集，第 22–33 页。爱德华·埃尔加出版社，1996。Andrzej S Nowak 和 TES Raghavan。折扣随机博弈中具有对称信息的平稳相关均衡的存在性。运筹学数学，17(3):519–526，1992。Doina Precup。用于离策略策略评估的资格迹。计算机科学系教员出版物系列，第 80 页，2000。Paria Rashidinejad、Banghua Zhu、Cong Ma、Jiantao Jiao 和 Stuart Russell。连接离线强化学习与模仿学习：一个关于悲观主义的故事。神经信息处理系统进展，34:11702–11716，2021。Lloyd S Shapley。随机博弈。美国国家科学院院刊，39(10):1095–1100，1953。Laixi Shi、Gen Li、Yuting Wei、Yuxin Chen 和 Yuejie Chi。面向离线强化学习的悲观 Q 学习：迈向最优样本复杂度。*arXiv* 预印本 *arXiv:2202.13890*，2022。Ziang Song、Song Mei 和 Yu Bai。何时能够以样本高效的方式学习拥有大量玩家的一般和马尔可夫博弈？*arXiv* 预印本 *arXiv:2110.04184*，2021。Masatoshi Uehara 和 Wen Sun。部分覆盖下的悲观基于模型离线强化学习。*arXiv* 预印本 *arXiv:2107.06226*，2021。Ruosong Wang、Ruslan Salakhutdinov 和 Lin F Yang。通过一般价值函数近似实现可证明高效的强化学习。*arXiv* 预印本 *arXiv:2005.10804*，2020。

Chen-Yu Wei、Yi-Te Hong 和 Chi-Jen Lu。随机博弈中的在线强化学习。神经信息处理系统进展，30，2017。Qiaomin Xie、Yudong Chen、Zhaoran Wang 和 Zhuoran Yang。利用函数逼近和相关均衡学习零和同时行动马尔可夫博弈。载于学习理论会议，第 3674–3682 页。PMLR，2020。Tengyang Xie 和 Nan Jiang。批量强化学习的 Q* 逼近方案：理论比较。载于人工智能不确定性会议，第 550–559 页。PMLR，2020。Tengyang Xie 和 Nan Jiang。仅凭可实现性的批量价值函数近似。载于国际机器学习会议，第 11404–11413 页。PMLR，2021。Tengyang Xie、Ching-An Cheng、Nan Jiang、Paul Mineiro 和 Alekh Agarwal。离线强化学习中的贝尔曼一致悲观性。神经信息处理系统进展，34：6683–6694，2021a。Tengyang Xie、Nan Jiang、Huan Wang、Caiming Xiong 和 Yu Bai。策略微调：连接样本高效的离线与在线强化学习。神经信息处理系统进展，34：27395–27407，2021b。Wei Xiong、Han Zhong、Chengshuai Shi、Cong Shen、Liwei Wang 和 Tong Zhang。具有线性函数逼近的近乎极小极大最优离线强化学习：单智能体马尔可夫决策过程与马尔可夫博弈。*arXiv* 预印本 *arXiv:2205.15512*，2022。Yuling Yan、Gen Li、Yuxin Chen 和 Jianqing Fan。基于模型的强化学习对于离线零和马尔可夫博弈是极小极大最优的。*arXiv* 预印本 *arXiv:2206.04044*，2022。Ming Yin 和 Yu-Xiang Wang。迈向具有悲观主义的实例最优离线强化学习。神经信息处理系统进展，34：4065–4078，2021。Ming Yin、Yu Bai 和 Yu-Xiang Wang。通过双重方差缩减实现近乎最优的离线强化学习。神经信息处理系统进展，34：7677–7688，2021a。Ming Yin、Yu Bai 和 Yu-Xiang Wang。强化学习离线策略评估中近乎最优且可证明的一致收敛。载于国际人工智能与统计会议，第 1567–1575 页。PMLR，2021b。Ming Yin、Yaqi Duan、Mengdi Wang 和 Yu-Xiang Wang。具有线性表示的近乎最优离线强化学习：利用方差信息并采用悲观主义。*arXiv* 预印本 *arXiv:2203.05804*，2022。Andrea Zanette、Martin J Wainwright 和 Emma Brunskill。演员-评论家方法在离线强化学习中的可证明优势。神经信息处理系统进展，34：13626–13640，2021。Wenhao Zhan、Baihe Huang、Audrey Huang、Nan Jiang 和 Jason Lee。具有可实现性和单策略集中性的离线强化学习。载于学习理论会议，第 2730–2775 页。PMLR，2022。Han Zhong、Wei Xiong、Jiyuan Tan、Liwei Wang、Tong Zhang、Zhaoran Wang 和 Zhuoran Yang。悲观极小极大值迭代：从离线数据集中高效学习均衡的可证明方法。*arXiv* 预印本 *arXiv:2202.07511*，2022。

类内差距与实际差距之间的联系

本文考虑“类内”均衡差距，其定义是相对于某些偏离策略类而言的（第 3 节）。人们也常考虑更强的均衡差距概念，我们将其简记为间隙^{EQ}，此时偏离策略不受限制，例如，对于纳什均衡和粗相关均衡，偏离策略可以是任意策略 [Nash Jr，1996，Aumann，1974]。

为了建立类内差距与更强的差距概念之间的联系，我们作出如下策略完备性假设：

假设C （策略完备性）。对于任意玩家 $i \in [m]$ 以及任意均衡 $\in$ {纳什均衡，粗相关均衡}，我们有 e

$$\max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} V_i^{\pi^\dagger}(s_0) \geq \max_{\pi' : \pi'_{-i} = \pi_{-i}} V_i^{\pi'}(s_0) - \varepsilon_\Pi.$$

对于相关均衡，我们有

$$\max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{CE}}(\pi)} V_i^{\pi^\dagger}(s_0) \geq \max_{\phi_i} V_i^{(\phi_i \diamond \pi_i) \odot \pi_{-i}}(s_0) - \varepsilon_\Pi.$$

假设 C 要求（不受限的）最佳响应策略包含于 Π （相关均衡对应的最佳响应策略则包含于 Φ）。在假设 C下，显然，对于任意均衡 $\in$ {纳什均衡、相关均衡、粗相关均衡}以及 π，间隙$^{\mathrm{EQ}}(\pi) \leq$ 间隙$^{\Pi^\dagger,\mathrm{EQ}}(\pi) + \varepsilon_\Pi$。

B A 双人零和博弈中的联系

在本文的大部分内容中，我们考虑的是多人一般和马尔可夫博弈的一般情形。当处于双人零和马尔可夫博弈（2p0s）等特殊设定时，我们通常可以利用其特殊结构，提出替代算法 [Yan 等人，2022，Jin 等人，2022]。

具体而言，Cui 和 Du [2022b，第 3] 节为表格型设定设计了一种离线双人零和算法，而将其算法扩展到函数逼近设定（使用本文中的不确定性量化技术）后，得到的算法表面上看起来与我们的算法 1非常不同。然而，下面我们将展示，尽管表面上存在差异，这两种算法实际上非常相似，并且可以像我们的算法 1一样，以相同的方式通过乐观/悲观原则推导出来，唯一的细微差别在于最小化对偶间隙，而不是我们的 间隙^{NE}。因此，对于他们的算法，我们可以通过对我们的算法和分析进行略微调整，给出类似于定理3的保证。

双人零和设定 我们现在介绍一些专用于双人零和博弈的记号。我们考虑两个玩家，其中 x- 玩家旨在最大化总奖励，而 y-玩家旨在将其最小化。 x-玩家和y-玩家的策略集分别记为 $\Pi^{\max}$ 和 $\Pi^{\min}$ 。我们考虑策略收益 $V \in [-1,1]^{|\Pi^{\max}| \times |\Pi^{\min}|}$，其中 $V^{\mu,\ \nu}$ 分别表示当 x-玩家和y-玩家分别遵循策略 μ 和策略 ν 时的效用/损失。我们用 $\overline{V}$和 $\underline{V}$分别表示 V的置信上界和置信下界估计。为将这些符号与正文中的符号对应起来，假设玩家 1 是最大化玩家 x ，玩家 2 是最小化玩家 y，则 $V^{\mu,\ \nu}$ 本质上就是 $V_1^\pi(s_0)(= -V_2^\pi(s_0))$（对于 $\pi = \mu \times \nu$）。此外，由于博弈具有零和性质，我们有 $V^{\mu,\ \overline{\nu}} = V_1^\pi(\overline{s_0}) = -V_2^\pi(\underline{s_0})$，以及 $\underline{V}^{\mu,\ \nu} = \underline{V}_1^\pi(s_0) = -\overline{V}_2^\pi(s_0)$。

对偶间隙 对于双人零和博弈，一个常见的学习目标是对偶间隙，其定义如下：

$$\text{Dual-Gap}(\mu,\nu) = \max_{\mu^\dagger \in \Pi^{\max}} V^{\mu^\dagger,\nu} - \min_{\nu^\dagger \in \Pi^{\min}} V^{\mu,\nu^\dagger}.$$

由于 $\mu^\dagger$ 和 $\nu^\dagger$ 可以分别取为 μ 和 ν，因此对偶间隙始终是非负的。它衡量策略与纳什均衡策略的接近程度，而纳什均衡策略的对偶间隙始终为零。受 Cui 和 Du [2022b], 提出的表格型双人零和算法启发，可以在悲观估计下设计一种离线算法，在存在对抗性对手的情况下 独立地选择两个策略：

$$\mu = \operatorname*{argmax}_{\mu} \min_{\nu^\dagger \in \Pi^{\min}} \underline{V}^{\mu,\nu^\dagger} \text{ and } \nu = \operatorname*{argmin}_{\nu} \max_{\mu^\dagger \in \Pi^{\max}} \overline{V}^{\mu^\dagger,\nu}. \tag{11}$$

类似的想法也可以在 Jin 等人 [2022], 为双人零和博弈设计的在线算法中找到。将他们的乐观性（针对在线场景）转变为悲观主义（针对离线场景），我们同样可以得到 (11)。

在我们的算法框架中恢复 (11) 看起来与我们的算法 (11) 乍看之下大不相同，因为 1 是独立选择各玩家的策略，而我们的算法 1则需要进行联合优化。不过，我们现在将证明，它只是我们算法的一个小变体，因此我们的分析和保证可以直接扩展到该变体。

首先请注意，在专用于双人零和博弈时，对偶间隙与我们的目标 $\mathrm{Gap}^{\Pi^\dagger,\mathrm{EQ}}(\pi)$ 并不相同。回想一下，

$$\mathrm{Gap}^{\Pi^\dagger,\mathrm{EQ}}(\pi) := \max_{i\in[m]} \max_{\pi^\dagger \in \Pi^{\dagger,\mathrm{EQ}}(\pi)} V_i^{\pi^\dagger}(s_0) - V_i^{\pi}(s_0).$$

要恢复对偶间隙，只需将上述目标中的 $\max_i$ 替换为 $\sum_i$，即可在双人零和情形下得到：

$$\begin{aligned}\sum_{i\in[m]} \max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} V_i^{\pi^\dagger}(s_0) - V_i^{\pi}(s_0) &= (\max_{\mu^\dagger \in \Pi^{\max}} \overline{V}^{\mu^\dagger,\nu} - V^{\mu,\nu}) + (V^{\mu,\nu} - \min_{\nu^\dagger \in \Pi^{\min}} \underline{V}^{\mu,\nu^\dagger}) \\ &= \max_{\mu^\dagger \in \Pi^{\max}} V^{\mu^\dagger,\nu} - \min_{\nu^\dagger \in \Pi^{\min}} V^{\mu,\nu^\dagger} = \text{Dual-Gap}(\mu,\nu).\end{aligned}$$

从上式可以看出，我们算法 1 中的目标与对偶间隙几乎相同，相差一个乘法因子 2；对于非负的 a, b 我们有 $\max(a,\ b) \le a+b \le 2\max(a,\ b)$。因此，我们的算法1 直接享有对偶间隙保证。

然而，请记住，我们的目标是恢复(11)，因此我们选择直接处理对偶间隙，并按照算法 1 中的相同思路对其进行松弛：由于 $V \le V$ 且 $-V \le -V$，我们有

$$\text{Dual-Gap}(\mu,\nu) = \max_{\mu^\dagger \in \Pi^{\max}} V^{\mu^\dagger,\nu} - \min_{\nu^\dagger \in \Pi^{\min}} V^{\mu,\nu^\dagger} \le \max_{\mu^\dagger \in \Pi^{\max}} \overline{V}^{\mu^\dagger,\nu} - \min_{\nu^\dagger \in \Pi^{\min}} \underline{V}^{\mu,\nu^\dagger}. \tag{12}$$

现在，只需注意到 (11) 即可恢复： μ 和 ν 可以在 (12) 右侧独立优化，且其最优解恰好是(11)。

我们还为上述算法提供了如下保证：

命题 7。 考虑一个策略收益为 $V \in [-1,1]^{|\Pi^{\max}|\times|\Pi^{\min}|}$ 的双人零和马尔可夫博弈，令

$$J(\mu,\nu) = \max_{\mu^\dagger \in \Pi^{\max}} \overline{V}^{\mu^\dagger,\nu} - \min_{\nu^\dagger \in \Pi^{\min}} \underline{V}^{\mu,\nu^\dagger}.$$

令 $\widehat{\mu},\widehat{\nu} = \operatorname{argmin} J(\mu,\nu)$，以高概率有

$$\text{Dual-Gap}(\widehat{\mu},\widehat{\nu}) \le \min_{\widetilde{\mu},\widetilde{\nu}\in\Pi^{\max}\times\Pi^{\min}} \Delta^{\widetilde{\mu},\nu^*} + \Delta^{\mu^*,\widetilde{\nu}} + \mathrm{subopt}^{\pi^\star}(\widetilde{\mu}) + \mathrm{subopt}^{\pi^\star}(\widetilde{\nu}),$$

其中 $\Delta^{\mu,\ \nu} := \overline{V}^{\mu,\ \nu} - \underline{V}^{\mu,\ \nu}$，次优$^{\pi^\star}(\widetilde{\mu}) := $ 最大值$_{\mu^\dagger\in\Pi^{\max}} V^{\mu^\dagger,\ \nu^*} - V^{\widetilde{\mu},\ \nu^*}$ 以及次优$^{\pi^\star}(\widetilde{\nu}) := V^{\mu^*,\ \widetilde{\nu}} -$ 最小值$_{\nu^\dagger\in\Pi^{\min}} V^{\mu^*,\ \nu^\dagger}$。

证明。 通过标准的集中度分析，我们可以保证以高概率， $V^{\mu,\,\nu} \geq V^{\mu,\,\nu}$ 以及 $V^{\mu,\,\nu} \leq V^{\mu,\,\nu}$ 对任意 $\mu,\ \nu \in \Pi^{\max} \times \Pi^{\min}$ 均成立。这意味着，对于任意 $\mu,\ \nu \in \Pi^{\max} \times \Pi^{\min}$，$\text{Dual-Gap}(\mu,\ \nu) \leq J(\mu,\ \overline{\nu})$。对于纳什策略 $\pi^\star = (\mu^*,\ \nu^*)$，令 $\mu^\dagger = \operatorname{argmax}_{\mu^\dagger \in \Pi^{\max}} V^{\mu^\dagger,\,\nu^*}$ 以及 $\nu^\dagger = \operatorname{argmin}_{\nu^\dagger \in \Pi^{\min}} V^{\mu^*,\,\nu^\dagger}$。我们有

$$\begin{aligned} J(\mu^*,\nu^*) - \text{Dual-Gap}(\mu^*,\nu^*) &= \overline{V}^{\mu^\dagger,\nu^*} - \max_{\mu\in\Pi^{\max}} V^{\mu,\nu^*} + \min_{\nu\in\Pi^{\min}} V^{\mu^*,\nu} - \underline{V}^{\mu^*,\nu^\dagger} \\ &\leq (\overline{V}^{\widetilde{\mu},\nu^*} - V^{\widetilde{\mu},\nu^*}) + (V^{\mu^*,\widetilde{\nu}} - \underline{V}^{\mu^*,\widetilde{\nu}}) + \text{subopt}^{\pi^\star}(\widetilde{\mu}) + \text{subopt}^{\pi^\star}(\widetilde{\nu}) \\ &\leq \Delta^{\widetilde{\mu},\nu^*} + \Delta^{\mu^*,\widetilde{\nu}} + \text{subopt}^{\pi^\star}(\widetilde{\mu}) + \text{subopt}^{\pi^\star}(\widetilde{\nu}), \end{aligned} \tag{13}$$

其中 $\widetilde{\mu}$ 和 $\widetilde{\nu}$ 分别是 $\Pi^{\max}$ 和 $\Pi^{\min}$ 中的任意策略。根据 $(\widehat{\mu},\widehat{\nu})$ 的最优性以及 (13)，我们得到

$$\text{Dual-Gap}(\widehat{\mu},\widehat{\nu}) \leq J(\widehat{\mu},\widehat{\nu}) \leq J(\mu^*,\nu^*) \leq \min_{\widetilde{\mu},\widetilde{\nu}\in\Pi^{\max}\times\Pi^{\min}} \Delta^{\widetilde{\mu},\nu^*} + \Delta^{\mu^*,\widetilde{\nu}} + \text{subopt}^{\pi^\star}(\widetilde{\mu}) + \text{subopt}^{\pi^\star}(\widetilde{\nu}).$$

证明完成。 □

C 第4节的证明

本节证明定理 3。我们首先给出一些集中度结果。

引理8. 以至少 $1-\delta$ 的概率，对于任意玩家 $i \in [m]$、任意 f_i、g_1、$g_2 \in \mathcal{F}_i$ 以及任意 $\pi \in \Pi_i^{\text{ext}}$，均有

$$\begin{aligned} &\Big| \|g_1 - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 - \|g_2 - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 \\ &\quad - \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \big(g_1(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\big)^2 + \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \big(g_2(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\big)^2 \Big| \\ &\leq 2V_{\max}\|g_1 - g_2\|_{2,d_D}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2 \log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}. \end{aligned}$$

证明。 对于玩家 i，我们注意到

$$\begin{aligned} &\|g_1 - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 - \|g_2 - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 \\ &= \mathbb{E}_{d_D}\left[(g_1(s,\mathbf{a}) - (\mathcal{T}_i^\pi f_i)(s,\mathbf{a}))^2\right] - \mathbb{E}_{d_D}\left[(g_2(s,\mathbf{a}) - (\mathcal{T}_i^\pi f_i)(s,\mathbf{a}))^2\right] \\ &= \mathbb{E}_{d_D}\left[(g_1(s,\mathbf{a}) - g_2(s,\mathbf{a}))(g_1(s,\mathbf{a}) + g_2(s,\mathbf{a}) - 2(\mathcal{T}_i^\pi f)(s,\mathbf{a}))\right] \\ &= \mathbb{E}_{d_D}\left[(g_1(s,\mathbf{a}) - g_2(s,\mathbf{a}))\mathbb{E}_{s'\sim P(\cdot|s,\mathbf{a})}\left[g_1(s,\mathbf{a}) + g_2(s,\mathbf{a}) - 2\mathbf{r}_i - 2\gamma f(s',\pi)|s,\mathbf{a}\right]\right] \\ &= \mathbb{E}_{d_D\times P}\left[(g_1(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi))^2\right] - \mathbb{E}_{d_D\times P}\left[(g_2(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi))^2\right]. \end{aligned} \tag{14}$$

设随机变量 $X = (g_1(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi))^2 - (g_2(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi))^2$，$X$ 服从 $d_D \times P$。我们知道 $\mathbb{E}_{d_D\times P}[X] = \mathbb{E}_{d_D\times P}\left[(g_1(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi))^2\right] - \mathbb{E}_{d_D\times P}\left[(g_2(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi))^2\right]$。对于方差，我们有

$$\begin{aligned} \mathbb{V}_{d_D\times P}[X] &\leq \mathbb{E}_{d_D\times P}[X^2] \\ &\leq \mathbb{E}_{d_D\times P}[(g_1(s,\mathbf{a}) - g_2(s,\mathbf{a})^2(g_1(s,\mathbf{a}) + g_2(s,\mathbf{a}) - 2\mathbf{r}_i - 2\gamma f_i(s',\pi))^2] \\ &\leq 4V_{\max}^2\mathbb{E}_{d_D}\left[(g_1(s,\mathbf{a}) - g_2(s,\mathbf{a}))^2\right]. \end{aligned}$$

我们按如下方式进行

$$
\begin{aligned}
&\Big|\|g_1-\mathcal{T}_i^\pi f_i\|_{2,d_D}^2-\|g_2-\mathcal{T}_i^\pi f_i\|_{2,d_D}^2 \\
&-\frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}}\big(g_1(s,\mathbf{a})-\mathbf{r}_i-\gamma f_i(s',\pi)\big)^2+\frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}}\big(g_2(s,\mathbf{a})-\mathbf{r}_i-\gamma f_i(s',\pi)\big)^2\Big| \\
&=\left|\mathbb{E}_{d_D\times P}\left[(g_1(s,\mathbf{a})-\mathbf{r}_i-\gamma f_i(s',\pi))^2\right]-\mathbb{E}_{d_D\times P}\left[(g_2(s,\mathbf{a})-\mathbf{r}_i-\gamma f_i(s',\pi))^2\right]-\frac{1}{n}\sum_{j=1}^n X_j\right| \\
&\qquad\qquad\text{(By (14) and definition of } X\text{)} \\
&\le\sqrt{\frac{4V_{\max}^2\|g_1-g_2\|_{d_D}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}}+\frac{V_{\max}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}. \qquad\text{(By Freedman's inequality)}
\end{aligned}
$$

对 $i\in[m]$ 使用并集界即可完成证明。 □

对于任意玩家 $i\in[m]$ 和 $\pi\in\Pi_i^{\text{ext}}$，定义

$$f_i^\pi\coloneqq\operatorname*{argmin}_{f\in\mathcal{F}_i}\sup_{\text{admissible } d}\|f-\mathcal{T}_i^\pi f\|_{2,d}^2 \tag{15}$$

$$g_i^\pi\coloneqq\operatorname*{argmin}_{g\in\mathcal{F}_i}\frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}}\big(g(s,\mathbf{a})-\mathbf{r}_i-\gamma f_i^\pi(s',\pi)\big)^2. \tag{16}$$

我们按如下方式界定 $\|f_i^\pi-g_i^\pi\|_{2,\ d_D}$。

引理9。 设 f_i^π和 g_i^π如公式 (15) 和 (16)中所定义。在引理 *8*的成功事件下，对于任意玩家 $i\in[m]$ 和 $\pi\in\Pi_i^{\text{ext}}$，有

$$\|f_i^\pi-g_i^\pi\|_{2,d_D}\le 6V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}}+2\sqrt{\varepsilon_{\mathcal{F}}}.$$

证明。 我们知道

$$
\begin{aligned}
&\|f_i^\pi-g_i^\pi\|_{2,d_D}^2 \\
&\le 2\|f_i^\pi-\mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2+2\|g_i^\pi-\mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 \\
&=2\|g_i^\pi-\mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2-2\|f_i^\pi-\mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2+4\|f_i^\pi-\mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 \\
&\le 2\|g_i^\pi-\mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2-2\|f_i^\pi-\mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2+4\varepsilon_{\mathcal{F}} \qquad\text{(根据假设A)} \\
&\overset{\text{(a)}}{\le} 4V_{\max}\sqrt{\frac{\|g_i^\pi-f_i^\pi\|_{2,d_D}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}}+\frac{2V_{\max}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}+4\varepsilon_{\mathcal{F}},
\end{aligned}
\tag{17}
$$

其中（a）来源于

$$
\begin{aligned}
&\|g_i^\pi-\mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2-\|f_i^\pi-\mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 \\
&\le\frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}}\big(g_i^\pi(s,\mathbf{a})-\mathbf{r}_i-\gamma f_i^\pi(s',\pi)\big)^2-\frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}}\big(f_i^\pi(s,\mathbf{a})-\mathbf{r}_i-\gamma f_i^\pi(s',\pi)\big)^2 \\
&\quad+2V_{\max}\sqrt{\frac{\|g_i^\pi-f_i^\pi\|_{2,d_D}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}}+\frac{V_{\max}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \qquad\text{(by Lem}^{\text{ma 8}}\text{)}
\end{aligned}
$$

$$\leq 2V_{\max}\sqrt{\frac{\|g_i^\pi - f_i^\pi\|_{2,d_D}^2 \log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2 \log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \qquad \text{（根据 } g\text{的最优性）}$$

求解（17）即可完成证明。 □

在 下一个引理中，我们证明 Q_i^π的最佳近似包含于 $\mathcal{F}^{\pi,\ \beta_f}_{\ i}$。

引理10。 在引理8的成功事件下，对于任意玩家 $i \in [m]$ 和 $\pi \in \Pi_i^{\text{ext}}$，以下关于 $\mathcal{E}_i(f_i^\pi,\ \pi;\ \mathcal{D})$ 成立

$$\mathcal{E}_i(f_i^\pi, \pi; \mathcal{D}) \leq \frac{80V_{\max}^2 \log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} + 30\varepsilon_{\mathcal{F}} =: \beta_f.$$

证明。 应用引理8 和引理9，我们得到

$$\begin{aligned}
&\Big| \|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 - \|g_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 - \\
&\frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \left(f_i^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i^\pi(s',\pi)\right)^2 + \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \left(g_i^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i^\pi(s',\pi)\right)^2 \Big| \\
&\leq 2V_{\max}\|f_i^\pi - g_i^\pi\|_{2,d_D}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2 \log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \\
&\leq 4V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F}}} + \frac{13V_{\max}^2 \log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}. \qquad (18)
\end{aligned}$$

接下来，我们对$\|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,\ d_D}^2 - \|g_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,\ d_D}^2$作如下界定，

$$\begin{aligned}
&\|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 - \|g_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 \\
&\leq \left(\|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D} + \|g_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}\right)\left|\|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D} - \|g_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}\right| \\
&\leq \left(2\|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D} + \|f_i^\pi - g_i^\pi\|_{2,d_D}\right)\|f_i^\pi - g_i^\pi\|_{2,d_D} \qquad \text{(By triangle inequality)} \\
&\leq 36V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F}}} + 36V_{\max}^2\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} + 8\varepsilon_{\mathcal{F}}. \qquad \text{(By Assumption A and Lemma 9)}
\end{aligned}$$

将其与(18)结合，得到

$$\begin{aligned}
&\frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \left(f_i^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i^\pi(s',\pi)\right)^2 - \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \left(g_i^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i^\pi(s',\pi)\right)^2 \\
&\leq \|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 - \|g_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 + 4V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F}}} + \frac{13V_{\max}^2 \log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \\
&\leq 40V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F}}} + \frac{59V_{\max}^2 \log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} + 8\varepsilon_{\mathcal{F}} \\
&\leq \frac{80V_{\max}^2 \log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} + 30\varepsilon_{\mathcal{F}}. \qquad (\sqrt{ab} \leq \tfrac{a+b}{2})
\end{aligned}$$

□

接下来，我们证明$|f_i^\pi(s_0,\pi)-V_i^\pi(s_0)|$具有如下上界

引理 11. 对于任意玩家 $i\in[m]$ 和$\pi\in\Pi_i^{\text{ext}}$，令f_i^π按*(15)*中的定义给出，我们有 e

$$|f_i^\pi(s_0,\pi)-V_i^\pi(s_0)|\le\frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.$$

证明. 应用引理20，得到

$$\begin{aligned}|f_i^\pi(s_0,\pi)-V_i^\pi(s_0)|&\le\frac{|\mathbb{E}_{s,\mathbf{a}\sim d^\pi}[f(s,\mathbf{a})-(\mathcal{T}_i^\pi f)(s,\mathbf{a})]|}{1-\gamma}\\&\le\frac{\|f-\mathcal{T}_i^\pi f\|_{2,d^\pi}}{1-\gamma}\le\frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.\end{aligned}$$

第二个不等式来自詹森不等式，最后一个不等式则由假设A得出。 □

对于版本空间 $\mathcal{F}_i^{\pi,\beta_f}$，我们定义

$$\begin{aligned}f_i^{\pi,\max}&:=\operatorname*{argmax}_{f_i\in\mathcal{F}_i^{\pi,\beta_f}}f_i(s_0,\pi)\\f_i^{\pi,\min}&:=\operatorname*{argmin}_{f_i\in\mathcal{F}_i^{\pi,\beta_f}}f_i(s_0,\pi).\end{aligned}$$

$f_i^{\pi,\max}(s_0,\pi)$ 和 $f_i^{\pi,\min}(s_0,\pi)$ 分别是价值函数$V_i^\pi(s_0)$ 的上界和下界。

引理 12. 在引理*8*的成功事件下，对于任意玩家 $i\in[m]$ 及任意 $\pi\in\Pi_i^{\text{ext}}$，以下两个不等式成立

$$\begin{aligned}f_i^{\pi,\max}(s_0,\pi)&\ge V_i^\pi(s_0)-\frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}\\f_i^{\pi,\min}(s_0,\pi)&\le V_i^\pi(s_0)+\frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.\end{aligned}$$

证明。 根据引理 10，我们知道，在引理 8的成功事件下， $f_i^\pi\in\mathcal{F}_i^{\pi,\beta_f}$。然后，我们调用引理 11 并得到

$$f_i^{\pi,\min}(s_0,\pi)\le f_i^\pi(s_0,\pi)\le Q_i^\pi(s_0,\pi)+\frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}=V_i^\pi(s_0)+\frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.$$

同理，我们有

$$f_i^{\pi,\max}(s_0,\pi)\ge f_i^\pi(s_0,\pi)\ge Q_i^\pi(s_0,\pi)-\frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}=V_i^\pi(s_0)-\frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.$$

□

现在我们证明， $\mathcal{E}_i(f_i,\ \pi;\ \mathcal{D})$ 可以有效地估计 $\|f_i-\mathcal{T}_i^\pi f_i\|_{2,\ d_D}^2$。

引理 13。 在引理 *8*的成功事件下，对于任意玩家 $i\in[m]$ 以及任意 $\pi\in\Pi_i^{\text{ext}}$，给定 $\varepsilon>0$，若$f_i\in\mathcal{F}_i$满足 $\mathcal{E}_i(f_i,\ \pi;\ \mathcal{D})\le\varepsilon$，则有

$$\|f_i-\mathcal{T}_i^\pi f_i\|_{2,d_D}\le 8V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}}+4\sqrt{\varepsilon_{\mathcal{F},\mathcal{F}}}+\sqrt{\varepsilon}.$$

证明。 令 g_i^π如 (16) 中所定义，我们先对项 $\|g_i^\pi - \mathcal{T}_i^\pi f_i\|_{2,\ d_D}$ 求上界。定义

$$f_{i,d_D}^\pi \coloneqq \operatorname*{argmin}_{f_i' \in \mathcal{F}_i} \left\|f_i' - \mathcal{T}_i^\pi f_i\right\|_{2,d_D}^2 .$$

利用引理8，我们得到

$$\begin{aligned}
&\Bigg| \|g_i^\pi - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 - \left\|f_{i,d_D}^\pi - \mathcal{T}_i^\pi f_i\right\|_{2,d_D}^2 - \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s') \in \mathcal{D}} \left(g(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2 \\
&\quad + \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s') \in \mathcal{D}} \left(f_{i,d_D}^\pi(s,\mathbf{a}) - r - \gamma f_i(s',\pi)\right)^2 \Bigg| \\
&\leq 2V_{\max} \|g_i^\pi - f_{i,d_D}^\pi\|_{2,d_D} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}.
\end{aligned}$$

重新整理各项后，我们有

$$\begin{aligned}
&\|g_i^\pi - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 \\
&\leq \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s') \in \mathcal{D}} \left(g_i^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2 - \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s') \in \mathcal{D}} \left(f_{i,d_D}^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2 \\
&\quad + \left\|f_{i,d_D}^\pi - \mathcal{T}_i^\pi f_i\right\|_{2,d_D}^2 + 2V_{\max} \|g_i^\pi - f_{i,d_D}^\pi\|_{2,d_D} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \\
&\leq \left\|f_{i,d_D}^\pi - \mathcal{T}_i^\pi f_i\right\|_{2,d_D}^2 + 2V_{\max} \|g_i^\pi - f_{i,d_D}^\pi\|_{2,d_D} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \\
&\leq \varepsilon_{\mathcal{F},\mathcal{F}} + 2V_{\max} \|g_i^\pi - \mathcal{T}_i^\pi f_i\|_{2,d_D} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + 2V_{\max} \sqrt{\varepsilon_{\mathcal{F},\mathcal{F}}} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \\
&\leq 2V_{\max} \|g_i^\pi - \mathcal{T}_i^\pi f_i\|_{2,d_D} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{2V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} + 2\varepsilon_{\mathcal{F},\mathcal{F}}. \qquad (19)
\end{aligned}$$

第二个不等式来自 g_i^π的最优性。第三个不等式由假设B 以及$\|g_i^\pi - f_{i,d_D}^\pi\|_{2,d_D} \leq \|g_i^\pi - \mathcal{T}^\pi f_i\|_{2,\ d_D} + \|f_{i,d_D}^\pi - \mathcal{T}^\pi f_i\|_{2,d_D}$得到。最后一个不等式来自 $\sqrt{ab} \leq \frac{a+b}{2}$。解(19)，可得

$$\|g_i^\pi - \mathcal{T}_i^\pi f_i\|_{2,d_D} \leq 3V_{\max} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \sqrt{2\varepsilon_{\mathcal{F},\mathcal{F}}}. \qquad (20)$$

接着，我们对引理 8 应用 f_i^π

$$\begin{aligned}
&\Bigg| \|f_i - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 - \|g_i^\pi - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 \\
&\quad - \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s') \in \mathcal{D}} \left(f_i(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2 + \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s') \in \mathcal{D}} \left(g_i^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2 \Bigg| \\
&\leq 2V_{\max} \|f_i - g_i^\pi\|_{2,d_D} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}
\end{aligned}$$

$$\leq 2V_{\max}\left(\|f_i-\mathcal{T}_i^\pi f_i\|_{2,d_D}+\|g_i^\pi-\mathcal{T}_i^\pi f_i\|_{2,d_D}\right)\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}}+\frac{V_{\max}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}$$

$$\leq 2V_{\max}\|f_i-\mathcal{T}_i^\pi f_i\|_{2,d_D}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}}+3V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F},\mathcal{F}}}+\frac{7V_{\max}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}. \quad \text{(By (20))}$$

重新排列各项后，我们得到

$$\begin{aligned}
&\|f_i-\mathcal{T}_i^\pi f_i\|_{2,d_D}^2\\
\leq\ &\|g_i^\pi-\mathcal{T}_i^\pi f_i\|_{2,d_D}^2+\frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}}\left(f_i(s,\mathbf{a})-\mathbf{r}_i-\gamma f_i(s',\pi)\right)^2-\frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}}\left(g_i^\pi(s,\mathbf{a})-\mathbf{r}_i-\gamma f_i(s',\pi)\right)^2\\
&+2V_{\max}\|f_i-\mathcal{T}_i^\pi f_i\|_{2,d_D}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}}+3V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F},\mathcal{F}}}+\frac{7V_{\max}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}\\
\leq\ &\left(3V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}}+\sqrt{2\varepsilon_{\mathcal{F},\mathcal{F}}}\right)^2+\varepsilon \qquad \text{(By (20) and } \mathcal{E}(f,\pi;\mathcal{D})\leq\varepsilon)\\
&+2V_{\max}\|f_i-\mathcal{T}_i^\pi f_i\|_{2,d_D}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}}+3V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F},\mathcal{F}}}+\frac{7V_{\max}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}\\
=\ &2V_{\max}\|f_i-\mathcal{T}_i^\pi f_i\|_{2,d_D}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}}+12V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F},\mathcal{F}}}+\frac{16V_{\max}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}+2\varepsilon_{\mathcal{F},\mathcal{F}}+\varepsilon.
\end{aligned} \tag{21}$$

求解 (21) 并使用算术-几何平均不等式，即可完成证明。 □

我们对 Δ_i^π作如下上界估计。

命题 2（区间宽度的界）。以至少 $1-\delta$ 的概率，对于任意玩家 $i\in[m]$ 和任意$\pi\in\Pi_i^{\text{ext}}$，我们有

$$\Delta_i^\pi\leq\min_d\frac{1}{1-\gamma}\sqrt{\mathscr{C}(d;d_D,\mathcal{F}_i,\pi)}\varepsilon_{\text{apx}}+\frac{1}{1-\gamma}\sum_{s,\mathbf{a}}(d^\pi\backslash d)(s,\mathbf{a})\left[\Delta f_i^\pi(s,\mathbf{a})-\gamma(P^\pi\Delta f_i^\pi)(s,\mathbf{a})\right], \tag{7}$$

where $\varepsilon_{\text{apx}}=\mathcal{O}\left(V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}}+\sqrt{\varepsilon_{\mathcal{F}}+\varepsilon_{\mathcal{F},\mathcal{F}}}\right)$, $(d^\pi\setminus d)(s,\mathbf{a}):=\max(d^\pi(s,\mathbf{a})-d(s,\mathbf{a}),0)$, $\Delta f_i^\pi(s,\mathbf{a}):=f_i^{\pi,\max}(s,\mathbf{a})-f_i^{\pi,\min}(s,\mathbf{a})$, *and* $(P^\pi f)(s,\mathbf{a})=\mathbb{E}_{s'\sim P(\cdot|s,\mathbf{a})}[f(s',\pi)]$.

证明。我们对 $f^{\pi,\max}_i$ 和 $f^{\pi,\min}_i$ 应用引理 20，得到

$$\begin{aligned}
&f_i^{\pi,\max}(s_0,\pi)-f_i^{\pi,\min}(s_0,\pi)\\
=\ &f_i^{\pi,\max}(s_0,\pi)-V_i^\pi(s_0)+V_i^\pi(s_0)-f_i^{\pi,\min}(s_0,\pi).\\
=\ &\frac{1}{1-\gamma}\left(\mathbb{E}_{d_\pi}\left[f_i^{\pi,\max}-\mathcal{T}_i^\pi f_i^{\pi,\max}\right]-\mathbb{E}_{d_\pi}\left[f_i^{\pi,\min}-\mathcal{T}_i^\pi f_i^{\pi,\min}\right]\right) \qquad \text{(By Lemma 20)}\\
=\ &\frac{1}{1-\gamma}\Big(\mathbb{E}_d\left[(f_i^{\pi,\max}-\mathcal{T}_i^\pi f_i^{\pi,\max})-\left(f_i^{\pi,\min}-\mathcal{T}_i^\pi f_i^{\pi,\min}\right)\right]\\
&+\mathbb{E}_{d_\pi}\left[(f_i^{\pi,\max}-\mathcal{T}_i^\pi f_i^{\pi,\max})-(f_i^{\pi,\min}-\mathcal{T}_i^\pi f_i^{\pi,\min})\right]
\end{aligned}$$

$$
\begin{aligned}
&\quad - \mathbb{E}_d\left[(f_i^{\pi,\max} - \mathcal{T}_i^\pi f_i^{\pi,\max}) - (f_i^{\pi,\min} - \mathcal{T}_i^\pi f_i^{\pi,\min})\right])\\
&= \frac{1}{1-\gamma}\underbrace{\mathbb{E}_d\left[(f_i^{\pi,\max} - \mathcal{T}_i^\pi f_i^{\pi,\max}) - \left(f_i^{\pi,\min} - \mathcal{T}_i^\pi f_i^{\pi,\min}\right)\right]}_{\text{(I)}}\\
&\quad + \frac{1}{1-\gamma}\underbrace{(\mathbb{E}_{d_\pi}\left[\Delta f_i^\pi - \gamma P^\pi \Delta f_i^\pi\right] - \mathbb{E}_d\left[\Delta f_i^\pi - \gamma P^\pi \Delta f_i^\pi\right])}_{\text{(II)}}, \qquad (\Delta f_i^\pi := f_i^{\pi,\max} - f_i^{\pi,\min})
\end{aligned}
$$

其中 $d \in \Delta(\mathcal{S}\times\mathcal{A})$ 是任意分布。对于项 (I)，我们有

$$
\begin{aligned}
\text{(I)} &\le \left|\mathbb{E}_d\left[(f_i^{\pi,\max} - \mathcal{T}_i^\pi f_i^{\pi,\max})\right]\right| + \left|\mathbb{E}_d\left[\left(f_i^{\pi,\min} - \mathcal{T}_i^\pi f_i^{\pi,\min}\right)\right]\right|\\
&\le \|f_i^{\pi,\max} - \mathcal{T}_i^\pi f_i^{\pi,\max}\|_{2,d} + \|f_i^{\pi,\min} - \mathcal{T}_i^\pi f_i^{\pi,\min}\|_{2,d} \qquad \text{(By Jensen's inequality)}\\
&\le \sqrt{\mathscr{C}(d; d_D, \mathcal{F}_i, \pi)}\left(\|f_i^{\pi,\max} - \mathcal{T}_i^\pi f_i^{\pi,\max}\|_{2,d_D} + \|f_i^{\pi,\min} - \mathcal{T}_i^\pi f_i^{\pi,\min}\|_{2,d_D}\right).
\end{aligned}
$$

回忆一下，$f_i^{\pi,\max} := \operatorname{argmax}_{f_i\in\mathcal{F}^\pi,\ \beta_f} f_i(s_0,\pi)$ 和 $f_i^{\pi,\min} := \operatorname{argmin}_{f_i\in\mathcal{F}^\pi,\ \beta_f} f_i(s_0,\pi)$，以及 $\beta_f = 80V^2\max\log\frac{|\mathcal{F}||\Pi|}{\delta}/n + \text{ext}30\varepsilon_{\mathcal{F}}$。我们应用引理 13，得到

$$
\text{(I)} \le \sqrt{\mathscr{C}(d; d_D, \mathcal{F}_i, \pi)}\mathcal{O}\left(V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}\right). \tag{22}
$$

对于项 (II)，我们有

$$
\begin{aligned}
\text{(II)} &\le \sum_{(s,\mathbf{a})\in\mathcal{S}\times\mathcal{A}} (d_\pi \setminus d)(s,\mathbf{a})\left[\Delta f_i^\pi(s,\mathbf{a}) - \gamma(P^\pi\Delta f_i^\pi)(s,\mathbf{a})\right]\\
&\quad + \sum_{(s,\mathbf{a})\in\mathcal{S}\times\mathcal{A}} \mathbb{I}(d(s,\mathbf{a}) > d_\pi(s,\mathbf{a}))\left[d(s,\mathbf{a}) - d_\pi(s,\mathbf{a})\right]\left|\Delta f_i^\pi(s,\mathbf{a}) - \gamma(P^\pi\Delta f_i^\pi)(s,\mathbf{a})\right|\\
&\le \sum_{(s,\mathbf{a})\in\mathcal{S}\times\mathcal{A}} (d_\pi \setminus d)(s,\mathbf{a})\left[\Delta f_i^\pi(s,\mathbf{a}) - \gamma(P^\pi\Delta f_i^\pi)(s,\mathbf{a})\right]\\
&\quad + \mathbb{E}_d\left[|f_i^{\pi,\max} - \mathcal{T}_i^\pi f_i^{\pi,\max}| + \left|f_i^{\pi,\min} - \mathcal{T}_i^\pi f_i^{\pi,\min}\right|\right]\\
&\le \sum_{(s,\mathbf{a})\in\mathcal{S}\times\mathcal{A}} (d_\pi \setminus d)(s,\mathbf{a})\left[\Delta f_i^\pi(s,\mathbf{a}) - \gamma(P^\pi\Delta f_i^\pi)(s,\mathbf{a})\right]\\
&\quad + \sqrt{\mathscr{C}(d; d_D, \mathcal{F}_i, \pi)}\mathcal{O}\left(V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}\right).
\end{aligned} \tag{23}
$$

最后一步由对项 (I) 的分析得到。结合 (22) 和 (23)，可得

$$
\begin{aligned}
f_i^{\pi,\max}(s_0,\pi) - f_i^{\pi,\min}(s_0,\pi) &\le \min_d \frac{1}{1-\gamma}\sqrt{\mathscr{C}(d; d_D, \mathcal{F}_i, \pi)}\mathcal{O}\left(V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}\right)\\
&\quad + \frac{1}{1-\gamma}\sum_{(s,\mathbf{a})\in\mathcal{S}\times\mathcal{A}} (d_\pi \setminus d)(s,\mathbf{a})\left[\Delta f_i^\pi(s,\mathbf{a}) - \gamma(P^\pi\Delta f_i^\pi)(s,\mathbf{a})\right].
\end{aligned}
$$

证明完成。 □

接下来，我们证明 $\mathrm{Gap}^{\Pi^\dagger,\mathrm{EQ}}(\pi)$ 的上界为估计间隔 $\widehat{\mathrm{Gap}}^{\mathrm{EQ}}(\pi)$。

引理14。 在引理 *8*的成功事件下，对于任意 $\pi \in \Pi$，我们有

$$\mathrm{Gap}^{\Pi^\dagger,\mathrm{EQ}}(\pi) \le \widehat{\mathrm{Gap}}_{\mathrm{EQ}}(\pi) + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.$$

证明。 令 $\pi_i^\dagger = \mathrm{argmax}_{\pi^\dagger \in \Pi_i^\dagger(\pi)} V_i^{\pi^\dagger}(s_0)$。根据引理 12，对于任意玩家 $i \in [m]$，以至少 $1-\delta$ 的概率，

$$V_i^{\pi_i^\dagger}(s_0) \le \max_{f_i \in \mathcal{F}_i^{\pi_i^\dagger,\beta_f}} f_i(s_0, \pi_i^\dagger) + \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.$$

回顾间隙Gap的定义$\mathrm{Gap}^{\Pi^\dagger,\mathrm{EQ}}(\pi)$，我们得到

$$\begin{aligned}
\mathrm{Gap}^{\Pi^\dagger,\mathrm{EQ}}(\pi) &= \max_{i\in[m]} \max_{\pi^\dagger \in \Pi_i^\dagger(\pi)} V_i^{\pi^\dagger}(s_0) - V_i^{\pi}(s_0) \\
&\le \max_{i\in[m]} \left(\max_{f_i \in \mathcal{F}_i^{\pi_i^\dagger,\beta_f}} f_i(s_0,\pi_i^\dagger) - V_i^{\pi}(s_0) \right) + \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}. \\
&\le \max_{i\in[m]} \left(\max_{\pi^\dagger \in \Pi_i^\dagger(\pi)} \overline{V}_i^{\pi^\dagger}(s_0) - V_i^{\pi}(s_0) \right) + \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} \\
&\qquad \text{(By definition of } \overline{V}_i^{\pi^\dagger}(s_0)\text{)} \\
&\le \max_{i\in[m]} \left(\max_{\pi^\dagger \in \Pi_i^\dagger(\pi)} \overline{V}_i^{\pi^\dagger}(s_0) - \underline{V}_i^{\pi}(s_0) \right) + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} \\
&\qquad \text{(By definition of } \underline{V}_i^{\pi}(s_0) \text{ and Lemma 12)} \\
&= \widehat{\mathrm{Gap}}_{\mathrm{EQ}}(\pi) + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.
\end{aligned}$$

□

现在我们可以证明定理3。

定理3。 以至少 $1-\delta$ 的概率，对于任意 $\pi \in \Pi$ 和 $\mathrm{EQ} \in$ {纳什均衡、相关均衡、粗相关均衡}，算法*1* 的输出策略 $\widehat{\pi}$满足

$$\mathrm{Gap}^{\Pi^\dagger,\mathrm{EQ}}(\widehat{\pi}) \le \mathrm{Gap}^{\Pi^\dagger,\mathrm{EQ}}(\pi) + \frac{4\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} + \max_{i\in[m]} \min_{\widetilde{\pi}_i \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} \left(\Delta_i^{\widetilde{\pi}_i} + \Delta_i^{\pi} + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i) \right),$$

其中 次优$_i^{\pi}(\widetilde{\pi}_i)$：$=$ 最大值$_{\pi^\dagger \in \Pi_i^\dagger(\pi)} V_i^{\pi^\dagger}(s_0) - V_i^{\widetilde{\pi}_i}(s_0)$。

证明。 令 $\pi_i^\dagger = \mathrm{argmax}_{\pi^\dagger \in \Pi_i^\dagger(\pi)} V_i^{\pi^\dagger}(s_0)$。以至少 $1-\delta$ 的概率，对于每个玩家 $i \in [m]$，我们将 $V_i^{\pi_i^\dagger}(s_0) - V_i^{\pi}(s_0)$ 的上界估计为

$$\begin{aligned}
\overline{V}_i^{\pi_i^\dagger}(s_0) - \underline{V}_i^{\pi}(s_0) &= \overline{V}_i^{\widetilde{\pi}_i}(s_0) - \underline{V}_i^{\pi}(s_0) + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i) \qquad (\widetilde{\pi}_i \text{ is an arbitrary policy from } \Pi_i^{\mathrm{ext}}(\pi)) \\
&= f_i^{\widetilde{\pi}_i,\max}(s_0) - f_i^{\pi,\min}(s_0) + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i)
\end{aligned}$$

$$\leq V_i^{\widetilde{\pi}_i}(s_0) + \Delta_i^{\widetilde{\pi}_i} - f_i^{\pi,\max}(s_0) + \Delta_i^{\pi} + \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i)$$

(By definition of $\Delta_i^{\widetilde{\pi}_i}$ and Lemma 12)

$$\leq V_i^{\widetilde{\pi}_i}(s_0) - V_i^{\pi}(s_0) + \Delta_i^{\widetilde{\pi}_i} + \Delta_i^{\pi} + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i) \qquad \text{(By Lemma 12)}$$

$$\leq \max_{\pi^\dagger \in \Pi_i^\dagger(\pi)} V_i^{\pi^\dagger}(s_0) - V_i^{\pi}(s_0) + \Delta_i^{\widetilde{\pi}_i} + \Delta_i^{\pi} + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i).$$

这直接意味着

$$\widehat{\mathrm{Gap}}_{\mathrm{EQ}}(\pi) \leq \mathrm{Gap}^{\Pi^\dagger,\mathrm{EQ}}(\pi) + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} + \max_{i\in[m]} \min_{\widetilde{\pi}_i \in \Pi_i^\dagger(\pi)} \left(\Delta_i^{\widetilde{\pi}_i} + \Delta_i^{\pi} + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i)\right). \tag{24}$$

根据 $\widehat{\pi}$ 的最优性，对于任意 $\pi \in \Pi$，我们有

$$\begin{aligned}
\mathrm{Gap}^{\Pi^\dagger,\mathrm{EQ}}(\widehat{\pi}) &\leq \widehat{\mathrm{Gap}}_{\mathrm{EQ}}(\widehat{\pi}) + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} \\
&\leq \widehat{\mathrm{Gap}}_{\mathrm{EQ}}(\pi) + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} \\
&\leq \mathrm{Gap}^{\Pi^\dagger,\mathrm{EQ}}(\pi) + \frac{4\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} + \max_{i\in[m]} \min_{\widetilde{\pi}_i \in \Pi_i^\dagger(\pi)} \left(\Delta_i^{\widetilde{\pi}_i} + \Delta_i^{\pi} + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i)\right). \qquad \text{(By (24))}
\end{aligned}$$

证明至此完成。 □

D 第5节的证明

本节中，我们证明定理 5。我们首先给出一些集中度结果。

Le引理 15. 以至少 $1-\delta$ 的概率，对于任意 g_1、g_2、$h \in \mathcal{G}_i$ 和 $\pi \in \Pi_i^{\mathrm{ext}}$，我们有

$$\begin{aligned}
&\Big| \|g_1 - \mathcal{T}_i^\pi h\|_{2,d_S}^2 - \|g_2 - \mathcal{T}_i^\pi h\|_{2,d_S}^2 \\
&\quad - \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(g_1(s) - \mathbf{r}_i - \gamma h(s')\right)^2 + \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(g_2(s) - \mathbf{r}_i - \gamma h(s')\right)^2 \Big| \\
&\leq 2V_{\max}\|g_1 - g_2\|_{2,d_S} \sqrt{C_A(\pi) \frac{\log \frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n}} + \frac{C_A(\pi) V_{\max}^2 \log \frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n},
\end{aligned}$$

其中 $C_A(\pi) :=$ 最大值$_{s,\ \mathbf{a}\ }$ $\frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}$。

证明。 首先，我们观察到

$$\begin{aligned}
&\|g_1 - \mathcal{T}_i^\pi h\|_{2,d_S}^2 - \|g_2 - \mathcal{T}_i^\pi h\|_{2,d_S}^2 \\
&= \mathbb{E}_{s\sim d_S, \mathbf{a}\sim\pi(\cdot|s), s'\sim P(\cdot|s,\mathbf{a})}[(g_1(s) - r_i(s,\mathbf{a}) - \gamma h(s'))^2] - \mathbb{E}_{s\sim d_S, \mathbf{a}\sim\pi(\cdot|s), s'\sim P(\cdot|s,\mathbf{a})}[(g_2(s) - r_i(s,\mathbf{a}) - \gamma h(s'))^2].
\end{aligned}$$

令随机变量 $X = \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}(g_1(s_i) - r_i(s_i,\mathbf{a}) - \gamma h(s_i'))^2 - \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}(g_2(s_i) - r_i(s_i,\mathbf{a}) - \gamma h(s_i'))^2$，$X$ 取自 $d_S \times d_A \times \mathcal{P}$。于是得到

$$\Big| \|g_1 - \mathcal{T}_i^\pi f\|_{2,d_S}^2 - \|g_2 - \mathcal{T}_i^\pi f\|_{2,d_S}^2$$

$$
\left. - \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s') \in \mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(g_1(s) - \mathbf{r}_i - \gamma h(s')\right)^2 + \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s') \in \mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(g_2(s) - \mathbf{r}_i - \gamma h(s')\right)^2 \right|
$$

$$
= \left| \mathbb{E}_{d_S \times \pi \times \mathcal{P}} \left[(g_1(s) - r_i(s,\mathbf{a}) - \gamma h(s'))^2 \right] - \mathbb{E}_{d_S \times \pi \times \mathcal{P}} \left[(g_2(s) - r_i(s,\mathbf{a}) - \gamma h(s'))^2 \right] - \frac{1}{n} \sum_{i=1}^{n} X_i \right|
$$

（根据 X的定义）

此处 $\mathbb{E}_{d_S \times d_A \times \mathcal{P}}[X] = \mathbb{E}_{d_S \times \pi \times \mathcal{P}} \left[(g_1(s) - r_i(s, \mathbf{a}) - \gamma h(s'))^2 \right] - \mathbb{E}_{d_S \times \pi \times \mathcal{P}} \left[(g_2(s) - r_i(s, \mathbf{a}) - \gamma h(s'))^2 \right]$。对于方差，我们有

$$
\begin{aligned}
& \mathbb{V}_{d_S \times d_A \times \mathcal{P}}[X] \\
& \leq \mathbb{E}_{d_S \times d_A \times \mathcal{P}} \left[\frac{\pi(\mathbf{a}|s)^2}{d_A(\mathbf{a}|s)^2} (g_1(s) - g_2(s))^2 (g_1(s) + g_2(s) - 2r_i(s,\mathbf{a}) - 2\gamma h(s'))^2 \right] \\
& \leq 4V_{\max}^2 \mathbb{E}_{d_S \times d_A} \left[\frac{\pi(\mathbf{a}|s)^2}{d_A(\mathbf{a}|s)^2} (g_1(s) - g_2(s))^2 \right] \\
& \leq 4V_{\max}^2 \max_{s,\mathbf{a}} \frac{\pi(\mathbf{a}|s)^2}{d_A(\mathbf{a}|s)} \mathbb{E}_{d_S} \left[(g_1(s) - g_2(s))^2 \right]
\end{aligned}
$$

令 $C_A(\pi) := \max_{s,\ \mathbf{a}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}$。根据Freedman不等式和并集界，以至少 $1-\delta$ 的概率，我们有

$$
\left\| \mathbb{E}[X] - \frac{1}{n} \sum_{i=1}^{n} X_i \right\| \leq \sqrt{\frac{4V_{\max}^2 C_A(\pi) \|g_1 - g_2\|_{d_S}^2 \log \frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{C_A(\pi) V_{\max}^2 \log \frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}.
$$

□

对于任意玩家 $i \in [m]$ 和 $\pi \in \Pi_i^{\text{ext}}$，定义

$$
g_i^\pi := \operatorname*{argmin}_{g \in \mathcal{G}_i} \sup_{\text{admissible } d} \|g - \mathcal{T}_i^\pi g\|_{2,d}^2 \tag{25}
$$

$$
h_i^\pi := \operatorname*{argmin}_{h \in \mathcal{G}_i} \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s') \in \mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(h(s) - \mathbf{r}_i - \gamma g_i^\pi(s')\right)^2. \tag{26}
$$

我们如下界定 $\|g_i^\pi - h_i^\pi\|_{2,\ d_S}$。

引理16。 令 g_i^π和 h_i^π按照公式 (25) 和 (26)定义。在引理 *15*的成功事件下，对于任意玩家 $i \in [m]$ 和 $\pi \in \Pi_i^{\text{ext}}$，有

$$
\|g_i^\pi - h_i^\pi\|_{2,d_S} \leq 6V_{\max} \sqrt{C_A(\pi) \frac{\log \frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}} + 2\sqrt{\varepsilon_{\mathcal{F}}}.
$$

证明通过调用引理 15 来处理 g_i^π和 h_i^π，计算过程与引理 9相同。与引理 10类似，我们证明 V_i^π的最佳近似包含在 $\mathcal{G}^{\pi}{}_{,i}^{\ \beta_g}$ 中。

引理17. 在引理*15*的成功事件下，对于任意玩家 $i \in [m]$ 和 $\pi \in \Pi_i^{\text{ext}}$，以下关于 $\mathcal{E}_i(g_i^\pi, \pi; \mathcal{D})$ 的不等式成立

$$
\mathcal{E}_i(g_i^\pi, \pi; \mathcal{D}) \leq \frac{80 C_A(\pi) V_{\max}^2 \log \frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n} + 30\varepsilon_{\mathcal{F}} =: \beta_g.
$$

证明. 应用引理15 和引理16，可得

$$\begin{aligned}
&\Big| \|g_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}^2 - \|h_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}^2 - \\
&\frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(g_i^\pi(s) - \mathbf{r}_i - \gamma g_i^\pi(s')\right)^2 + \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(h_i^\pi(s) - \mathbf{r}_i - \gamma g_i^\pi(s')\right)^2 \Big| \\
\le\ & 4V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F}}} + \frac{13C_A(\pi)V_{\max}^2\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}. \qquad (27)
\end{aligned}$$

与引理10类似，我们对$\|g_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,\ d_S}^2 - \|h_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,\ d_S}^2$作如下界定，

$$\begin{aligned}
&\|g_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}^2 - \|h_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}^2 \\
\le\ & \left(\|g_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S} + \|h_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}\right)\left|\|g_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S} - \|h_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}\right| \\
\le\ & 36V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F}}} + 36V_{\max}^2\frac{C_A(\pi)\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n} + 8\varepsilon_{\mathcal{F}}. \qquad \text{(By Lemma 16)}
\end{aligned}$$

将其与(27)结合，可得

$$\begin{aligned}
&\frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(g_i^\pi(s) - \mathbf{r}_i - \gamma g_i^\pi(s')\right)^2 - \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(h_i^\pi(s) - \mathbf{r}_i - \gamma g_i^\pi(s')\right)^2 \\
\le\ & \|g_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}^2 - \|h_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}^2 + 4V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F}}} + \frac{13C_A(\pi)V_{\max}^2\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n} \\
\le\ & \frac{80C_A(\pi)V_{\max}^2\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n} + 30\varepsilon_{\mathcal{F}}. \qquad \text{(By AM-GM inequality)}
\end{aligned}$$

□

接下来我们证明，$g^{\pi,\ i}{}^{\max}(s_0)$和$g^{\pi,\ i}{}^{\min}(s_0)$分别是价值函数$V_i^\pi(s_0)$的上界和下界。

引理18。 在引理15的成功事件下，对任意玩家$i\in[m]$以及任意$\pi\in\Pi_i^{\text{ext}}$，以下两个不等式成立

$$\begin{aligned}
g_i^{\pi,\max}(s_0) &\ge V_i^\pi(s_0) - \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} \\
g_i^{\pi,\min}(s_0) &\le V_i^\pi(s_0) + \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.
\end{aligned}$$

证明。 令g_i^π如(25)所定义。根据引理21，可得

$$\begin{aligned}
|g_i^\pi(s_0) - V_i^\pi(s_0)| &\le \frac{\mathbb{E}_{s,\mathbf{a}\sim d^\pi, s'\sim P(\cdot|s,\mathbf{a})}\left[g(s) - r_i(s,\mathbf{a}) - \gamma g(s')\right]}{1-\gamma} \\
&\le \frac{\|g - \mathcal{T}_i^\pi g\|_{2,d^\pi}}{1-\gamma} \le \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.
\end{aligned}$$

根据引理17，我们知道$g_i^\pi\in\mathcal{G}^{\pi,\ i}{}^{\beta_g}$。于是得到

$$g_i^{\pi,\max}(s_0) \ge g_i^\pi(s_0) \ge V_i^\pi(s_0) - \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.$$

关于$g^{\pi,\ i}{}^{\min}$的情况也类似。 □

现在我们证明， $\mathcal{E}_i(g_i,\ \pi;\ \mathcal{D})$能够有效估计$\|g_i - \mathcal{T}_i^\pi g_i\|_{2,\ d_S}^2$。

引理19。 在引理15的成功事件下，对于任意玩家 $i \in [m]$ 以及任意 $\pi \in \Pi_i^{\text{ext}}$，给定 $\varepsilon > 0$，若$g_i \in \mathcal{G}_i$满足 $\mathcal{E}_i(g_i,\ \pi;\ \mathcal{D}) \le \varepsilon$，则有

$$\|g_i - \mathcal{T}_i^\pi g_i\|_{2,d_S} \le 8V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}} + 4\sqrt{\varepsilon_{\mathcal{F},\mathcal{F}}} + \sqrt{\varepsilon}.$$

证明。 令 h_i^π如(26)中所定义，并定义

$$g_{i,d_S}^\pi \coloneqq \operatorname*{argmin}_{g_i' \in \mathcal{G}_i} \left\|g_i' - \mathcal{T}_i^\pi g_i\right\|_{2,d_S}^2.$$

与引　　理13类似，我们首先对$\|h_i^\pi - \mathcal{T}_i^\pi g_i\|_{2,\ d_S}$求上界。利用引理15，可得:

$$\begin{aligned}
&\Bigg| \|h_i^\pi - \mathcal{T}_i^\pi g_i\|_{2,d_S}^2 - \left\|g_{i,d_S}^\pi - \mathcal{T}_i^\pi g_i\right\|_{2,d_S}^2 - \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}\left(h_i^\pi(s) - \mathbf{r}_i - \gamma g_i(s')\right)^2 \\
&\quad + \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}\left(g_{i,d_S}^\pi(s) - r - \gamma g_i(s')\right)^2 \Bigg| \\
&\le 2V_{\max}\|h_i^\pi - g_{i,d_S}^\pi\|_{2,d_S}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{C_A(\pi)V_{\max}^2\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}.
\end{aligned}$$

重新整理各项，并按照引理13中的类似计算，可得

$$\begin{aligned}
&\|h_i^\pi - \mathcal{T}_i^\pi g_i\|_{2,d_S}^2 \\
&\le 2V_{\max}\|h_i^\pi - \mathcal{T}_i^\pi g_i\|_{2,d_S}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{2C_A(\pi)V_{\max}^2\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n} + 2\varepsilon_{\mathcal{F},\mathcal{F}}. \qquad (28)
\end{aligned}$$

解(28)，可得

$$\|h_i^\pi - \mathcal{T}_i^\pi g_i\|_{2,d_S} \le 3V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}} + \sqrt{2\varepsilon_{\mathcal{F},\mathcal{F}}}. \qquad (29)$$

然后，我们调用引理15 来处理 g_i^π并得到

$$\begin{aligned}
&\Bigg| \|g_i - \mathcal{T}_i^\pi g_i\|_{2,d_S}^2 - \|g_i^\pi - \mathcal{T}_i^\pi g_i\|_{2,d_S}^2 \\
&\quad - \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}\left(g_i(s) - \mathbf{r}_i - \gamma g_i(s')\right)^2 + \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}\left(g_i^\pi(s) - \mathbf{r}_i - \gamma g_i(s')\right)^2 \Bigg| \\
&\le 2V_{\max}\|g_i - \mathcal{T}_i^\pi g_i\|_{2,d_S}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}} + 3V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F},\mathcal{F}}} + \frac{7C_A(\pi)V_{\max}^2\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}.
\end{aligned}$$

通过与引理13类似的计算，我们整理各项并得到

$$\begin{aligned}
&\|g_i - \mathcal{T}_i^\pi g_i\|_{2,d_S}^2 \\
&= 2V_{\max}\|g_i - \mathcal{T}_i^\pi g_i\|_{2,d_S}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}} + 12V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F},\mathcal{F}}} + \frac{16C_A(\pi)V_{\max}^2\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n} + 2\varepsilon_{\mathcal{F},\mathcal{F}} \qquad (30)
\end{aligned}$$

解出（21）并使用算术-几何平均不等式即可完成证明。 □

现在我们可以证明定理5。

定理 5 （V 型保证）。 以至少 $1-\delta$ 的概率，对于任意 $\pi \in \Pi$ 和 均衡 $\in$ {纳什均衡、相关均衡、粗相关均衡}，由式 $\widehat{\pi}$ 得到的输出策略*(9)* 满足

$$\mathrm{Gap}^{\Pi^{\dagger},\mathrm{EQ}}(\widehat{\pi}) \le \mathrm{Gap}^{\Pi^{\dagger},\mathrm{EQ}}(\pi) + \frac{4\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} + \max_{i\in[m]} \min_{\widetilde{\pi}_i \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} \left(\Delta_i^{\widetilde{\pi}_i} + \Delta_i^{\pi} + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i)\right),$$

其中 $\Delta_i^{\pi} = g^{\pi,\,\max}_{i}(s_0) - g^{\pi,\,\min}_{i}(s_0)$ 以及 次优$_i^{\widetilde{\pi}_i}$ $=$ 最大值$_{\pi^{\dagger}\in\Pi^{\dagger,\,\mathrm{EQ}}_{i}(\pi)}$ $g^{\pi^{\dagger},\ \text{最大值}\,i}(s_0)$ $- g^{\widetilde{\pi}_i,\ \text{最大值}\,i}(s_0)$。此外，以至少 $1-\delta$ 的概率，对于任意玩家 $i \in [m]$ 和任意 $\pi \in \Pi_i^{\mathrm{ext}}$，有

$$\Delta_i^{\pi} \le \min_{d\in\Delta(S)} \frac{1}{1-\gamma}\sqrt{\mathscr{C}(d; d_S, \mathcal{G}_i, \pi)}\varepsilon_{\mathrm{apx}} + \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}}(d^{\pi} \setminus d)(s)\left[\Delta g_i^{\pi}(s) - \gamma(P_i^{\pi}\Delta g_i^{\pi})(s)\right],$$

where $\varepsilon_{\mathrm{apx}} = \mathcal{O}\left(V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}\right)$, $(d^{\pi}\setminus d)(s) \coloneqq \max(d^{\pi}(s) - d(s), 0)$, $\Delta g_i^{\pi}(s) \coloneqq g_i^{\pi,\max}(s) - g_i^{\pi,\min}(s)$, *and* $(P_i^{\pi}g)(s) = \mathbb{E}_{\mathbf{a}\sim\pi(\cdot|s),s'\sim P(\cdot|s,\mathbf{a})}[r_i(s,\mathbf{a}) + g(s')]$.

证明。 第一部分的证明与定理 3相同。对于第二部分，我们对$g^{\pi,\,\min}_{i}$ 和 $g^{\pi,\,\max}_{i}$ 应用引理 21

$$\begin{aligned}
&g_i^{\pi,\max}(s_0) - g_i^{\pi,\min}(s_0) \\
&= \frac{1}{1-\gamma}\underbrace{\mathbb{E}_d\left[(g_i^{\pi,\max} - \mathcal{T}_i^{\pi}g_i^{\pi,\max}) - \left(g_i^{\pi,\min} - \mathcal{T}_i^{\pi}g_i^{\pi,\min}\right)\right]}_{\text{(I)}} \\
&\quad + \frac{1}{1-\gamma}\underbrace{\left(\mathbb{E}_{d_\pi}\left[\Delta g_i^{\pi} - \gamma P_i^{\pi}\Delta g_i^{\pi}\right] - \mathbb{E}_d\left[\Delta g_i^{\pi} - \gamma P_i^{\pi}\Delta g_i^{\pi}\right]\right)}_{\text{(II)}}, \qquad (\Delta g_i^{\pi} \coloneqq g_i^{\pi,\max} - g_i^{\pi,\min})
\end{aligned}$$

其中 $d \in \Delta(\mathcal{S})$是任意分布。对于项 (I)，有

$$\begin{aligned}
\text{(I)} &\le \left|\mathbb{E}_d\left[(g_i^{\pi,\max} - \mathcal{T}_i^{\pi}g_i^{\pi,\max})\right]\right| + \left|\mathbb{E}_d\left[\left(g_i^{\pi,\min} - \mathcal{T}_i^{\pi}g_i^{\pi,\min}\right)\right]\right| \\
&\le \|g_i^{\pi,\max} - \mathcal{T}_i^{\pi}g_i^{\pi,\max}\|_{2,d} + \|g_i^{\pi,\min} - \mathcal{T}_i^{\pi}g_i^{\pi,\min}\|_{2,d} \qquad \text{(By Jensen's inequality)} \\
&\le \sqrt{\mathscr{C}(d; d_S, \mathcal{G}_i, \pi)}\left(\|g_i^{\pi,\max} - \mathcal{T}_i^{\pi}g_i^{\pi,\max}\|_{2,d_S} + \|g_i^{\pi,\min} - \mathcal{T}_i^{\pi}g_i^{\pi,\min}\|_{2,d_S}\right).
\end{aligned}$$

回想一下 $\beta_g = \frac{80C_{A_{(\pi)V^2\text{最大}}\text{对数}\frac{|\mathcal{G}||\Pi}{\delta}}}{n}$ $+ 30\varepsilon_{\mathcal{F}}$ 极值。我们调用引理 13 并得到其成立的概率至少为 $1-\delta$

$$\text{(I)} \le \sqrt{\mathscr{C}(d; d_S, \mathcal{G}_i, \pi)}\mathcal{O}\left(V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}\right). \tag{31}$$

对于项（II），我们有

$$\begin{aligned}
\text{(II)} \le &\sum_{s\in\mathcal{S}}(d_\pi \setminus d)(s)\left[\Delta g_i^{\pi}(s) - \gamma(P_i^{\pi}\Delta g_i^{\pi})(s)\right] \\
&+ \sum_{(s)\in\mathcal{S}}\mathbb{I}(d(s) > d_\pi(s))\left[d(s) - d_\pi(s)\right]\left|\Delta g_i^{\pi}(s) - \gamma(P_i^{\pi}\Delta g_i^{\pi})(s)\right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{s\in\mathcal{S}}(d_\pi \setminus d)(s)\left[\Delta g_i^\pi(s) - \gamma(P_i^\pi \Delta g_i^\pi)(s)\right] \\
&\quad + \mathbb{E}_d\left[|g_i^{\pi,\max} - \mathcal{T}_i^\pi g_i^{\pi,\max}| + \left|g_i^{\pi,\min} - \mathcal{T}_i^\pi g_i^{\pi,\min}\right|\right] \\
&\leq \sum_{s\in\mathcal{S}}(d_\pi \setminus d)(s)\left[\Delta g_i^\pi(s) - \gamma(P_i^\pi \Delta g_i^\pi)(s)\right] \\
&\quad + \sqrt{\mathscr{C}(d; d_S, \mathcal{G}_i, \pi)}\mathcal{O}\left(V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}\right). \qquad (32)
\end{aligned}
$$

最后一步来自对项（I）的分析。结合（31）和（32），可得

$$
\begin{aligned}
g_i^{\pi,\max}(s_0) - g_i^{\pi,\min}(s_0) \leq \min_d \frac{1}{1-\gamma}\sqrt{\mathscr{C}(d; d_S, \mathcal{G}_i, \pi)}\mathcal{O}\left(V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}\right) \\
+ \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}}(d_\pi \setminus d)(s)\left[\Delta g_i^\pi(s) - \gamma(P_i^\pi \Delta g_i^\pi)(s)\right].
\end{aligned}
$$

证明至此完成。 □

E 辅助引理

引理 20 (Q函数评估误差引理)。 对于任意玩家 $i \in [m]$ 以及任意 $\pi \in \Pi_i^{\text{ext}}$，以及任意$f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$

$$
f(s_0, \pi) - V_i^\pi(s_0) = \frac{\mathbb{E}_{s,\mathbf{a}\sim d^\pi, s'\sim P(\cdot|s,\mathbf{a})}\left[f(s,\mathbf{a}) - r_i(s,\mathbf{a}) - \gamma f(s',\pi)\right]}{1-\gamma}
$$

证明。 我们注意到

$$
\begin{aligned}
&\sum_{s,\mathbf{a}}\sum_{t=0}^{\infty}\gamma^{t+1}\Pr(s_t = s, \mathbf{a}_t = \mathbf{a}|s_0,\pi)\sum_{s'}\Pr(s_{t+1} = s'|s_t = s, \mathbf{a}_t = \mathbf{a})f(s',\pi) \\
=& \sum_{s,\mathbf{a}}\sum_{t=1}^{\infty}\gamma^{t}\Pr(s_t = s, \mathbf{a}_t = \mathbf{a}|s_0,\pi)f(s,\mathbf{a})
\end{aligned}
$$

于是，我们有

$$
\begin{aligned}
&\frac{\mathbb{E}_{s,\mathbf{a}\sim d^\pi, s'\sim P(\cdot|s,\mathbf{a})}\left[f(s,\mathbf{a}) - \gamma f(s',\pi)\right]}{1-\gamma} \\
=& \sum_{s,\mathbf{a}}\sum_{t=0}^{\infty}\gamma^t\Pr(s_t = s, \mathbf{a}_t = \mathbf{a}|s_0,\pi)f(s,\mathbf{a}) - \sum_{s,\mathbf{a}}\sum_{t=1}^{\infty}\gamma^t\Pr(s_t = s, \mathbf{a}_t = \mathbf{a}|s_0,\pi)f(s,\mathbf{a}) \\
=& \sum_{\mathbf{a}}\Pr(\mathbf{a}_0 = \mathbf{a}|s_0,\pi)f(s_0,\mathbf{a}) = f(s_0,\pi).
\end{aligned}
$$

由于 $V_i^\pi(s_0) = \frac{\mathbb{E}_{d^\pi}[r_i(s,\mathbf{a})]}{1-\gamma}$，重新整理各项即可完成证明。 □

引理 21 (价值函数评估误差引理)。 对于任意玩家 $i \in [m]$ 以及任意 $\pi \in \Pi_i^{\text{ext}}$，以及任意$f \in \mathbb{R}^{\mathcal{S}}$

$$
f(s_0) - V_i^\pi(s_0) = \frac{\mathbb{E}_{s,\mathbf{a}\sim d^\pi, s'\sim P(\cdot|s,\mathbf{a})}\left[f(s) - r_i(s,\mathbf{a}) - \gamma f(s')\right]}{1-\gamma}
$$

证明。 我们注意到

$$\begin{aligned}
&\sum_{s,\mathbf{a}}\sum_{t=0}^{\infty}\gamma^{t+1}\Pr(s_t=s,\mathbf{a}_t=\mathbf{a}|s_0,\pi)\sum_{s'}\Pr(s_{t+1}=s'|s_t=s,\mathbf{a}_t=\mathbf{a})f(s') \\
=&\sum_{s,\mathbf{a}}\sum_{t=1}^{\infty}\gamma^{t}\Pr(s_t=s,\mathbf{a}_t=\mathbf{a}|s_0,\pi)f(s)
\end{aligned}$$

于是，我们有

$$\begin{aligned}
&\frac{\mathbb{E}_{s\sim d^\pi,s'\sim P(\cdot|s,\mathbf{a})}\left[f(s)-\gamma f(s')\right]}{1-\gamma} \\
=&\sum_{s,\mathbf{a}}\sum_{t=0}^{\infty}\gamma^{t}\Pr(s_t=s,\mathbf{a}_t=\mathbf{a}|s_0,\pi)f(s)-\sum_{s,\mathbf{a}}\sum_{t=1}^{\infty}\gamma^{t}\Pr(s_t=s,\mathbf{a}_t=\mathbf{a}|s_0,\pi)f(s) \\
=&\sum_{\mathbf{a}}\Pr(\mathbf{a}_0=\mathbf{a}|s_0,\pi)f(s_0)=f(s_0).
\end{aligned}$$

由于 $V_i^\pi(s_0)=\frac{\mathbb{E}_{d^\pi}[r_i(s,\mathbf{a})]}{1-\gamma}$，重新整理各项即可完成证明。 □