arXiv:2305.04819v1 [cs.LG] 8 May 2023

多智能体强化学习中的局部优化实现全局最优

Yulai Zhao* Zhuoran Yang† Zhaoran Wang‡ Jason D. Lee§

摘要

采用函数逼近的策略优化方法广泛应用于多智能体强化学习。然而，如何设计具有统计保证的此类算法仍未得到解决。借助刻画多智能体策略优化问题景观的多智能体性能差异引理，我们发现，局部动作价值函数是每个局部策略的理想下降方向。受此观察启发，我们提出了一种多智能体 PPO 算法，其中每个智能体的局部策略都以类似原始 PPO 的方式进行更新。我们证明，在马尔可夫博弈满足标准正则性条件，且问题相关量满足相应条件时，我们的算法以次线性速率收敛至全局最优策略。我们将算法扩展到离策略设定，并在策略评估中引入悲观性，这与实验结果相符。据我们所知，这是首个在合作马尔可夫博弈中具有可证明收敛性的多智能体 PPO 算法。

1 引言

近年来，多智能体强化学习（MARL）取得了许多实证成果，例如围棋等热门策略游戏 [Silver 等人， 2016], 星际争霸 II [Vinyals 等人， 2019], 以及扑克 [Brown 和 Sandholm， 2018]。与仅关注单个智能体如何最大化总奖励的基础强化学习（RL）不同，多智能体强化学习研究多个智能体如何与共享环境及其他智能体互动。

策略优化方法广泛应用于多智能体强化学习中。这些算法通常使用某个函数类对策略进行参数化，并利用策略梯度定理计算累积奖励的梯度 [Sutton 等人， 1999] 或其变体（例如自然策略梯度Kakade[2001] 和近端策略优化 [Schulman 等人， 2017]）来更新策略参数。

尽管在实证研究中取得了成功，但关于多智能体强化学习中策略优化的理论研究仍然非常有限。即使在智能体共享同一目标（即最大化总奖励函数）的合作场景下，也会面临诸多挑战 [Zhang 等人， 2021]。（1）非平稳性：一个智能体采取的每个动作都会影响总奖励和状态转移。因此，每个学习智能体都必须学会适应由其他智能体引起的环境变化。从优化的角度来看，多智能体策略优化问题的几何结构变得难以明确。由于缺乏平稳马尔可夫性质（即未来的演化仅取决于前一状态和个体动作），直接应用传统的单智能体分析方法会变得模糊。（2）可扩展性：考虑到其他智能体的存在，每个独立智能体都将面临联合动作空间，其维度呈指数级增长

*普林斯顿大学；yulaiz@princeton.edu
†耶鲁大学；zhuoran.yang@yale.edu
‡西北大学；zhaoranwang@gmail.com§普林斯顿大学；jasonlee@princeton.edu

随着智能体数量的增加，环境中的智能体过多会使多智能体强化学习的理论分析复杂化。(3) 函数逼近：与可扩展性问题密切相关，多智能体强化学习中的状态空间和联合动作空间通常极其庞大，因此函数逼近成为多智能体强化学习中不可或缺的组成部分，有助于降低计算和统计分析的难度。

本文旨在回答以下这一基本问题：

我们能否在采用函数逼近的合作场景下，设计一种具有可证明收敛性的多智能体策略优化算法？

我们对上述问题给出了肯定的回答。我们提出了一种多智能体 PPO 算法，其中每个智能体的局部策略都按照与原始 PPO 算法类似的方式进行更新，**且更新过程是顺序进行的**。 [Schulman 等人， 2017]。具体而言，我们利用多智能体性能差异引理（参见引理 4.1），并假设联合策略被分解为条件相关的策略。该引理刻画了策略优化的景观，表明使用局部动作价值函数作为每个局部策略的改进方向具有优越性。这种因子化结构从根本上规避了非平稳性和可扩展性方面的问题。为了解决大规模状态空间带来的挑战，我们采用对数线性参数化来表示每个局部策略，并提出通过 KL 散度正则化的镜像下降来更新策略参数，其中下降方向由各自独立的估计过程确定。结合上述结果，我们得到了所提出的多智能体 PPO 算法。我们证明，多智能体 PPO 算法以次线性速率收敛至全局最优策略。此外，我们将多智能体 PPO 扩展至离策略设定，在该设定中，策略使用根据数据分布 μ收集的样本进行评估。在采样分布满足覆盖性假设的条件下，我们证明了类似的理论保证。

我们总结如下贡献。

我们的贡献。 首先，通过聚焦于因子化策略，我们证明了性能差异引理的多智能体版本，表明动作价值函数是局部策略的理想下降方向。这种几何刻画有助于解决非平稳性问题，并启发我们提出多智能体近端策略优化算法。

其次，我们对策略采用对数线性函数近似。我们证明，多智能体近端策略优化以次线性 $\mathcal{O}\left(\frac{N}{1-\gamma}\sqrt{\frac{\text{对数}\,|\mathcal{A}|}{K}}\right)$ 速率收敛，误差上限为策略评估和改进过程中产生的某些统计误差，其中 K 为迭代次数， N为智能体数量， $|\mathcal{A}|$为每个智能体的动作空间。样本复杂度关于 N呈多项式增长，从而打破了可扩展性瓶颈。

第三，我们提出多智能体 PPO 算法的一种离策略变体，并在策略评估中引入悲观性。该算法也以次线性速度收敛至全局最优策略，误差上限为统计误差 $\widetilde{\mathcal{O}}(n^{-\frac{1}{3}})$。其中， n 为用于估计评论家的样本数量。[1]样本复杂度界的一个关键特点是，它只要求单策略集中性。

据我们所知，这是首个在函数逼近下、适用于合作型马尔可夫博弈且具有可证明收敛性的多智能体 PPO 算法。

组织结构。本文结构安排如下：第2节回顾相关文献。第3节正式描述问题设定并介绍必要的定义。第4节详细阐述主要的多智能体 PPO 算法。第5节将结果进一步扩展至离策略设定。最后，我们在第6节总结全文，并将证明放在附录中。

[1] $\widetilde{\mathcal{O}}(\cdot)$ 隐藏了对数因子。

2 相关工作

策略优化 许多实证研究已证明策略优化方法在博弈及其他应用中的有效性和效率 [Silver et al., 2016, 2017, Guo et al., 2016, Tian et al., 2019]。这些研究通常采用 Sutton 等人开创性的策略梯度（PG）定理[1999], ，或采用为提升传统 PG 方法实证表现而提出的多种 PG 变体，在参数化空间中更新策略参数。具体而言，Kakade [2001] 提出了自然策略梯度（NPG）算法，该算法基于 KL 散度正则化的思想，在参数空间中寻找最陡的下降方向。基于信赖域学习的算法在实践中通常被视为先进的策略优化方法 [Lillicrap et al., 2015, Duan et al., 2016], ，其更新稳定且表现优异。具体来说，TRPO [Schulman et al., 2015]和 PPO [Schulman et al., 2017] 可视为受 KL 散度约束的 NPG 变体。这些算法的一个优良特性是能够保证期望回报单调提升。

尽管取得了丰硕的实证研究成果，但缺乏凸性往往阻碍了策略优化方法相关理论的发展。记 K 和 T 分别表示迭代次数和样本数量。 Agarwal et al. [2020] 证明，对于采用函数逼近的在线自然策略梯度（NPG），其迭代复杂度为 $\mathcal{O}(K^{-\frac{1}{2}})$，样本复杂度为 $\mathcal{O}(T^{-\frac{1}{4}})$。Shaniet al. [2020] 研究了基于样本的信赖域策略优化（ TRPO ），并证明其以 $\tilde{\mathcal{O}}(T^{-\frac{1}{2}})$ 的速率收敛至全局最优解；加入正则化后，该速率可提升至 $\mathcal{O}(1/T)$。通过对原始 PPO 算法进行少量修改，Liuet al. [2019] 在策略和函数均采用神经网络参数化时，给出了以 $\mathcal{O}(K^{-\frac{1}{2}})$ 的速率收敛至全局最优解的结果。其分析的关键在于无限维镜像下降中理想的单点单调性，该性质有助于在缺乏凸性的情况下刻画策略更新。我们在分析多智能体 PPO 算法时，也利用了类似的单点性质。

多智能体强化学习 马尔可夫博弈（MG）是刻画多智能体决策过程的常用模型 [Shapley, 1953, Littman, 1994], 它可以被视为马尔可夫决策过程（MDP）的多智能体扩展。基于策略的算法可以通过函数逼近推广到大规模状态空间。近年来，人们对开发具有可证明效率的马尔可夫博弈算法越来越感兴趣 [Daskalakis 等人, 2020, Cen 等人, 2021年, Zhao 等人, 2022, Ding 等人, 2022, Cen 等人, 2022]。这些研究通常考察竞争性强化学习场景，例如零和博弈。其收敛速率通常取决于各种集中度系数，并且在最坏情况下可能无法紧致地缩放。

多智能体强化学习的策略优化 由于每个智能体面临的环境具有非平稳性，因此在多智能体强化学习场景中应用策略优化方法，比在单智能体场景中更加复杂 [Zhang 等人, 2021]。实践中常采用一种称为集中式训练与分散式执行（CTDE）的学习范式 [Kraemer 和 Banerjee, 2016, Lowe 等人, 2017年, Foerster 等人, 2018年, Yang 等人, 2018年, Wen 等人, 2019, Zhang 等人, 2020]。在集中式训练与分散式执行中，联合集中式价值函数有助于解决由其他智能体引起的非平稳性问题。训练期间，每个智能体都能访问全局状态和其他智能体的动作，因此可以单独调整其策略参数。例如，Lowe 等人 [2017]提出了一种多智能体策略梯度算法，其中智能体根据所有智能体的观测和动作学习一个集中式评论家。

信赖域学习 [Schulman 等人, 2015] 近期已与集中式训练与分散式执行范式相结合，以确保单调改进。具体而言，IPPO [de Witt 等人,

2020] 和 MAPPO [Yu 等人， 2021]展示了基于近端策略优化的方法在合作场景中的出色表现。这些方法在实践中的有效性通常受到同质性假设的限制，即智能体共享相同的动作空间和策略参数。从理论上讲，为多智能体强化学习中的策略优化算法提供统计保证，比单智能体场景更加复杂 [Zhang 等人， 2021]。在马尔可夫博弈中，每个智能体所面对的非平稳环境，使得单智能体收敛性分析无法直接应用。Kuba 等人近期提出了多智能体强化学习中的首批信赖域学习算法；在不假设智能体同质、也不采用价值函数分解规则的情况下，这些算法仍具有单调改进保证。他们得出这些结果的关键观察，是多智能体优势函数分解规则能够构建顺序策略更新结构。然而，他们并未给出收敛速率。本文利用这种条件依赖结构，为完全合作马尔可夫博弈设计了一种新的、可证明收敛的近端策略优化算法；该算法以次线性速率收敛至全局最优策略。

基于悲观性的强化学习方法尽管函数逼近能够处理较大的状态/动作空间，但它也有自身的缺点。使用函数逼近器时出现的一个重要问题是，价值函数通常会产生正偏差Thrun 和 Schwartz [1993]。学习者可能无法获得准确的评估。许多实证研究利用悲观性原则来修正此类高估 [Fujimoto 等人，2018年，Laskin 等人，2020，Lee 等人，2020，Moskovitz 等人， 2021]。例如，为了减少函数逼近带来的评估偏差，Fujimoto 等人 [2018] 通过从两个价值估计中取较小值，构造 Bellman 目标，将其作为直观的估计下界。他们的方法对价值函数采取了悲观的观点。

在理论方面，离线强化学习领域日益增多的研究也聚焦于悲观性，以应对数据集缺乏数据覆盖的问题 [Liu et al., 2020, Jin et al., 2021, Uehara and Sun, 2021, Rashidinejad et al., 2021, Zhan et al., 2022]。从技术角度来看，这些研究旨在最大化训练后的智能体能够获得的最坏情况奖励。与依赖数据集覆盖假设的方法不同 [Munos, 2003, Munos and Szepesv´ari, 2008, Chen 和 Jiang, 2019], 这些方法提供了依赖于数据集的性能界限，从而为缺乏探索的数据集提供了稳健的结果，而传统方法无法应用于此类数据集。我们在第5节中重点研究离策略设定，并利用贝尔曼一致悲观主义 [Xie et al., 2021]。我们通过假设存在一个采样预言机，提供用于近似动作价值函数的奖励和状态转移估计，给出了线性函数近似下的具体界限。

3 预备知识

在本节中，我们介绍必要的符号、问题设定，以及一些将在本文中频繁使用的有用量。

3.1 设置与符号约定

设置 我们考虑一个完全合作的 马尔可夫博弈 [Shapley， 1953， Littman， 1994],其定义为一个元组（$\mathcal{N}$， $\mathcal{S}$， $\boldsymbol{\mathcal{A}}$ $\mathcal{P}$， r， γ）。其中， $\mathcal{N} = \{1, \ldots, N\}$ 表示智能体集合， $\mathcal{S}$ 表示有限状态空间， $\boldsymbol{\mathcal{A}} = \mathcal{A}^N$ 表示所有智能体的有限动作空间的乘积（即联合动作空间）， $\mathcal{P}$: $\mathcal{S} \times \boldsymbol{\mathcal{A}} \times \mathcal{S} \to [0, 1]$ 决定状态转移

方案、奖励函数 $r:\ \mathcal{S}\times\mathsf{A}\ \to[0,1]$，以及 $\gamma\in[0,1)$ 是折扣因子。[2]智能体按照以下协议与环境交互：在时间步 t，智能体处于状态 $s_t\in\mathcal{S}$；每个智能体 i 根据其策略 $\pi^i(\cdot|s_t)$采取动作 $a_t^i\in\mathcal{A}$，该动作与其他智能体的动作共同构成联合动作 $\mathbf{a}_t=(a_t^1,\ .....,\ a_t^N)\in\mathsf{A}$，该联合动作根据联合策略 $\boldsymbol{\pi}(\cdot|s_t)=\prod_{i=1}^N\pi^i(\cdot|s_t)$抽取；智能体获得联合奖励 $r_t=r(s_t,\ \mathbf{a}_t)\in\mathbb{R}$，并移动到 $s_{t+1}\sim\mathcal{P}(\cdot|s_t,\ \mathbf{a}_t)$。给定联合策略$\boldsymbol{\pi}$、状态转移概率函数$\mathcal{P}$和初始状态分布 ρ，我们将折扣占用状态-动作分布定义为

$$d_{\boldsymbol{\pi}}(s,\mathbf{a})=(1-\gamma)\mathbb{E}\sum_{t=0}^{\infty}\Pr^{\boldsymbol{\pi}}(s_t=s,a_t=\mathbf{a}|s_0\sim\rho).$$

标准价值函数和动作价值函数定义为

$$V_{\boldsymbol{\pi}}(s)\triangleq\mathop{\mathbb{E}}_{\mathbf{a}_{0:\infty}\sim\boldsymbol{\pi},s_{1:\infty}\sim\mathcal{P}}\left[\sum_{t=0}^{\infty}\gamma^t r_t\Big|\ s_0=s\right],$$

$$Q_{\boldsymbol{\pi}}(s,\boldsymbol{a})\triangleq\mathop{\mathbb{E}}_{s_{1:\infty}\sim\mathcal{P},\mathbf{a}_{1:\infty}\sim\boldsymbol{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t r_t\Big|\ s_0=s,\ \mathbf{a}_0=\boldsymbol{a}\right].$$

考虑所有智能体的标准优势函数表示为 $A_{\boldsymbol{\pi}}(s,\boldsymbol{a})\ \triangleq Q_{\boldsymbol{\pi}}(s,\boldsymbol{a})-V_{\boldsymbol{\pi}}(s)$。稍后，我们将介绍针对特定智能体的优势函数。

令 $\nu_{\boldsymbol{\pi}}(s)$ 和 $\sigma_{\boldsymbol{\pi}}(s,\mathbf{a})=\boldsymbol{\pi}(\mathbf{a}|s)\cdot\nu_{\boldsymbol{\pi}}(s)$ 分别表示与联合策略 $\boldsymbol{\pi}$ 相关的平稳状态分布和平稳状态-动作分布。将底层最优策略定义为 $\boldsymbol{\pi}_*$。为简便起见，本文使用 ν_* 和 σ_* 分别表示 $\nu_{\boldsymbol{\pi}_*}$ 和$\sigma_{\boldsymbol{\pi}_*}$。

在本文中，我们密切关注不同智能体子集对整个团队表现的贡献。在继续介绍多智能体定义之前，我们先引入以下多智能体记号。

记号在本文中，我们用从 1 到 N 的整数为 N 个智能体编号，并用集合 $\mathcal{N}=\{i|i=1,\cdots,N\}$ 表示所有智能体。我们用 $m\in\mathcal{N}$ 表示第 m 个特定智能体。特别地，某个项上标中的集合记号表示与该集合中智能体相关的量。例如，$\mathbf{a}^{\{1,2,3\}}$ 表示智能体 1、2 和 3 的联合动作。指代第 k 个特定智能体时，我们可能会在上标中写出索引 k。不带任何上标的粗体符号（例如 $\mathbf{a}$）表示所有智能体。为简洁起见，令 $(m:\ m')$ 表示集合 $\{i|m\le i\le m',i\in\mathcal{N}\}$ 的简写。例如，$\boldsymbol{\pi}^{1:m}(\cdot|s)$ 表示考虑智能体 1、$2\cdots$、m 的联合策略。

现在，我们介绍多智能体动作价值函数和优势函数，用于刻画特定子智能体的贡献。

定义 3.1. 设 P是 $\mathcal{N}$的一个子集。与 P 中的智能体相关联的多智能体动作价值函数为

$$Q_{\boldsymbol{\pi}}^P\left(s,\mathbf{a}^P\right)\triangleq\mathbb{E}_{\tilde{\mathbf{a}}\sim\tilde{\boldsymbol{\pi}}}\left[Q_{\boldsymbol{\pi}}\left(s,\mathbf{a}^P,\tilde{\mathbf{a}}\right)\right],$$

这里，我们在符号上方加波浪线来表示补集智能体，即 $\tilde{\mathbf{a}}=\{a^i|i\notin P,i\in\mathcal{N}\}$。

设 $P,P'\subseteq\mathcal{N}$是智能体的两个不相交子集。多智能体优势函数定义如下。它本质上衡量在设置智能体 $\mathbf{a}^{P'}$ 的基础上设置智能体 $\mathbf{a}^P$所带来的改进，同时所有其他智能体遵循 $\boldsymbol{\pi}$。

$$A_{\boldsymbol{\pi}}^{P'}\left(s,\mathbf{a}^P,\mathbf{a}^{P'}\right)\triangleq Q_{\boldsymbol{\pi}}^{P\cup P'}\left(s,\mathbf{a}^P,\mathbf{a}^{P'}\right)-Q_{\boldsymbol{\pi}}^P\left(s,\mathbf{a}^P\right).$$

[2]为清晰起见，我们假设 N 智能体共享同一组动作。将我们的结果推广到动作集合不同的情形是很直接的。请参见第4节。

多智能体贝尔曼算子通过推广经典版本来定义。

定义 3.2。对于 $m \in \mathcal{N}$ 和任意函数 $f: \mathcal{S} \times \mathcal{A}^m \to \mathbb{R}$，我们定义多智能体贝尔曼算子 $\mathcal{T}_{\boldsymbol{\pi}}^{1:m}: \mathbb{R}^{\mathcal{S}\times\mathcal{A}^m} \mapsto \mathbb{R}^{\mathcal{S}\times\mathcal{A}^m}$ 为

$$\mathcal{T}_{\boldsymbol{\pi}}^{1:m} f(s, \mathbf{a}^{1:m}) := \mathop{\mathbb{E}}_{\tilde{\mathbf{a}}\sim\tilde{\boldsymbol{\pi}}} r(s, \mathbf{a}^{1:m}, \tilde{\mathbf{a}}) + \gamma \mathop{\mathbb{E}}_{\substack{\tilde{\mathbf{a}}\sim\tilde{\boldsymbol{\pi}} \\ s'\sim\mathcal{P}(\cdot|s,\mathbf{a}^{1:m},\tilde{\mathbf{a}})}} f(s', \boldsymbol{\pi}^{1:m})$$

其中，$f(s', \boldsymbol{\pi}^{1:m})$ 是 $\mathbb{E}_{\mathbf{a}'\sim\boldsymbol{\pi}^{1:m}(\cdot|s')} f(s', \mathbf{a}')$ 的简写。

不难看出，$Q_{\boldsymbol{\pi}}^{1:m}$ 是 $\mathcal{T}_{\boldsymbol{\pi}}^{1:m}$ 的唯一不动点，它对应于经典的单智能体贝尔曼算子。

3.2 KL散度正则化镜像下降

我们回顾可证明的单智能体近端策略优化算法中的镜像下降表述 [Liu 等人， 2019]。在第 k次迭代中，策略参数 θ 通过以下方式更新

$$\theta_{k+1} \leftarrow \arg\max_{\theta} \hat{\mathbb{E}}\big[\langle A_k(s,\cdot), \pi_\theta(\cdot|s)\rangle - \beta_k KL\left(\pi_\theta(\cdot\|s)\|\pi_{\theta_k}(\cdot|s)\right)\big]. \tag{1}$$

下文我们将用 $\langle\cdot,\cdot\rangle$ 表示 $\mathcal{A}$上的内积。期望ˆ取自 $\mathbb{E}$，它是平稳状态-动作分布 $\nu_{\pi_{\theta_k}}$的经验估计，而 A_k 是优势函数 $A^{\pi_{\theta_k}}$的估计。

采用 KL散度，(1) 与 NPG [Kakade、 2001] 更新密切相关。作为一种变体，这种表述与原始 PPO [Schulman 等人，2017]：此处使用 $KL(\pi_\theta(\cdot\|s)\|\pi_{\theta_k}(\cdot|s))$，而非 $KL(\pi_{\theta_k}(\cdot\|s)\|\pi_\theta(\cdot|s))$。这种变化对于给出可证明的保证至关重要，下一节将对此进行说明。

4 多智能体近端策略优化

回顾一下，ν_* 是 $\boldsymbol{\pi}_*$的平稳状态分布。本节旨在最大化分布 ν_*下的期望价值函数：$J(\boldsymbol{\pi}) \triangleq \mathbb{E}_{s\sim\nu_*} V_{\boldsymbol{\pi}}(s)$。

本文旨在提出一种信赖域学习多智能体算法，并为其提供严格的收敛理论。正如我们所提到的，合作型多智能体强化学习中的策略优化颇具挑战，因为策略优化问题会变成涉及所有智能体的联合优化问题。目前尚不清楚：（a）联合策略对应的总奖励多变量函数呈现怎样的景观；（b）每个智能体适合采用哪些策略下降方向。我们通过利用下文所述的多智能体强化学习中性能差异引理的串行分解，提出了一种刻画该景观的方法。

引理 4.1。对于任意联合策略 $\boldsymbol{\pi}$，我们有

$$J(\boldsymbol{\pi}_*) - J(\boldsymbol{\pi}) = \frac{1}{1-\gamma} \sum_{m=1}^{N} \mathop{\mathbb{E}}_{\substack{s\sim\nu_* \\ \mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*^{1:m-1}}} \left\langle Q_{\boldsymbol{\pi}}^{1:m}(s, \mathbf{a}^{1:m-1}, \cdot), \pi_*^m(\cdot|s, \mathbf{a}^{1:m-1}) - \pi^m(\cdot|s, \mathbf{a}^{1:m-1}) \right\rangle$$

其中内积是关于 $a^m \in \mathcal{A}$ 的。

借助这一几何刻画，我们可以证明，使用 $Q_{\boldsymbol{\pi}}^m(s, \mathbf{a}^{1:m-1}, a^m)$ 作为下降方向，并对**每个** 智能体$m \in \mathcal{N}$ 执行 KL 散度正则化的镜像下降，可以获得更高的**总** 奖励，从而实现串行优化过程。下面我们将详细介绍该算法。

为了表示条件策略，我们采用对数线性参数化。

参数化对于第 m 个智能体（$m \in \mathcal{N}$），其条件策略取决于所有先前排序的智能体 $\mathbf{a}^{1:m-1}$。给定系数向量 $\theta^m \in \Theta$，其中 $\Theta = \{\|\theta\| \leq R | \theta \in \mathbb{R}^d\}$ 是一个凸的、受范数约束的集合。状态s 下选择动作 a^m 的概率为

$$\pi_{\theta^m}(a^m|s, \mathbf{a}^{1:m-1}) = \frac{\exp\left(\phi^\top(s, \mathbf{a}^{1:m-1}, a^m)\theta^m\right)}{\sum_{a^m \in \mathcal{A}} \exp\left(\phi^\top(s, \mathbf{a}^{1:m-1}, a^m)\theta^m\right)} \tag{2}$$

其中 ϕ 是一组特征向量表示。不失一般性，我们施加一个正则性条件，使得每个 $\|\phi\|_2 \leq 1$。这种参数化已广泛应用于强化学习文献中 [Branavan 等人，2009，Gimpel 和 Smith，2010，Heess 等人，2013，Agarwal et al.，2020，Zhao 等人，2022]。[3]

4.1 策略改进与评估

在第 k 次迭代中，我们有当前策略 $\boldsymbol{\pi}_{\theta_k}$，需要：（1）进行**策略评估**，以获得动作价值函数估计值 $Q_{\boldsymbol{\pi}_{\theta_k}}$，用于确定 $\boldsymbol{\pi}_{\theta_k}$ 的质量。（2）进行**策略改进**，将策略更新为 $\boldsymbol{\pi}_{\theta_{k+1}}$。

为简化符号表示，我们使用 ν_k 和 σ_k 分别表示由$\boldsymbol{\pi}_{\theta_k}$诱导的平稳状态分布$\nu_{\boldsymbol{\pi}_\theta^k}$和平稳状态-动作分布$\sigma_{\boldsymbol{\pi}_\theta^k}$。

策略改进在第 k 次迭代中，我们将基于 $\hat{Q}^{1:m}_{\boldsymbol{\pi}_{\theta_k}}$（针对智能体 $m \in \mathcal{N}$）的理想更新定义为 $\hat{\pi}^m_{k+1}$，其中 $\hat{Q}^{1:m}_{\boldsymbol{\pi}_{\theta_k}}$ 是 $Q^{1:m}_{\boldsymbol{\pi}_{\theta_k}}$ 的估计量。理想更新通过以下更新得到

$$\hat{\pi}^m_{k+1} \leftarrow \arg\max_{\pi^m} \hat{F}(\pi^m) \tag{3}$$

$$\hat{F}(\pi^m) = \mathop{\mathbb{E}}_{\sigma_k}\left[\langle \hat{Q}^{1:m}_{\boldsymbol{\pi}_{\theta_k}}(s, \mathbf{a}^{1:m-1}, \cdot), \pi^m(\cdot|s, \mathbf{a}^{1:m-1})\rangle - \beta_k KL\left(\pi^m(\cdot|s, \mathbf{a}^{1:m-1}) \| \pi_{\theta^m_k}(\cdot|s, \mathbf{a}^{1:m-1})\right)\right]$$

其中，θ^m_k 是第 m 个智能体当前条件策略的参数。在上述公式中，分布是关于 $(s, \mathbf{a}^{1:m-1}) \sim \nu_k \boldsymbol{\pi}^{1:m-1}_{\theta_k}$ 取的，为简洁起见，我们记作 σ_k。在对数线性参数化下：$\pi_{\theta^m_k} \propto \exp\{\phi^\top \theta^m_k\}$，我们得到如下理想策略更新的闭式解。

命题 4.2。给定一个估计量 $\hat{Q}^{1:m}_{\boldsymbol{\pi}_{\theta_k}}$，经 *KL*散度 正则化的更新 (3) 具有如下显式解

$$\hat{\pi}^m_{k+1}(\cdot|s, \mathbf{a}^{1:m-1}) \propto \exp\left\{\beta_k^{-1} \hat{Q}^{1:m}_{\boldsymbol{\pi}_{\theta_k}}(s, \mathbf{a}^{1:m-1}, \cdot) + \phi^\top(s, \mathbf{a}^{1:m-1}, \cdot)\theta^m_k\right\}.$$

通过将约束 $\sum_{a^m \in \mathcal{A}} \pi^m(\cdot) = 1$ 作为拉格朗日乘子添加到 $F(\pi^m$，即可轻松证明。详情见附录 B。

为使用参数化的 $\pi_{\theta^m_{k+1}} \propto \exp\{\phi^\top \theta^m_{k+1}\}$ 近似理想的 $\hat{\pi}^m_{k+1}$，我们将以下均方误差（MSE）作为一个子问题进行最小化

$$\theta^m_{k+1} \leftarrow \arg\min_{\theta^m \in \Theta} L(\theta^m) \tag{4}$$

其中，$L(\theta^m)$ 定义为

$$L(\theta^m) = \mathop{\mathbb{E}}_{\sigma_k}\left((\theta^m - \theta^m_k)^\top \phi(s, \mathbf{a}^{1:m-1}, a^m) - \frac{\hat{Q}^{1:m}_{\boldsymbol{\pi}_{\theta_k}}(s, \mathbf{a}^{1:m-1}, a^m)}{\beta_k}\right)^2$$

[3]为便于说明，我们假设所有参与者共享同一参数集。要将我们的结果扩展到 N 智能体具有不同能力的情形，只需对分析进行少量修改。具体而言，我们只需分别处理每个智能体的更新范数界（R）、特征上的正则性条件以及 β。

直观而言，较小的 $L(\theta)$ 表明 π_{θ^m} 接近理想更新 $\hat{}\pi_{k+1}^m$。此外，如果 $\hat{}\pi_{k+1}^m$ 恰好属于对数线性函数类，即存在一个 $\vartheta \in \Theta$ 使得 $\hat{}\pi_{k+1}^m \propto \exp\{\phi^\top \vartheta\}$。那么我们有 $L(\vartheta) = 0$。

为解决均方误差（MSE）最小化问题（4），我们采用经典的随机梯度下降（SGD）更新方法。设步长为 η，在每一步 $t = 0, 1, \cdots, T-1$，参数 θ 通过以下方式进行更新：

$$\theta(t+\frac{1}{2}) \leftarrow \theta(t) - 2\eta\phi\left((\theta(t) - \theta_k^m)^\top \phi - \beta_k^{-1}\hat{Q}^{1:m}_{\boldsymbol{\pi}_{\theta_k}})\right)$$
$$\theta(t+1) \leftarrow \Pi_\Theta \theta(t+\frac{1}{2})$$

为简洁起见，我们省略了从 s、$\mathbf{a}^{1:m-1}$、a^m）中采样的内容。详细求解器请参见算法 3。

策略评估 在这一步中，我们旨在检验所获得策略的质量。因此，需要一个 Q-函数估计器。我们作出如下假设。

ˆ假设 4.3。假设我们可以访问一个 Q 函数的估计器，该估计器返回 Q。返回的 ˆ Q 对于第 k 次迭代时所有 $m \in \mathcal{N}$ 均满足以下条件

$$\left[\mathop{\mathbb{E}}_{\sigma_k}\left(\hat{Q}^{1:m}_{\boldsymbol{\pi}_{\theta_k}}(s, \mathbf{a}^{1:m-1}, a^m) - Q^{1:m}_{\boldsymbol{\pi}_{\theta_k}}(s, \mathbf{a}^{1:m-1}, a^m)\right)^2\right]^{1/2} \leq \xi_k^m.$$

我们还对估计器提出了一个正则性条件：存在一个正常数 B，使得对于任意 $m \in \mathcal{N}$ 和 $(s, \mathbf{a}^{1:m-1}, a^m) \in \mathcal{S} \times \mathcal{A}^{m-1} \times \mathcal{A}$，

$$\left|\hat{Q}^{1:m}_{\boldsymbol{\pi}_{\theta_k}}(s, \mathbf{a}^{1:m-1}, a^m)\right| \leq B.$$

在强化学习实践中，这类估计器通常由深度神经网络（DNN）构建 [Mnih 等人, 2015]。尽管近期已有研究关注将 DNN 作为函数逼近器时的理论保证 [Fan 等人, 2020]，但为确保算法具有普适性，我们假设可以使用这类估计器。我们指出，诸如通过展开轨迹来进行估计的 episodic sampling oracle [Agarwal et al., 2020] 或神经网络 [Mnih 等人，2015年，Liu 等人, 2019] 都可以作为可行的选择。作为一种推广，我们在第 5 节引入一种特定的价值函数逼近设定，其中假设所有 Q-函数都属于线性类 $\mathcal{F}$。为更好地进行探索，我们还采用悲观性原则。

算法 有了用于策略改进的子问题求解器和 Q-函数估计器，我们便可以给出具有可证明保证的多智能体 PPO 算法。伪代码见算法 1。该算法运行 K 次迭代。在第 k 次迭代中，我们通过估计器（参见假设 4.3）估计每个智能体的 Q-函数 $m \in \mathcal{N}$，以衡量 $\boldsymbol{\pi}_{\theta_k}$ 的质量。这些估计还将作为策略改进的理想下降方向。由于我们采用的是受约束的参数化策略类，因此通过求解均方误差（MSE）问题（4），找到最佳策略参数 $\theta \in \Theta$，以近似理想更新；该问题通过 SGD 运行 T 次迭代（参见算法 3）。得益于几何刻画（参见引理 4.1），依次更新每个智能体即可保证总奖励得到全局改进。

4.2 理论分析

我们的分析依赖于问题相关的量。我们将状态空间 $\mathcal{X}$ 上函数 f 的加权 L_p-范数记为 $\|f\|_{p,\ \rho} = \left(\sum_{x\in\mathcal{X}} \rho(x)|f(x)|^p\right)^{\frac{1}{p}}$

算法 1多智能体近端策略优化

输入： 马尔可夫博弈（$\mathcal{N}$，$\mathcal{S}$，$\boldsymbol{\mathcal{A}}$ $\mathcal{P}$，r，γ），惩罚参数 β，子问题的步长 η SGD 迭代次数 T ，迭代次数 K。

输出： 均匀地从 0，1，$\cdots K-1$ 中采样 k，返回 $\bar{\boldsymbol{\pi}} = \boldsymbol{\pi}_{\theta_k}$。

为每个 $m \in \mathcal{N}$ 初始化 $\theta_0^m = 0$。

</style}对 $k = 0, 1, \ldots, K-1$ **执行**

  设置参数 $\beta_k \leftarrow \beta\sqrt{K}$

  对 $m = 1, \cdots, N$ **do**

    从 $\sigma_k = \nu_k \boldsymbol{\pi}_{\theta_k}$采样 $\{s_t, \mathbf{a}_t^{1:m-1}, a_t^m\}_{t=0}^{T-1}$。

    获取 $\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s, \mathbf{a}^{1:m-1}, a^m)$ 用于每个样本。

    将样本输入算法3，获得 θ_{k+1}^m。

  结束循环

结束循环

定义 4.4. 在第 k 次迭代时，对于 $m \in \mathcal{N}$，我们使用拉东–尼科迪姆导数定义以下问题相关的量

$$\phi_k^m = \left\| \frac{d(\nu_* \boldsymbol{\pi}_*^{1:m})}{d(\nu_k \boldsymbol{\pi}_{\theta_k}^{1:m})} \right\|_{2,\sigma_k}$$

这些条件是著名的集中性系数 [Munos，2003，Farahmand 等人，2010，Chen 和 Jiang，2019]，适用于因子化策略。不过，我们的条件在结构上更简单、限制也更弱，因为它们仅涉及平稳状态-动作分布之间的密度比，无需展开轨迹。

现在，我们可以给出刻画全局收敛速度的主定理。

定理 4.5。在假设 *4.3* 下，对于完全合作马尔可夫博弈中由算法 $\bar{\boldsymbol{\pi}}$ *1*得到的输出策略，令 $\eta = \frac{R}{G\sqrt{T}}$ 且

$$\beta = \sqrt{\frac{NB^2/2}{N \log |\mathcal{A}| + \sum_{m=1}^{N} \sum_{k=0}^{K-1} (\Delta_k^m + \delta_k^m)}}.$$

经过 K次迭代后，我们有 $J(\boldsymbol{\pi}_*) - J(\bar{\boldsymbol{\pi}})$ 的上界为

$$\mathcal{O}\left(\frac{B\sqrt{N}}{1-\gamma} \sqrt{\frac{N \log |\mathcal{A}| + \sum_{m=1}^{N} \sum_{k=0}^{K-1} (\Delta_k^m + \delta_k^m)}{K}} \right)$$

其中 $\Delta_k^m = \sqrt{2}(\phi_k^m + \phi_k^{m-1}) \cdot \left(\epsilon_k^m + \frac{\xi_k^m}{\beta_k} \right)$ 且 $\delta_k^m = 2\phi_k^{m-1} \epsilon_k^m$。这里，$\epsilon_k^m$是一次近端策略优化迭代的统计误差：对于智能体 $m \in \mathcal{N}$，

$$\mathbb{E}_{\sigma_k}\left((\theta_{k+1}^m - \theta_k^m)^\top \phi - \beta_k^{-1} \hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m} \right)^2 \leq (\epsilon_k^m)^2$$

为简洁起见，我们省略 $(s, \mathbf{a}^{1:m-1}, a^m)$。

令 ϵ_{approx} 表示我们采用的对数线性策略类的近似能力，则$\epsilon_k^m = \epsilon_{approx} + \mathcal{O}(T^{-\frac{1}{4}})$。

定理 4.5明确刻画了输出 $\bar{\boldsymbol{\pi}}$ 的性能与迭代次数和迭代误差之间的关系。当 PPO 更新是理想的，即，

观察 δ'_k， Δ_k^m对于任意 $m \in \mathcal{N}$，并且 $k < K$，收敛速率简化为 $\mathcal{O}\left(\frac{NB}{1-\gamma}\sqrt{\frac{\log|\mathcal{A}|}{K}}\right)$。对迭代 K的依赖为 $\mathcal{O}(K^{-\frac{1}{2}})$，与基于样本的单智能体自然策略梯度分析具有相同的收敛速率 [Agarwal etal.，2020，Liuetal.， 2019]。

定理 4.5 的证明还需要以下部分：文献 [Liu 等人中使用的镜像下降更新分析、2019] 以及引入智能体之间序列依赖结构的引理 4.1 。完整证明见附录 B。

4.3 与独立学习的比较

在多智能体强化学习中，独立学习是指一类独立训练多个智能体的算法。在这些方法中，每个智能体都有自己的策略函数，将该智能体的观测映射为其动作。策略通过策略梯度方法以去中心化方式进行优化，不进行显式通信或协调，也不对其他智能体的行为进行显式建模。由于具有较强的性能和效率，独立学习方法在多智能体强化学习中得到了广泛应用。

在本小节中，我们将详细比较我们的算法与以往关于独立学习的研究结果（包括实验和理论结果）。我们还进行了仿真研究，以展示我们的序列策略更新结构相较于朴素的独立策略梯度更新的优越性。

实验一些实证尝试表明，独立策略梯度学习在多智能体强化学习中能够取得出乎意料的优异表现，例如 MAPPO [Yu 等人，2021], IPPO [de Witt 等人， 2020],以及 [Papoudakis 等人， 2021]。

尽管在实证上取得了成功，这些方法仍存在若干缺点。IPPO 和 MAPPO 假设同质性（智能体共享相同的动作空间和策略参数）。因此，需要进行参数共享。尽管可以关闭参数共享，但它们仍然无法保证单调改进，尽管被称为基于 PPO 的算法。回顾一下，原始 TRPO 的主要优点 [Schulman 等人， 2015] 是单调性。此外，这些方法也不具备任何收敛保证。当关闭参数共享时，收敛问题会变得更加严重。Kuba 等人 [Kuba 等人，2022，命题 1] 表明，参数共享可能导致指数级更差的次优结果。

得益于序列式智能体结构以及新颖的多智能体镜像下降分析，我们提出了首个以次线性速率收敛的多智能体强化学习算法。请注意，我们的结果既不依赖于智能体的同质性，也不依赖于价值函数分解规则。

理论 已有多项理论研究从数学角度探讨了多智能体强化学习中独立策略优化算法收敛至纳什均衡（NE）策略的收敛保证 [Daskalakis 等人，2020，Leonardos 等人，2022，Fox 等人，2022，Ding 等人， 2022]。具体而言，Daskalakis 等人 [2020] 研究了竞争性强化学习。其他研究则探讨了在马尔可夫势博弈（完全合作博弈的一种扩展）中收敛至纳什均衡策略。然而，我们认为，纳什均衡策略在价值函数方面未必是最优的。

与他们的工作不同，我们提出了首个具有理论保证的多智能体策略优化算法，该算法能够找到价值函数接近全局最优的策略，且收敛速率为次线性。

仿真 为进一步验证我们算法在理论和实验上的优势，我们进行了数值仿真，以展示该算法的优越性

与朴素的独立策略梯度更新相比，采用顺序更新结构。我们考虑冯·诺依曼比率博弈，这是一种简单的随机博弈，Daskalakis 等人也曾使用过Daskalakis 等人 [2020]。仿真结果表明，与我们的算法不同，独立学习方法很难摆脱驻点。此外，在最大化价值函数方面，我们的算法始终优于独立学习。详细的设置和结果见第 E节。

5 悲观多智能体近端策略优化与线性函数近似

在本节中，我们研究离策略设定，使用来自数据分布的样本μ 来评估 $Q_{\boldsymbol{\pi}}$。在实验中，函数逼近器通常会导致价值函数出现正偏差Thrun 和 Schwartz [1993], ，因此许多深度离策略行动者-评论家算法会引入悲观性，以减少此类高估 [Fujimoto 等人，2018年，Laskin 等人， 2020]。在此设定下，我们也采用悲观策略评估，与实验研究保持一致。

我们关注价值函数和策略采用线性参数化的设定。我们的结果可以推广到一般函数逼近设定，详见附录D。

定义 5.1 (线性函数近似)。令 ϕ 为一组条件构造的特征映射，其定义与第4节相同。将动作价值函数类定义为$\mathcal{F}^m = \{\phi^\top\omega\colon\ \omega \in \mathbb{R}^d,\ \|\omega\|_2 \le L,\ \phi^\top\omega \in [0, 1/1-\gamma]\}$。策略类仍采用对数线性参数化：$\Pi^m = \{\pi \propto \exp(\phi^\top\theta)\colon\ \theta \in \mathbb{R}^d,\ \|\theta\|_2 \le R\}$ (参见第4节)。

注 5.2。根据该定义，对于任意 $m \in \mathcal{N}$ 和策略 $\boldsymbol{\pi}$，都必须存在满足以下条件的参数 $\omega \in \mathbb{R}^d$

$$Q_{\boldsymbol{\pi}}^{1:m}(s, \mathbf{a}^{1:m}) = \phi(s, \mathbf{a}^{1:m})^\top\omega$$

在本节中，我们将初始状态固定为某个 s_0。因此，我们旨在最大化的期望奖励定义为

$$J(\boldsymbol{\pi}) \triangleq V_{\boldsymbol{\pi}}(s_0).$$

请注意，在单智能体离线强化学习中，只有一个策略会影响特定状态下的动作，因此我们可以利用由状态、动作、奖励和状态转移组成的离线数据集来衡量价值函数估计的质量$\mathcal{D}$ 。直观地说，当以下L_0 趋近于 0 时，我们可以说 f 是 Q-函数 [Xie 等人的一个良好近似器， 2021]。

$$L_0 = \frac{1}{n}\sum_{(s,a,r,s')\sim\mathcal{D}}\big(f(s,a) - r - \gamma f(s', \pi)\big)^2$$

其中， $f(s', \pi)$ 是 $\sum_{a'} f(s', a')\pi(a'|s')$ 的简写形式，本节中将频繁使用该记号。

然而，在多智能体环境中，复杂的依赖结构使得此类离线数据集无法应用。具体而言，对于第 m 个智能体及策略 $\boldsymbol{\pi}$，估计多智能体价值函数 $Q_{\boldsymbol{\pi}}^{1:m}$要求所有不在 $\{1$ 中的智能体： $m\}$都必须遵循 $\boldsymbol{\pi}$ （参见定义 3.1），而离线数据集无法保证这一点。

因此，在我们研究的多智能体设定中，在线交互是不可避免的。下面我们将对样本生成协议作出说明。

我们将从固定的数据分布 $\mu = \mu_s\mu_a \in \Delta(\mathcal{S} \times \mathcal{A})$ 中收集状态-动作样本。在良性情形下，良好的覆盖 μ 可确保在整个状态空间和动作空间中进行充分探索。假设我们可以使用标准的强化学习预言机

定义 5.3（采样预言机）。 该预言机可以从 $s \sim \mu_s$出发，采取任意动作$\mathbf{a} \in \boldsymbol{\mathcal{A}}$，并获得下一个状态 $s' \sim \mathcal{P}(\cdot|s,\mathbf{a})$ 以及奖励 $r(s,\mathbf{a})$。

我们的查询预言机与经典的 **在线采样预言机**（用于马尔可夫决策过程）[Kakadeand Langford，2002，Du et al.，2019，Agarwalet al.，2020]。区别在于，我们只进行一步状态转移，而经典在线模型通常会在每个回合结束时终止。我们还注意到，我们的预言机弱于 **生成模型** [Kearns and Singh，2002，Kakade，2003，Sidford et al.，2018年，Li et al.，2020]，后者假设智能体可以转移到 **任意** 状态，因此大大降低了对显式探索的需求。而我们的预言机则从一个固定的 μ_s开始。[4]

我们通过以下步骤利用采样器，获得在多智能体贝尔曼算子下误差较小的动作价值函数（参见定义 5.1）。对于智能体 $m \in \mathcal{N}$ 和 $\boldsymbol{\pi}$，**(1)** 获得 $s \sim \mu_s$；**(2)** 获得 **一个** $\sim \mu_a$ 和 **一个**$' \sim \boldsymbol{\pi}^{m+1:N}(\cdot|s)$；**(3)** 将（$\mathbf{a}^{1:m}$，$\mathbf{a}'$）作为联合动作来查询预言机，其中 $\mathbf{a}^{1:m}$ 表示 $\{1:\ m\}$的子集 $\mathbf{a}$。预言机返回（r，s'），并保证满足：

$$r \sim \mathop{\mathbb{E}}_{\tilde{\mathbf{a}}\sim\boldsymbol{\pi}^{m+1:N}} R(s,\mathbf{a}^{1:m},\tilde{\mathbf{a}}), \quad s' \sim \mathop{\mathbb{E}}_{\tilde{\mathbf{a}}\sim\boldsymbol{\pi}^{m+1:N}} \mathcal{P}(\cdot|s,\mathbf{a}^{1:m},\tilde{\mathbf{a}}).$$

重复这些步骤 n 次。这共同构成数据集 $\mathcal{D}^m = \{(s_i,\mathbf{a}_i^{1:m},r_i,s_i')|i=1,2,\cdots n\}$。定义

$$L^{1:m}(f',f,\boldsymbol{\pi}) \coloneqq \frac{1}{n}\sum_{\mathcal{D}^m}\left(f'(s,\mathbf{a}^{1:m}) - r - \gamma f(s',\boldsymbol{\pi}^{1:m})\right)^2$$

其中 $f \in \mathcal{F}^m$（参见定义 5.1），求和遍历 n 个四元组（s，$\mathbf{a}^{1:m}$，r，s'）。

我们需要以下贝尔曼误差来评估 f 的质量。

$$\mathcal{E}^{1:m}(f,\boldsymbol{\pi}) = L^{1:m}(f,f,\boldsymbol{\pi}) - \min_{f'\in\mathcal{F}^m} L^{1:m}(f',f,\boldsymbol{\pi}). \tag{5}$$

直观而言，当该量较小时，我们将 f 视为 $Q_{\boldsymbol{\pi}}^{1:m}(s,\mathbf{a}^{1:m})$ 的一种良好逼近。该形式也适用于一般函数逼近。详情请参见附录 D。

我们需要一种能够考虑分布失配的集中度度量。

定义 5.4（集中度）。 以下条件刻画了从 $d_{\boldsymbol{\pi}_*}$ 到采样分布的分布偏移。

$$\mathcal{C}_\mu^{d_{\boldsymbol{\pi}_*}} = \sup_{m\in\mathcal{N},f\in\mathcal{F}^m,\boldsymbol{\pi}\in\Pi^m} \frac{\left\|f - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f\right\|_{2,d_{\boldsymbol{\pi}_*}}}{\left\|f - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f\right\|_{2,\mathcal{D}^m}}.$$

回顾一下，$\|\cdot\|_{2,\ \rho}$ 是加权 L_2范数。在分子中，求和遍历（s、$\mathbf{a}^{1:m}$）$\sim d_{\boldsymbol{\pi}_*}$。而在分母中，求和遍历（$s$、$\mathbf{a}^{1:m}$），其中 $\mathcal{D}^m$ 作为 μ的经验版本。该概念在文献中发挥着与集中性系数类似的作用 [Munos，2003，Agarwal et al.，2020]：它衡量底层最优分布与我们所使用的样本分布之间的分布失配。

策略评估 在第 k 次迭代中，我们拥有当前策略 $\boldsymbol{\pi}_{\theta_k}$。我们通过正则化进行悲观策略评估，以减少评估 $Q_{\boldsymbol{\pi}_{\theta_k}}^{1:m}$ 时产生的价值偏差。

$$\omega_k^m \leftarrow \arg\min_{\omega}\left(f(s_0,\boldsymbol{\pi}_k^{1:m}) + \lambda\mathcal{E}^{1:m}(f,\boldsymbol{\pi}_{\theta_k})\right).$$

[4]在马尔可夫决策过程中，这种预言机称为 μ-重置模型 [Kakade and Langford，2002]。

这里 $\mathcal{E}$ 是式（5）中定义的贝尔曼误差。我们得到 $f_k^m = \phi^\top \omega_k^m$作为 $Q_{\boldsymbol{\pi}_{\theta_k}}^{1:m}$的悲观估计。在进行线性函数近似时，该更新有闭式解（参见定义5.1）。此外，在线性函数近似下，由于右侧目标函数对 ω呈二次依赖，因此可以高效地通过计算求解右侧的最小化问题。详情见附录C

策略改进 当价值函数和策略均采用线性参数化时（参见定义 5.1），对于任意（s，$\mathbf{a}^{1:m}$），$\in \mathcal{S} \times \mathcal{A}^m$

$$\pi_{k+1}^m(a^m|s, \mathbf{a}^{1:m-1}) \propto \pi_k^m(a^m|s, \mathbf{a}^{1:m-1}) \cdot \exp(\eta f_k^m(s, \mathbf{a}^{1:m})) \tag{6}$$

可进一步简化为参数更新，形式如下： $\mathbb{R}^d$

$$\theta_{k+1}^m = \theta_k^m + \eta \omega_k^m.$$

这一发现表明，在此设定下，策略改进比第 4 节中的策略改进简单得多。对于第 k 次迭代和智能体 $m \in \mathcal{N}$，我们只需将 $\eta\omega_k^m$ 加到策略参数 θ_k^m中，即可改进策略。

算法 通过悲观策略评估和直观的策略改进，我们提出了多智能体近端策略优化的悲观变体，见算法 2。

算法 2采用线性函数近似的悲观多智能体近端策略优化**输入：** 正则化系数 λ

。**输出：从 0,** $1 \cdots K-1$ **中均匀采样** k，**返回** $\bar{\boldsymbol{\pi}} = \boldsymbol{\pi}_{\theta_k}$。**1：为每个** $m \in \mathcal{N}$ **初始化** $\theta_0^m = 0$。**2：对于** $k = 0, \mathbf{1}, \ldots, K-1$ **执行 3：对于** $m = 1, \mathbf{2}, \cdots, N$ **执行 4：悲观策略评估：** $\omega_k^m \leftarrow \arg\min_{\omega} \left(f(s_0, \boldsymbol{\pi}_{\theta_k}^{1:m}) + \lambda \mathcal{E}^{1:m}(f, \boldsymbol{\pi}_{\theta_k}) \right)$。**5：策略改进：** $\theta_{k+1}^m = \theta_k^m + \eta\omega_k^m$。**6：结束循环 7：结束循环**

现在我们已经准备好给出本节的主定理。

定理 5.5。对于完全合作马尔可夫博弈中由算法 2得到的输出策略 $\bar{\boldsymbol{\pi}}$，设 $\eta = (1-\gamma)\sqrt{\frac{\log_2 |\mathcal{A}|}{K}}$ 且 $\lambda = (1-\gamma)^{-1}\left(\frac{d \log \frac{nLR}{\delta}}{n}\right)^{-2/3}$。经过 K 次迭代后，以至少 $1-\delta$ 的概率，我们有 $J(\boldsymbol{\pi}_*) - J(\bar{\boldsymbol{\pi}})$上界为

$$\mathcal{O}\left(\frac{N}{(1-\gamma)^2} \sqrt{\frac{\log |\mathcal{A}|}{K}} + \frac{\mathcal{C}_\mu^{d_{\boldsymbol{\pi}_*}}}{(1-\gamma)^2} \sqrt[3]{\frac{d \log \frac{nLR}{\delta}}{n}} \right)$$

为理解这一界，第一项表示镜像下降更新（6）所累积的优化误差。第一项对折扣因子具有 $(1-\gamma)^{-2}$ 依赖性，这一依赖性可能并不紧，我们将改进它留作未来工作。第二项表示训练过程中累积的估计误差。我们使用来自 μ 的状态-动作对和采样预言机来最小化 $\mathcal{E}^{1:m}$（f，$\boldsymbol{\pi}$），从而引入分布失配，其由 $\mathcal{C}_\mu^{d_{\boldsymbol{\pi}_*}}$表示。请注意，这种单策略集中性已经弱于传统的集中性系数 [Munos，2003，Farahmand 等人，2010，Perolat 等人，2015]。直观而言，较小的集中度值

要求数据分布 μ 接近 d_{π_*}，后者是最优策略未知的占用分布。另一方面，如果 $\mathcal{C}_{\mu}^{d_{\pi_*}}$较大，那么该界就会变松。我们在附录中为一般函数逼近给出了类似结果（参见定理D.7）。

该定理没有明确依赖状态空间 $\mathcal{S}$。因此，即使在无限状态设定下，该在线算法对函数逼近也能给出良好的保证。

为证明定理 5.5，关于贝尔曼一致悲观主义的定量分析 [Xie 等人， 2021] 很有帮助。我们利用合作马尔可夫博弈的条件依赖结构，获得了统计保证和收敛保证。详情请参阅附录 C。

6 结论

在这项工作中，我们提出了一种新的多智能体 PPO 算法，该算法以次线性速率收敛到全局最优策略。该算法的关键是多智能体性能差异引理，它使顺序局部策略更新成为可能。作为推广，我们将该算法扩展到离策略设定，并给出了类似的收敛保证。据我们所知，这是合作马尔可夫博弈中首个具有可证明保证的多智能体 PPO 算法。

致谢

JDL 感谢美国陆军研究办公室通过 MURI 奖 W911NF-11-1-0304、斯隆研究奖学金、NSF CCF 2002272、NSF IIS 2107304、NSF CIF 2212262、ONR 青年研究者奖以及 NSF CAREER 奖 2144994 提供的支持。

参考文献

A. Agarwal、S. M. Kakade、J. D. Lee 和 G. Mahajan。马尔可夫决策过程中策略梯度方法的最优性与近似。在学习理论会议，第 64–66 页。PMLR，2020。A. Antos、C. Szepesv´ari 和 R. Munos。基于贝尔曼残差最小化的拟合策略迭代与单一样本路径学习近最优策略。机器学习，71(1):89–129，2008。S. R. K. Branavan、H. Chen、L. S. Zettlemoyer 和 R. Barzilay。通过强化学习将指令映射为动作。在*ACL* 第 *47* 届年会与 *AFNLP* 第 *4* 届国际自然语言处理联合会议论文集：第 *1* 卷，ACL '09，第 82–90 页，美国，2009。计算语言学协会。N. Brown 和 T. Sandholm。用于单挑无限注扑克的超人类 AI：Libratus 击败顶尖职业选手。《科学》，359(6374):418–424，2018。S. Cen、Y. Wei 和 Y. Chi。带熵正则化的竞争博弈快速策略外梯度方法。在神经信息处理系统进展，第 27952–27964 页。Curran Associates, Inc.，2021。S. Cen、Y. Chi、S. S. Du 和 L. Xiao。零和马尔可夫博弈中策略优化的更快最后迭代收敛。*arXiv* 预印本 *arXiv:2210.01050*，2022。

J. Chen 和 N. Jiang。批量强化学习中的信息论考量。在国际机器学习会议，第 1042–1051 页。PMLR，2019。C. Daskalakis、D. J. Foster 和 N. Golowich。用于竞争性强化学习的独立策略梯度方法。神经信息处理系统进展，33:5527–5540，2020。C. S. de Witt、T. Gupta、D. Makoviichuk、V. Makoviychuk、P. H. Torr、M. Sun 和 S. Whiteson。StarCraft 多智能体挑战中，独立学习是否就足够？*arXiv* 预印本 *arXiv:2011.09533*，2020。D. Ding、C.-Y. Wei、K. Zhang 和 M. Jovanovic。面向大规模马尔可夫势博弈的独立策略梯度：更优收敛速率、函数逼近与无关博弈的收敛性。在国际机器学习会议，第 5166–5220 页。PMLR，2022。S. S. Du、S. M. Kakade、R. Wang 和 L. F. Yang。良好的表示是否足以实现样本高效的强化学习？*arXiv* 预印本 *arXiv:1910.03016*，2019。Y. Duan、X. Chen、R. Houthooft、J. Schulman 和 P. Abbeel。连续控制深度强化学习基准测试。在国际机器学习会议，第 1329–1338 页。PMLR，2016。J. Fan、Z. Wang、Y. Xie 和 Z. Yang。深度 Q 学习的理论分析。在动力学与控制学习，第 486–489 页。PMLR，2020。A.-m. Farahmand、C. Szepesv´ari 和 R. Munos。近似策略迭代和值迭代中的误差传播。神经信息处理系统进展，23，2010。J. Foerster、G. Farquhar、T. Afouras、N. Nardelli 和 S. Whiteson。反事实多智能体策略梯度。在*AAAI*人工智能会议论文集，第 32 卷，2018。R. Fox、S. M. Mcaleer、W. Overman 和 I. Panageas。独立自然策略梯度在马尔可夫势博弈中总能收敛。在人工智能与统计国际会议，第 4414–4425 页。PMLR，2022。S. Fujimoto、H. Hoof 和 D. Meger。解决行动者-评论家方法中的函数逼近误差。在国际机器学习会议，第 1587–1596 页。PMLR，2018。K. Gimpel 和 N. A. Smith。Softmax-margin CRF：用代价函数训练对数线性模型。在人类语言技术：北美计算语言学协会第 *2010* 届年会，第 733–736 页，2010。X. Guo、S. Singh、R. Lewis 和 H. Lee。通过深度学习设计奖励，以改进 Atari 游戏中的蒙特卡洛树搜索。*arXiv* 预印本 *arXiv:1604.07095*，2016。N. Heess、D. Silver 和 Y. W. Teh。采用基于能量的策略的行动者-评论家强化学习。在欧洲强化学习研讨会，第 45–58 页。PMLR，2013。C. Jin、R. Ge、P. Netrapalli、S. M. Kakade 和 M. I. Jordan。如何高效逃离鞍点。在国际机器学习会议，第 1724–1732 页。PMLR，2017。

Y. Jin、Z. Yang 和 Z. Wang。《对于离线强化学习而言，悲观性是否能够被证明是高效的？》载于《国际机器学习会议》，第5084–5096页。PMLR，2021年。S. Kakade 和 J. Langford。《近似最优的近似强化学习》。载于《第19届国际机器学习大会论文集》。Citeseer，2002年。S. M. Kakade。《自然策略梯度》。载于《神经信息处理系统进展》，第14卷，2001年。S. M. Kakade。《强化学习的样本复杂度》。伦敦大学、伦敦大学学院（英国），2003年。M. Kearns 和 S. Singh。《多项式时间内的近似最优强化学习》。Machine Learning，49(2):209–232，2002年。L. Kraemer 和 B. Banerjee。《多智能体强化学习作为分散式规划的演练》。神经计算，190:82–94，2016年。J. G. Kuba、R. Chen、M. Wen、Y. Wen、F. Sun、J. Wang 和 Y. Yang。《多智能体强化学习中的信赖域策略优化》。载于《国际学习表征会议》，2022年。URL https://openreview.net/forum?id=EcGGFkNTxdJ。M. Laskin、K. Lee、A. Stooke、L. Pinto、P. Abbeel 和 A. Srinivas。《使用增强数据进行强化学习》。载于《神经信息处理系统进展》，33:19884–19895，2020年。K.-H. Lee、I. Fischer、A. Liu、Y. Guo、H. Lee、J. Canny 和 S. Guadarrama。《预测信息加速强化学习》。载于《神经信息处理系统进展》，33:11890–11901，2020年。S. Leonardos、W. Overman、I. Panageas 和 G. Piliouras。《马尔可夫势博弈中多智能体策略梯度的全局收敛》。载于《国际学习表征会议》，2022年。URL https://openreview.net/forum?id=gfwON7rAm4。G. Li、Y. Wei、Y. Chi、Y. Gu 和 Y. Chen。《利用生成模型打破基于模型的强化学习中的样本量壁垒》。载于《神经信息处理系统进展》，33:12861–12872，2020年。T. P. Lillicrap、J. J. Hunt、A. Pritzel、N. Heess、T. Erez、Y. Tassa、D. Silver 和 D. Wierstra。《使用深度强化学习进行连续控制》。arXiv预印本 arXiv:1509.02971，2015年。M. L. Littman。《马尔可夫博弈作为多智能体强化学习的框架》。载于《机器学习论文集1994》，第157–163页。爱思唯尔，1994年。B. Liu、Q. Cai、Z. Yang 和 Z. Wang。《神经信赖域／近端策略优化实现全局最优策略》。载于《神经信息处理系统进展》，32，2019年。Y. Liu、A. Swaminathan、A. Agarwal 和 E. Brunskill。《无需广泛探索即可获得理论保证良好的批量离策略强化学习》。载于《神经信息处理系统进展》，33:1264–1274，2020年。R. Lowe、Y. I. Wu、A. Tamar、J. Harb、O. Pieter Abbeel 和 I. Mordatch。《混合合作-竞争环境中的多智能体行动者-评论家》。载于《神经信息处理系统进展》，30，2017年。

V. Mnih、K. Kavukcuoglu、D. Silver、A. A. Rusu、J. Veness、M. G. Bellemare、A. Graves、M. Riedmiller、A. K. Fidjeland、G. Ostrovski 等。《通过深度强化学习实现人类水平的控制》。《自然》，518(7540):529–533，2015年。T. Moskovitz、J. Parker-Holder、A. Pacchiano、M. Arbel 和 M. Jordan。《深度强化学习中的战术乐观性与悲观性》。《神经信息处理系统进展》，34:12849–12863，2021年。R. Munos。《近似策略迭代的误差界》。载于《国际机器学习会议》，第560–567页，2003年。R. Munos 和 C. Szepesv´ari。《拟合值迭代的有限时间界》。《机器学习研究杂志》，9(5)，2008年。Y. Nesterov。《凸优化导论讲义：基础课程》，第87卷。施普林格科学与商业媒体，2003年。G. Papoudakis、F. Christianos、L. Sch¨afer 和 S. V. Albrecht。《合作任务中的多智能体深度强化学习算法基准测试》。载于《第三十五届神经信息处理系统会议数据集与基准赛道（第一轮）》，2021年。J. Perolat、B. Scherrer、B. Piot 和 O. Pietquin。《双人零和马尔可夫博弈的近似动态规划》。载于《国际机器学习会议》，第1321–1329页，2015年。P. Rashidinejad、B. Zhu、C. Ma、J. Jiao 和 S. Russell。《连接离线强化学习与模仿学习：一个关于悲观性的故事》。《神经信息处理系统进展》，34:11702–11716，2021年。J. Schulman、S. Levine、P. Abbeel、M. Jordan 和 P. Moritz。《信赖域策略优化》。载于《国际机器学习会议》，第1889–1897页。PMLR，2015年。J. Schulman、F. Wolski、P. Dhariwal、A. Radford 和 O. Klimov。《近端策略优化算法》。*arXiv*预印本 *arXiv:1707.06347*，2017年。S. Shalev-Shwartz 和 S. Ben-David。《理解机器学习：从理论到算法》。剑桥大学出版社，2014年。L. Shani、Y. Efroni 和 S. Mannor。《自适应信赖域策略优化：正则化马尔可夫决策过程的全局收敛与更快收敛率》。载于《*AAAI*人工智能会议论文集》，第5668–5675页，2020年。L. S. Shapley。《随机博弈》。《美国国家科学院院刊》，39(10):1095–1100，1953年。A. Sidford、M. Wang、X. Wu、L. F. Yang 和 Y. Ye。《使用生成模型求解折扣马尔可夫决策过程的近似最优时间复杂度与样本复杂度》。*arXiv*预印本 *arXiv:1806.01492*，2018年。D. Silver、A. Huang、C. J. Maddison、A. Guez、L. Sifre、G. Van Den Driessche、J. Schrit-twieser、I. Antonoglou、V. Panneershelvam、M. Lanctot 等。《利用深度神经网络和树搜索掌握围棋》。《自然》，529(7587):484–489，2016年。D. Silver、J. Schrittwieser、K. Simonyan、I. Antonoglou、A. Huang、A. Guez、T. Hubert、L. Baker、M. Lai、A. Bolton 等。《无需人类知识掌握围棋》。《自然》，550(7676):354–359，2017年。

R. S. Sutton、D. McAllester、S. Singh 和 Y. Mansour。函数逼近下强化学习的策略梯度方法。神经信息处理系统进展，12，1999。S. Thrun 和 A. Schwartz。使用函数逼近进行强化学习时面临的问题。载于第四届联结主义模型暑期学校论文集，第 255 卷，第 263 页。新泽西州希尔斯代尔，1993。Y. Tian、J. Ma、Q. Gong、S. Sengupta、Z. Chen、J. Pinkerton 和 L. Zitnick。Elf OpenGo：对 AlphaZero 的分析及开放式重新实现。载于国际机器学习会议，第 6244–6253 页。PMLR，2019。M. Uehara 和 W. Sun。悲观模型式离线强化学习：部分覆盖下的 PAC 界与后验采样。*arXiv* 电子预印本，第 arXiv–2107 页，2021。O. Vinyals、I. Babuschkin、W. M. Czarnecki、M. Mathieu、A. Dudzik、J. Chung、D. H. Choi、R. Powell、T. Ewalds、P. Georgiev 等。运用多智能体强化学习在星际争霸 II 中达到宗师级水平。《自然》，575(7782)：350–354，2019。Y. Wen、Y. Yang、R. Luo、J. Wang 和 W. Pan。多智能体强化学习的概率递归推理。*arXiv* 预印本 *arXiv:1901.09207*，2019。T. Xie 和 N. Jiang。批量强化学习的 Q* 近似方案：理论比较。载于人工智能不确定性会议，第 550–559 页。PMLR，2020。T. Xie、C.-A. Cheng、N. Jiang、P. Mineiro 和 A. Agarwal。离线强化学习的贝尔曼一致悲观主义。神经信息处理系统进展，34：6683–6694，2021。Y. Yang、R. Luo、M. Li、M. Zhou、W. Zhang 和 J. Wang。平均场多智能体强化学习。载于国际机器学习会议，第 5571–5580 页。PMLR，2018。C. Yu、A. Velu、E. Vinitsky、Y. Wang、A. Bayen 和 Y. Wu。近端策略优化在协作式多智能体博弈中的惊人效果。*arXiv* 预印本 *arXiv:2103.01955*，2021。W. Zhan、B. Huang、A. Huang、N. Jiang 和 J. 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。H. Zhang、W. Chen、Z. Huang、M. Li、Y. Yang、W. Zhang 和 J. Wang。多智能体协调的双层行动者-评论家方法。载于*AAAI*人工智能会议论文集，第 7325–7332 页，2020。K. Zhang、Z. Yang 和 T. Ba̧sar。多智能体强化学习：理论与算法精选综述。强化学习与控制手册，第 321–384 页，2021。Y. Zhao、Y. Tian、J. Lee 和 S. Du。两人零和马尔可夫博弈中可证明高效的策略优化。载于人工智能与统计国际会议，第 2736–2761 页。PMLR，2022。

第 A 节的子问题求解器4

算法 3MA-PPO 的策略改进求解器

输入： MG $(\mathcal{N}, \mathcal{S}, \boldsymbol{\mathcal{A}}\ \mathcal{P}, r, \gamma)$, 迭代 T 次，步长 η，样本数 $\{s_t, \mathbf{a}_t^{1:m-1}, a_t^m\}_{t=0}^{T-1}$.**输出：** 策略更新 θ。

1: Initialize $\theta_0 = 0$.
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: Let $(s, \mathbf{a}^{1:m-1}, a) \leftarrow (s_t, \mathbf{a}_t^{1:m-1}, a_t^m)$.
4: $\theta(t+\frac{1}{2}) \leftarrow \theta(t) - 2\eta\phi(s, \mathbf{a}^{1:m-1}, a)\left((\theta(t) - \theta_k^m)^\top \phi(s, \mathbf{a}^{1:m-1}, a^m) - \beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_k}^{1:m}(s, \mathbf{a}^{1:m-1}, a^m)\right)$.
5: $\theta(t+1) \leftarrow \Pi_\Theta \theta(t+\frac{1}{2})$
6: **end for**
7: Calculate average: $\bar{\theta} \leftarrow \frac{1}{T}\sum_{t=1}^{T}\theta_t$.

B 第 4 节的证明4

首先，我们注意到，使用 SGD 更新来求解 MSE 问题具有以下保证。

引理 B.1 （平均策略）。 对于凸目标函数 $F(\theta)$，假设梯度以 G 为界，且输出 θ 收敛到该类别中的最优函数 *at*

$$F(\bar{\theta}) - \min_{\|\theta\| \le R} F(\theta) \le \frac{GR}{\sqrt{T}}$$

其中我们设定 $\eta = \frac{R}{G\sqrt{T}}$。

证明。 请参阅定理 14。8 [Shalev-Shwartz 和 Ben-David, 2014]。 □

现在我们转向算法 3，在该算法中，我们将样本 $\{s_t, \mathbf{a}_t^{1:m-1}, a_t^m\}_{t=0}^{T-1}$来自 $\sigma_k = \nu_k \boldsymbol{\pi}_{\theta_k}$ 输入算法（对于 $m \in \mathcal{N}$），以最小化

$$L(\theta^m) = \mathbb{E}_{\sigma_k}\left((\theta^m)^\top \phi(s, \mathbf{a}^{1:m-1}, a^m) - (\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s, \mathbf{a}^{1:m-1}, a^m) + (\theta_k^m)^\top \phi(s, \mathbf{a}^{1:m-1}, a^m))\right)^2.$$

我们对该算法有如下理论保证。

引理 B.2 (策略改进误差)。 在第 k 次外循环中，算法 θ_{k+1}的输出策略 *3* 满足

$$\sqrt{L(\theta_{k+1}^m)} \le \epsilon_k^m$$

其中 $\epsilon_k^m = \epsilon_{approx} + \mathcal{O}(T^{-\frac{1}{+}})$。

证明。 对于算法 3，我们有 $\|\phi\|_2 \le 1$ 且 $\Theta = \{\|\theta\| \le R | \theta \in \mathbb{R}^d\}$。因此， $L(\theta)$ 的梯度有界于

$$G = 2\left(R + \frac{1}{(1-\gamma)\beta_k}\right).$$

由引理 B.1可知，当外层迭代为 $k < K$ 时，以下保证成立

$$L(\theta_{k+1}^m) \le = \min_\theta L(\theta) + \mathcal{O}(1/\sqrt{T}),$$

当我们设定 $\eta = \frac{R}{G\sqrt{T}}$ 时。因此

$$\epsilon_k^m = \sqrt{\min_\theta L(\theta)} + \mathcal{O}(T^{-\frac{1}{4}}) = \epsilon_{approx} + \mathcal{O}(T^{-\frac{1}{4}}).$$

证明完成。 □

引理 B.3（多智能体优势分解）。在合作马尔可夫博弈中，对于任意联合策略 π、状态 s 和智能体 $1:m$，

$$A_{\boldsymbol{\pi}}^{1:m}(s, \mathbf{a}^{1:m}) = \sum_{i=1}^{m} A_{\boldsymbol{\pi}}^{i}(s, \mathbf{a}^{1:i-1}, a^i)$$

证明。 请参阅引理 1 [Kuba 等人，2022]. □

引理的证明4.1

证明。根据经典性能差异引理 [Kakade and Langford，2002，引理 6。1] 我们有

$$J(\boldsymbol{\pi}_*) - J(\boldsymbol{\pi}) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\sigma_*} A_{\boldsymbol{\pi}}^{1:N}(s, \mathbf{a}^{1:N})$$

通过多智能体优势分解引理（参见引理），将全体智能体优势函数分解为个体贡献，我们有 B.3。

$$\begin{aligned} J(\boldsymbol{\pi}_*) - J(\boldsymbol{\pi}) &= \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\sigma_*} A_{\boldsymbol{\pi}}^{1:N}(s, \mathbf{a}^{1:N}) \\ &= \frac{1}{1-\gamma} \sum_{m=1}^{N} \mathbb{E}_{s\sim\nu_*} \mathbb{E}_{\mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*^{1:m-1}} \left\langle A_{\boldsymbol{\pi}}^{m}(s, \mathbf{a}^{1:m-1}, a^m), \pi_*^m(\cdot|s, \mathbf{a}^{1:m-1}) \right\rangle. \end{aligned}$$

注意，我们有 $\sum_a \pi^m(a|s, a^{1:m-1}) A_{\boldsymbol{\pi}}^m(s, \mathbf{a}^{1:m-1}, a) = 0$，那么

$$\begin{aligned} &\frac{1}{1-\gamma} \sum_{m=1}^{N} \mathbb{E}_{s\sim\nu_*} \mathbb{E}_{\mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*^{1:m-1}} \left\langle A_{\boldsymbol{\pi}}^{m}(s, \mathbf{a}^{1:m-1}, \cdot), \pi_*^m(\cdot|s, \mathbf{a}^{1:m-1}) - \pi^m(\cdot|s, a^{1:m-1}) \right\rangle \\ = &\frac{1}{1-\gamma} \sum_{m=1}^{N} \mathbb{E}_{s\sim\nu_*} \mathbb{E}_{\mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*^{1:m-1}} \left\langle Q_{\boldsymbol{\pi}}^{1:m}(s, \mathbf{a}^{1:m-1}, \cdot), \pi_*^m(\cdot|s, \mathbf{a}^{1:m-1}) - \pi^m(\cdot|s, a^{1:m-1}) \right\rangle \end{aligned}$$

其中，最后一行成立是因为 $A_{\boldsymbol{\pi}}^m(s, \mathbf{a}^{1:m-1}, a^m) = Q_{\boldsymbol{\pi}}^{1:m}(s, \mathbf{a}^{1:m-1}, a^m) - Q_{\boldsymbol{\pi}}^{1:m-1}(s, \mathbf{a}^{1:m-1})$ 和 $Q_{\boldsymbol{\pi}}^{1:m-1}(s, \mathbf{a}^{1:m-1})$ 可以省略，因为它不随 a^m 变化。

□

命题的证明4.2。

证明。 对于任意 $(s, \mathbf{a}^{1:m-1}) \in \mathcal{S} \times \mathcal{A}^{m-1}$，策略 $\hat{}\pi_{k+1}^m(\cdot|s, \mathbf{a}^{1:m-1})$ 通过以下方式获得。

$$\begin{aligned} \max_{\pi^m} \mathbb{E}_{\nu_k} \Big[\langle \hat{Q}_{\boldsymbol{\pi}_k}^{1:m}(s, \mathbf{a}^{1:m-1}, \cdot), \pi^m(\cdot|s, \mathbf{a}^{1:m-1}) \rangle - \beta_k KL\left(\pi^m(\cdot|s, \mathbf{a}^{1:m-1}) \| \pi_{\theta_k}(\cdot|s, \mathbf{a}^{1:m-1})\right) \Big] \\ \text{s.t.} \quad \sum_{a^m \in \mathcal{A}} \pi^m(a^m|s, \mathbf{a}^{1:m-1}) = 1 \end{aligned}$$

将约束作为拉格朗日乘子加入后，我们得到

$$\int_{\mathcal{S}\times\mathcal{A}^{m-1}} \left[\langle \hat{Q}_{\boldsymbol{\pi}_k}^{1:m}(s, \mathbf{a}^{1:m-1}, \cdot), \pi^m(\cdot|s, \mathbf{a}^{1:m-1}) \rangle - \beta_k KL\left(\pi^m(\cdot|s, \mathbf{a}^{1:m-1}) \| \pi_{\theta_k^m}(\cdot|s, \mathbf{a}^{1:m-1})\right) \right] \sigma_k ds d\mathbf{a}^{1:m-1}$$

$$+\int_{\mathcal{S}\times\mathcal{A}^{m-1}}\left(\sum_{a^m\in\mathcal{A}}\pi^m(a^m|s,\mathbf{a}^{1:m-1})-1\right)dsd\mathbf{a}^{1:m-1}$$

注意，$\pi_{\theta_k^m}\propto\exp\{\theta_k^{m\top}\phi\}$，最优性条件给出

$$\hat{\pi}_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1})\propto\exp\{\beta_k^{-1}\hat{Q}_{\pi_k}^{1:m}(s,\mathbf{a}^{1:m-1},\cdot)+\theta_k^{m\top}\phi(s,\mathbf{a}^{1:m-1},a)\}.$$

□

引理 B.4。假设对于任意智能体 $m\in\mathcal{N}$，策略改进误差和策略评估误差满足

$$\mathbb{E}_{\sigma_k}\left(\theta_{k+1}^{m\top}\phi(s,\mathbf{a}^{1:m-1},a^m)-(\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s,\mathbf{a}^{1:m-1},a^m)+\theta_k^{m\top}\phi(s,\mathbf{a}^{1:m-1},a^m))\right)^2\leq(\epsilon_k^m)^2, \tag{7}$$

$$\mathbb{E}_{\sigma_k}\left(\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s,\mathbf{a}^{1:m-1},a^m)-Q_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s,\mathbf{a}^{1:m-1},a^m)\right)^2\leq(\xi_k^m)^2. \tag{8}$$

考虑 $\theta_{k+1}^{m\top}\phi-\theta_k^{m\top}\phi$ 的 L_∞-范数，我们有

$$\mathbb{E}_{s\sim\nu_*,\mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*}\left\|(\theta_{k+1}^m-\theta_k^m)^\top\phi(s,\mathbf{a}^{1:m-1},\cdot)-\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s,\mathbf{a}^{1:m-1},\cdot)\right\|_\infty\leq\frac{\delta_k^m}{2}$$

其中 $\delta_k^m=2\phi_k^{m-1}\epsilon_k^m$。

证明。该证明是显而易见的，

$$\begin{aligned}&\mathbb{E}_{s\sim\nu_*,\mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*^{1:m-1}}\left\|(\theta_{k+1}^m-\theta_k^m)^\top\phi(s,\mathbf{a}^{1:m-1},\cdot)-\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s,\mathbf{a}^{1:m-1},\cdot)\right\|_\infty\\\leq&\mathop{\mathbb{E}}_{\substack{s\sim\nu_*\\\mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*}}\left\|(\theta_{k+1}^m-\theta_k^m)^\top\phi(s,\mathbf{a}^{1:m-1},a^m)-\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s,\mathbf{a}^{1:m-1},a^m)\right\|.\end{aligned}$$

我们从 σ_* 转向 σ_k，并引入集中性系数来衡量分布偏移。

$$\begin{aligned}&\mathop{\mathbb{E}}_{\substack{s\sim\nu_k\\\mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_{\theta_k}}}\left\|(\theta_{k+1}^m-\theta_k^m)^\top\phi(s,\mathbf{a}^{1:m-1},a^m)-\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s,\mathbf{a}^{1:m-1},a^m)\right\|\cdot\frac{\nu_*\boldsymbol{\pi}_*^{1:m-1}}{\nu_k\boldsymbol{\pi}_{\theta_k}^{1:m-1}}\\\leq&\left[\mathbb{E}_{\sigma_k}\left((\theta_{k+1}^m-\theta_k^m)^\top\phi-\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}\right)^2\right]^{1/2}\cdot\left[\mathbb{E}_{\sigma_k}\left|\frac{d(\nu_*\boldsymbol{\pi}_*^{1:m-1})}{d(\nu_k\boldsymbol{\pi}_{\theta_k}^{1:m-1})}\right|^2\right]^{1/2}\\=&\epsilon_k^m\phi_k^{m-1}\end{aligned}$$

其中，第二行使用了柯西-施瓦茨不等式。

证明完成。 □

回顾一下，我们将基于 $\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}$ 的理想更新策略定义为 $\hat{\pi}_{k+1}^m$。相应地，我们将基于精确价值函数$Q_{\boldsymbol{\pi}_{\theta_k}}^{1:m}$ 的理想更新定义为

$$\begin{aligned}&\pi_{k+1}^m\leftarrow\arg\max_{\pi^m}F(\pi^m)\\&F(\pi^m)=\mathbb{E}_{\sigma_k}\left[\langle Q_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s,\mathbf{a}^{1:m-1},\cdot),\pi^m(\cdot|s,\mathbf{a}^{1:m-1})\rangle-\beta_kKL\left(\pi^m(\cdot|s,\mathbf{a}^{1:m-1})\|\pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1})\right)\right].\end{aligned}$$

在对数线性参数化下：$\pi_{\theta_k^m}\propto\exp\{\phi^\top\theta_k^m\}$，类似地我们有

$$\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1})\propto\exp\left\{\beta_k^{-1}Q_{\pi_k}^{1:m}(s,\mathbf{a}^{1:m-1},\cdot)+\phi^\top(s,\mathbf{a}^{1:m-1},\cdot)\theta_k^m\right\}.$$

引理 B.5 (误差传播)。 假设对于任意智能体 $m \in \mathcal{N}$ 和 $(s,\mathbf{a}^{1:m-1}) \in \mathcal{S}\times\mathcal{A}^{m-1}$，策略改进误差和策略评估误差满足，

$$\mathbb{E}_{\sigma_k}\left((\theta_{k+1}^m-\theta_k^m)^\top\phi-\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}\right)^2\le(\epsilon_k^m)^2,$$
$$\mathbb{E}_{\sigma_k}\left(\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}-Q_{\boldsymbol{\pi}_{\theta_k}}^{1:m}\right)^2\le(\xi_k^m)^2$$

为简洁起见，我们省略$(s,\mathbf{a}^{1:m-1},a^m)$。

比较统计误差，可得

$$\left|\mathbb{E}_{s\sim\nu_*,\mathbf{a}\sim\boldsymbol{\pi}_*}\left\langle\log\frac{\pi_{\theta_{k+1}^m}(\cdot|s,\mathbf{a}^{1:m-1})}{\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1})},\pi_*^m(\cdot|s,\mathbf{a}^{1:m-1})-\pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1})\right\rangle\right|\le\Delta_k^m.$$

其中 $\Delta_k^m=\sqrt{2}(\phi_k^m+\phi_k^{m-1})\cdot\left(\epsilon_k^m+\frac{\xi_k^m}{\beta_k}\right)$

引理B.5给出了实际参数化$\pi_{\theta_{k+1}^m}$ 基于 $\hat{}\, Q^{1:m}$ 的策略与理想策略 π_{k+1}^m 基于精确价值函数 $Q^{1:m}$之间的定量差异。

证明。 首先，根据定义，对于任意 $m\in\mathcal{N}$和 s，$\mathbf{a}^{1:m-1}\in\mathcal{S}\times\mathcal{A}^{m-1}$，我们有

$$\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1})=\exp\left\{\beta_k{}^{-1}Q_{\boldsymbol{\pi}_{\theta_k}}+\theta_k^{m\top}\phi(s,\mathbf{a}^{1:m-1},\cdot)\right\}/W(s,\mathbf{a}^{1:m-1}),\tag{9}$$
$$\pi_{\theta_{k+1}^m}(\cdot|s,\mathbf{a}^{1:m-1})=\exp\left\{\theta_{k+1}^m{}^\top\phi(s,\mathbf{a}^{1:m-1},\cdot)\right\}/M(s,\mathbf{a}^{1:m-1}).\tag{10}$$

将其代入表达式，可得

$$\begin{aligned}\text{LHS}&=\left\langle\log\pi_{\theta_{k+1}^m}-\log\pi_{k+1}^m,\pi_*^m-\pi_{\theta_k^m}\right\rangle\\&=\left\langle\theta_{k+1}^m{}^\top\phi-(\beta_k^{-1}Q_{\boldsymbol{\pi}_{\theta_k}}+\theta_k^{m\top}\phi),\pi_*^m-\pi_{\theta_k^m}\right\rangle\\&=\underbrace{\left\langle\theta_{k+1}^m{}^\top\phi-(\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}+\theta_k^{m\top}\phi),\pi_*^m-\pi_{\theta_k^m}\right\rangle}_{(a)}+\underbrace{\left\langle\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}-\beta_k^{-1}Q_{\boldsymbol{\pi}_{\theta_k}},\pi_*^m-\pi_{\theta_k^m}\right\rangle}_{(b)}\end{aligned}$$

在第二行中，我们使用了以下事实：$\left\langle\log\frac{W}{M},\pi_*^m-\pi_{\theta_k^m}\right\rangle=\log\frac{W}{M}\sum_{a^m}\left(\boldsymbol{\pi}_*^m(\cdot)-\pi_{\theta_k^m}(\cdot)\right)=0$。

分别对这两项进行界定，可得

- 对于（a），对 $\mathcal{S}\times\mathcal{A}^{m-1}$ 取期望可得

$$\begin{aligned}&\left|\mathbb{E}_{s\sim\nu_*}\mathbb{E}_{\mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*}\left\langle\theta_{k+1}^m{}^\top\phi-(\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}+\theta_k^{m\top}\phi),\pi_*^m-\pi_{\theta_k^m}\right\rangle\right|\\&=\left|\int_{\mathcal{S}\times\mathcal{A}^{m-1}\times\mathcal{A}}\left[(\theta_{k+1}^m-\theta_k^m)^\top\phi-\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}\right]\cdot\left(\pi_*^m-\pi_{\theta_k^m}\right)da^m\cdot\boldsymbol{\pi}_*^{1:m-1}d(\mathbf{a}^{1:m-1})\cdot\nu_*(s)ds\right|\end{aligned}$$

这里的期望是关于 $\sigma*$ 计算的，我们通过引入集中性系数，将其改为对 σ_k 的期望

$$\begin{aligned}&\left|\int_{\mathcal{S}\times\mathcal{A}^{m-1}\times\mathcal{A}}\left[(\theta_{k+1}^m-\theta_k^m)^\top\phi-\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}\right]\cdot\left(\frac{\boldsymbol{\pi}_*^{1:m}}{\boldsymbol{\pi}_{\theta_k}^{1:m}}-\frac{\boldsymbol{\pi}_*^{1:m-1}\pi_{\theta_k^m}}{\boldsymbol{\pi}_{\theta_k}^{1:m}}\right)\boldsymbol{\pi}_{\theta_k}^{1:m}(\mathbf{a}^{1:m}|s)d(\mathbf{a}^{1:m})\cdot\nu_*(s)ds\right|\\&=\left|\int_{\mathcal{S}\times\mathcal{A}^{m-1}\times\mathcal{A}}\left[(\theta_{k+1}^m-\theta_k^m)^\top\phi-\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}\right]\cdot\frac{\nu_*(s)}{\nu_k(s)}\left(\frac{\boldsymbol{\pi}_*^{1:m}}{\boldsymbol{\pi}_{\theta_k}^{1:m}}-\frac{\boldsymbol{\pi}_*^{1:m-1}\pi_{\theta_k^m}}{\boldsymbol{\pi}_{\theta_k}^{1:m}}\right)d\sigma_k\right|\end{aligned}$$

$$\overset{(i)}{\leq} \left[\mathbb{E}_{\sigma_k}\left((\theta_{k+1}^m-\theta_k^m)^\top\phi-\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}\right)^2\right]^{1/2}\cdot\left[\mathbb{E}_{\sigma_k}\left|\frac{d(\nu_*\boldsymbol{\pi}_*^{1:m})}{d(\nu_k\boldsymbol{\pi}_{\theta_k}^{1:m})}-\frac{d(\nu_*\boldsymbol{\pi}_*^{1:m-1})}{d(\nu_k\boldsymbol{\pi}_{\theta_k}^{1:m-1})}\right|^2\right]^{1/2}$$

$$\overset{(ii)}{\leq} \sqrt{2}\epsilon_k^m(\phi_k^m+\phi_k^{m-1})$$

(*i*)：这是因为柯西-施瓦茨不等式。
(*ii*)：这是因为：$\sqrt{\int|f-g|^2d\sigma}\leq\sqrt{\int 2|f|^2d\sigma+\int 2|g|^2d\sigma}\leq\sqrt{2}(\|f\|_{2,\sigma}+\|g\|_{2,\sigma})$。

- 对于（b），对 $\mathcal{S}\times\mathcal{A}^{m-1}$ 取期望，我们有

$$\left|\mathbb{E}_{s\sim\nu_*}\mathbb{E}_{\mathbf{a}^{1:m-1}\sim\boldsymbol{\pi}_*}\left\langle\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}-\beta_k^{-1}Q_{\boldsymbol{\pi}_{\theta_k}},\pi_*^m-\pi_{\theta_k^m}\right\rangle\right|$$

$$=\left|\int_{\mathcal{S}\times\mathcal{A}^{m-1}\times\mathcal{A}}\left[\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}-\beta_k^{-1}Q_{\boldsymbol{\pi}_{\theta_k}}\right]\cdot\left(\pi_*^m-\pi_{\theta_k^m}\right)da^m\cdot\boldsymbol{\pi}_*^{1:m-1}d(\mathbf{a}^{1:m-1})\cdot\nu_*(s)ds\right|$$

类似地，我们将对 $\sigma*$ 的期望替换为对 σ_k 的期望

$$\left|\int_{\mathcal{S}\times\mathcal{A}^{m-1}\times\mathcal{A}}\left[\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}-\beta_k^{-1}Q_{\boldsymbol{\pi}_{\theta_k}}\right]\cdot\frac{\nu_*(s)}{\nu_k(s)}\left(\frac{\boldsymbol{\pi}_*^{1:m}}{\boldsymbol{\pi}_{\theta_k}^{1:m}}-\frac{\boldsymbol{\pi}_*^{1:m-1}\pi_{\theta_k^m}}{\boldsymbol{\pi}_{\theta_k}^{1:m}}\right)\boldsymbol{\pi}_{\theta_k}^{1:m}(\mathbf{a}^{1:m}|s)d(\mathbf{a}^{1:m})\cdot\nu_k(s)ds\right|$$

$$=\left|\int_{\mathcal{S}\times\mathcal{A}^{m-1}\times\mathcal{A}}\left[\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}-\beta_k^{-1}Q_{\boldsymbol{\pi}_{\theta_k}}\right]\cdot\frac{\nu_*(s)}{\nu_k(s)}\left(\frac{\boldsymbol{\pi}_*^{1:m}}{\boldsymbol{\pi}_{\theta_k}^{1:m}}-\frac{\boldsymbol{\pi}_*^{1:m-1}}{\boldsymbol{\pi}_{\theta_k}^{1:m-1}}\right)d\sigma_k\right|$$

$$\overset{(i)}{\leq}\left[\mathbb{E}_{\sigma_k}\left(\beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}-\beta_k^{-1}Q_{\boldsymbol{\pi}_{\theta_k}}\right)^2\right]^{1/2}\cdot\left[\mathbb{E}_{\sigma_k}\left|\frac{d(\nu_*\boldsymbol{\pi}_*^{1:m})}{d(\nu_k\boldsymbol{\pi}_{\theta_k}^{1:m})}-\frac{d(\nu_*\boldsymbol{\pi}_*^{1:m-1})}{d(\nu_k\boldsymbol{\pi}_{\theta_k}^{1:m-1})}\right|^2\right]^{1/2}$$

$$\overset{(ii)}{\leq}\frac{\sqrt{2}}{\beta_k}\xi_k^m(\phi_k^m+\phi_k^{m-1})$$

（i）：这是由柯西-施瓦茨不等式得出的。
(*ii*)：这是因为：$\sqrt{\int|f-g|^2d\sigma}\leq\sqrt{\int 2|f|^2d\sigma+\int 2|g|^2d\sigma}\leq\sqrt{2}(\|f\|_{2,\sigma}+\|g\|_{2,\sigma})$。

结合这些界，我们得出了引理 B.5 的证明。

$$\Delta_k^m=\sqrt{2}(\phi_k^m+\phi_k^{m-1})\cdot\left(\epsilon_k^m+\frac{\xi_k^m}{\beta_k}\right).$$

□

下面我们介绍一个在多智能体近端策略优化分析中至关重要的引理。原始版本被广泛采用，并已证明对镜像下降分析非常有用 [Nesterov，2003]。

引理 B.6 (单步下降)。对于理想更新策略 π_{k+1}^m、实际更新策略 $\pi_{\theta_{k+1}^m}$ 和当前策略 $\pi_{\theta_k^m}$，我们有，对于任意 $(s,\mathbf{a}^{1:m-1})\in\mathcal{S}\times\mathcal{A}^{m-1}$，

$$KL\left(\pi_*^m(\cdot|s,\mathbf{a}^{1:m-1})\|\pi_{\theta_{k+1}^m}(\cdot|s,\mathbf{a}^{1:m-1})\right)-KL\left(\pi_*^m(\cdot|s,\mathbf{a}^{1:m-1})\|\pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1})\right)$$

$$
\begin{aligned}
&\leq \left\langle \log \frac{\pi_{\theta_{k+1}^m}(\cdot|s,\mathbf{a}^{1:m-1})}{\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1})}, \pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1}) - \pi_*^m(\cdot|s,\mathbf{a}^{1:m-1}) \right\rangle \\
&+ \frac{1}{\beta_k} \left\langle A^m_{\pi_{\theta_k}}(s,\mathbf{a}^{1:m-1},\cdot), \pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1}) - \pi_*^m(\cdot|s,\mathbf{a}^{1:m-1}) \right\rangle - \frac{1}{2} \left\| \pi_{\theta_{k+1}^m}(\cdot|s,\mathbf{a}^{1:m-1}) - \pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1}) \right\|_1^2 \\
&- \left\langle (\theta_{k+1}^m - \theta_k^m)^\top \phi(s,\mathbf{a}^{1:m-1},\cdot), \pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1}) - \pi_{\theta_{k+1}^m}(\cdot|s,\mathbf{a}^{1:m-1}) \right\rangle
\end{aligned}
$$

证明。在证明中，为避免滥用记号，我们在不引起歧义时直接省略 $(s,\mathbf{a}^{1:m-1})$。根据定义，我们有

$$
\begin{aligned}
& KL\left(\pi_*^m \| \pi_{\theta_k^m}\right) - KL\left(\pi_*^m \| \pi_{\theta_{k+1}^m}\right) \\
&= \left\langle \log \frac{\pi_{\theta_{k+1}^m}}{\pi_{\theta_k^m}}, \pi_*^m \right\rangle \\
&= \left\langle \log \frac{\pi_{\theta_{k+1}^m}}{\pi_{\theta_k^m}}, \pi_*^m - \pi_{\theta_{k+1}^m} \right\rangle + KL\left(\pi_{\theta_{k+1}^m} \| \pi_{\theta_k^m}\right) \qquad (11)
\end{aligned}
$$

回顾 π_{k+1}^m （9）和$^{\pi_{\theta_{k+1}^m}}$ （10）的定义，我们有以下两个等式

$$
\left\langle \log \frac{\pi_{\theta_{k+1}^m}}{\pi_{\theta_k^m}}, \pi_{\theta_k^m} - \pi_{\theta_{k+1}^m} \right\rangle = \left\langle (\theta_{k+1}^m - \theta_k^m)^\top \phi, \pi_{\theta_k^m} - \pi_{\theta_{k+1}^m} \right\rangle, \qquad (12)
$$

$$
\left\langle \beta_k^{-1} Q_{\pi_{\theta_k}}, \pi_*^m - \pi_{\theta_k^m} \right\rangle = \left\langle \log \frac{\pi_{k+1}^m}{\pi_{\theta_k^m}}, \pi_*^m - \pi_{\theta_k^m} \right\rangle. \qquad (13)
$$

将这些结果（12）、（13）代入（11）的右端项，我们得到

$$
\begin{aligned}
& KL\left(\pi_*^m \| \pi_{\theta_k^m}\right) - KL\left(\pi_*^m \| \pi_{\theta_{k+1}^m}\right) \\
&= \left\langle \log \frac{\pi_{\theta_{k+1}^m}}{\pi_{k+1}^m}, \pi_*^m - \pi_{\theta_k^m} \right\rangle + \left\langle \beta_k^{-1} Q^{1:m}_{\pi_{\theta_k}}, \pi_*^m - \pi_{\theta_k^m} \right\rangle + \left\langle (\theta_{k+1}^m - \theta_k^m)^\top \phi, \pi_{\theta_k^m} - \pi_{\theta_{k+1}^m} \right\rangle + KL\left(\pi_{\theta_{k+1}^m} \| \pi_{\theta_k^m}\right) \\
&\geq \left\langle \log \frac{\pi_{\theta_{k+1}^m}}{\pi_{k+1}^m}, \pi_*^m - \pi_{\theta_k^m} \right\rangle + \left\langle \beta_k^{-1} A^m_{\pi_{\theta_k}}, \pi_*^m - \pi_{\theta_k^m} \right\rangle + \left\langle (\theta_{k+1}^m - \theta_k^m)^\top \phi, \pi_{\theta_k^m} - \pi_{\theta_{k+1}^m} \right\rangle + \frac{1}{2} \left\| \pi_{\theta_{k+1}^m} - \pi_{\theta_k^m} \right\|_1^2
\end{aligned}
$$

最后一行中：（1）根据多智能体优势函数的定义3.1 ，我们有

$$
\langle Q^{1:m}_{\pi_{\theta_k}} - A^m_{\pi_{\theta_k}}, \pi_*^m - \pi_{\theta_k^m} \rangle = 0,
$$

以及（2）信息论中的 Pinsker不等式给出了 KL-散度的下界。

$$
KL\left(\pi_{\theta_{k+1}^m} \| \pi_{\theta_k^m}\right) \geq \frac{1}{2} \left\| \pi_{\theta_{k+1}^m} - \pi_{\theta_k^m} \right\|_1^2,
$$

将这些结果代入表达式，即完成证明。 □

有了这些结果，我们可以给出主定理的证明。

定理的证明4.5。

证明。 根据引理 B.6，对 $s \sim \nu_*$ 和 $\mathbf{a} \sim \boldsymbol{\pi}_*$取期望，我们有 ve

$$
\frac{1}{\beta_k} \mathop{\mathbb{E}}_{\sigma_*} \left\langle A^m_{\pi_{\theta_k}}(s,\mathbf{a}^{1:m-1},\cdot), \pi_*^m(\cdot|s,\mathbf{a}^{1:m-1}) - \pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1}) \right\rangle
$$

$$
\begin{aligned}
&\leq \mathop{\mathbb{E}}_{\sigma_*}\left[KL\left(\pi_*^m(\cdot|s,\mathbf{a}^{1:m-1})\|\pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1})\right)-KL\left(\pi_*^m(\cdot|s,\mathbf{a}^{1:m-1})\|\pi_{\theta_{k+1}^m}(\cdot|s,\mathbf{a}^{1:m-1})\right)\right]\\
&-\mathop{\mathbb{E}}_{\sigma_*}\left\langle\log\frac{\pi_{\theta_{k+1}^m}(\cdot|s,\mathbf{a}^{1:m-1})}{\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1})},\pi_*^m(\cdot|s,\mathbf{a}^{1:m-1})-\pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1})\right\rangle\\
&-\frac{1}{2}\mathop{\mathbb{E}}_{\sigma_*}\left\|\pi_{\theta_{k+1}^m}(\cdot|s,\mathbf{a}^{1:m-1})-\pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1})\right\|_1^2\\
&-\mathop{\mathbb{E}}_{\sigma_*}\left\langle{\theta_{k+1}^m}^\top\phi(s,\mathbf{a}^{1:m-1},\cdot)-{\theta_k^m}^\top\phi(s,\mathbf{a}^{1:m-1},\cdot),\pi_{\theta_k^m}(\cdot|s,\mathbf{a}^{1:m-1})-\pi_{\theta_{k+1}^m}(\cdot|s,\mathbf{a}^{1:m-1})\right\rangle.
\end{aligned}
$$

与上一节类似，为简洁起见，在不造成符号混淆的情况下，以下推导中省略 $(s,\mathbf{a}^{1:m-1})$。将引理 B.5代入上述不等式后，可将其整理为

$$
\begin{aligned}
&\frac{1}{\beta_k}\mathop{\mathbb{E}}_{\sigma_*}\left\langle A_{\pi_{\theta_k}}^m,\pi_*^m-\pi_{\theta_k^m}\right\rangle\\
&\leq \mathop{\mathbb{E}}_{\sigma_*}\left[KL\left(\pi_*^m\|\pi_{\theta_k^m}\right)-KL\left(\pi_*^m\|\pi_{\theta_{k+1}^m}\right)\right]\\
&\qquad\qquad-\mathop{\mathbb{E}}_{\sigma_*}\left[\left\langle\log\frac{\pi_{\theta_{k+1}^m}}{\pi_{k+1}^m},\pi_*^m-\pi_{\theta_k^m}\right\rangle+\frac{1}{2}\left\|\pi_{\theta_{k+1}^m}-\pi_{\theta_k^m}\right\|_1^2+\left\langle(\theta_{k+1}^m-\theta_k^m)^\top\phi,\pi_{\theta_k^m}-\pi_{\theta_{k+1}^m}\right\rangle\right]\\
&\leq \underbrace{\mathop{\mathbb{E}}_{\sigma_*}\left[KL\left(\pi_*^m\|\pi_{\theta_k^m}\right)-KL\left(\pi_*^m\|\pi_{\theta_{k+1}^m}\right)\right]}_{(i)}+\Delta_k^m\underbrace{-\mathop{\mathbb{E}}_{\sigma_*}\left[\frac{1}{2}\left\|\pi_{\theta_{k+1}^m}-\pi_{\theta_k^m}\right\|_1^2+\left\langle(\theta_{k+1}^m-\theta_k^m)^\top\phi,\pi_{\theta_k^m}-\pi_{\theta_{k+1}^m}\right\rangle\right]}_{(ii)}
\end{aligned}
\tag{14}
$$

对两边分别对 $m=1,\cdots,N$ 和 $k=0,\cdots,K-1$ 求和，注意到

左端项 这等于 $(1-\gamma)\sum_{k=0}^{K-1}\frac{1}{\beta_k}(J(\boldsymbol{\pi}_*)-J(\boldsymbol{\pi}_{\theta_k}))$，由性能差异引理可得（参见引理 4.1）。

右端项 (i)：对 k 和 m 求和后，由于对于任意 $m\in\mathcal{N}$，我们有 $\mathop{\mathbb{E}}_{\sigma_*}KL\left(\pi_*^m\|\pi_{\theta_0}^m\right)\leq\log|\mathcal{A}|$，因此该项的上界为 $N\log\mathcal{A}$；这是因为初始策略 π_{θ_0} 在动作空间上均匀分布。

(ii)：根据 H¨older 不等式，我们有

$$
-\mathop{\mathbb{E}}_{\sigma_*}\left\langle(\theta_{k+1}^m-\theta_k^m)^\top\phi,\pi_{\theta_k^m}-\pi_{\theta_{k+1}^m}\right\rangle\leq\mathop{\mathbb{E}}_{\sigma_*}\left\|(\theta_{k+1}^m-\theta_k^m)^\top\phi\right\|_\infty\cdot\left\|\pi_{\theta_k^m}-\pi_{\theta_{k+1}^m}\right\|_1
$$

利用三角不等式，可将其上界估计为

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\sigma_*}\left\|(\theta_{k+1}^m-\theta_k^m)^\top\phi-\beta_k^{-1}\hat{Q}\right\|_\infty\cdot\left\|\pi_{\theta_k^m}-\pi_{\theta_{k+1}^m}\right\|_1+\mathop{\mathbb{E}}_{\sigma_*}\left\|\beta_k^{-1}\hat{Q}\right\|_\infty\cdot\left\|\pi_{\theta_k^m}-\pi_{\theta_{k+1}^m}\right\|_1\\
&\leq\delta_k^m+\mathop{\mathbb{E}}_{\sigma_*}\left\|\beta_k^{-1}\hat{Q}\right\|_\infty\cdot\left\|\pi_{\theta_k^m}-\pi_{\theta_{k+1}^m}\right\|_1,
\end{aligned}
$$

其中代入引理 B.4 以及：$\left\|\pi_{\theta_k^m}-\pi_{\theta_{k+1}^m}\right\|_1\leq\left\|\pi_{\theta_k^m}\right\|_1+\left\|\pi_{\theta_{k+1}^m}\right\|_1=2$。我们得到

$$
\begin{aligned}
&-\mathop{\mathbb{E}}_{\sigma_*}\left[\frac{1}{2}\left\|\pi_{\theta_{k+1}^m}-\pi_{\theta_k^m}\right\|_1^2\right]+\mathop{\mathbb{E}}_{\sigma_*}\left\|\beta_k^{-1}\hat{Q}\right\|_\infty\cdot\left\|\pi_{\theta_k^m}-\pi_{\theta_{k+1}^m}\right\|_1\\
&\leq\mathop{\mathbb{E}}_{\sigma_*}\frac{1}{2}\left\|\beta_k^{-1}\hat{Q}\right\|_\infty^2,\quad\text{because }\forall x,y\text{ it holds: }-\frac{1}{2}x^2+yx\leq\frac{1}{2}y^2.\\
&\leq\frac{B^2}{2\beta_k^2}
\end{aligned}
$$

最后，结合这些结果，我们重新整理（14）并得到

$$(1-\gamma)\sum_{k=0}^{K-1}\frac{1}{\beta_k}(J(\boldsymbol{\pi}_*)-J(\boldsymbol{\pi}_{\theta_k}))\leq N\log|\mathcal{A}|+\sum_{m=1}^{N}\sum_{k=0}^{K-1}(\Delta_k^m+\delta_k^m)+\sum_{m=1}^{N}\sum_{k=0}^{K-1}\frac{B^2}{2\beta_k^2}.$$

设定惩罚参数 $\beta_k=\beta\sqrt{K}$，并注意到 $\bar{\boldsymbol{\pi}}$ 是从 $\boldsymbol{\pi}_{\theta_k}$，$k=1,\ 2\cdots K-1$ 中均匀采样的，我们有

$$J(\boldsymbol{\pi}_*)-J(\bar{\boldsymbol{\pi}})\leq\frac{N\beta^2\log|\mathcal{A}|+NB^2/2+\beta^2\sum_{m=1}^{N}\sum_{k=0}^{K-1}(\Delta_k^m+\delta_k^m)}{(1-\gamma)\beta\sqrt{K}}.$$

考虑到策略改进/评估误差，β 的最优选择是

$$\beta=\sqrt{\frac{NB^2/2}{N\log|\mathcal{A}|+\sum_{m=1}^{N}\sum_{k=0}^{K-1}(\Delta_k^m+\delta_k^m)}}$$

then

$$J(\boldsymbol{\pi}_*)-J(\bar{\boldsymbol{\pi}})\leq\mathcal{O}\left(\frac{B\sqrt{N}}{1-\gamma}\sqrt{\frac{N\log|\mathcal{A}|+\sum_{m=1}^{N}\sum_{k=0}^{K-1}(\Delta_k^m+\delta_k^m)}{K}}\right).$$

证明完成。 □

第 5 节的证明

C.1 计算效率

观察悲观评估

$$\omega_k^m\leftarrow\arg\min_{\omega}\left(f(s_0,\boldsymbol{\pi}_k^{1:m})+\lambda\mathcal{E}^{1:m}(f,\boldsymbol{\pi}_{\theta_k})\right).$$

在线性函数近似下，$f(s_0,\ \boldsymbol{\pi}^{1:m})$ 被实例化为 $\phi(s_0,\ \boldsymbol{\pi}^{1:m})^\top\omega$，于是

$$L^{1:m}(f',f,\boldsymbol{\pi})=\frac{1}{n}\sum_{\mathcal{D}^m}\left(\phi(s,\mathbf{a}^{1:m})^\top\omega'-r-\gamma\phi(s',\boldsymbol{\pi}^{1:m})^\top\omega\right)^2,$$

因此，我们将 $\mathcal{E}^{1:m}(f,\ \boldsymbol{\pi})$ 定义为

$$\sum_{\mathcal{D}_m}(\phi(s,\mathbf{a}^{1:m})^\top\omega-r-\gamma\phi(s',\boldsymbol{\pi}^{1:m})^\top\omega)^2-\min_{\omega'}\sum_{\mathcal{D}_m}(\phi(s,\mathbf{a}^{1:m})^\top\omega'-r-\gamma\phi(s',\boldsymbol{\pi}^{1:m})^\top\omega)^2,$$

其中，求和是在来自 $\mathcal{D}^m$ 的样本（$s,\ \mathbf{a}^{1:m},\ r,\ s'$）上进行的。

因此，贝尔曼误差与价值函数参数 ω 呈二次型依赖关系，从而可以应用许多高效的数值求解器。

C.2 证明

线性函数近似直接意味着**可实现性**和**完备性**条件：对于任意 $m\in\mathcal{N},\boldsymbol{\pi}\in\Pi^m$,

$$\inf_{f\in\mathcal{F}^m}\sup_{\text{admissable }\nu}\|f-\mathcal{T}_{\boldsymbol{\pi}}^{1:m}f\|_{2,\nu}^2=0,$$

and

$$\sup_{f\in\mathcal{F}^m}\inf_{f\in\mathcal{F}^m}\|f'-\mathcal{T}_{\boldsymbol{\pi}}^{1:m}f\|_{2,\mu}^2=0.$$

这些条件成立，是因为我们假设状态-动作价值函数具有线性结构：$Q_{\boldsymbol{\pi}}^{1:m} \in \mathcal{F}^m$ （参见定义 5.1）。

首先，我们考察 Xie et al. 在线性函数近似中的集中分析 [，2021]。

引理 C.1。对于任意 $m \in \mathcal{N}$、$\boldsymbol{\pi} \in \Pi^m$，以至少 $1-\delta$ 的概率，有

$$\mathcal{E}^{1:m}(Q^{\boldsymbol{\pi}}, \pi) \leq \mathcal{O}\left(\frac{d \log \frac{nLR}{\delta}}{(1-\gamma)^2 n}\right) = \varepsilon_r.$$

引理 C.2。对于任意 $m \in \mathcal{N}$，$\boldsymbol{\pi} \in \Pi^m$，$f \in \mathcal{F}^m$ （参见定义 5.1），若 $\mathcal{E}^{1:m}(f, \boldsymbol{\pi}) \leq \varepsilon$，则以至少 $1-\delta$ 的概率，有

$$\left\|f - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f\right\|_{2,\mathcal{D}^m} \leq \mathcal{O}\left(\sqrt{\frac{d \log \frac{nLR}{\delta}}{(1-\gamma)^2 n}}\right) + \sqrt{\varepsilon}.$$

为简单起见，下面定义

$$R(s, \mathbf{a}^{1:m}) = \mathbb{E}_{\tilde{\mathbf{a}} \sim \tilde{\boldsymbol{\pi}}_k} r(s, \mathbf{a}^{1:m}, \tilde{\mathbf{a}}).$$

在以下引理中，我们证明，在算法 k 的每次迭代中，2都存在一个马尔可夫博弈 $\mathcal{M}_k$，其多智能体价值函数恰好等于 f_k^m，$m \in \mathcal{N}$。此外，$\mathcal{M}_k$ 的转移动态与原始 $\mathcal{M}$相同。我们有以下理论保证，用于控制 $\mathcal{M}$ 的奖励与 $\mathcal{M}_k$的奖励之间的差异。

引理C.3。在每次迭代 k 中，都存在一个马尔可夫博弈 $\mathcal{M}_k$，其动态与原始的 $\mathcal{M}$ 相同。设 $\mathcal{M}_k$ 的奖励函数为 R_k，则

$$\left\|R_k^{1:m}(s, \mathbf{a}^{1:m}) - R(s, \mathbf{a}^{1:m})\right\|_{2,\mathcal{D}^m}^2 \leq \varepsilon_r.$$

证明。令 $R_k = (I - \gamma \mathcal{P}) f_k$。这直接意味着

$$\begin{aligned} f_k^m &= \mathcal{T}_{\boldsymbol{\pi}_k, \mathcal{M}_k}^{1:m}(s.\mathbf{a}^{1:m}) \\ &= R_k^{1:m}(s, \mathbf{a}^{1:m}) + \gamma \mathbb{E} f_k^m(s', \mathbf{a}'). \end{aligned}$$

因此

$$\left\|R_k^{1:m}(s, \mathbf{a}^{1:m}) - R(s, \mathbf{a}^{1:m})\right\|_{2,\mathcal{D}^m}^2 = \|f_k^m - \mathcal{T}_{\boldsymbol{\pi}_k}^{1:m} f_k^m\|_{2,\mathcal{D}^m}^2 \leq \varepsilon_r$$

□

引理 C.4。对于任意条件策略 $\pi: \mathcal{S} \times \mathcal{A}^{m-1} \to \Delta(\mathcal{A})$，以及 $s \in \mathcal{S}, \mathbf{a}^{1:m-1} \in \mathcal{A}^{m-1}$,

$$\begin{aligned} &\sum_{k=1}^{K} \left\langle \pi_{k+1}(\cdot|s, \mathbf{a}^{1:m-1}), f_k^m(s, \mathbf{a}^{1:m-1}, \cdot) \right\rangle - \ell_{sa}(\pi_1(\cdot|s, \mathbf{a}^{1:m-1})) \\ \geq &\sum_{k=1}^{K} \left\langle \pi(\cdot|s, \mathbf{a}^{1:m-1}), f_k^m(s, \mathbf{a}^{1:m-1}, \cdot) \right\rangle - \ell_{sa}(\pi) \end{aligned}$$

其中我们假设 $\ell_{sa}(\pi) = \frac{1}{\eta} \sum_{a \in \mathcal{A}} \pi(a|s, \mathbf{a}^{1:m-1}) \cdot \log \pi(a|s, \mathbf{a}^{1:m-1})$

证明。只需注意到这一事实 $\theta_{k+1}^m = \theta_k^m + \eta w_k^m$，证明显而易见。 □

引理 C.5。 对于任意条件策略 π： $\mathcal{S}\times\mathcal{A}^{m-1}\to\Delta(\mathcal{A})$，以及 $s\in\mathcal{S}$,$\mathbf{a}^{1:m-1}\in\mathcal{A}^{m-1}$,

$$\begin{aligned}&\sum_{k=1}^{K}\left\langle\pi(\cdot|s,\mathbf{a}^{1:m-1})-\pi_k(\cdot|s,\mathbf{a}^{1:m-1}),f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\rangle\\ \leq&\sum_{k=1}^{K}\left\langle\pi_{k+1}(\cdot|s,\mathbf{a}^{1:m-1})-\pi_k(\cdot|s,\mathbf{a}^{1:m-1}),f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\rangle-\ell_{sa}(\pi_1(\cdot|s,\mathbf{a}^{1:m-1}))\end{aligned}$$

证明。 这些结果可以通过应用引理C.4获得。 □

引理 C.6。 对于任意条件策略 π： $\mathcal{S}\times\mathcal{A}^{m-1}\to\Delta(\mathcal{A})$，且 $s\in\mathcal{S}$，$\mathbf{a}^{1:m-1}\in\mathcal{A}^{m-1}$，若将步长设为 $\eta=\sqrt{\frac{\log|\mathcal{A}|}{2K}}$

$$\sum_{k=1}^{K}\left\langle\pi(\cdot|s,\mathbf{a}^{1:m-1})-\pi_k(\cdot|s,\mathbf{a}^{1:m-1}),f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\rangle\leq 2\sqrt{2\log|\mathcal{A}|K}$$

证明。 定义 $\mathcal{L}_{sa,k}^m=\sum_{k'=1}^{k}\langle\pi(\cdot|s,\mathbf{a}^{1:m-1})\ f_{k'}(s,\mathbf{a}^{1:m-1},\cdot)\rangle-\ell_{sa}(\pi)$。令 $B_{\mathcal{L}_{sa,k}^m}(\cdot\|\cdot)$ 和$B_{\ell_{sa}}(\cdot\|\cdot)$ 分别表示关于损失 $\mathcal{L}_{sa,k}^m$ 和 ℓ_{sa}的 Bregman 散度。利用散度的性质，我们有

$$\begin{aligned}\mathcal{L}_{sa,k}^m(\pi_k^m)&\leq\mathcal{L}_{sa,k}^m(\pi_{k+1}^m)+B_{\mathcal{L}_{sa,k}^m}(\pi_k^m(\cdot|s,\mathbf{a}^{1:m-1})\|\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1}))\\&=\mathcal{L}_{sa,k}^m(\pi_{k+1}^m)-B_{\ell_{sa}}(\pi_k^m(\cdot|s,\mathbf{a}^{1:m-1})\|\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1})).\end{aligned}$$

重新整理后，我们得到

$$B_{\ell_{sa}}(\pi_k^m(\cdot|s,\mathbf{a}^{1:m-1})\|\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1}))\leq\mathcal{L}_{sa,k}^m(\pi_{k+1}^m)-\mathcal{L}_{sa,k}^m(\pi_k^m).$$

上述表达式的右端项不大于

$$\left\langle\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1})-\pi_k^m(\cdot|s,\mathbf{a}^{1:m-1}),f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\rangle.$$

Then

$$\begin{aligned}&\left\langle\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1})-\pi_k^m(\cdot|s,\mathbf{a}^{1:m-1}),f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\rangle\\\leq&\sqrt{2\eta B_{\ell_{sa}}(\pi_k^m(\cdot|s,\mathbf{a}^{1:m-1})\|\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1}))}\cdot\left\|f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\|_\infty\\\leq&\sqrt{2\eta}\sqrt{\langle\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1})-\pi_k^m(\cdot|s,\mathbf{a}^{1:m-1}),f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\rangle}\frac{1}{1-\gamma}.\end{aligned}$$

因此我们有 $\left\langle\pi_{k+1}^m(\cdot|s,\mathbf{a}^{1:m-1})-\pi_k^m(\cdot|s,\mathbf{a}^{1:m-1})\ f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\rangle\leq\frac{2\eta}{(1-\gamma)^2}$。将其代入引理C.5 可得

$$\begin{aligned}&\sum_{k=1}^{K}\left\langle\pi(\cdot|s,\mathbf{a}^{1:m-1})-\pi_k(\cdot|s,\mathbf{a}^{1:m-1}),f_k^m(s,\mathbf{a}^{1:m-1},\cdot)\right\rangle\\\leq&\frac{2\eta K}{(1-\gamma)^2}+\frac{\log|\mathcal{A}|}{\eta}\\\leq&\frac{2\sqrt{2\log|\mathcal{A}|K}}{1-\gamma}\end{aligned}$$

其中，最后一行是通过设置 $\eta=(1-\gamma)\sqrt{\frac{\log|\mathcal{A}|}{2K}}$ 得到的。 □

引理 C.7. 对于任意条件策略 $\pi: \mathcal{S} \times \mathcal{A}^{m-1} \to \Delta(\mathcal{A})$,

$$Q_{\boldsymbol{\pi}}^{1:m}(s_0, \boldsymbol{\pi}^{1:m}) \geq \min_{f \in \mathcal{F}^m} \left(f(s_0, \boldsymbol{\pi}_k^{1:m}) + \lambda \mathcal{E}^{1:m}(f, \boldsymbol{\pi}_k) \right) - \lambda \varepsilon_r.$$

证明。 对于任意条件策略 $\boldsymbol{\pi}^{1:m}$，以及 $\forall m \in \mathcal{N}$。在 **可实现性** 假设下，我们有 $Q_{\boldsymbol{\pi}}^{1:m} = \arg\min_f \sup_\nu \|f - \mathcal{T}^{1:}{}_{\boldsymbol{\pi}}{}^m f\|_{2,\nu}^2$ 对于任意可行的 ν

$$\begin{aligned}
J(\boldsymbol{\pi}) &= Q_{\boldsymbol{\pi}}^{1:m}(s_0, \boldsymbol{\pi}^{1:m}) \\
&= Q_{\boldsymbol{\pi}}^{1:m}(s_0, \boldsymbol{\pi}^{1:m}) - \left(Q_{\boldsymbol{\pi}}^{1:m}(s_0, \boldsymbol{\pi}^{1:m}) + \lambda \mathcal{E}^{1:m}(Q_{\boldsymbol{\pi}}^{1:m}, \boldsymbol{\pi}) \right) + \left(Q_{\boldsymbol{\pi}}^{1:m}(s_0, \boldsymbol{\pi}^{1:m}) + \lambda \mathcal{E}^{1:m}(Q_{\boldsymbol{\pi}}^{1:m}, \boldsymbol{\pi}) \right) \\
&\geq \min_{f \in \mathcal{F}^m} \left(f(s_0, \boldsymbol{\pi}^{1:m}) + \lambda \mathcal{E}^{1:m}(f, \boldsymbol{\pi}) \right) - \lambda \varepsilon_r,
\end{aligned}$$

其中，最后一行使用了引理 C.1。 □

定理的证明2。

证明。 使用引理C.7，在第 k 次迭代时，我们有

$$\begin{aligned}
J(\boldsymbol{\pi}_k) &\geq \min_{f \in \mathcal{F}^m} \left(f(s_0, \boldsymbol{\pi}_k^{1:m}) + \lambda \mathcal{E}^{1:m}(f, \boldsymbol{\pi}_k) \right) - \lambda \varepsilon_r \\
&\geq f_k^m(s_0, \boldsymbol{\pi}_k^{1:m}) - \lambda \varepsilon_r \\
&= J_{\mathcal{M}_k}(\boldsymbol{\pi}_k) - \lambda \varepsilon_r
\end{aligned}$$

其中， f_k^m是马尔可夫博弈 $\mathcal{M}_k$ 的多智能体价值函数（参见引理 C.3）。因此，我们有:

$$\begin{aligned}
J(\boldsymbol{\pi}_*) - J(\bar{\boldsymbol{\pi}}) &= \frac{1}{K} \sum_{k=1}^{K} (J(\boldsymbol{\pi}_*) - J(\boldsymbol{\pi}_k)) \\
&\leq \frac{1}{K} \sum_{k=1}^{K} (J(\boldsymbol{\pi}_*) - J_{\mathcal{M}_k}(\boldsymbol{\pi}_k)) + \lambda \varepsilon_r \\
&\leq \underbrace{\frac{1}{K} \sum_{k=1}^{K} (J_{\mathcal{M}_k}(\boldsymbol{\pi}_*) - J_{\mathcal{M}_k}(\boldsymbol{\pi}_k))}_{\text{I}} + \underbrace{\frac{1}{K} \sum_{k=1}^{K} (J(\boldsymbol{\pi}_*) - J_{\mathcal{M}_k}(\boldsymbol{\pi}_*))}_{\text{II}} + \lambda \varepsilon_r
\end{aligned}$$

术语 I. 应用性能差异引理，我们有

$$\begin{aligned}
&\frac{1}{K} \sum_{k=1}^{K} (J_{\mathcal{M}_k}(\boldsymbol{\pi}_*) - J_{\mathcal{M}_k}(\boldsymbol{\pi}_k)) \\
&= \frac{1}{1-\gamma} \frac{1}{K} \sum_{m=1}^{N} \mathbb{E}_{s \sim d_{\boldsymbol{\pi}_*}} \mathbb{E}_{\mathbf{a}^{1:m-1}} \sum_{k=1}^{K} \left(Q_{\mathcal{M}_k}^{\boldsymbol{\pi}_k}(s, \mathbf{a}^{1:m-1}, \boldsymbol{\pi}_*^m) - Q_{\mathcal{M}_k}^{\boldsymbol{\pi}_k}(s, \mathbf{a}^{1:m-1}, \boldsymbol{\pi}_k^m) \right) \\
&= \frac{1}{1-\gamma} \frac{1}{K} \sum_{m=1}^{N} \mathbb{E}_{s \sim d_{\boldsymbol{\pi}_*}, \mathbf{a}^{1:m-1} \sim \boldsymbol{\pi}_*} \sum_{k=1}^{K} \left\langle \boldsymbol{\pi}_*^m(\cdot | s, \mathbf{a}^{1:m-1}) - \boldsymbol{\pi}_k^m(\cdot | s, \mathbf{a}^{1:m-1}), f_k^m(s, \mathbf{a}^{1:m-1}, \cdot) \right\rangle
\end{aligned}$$

其中，最后一行是因为 f_k^m是 $\mathcal{M}_k$的多智能体状态-动作价值函数。然后，如果 $\eta = (1-\gamma)\sqrt{\frac{\log|\mathcal{A}|}{2K}}$，引理 C.6 给出

$$\text{Term I} \leq \frac{2N}{(1-\gamma)^2} \sqrt{\frac{2 \log |\mathcal{A}|}{K}}.$$

第II项。 以下分析对任意 $m \in \mathcal{N}$ 均成立，因此为简洁起见，我们省略 m 。对于这一项，我们再次使用引理 1 [Xie 和 Jiang, 2020] 将其转化为状态-动作分布上的范数

$$J(\boldsymbol{\pi}_*) - J_{\mathcal{M}_k}(\boldsymbol{\pi}_*) = Q_{\boldsymbol{\pi}_*}(s, \boldsymbol{\pi}_*) - J_{\mathcal{M}_k}(\boldsymbol{\pi}_*) \leq \frac{1}{1-\gamma} \|Q_{\boldsymbol{\pi}_*} - \mathcal{T}_{\boldsymbol{\pi}_*, \mathcal{M}_k} Q_{\boldsymbol{\pi}_*}\|_{2, d_{\boldsymbol{\pi}_*}}$$

注意辅助马尔可夫博弈的定义，其中 f^k 是其价值函数（参见引理C.3），我们有

$$\begin{aligned} & \frac{1}{1-\gamma} \|Q_{\boldsymbol{\pi}_*} - R_k - \gamma P_{\boldsymbol{\pi}_k} Q_{\boldsymbol{\pi}_*}\|_{2, d_{\boldsymbol{\pi}_*}} \qquad R_k = (I - \gamma \mathcal{P}) f_k \\ = & \frac{1}{1-\gamma} \|f_k - \mathcal{T}_{\boldsymbol{\pi}_k} f_k\|_{2, d_{\boldsymbol{\pi}_*}} \\ \leq & \frac{\mathcal{C}_\mu^{d_{\boldsymbol{\pi}_*}}}{1-\gamma} \|f_k - \mathcal{T}_{\boldsymbol{\pi}_k} f_k\|_{2, \mathcal{D}} \end{aligned}$$

它不大于 $\frac{\mathcal{C}_\mu^{d_{\boldsymbol{\pi}_*}}}{1-\gamma} \left(\sqrt{\varepsilon_r + \sqrt{\lambda^{-1}(1-\gamma)}} \right)$，因为 $\mathcal{E}(f_k, \boldsymbol{\pi}_k) \leq \varepsilon_r + \lambda^{-1}(1-\gamma)\lambda$：

$$\begin{aligned} f_k(s_0, \boldsymbol{\pi}_k) + \lambda \mathcal{E}(f_k, \boldsymbol{\pi}_k) &= \min_f \left(f(s_0, \boldsymbol{\pi}_k) + \lambda \mathcal{E}(f, \boldsymbol{\pi}_k) \right) \\ &\leq Q_{\boldsymbol{\pi}_k}(s_0, \boldsymbol{\pi}_k) + \lambda \mathcal{E}(Q_{\boldsymbol{\pi}_k}, \boldsymbol{\pi}_k), \quad \text{Lemma C.2} \\ &\leq \frac{1}{1-\gamma} + \lambda \varepsilon_r. \end{aligned}$$

将 λ 代入原表达式即可完成证明。 □

D 基于一般函数逼近的悲观多智能体近端策略优化

本节将线性函数近似的结果推广到一般函数逼近（参见第5节）。

算法 4 基于一般函数逼近的悲观多智能体近端策略优化 输入：正则化系数 λ。输出：从 0, $1 \cdots K-1$ 中均匀采样 k，返回 $\bar{\pi} = \boldsymbol{\pi}_k$。1：均匀初始化：对每个 $m \in \mathcal{N}$，设置 $\theta_0^m = 0$。2：对于 $k = 0$=1，...，$K-1$，执行 3：对于 $m = 1$ =2，$\cdots$，...，N，执行 4：求得悲观估计：$f_k^m \leftarrow \arg\min_{f \in \mathcal{F}^m} \left(f(s_0, \boldsymbol{\pi}_k^{1:m}) + \lambda \mathcal{E}^{1:m}(f, \boldsymbol{\pi}_k) \right)$。5：策略改进：对于任意 $(s, \mathbf{a}^{1:m}) \in \mathcal{S} \times \mathcal{A}^m$，6：结束循环 7：结束循环

$$\pi_{k+1}^m(a^m | s, \mathbf{a}^{1:m-1}) \propto \pi_{k+1}^m(a^m | s, \mathbf{a}^{1:m-1}) \cdot \exp\left(\eta f_k^m(s, \mathbf{a}^{1:m-1}, a^m)\right).$$

在此设定中，价值函数在有限集合 $\mathcal{F}^m$ 上进行搜索。我们对一般函数类施加以下正则性条件

假设 D.1. 对于任意 $m \in \mathcal{N}$, $f \in \mathcal{F}^m$ 和 $(s, \mathbf{a}^{1:m}) \in \mathcal{S} \times \mathcal{A}^m$，均有

$$|f(s, \mathbf{a}^{1:m-1}, a^m)| \leq \frac{1}{1-\gamma}, \tag{15}$$

$$|\mathcal{F}^m| \leq |\mathcal{F}| \tag{16}$$

其中，$|\mathcal{F}|$ 是一个确定的正数。

与预先定义的固定策略类 Π 不同，现在策略改进是在 $\mathcal{F}^m$ 的基础上进行的，形式化地，对于 $m \in \mathcal{N}$ 和 $(s, \mathbf{a}^{1:m-1}) \in \mathcal{S} \times \mathcal{A}^{m-1}$

$$\Pi^m = \left\{ \pi^m(\cdot|s, \mathbf{a}^{1:m-1}) \propto \exp\left(\eta \sum_{j=0}^{k} f_j(s, \mathbf{a}^{1:m-1}, \cdot) \right) : f_j \in \mathcal{F}^m, 0 \leq k \leq K-1 \right\}. \tag{17}$$

另请注意，在一般函数逼近下，策略改进必须针对每个 $(s,\mathbf{a}^{1:m})$ 分别进行，而状态空间极其庞大时，这可能会带来麻烦。

对于一般函数逼近，通常需要对 $\mathcal{F}$ 作出两项表达能力假设 [Antos 等人，2008年，Xie 等人， 2021]。

假设 D.2 （可实现性）。对于任意 $m \in \mathcal{N}$, $\boldsymbol{\pi} \in \Pi^m$,

$$\inf_{f \in \mathcal{F}^m} \sup_{\text{admissable } \nu} \|f - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f\|_{2,\nu}^2 = \zeta_{\mathcal{F}}.$$

其中，ν 可以是 $\mathcal{S} \times \mathcal{A}^{1:m}$上的任意可接受分布，并且

假设 D.3 （完备性）。对于任意 $m \in \mathcal{N}$、 $\boldsymbol{\pi} \in \Pi^m$,

$$\sup_{f \in \mathcal{F}^m} \inf_{f \in \mathcal{F}^m} \|f' - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f\|_{2,\mu}^2 = \zeta'_{\mathcal{F}}.$$

对于一般函数逼近，我们有以下集中保证 [Xie 等人， 2021]。

引理 D.4. 对于任意 $m \in \mathcal{N}$, $\boldsymbol{\pi} \in \Pi^m$，令

$$f_{\boldsymbol{\pi}}^{1:m} = \arg\min_{f \in \mathcal{F}^m} \sup_{\text{admissable } \nu} \|f - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f\|_{2,\nu}^2,$$

其成立的概率至少为 $1-\delta$

$$\begin{aligned} \mathcal{E}^{1:m}(f_{\boldsymbol{\pi}}^{1:m}, \pi) &\leq \mathcal{O}\left(\frac{\log \frac{|\mathcal{F}^m||\Pi^m|}{\delta}}{n(1-\gamma)^2} + \zeta_{\mathcal{F}} \right) \\ &= \mathcal{O}\left(\frac{K \log \frac{|\mathcal{F}^m|}{\delta}}{n(1-\gamma)^2} + \zeta_{\mathcal{F}} \right) = \varepsilon_r, \end{aligned} \tag{18}$$

其中我们注意到 $|\Pi^m| \leq |\mathcal{F}^m|^K$ 来自 (17)。

引理 D.5。 对于任意 $m \in \mathcal{N}$、 $\boldsymbol{\pi} \in \Pi^m$、 $f \in \mathcal{F}^m$，若 $\mathcal{E}^{1:m}(f, \boldsymbol{\pi}) \leq \varepsilon$，则以至少 $1-\delta$ 的概率有

$$\left\|f - \mathcal{T}_{\boldsymbol{\pi}}^{1:m} f\right\|_{2,\mathcal{D}^m} \leq \mathcal{O}\left(\frac{1}{1-\gamma} \sqrt{\frac{K \log \frac{|\mathcal{F}^m|}{\delta}}{n}} \right) + \sqrt{\zeta'_{\mathcal{F}}} + \sqrt{\zeta'_{\mathcal{F}} + \varepsilon}.$$

引理 D.6。 对于任意条件策略 $\pi:\ \mathcal{S}\times\mathcal{A}^{m-1}\to\Delta(\mathcal{A})$,

$$Q_{\boldsymbol{\pi}}^{1:m}(s_0,\boldsymbol{\pi}^{1:m})\geq\min_{f\in\mathcal{F}^m}\left(f(s_0,\boldsymbol{\pi}_k^{1:m})+\lambda\mathcal{E}^{1:m}(f,\boldsymbol{\pi}_k)\right)-\frac{\sqrt{\zeta_{\mathcal{F}}}}{1-\gamma}-\lambda\varepsilon_r.$$

证明。 对于任意条件策略 $\boldsymbol{\pi}^{1:m}\quad\forall m\in\mathcal{N}\quad f_{\boldsymbol{\pi}}^{1:m}=\quad f\quad\nu\|f-,$ $\mathcal{T}_{\boldsymbol{\pi}}^{1\ m}f\|_{2\ \nu}^2\qquad\nu$ 以及 。令 arg min sup: ， 对于任意可行的

$$\begin{aligned}
J(\boldsymbol{\pi})&=Q_{\boldsymbol{\pi}}^{1:m}(s_0,\boldsymbol{\pi}^{1:m})\\
&=Q_{\boldsymbol{\pi}}^{1:m}(s_0,\boldsymbol{\pi}^{1:m})-\left(f_{\boldsymbol{\pi}}^{1:m}(s_0,\boldsymbol{\pi}^{1:m})+\lambda\mathcal{E}^{1:m}(Q_{\boldsymbol{\pi}}^{1:m},\boldsymbol{\pi})\right)+\left(f_{\boldsymbol{\pi}}^{1:m}(s_0,\boldsymbol{\pi}^{1:m})+\lambda\mathcal{E}^{1:m}(f_{\boldsymbol{\pi}}^{1:m},\boldsymbol{\pi})\right)\\
&\geq\min_{f\in\mathcal{F}^m}\left(f(s_0,\boldsymbol{\pi}^{1:m})+\lambda\mathcal{E}^{1:m}(f,\boldsymbol{\pi})\right)-\frac{\sqrt{\zeta_{\mathcal{F}}}}{1-\gamma}-\lambda\varepsilon_r,
\end{aligned}$$

其中在最后一行中，我们使用了引理 C.1 和可实现性假设

$$f_{\boldsymbol{\pi}}^{1:m}-Q_{\boldsymbol{\pi}}^{1:m}\leq\frac{\|f_{\boldsymbol{\pi}}^{1:m}-\mathcal{T}_{\boldsymbol{\pi}}^{1:m}f_{\boldsymbol{\pi}}^{1:m}\|_{2,d_{\boldsymbol{\pi}}}}{1-\gamma}\leq\frac{\sqrt{\zeta_{\mathcal{F}}}}{1-\gamma}.$$

□

有了这些有用的引理，我们现在可以继续讨论一般函数逼近情形下的主定理

定理 D.7。 回顾 ε_r (18) 的定义，对于完全合作马尔可夫博弈中由算法 4得到的输出策略 $\bar{\boldsymbol{\pi}}$，令 $\eta=(1-\gamma)\sqrt{\log_{2K}|\mathcal{A}|}$，且 $\lambda=(1-\gamma)^{-1}\varepsilon_r^{-\frac{2}{3}}$。经过 K 次迭代后，以至少 $1-\delta$ 的概率，我们有 -

$$J(\boldsymbol{\pi}_*)-J(\bar{\boldsymbol{\pi}})\leq\mathcal{O}\left(\frac{N}{(1-\gamma)^2}\sqrt{\frac{\log|\mathcal{A}|}{K}}+\frac{\mathcal{C}_\mu^{d_{\boldsymbol{\pi}_*}}}{1-\gamma}\cdot\left(\frac{1}{1-\gamma}\sqrt[3]{\frac{K\log\frac{|\mathcal{F}|}{n}}{n}}+\sqrt{\zeta_{\mathcal{F}}+\zeta'_{\mathcal{F}}}+\sqrt[3]{\zeta_{\mathcal{F}}}\right)\right)$$

证明。 使用引理D.6，在第 k 次迭代时，我们有

$$\begin{aligned}
J(\boldsymbol{\pi}_k)&\geq\min_{f\in\mathcal{F}^m}\left(f(s_0,\boldsymbol{\pi}_k^{1:m})+\lambda\mathcal{E}^{1:m}(f,\boldsymbol{\pi}_k)\right)-\frac{\sqrt{\zeta_{\mathcal{F}}}}{1-\gamma}-\lambda\varepsilon_r\\
&\geq f_k^m(s_0,\boldsymbol{\pi}_k^{1:m})-\frac{\sqrt{\zeta_{\mathcal{F}}}}{1-\gamma}-\lambda\varepsilon_r\\
&=J_{\mathcal{M}_k}(\boldsymbol{\pi}_k)-\frac{\sqrt{\zeta_{\mathcal{F}}}}{1-\gamma}-\lambda\varepsilon_r
\end{aligned}$$

与附录 C类似，我们有

$$\begin{aligned}
J(\boldsymbol{\pi}_*)-J(\bar{\boldsymbol{\pi}})&=\frac{1}{K}\sum_{k=1}^{K}(J(\boldsymbol{\pi}_*)-J(\boldsymbol{\pi}_k))\\
&\leq\frac{1}{K}\sum_{k=1}^{K}(J(\boldsymbol{\pi}_*)-J_{\mathcal{M}_k}(\boldsymbol{\pi}_k))+\frac{\sqrt{\zeta_{\mathcal{F}}}}{1-\gamma}+\lambda\varepsilon_r\\
&\leq\underbrace{\frac{1}{K}\sum_{k=1}^{K}(J_{\mathcal{M}_k}(\boldsymbol{\pi}_*)-J_{\mathcal{M}_k}(\boldsymbol{\pi}_k))}_{\text{I}}+\underbrace{\frac{1}{K}\sum_{k=1}^{K}(J(\boldsymbol{\pi}_*)-J_{\mathcal{M}_k}(\boldsymbol{\pi}_*))}_{\text{II}}+\frac{\sqrt{\zeta_{\mathcal{F}}}}{1-\gamma}+\lambda\varepsilon_r
\end{aligned}$$

项 I. 关于优化项的分析与附录 C 中的分析相同。如果$\eta=(1-\gamma)\sqrt{\log_{2K}|\mathcal{A}|}$，引理 C.6 给出

$$\text{Term I}\leq\frac{2N}{(1-\gamma)^2}\sqrt{\frac{2\log|\mathcal{A}|}{K}}.$$

第II项。 与附录类似C

$$\begin{aligned} J(\boldsymbol{\pi}_*) - J_{\mathcal{M}_k}(\boldsymbol{\pi}_*) &= Q_{\boldsymbol{\pi}_*}(s, \boldsymbol{\pi}_*) - J_{\mathcal{M}_k}(\boldsymbol{\pi}_*) \\ &\leq \frac{\mathcal{C}_\mu^{d_{\boldsymbol{\pi}_*}}}{1-\gamma} \left\| f_k - \mathcal{T}_{\boldsymbol{\pi}_k} f_k \right\|_{2,\mathcal{D}} \end{aligned}$$

不大于

$$\frac{\mathcal{C}_\mu^{d_{\boldsymbol{\pi}_*}}}{1-\gamma} \left(\mathcal{O}\left(\frac{1}{1-\gamma} \sqrt{\frac{K \log \frac{|\mathcal{F}^m|}{\delta}}{n}} \right) + \sqrt{\zeta'_{\mathcal{F}}} + \sqrt{\zeta'_{\mathcal{F}} + \varepsilon_r} + \sqrt{\frac{1}{(1-\gamma)\lambda}} \right),$$

因为我们有 $\mathcal{E}(f_k, \boldsymbol{\pi}_k) \leq \varepsilon_r + {}^{1}\!/_{(1-\gamma)\lambda}$：

$$\begin{aligned} f_k(s_0, \boldsymbol{\pi}_k) + \lambda \mathcal{E}(f_k, \boldsymbol{\pi}_k) &= \min_f \left(f(s_0, \boldsymbol{\pi}_k) + \lambda \mathcal{E}(f, \boldsymbol{\pi}_k) \right) \\ &\leq Q_{\boldsymbol{\pi}_k}(s_0, \boldsymbol{\pi}_k) + \lambda \mathcal{E}(Q_{\boldsymbol{\pi}_k}, \boldsymbol{\pi}_k), \quad \text{Lemma C.2} \\ &\leq \frac{1}{1-\gamma} + \lambda \varepsilon_r. \end{aligned}$$

将 λ 代入原始表达式，即可完成证明。 □

E 仿真

在本节中，我们通过一个简单示例来展示，与朴素的独立策略梯度更新相比，我们的序贯更新结构具有优越性。我们考虑 von Neumann 比率博弈，这是一种简单的随机博弈， [Daskalakis 等人也 2020]使用过。

在这个博弈中，只有两个智能体，且每个智能体的动作空间大小为 2。这里只有一个状态，即不存在状态转移。选择动作 (a, b) 时的即时奖励为 $R(a, b)$，每轮停止的概率为 $S(a, b)$。此博弈的价值函数$V(\pi_x,\pi_y)$ 为】【。

$$V = \frac{\pi_x^\top R \pi_y}{\pi_x^\top S \pi_y}.$$

两个智能体相互协作，以最大化价值函数。从现在起，我们用 $(x, \ 1-x)$ 和 $(y, \ 1-y)$ 表示这两个策略。我们将参数设为

$$R = \begin{bmatrix} 1 & 0.5 \\ -0.5 & 1 \end{bmatrix}, \quad \text{and} \quad R = \begin{bmatrix} 1 & 1 \\ 0.1 & 0.1 \end{bmatrix}.$$

将价值函数视为变量 x 和 y 的函数，则驻点位于 $(x, \ y) = (0.5, \ 0)$ 附近，此处的价值函数为 $V \approx 0.46$，小于全局最大值 $V = 1$。

为了解决这一问题，我们采用两种算法：(1) 我们的顺序更新算法；(2) 独立的（策略梯度）学习方法。在这两种算法中，我们都使用 softmax 参数化来表示策略。具体而言，通过将 2 中的对数线性参数化设置为 ϕ 的独热表示，即对于动作 a， $\phi(a)^\top \theta = \theta_a$ 其中 θ_a 表示 θ 中与 a 对应的特定条目。

结果我们在不同设置下，使用顺序梯度更新和独立学习方法测试我们的算法。结果如图1所示。 [5] 下面我们将讨论这项模拟研究的实证发现。

首先，我们发现，独立策略优化方法在驻点附近往往会遇到困难（见 (a)-(c)），而这些驻点不一定是全局最优点。在这个例子中，较大的步长有助于缓解这一问题（例如，在 (c) 中，独立 PG 在经过 3000 次迭代后逃离了驻点）。然而，与我们的方法相比，其收敛到全局最优解的速度仍然较慢。我们注意到，噪声可能有助于逃离驻点 [Jin 等人，2017]。我们的发现与前文所做的理论比较结果一致。即使独立 PG 方法没有陷入驻点，从 (d)-(f) 可以看出，在最大化价值函数方面，我们的算法始终表现更优。

在这个玩具示例中，优化景观相当简单：只有两个智能体参与，每个智能体仅有两种可能的动作。不允许发生状态转移，而这正是执行序贯决策时面临的主要难点。我们指出，从全局来看，该问题只有一个驻点。在这种相对简单的情况下，我们的算法在两个主要方面始终优于独立策略梯度（PG）。首先，当当前策略接近梯度信息较少的驻点时，我们的算法不像独立策略梯度那样难以继续优化。其次，我们的算法表现出较快的收敛速度，能够迅速达到全局最大价值函数。因此，当环境的复杂性显著增加时，例如：（1）多个异构智能体相互作用，并与未知环境进行交互；（2）采用复杂的函数逼近器（例如深度神经网络），使用独立策略梯度在寻找全局最优解方面将面临更大的困难，因为优化景观中会出现更多驻点。

我们的研究结果展现了条件依赖结构的必要性和实用性，它有助于我们找到一种策略，使其价值函数在全局范围内达到次优。

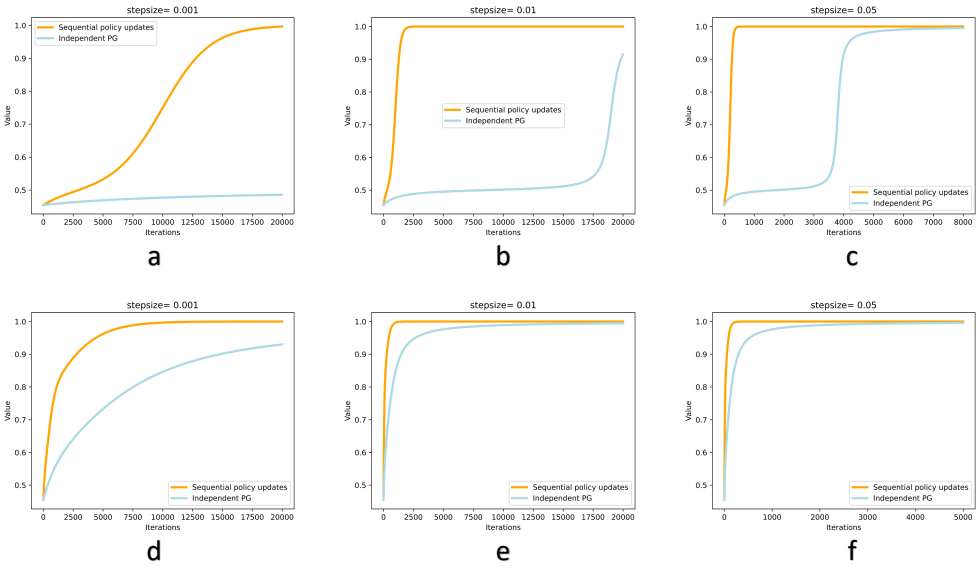

图 1：我们的算法与独立学习方法的表现。(a)–(c) 中：策略初始化在驻点附近。(d)–(f) 中：两种策略均采用均匀初始化。

[5] 实现代码请见 https://github.com/zhaoyl18/ratio_游戏。