零和马尔可夫博弈中的可微仲裁

Jing Wang*，†纽约大学，美国纽约
jw5665@nyu.edu

Meichen Song*，†石溪大学，美国纽约
m
eichen.song@stonybrook.edu

Feng Gao‡交叉信息研究院，清华大学，中国北京
feng.gao220@gmail.com

Boyi Liu 西北大学，美国伊利诺伊州
boyiliu2018@u.northwestern.edu

Zhaoran Wang 西北大学，美国伊利诺伊州
zhaoran.wang@northwestern.edu

Yi Wu 交叉信息研究院，清华大学；上海棋院
jxwuyi@gmail.com

arXiv:2302.10058v1 [cs.MA] 20 Feb 2023

摘要

我们首次研究如何扰动双人零和马尔可夫博弈中的奖励，以诱导出理想的纳什均衡，即进行仲裁。此类问题可表述为双层优化问题。下层需要在给定奖励函数的条件下求解纳什均衡，这使得整体问题难以通过端到端方式进行优化。我们提出了一种反向传播方案，通过纳什均衡进行微分，为上层提供梯度反馈。具体而言，我们的方法只需要一个黑盒求解器来求解（正则化的）纳什均衡（NE）。我们针对采用适当黑盒纳什均衡求解器的所提框架进行了收敛性分析，并在两个多智能体强化学习（MARL）环境中验证了其经验效果。

关键词

零和博弈；均衡精化；双层优化

ACM 引用格式：

Jing Wang*、MeichenSong*、†、FengGao‡、Boyi Liu、Zhaoran Wang 和 Yi Wu。2023 年。零和马尔可夫博弈中的可微仲裁。载于第 *22* 届国际自主智能体与多智能体系统会议（*AAMAS 2023*）论文集，英国伦敦，*2023*年*5*月*29*日至*6*月*2*日，IFAAMAS，29 页。

1 引言

在人类社会的发展过程中，调和自利与集体利益之间的冲突无处不在 [32]。我们将马尔可夫博弈 [53] 视为人类社会的抽象模型。在马尔可夫博弈中，纳什均衡（NE） [8], 指任何玩家都无法通过单方面改变其策略而获益，是一个重要概念。尽管由马尔可夫博弈建模的玩家会理性地行动以最大化自身奖励，但缺乏集体考量可能导致不理想的纳什均衡，从系统层面来看，这会损害整体福利 [23]。

激励设计 [73], 最初为多臂老虎机场景而提出，旨在通过扰动奖励来调整纳什均衡，从而化解此类冲突，使利己玩家达到理想的均衡，例如，实现更高的社会福利或更高的探索率。更具体地说，激励设计问题自然可建模为一种双层结构 [19, 58]。上层设计者旨在确定最优激励，而下层玩家则在激励扰动奖励下采取行动。为此，设计者需要预判下层玩家在马尔可夫博弈中对所设定激励的反应。然而，穿过双层结构求导十分困难 [27], , 尤其是在下层包含多个智能体时。具体而言，对于上层设计者来说，下层均衡似乎是一个黑箱，这使其难以评估所设定的激励将如何影响下层均衡。因此，上层设计者缺乏梯度反馈，无法有效更新其用于设定激励的策略 [51]。

已有一些研究尝试绕开求解双层优化时导数带来的难题。一种思路是将问题转化为多目标问题，从而摆脱双层问题结构：也就是说，低层玩家通过探索多个纳什均衡，同时优化博弈奖励和设计者的目标 [15, 22, 38, 49, 71, 81]。简化后的问题可以直接由纳什均衡求解器求解。然而，对于能否找到符合设计者目标的最优纳什均衡，目前尚无理论保证 [21]。另一种思路是保留双层问题结构，并在上层应用无梯度优化器来规避导数问题 [64]。然而，这种方法计算成本很高，因为零阶方法通常需要多次查询低层求解器，才能得出所需的纳什均衡。因此，亟需一种能够应用于双层结构下复杂环境、且具有收敛保证的一阶方法，以解决样本效率问题。

我们将激励设计问题扩展到多智能体强化学习场景，并构建了一个具有可证明性质的可微仲裁 （DA）框架。我们推导了 DA 框架中通过纳什均衡的导数，使上层能够利用来自下层的策略梯度反馈，获得设计者目标的梯度。通过选择具有收敛保证的适当纳什均衡求解器，例如策略额外梯度法（PEM） [14]和熵正则化 OMWU [13]，我们针对双人零和博弈场景给出了 DA 的理论收敛性证明。在实践中，我们使用软演员-评论家算法的多智能体变体（MASAC） [30], 实现 DA 框架，从而得到一种实用算法 DASAC。我们在两个双人零和多智能体强化学习环境中评估了该算法。实验结果表明

*这些作者对本研究的贡献相同 †本研究是在上海棋院实习期间完成的。 ‡Feng Gao 完成了本研究的全部实验部分。

《第*22*届自主智能体与多智能体系统国际会议论文集》（*AAMAS 2023*），*A. Ricci*、*W. Yeoh*、*N. Agmon*、*B. An*（编），*2023*年*5*月*29*日至*6*月*2*日，英国伦敦。© 2023年国际自主智能体与多智能体系统基金会（www.ifaamas.org）。保留所有权利。

设计者提出的激励能够很好地仲裁冲突：也就是说，上层设计者的损失函数降低，同时下层自利玩家仍能达到纳什均衡。每种场景中涌现出的仲裁行为都符合人类直觉。此外，与零阶方法相比，上层的样本效率显著提升。

贡献。**i)** 我们研究多智能体强化学习场景中的激励设计问题，并着手解决如何通过双层结构推导上层目标梯度这一挑战。**ii)**我们提出了首个 可微的一阶 框架可微仲裁 （双层近似），可在马尔可夫博弈中仲裁冲突，从而获得理想的纳什均衡。**iii)** 我们从理论上证明，在采用恰当的纳什均衡求解器时，双人零和马尔可夫博弈中的双层近似具有收敛性。**iv)** 我们通过实验证明，在两个多智能体强化学习环境中，双层近似比零阶方法更高效地找到具有可解释行为的理想纳什均衡。

2 相关工作

马尔可夫博弈中的纳什均衡。 关于在多智能体马尔可夫博弈中寻找纳什均衡的研究已相当丰富。Claus 和 Boutilier 提出了虚拟博弈 [18] （FS），即假设所有玩家的对手均采用固定策略，以在保证收敛至纳什均衡的前提下进行均衡选择。FS 最初是为标准式博弈提出的，因此在虚拟自我博弈 （FSP） [33] 提出之前，并未广泛应用于复杂场景 [34, 45, 61] 。 [35] 将扩展式博弈中的 SP 推广至不完全信息博弈。 [55]提出了平滑虚拟自我博弈算法，将 PPO 算法 [74] 从单智能体扩展至多智能体零和马尔可夫博弈，并提供了收敛保证。在多智能体强化学习领域， [57] 中的 MADDPG 和 [39]中的 MAAC 将演员-评论家 [43] 和软演员-评论家 [30]算法从单智能体扩展至复杂的多智能体场景，并通过实验证明，在寻找纳什均衡方面，它们的表现全面优于去中心化多智能体强化学习方法。尽管 MADDPG 和 MAAC 在实证中取得了成功，但它们并无理论收敛保证。另一方面，许多研究为双人零和马尔可夫博弈中的基于价值的方法 [13, 14, 16]和基于策略的方法 [3, 94] 提供了收敛保证。Cen [13, 14] 将熵正则化项引入双人零和马尔可夫博弈，并提出策略外梯度方法和熵正则化在线镜像下降与乘法权重更新方法；在双人零和马尔可夫博弈中，即使不假设纳什均衡唯一，这些方法也具有收敛保证。

均衡选择／精炼。 由于即使是团队马尔可夫博弈（其中玩家具有共同利益）也可能存在多个纳什均衡（NE），因此博弈论领域中关于NE选择的研究成果十分丰富 [66, 75], 例如可容许性 [7], 子博弈完美性 [75],帕累托效率 [10] 或针对对手偏离最佳响应的稳定性 [24]。 [73] 着重于自适应激励设计，通过额外激励调整内在奖励，以平衡个体的自身利益与系统的社会福利，从而实现理想的NE选择。激励设计最早由经济学领域提出 [70] , 并因大量后续研究而受到广泛关注 [11, 44, 88]。过去几十年间，人们探索了经济激励在多个领域中的应用，包括能源 [12, 65], 交通运输 [42, 62], 医疗保健 [5, 59], 和教育 [60]。近年来，激励设计也扩展到了博弈论和强化学习领域。 [54] 将激励设计应用于多臂老虎机问题，并证明，对于一大类博弈，所提出的算法以次线性速率收敛至全局最优解。 [64] 为非合作多智能体系统提出了一种激励设计机制，并使用贝叶斯优化这一种样本高效优化算法，优化上层激励目标，而不是采用基于梯度的方法, 因为下层多智能体强化学习问题是一个黑箱。 [90] 提出了一种分散式激励机制，使每个个体能够直接向其他个体提供奖励，并分别学习自身的激励函数。此外， [78] 针对单个下层玩家提出了一种双层激励机制，这可以视为我们所研究设置的特例。斯塔克尔伯格博弈 [48, 52] 是经济学中的一种策略博弈，其中领导者先行动，跟随者依次行动。将激励设计目标视为领导者后，激励设计可以重新表述为斯塔克尔伯格博弈。 [86] 提出了一种梯度下降算法，用于求解斯塔克尔伯格老虎机问题中的 NE。 [95] 提出了一种值迭代方法，用于求解具有收敛保证的斯塔克尔伯格马尔可夫博弈。 [85] 研究了在序贯决策问题上通过策略梯度进行的特征微分，但不提供全局收敛保证。

双层优化。在机器学习中，大量任务都可以表述为双层优化问题，例如对抗学习 [40], 元学习 [26], 超参数优化 [9], 以及端到端学习 [4]。许多基于梯度的算法被用于求解这些问题。其中大多数是基于一阶梯度的方法 [9, 26, 28, 67] ，这得益于一阶梯度的计算效率以及快速的经验收敛性。相比之下，零阶方法 [79, 87]需要大量数据采样，因而计算负担沉重，需要消耗大量计算资源。 [69] 提出了一种二阶方法。然而，该算法需要计算海森矩阵的逆，这使其仅适用于简单的表格型设置。基于梯度的方法求解双层优化问题时，最具挑战性的部分是推导下层解相对于上层激励变量的导数。Feng [25] 和 Yang [91] 将下层优化过程展开为一系列梯度更新，并直接计算整个优化过程中的梯度。尽管这种方法简单，但它会因内存爆炸和数值问题而面临巨大的计算困难。因此，必须采用滑动窗口截断梯度等近似技术。相比之下，我们的 DA 框架为下层最优解的精确梯度计算提供了准确的一阶公式。Wang [85] 构建了一个双层问题，通过学习预测模型来推断马尔可夫决策过程（MDP）中缺失的参数。这篇论文采用单智能体设置，其中下层目标是最小化一个明确的性能指标，而我们的设置则更为复杂，因为它要求计算纳什均衡（NE）。

3 问题表述

本节安排如下。第 3.1 节首先给出双人零和马尔可夫博弈 G 的一个典型定义。为了说明

设计者如何诱导玩家的行为，第 3.2 节将激励参数 θ 引入双人零和马尔可夫博弈的奖励中，并构建激励马尔可夫博弈 G_θ。第 3.3 节通过在 G_θ 的奖励中加入熵正则化项，设计熵正则化马尔可夫博弈 $\mathrm{G}_{\pi\theta}$，并证明 $\mathrm{G}_{\pi\theta}$ 存在唯一的纳什均衡。第 3.4 节基于具有激励扰动奖励的正则化马尔可夫博弈，将仲裁系统构建为双层优化方案。

3.1 双人零和马尔可夫博弈

考虑一个双人零和马尔可夫博弈 $\mathrm{G}=(\mathrm{S},\ \{\mathrm{A}^i\}_{i\in\{1,\ 2\}},\mathrm{P},\ \{r^i\}_{i\in\{1,\ 2\}}\ \gamma)$，其中 S 是所有玩家观测到的状态空间，$\mathrm{A}^i$ 是玩家 i 的动作空间，$\mathrm{A}:=\mathrm{A}^1\times\mathrm{A}^2$ 是两个玩家的联合动作空间；$\mathrm{P}:\mathrm{S}\times\mathrm{A}\times\mathrm{S}\to[0,1]$ 表示采取联合动作 $a\in\mathrm{A}$ 时，状态 $s\in\mathrm{S}$ 转移到状态 $s\prime\in\mathrm{S}$ 的转移概率。$r^i:\mathrm{S}\times\mathrm{A}\to[-1,1]$ 是玩家 i 的即时奖励函数，这意味着在零和博弈中 $r^1=-r^2$。$\gamma\in[0,1)$ 是折扣因子。] 为简化记号，我们考虑双人零和情形。对更一般情形的扩展将在附录 C 中讨论。

3.2 激励扰动奖励下的马尔可夫博弈

设 $\mathrm{G}_\theta=(\mathrm{S},\{\mathrm{A}^i\}_{i\in\{1,2\}},\mathrm{P},\{r^i(\cdot;\theta)\}_{i\in\{1,2\}}\ \gamma)$ 为一个双人零和马尔可夫博弈，其激励扰动奖励函数 $r^i(\cdot;\theta)$ 由参数 θ 显式表示，其中 $\theta\in\mathbb{R}^m$ 表示设计者添加的激励。我们假设 $r^i(\cdot;\theta):\mathrm{S}\times\mathrm{A}\to[-1,1]$ 关于 θ 一致有界，且保持零和。对于给定的轨迹 $\tau=(s_0,\{a_0^i\}_{i\in\{1,2\}}\ s_1,\ldots,\{a_{T-1}^i\}_{i\in\{1,2\}}\ s_T)$，我们将玩家 $i\in\{1,2\}$ 的总折扣奖励记为

$$R^i(\tau;\theta)=\sum_{t=0}^{T-1}\gamma^t r^i(s_t,a_t;\theta).$$

定义 $\pi^i:\mathrm{S}\times\mathrm{A}^i\to[0,1]$ 为玩家 i 的策略概率分布，$\pi=(\pi^1,\ \pi^2)$ 则表示联合策略的概率分布。

令 $D^\pi(\tau)$ 表示基于策略对 π 的轨迹 τ 的概率分布。即，

$$D^\pi(\tau):=\rho_0(s_0)\prod_{t=0}^{T-1}P(s_{t+1}|s_t,a_t)\left(\prod_{i=1}^{2}\pi^i(a_t^i|s_t)\right),$$

其中，$\rho_0(s_0)$ 是初始状态的分布。我们将 $\mathbb{E}_\pi[:\ ]$ 表示为对轨迹 $\tau\sim D^\pi(\tau)$ 的期望。然后，通过状态价值函数 $V_\pi^i:\mathrm{S}\to\mathbb{R}$ 评估该博弈 G_θ 中策略对 π 的表现，其定义为

$$\begin{aligned}V_\pi^i(s;\theta)&=\mathbb{E}_\pi\left[\sum_{t=0}^{T-1}\gamma^t\cdot r^i(s_t,a_t^i,a_t^{-i};\theta)|s_0=s\right]\\&=\mathbb{E}_{\tau\sim D^\pi}\left[R^i(\tau;\theta)\middle|s_0=s\right].\end{aligned}\tag{3.1}$$

纳什均衡（NE）是一个重要概念，指任何智能体都无法通过单独改变其策略获益。激励型双人零和马尔可夫博弈 G_θ 的纳什均衡是以下最小最大博弈的解：

$$\min_{\pi^2}\max_{\pi^1}V_\pi^1(s;\theta),\quad s\in\mathcal{S}.\tag{3.2}$$

Shapley [76] 证明，对于所有状态 $s\in\mathrm{S}$，极小极大博弈 3.2 都存在一个纳什均衡对，且极小极大值唯一 [84]。

3.3 熵正则化马尔可夫博弈

为确保纳什均衡唯一，我们遵循 [14, 20, 93]，并引入熵正则化版本 $\mathrm{G}_\theta^{'}=(\mathrm{S},\ \{\mathrm{A}^i\}_{i\in\{1,\ 2\}},\mathrm{P},\ \{r_{(i)}(\cdot;\theta)\}_{i\in\{1,\ 2\}}\ \gamma,\ \lambda)$，其中 $\lambda\geq 0$ 是正则化参数。熵正则化技术常用于强化学习，以鼓励探索，避免陷入次优解 [2, 29, 46] ，并使不同智能体的策略不受对手策略的过度影响 [56, 89]。

具体而言，玩家 i 的奖励函数被替换为其熵正则化奖励函数 $r_{(i)\pi}(\cdot;\theta):\mathrm{S}\times\mathrm{A}^i\times\mathrm{A}^{-i}\to\mathbb{R}$，其定义如下：

$$\begin{aligned}r_\pi^{(i)}(s,a^i,a^{-i};\theta)=&r^i(s,a^i,a^{-i};\theta)-\lambda\log(\pi^i(a^i|s))\\&+\lambda\log(\pi^{-i}(a^{-i}|s)).\end{aligned}\tag{3.3}$$

这里，我们稍微宽泛地使用“熵”一词，并对其含义作出说明。与策略对 π 相关的状态奖励函数和熵正则化状态奖励函数定义如下：

$$\begin{aligned}r_\pi^i(s;\theta)&=\mathbb{E}_{(a^i,a^{-i})\sim\pi}\left[r^i(s,a^i,a^{-i};\theta)\right],\\r_\pi^{(i)}(s;\theta)&=\mathbb{E}_{(a^i,a^{-i})\sim\pi}\left[r_\pi^{(i)}(s,a^i,a^{-i};\theta)\right]\\&=r_\pi^i(s;\theta)+\lambda H(\pi^i(\cdot|s))-\lambda H(\pi^{-i}(\cdot|s)),\end{aligned}\tag{3.4}$$

其中 $H(\pi^i(\cdot|s))=-\sum_{a^i\in\mathcal{A}_i}\pi^i(a^i|s)\log(\pi^i(a^i|s))$ 为香农熵。相应地，给定玩家 i 在博弈 $\mathrm{G}_{\pi\theta}$ 中的轨迹 τ 时，其总折扣奖励 $R_{(i)\pi}$ 以及激励参数为 V 的熵正则化状态价值函数 $_{(i)\pi}:\mathrm{S}\to\mathbb{R}$ 定义如下：

$$\begin{aligned}R_\pi^{(i)}(\tau;\theta)&=\sum_{t=0}^{T-1}\gamma^t r_\pi^{(i)}(s_t,a_t^i,a_t^{-i};\theta),\\V_\pi^{(i)}(s;\theta)&=\mathbb{E}_\pi\left[\sum_{t=0}^{T-1}\gamma^t\cdot r_\pi^{(i)}(s_t,a_t^i,a_t^{-i};\theta)|s_0=s\right]\\&=\mathbb{E}_{\tau\sim D^\pi}\left[R_\pi^{(i)}(\tau;\theta)\middle|s_0=s\right].\end{aligned}\tag{3.5}$$

假设带有激励扰动奖励的正则化马尔可夫博弈 $\mathrm{G}_{\pi\theta}$ 的纳什均衡为 $\pi_*(\theta)=(\pi_*^1(\theta)\ \pi_*^2(\theta))$。将玩家 -$i$记作玩家 i’ 的对手，则联合策略也可以改写为策略对 $\pi=(\pi^i,\pi^{-i})$，其中 $i\in\{1,2\}$。于是，纳什均衡可定义如下：对于 $\forall i\in\{1,2\}$，纳什均衡 π^* 满足

$$V_{\pi_*^{(i)}(\theta)}^{(i)}(s;\theta)=\max_{\pi^i}V_{(\pi^i,\pi_*^{-i}(\theta))}^{(i)}(s;\theta),\quad\forall s\in\mathcal{S}.\tag{3.6}$$

正则化马尔可夫博弈 $\mathrm{G}_{\pi\theta}$ 的纳什均衡等价于以下极小极大优化问题的解

$$\min_{\pi^2}\max_{\pi^1}V_\pi^{(1)}(s;\theta),\quad\forall s\in\mathcal{S}.\tag{3.7}$$

众所周知，[13, 14, 63], 对于极小极大优化问题 (3.7)，存在唯一的策略对 π_*。因此，熵正则化马尔可夫博弈存在唯一的纳什均衡。此外，[14] 证明，当正则化参数为 $\lambda\leq\frac{(1-\gamma\ \epsilon)}{(\log(|\mathcal{A}_1|)\log(|\mathcal{A}_2|))}$ 时，正则化马尔可夫博弈 $\mathrm{G}_{\pi\theta}$ 的 $\frac{2}{}$ 个 ϵ-最优纳什均衡也是原博弈的一个 $\epsilon/2$-最优纳什均衡，这表明正则化马尔可夫博弈的纳什均衡与原博弈的纳什均衡之间的差异可以通过正则化参数加以控制。其定义为

ϵ-最优纳什均衡在 [14], 中有明确的定义，即策略的最优性差距小于 ϵ。

3.4 双层优化方案

我们的最终目标是通过引导所有个体达到理想的、更优的纳什均衡，从而在利己玩家与集体利益设计者之间仲裁冲突，并提升系统福利。第 3.2 节介绍了激励参数 θ，用于修改马尔可夫博弈的内在奖励 $r^i(\cdot;\theta)$，并将其作为设计系统目标、优化纳什均衡以实现仲裁的工具。本节针对这种仲裁系统所引发的均衡约束，构建双层优化方案。在该双层优化方案中，对于下层玩家，我们旨在为给定的激励马尔可夫博弈 $G_r\theta$寻找纳什均衡。对于上层设计者，我们定义一个系统仲裁目标，以 θ的最优值为目标，使其能够引导所有个体趋向理想的纳什均衡。

为了在第 4 节进行梯度计算以及在第 7 节开展实验，我们用 $\phi^i \in \mathbb{R}^{d_i}$ 参数化玩家 i（$i \in \{1,2\}$）的策略 $\pi^i_{\phi^i}$，并将策略对记为 $\pi_\phi = (\pi^1_{\phi^1}, \pi^2_{\phi^2})$，其中 $\phi = (\phi^1, \phi^2) \in \mathbb{R}^d, d = d_1 + d_2$。为避免下标滥用，我们省略 $\pi^i_{\phi^i}$中 ϕ^i的 i（关于 π^i_ϕ）。这种简化不会造成混淆，因为单个玩家 i的策略 π^i仅由 ϕ^i决定。因此，联合策略也可改写为策略对 $\pi_\phi = (\pi^i_\phi, \pi^{-i}_\phi)$，其中 $i \in \{1, 2\}$。具有激励扰动奖励的正则化马尔可夫博弈 G'_θ的纳什均衡记为 $\pi_{\phi_*}(\theta) = (\pi^1_{\phi_*}(\theta)\ \pi^2_{\phi_*}(\theta))$，其中 $\phi_*(\theta) = (\phi^1_*(\theta)\ \phi^2_*(\theta))$ 是激励 θ下纳什均衡的参数。式 (3.6) 中正则化马尔可夫博弈的纳什均衡定义可改写为：

$$V^{(i)}_{\pi_{\phi_*(\theta)}}(s;\theta) = \max_{\phi^i \in \mathbb{R}^{d_i}} V^{(i)}_{(\pi^i_\phi, \pi^{-i}_{\phi_*(\theta)})}(s;\theta), \quad \forall s \in \mathcal{S}.$$

仲裁系统。假设博弈系统 $G_r\theta$ 的仲裁目标仅由激励参数 θ 和联合策略 π_ϕ决定，可表示为映射 $f(\theta,\phi): \Theta \times \mathbb{R}^d \to \mathbb{R}$。该系统旨在使 G_θ的纳什均衡处的仲裁目标损失最小化 $f(\theta, \phi_*(\theta))$。因此，仲裁系统可表述为一个双层优化问题：

$$\begin{aligned} \min_{\theta\in\Theta} \quad & f_*(\theta) = f(\theta, \phi_*(\theta)) \\ s.t. \quad & \mathbb{E}_{\nu^*}\left[V^{(i)}_{\pi_{\phi_*(\theta)}}(s;\theta)\right] \\ & = \max_{\phi^i \in \mathbb{R}^{d_i}} \mathbb{E}_{\nu^*}\left[V^{(i)}_{(\pi^i_\phi, \pi^{-i}_{\phi_*(\theta)})}(s;\theta)\right], \ \forall i \in \{1,2\} \end{aligned} \tag{3.8}$$

备注：公式 3.8 是我们的主要目标。我们主要关注如何高效地学习上层激励参数，同时将下层视为黑盒纳什均衡求解器。在接下来的章节中，我们将重点推导上层损失函数 $f_*(\theta)$ 的一阶梯度，然后基于此建立我们的可微仲裁（DA）框架。

4 方法

带有纳什均衡约束的双层仲裁系统（式 3.8）通过激励参数 θ间接引导玩家，以保障仲裁目标 f。对于下层，我们试图确定特定激励环境 $G_r\theta$下纳什均衡的策略参数，记为 $\phi^*(\theta)$。对于上层设计者，我们旨在推导一阶算法，以寻找最优激励参数 θ，使目标 $f_*(\theta) = f(\theta,\ \phi_*(\theta))$ 得到优化，同时玩家达到纳什均衡。在此类双层问题中，求取 f_* 的梯度颇具挑战，尤其是在下层求解正则化纳什均衡的求解器属于黑箱的情况下，因为激励参数 θ 会通过改进玩家的正则化纳什均衡，隐式地影响目标 f_*。我们的目标是在纳什均衡约束下推导 f_* 关于 θ 的梯度，然后制定一种反向传播方案，利用下层策略梯度的反馈，求得正则化纳什均衡处激励目标函数的梯度 $\nabla_\theta f(\theta,\ \phi^*(\theta))$。具体而言，梯度 $\nabla_\theta f(\theta,\ \phi^*(\theta))$ 可推导如下：

$$\nabla_\theta f_* = \left[\nabla_\theta f(\theta,\phi) + \nabla_\theta \phi^*(\theta)^\intercal \nabla_\phi f(\theta,\phi)\right]\Big|_{\phi=\phi^*(\theta).} \tag{4.1}$$

其中，$\nabla_\theta f(\theta,\phi)$ 和 $\nabla_\phi f(\theta,\phi)$ 已知，而纳什均衡梯度$\nabla_\theta \phi_*(\theta)$则需要基于下层的策略梯度反馈推导。纳什均衡梯度 $\nabla_\theta \phi^*(\theta)$ 的计算将在第 4.1 节中详细讨论，随后将在第 4.2 节提出一个适用于正则化纳什均衡的通用反向传播框架。

4.1 激励的梯度

本节将推导纳什均衡 $\nabla_\theta f(\theta,\ \phi^*(\theta))$ 处的梯度，其中关键在于推导纳什均衡梯度 $\nabla_\theta \phi^*(\theta)$。为简化记号，我们将 $u_{(i)\theta}(\phi;\ s)$ 记作 $V_{(i)}\phi(s;\ \theta)$ 关于 ϕ^i 的梯度。然后定义

$$u_\theta(\phi) := \mathbb{E}_{\nu^{(*)}}[u_\theta(\phi;s)], \tag{4.2}$$

$$u_\theta(\phi;s) := (u^{(1)}_\theta(\phi;s), u^{(2)}_\theta(\phi;s)) = (\nabla_{\phi^1} V^{(1)}_{\pi_\phi}(s;\theta), \nabla_{\phi^2} V^{(2)}_{\pi_\phi}(s;\theta)).$$

那么，$u_\theta(\phi)$ 关于 θ 和 ϕ 的梯度可直接写为：

$$\nabla_\theta u_\theta(\phi) = \begin{bmatrix} \mathbb{E}_{\nu^{(*)}} \nabla^2_{\theta\phi^1} V^{(1)}_{\pi_\phi}(s;\theta) \\ \mathbb{E}_{\nu^{(*)}} \nabla^2_{\theta\phi^2} V^{(2)}_{\pi_\phi}(s;\theta) \end{bmatrix} \in \mathbb{R}^{d\times m}, \tag{4.3}$$

$$\nabla_\phi u_\theta(\phi) = \begin{bmatrix} \mathbb{E}_{\nu^{(*)}} \nabla^2_{\phi\phi^1} V^{(1)}_{\pi_\phi}(s;\theta) \\ \mathbb{E}_{\nu^{(*)}} \nabla^2_{\phi\phi^2} V^{(2)}_{\pi_\phi}(s;\theta) \end{bmatrix} \in \mathbb{R}^{d\times d}. \tag{4.4}$$

下述引理 4.1 表明，$\nabla^2_{\theta\phi^i} V_{(i)}\pi_\phi(s;\theta)$ 和 $\nabla^2_{\phi\phi^i} V_{(i)}\pi_\phi(s;\theta)$ 的计算可基于策略梯度显式推导；这些梯度由下层的纳什均衡求解器反向传播得到。受篇幅限制，引理 4.1 中梯度和海森矩阵的计算细节在附录的引理 A.1 中给出。根据式 (4.3–4.4) 和引理 4.1，$\nabla_\theta \phi^*(\theta)$ 可表示为策略梯度，其推导见下述引理 4.2，由此得到纳什均衡 $\nabla_\theta f_*(\theta)$ 处目标梯度的最终表达式。

引理 4.1. 在激励正则化马尔可夫博弈 $\mathcal{G}'_\theta$中，为简化记号，我们记作$\pi^i_{\phi^i} :=\ \pi^i_{\phi^i}(a^i_t|s_t)$、$\pi^{-i}_{\phi^{-i}} :=\ \pi^{-i}_{\phi^{-i}}(a^{-i}_t|s_t)$ 和 $\pi_\phi :=\ \pi_\phi(a_t|s_t)$，则有

$$\nabla_\theta V^{(i)}_{\pi_\phi}(s;\theta) = \nabla_\theta V^i_{\pi_\phi}(s;\theta); \quad \nabla^2_{\theta\phi^i} V^{(i)}_{\pi_\phi}(s;\theta) = \nabla^2_{\theta\phi^i} V^i_{\pi_\phi}(s;\theta);$$

$$\nabla_{\phi^i} V^{(i)}_{\pi_\phi}(s;\theta) = \nabla_{\phi^i} V^i_{\pi_\phi}(s;\theta) - \lambda \mathbb{E}_{D^{\pi_\phi}} \Bigg[\sum_{t=0}^{T-1} \gamma^t \nabla_{\phi^i} \log \pi^i_{\phi^i} + \left(\sum_{t=0}^{T-1} \gamma^t \left(\log \pi^i_{\phi^i} - \log \pi^{-i}_{\phi^{-i}} \right) \right) \left(\sum_{t=0}^{T-1} \nabla_{\phi^i} \log \pi^i_{\phi^i} \right) \Bigg];$$

$$\begin{aligned}
\nabla^2_{\phi\phi^i} V^{(i)}_{\pi_\phi}(s;\theta) &= \nabla^2_{\phi\phi^i} V^i_{\pi_\phi}(s;\theta) \\
&- \lambda \mathbb{E}_{D^{\pi_\phi}} \Bigg[\left(\sum_{t=0}^{T-1} \nabla_{\phi^i} \log \pi^i_{\phi^i} \right) \left(\sum_{t=0}^{T-1} \gamma^t \left(\nabla_\phi \log \pi^i_{\phi^i} - \nabla_\phi \log \pi^{-i}_{\phi^{-i}} \right) \right)^\top \\
&+ \left(\sum_{t=0}^{T-1} \gamma^t \left(\log \pi^i_{\phi^i} - \log \pi^{-i}_{\phi^{-i}} \right) \right) \left(\sum_{t=0}^{T-1} \nabla_{\phi^i} \log \pi^i_{\phi^i} \right) \left(\sum_{t=0}^{T-1} \nabla_\phi \log \pi_\phi \right)^\top \\
&+ \left(\sum_{t=0}^{T-1} \gamma^t \left(\log \pi^i_{\phi^i} - \log \pi^{-i}_{\phi^{-i}} \right) \right) \left(\sum_{t=0}^{T-1} \nabla^2_{\phi\phi^i} \log \pi_\phi \right)^\top \\
&+ \left(\sum_{t=0}^{T-1} \gamma^t \nabla_{\phi^i} \log \pi^i_{\phi^i} \right) \left(\sum_{t=0}^{T-1} \nabla_\phi \log \pi_\phi \right)^\top + \left(\sum_{t=0}^{T-1} \gamma^t \nabla^2_{\phi\phi^i} \log \pi_\phi \right) \Bigg],
\end{aligned}$$

其中，$\mathbb{E}_{D^{\pi_\phi}}[\,]$ 表示对轨迹 $\tau \sim D^{\pi_\phi}(\tau)$ 取期望，而 $\nabla_\theta V^i_{\pi_\phi}(s;\ \theta)$、$\nabla_{\phi^i} V^i_{\pi_\phi}(s;\ \theta)$、$\nabla^2_{\theta\phi^i} V^i_\phi(s;\ \theta)$ 和 $\nabla^2_{\phi\phi^i} V^i_\phi(s;\ \theta)$ 均由附录中的引理 *A.1* 计算得出。

引理 4.1 利用纳什均衡求解器中易于获取的策略梯度信息，描述利己玩家如何响应激励参数的调整。该引理易于证明，但对于提供策略梯度中所蕴含的低层信息至关重要，从而能够在集体损失较高时指导激励参数更新的方向。

鉴于引理 4.1，引理 4.2 给出了纳什均衡关于激励参数的梯度，这对于利用从低层反向传播的信息来更新激励参数至关重要。它描述了玩家之间的均衡如何受到设计者激励的影响，并表明该信息与纳什均衡处的 $V^{(i)}_{\pi_\phi}(s;\theta)$ 的海森矩阵有关。

引理 4.2。对于激励正则化马尔可夫博弈 $\mathcal{G}'_\theta$ 的纳什均衡 $\pi_{\phi^*(\theta)}$ 的策略参数 $\phi^{(*)}(\theta)$，以下公式成立：

$$\nabla_\theta \phi^{(*)}(\theta) = -\left[\nabla_\phi u_\theta(\phi)\right]^{-1} \nabla_\theta u_\theta(\phi)\big|_{\phi=\phi^*(\theta)}. \tag{4.5}$$

然后将公式 (4.5) 代入公式 (4.1)，即可得到 $\nabla_\theta f_*(\theta)$ 的最终显式表达式：

$$\begin{aligned}
&\nabla_\theta f_*(\theta, \phi^*(\theta)) \\
&= \left[\nabla_\theta f(\theta,\phi) - \nabla_\theta u_\theta(\phi)^\top \left[\nabla_\phi u_\theta(\phi)\right]^{-1} \nabla_\phi f(\theta,\phi) \right] \big|_{\phi=\phi^*(\theta)}.
\end{aligned} \tag{4.6}$$

4.2 可微仲裁

基于第 4.1 节中的梯度计算，我们提出可微仲裁 （DA）框架来求解仲裁系统（式 3.8）。DA 通过纳什均衡对上层设计者目标函数进行微分，并将策略的梯度信息从下层反向传播，以求得上层梯度。在这一方案中，下层用于求解正则化纳什均衡的黑盒求解器可以是任何包含策略梯度信息的方法，例如 SAC [30],MADDPG [57], MAAC [39], 策略外梯度法 [14]和熵正则化 OMWU [13]。

框架 1双层近似：多智能体强化学习中的可微仲裁。

输入： β_k：上层迭代的学习率
输出： θ：激励参数； ϕ：策略 π_ϕ 的参数
初始化激励参数 $\theta = \theta_0$。
对于 k=0,1,...**执行**
  初始化博弈 G_{θ_k}的策略π_{ϕ_0} 参数 $\phi = \phi_0$ 。
  对于 t=0,1... **执行**
    使用纳什均衡求解器更新 $\phi = \phi_t$ ，直至对应策略 π_{ϕ_t} 收敛到博弈 G_{θ_k} 的纳什均衡，关于 $\phi^*(\theta_k)$。
  结束循环
  更新激励参数

$$\theta = \theta_{k+1} \leftarrow \theta_k - \beta_k \nabla f_*(\theta_k),$$

  其中， $\nabla f_*(\theta_k)$ 定义见式 (4 6)．．
结束循环

DA 的工作流程如框架 1 所示。在本文中，我们专门为采用具有收敛保证的下层纳什均衡求解器的 DA 框架推导了收敛保证，并使用多智能体版SAC实现了 DA 框架，具体内容将在第6节中详细介绍。

DA 框架是一阶方法。这是因为上层激励 θ 的更新只依赖一阶梯度 $\nabla f_*(\theta)$，无需任何关于 θ 的海森信息。引理 4.1 中定义的二阶梯度是底层价值函数 V 的海森矩阵，并用于计算一阶梯度 $\nabla f_*(\theta)$。

扩展到一般设置。尽管 DA 框架是在双人零和完全可观测马尔可夫博弈的背景下描述的，但它可以扩展到更一般的设置：N 个玩家、一般和以及部分可观测马尔可夫决策过程 （*POMDP*）。对于 N 人一般和马尔可夫博弈，需要博弈 G_θ 满足单调性，以确保熵正则化博弈 G_{θ} 具有强单调性，从而存在唯一的纳什均衡。随后，DA 框架可以利用合适的底层黑盒纳什均衡求解器求解（正则化）纳什均衡。对于 POMDP，只需进行一些小幅调整，将基于状态的策略 $\pi^i(a^i|s)$ 替换为基于观测的策略 $\pi^i(a^i|o^i)$，其中 $o^i = O^i(s)$ 是玩家 i 在状态 s 下的观测。这些调整不会改变梯度计算，而这对 DA 框架至关重要。一般设置下的问题表述和梯度推导详见附录 C。

5 收敛性分析

本节将在双人零和设置下，针对采用适当下层纳什均衡求解器的 DA 框架，开展收敛性分析。第 5.1 节讨论文献中具有收敛保证、可应用于我们的 DA 框架的下层纳什均衡求解器。第 5.2 节在下层纳什均衡求解器收敛的条件下，保证双层 DA 框架中上层激励参数 θ 的收敛性。

5.1 下层的收敛性

双层 DA 框架的收敛性要求下层纳什均衡求解器能够保证收敛至熵正则化马尔可夫博弈的纳什均衡 π_*

熵正则化马尔可夫博弈 $G_{\mathcal{M}\theta}$。在实际博弈中，有多种纳什均衡求解器表现良好 [30, 39, 57]。一些已提出的纳什均衡求解器已有理论收敛性证明 [37, 77, 80]。为了说明 DA 框架的收敛性，我们选取两种在熵正则化下提出且与我们的设定直接相关的纳什均衡求解器。已知策略外梯度法（PEM） [14] 和熵正则化 OMWU [13] 在我们的设定下都具有收敛保证。本节重述 PEM 和熵正则化 OMWU 的主要收敛结果，以完整说明 DA 框架的收敛保证。

命题 5.1 （ [14]中的定理 3）。 假设 $|\mathcal{S}_1| \geq |\mathcal{S}_2|$ 且$\lambda \leq 1$。将学习率设为 $\eta_t = \eta = \frac{1-\gamma}{2(1+\lambda)}$，策略外梯度法（补充材料中的算法 *4*）最多经过 $\widetilde{\mathcal{O}}\left(\lambda^{-1}(1-\gamma)^{-2}\log^2\left(\frac{1}{\epsilon}\right)\right)$ 次迭代，即可得到满足以下条件的策略对 $\pi = (\pi_1, \pi_2)$：

$$\max_{\pi_1', \pi_2'} \mathbb{E}_{s\sim\rho}\left[\left(V^{(1)}_{\pi_1', \pi_2}(s;\theta) - V^{(1)}_{\pi_1, \pi_2'}(s;\theta)\right)\right] \leq \epsilon,$$

其中，ρ 是状态空间$\mathcal{S}$上的任意分布。

命题 5.2 （ [13]中的定理 1）。 将 $0 < \eta \leq \left(\frac{1-\gamma}{32000|\mathcal{S}|}\right)^3$且 $\alpha_t = \eta\lambda$设定后，熵正则化 *OMWU*（补充材料中的算法 *5*）最多经过 $\widetilde{\mathcal{O}}\left(\frac{1}{1-\gamma}\left(\eta\lambda\log\frac{1}{\epsilon}\right)\right)$次迭代，即可得到满足以下条件的策略对 $\pi = (\pi_1, \pi_2)$：

$$\max_{s\in\mathcal{S}, \pi_1', \pi_2'} \left(V^{(1)}_{\pi_1', \pi_2}(s;\theta) - V^{(1)}_{\pi_1, \pi_2'}(s;\theta)\right) \leq \epsilon.$$

命题 5.1 和命题 5.2 表明，PEM 和熵正则化 OMWU 算法能够以线性速率收敛到下层激励型熵正则化马尔可夫博弈 $G_{\mathcal{M}\theta}$ 的纳什均衡。

5.2 上层的收敛性

假设 $\theta^* = \arg\min_\theta f_*(\theta)$ 表示仲裁系统 (3.8) 中激励参数 θ的最优点。本节将说明，在某些假设下，框架 1 能保证激励参数 θ收敛到 θ^*。

假设 5.3。 我们假设 $\nabla_\theta f_*(\theta)$ 关于 $\theta \in \Theta$是 L-Lipschitz 连续（L-光滑）的，这意味着

$$\left\|\nabla_\theta f_*(\theta) - \nabla_\theta f_*(\theta')\right\| \leq L\left\|\theta - \theta'\right\|, \quad \forall \theta, \theta' \in \Theta$$

我们还假设 $f_*(\theta)$ 关于 $\theta \in \Theta$有界，这意味着存在 $M > 0$ 使得

$$|f_*(\theta)| \leq M, \quad \forall \theta \in \Theta$$

L-光滑假设是不同算法 [1, 41, 50, 72, 92] 的收敛性证明中常见的假设。我们还在附录 E 中给出一个具体示例，以说明假设 5.3 在实际情形下是合理的。假设 5.3 不要求复合函数 $f_*(\theta) = f(\theta, \phi^*(\theta))$ 为凸函数，而且上层的收敛保证甚至适用于非凸情形。

定理 5.4。 假设假设 *5.3* 成立，且下层纳什均衡求解器具有收敛保证。在框架 *1* 中，设激励参数 θ_k 的更新规则 $\theta_{k+1} = \theta_k - \beta_k\nabla_\theta f_*(\theta_k)$ 以学习率 $\beta_k = \frac{1}{L}$ 运行 T 次迭代，则有

$$\min_{k=0,\ldots,T}\|\nabla_\theta f_*(\theta_k)\|^2 \leq \frac{2L(f_*(\theta_0) - f_*(\theta^*))}{T+1} \leq \frac{4LM}{T+1}.$$

定理 5.4 意味着，在梯度范数序列 $\{\|\nabla_\theta f_*(\theta_k)\|\}_{k=0}^{\infty}$ 中存在一个收敛至零的子序列。因此，定理 5.4 保证激励参数 θ 在上层总能以次线性收敛速率收敛至某个次优点。容易看出，若复合函数 $f^*(\theta)$ 是凸函数，则其会收敛至全局最优点。详细证明见附录 F。

6 实现

尽管 PEM 和熵正则化 OMWU 具有理论收敛保证，但它们的直接实现并不实用/高效。PEM 是一种双循环算法，每次内循环都需要求解一个熵正则化矩阵博弈的纳什均衡。熵正则化 OMWU 虽然是一种采用双时间尺度迭代的单循环算法，但其更新规则适用于表格化设置，即按状态逐一更新策略。因此，我们借鉴多智能体软演员-评论家（MASAC）的一些技术，对熵正则化 OMWU 的原始算法进行改进，并提出 DA-SAC 算法（算法 6）。我们将这些改进归纳如下：

i) 我们不采用熵正则化 OMWU 中逐状态的策略更新，而是通过最小化独立策略模型与指数能量函数之间的 KL 散度来优化该模型，即采用 SAC 中广泛接受的演员-评论家结构。ii) 我们采用在实践中效果显著、并被大多数多智能体强化学习实用算法（如 SAC [30], MADDPG [57] 和 MAAC [39], ）所采用的单步梯度更新，而非熵正则化 OMWU 中引入的外梯度技术。iii) 我们遵循 SAC，将能量函数设为 Q 函数的倍数，其中缩放因子为自动调节的温度系数 [31] 的倒数。

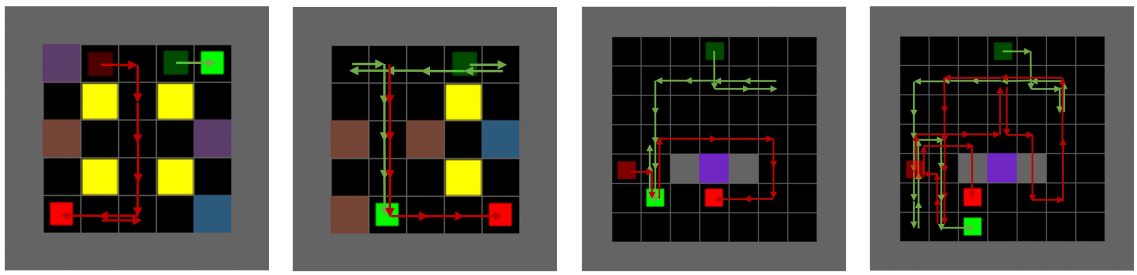
更新后的伪代码及所有其余细节均见附录 G。

7 实验

我们在两个双玩家零和马尔可夫博弈及其激励变体中，评估 DA 框架上层的样本效率。实验采用剪刀石头布（RWS） [83] ，该环境包含循环强化学习挑战 [6, 47, 68] ，以及标准的捕食者-猎物博弈；两者均在网格世界环境中实现 [17]。

7.1 评估环境

剪刀游戏（RWS）。在 5×5 网格世界剪刀游戏（RWS）中，资源（石头、剪刀或布）分布在资源池中（图2）。其中三个是确定性资源池，资源位置固定；另外六个是非确定性资源池，里面随机填充资源。两名玩家随机出现在空闲网格中，并能完全观察环境以及对手拥有的资源。每名



(a) RWS (θ_0)　(b) RWS (θ^*)　(c) PP (θ_0)　(d) PP (θ^*)

图1：激励型 RWS 和激励型 PP 在不同激励参数 θ_0 和 θ^* 下的轨迹（从左到右）。θ_0 是初始激励参数，θ^* 是采用 DA 框架优化后的参数。与未进行仲裁的初始纳什均衡（图 (a) 和 (c)）相比，仲裁后的纳什均衡（图 (b) 和 (d)）在网格世界游戏中具有更高的探索率。

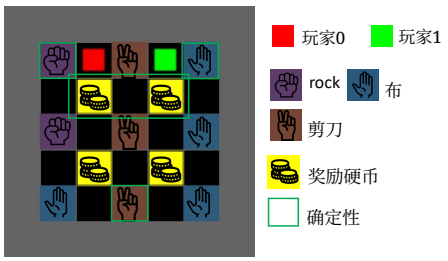


图2：石头-水-剪刀（RWS）的配置。两名玩家随机出现在空闲网格中，并收集资源以进行对抗。奖励硬币是设计者添加的激励，用于鼓励探索。

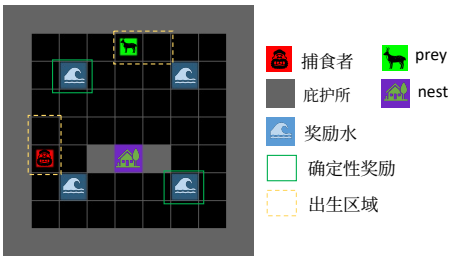


图3：捕食者-猎物（PP）的配置。捕食者的目标是捕捉猎物，而猎物的目标是返回巢穴。它们都随机出生在出生区域。设计者将水池设为激励。

玩家每一步可以选择向四个方向之一移动一格。当玩家 i 走到有资源的网格时，就会将该资源收集到自己的库存 v^i中，而该资源会从网格上移除。在初始状态，四个石头、布和剪刀会随机分配给两名玩家，并确保每名玩家至少拥有一种各类资源。经过 25 步后，双方展开对抗，并根据标准反对称矩阵 M 36 计算每名玩家的收益，r^0 和 r^1，具体如下

y　[ ]

$$r^0 = \frac{v^0}{\|v^0\|} M \left(\frac{v^1}{\|v^1\|}\right)^\top = -r^1, \quad \text{where } M = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix}.$$

此外，假设玩家对游戏规则一无所知，因此必须通过探索来发现这些规则。

RWS 是经典矩阵博弈“石头-布-剪刀”（RPS）的扩展，复杂度逐渐增加，并且保留了 RPS 的博弈论结构。由于纳什均衡意味着玩家库存中每种资源的数量保持不变，[82], 玩家在达到纳什均衡后往往会停留在某个网格附近，而不再继续探索；上层设计者则希望玩家能够探索更多网格。面对这种冲突，设计者通过散布金币来干预奖励，以此作为激励，引导玩家探索更多网格。

有四个网格会生成金币。其中两个填充了固定奖励 0.5，其余的填充了可调奖励，并标记为激励参数 $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$。一旦玩家 i收集金币后，相应的奖励会计入其总奖励。反之，玩家 i 收集到的奖励将从其对手 $-i$的总奖励中扣除。玩家 i的总奖励为

$$r^i_{total} = r^i + \sum_{j=1,2} \theta_j (I^i_j - I^{-i}_j) + 0.5(N^i_{\text{fixed}} - N^{-i}_{\text{fixed}}),$$

其中，$I^i_j = 1$ 若玩家 i 收集了带有激励奖励 θ_j 的金币，则 $I^i_j = 0$，否则 $I^i_j = 0$；N^ifixed 表示玩家 i 收集的固定金币数量。然后，探索率（ER）可定义为

$$ER(\theta) = \frac{\sum_i G^i_e(\theta)}{G_{\text{total}}}, \tag{7.1}$$

其中，$G^i_e(\theta)$ 表示玩家 i 已探索的网格数，Gtotal 表示网格总数。因此，设计者的损失定义为 $f_*(\theta) = 1 - ER(\theta)$。

捕食者-猎物（PP）。捕食者-猎物所处的环境设定在如图 3所示的 7×7 网格世界中，其中包含猎物巢穴、两个庇护所和水池。捕食者和猎物都不能停留在庇护所中。每一步，猎物可以移动一个网格，捕食者可以沿四个方向移动一到两个网格。博弈在以下三种情况下结束：i）回合达到最大长度25；ii）猎物返回巢穴，此时猎物的奖励 $r$prey 为 $+1$，捕食者的奖励 $r$pred 为 -1；iii）捕食者抓住猎物（双方同时到达同一网格），此时猎物的奖励 $r$prey 为 -1，而

捕食者的奖励 r_{pred} 为 $+1$。从整个生态系统的角度来看，例如植物种子的传播，设计者通过在四个固定网格中设置四个水量各异的水池，促使它们探索更多地方。其中两个水池注入固定水量，并附加0.1的奖励；另外两个水池注入可调水量，其奖励记为激励参数 $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$。玩家找到并喝掉水池中的水后，会获得相应的奖励，而其对手则会受到相应的惩罚。因此，捕食者的总收益为

$$r_{\text{捕食者}} = I_{\text{捕获}} - I_{\text{巢穴}} + \sum_{j=1,\ 2} \theta_j \left(I_{\text{捕食者 } j} - I_{\text{猎物 } j} \right) + 0.5 \left(N_{\text{捕食者固定}} - N_{\text{猎物固定}} \right),$$

其中，$I_{\text{捕获}}$和$I_{\text{巢穴}}$是两个指示变量，分别表示捕食者捕获猎物或猎物返回巢穴；$I_{\text{捕食者 } j}$（或$I_{\text{猎物 } j}$）是指示函数，用于表示捕食者（或猎物）是否发现了有 θ_j 水量的水池；$N_{\text{捕食者固定}}$（或 $N_{\text{猎物固定}}$）表示捕食者（或猎物）发现的固定水池数量。此外，我们将猎物的奖励定义为$r_{\text{猎物}} = -r_{\text{捕食者}}$。PP 的仲裁目标与 RWS 中定义的相同。

7.2 结果与分析

我们使用*GridSearch(M)* 表示零阶方法，该方法将 M 个点分布到激励参数 θ的可行域中，作为 DASAC 的竞争方法。此外，我们采用贝叶斯优化来提高零阶方法的效率，将其记为 *BayesOpt*，该方法也用于 [64]。GridSearch(M) 和 BayesOpt 的实验细节见附录 G.2。图 4 展示了在环境中应用 DASAC 后，仲裁目标f_* 的变化趋势。图 4 中的绿色点线表示 GridSearch(100) 的最佳目标得分。一方面，图 4 表明，我们提出的可微一阶框架能够以更低的目标损失获得更优的激励参数θ。另一方面，图 4 表明，GridSearch(M) 需要 100 个样本，换言之，需要将底层马尔可夫博弈求解 100 次，才能达到与 DASAC 相当的性能。尽管 BayesOpt（12 次外循环迭代，>2 比我们慢）提高了 GridSearch 的效率（100 次外循环迭代），DASAC 仍能比这些零阶方法实现更高的效率（5 次外循环迭代）。考虑梯度计算后，DASAC 完成一次迭代需要 3.5 小时，而零阶方法（网格搜索和

在相同的图形处理器配置下，贝叶斯优化每次迭代需要 3 小时。我们的方法比贝叶斯优化快 2 倍，比网格搜索（100）快 17 倍。这意味着，我们的一阶框架只需进行更少的评估，上层即可获得理想的纳什均衡策略，从而提高样本效率。

此外，我们将通过分析初始激励 θ_0 和最优激励 θ^* 下的纳什均衡行为（图 4 中标出），说明我们的 DASAC 倾向于达到一个理想的纳什均衡。图1展示了在 RWS 和 PP 环境中，玩家分别采用初始激励 θ_0 和最优激励θ^* 时纳什均衡策略所产生的轨迹。这表明，最优 θ^* 会引导系统走向更好的纳什均衡，使玩家更倾向于探索。具体来说，在初始激励参数 θ_0下，尽管玩家开始考虑探索率，并跳出无激励系统中的习惯性行为，

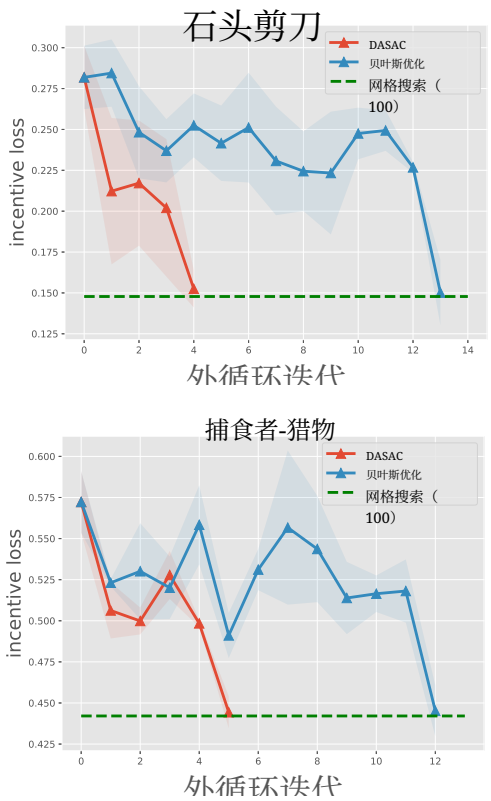


图 4：RWS（上）和捕食者-猎物（下）中 DASAC、贝叶斯优化和网格搜索（100）的仲裁目标损失 f_*。绿色虚线表示网格搜索（100）的最佳目标得分。我们的 DASAC 大幅优于两种零阶基线方法——网格搜索（100）和贝叶斯优化。这表明，我们的一阶框架只需进行更少的评估，上层即可获得理想的纳什均衡策略，从而提高样本效率。

例如，在 RWS 中几步之后停滞不前却不收集资源，或在 PP 中相对地处于对角位置而陷入僵局，但经过几次探索后，他们仍倾向于维持习惯性行为。随着激励参数的训练不断进行，无激励系统中的习惯性行为逐渐被抛弃，玩家也形成了新的行为模式，从而提高探索率，同时确保实现自身目标。例如，在 PP 中，捕食者仍尽力捕捉猎物，而猎物则尽可能迅速地逃离。不同之处在于，猎物倾向于四处游走，探索水源以补充能源，而不是待在原地并避免冒险行动。

8 结论

我们的工作首次在多智能体强化学习背景下提出了一个可证明的可微框架，用于解决双层仲裁问题。我们给出了收敛性证明，并通过实验证明 DASAC 在两个马尔可夫博弈中的仲裁效果。只要采用收敛性已通过实验证明的适当纳什均衡求解器，我们的工作就可以扩展到多玩家和一般和设置。因此，我们将在实验中测试

DA 框架在更一般情形下的性能。

参考文献

[1] Hadi Abbaszadehpeivasti、Etienne de Klerk 和 Moslem Zamani。2022。L-光滑函数在固定步长下梯度法的精确最坏情形收敛速率。《Optimization Letters》16，6（2022），1649–1661。[2] Zafarali Ahmed、Nicolas Le Roux、Mohammad Norouzi 和 Dale Schuurmans。2019。理解熵对策略优化的影响。载于《国际机器学习会议》。PMLR，151–160。[3] Ahmet Alacaoglu、Luca Viano、Niao He 和 Volkan Cevher。2022。零和马尔可夫博弈的自然演员-评论家框架。载于《国际机器学习会议》。PMLR，307–366。[4] Brandon Amos 和 J Zico Kolter。2017。Optnet：作为神经网络一层的可微优化。载于《国际机器学习会议》。PMLR，136–145。[5] Haris Aziz 和 Florian Brandl。2021。高效、公平且符合激励相容原则的医疗资源配给。arXiv 预印本 arXiv:2102.04384（2021）。[6] David Balduzzi、Karl Tuyls、Julien Perolat 和 Thore Graepel。2018。重新评估评估。arXiv 预印本 arXiv:1806.02643（2018）。[7] Jeffrey S Banks 和 Joel Sobel。1987。信号博弈中的均衡选择。《Econometrica: Journal of the Econometric Society》（1987），647–661。[8] Tamer Başar 和 Geert Jan Olsder。1998。动态非合作博弈论。SIAM。[9] Atilim Gunes Baydin、Robert Cornish、David Martinez Rubio、Mark Schmidt 和 Frank Wood。2018。在线学习率超梯度下降自适应。载于《国际学习表征会议》。[10] B Douglas Bernheim 和 Michael D Whinston。1987。联盟稳健纳什均衡 II：应用。《经济理论杂志》42，1（1987），13–29。[11] Patrick Bolton 和 Mathias Dewatripont。2004。契约理论。MIT 出版社。[12] Carlo Cambini 和 Laura Rondi。2010。激励规制与投资：来自欧洲能源公用事业的证据。《规制经济学杂志》38，1（2010），1–26。[13] Shicong Cen、Yuejie Chi、Simon S Du 和 Lin Xiao。2022。零和马尔可夫博弈中策略优化的末次迭代更快收敛。arXiv 预印本 arXiv:2210.01050（2022）。[14] Shicong Cen、Yuting Wei 和 Yuejie Chi。2021。具有熵正则化的竞争博弈快速策略外梯度法。《神经信息处理系统进展》34（2021），27952–27964。[15] Georgios Chalkiadakis 和 Craig Boutilier。2003。多智能体强化学习中的协调：贝叶斯方法。载于第二届自主智能体与多智能体系统国际联合会议论文集（澳大利亚墨尔本）（AAMAS '03）。计算机协会，纽约，NY，美国，

709 716 10 1145 860575 860689 – 。[16] 2021 https://doi.org/ 。/ 。Ziyi Chen、Shaocong Ma 和 Yi Zhou。 。零和马尔可夫博弈的样本高效随机策略外梯度算法。载于《国际学习表征会议》。[17] Maxime Chevalier-Boisvert、Lucas Willems 和 Suman Pal。2018。适用于 OpenAI Gym 的极简网格世界环境。https://github.com/maximecb/gym-minigrid。[18] Caroline Claus 和 Craig Boutilier。1998。合作式多智能体系统中强化学习的动态过程。AAAI/IAAI 1998，746-752（1998），2。[19] Benoît Colson、Patrice Marcotte 和 Gilles Savard。2007。双层优化概述。《运筹学年鉴》153，1（2007），235–256。[20] Bo Dai、Albert Shaw、Lihong Li、Lin Xiao、Niao He、Zhen Liu、Jianshu Chen 和 Le Song。2018。SBEED：采用非线性函数逼近的收敛强化学习。载于第 35 届国际机器学习会议论文集（《机器学习研究论文集》，第 80 卷），Jennifer Dy 和 Andreas Krause（编）。PMLR，1125–1134。https://proceedings.mlr.press/v80/dai18c.html[21] Sam Devlin 和 Daniel Kudenko。2011。多智能体系统中基于势函数的奖励塑形的理论考量。载于第 10 届自主智能体与多智能体系统国际会议。ACM，225–232。[22] Yali Du、Lei Han、Meng Fang、Ji Liu、Tianhong Dai 和 Dacheng Tao。2019。Liir：在多智能体强化学习中学习个体内在奖励。（2019）。[23] Pradeep Dubey。1986。纳什均衡的低效性。《运筹学数学》

11, 1 1986 1 8 3690047 ( ), –。[24] 2013 http://www.jstor.org/stable/ Fei Fang、Albert Xin Jiang 和 Milind Tambe。 。利用多个移动资源保护移动目标。《人工智能研究杂志》48（2013），583–634。[25] Xidong Feng、Oliver Slumbers、Ziyu Wan、Bo Liu、Stephen McAleer、Ying Wen、Jun Wang 和 Yaodong Yang。2021。双人零和博弈中的神经自适应课程。《神经信息处理系统进展》34（2021），3504–3517。[26] Chelsea Finn、Pieter Abbeel 和 Sergey Levine。2017。模型无关的元学习：实现深度网络的快速适应。载于《国际机器学习 g° PMLR，1126–1135。[27] Terry L Friesz、Roger L Tobin、Hsun-Jung Cho 和 Nihal J Mehta。1990。基于敏感性分析的带变分不等式约束的数学规划启发式算法。《运筹学数学》48，1（1990），265–284。[28] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。2020。生成对抗网络。《ACM通讯》63，11（2020），139–144。[29] Tuomas Haarnoja、Haoran Tang、Pieter Abbeel 和 Sergey Levine。2017。使用深度能量策略的强化学习。在《国际机器学习会议》。PMLR，1352–1361。[30] Tuomas Haarnoja、Aurick Zhou、Pieter Abbeel 和 Sergey Levine。2018。软演员-评论家：使用随机演员的离策略最大熵深度强化学习。在《国际机器学习会议》。PMLR，1861–1870。[31] Tuomas Haarnoja、Aurick Zhou、Kristian Hartikainen、George Tucker、Sehoon Ha、Jie Tan、Vikash Kumar、Henry Zhu、Abhishek Gupta、Pieter Abbeel 等。2018。软演员-评论家算法及其应用。arXiv预印本 arXiv:1812.05905（2018）。[32] Yuval Noah Harari。2014。《人类简史》。兰登书屋。[33] Johannes Heinrich、Marc Lanctot 和 David Silver。2015。扩展式博弈中的虚拟自我博弈。在《第32届国际机器学习会议论文集》（《机器学习研究论文集》，第37卷），Francis Bach 和 David Blei（编）。PMLR，法国里尔，805–813。https://proceedings.mlr.press/v37/heinrich15.html[34] Johannes Heinrich 和 David Silver。2015。计算机扑克中的光滑UCT搜索。在《第二十四届国际人工智能联合会议》。[35] Johannes Heinrich 和 David Silver。2016。不完全信息博弈中的自我博弈深度强化学习。arXiv预印本 arXiv:1603.01121（2016）。[36] Josef Hofbauer 和 Karl Sigmund。2003。演化博弈动力学。《美国数学学会公报》40，4（2003），479–519。[37] Junling Hu 和 Michael P Wellman。2003。一般和随机博弈的纳什Q学习。《机器学习研究》4，11月（2003），1039–1069。[38] Aly Ibrahim、Anirudha Jitani、Daoud Piracha 和 Doina Precup。 [不详。]多智能体强化学习中的奖励重新分配机制。（[不详]）。[39] Shariq Iqbal 和 Fei Sha。2019。多智能体强化学习的演员-注意力-评论家。在《国际机器学习会议》。PMLR，2961–2970。[40] Haoming Jiang、Zhehui Chen、Yuyang Shi、Bo Dai 和 Tuo Zhao。2021。通过学习攻击来学习防御。在《国际人工智能与统计会议》。PMLR，577–585。[41] Diederik P Kingma 和 Jimmy Ba。2014。Adam：一种随机优化方法。arXiv预印本 arXiv:1412.6980（2014）。[42] Zafeiris Kokkinogenis、Nuno Monteiro、Rosaldo JF Rossetti、Ana LC Bazzan 和 Pedro Campos。2014。策略与激励设计评估：面向社会的人工交通系统框架。在第17届国际IEEE智能交通系统会议（ITSC）。IEEE，151–156。[43] Vijay R Konda 和 John N Tsitsiklis。2000。演员-评论家算法。在《神经信息处理系统进展》。1008–1014。[44] Jean-Jacques Laffont 和 David Martimort。2009。《激励理论》。普林斯顿大学出版社。[45] Theodore J Lambert Iii、Marina A Epelman 和 Robert L Smith。2005。大规模优化的虚拟博弈方法。《运筹学》53，3（2005），477–489。[46] Kyungjae Lee、Sungjoon Choi 和 Songhwai Oh。2018。具有因果稀疏Tsallis熵正则化的稀疏马尔可夫决策过程及其强化学习。《IEEE机器人与自动化快报》3，3（2018），1466–1473。[47] Joel Z Leibo、Edward Hughes、Marc Lanctot 和 Thore Graepel。2019。自动课程与社会互动中创新的涌现：多智能体智能研究宣言。arXiv预印本 arXiv:1903.00742（2019）。[48] George Leitmann。1978。关于广义斯塔克尔伯格策略。《优化理论与应用杂志》26，4（1978），637–643。[49] Chenghao Li、Chengjie Wu、Tonghan Wang、Jun Yang、Qianchuan Zhao 和 Chongjie Zhang。2021。庆祝共享多智能体强化学习中的多样性。arXiv预印本 arXiv:2106.02195（2021）。[50] Huan Li、Cong Fang、Wotao Yin 和 Zhouchen Lin。2018。分布式加速梯度方法的锐利收敛速率分析。arXiv预印本 arXiv:1810.01053（2018）。[51] Jiayang Li、Jing Yu、Yu Nie 和 Zhaoran Wang。2020。博弈中的端到端学习与干预。《神经信息处理系统进展》33（2020）。[52] Tao Li 和 Suresh P Sethi。2017。动态斯塔克尔伯格博弈模型综述。《离散与连续动力系统-B》22，1（2017），125。[53] Michael L Littman。1994。作为多智能体强化学习框架的马尔可夫博弈。在《机器学习论文集1994》。Elsevier，157–163。[54] Boyi Liu、Jiayang Li、Zhuoran Yang、Hoi-To Wai、Mingyi Hong、Yu Marco Nie 和 Zhaoran Wang。2021。通过激励诱导均衡：同步设计与博弈找到全局最优解。arXiv预印本 arXiv:2110.01212（2021）。

[55] Boyi Liu、Zhuoran Yang 和 Zhaoran Wang。2020。零和马尔可夫博弈中的策略优化：虚拟自我博弈可证明达到纳什均衡。（2020）。[56] Jingbin Liu、Xinyang Gu 和 Shuai Liu。2019。具有熵正则化的策略优化强化学习。arXiv预印本 arXiv:1912.01557（2019）。[57] Ryan Lowe、Yi Wu、Aviv Tamar、Jean Harb、Pieter Abbeel 和 Igor Mordatch。2017。混合合作-竞争环境中的多智能体演员-评论家。在《第31届神经信息处理系统国际会议论文集》。6382–6393。[58] Zhi-Quan Luo、Jong-Shi Pang 和 Daniel Ralph。1996。带均衡约束的数学规划。剑桥大学出版社。[59] Xiaochen Ma、Hong Wang、Li Yang、Leiyu Shi 和 Xiaoyun Liu。2019。重新调整中国基层医疗卫生服务提供者的激励体系。《英国医学杂志》365（2019）。[60] Hugh Macartney、Robert McMillan 和 Uros Petronijevic。2016。教育中的激励设计：一项实证分析。技术报告。美国国家经济研究局。[61] H Brendan McMahan 和 Geoffrey J Gordon。2007。扑克及其他凸博弈的快速基于束的随时算法。在《人工智能与统计》。PMLR，323–330。[62] Joy Melnikow、Mary Paliescheskey 和 Gary K Stewart。1997。交通激励对首次产前预约依从性的影响：一项随机试验。《妇产科学》89，6（1997），1023–1027。[63] Panayotis Mertikopoulos 和 William H Sandholm。2016。通过强化与正则化进行博弈学习。《运筹学数学》41，4（2016），1297–1324。[64] David Mguni、Joel Jennings、Emilio Sison、Sergio Valcarcel Macua、Sofia Ceppi 和 Enrique Munoz de Cote。2019。协调群体：在非合作系统中诱导理想均衡。在《第18届自主智能体与多智能体系统国际会议论文集》。386–394。[65] Amir-Hamed Mohsenian-Rad、Vincent WS Wong、Juri Jatskevich 和 Robert Schober。2010。智能电网的最优自主基于激励的能源消耗调度算法。在《2010年创新智能电网技术会议》（ISGT）。IEEE，1–6。[66] Roger B Myerson。1978。纳什均衡概念的精化。《国际博弈论杂志》7，2（1978），73–80。[67] Alex Nichol、Joshua Achiam 和 John Schulman。2018。关于一阶元学习算法。arXiv预印本 arXiv:1803.02999（2018）。[68] Shayegan Omidshafiei、Christos Papadimitriou、Georgios Piliouras、Karl Tuyls、Mark Rowland、Jean-Baptiste Lespiau、Wojciech M Czarnecki、Marc Lanctot、Julien Perolat 和 Remi Munos。2019。 α-排名：通过演化进行多智能体评估。《科学报告》9，1（2019），1–29。[69] Jack Parker-Holder、Luke Metz、Cinjon Resnick、Hengyuan Hu、Adam Lerer、Alistair Letcher、Alex Peysakhovich、Aldo Pacchiano 和 Jakob Foerster。2020。Ridge Rider：通过沿海森矩阵的特征向量寻找多样化解。arXiv预印本 arXiv:2011.06505（2020）。[70] A.C. Pigou 和 N. Aslanbeigui。1920。《福利经济学》（第1版）。劳特利奇出版社。[71] Aida Rahmattalabi、Jen Jen Chung、Mitchell Colby 和 Kagan Tumer。2016。D++：紧密耦合多智能体领域中的结构化信用分配。在《2016年IEEE/RSJ智能机器人与系统国际会议》（IROS）。IEEE，4424–4429。[72] Ali Ramezani-Kebrya、Ashish Khisti 和 Ben Liang。2018。关于带动量的随机梯度下降的泛化。arXiv预印本 arXiv:1809.04564（2018）。[73] Lillian J Ratliff、Roy Dong、Shreyas Sekar 和 Tanner Fiez。2019。激励设计视角：挑战与机遇。《控制、机器人与自主系统年度评论》2（2019），305–338。[74] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。2017。近端策略优化算法。arXiv预印本 arXiv:1707.06347（2017）。[75] Reinhard Selten。1965。具有需求迟滞的寡头模型的博弈论处理：第一部分：动态价格均衡的确定。《国家科学总汇杂志》/《制度与理论经济学杂志》H. 2（1965），301–324。[76] Lloyd S Shapley。1953。随机博弈。《美国国家科学院院刊》39，10（1953），1095–1100。[77] Satinder Singh、Tommi Jaakkola、Michael L Littman 和 Csaba Szepesvári。2000。单步同策略强化学习算法的收敛性结果。《机器学习》38，3（2000），287–308。[78] Chen Siyu、Yang Donglin、Li Jiayang、Wang Senmiao、Yang Zhuoran 和 Wang Zhaoran。2022。马尔可夫决策过程的自适应模型设计。技术报告。[79] Xingyou Song、Wenbo Gao、Yuxiang Yang、Krzysztof Choromanski、Aldo Pacchiano 和 Yunhao Tang。2020。ES-MAML：海森矩阵自由元学习的简单方法。在《国际学习表征会议》。[80] Csaba Szepesvári 和 Michael L Littman。1999。基于价值函数的强化学习算法的统一分析。《神经计算》11，8（1999），2017–2060。

[81] Zhenggang Tang、Chao Yu、Boyuan Chen、Huazhe Xu、Xiaolong Wang、Fei Fang、Simon Shaolei Du、Yu Wang 和 Yi Wu。2021。《通过奖励随机化发现多样化的多智能体策略行为》。载于国际学习表征会议。https://openreview.net/forum?id=lvRTC669EY

[82] 2007 Anne van den Nouweland。。《石头剪刀布；一种新颖而优雅的证明》。（2007）。[83] Alexander Vezhnevets、Yuhuai Wu、Maria Eckstein、Rémi Leblond 和 Joel Z Leibo。2020。《选项作为响应：在多智能体强化学习中为行为层级奠定基础》。载于国际机器学习会议。PMLR，9733–9742。[84] John Von Neumann 和 Oskar Morgenstern。2007。《博弈论与经济行为》。载于博弈论与经济行为。普林斯顿大学出版社。[85] Kai Wang、Sanket Shah、Haipeng Chen、Andrew Perrault、Finale Doshi-Velez 和 Milind Tambe。2021。《从特征中学习马尔可夫决策过程：通过强化学习进行先预测后优化的序贯决策》。神经信息处理系统进展34（2021），8795–8806。[86] Kai Wang、Lily Xu、Andrew Perrault、Michael K Reiter 和 Milind Tambe。2022。《协调跟随者以达成更优均衡：面向斯塔克尔伯格博弈的端到端梯度下降》。载于*AAAI*人工智能会议论文集，第36卷。5219–5227。[87] Xiaoxing Wang、Wenxuan Guo、Junchi Yan、Jianlin Su 和 Xiaokang Yang。2021。《ZARTS：神经架构搜索的零阶优化》。*arXiv*预印本 *arXiv:2110.04743*（2021）。[88] Thomas A Weber 和 AV Kryazhimskiy。2011。《最优控制理论及其在经济学中的应用》。第10卷。MIT 出版社，马萨诸塞州剑桥。[89] Changnan Xiao、Haosen Shi、Jiajun Fan 和 Shihong Deng。2021。《基于策略的强化学习的无熵正则化机制》。*arXiv*预印本 *arXiv:2106.00707*（2021）。[90] Jiachen Yang、Ang Li、Mehrdad Farajtabar、Peter Sunehag、Edward Hughes 和 Hongyuan Zha。2020。《学习激励其他学习智能体》。神经信息处理系统进展33（2020）。[91] Jiachen Yang、Ethan Wang、Rakshit Trivedi、Tuo Zhao 和 Hongyuan Zha。2022。《通过多智能体元梯度强化学习进行自适应激励设计》。载于第*21*届国际自主智能体与多智能体系统会议论文集。1436–1445。[92] Tianbao Yang、Qihang Lin 和 Zhe Li。2016。《凸优化与非凸优化中随机动量方法的统一收敛性分析》。*arXiv*预印本 *arXiv:1604.03257*（2016）。[93] Rui Zhao、Xudong Sun 和 Volker Tresp。2019。《最大熵正则化多目标强化学习》。载于第*36*届国际机器学习会议论文集（《机器学习研究论文集》，第*97*卷），Kamalika Chaudhuri 和 Ruslan Salakhutdinov（编）。PMLR，7553–7562。https://proceedings.mlr.press/v97/zhao19d.html [94] Yulai Zhao、Yuandong Tian、Jason Lee 和 Simon Du。2022。《双人零和马尔可夫博弈中策略优化的可证明效率》。载于国际人工智能与统计会议。PMLR，2736–2761。[95] Han Zhong、Zhuoran Yang、Zhaoran Wang 和 Michael I Jordan。2021。《在具有短视跟随者的一般和马尔可夫博弈中，强化学习能否找到斯塔克尔伯格-纳什均衡？》*arXiv*预印本 *arXiv:2112.13521*（2021）。

零和马尔可夫博弈中的可微仲裁

目录

section*.1section.1section.2section.3subsection.3.1subsection.3.2subsection.3.3subsection.3.4section.4subsection.4.1subsection.4.2section.5subsection.5.1subsecti

A 引理 4.1 的证明 11
B 引理 4.2 的证明 15
C DA 框架的扩展 15
D PEM 和熵正则化 OMWU 22
D.1 策略外梯度法（PEM） 22
D.2 熵正则化 OMWU 22
E 假设 5.3 的说明 23
E.1 引理 E.3 的证明 25
F 定理 5.4 的证明 27
G 实验细节 28
G.1 DASAC 28
G.2 GridSearch(M) 和贝叶斯优化基线 29
G.3 实现细节 29

引理 4.1 的证明

本节中，我们给出梯度 $\nabla_\theta V^{i}_{\pi_\phi}(s;\,\theta)$ $\nabla_{\phi^i} V^{i}_{\pi_\phi}(s;\,\theta)$ $\nabla^2_{\theta\phi^i} V^{i}_{\pi_\phi}(s;\,\theta)$ 和 $\nabla^2_{\phi\phi^i} V^{i}_{\pi_\phi}(s;\,\theta)$ 的表述。我们首先在下述引理 A.1 中介绍 $V^i_{\pi_\phi}(s;\,\theta)$ 的梯度推导，随后证明引理 4.1。

引理 A.1. 对于激励马尔可夫博弈 $\mathcal{G}_\theta = (\mathcal{N}、\mathcal{S}、\{\mathcal{A}^i\}_{i\in\{1,2\}}、\mathcal{P}、\{r^i(\cdot;\;\theta)\}_{i\in\{1,2\}}\ \gamma)$，令 $R^i(\tau;\;\theta)$ 表示从初始状态 $s_0 = s$ 出发，沿着 π_ϕ 进行 T 步所得到的样本轨迹 τ 的总奖励。我们有

$$
\begin{aligned}
\nabla_\theta V^i_{\pi_\phi}(s;\theta) &= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\nabla_\theta R^i(\tau;\theta)|s_0 = s\right],\\
\nabla_{\phi^i} V^i_{\pi_{\phi^i}}(s;\theta) &= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[R^i(\tau;\theta)\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a_t|s_t)\Big|s_0 = s\right],\\
\nabla^2_{\theta\phi^i} V^i_{\pi_\phi}(s;\theta) &= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a_t|s_t)\right)\nabla_\theta R^i(\tau;\theta)^\intercal\Big|s_0 = s\right],\\
\nabla^2_{\phi\phi^i} V^i_{\pi_\phi}(s;\theta) &= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[R^i(\tau;\theta)\left((\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a_t|s_t))(\sum_{t=0}^{T-1}\nabla_{\phi}\log\pi_{\phi}(a_t|s_t))^\intercal\right.\right.\\
&\qquad\left.\left.+\sum_{t=0}^{T-1}\nabla^2_{\phi\phi^i}\log\pi_\phi(a_t|s_t)\right)\Big|s_0 = s\right],
\end{aligned}
$$

其中

$$
\begin{aligned}
\nabla_\phi\log\pi_\phi(a_t|s_t) &= (\nabla_{\phi^1}\log\pi^1_{\phi^1}(a_t|s_t), \nabla_{\phi^2}\log\pi^2_{\phi^2}(a_t|s_t))\\
\nabla^2_{\phi\phi^i}\log\pi_\phi(a_t|s_t) &= (\nabla^2_{\phi^1\phi^i}\log\pi^1_{\phi^1}(a_t|s_t), \nabla^2_{\phi^2\phi^i}\log\pi^2_{\phi^2}(a_t|s_t)).
\end{aligned}
$$

引理 A.1 的证明。首先，我们求出 $\nabla_{\phi^i}\log\ D^{\pi_\phi}(\tau|s_0 = s)$：

$$
\begin{aligned}
\nabla_{\phi^i}\log D^{\pi_\phi}(\tau|s_0 = s) &= \nabla_{\phi^i}\log\left[\prod_{t=0}^{T-1}P(s_{t+1}|s_t,a_t)\left(\prod_{i=1}^{2}\pi^i_{\phi^i}(a^i_t|s_t)\right)\right]\\
&= \nabla_{\phi^i}\left[\sum_{t=0}^{T-1}\log P(s_{t+1}|s_t,a_t) + \sum_{t=0}^{T-1}\log\pi^i_{\phi^i}(a_t|s_t) + \sum_{t=0}^{T-1}\log\pi^{-i}_{\phi^{-i}}(a_t|s_t)\right] = \sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a_t|s_t).
\end{aligned}
$$

类似地，我们有 $\nabla_\phi \log D^{\pi_\phi}(\tau|s_0=s) = \sum_{t=0}^{T-1} \nabla_\phi \log \pi_\phi(a_t|s_t)$，其中 $\nabla_\phi \log \pi_\phi(a_t|s_t) = (\nabla_{\phi^1} \log \pi^1_{\phi^1}(a_t|s_t)\, \nabla_{\phi^2} \log \pi^2_{\phi^2}(a_t|s_t))$。此外，回想一下 $\nabla_{\phi\phi^i} \log \pi_\phi(a_t|s_t) = (\nabla_{\phi^1\phi^i} \log \pi^1_{\phi^1}(a_t|s_t)\, \nabla_{\phi^2\phi^i} \log \pi^2_{\phi^2}(a_t|s_t))$，我们可以推导出 $\nabla_\theta V^i_{\pi_\phi}(s;\theta)$、$\nabla^2_{\phi^i\theta} V^i_{\pi_\phi}(s;\theta)$、$\nabla_{\phi^i} V^i_{\pi_\phi}(s;\theta)$ 和 $\nabla^2_{\phi\phi^i} V^i_{\pi_\phi}(s;\theta)$，如下所示，

$$
\begin{aligned}
\nabla_\theta V^i_{\pi_\phi}(s;\theta) &= \nabla_\theta \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[R^i(\tau;\theta)|s_0=s\right] = \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\nabla_\theta R^i(\tau;\theta)|s_0=s\right], \\
\nabla_{\phi^i} V^i_{\pi_\phi}(s;\theta) &= \nabla_{\phi^i} \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[R^i(\tau;\theta)|s_0=s\right] = \nabla_{\phi^i}\left[\sum_{\tau\sim D^{\pi_\phi}} D^{\pi_\phi}(\tau|s_0=s) R^i(\tau;\theta)\right] \\
&= \sum_{\tau\sim D^{\pi_\phi}} D^{\pi_\phi}(\tau|s_0=s) \nabla_{\phi^i} \log D^{\pi_\phi}(\tau|s_0=s) R^i(\tau;\theta) \\
&= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\nabla_{\phi^i} \log D^{\pi_\phi}(\tau|s_0=s) R^i(\tau;\theta)|s_0=s\right] \\
&= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[R^i(\tau;\theta) \sum_{t=0}^{T-1} \nabla_{\phi^i} \log \pi^i_{\phi^i}(a_t|s_t)|s_0=s\right],
\end{aligned}
$$

其中第三个等式是由 $\nabla_{\phi^i} D^{\pi_\phi} = D^{\pi_\phi} \frac{\nabla_{\phi^i} D^{\pi_\phi}}{D^{\pi_\phi}} = D^{\pi_\phi} \underline{\nabla_{\phi^i} \log} D^{\pi_\phi}$ 以及交换梯度与求和的顺序得到的。

$$
\begin{aligned}
\nabla^2_{\theta\phi^i} V^i_{\pi_\phi}(s;\theta) &= \nabla_\theta \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[R^i(\tau;\theta) \sum_{t=0}^{T-1} \nabla_{\phi^i} \log \pi^i_{\phi^i}(a_t|s_t)|s_0=s\right] \\
&= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\left(\sum_{t=0}^{T-1} \nabla_{\phi^i} \log \pi^i_{\phi^i}(a_t|s_t)\right) \nabla_\theta R^i(\tau;\theta)^\intercal |s_0=s\right], \\
\nabla^2_{\phi\phi^i} V^i_{\pi_\phi}(s;\theta) &= \nabla_{\phi\phi^i}\left[\sum_{\tau\sim D^{\pi_\phi}} D^{\pi_\phi}(\tau|s_0=s) R^i(\tau;\theta)\right] \\
&= \nabla_\phi\left[\sum_{\tau\sim D^{\pi_\phi}} D^{\pi_\phi}(\tau|s_0=s) \nabla_{\phi^i} \log D^{\pi_\phi}(\tau|s_0=s) R^i(\tau;\theta)\right] \\
&= \nabla_\phi\left[\sum_{\tau\sim D^{\pi_\phi}} D^{\pi_\phi}(\tau|s_0=s)\left(\sum_{t=0}^{T-1} \nabla_{\phi^i} \log \pi^i_{\phi^i}(a_t|s_t)\right) R^i(\tau;\theta)\right] \\
&= \sum_{\tau\sim D^{\pi_\phi}} R^i(\tau;\theta)\left[\left(\sum_{t=0}^{T-1} \nabla_{\phi^i} \log \pi^i_{\phi^i}(a_t|s_t)\right) \nabla_\phi D^{\pi_\phi}(\tau|s_0=s)^\intercal \right. \\
&\qquad \left. + D^{\pi_\phi}(\tau|s_0=s)\left(\sum_{t=0}^{T-1} \nabla_{\phi\phi^i} \log \pi_\phi(a_t|s_t)\right)\right] \\
&= \sum_{\tau\sim D^{\pi_\phi}} R^i(\tau;\theta) D^{\pi_\phi}(\tau|s_0=s)\left[\left(\sum_{t=0}^{T-1} \nabla_{\phi^i} \log \pi^i_{\phi^i}(a_t|s_t)\right) \nabla_\phi \log D^{\pi_\phi}(\tau|s_0=s)^\intercal \right. \\
&\qquad \left. + \sum_{t=0}^{T-1} \nabla_{\phi\phi^i} \log \pi_\phi(a_t|s_t)\right] \\
&= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[R^i(\tau;\theta)\left((\sum_{t=0}^{T-1} \nabla_{\phi^i} \log \pi^i_{\phi^i}(a_t|s_t))(\sum_{t=0}^{T-1} \nabla_\phi \log \pi_\phi(a_t|s_t))^\intercal \right.\right. \\
&\qquad \left.\left. + \sum_{t=0}^{T-1} \nabla^2_{\phi\phi^i} \log \pi_\phi(a_t|s_t)\right)\Big| s_0=s\right].
\end{aligned}
$$

□

在本节剩余部分，我们将完成正文中引理 4.1 的证明。

引理 4.1 的证明。回顾式 (3.5) 中 $V_{(i)^{\pi_\phi}}(s;\,\theta)$ 和式 (3.3) 中 $r_\pi^i(s,a^i,a^{-i};\,\theta)$ 的定义，可得，

$$\begin{aligned}
V_{\pi_\phi}^{(i)}(s;\theta) &= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\sum_{t=0}^{T-1}\gamma^t\cdot r_\pi^{(i)}(s_t,a_t^i,a_t^{-i};\theta)|s_0=s\right] \\
&= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\sum_{t=0}^{T-1}\gamma^t\cdot r_\pi^{i}(s_t,a_t^i,a_t^{-i};\theta) - \lambda\sum_{t=0}^{T-1}\gamma^t\left(\log\pi_{\phi^i}^i(a_t^i|s_t)-\log\pi_{\phi^{-i}}^{-i}(a_t^{-i}|s_t)\right)|s_0=s\right] \\
&= V_{\pi_\phi}^i(s;\theta) - \lambda\mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\sum_{t=0}^{T-1}\gamma^t\left(\log\pi_{\phi^i}^i(a_t^i|s_t)-\log\pi_{\phi^{-i}}^{-i}(a_t^{-i}|s_t)\right)|s_0=s\right]. \qquad \text{(A.1)}
\end{aligned}$$

为便于记号表示，我们定义

$$U_{\pi_\phi}^i(s) = \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\sum_{t=0}^{T-1}\gamma^t\left(\log\pi_{\phi^i}^i(a_t^i|s_t)-\log\pi_{\phi^{-i}}^{-i}(a_t^{-i}|s_t)\right)|s_0=s\right].$$

于是我们有，

$$\nabla_\theta U_{\pi_\phi}^i(s) = 0; \qquad \text{(A.2)}$$

$$\begin{aligned}
\nabla_{\phi^i}U_{\pi_\phi}^i(s) =& \nabla_{\phi^i}\left[\sum_\tau D^{\pi_\phi}(\tau|s_0=s)\sum_{t=0}^{T-1}\gamma^t\left(\log\pi_{\phi^i}^i(a_t^i|s_t)-\log\pi_{\phi^{-i}}^{-i}(a_t^{-i}|s_t)\right)\right] \\
=& \sum_\tau D^{\pi_\phi}(\tau|s_0=s)\left[\nabla_{\phi_i}\log D^{\pi_\phi}(\tau|s_0=s)\sum_{t=0}^{T-1}\gamma^t\left(\log\pi_{\phi^i}^i(a_t^i|s_t)-\log\pi_{\phi^{-i}}^{-i}(a_t^{-i}|s_t)\right)\right. \\
&\left.+\sum_{t=0}^{T-1}\gamma^t\nabla_{\phi^i}\log\pi_{\phi^i}^i(a_t^i|s_t)\right] \\
=& \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi_{\phi^i}^i(a_t^i|s_t)-\log\pi_{\phi^{-i}}^{-i}(a_t^{-i}|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi_{\phi^i}^i(a_t^i|s_t)\right)\right. \\
&\left.+\sum_{t=0}^{T-1}\gamma^t\nabla_{\phi^i}\log\pi_{\phi^i}^i(a_t^i|s_t)\right]; \qquad \text{(A.3)}
\end{aligned}$$

$$\nabla_{\theta\phi^i}U_{\pi_\phi}^i(s) = 0; \qquad \text{(A.4)}$$

$$\begin{aligned}
\nabla_{\phi\phi^i}U_{\pi_\phi}^i(s) =& \overbrace{\nabla_\phi\left[\sum_\tau D^{\pi_\phi}(\tau|s_0=s)\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi_{\phi^i}^i(a_t^i|s_t)-\log\pi_{\phi^{-i}}^{-i}(a_t^{-i}|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi_{\phi^i}^i(a_t^i|s_t)\right)\right]}^{:=(I)} \\
&+\underbrace{\nabla_\phi\left[\sum_\tau D^{\pi_\phi}(\tau|s_0=s)\sum_{t=0}^{T-1}\gamma^t\nabla_{\phi^i}\log\pi_{\phi^i}^i(a_t^i|s_t)\right]}_{:=(II)}.
\end{aligned}$$

接下来我们将计算（I）和（II）。

$$\begin{aligned}
(I) =& \sum_\tau D^{\pi_\phi}(\tau|s_0=s)\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi_{\phi^i}^i(a_t^i|s_t)-\log\pi_{\phi^{-i}}^{-i}(a_t^{-i}|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi_{\phi^i}^i(a_t^i|s_t)\right)\nabla_\phi\log D^{\pi_\phi}(\tau|s_0=s)^\intercal \\
&+\sum_\tau D^{\pi_\phi}(\tau|s_0=s)\nabla_\phi\left[\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi_{\phi^i}^i(a_t^i|s_t)-\log\pi_{\phi^{-i}}^{-i}(a_t^{-i}|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi_{\phi^i}^i(a_t^i|s_t)\right)\right]
\end{aligned}$$

$$
\begin{aligned}
(I) =& \sum_{\tau} D^{\pi_\phi}(\tau|s_0=s)\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t)-\log\pi^{-i}_{\phi^{-i}}(a^{-i}_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|s_t)\right)^\intercal \\
&+\sum_{\tau} D^{\pi_\phi}(\tau|s_0=s)\left[\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\left(\sum_{t=0}^{T-1}\gamma^t\left(\nabla_\phi\log\pi^i_{\phi^i}(a^i_t|s_t)-\nabla_\phi\log\pi^{-i}_{\phi^{-i}}(a^{-i}_t|s_t)\right)\right)^\intercal\right. \\
&\qquad\left.+\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t)-\log\pi^{-i}_{\phi^{-i}}(a^{-i}_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla^2_{\phi\phi^i}\log\pi_\phi(a^i_t|s_t)\right)\right] \\
=&\mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t)-\log\pi^{-i}_{\phi^{-i}}(a^{-i}_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|s_t)\right)^\intercal\right. \\
&+\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\left(\sum_{t=0}^{T-1}\gamma^t\left(\nabla_\phi\log\pi^i_{\phi^i}(a^i_t|s_t)-\nabla_\phi\log\pi^{-i}_{\phi^{-i}}(a^{-i}_t|s_t)\right)\right)^\intercal \\
&\left.+\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t)-\log\pi^{-i}_{\phi^{-i}}(a^{-i}_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla^2_{\phi\phi^i}\log\pi_\phi(a^i_t|s_t)\right)\right];
\end{aligned}
$$

$$
\begin{aligned}
(II) =& \sum_{\tau} D^{\pi_\phi}(\tau|s_0=s)\left[\left(\sum_{t=0}^{T-1}\gamma^t\nabla_{\phi^i}\log\pi^i_{\phi^i}(a_t|s_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|s_t)\right)^\intercal\right. \\
&\left.+\left(\sum_{t=0}^{T-1}\gamma^t\nabla^2_{\phi\phi^i}\log\pi_\phi(a_t|s_t)\right)\right] \\
=&\mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\left(\sum_{t=0}^{T-1}\gamma^t\nabla_{\phi^i}\log\pi^i_{\phi^i}(a_t|s_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|s_t)\right)^\intercal+\left(\sum_{t=0}^{T-1}\gamma^t\nabla^2_{\phi\phi^i}\log\pi_\phi(a_t|s_t)\right)\right].
\end{aligned}
$$

因此,

$$
\begin{aligned}
\nabla_{\phi\phi^i}U^i_{\pi_\phi}(s) =& \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t)-\log\pi^{-i}_{\phi^{-i}}(a^{-i}_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|s_t)\right)^\intercal\right. \\
&+\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\left(\sum_{t=0}^{T-1}\gamma^t\left(\nabla_\phi\log\pi^i_{\phi^i}(a^i_t|s_t)-\nabla_\phi\log\pi^{-i}_{\phi^{-i}}(a^{-i}_t|s_t)\right)\right)^\intercal \\
&+\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t)-\log\pi^{-i}_{\phi^{-i}}(a^{-i}_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla^2_{\phi\phi^i}\log\pi_\phi(a^i_t|s_t)\right) \\
&\left.+\left(\sum_{t=0}^{T-1}\gamma^t\nabla_{\phi^i}\log\pi^i_{\phi^i}(a_t|s_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|s_t)\right)^\intercal+\left(\sum_{t=0}^{T-1}\gamma^t\nabla^2_{\phi\phi^i}\log\pi_\phi(a_t|s_t)\right)\right]
\end{aligned}
\tag{A.5}
$$

由于 **(A.1)** 可以写成

$$
V^{(i)}_{\pi_\phi}(s;\theta)=V^i_{\pi_\phi}(s;\theta)-\lambda U^i_{\pi_\phi}(s),
$$

那么结合 (A.2-A.5)，我们有

$$\nabla_\theta V^{(i)}_{\pi_\phi}(s;\theta) = \nabla_\theta V^i_{\pi_\phi}(s;\theta);$$

$$\nabla_{\phi^i} V^{(i)}_{\pi_\phi}(s;\theta) = \nabla_{\phi^i} V^i_{\pi_\phi}(s;\theta) - \lambda \mathbb{E}_{\tau\sim D^{\pi_\phi}} \left[\left(\sum_{t=0}^{T-1} \gamma^t \left(\log \pi^i_{\phi^i}(a^i_t|s_t) - \log \pi^{-i}_{\phi^{-i}}(a^{-i}_t|s_t) \right) \right) \left(\sum_{t=0}^{T-1} \nabla_{\phi^i} \log \pi^i_{\phi^i}(a^i_t|s_t) \right) + \sum_{t=0}^{T-1} \gamma^t \nabla_{\phi^i} \log \pi^i_{\phi^i}(a^i_t|s_t) \right];$$

$$\nabla^2_{\theta\phi^i} V^{(i)}_{\pi_\phi}(s;\theta) = \nabla^2_{\theta\phi^i} V^i_{\pi_\phi}(s;\theta);$$

$$\begin{aligned}
\nabla^2_{\phi\phi^i} V^{(i)}_{\pi_\phi}(s;\theta) = \nabla^2_{\phi\phi^i} V^i_{\pi_\phi}(s;\theta) - \lambda \mathbb{E}_{\tau\sim D^{\pi_\phi}} \Bigg[& \left(\sum_{t=0}^{T-1} \gamma^t \left(\log \pi^i_{\phi^i}(a^i_t|s_t) - \log \pi^{-i}_{\phi^{-i}}(a^{-i}_t|s_t) \right) \right) \left(\sum_{t=0}^{T-1} \nabla_{\phi^i} \log \pi^i_{\phi^i}(a^i_t|s_t) \right) \left(\sum_{t=0}^{T-1} \nabla_\phi \log \pi_\phi(a_t|s_t) \right)^\top \\
& + \left(\sum_{t=0}^{T-1} \nabla_{\phi^i} \log \pi^i_{\phi^i}(a^i_t|s_t) \right) \left(\sum_{t=0}^{T-1} \gamma^t \left(\nabla_\phi \log \pi^i_{\phi^i}(a^i_t|s_t) - \nabla_\phi \log \pi^{-i}_{\phi^{-i}}(a^{-i}_t|s_t) \right) \right)^\top \\
& + \left(\sum_{t=0}^{T-1} \gamma^t \left(\log \pi^i_{\phi^i}(a^i_t|s_t) - \log \pi^{-i}_{\phi^{-i}}(a^{-i}_t|s_t) \right) \right) \left(\sum_{t=0}^{T-1} \nabla^2_{\phi\phi^i} \log \pi_\phi(a^i_t|s_t) \right) \\
& + \left(\sum_{t=0}^{T-1} \gamma^t \nabla_{\phi^i} \log \pi^i_{\phi^i}(a_t|s_t) \right) \left(\sum_{t=0}^{T-1} \nabla_\phi \log \pi_\phi(a_t|s_t) \right)^\top + \left(\sum_{t=0}^{T-1} \gamma^t \nabla^2_{\phi\phi^i} \log \pi_\phi(a_t|s_t) \right) \Bigg]
\end{aligned}$$

□

B 引理 4.2 的证明

本节基于引理 4.1 证明引理 4.2。

引理 4.2 的证明。由于 $\phi^{(*)}(\theta)$ 是正则化马尔可夫博弈中、激励奖励为 G_θ 的纳什均衡 $\pi_{\phi^{(*)}(\theta)}$ 的策略参数，因此 $u_\theta(\phi^{(*)}(\theta)) = (\mathbb{E}_{\nu^*} \nabla_{\phi^1} V^{\pi_\phi}_{(1)}(\theta)\ \mathbb{E}_{\nu^*} \nabla_{\phi^2} V^{\pi_\phi}_{(2)}(\theta))\big|_{\phi=\phi^{(*)}(\theta)} = 0$。然后，对等式两边关于 θ 求导，对于任意 $i \in \{1,2\}$，我们有

$$\mathbb{E}_{\nu^*} \nabla_{\theta\phi^i} V^{(i)}_{\pi_\phi}(\theta) + \mathbb{E}_{\nu^*} \nabla_{\phi\phi^i} V^{(i)}_{\pi_\phi}(\theta) [\nabla_\theta \phi^{(*)}(\theta)] = 0.$$

因此，定义

$$\nabla_\theta u_\theta(\phi) = \begin{bmatrix} \mathbb{E}_{\nu^*} \nabla_{\theta\phi^1} V^{(1)}_{\pi_\phi}(\theta) \\ \mathbb{E}_{\nu^*} \nabla_{\theta\phi^2} V^{(2)}_{\pi_\phi}(\theta) \end{bmatrix} \in \mathbb{R}^{d\times m},$$

$$\nabla_\phi u_\theta(\phi) = \begin{bmatrix} \mathbb{E}_{\nu^*} \nabla_{\phi\phi^1} V^{(1)}_{\pi_\phi}(\theta) \\ \mathbb{E}_{\nu^*} \nabla_{\phi\phi^2} V^{(2)}_{\pi_\phi}(\theta) \end{bmatrix} \in \mathbb{R}^{d\times d},$$

其中 ν^* 是 $\nu_{\pi_{\phi^{(*)}(\theta)}}(s)$。我们有，

$$\nabla_\theta \phi^{(*)}(\theta) = -\left[\nabla_\phi u_\theta(\phi)\right]^{-1} \nabla_\theta u_\theta(\phi)\big|_{\phi=\phi^*(\theta)}.$$

□

C DA 框架的扩展

请注意，只要下层纳什均衡求解器适当，DA 框架本身具有通用性。本节将详细介绍 DA 框架在两种一般情形下的扩展：i）N 个玩家 且对奖励不作限制；ii）部分观测。由于 DA 框架的核心部分是在纳什均衡处推导上层损失梯度 $\nabla f_*(\theta)$，因此下文将介绍在扩展设置下，如何根据下层提供的策略梯度信息计算 $\nabla f_*(\theta)$。

扩展到 N 玩家且不限制奖励。我们首先将包含 N 个玩家的激励马尔可夫博弈定义为 $G_\theta = (S, \{A^i, r^i\ \theta\ \lambda\ \tau\ \theta\ s\ \theta\ i \in 1 \dots N\ (\cdot;\theta)\}_{i\in\{1,\cdots N\}}, P\ \gamma)$。随后，对所有 $\{,\ ,\}$ 定义总折扣奖励 $R^i_\pi(;)$ 和价值函数 $V^i_\pi(;)$，如下所示：

$$R^i_\pi(\tau;\theta) = \sum_{t=0}^{T-1} \gamma^t r^i(s_t, a^i_t, a^{-i}_t;\theta), \tag{C.1}$$

$$\begin{aligned}
V^i_\pi(s;\theta) &= \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma^t \cdot r^i_\pi(s_t, a^i_t, a^{-i}_t;\theta) | s_0 = s \right] \\
&= \mathbb{E}_{\tau\sim D^\pi} \left[R^i_\pi(\tau;\theta) \big| s_0 = s \right].
\end{aligned} \tag{C.2}$$

然后定义熵正则化的 $\mathrm{G}_{r\theta} = (\mathrm{S},\ \{\mathrm{A}^i,\ r_{(i)\pi}(\cdot;\theta)\}_{i\in\{1,\ \cdots\ N\}}\lambda,\ \mathrm{P}\ \gamma)$，其中 $r_{(i)}(\cdot;\theta) : \mathrm{S}\ \times\mathrm{A}^i\times\mathrm{A}^{-i}\to\mathrm{R}$ 定义为：

$$r_\pi^{(i)}(\cdot;\theta) = r^i(\cdot;\theta) - \lambda\log(\pi^i(a^i|s)) + \frac{\lambda}{N-1}\sum_{\substack{j=1,\ldots,n\\ j\neq i}}\log(\pi^j(a^j|s)).$$

那么，熵正则化价值函数为

$$R_\pi^{(i)}(\tau;\theta) = \sum_{t=0}^{T-1}\gamma^t r_\pi^{(i)}(s_t, a_t^i, a_t^{-i};\theta), \tag{C.3}$$

$$\begin{aligned} V_\pi^{(i)}(s;\theta) &= \mathbb{E}_\pi\left[\sum_{t=0}^{T-1}\gamma^t\cdot r_\pi^{(i)}(s_t, a_t^i, a_t^{-i};\theta)|s_0 = s\right] \\ &= \mathbb{E}_{\tau\sim D^\pi}\left[R_\pi^{(i)}(\tau;\theta)\middle|s_0 = s\right]. \end{aligned} \tag{C.4}$$

除将零和约束改为 $r^i(\cdot;\theta)$ 外，其他定义与第 3.2 节中的定义相同。为简化计算，我们将联合策略 π 参数化为 ϕ。因此，$\pi_\phi = (\pi_{\phi^1}^1,\ \cdots\ \pi_{\phi^N}^N)$，其中 $\phi = (\phi^1,\ \cdots\ \phi^N) \in \mathbb{R}^d,\ d = \sum_{i=1}^N d_i$。在 N 人马尔可夫博弈中，我们记

$$\begin{aligned} \nabla_\phi\log\pi_\phi(a_t|s_t) &= (\nabla_{\phi^1}\log\pi_{\phi^1}^1(a_t|s_t), \cdots, \nabla_{\phi^N}\log\pi_{\phi^N}^N(a_t|s_t)) \\ \nabla_{\phi\phi^i}^2\log\pi_\phi(a_t|s_t) &= (\nabla_{\phi^1\phi^i}^2\log\pi_{\phi^1}^1(a_t|s_t), \cdots, \nabla_{\phi^N\phi^i}^2\log\pi_{\phi^N}^N(a_t|s_t)). \end{aligned}$$

引理 C.1 和引理 C.2 分别将引理 A.1 和引理 4.1 扩展到了 N 人一般和博弈。对于引理 4.2 的扩展，我们将式 (4.2) 和式 (4.3) 中的记号替换为

$$u_\theta(\phi;s) := (\nabla_{\phi^1}V_{\pi_\phi}^{(1)}(s;\theta), \cdots, \nabla_{\phi^N}V_{\pi_\phi}^{(N)}(s;\theta)), \quad u_\theta(\phi) := \mathbb{E}_{\nu^{(*)}}[u_\theta(\phi;s)], \tag{C.5}$$

$$\nabla_\theta u_\theta(\phi) = \begin{bmatrix}\mathbb{E}_{\nu^{(*)}}\nabla_{\theta\phi^1}^2V_{\pi_\phi}^{(1)}(s;\theta)\\ \cdots\\ \mathbb{E}_{\nu^{(*)}}\nabla_{\theta\phi^N}^2V_{\pi_\phi}^{(N)}(s;\theta)\end{bmatrix} \in \mathbb{R}^{d\times m}, \nabla_\phi u_\theta(\phi) = \begin{bmatrix}\mathbb{E}_{\nu^{(*)}}\nabla_{\phi\phi^1}^2V_{\pi_\phi}^{(1)}(s;\theta)\\ \cdots\\ \mathbb{E}_{\nu^{(*)}}\nabla_{\phi\phi^N}^2V_{\pi_\phi}^{(N)}(s;\theta)\end{bmatrix} \in \mathbb{R}^{d\times d}. \tag{C.6}$$

引理 4.2 中的式 (4.5) 以及式 (4.6) 中上层损失 $\nabla f_*(\theta)$ 的梯度保持不变。

引理 C.1。对于激励马尔可夫博弈 $\mathcal{G}_\theta = (\mathcal{S}, \{\mathcal{A}^i, \Omega^i, O^i, r^i(\cdot;\ \ \theta)\}_{i\in\{1,\cdots\ N\}}, \mathcal{P}\ \gamma)$，令 $R^i(\tau;\ \ \theta)$ 表示从初始状态 $s_0 = s$ 出发、遵循 π_ϕ 经历 T 步得到的样本轨迹 τ 的总奖励。我们有

$$\begin{aligned} \nabla_\theta V_{\pi_\phi}^i(s;\theta) &= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\nabla_\theta R^i(\tau;\theta)|s_0 = s\right], \\ \nabla_{\phi^i}V_{\pi_{\phi^i}}^i(s;\theta) &= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[R^i(\tau;\theta)\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi_{\phi^i}^i(a_t|s_t)\middle|s_0 = s\right], \\ \nabla_{\theta\phi^i}^2V_{\pi_\phi}^i(s;\theta) &= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi_{\phi^i}^i(a_t|s_t)\right)\nabla_\theta R^i(\tau;\theta)^\intercal\middle|s_0 = s\right], \\ \nabla_{\phi\phi^i}^2V_{\pi_\phi}^i(s;\theta) &= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[R^i(\tau;\theta)\left((\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi_{\phi^i}^i(a_t|s_t))(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|s_t))^\intercal\right.\right. \\ &\qquad \left.\left.+\sum_{t=0}^{T-1}\nabla_{\phi\phi^i}^2\log\pi_\phi(a_t|s_t)\right)\middle|s_0 = s\right], \end{aligned}$$

其中

$$\begin{aligned} \nabla_\phi\log\pi_\phi(a_t|s_t) &= (\nabla_{\phi^1}\log\pi_{\phi^1}^1(a_t|o_t^1), \cdots, \nabla_{\phi^N}\log\pi_{\phi^N}^N(a_t|o_t^N)), \\ \nabla_{\phi\phi^i}^2\log\pi_\phi(a_t|s_t) &= (\nabla_{\phi^1\phi^i}^2\log\pi_{\phi^1}^1(a_t|o_t^1), \cdots, \nabla_{\phi^N\phi^i}^2\log\pi_{\phi^N}^N(a_t|o_t^N)). \end{aligned}$$

引理 C.1 的证明。根据引理 A.1 的证明，$\nabla_{\phi^i} V^i_{\pi_\phi}(s;\,\theta)$ 和 $\nabla^2_{\phi\phi^i} V^i_{\pi_\phi}(s;\,\theta)$ 的推导基于 $\nabla_{\phi^i}\log D^{\pi_\phi}(\tau|s_0=s)$。回顾 $D^{\pi_\phi}(\tau) = \Pi_{t=0}^{T-1}\rho(s_0)P(s_{t+1}|a_t,\ s_t)\left[\Pi_{i=1}^{N}\pi^i_{\phi^i}(a^i_t|o^i)\right]$，我们有

$$\begin{aligned}\nabla_{\phi^i}\log D^{\pi_\phi}(\tau|s_0=s) &= \nabla_{\phi^i}\left[\sum_{t=0}^{T-1}\log P(s_{t+1}|s_t,a_t)+\sum_{t=0}^{T-1}\log\pi^i_{\phi^i}(a_t|s_t)+\sum_{t=0}^{T-1}\log\pi^{-i}_{\phi^{-i}}(a_t|o_t^{-i})\right]\\ &= \sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a_t|s_t).\end{aligned} \tag{C.7}$$

类似地，我们有 $\nabla_\phi\log D^{\pi_\phi}(\tau|s_0=s) = \sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|s_t)$。基于 (C.7) 的其余证明与引理 A.1 的证明完全相同。□

鉴于引理 C.1，对于基于激励奖励的正则化状态价值函数的梯度，我们用以下引理 C.2 替代引理 4.1。

引理 C.2。在激励正则化马尔可夫博弈 $\mathcal{G}^{'}_{\theta}$ 中，我们有

$$\begin{aligned}&\nabla_\theta V^{(i)}_{\pi_\phi}(s;\theta) = \nabla_\theta V^i_{\pi_\phi}(s;\theta);\\ &\nabla_{\phi^i} V^{(i)}_{\pi_\phi}(s;\theta) = \nabla_{\phi^i} V^i_{\pi_\phi}(s;\theta) - \lambda\mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\right.\\ &\qquad\qquad\left. + \sum_{t=0}^{T-1}\gamma^t\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right];\\ &\nabla^2_{\theta\phi^i} V^{(i)}_{\pi_\phi}(s;\theta) = \nabla^2_{\theta\phi^i} V^i_{\pi_\phi}(s;\theta);\end{aligned}$$

$$\begin{aligned}\nabla^2_{\phi\phi^i} V^{(i)}_{\pi_\phi}(s;\theta) &= \nabla^2_{\phi\phi^i} V^i_{\pi_\phi}(s;\theta) - \lambda\mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\left(\sum_{t=0}^{T-1}\gamma^t\left(\nabla_\phi\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\nabla_\phi\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right)^{\mathsf{T}}\right.\\ &+\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|s_t)\right)^{\mathsf{T}}\\ &+\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla^2_{\phi\phi^i}\log\pi_\phi(a^i_t|s_t)\right)\\ &\left.+\left(\sum_{t=0}^{T-1}\gamma^t\nabla_{\phi^i}\log\pi^i_{\phi^i}(a_t|s_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|s_t)\right)^{\mathsf{T}} + \left(\sum_{t=0}^{T-1}\gamma^t\nabla^2_{\phi\phi^i}\log\pi_\phi(a_t|s_t)\right)\right]\end{aligned}$$

引理 C.2 的证明。回顾 (C.3) 中 $V(i)^{\pi_\phi}(s;\theta)$ 的定义，可得，

$$\begin{aligned}V^{(i)}_{\pi_\phi}(s;\theta) &= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\sum_{t=0}^{T-1}\gamma^t\cdot r^{(i)}_\pi(s_t,a^i_t,a^{-i}_t;\theta)|s_0=s\right]\\ &= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\sum_{t=0}^{T-1}\gamma^t\cdot r^i_\pi(s_t,a^i_t,a^{-i}_t;\theta) - \lambda\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)|s_0=s\right]\\ &= V^i_{\pi_\phi}(s;\theta) - \lambda\mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)|s_0=s\right].\end{aligned} \tag{C.8}$$

为便于记号表示，我们定义

$$U^i_{\pi_\phi}(s) = \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)|s_0 = s\right].$$

于是可得，

$$\begin{aligned}
\nabla_\theta U^i_{\pi_\phi}(s) =& 0; \qquad (C.9)\\
\nabla_{\phi^i} U^i_{\pi_\phi}(s) =& \nabla_{\phi^i}\left[\sum_\tau D^{\pi_\phi}(\tau|s_0=s)\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right]\\
=& \sum_\tau D^{\pi_\phi}(\tau|s_0=s)\left[\nabla_{\phi_i}\log D^{\pi_\phi}(\tau|s_0=s)\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right.\\
&\left.+\sum_{t=0}^{T-1}\gamma^t\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right]\\
=& \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\right.\\
&\left.+\sum_{t=0}^{T-1}\gamma^t\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right]; \qquad (C.10)\\
& \qquad (C.11)
\end{aligned}$$

$$\begin{aligned}
\nabla_{\theta\phi^i} U^i_{\pi_\phi}(s) =& 0; \qquad (C.12)\\
\nabla_{\phi\phi^i} U^i_{\pi_\phi}(s) =& \overbrace{\nabla_\phi\left[\sum_\tau D^{\pi_\phi}(\tau|s_0=s)\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\right]}^{:=(I)}\\
&\underbrace{+\nabla_\phi\left[\sum_\tau D^{\pi_\phi}(\tau|s_0=s)\sum_{t=0}^{T-1}\gamma^t\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right]}_{:=(II)}.
\end{aligned}$$

接下来，我们将计算（I）和（II）。

$$
\begin{aligned}
(I) = &\sum_{\tau} D^{\pi_\phi}(\tau|s_0=s)\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\nabla_\phi \log D^{\pi_\phi}(\tau|s_0=s)^\intercal \\
&+\sum_{\tau} D^{\pi_\phi}(\tau|s_0=s)\nabla_\phi\left[\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\right] \\
=&\sum_{\tau} D^{\pi_\phi}(\tau|s_0=s)\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|s_t)\right)^\intercal \\
&+\sum_{\tau} D^{\pi_\phi}(\tau|s_0=s)\left[\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\left(\sum_{t=0}^{T-1}\gamma^t\left(\nabla_\phi\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\nabla_\phi\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right)^\intercal\right. \\
&\qquad\left.+\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla^2_{\phi\phi^i}\log\pi_\phi(a^i_t|s_t)\right)\right] \\
=&\mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|s_t)\right)^\intercal\right. \\
&\qquad+\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\left(\sum_{t=0}^{T-1}\gamma^t\left(\nabla_\phi\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\nabla_\phi\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right)^\intercal \\
&\qquad\left.+\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla^2_{\phi\phi^i}\log\pi_\phi(a_t|s_t)\right)\right]; \\
(II) = &\sum_{\tau} D^{\pi_\phi}(\tau|s_0=s)\left[\left(\sum_{t=0}^{T-1}\gamma^t\nabla_{\phi^i}\log\pi^i_{\phi^i}(a_t|s_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|s_t)\right)^\intercal\right. \\
&\qquad\left.+\left(\sum_{t=0}^{T-1}\gamma^t\nabla^2_{\phi\phi^i}\log\pi_\phi(a_t|s_t)\right)\right] \\
=&\mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\left(\sum_{t=0}^{T-1}\gamma^t\nabla_{\phi^i}\log\pi^i_{\phi^i}(a_t|s_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|s_t)\right)^\intercal+\left(\sum_{t=0}^{T-1}\gamma^t\nabla^2_{\phi\phi^i}\log\pi_\phi(a_t|s_t)\right)\right].
\end{aligned}
$$

因此，

$$
\begin{aligned}
\nabla_{\phi\phi^i}U^i_{\pi_\phi}(s) = &\mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|s_t)\right)^\intercal\right. \\
&+\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|s_t)\right)\left(\sum_{t=0}^{T-1}\gamma^t\left(\nabla_\phi\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\nabla_\phi\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right)^\intercal \qquad \text{(C.13)} \\
&+\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla^2_{\phi\phi^i}\log\pi_\phi(a^i_t|s_t)\right) \\
&\left.+\left(\sum_{t=0}^{T-1}\gamma^t\nabla_{\phi^i}\log\pi^i_{\phi^i}(a_t|s_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|s_t)\right)^\intercal+\left(\sum_{t=0}^{T-1}\gamma^t\nabla^2_{\phi\phi^i}\log\pi_\phi(a_t|s_t)\right)\right] \qquad \text{(C.14)}
\end{aligned}
$$

由于 (C.8) 可以写成

$$V_{\pi_\phi}^{(i)}(s;\theta) = V_{\pi_\phi}^{i}(s;\theta) - \lambda U_{\pi_\phi}^{i}(s),$$

结合（A.2-A.5），可得

$$\nabla_\theta V_{\pi_\phi}^{(i)}(s;\theta) = \nabla_\theta V_{\pi_\phi}^{i}(s;\theta);$$

$$\nabla_{\phi^i} V_{\pi_\phi}^{(i)}(s;\theta) = \nabla_{\phi^i} V_{\pi_\phi}^{i}(s;\theta) - \lambda \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\left(\sum_{t=0}^{T-1}\gamma^t\left(\log \pi_{\phi^i}^{i}(a_t^i|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log \pi_{\phi^j}^{j}(a_t^j|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log \pi_{\phi^i}^{i}(a_t^i|s_t)\right) + \sum_{t=0}^{T-1}\gamma^t \nabla_{\phi^i}\log \pi_{\phi^i}^{i}(a_t^i|s_t)\right];$$

$$\nabla^2_{\phi^i} V_{\pi_\phi}^{(i)}(s;\theta) = \nabla^2_{\phi^i} V_{\pi_\phi}^{i}(s;\theta);$$

$$\begin{aligned}\nabla^2_{\phi\phi^i} V_{\pi_\phi}^{(i)}(s;\theta) = \nabla^2_{\phi\phi^i} V_{\pi_\phi}^{i}(s;\theta) - \lambda \mathbb{E}_{\tau\sim D^{\pi_\phi}}\Bigg[&\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log \pi_{\phi^i}^{i}(a_t^i|s_t)\right)\left(\sum_{t=0}^{T-1}\gamma^t\left(\nabla_\phi \log \pi_{\phi^i}^{i}(a_t^i|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\nabla_\phi \log \pi_{\phi^j}^{j}(a_t^j|s_t)\right)\right)^\top\\ &+\left(\sum_{t=0}^{T-1}\gamma^t\left(\log \pi_{\phi^i}^{i}(a_t^i|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\log \pi_{\phi^j}^{j}(a_t^j|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log \pi_{\phi^i}^{i}(a_t^i|s_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi \log \pi_\phi(a_t|s_t)\right)^\top\\ &+\left(\sum_{t=0}^{T-1}\gamma^t\left(\log \pi_{\phi^i}^{i}(a_t^i|s_t) - \frac{1}{N-1}\sum_{\substack{j=1,\dots,N\\ j\neq i}}\nabla_\phi \log \pi_{\phi^j}^{j}(a_t^j|s_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla^2_{\phi\phi^i}\log \pi_\phi(a_t^i|s_t)\right)\\ &+\left(\sum_{t=0}^{T-1}\gamma^t\nabla_{\phi^i}\log \pi_{\phi^i}^{i}(a_t|s_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi \log \pi_\phi(a_t|s_t)\right)^\top + \left(\sum_{t=0}^{T-1}\gamma^t\nabla^2_{\phi\phi^i}\log \pi_\phi(a_t|s_t)\right)\Bigg]\end{aligned}$$

□

扩展到部分观测。我们首先构建带有激励奖励的多智能体部分可观测马尔可夫决策过程。略微滥用一下记号，我们记为 $G_\theta = (S, \{A^i, \Omega^i, O^i, r^i(\cdot;\theta)\}_{i\in\{1,\cdots N\}}, P, \gamma)$，其中 Ω^i是智能体 i的观测集，$O^i: S \to \Omega^i$是智能体 i的观测函数，并将状态 s下智能体 i的观测记为 $o^i = O^i(s)$。除将零和约束施加于 $r^i(\cdot;\theta)$ 外，元组中的其他元素与第 3.2 节相同。我们将玩家 i的基于观测的策略定义为 $\pi^i: \Omega^i \times A^i \to [0,1]$，并用 ϕ^i对玩家 i的策略进行参数化。然后，基于 G_θ，我们将带有激励奖励的正则化多智能体部分可观测马尔可夫决策过程构建为 $G'_\theta = (S, \{A^i, \Omega^i, O^i, r_{(i)}(\cdot;\theta)\}_{i\in\{1,\cdots N\}}, P, \gamma)$，其中 $r_{(i)}(s, a^i, a^{-i};\theta)$ 表示 $r_{(i)}^{\pi_\phi}(s, a^i, a^{-i};\theta) = r^i(s, a^i, a^{-i};\theta) - \lambda \log(\pi_{\phi^i}^i(a^i|o^i))$，且 $\lambda \geq 0$。对于轨迹 $\tau = (s_0, a_0, s_1, \cdots a_{T-1}, s_T)$，其中 $a_t = (a_t^1, \cdots a_t^N)$，$t = 0, \cdots T$，该轨迹的概率分布为：

$$D^{\pi_\phi}(\tau) := \prod_{t=0}^{T-1}\rho(s_0)P(s_{t+1}|a_t, s_t)\left[\prod_{i=1}^{N}\pi_{\phi^i}^{i}(a_t^i|o_t^i)\right],$$

其中 $o_t^i = O^i(s_t)$。在此设置下，对于状态价值函数的梯度，我们用以下引理 C.3 替代引理 C.1。

引理 C.3。 对于一个激励马尔可夫博弈 $\mathcal{G}_\theta = (\mathcal{S}, \{\mathcal{A}^i, \Omega^i, O^i, r^i(\cdot;\ \theta)\}_{i\in\{1,\cdots N\}}, \mathcal{P}\ \gamma)$，令 $R^i(\tau;\ \theta)$ 表示从初始状态 $s_0 = s$ 出发、按照 π_ϕ 经过 T 步的样本轨迹 τ 的总奖励。我们有

$$\begin{aligned}
\nabla_\theta V^i_{\pi_\phi}(s;\theta) &= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\nabla_\theta R^i(\tau;\theta)|s_0 = s\right],\\
\nabla_{\phi^i} V^i_{\pi_{\phi^i}}(s;\theta) &= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[R^i(\tau;\theta)\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a_t|o^i_t)\Big|s_0 = s\right],\\
\nabla^2_{\theta\phi^i} V^i_{\pi_\phi}(s;\theta) &= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a_t|o^i_t)\right)\nabla_\theta R^i(\tau;\theta)^\intercal\Big|s_0 = s\right],\\
\nabla^2_{\phi\phi^i} V^i_{\pi_\phi}(s;\theta) &= \mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[R^i(\tau;\theta)\left((\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a_t|o^i_t))(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|o_t))^\intercal\right.\right.\\
&\qquad\left.\left.+\sum_{t=0}^{T-1}\nabla^2_{\phi\phi^i}\log\pi_\phi(a_t|o_t)\right)\Big|s_0 = s\right],
\end{aligned}$$

其中

$$\begin{aligned}
\nabla_\phi\log\pi_\phi(a_t|o_t) &= (\nabla_{\phi^1}\log\pi^1_{\phi^1}(a_t|o^1_t),\cdots,\nabla_{\phi^N}\log\pi^N_{\phi^N}(a_t|o^N_t)),\\
\nabla^2_{\phi\phi^i}\log\pi_\phi(a_t|o_t) &= (\nabla^2_{\phi^1\phi^i}\log\pi^1_{\phi^1}(a_t|o^1_t),\cdots,\nabla^2_{\phi^N\phi^i}\log\pi^N_{\phi^N}(a_t|o^N_t)).
\end{aligned}$$

根据引理 C.3，对于基于部分可观测状态的正则化状态价值函数的梯度，我们用以下引理 C.4 替代引理 C.2。

引理 C.4。 对于激励奖励为 $\mathcal{G}_\theta = (\mathcal{S}, \{\mathcal{A}^i, \Omega^i, O^i, r_{(i)}(\cdot;\theta)\}_{i\in\{1,\cdots N\}}\lambda, \mathcal{P}\ \gamma)$的正则化马尔可夫博弈，我们有

$$\nabla_\theta V^{(i)}_{\pi_\phi}(s;\theta) = \nabla_\theta V^i_{\pi_\phi}(s;\theta);$$

$$\begin{aligned}
\nabla_{\phi^i} V^{(i)}_{\pi_\phi}(s;\theta) = \nabla_{\phi^i} V^i_{\pi_\phi}(s;\theta) - \lambda\mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|o^i_t) - \frac{1}{N-1}\sum_{\substack{j=1,\ldots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|o^j_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|o^i_t)\right)\right.\\
\left.+\sum_{t=0}^{T-1}\gamma^t\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|o^i_t)\right];
\end{aligned}$$

$$\nabla^2_{\theta\phi^i} V^{(i)}_{\pi_\phi}(s;\theta) = \nabla^2_{\theta\phi^i} V^i_{\pi_\phi}(s;\theta);$$

$$\begin{aligned}
\nabla^2_{\phi\phi^i} V^{(i)}_{\pi_\phi}(s;\theta) = \nabla^2_{\phi\phi^i} V^i_{\pi_\phi}(s;\theta) - \lambda\mathbb{E}_{\tau\sim D^{\pi_\phi}}\left[\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|o^i_t)\right)\left(\sum_{t=0}^{T-1}\gamma^t\left(\nabla_\phi\log\pi^i_{\phi^i}(a^i_t|o^i_t) - \frac{1}{N-1}\sum_{\substack{j=1,\ldots,N\\ j\neq i}}\nabla_\phi\log\pi^j_{\phi^j}(a^j_t|o^j_t)\right)\right)^\intercal\right.\\
+\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|o^i_t) - \frac{1}{N-1}\sum_{\substack{j=1,\ldots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|o^j_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla_{\phi^i}\log\pi^i_{\phi^i}(a^i_t|a^i_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|o_t)\right)^\intercal\\
+\left(\sum_{t=0}^{T-1}\gamma^t\left(\log\pi^i_{\phi^i}(a^i_t|o^i_t) - \frac{1}{N-1}\sum_{\substack{j=1,\ldots,N\\ j\neq i}}\log\pi^j_{\phi^j}(a^j_t|o^j_t)\right)\right)\left(\sum_{t=0}^{T-1}\nabla^2_{\phi\phi^i}\log\pi_\phi(a^i_t|o^i_t)\right)\\
\left.+\left(\sum_{t=0}^{T-1}\gamma^t\nabla_{\phi^i}\log\pi^i_{\phi^i}(a_t|o_t)\right)\left(\sum_{t=0}^{T-1}\nabla_\phi\log\pi_\phi(a_t|o_t)\right)^\intercal+\left(\sum_{t=0}^{T-1}\gamma^t\nabla^2_{\phi\phi^i}\log\pi_\phi(a_t|o_t)\right)\right]
\end{aligned}$$

引理 C.3 和 C.4 的证明与 C.1 和 C.2 的证明非常相似，只是符号略有变化。因此，本节不再讨论这些证明。

D PEM与熵正则化OMWU

本节将简要介绍纳什均衡求解器策略外梯度法（PEM） [14] 和熵正则化 OMWU [13] 的伪代码。

D.1 策略外梯度法（PEM）

首先，我们介绍熵正则化零和双人矩阵博弈。

$$\max_{\pi_1 \in \Delta(\mathcal{A}_1)} \min_{\pi_2 \in \Delta(\mathcal{A}_2)} f_\lambda(A; \pi_1, \pi_2) := \pi_1^\top A \pi_2 + \lambda \mathcal{H}(\pi_1) - \lambda \mathcal{H}(\pi_2), \tag{D.1}$$

其中， $A \in \mathbb{R}^{m \times n}$ 表示收益矩阵， $\pi_1 \in \Delta(A_1)$和 $\pi_2 \in \Delta(A_2)$分别表示每位玩家的混合/随机策略，它们被定义为概率单纯形 $\Delta(A_1)$和 $\Delta(A_2)$上的分布； $H(\pi) = -\sum_i \pi_i \log(\pi_i)$， λ为正则化参数。

Cen [14] 提出了两种外梯度法：PU和OMWU，它们具有线性收敛保证，可用于求解熵正则化零和双人矩阵博弈（D.1）。

算法2 PU方法

初始化： $\pi^{(0)}_1, \pi^{(0)}_2$。

参数： 学习率 η_t

对于t=0,1,2,... **执行**

更新 π_1和 π_2，针对每个动作 $a = (a_1, a_2) \in A = A_1 \times A_2$ 根据

$$\pi^{(t+1)}_1(a_1) \propto \pi^{(t)}_1(a_1)^{1-\eta_t\lambda} \exp(\eta_t [A\pi^{(t)}_2]_{a_1})$$

$$\pi^{(t+1)}_2(a_2) \propto \pi^{(t)}_2(a_2)^{1-\eta_t\lambda} \exp(-\eta_t [A^\top \pi^{(t)}_1]_{a_2})$$

更新每个动作 $a = (a_1, a_2)$ 对应的 π_1 和 $\pi_2 \in A = A_1 \times A_2$ 根据

$$\pi^{(t+1)}_1(a_1) \propto \pi^{(t)}_1(a_1)^{1-\eta_t\lambda} \exp(\eta_t [A\pi^{(t)}_2]_{a_1})$$

$$\pi^{(t+1)}_2(a_2) \propto \pi^{(t)}_2(a_2)^{1-\eta_t\lambda} \exp(-\eta_t [A^\top \pi^{(t)}_1]_{a_2})$$

结束循环

算法3OMWU 方法

初始化： $\pi^{(0)}_1 = \pi^{(0)}_1$ $\pi^{(0)}_2 = \pi^{(0)}_2$。

参数： 学习率 η_t

对于t=0,1,2,... **执行**

对每个动作 $a = (a_1, a_2) \in A = A_1 \times$，更新 π_1 和 π_2 根据

$$\pi^{(t+1)}_1(a_1) \propto \pi^{(t)}_1(a_1)^{1-\eta_t\lambda} \exp(\eta_t [A\pi^{(t)}_2]_{a_1})$$

$$\pi^{(t+1)}_2(a_2) \propto \pi^{(t)}_2(a_2)^{1-\eta_t\lambda} \exp(-\eta_t [A^\top \pi^{(t)}_1]_{a_2})$$

为每个动作 $a = (a_1, a_2)$ 更新 π_1 和 $\pi_2 \in A = A_1 \times$ A_2 根据

$$\pi^{(t+1)}_1(a_1) \propto \pi^{(t)}_1(a_1)^{1-\eta_t\lambda} \exp(\eta_t [A\pi^{(t)}_2]_{a_1})$$

$$\pi^{(t+1)}_2(a_2) \propto \pi^{(t)}_2(a_2)^{1-\eta_t\lambda} \exp(-\eta_t [A^\top \pi^{(t)}_1]_{a_2})$$

结束循环

策略外梯度法（PEM）基于 PU 和 OMWU 方法。下面提供 PEM 的伪代码（算法 4）供您参考。

算法 4PEM：策略外梯度法

初始化： $Q^{(0)} = 0$。

对于t=0,1,2,...,T主 **执行**

令 $Q^{(t)}$ 表示

$$Q^{(t)}(s, a_1, a_2) = r^1(s, a_1, a_2) + \lambda \mathbb{E}_{s' \sim P(\cdot|s,a_1,a_2)} V^{(t)}(s')$$

对每个状态 s，调用 PU（算法2）或 OMWU（算法3），迭代 Tsub 次，以求解以下熵正则化矩阵博弈，其中初始状态设为均匀分布

$$\max_{\pi_1(s) \in \Delta(\mathcal{A}_1)} \min_{\pi_2(s) \in \Delta(\mathcal{A}_2)} f_\lambda(Q^{(t)}(s); \pi_1(s), \pi_2(s)).$$

返回最后一次迭代结果 $\pi^{(t,T_{sub})}_1(s)$， $\pi^{(t,T_{sub})}_2(s)$。

设置 $V^{(t+1)}(s) = f_\lambda(Q^{(t)}(s); \pi^{(t,\ T_{sub})}_1(s)\ \pi^{(t,\ T_{sub})}_2(s))$ 。

结束循环

D.2 熵正则化 OMWU

本节给出熵正则化 OMWU 方法的伪代码 [13]。

算法 5熵正则化 OMWU

输入：正则化参数 $\lambda > 0$，策略更新的学习率 $\eta > 0$，价值更新的学习率 $\{\alpha_t\}_{t=0}^{\infty}$.

初始化：将 π_1^0、 π_1^0、$\overline{\pi}_2^0$ 和 π_2^0 设为均匀策略；并设置

$$Q^{(0)} = 0, \quad V^{(0)} = \lambda(\log|\mathcal{A}_1| + \log|\mathcal{A}_2|).$$

对于 t=0,1,2,...，对所有 $s \in \mathbf{S}$，执行：当 $t \geq 1$ 时，将策略对 $\pi = (\pi_1, \pi_2)$ 更新为：将策略对 $\pi = (\pi_1, \pi_2)$ 更新为：将 $Q(t+1)(s)$ 和 $V(t+1)(s)$ 更新为 $Q(t+1)(s, a_1, a_2) = r^1(s, a_1, a_2) + \lambda \mathbb{E}_{s' \sim P(\cdot|s,a_1,a_2)}[V(t)(s')]V(t+1)(s) = (1-\alpha_{t+1})V(t)(s) + \alpha_{t+1}[\pi(t+1)_1 \quad ^{\intercal} Q(t+1)(s)\pi(t+1)_2 + \lambda \mathbf{H}(\pi(t+1)\,1) - \lambda \mathbf{H}(\pi(t+1)\,2)$ **结束循环，结束循环**

$$\pi_1^{(t+1)}(a_1|s) \propto \pi_1^{(t)}(a_1|s)^{1-\eta_t\lambda} \exp(\eta[Q^{(t)}(s)\overline{\pi}_2^{(t)}(s)]_{a_1})$$

$$\pi_2^{(t+1)}(a_2|s) \propto \pi_2^{(t)}(a_2|s)^{1-\eta_t\lambda} \exp(-\eta[Q^{(t)}(s)\overline{\pi}_1^{(t)}(s)]_{a_2})$$

$$\overline{\pi}_1^{(t+1)}(a_1|s) \propto \pi_1^{(t)}(a_1|s)^{1-\eta_t\lambda} \exp(\eta[Q^{(t)}(s)\overline{\pi}_2^{(t)}(s)]_{a_1})$$

$$\overline{\pi}_2^{(t+1)}(a_2|s) \propto \pi_2^{(t)}(a_2|s)^{1-\eta_t\lambda} \exp(-\eta[Q^{(t)}(s)\overline{\pi}_1^{(t)}(s)]_{a_2})$$

E 假设 5.3 的说明

根据定义，$f_*(\theta) = f(\theta, \phi_*(\theta))$ $f_*(\theta)$ 是由 $f(\theta, \phi) : \Theta \times \Phi \to \mathbb{R}$ 和 $\phi_*(\theta) : \Theta \to \Phi$ 组成的复合函数，其中 Θ 和 Φ 分别是激励参数 θ 和策略参数 ϕ 的可行域。因此，直接对 $f_*(\theta)$ 作出假设可能过于激进，也难以验证。本节举例说明，如果方程 (3.3) 中定义的损失函数 $f(\theta, \phi)$ 和价值函数 $V(i)^{\pi_\phi}(s; \theta)$ 满足适当条件，那么当损失函数 $f(\theta, \phi)$ 和价值函数 $V(i)^{\pi_\phi}(s; \theta)$ 满足适当条件时，复合函数 $f_*(\theta)$ 满足假设 5.3。

假设 5.3 包含两个部分。首先，我们要求复合函数 $f_*(\theta)$ 有界。如果损失 $f(\theta, \phi)$ 有界，这一假设较为温和且容易满足。其次，我们要求 f_* 是 L-光滑的（$\nabla_\theta f_*(\theta)$ 是 L-Lipschitz 连续的）。这一假设需要验证。下面的命题 E.4 给出了一个具体例子，说明 $f_*(\theta)$ 既有界又是 L-光滑的。

假设 E.1。假设损失函数 $f(\theta, \phi) : \Theta \times \Phi \to \mathbb{R}$ 满足以下条件：

(1) **(有界)** $\exists M_f > 0$，使得 $|f(\theta, \phi)| \leq M_f, \forall(\theta, \phi) \in \Theta \times \Phi \subset \mathbb{R}^3$. (2)

(Lipschitz连续) $\exists G_f > 0$，使得 $\|\nabla f(\theta, \phi)\|_2 \leq G_f$. $\forall(\theta, \phi) \in \Theta \times \Phi \subset \mathbb{R}^3$.

(3) **(光滑)** $\exists L_f > 0$，使得对于 $\forall(\theta_1, \phi_1), (\theta_2, \phi_2) \in \Theta \times \Phi \subset \mathbb{R}^3$

$$\|\nabla f(\theta_1, \phi_1) - \nabla f(\theta_2, \phi_2)\|_2 \leq L_f \|(\theta_1, \phi_1) - (\theta_2, \phi_2)\|_2 .$$

假设 E.2。假设 θ、 ϕ_1 和 ϕ_2 的可行域均为一维，这意味着 $\theta \in \Theta \subset \mathbb{R}$、 $\phi_1 \in \Phi_1 \subset \mathbb{R}$、$\phi_2 \in \Phi_2 \subset \mathbb{R}$ 和 $\phi = (\phi_1, \phi_2) \in \Phi = \Phi_1 \times \Phi_1 \subset \mathbb{R}^2$。假设对于任意固定状态 $s \in \mathrm{S}$，价值函数 $V(i)^{\pi_\phi}(s; \theta) : \Theta \to \mathbb{R}$, $(i = 1, 2)$ 满足以下条件：

(4) **(二阶梯度有界)** $\exists G_v > 0$，使得对于 $\forall(\theta, \phi) \in \Theta \times \Phi \subset \mathbb{R}^3$ 以及 $i = 1, 2$,

$$\left|\frac{\partial^2 V_{\pi_\phi}^{(i)}(s;\theta)}{\partial\theta\partial\phi_i}\right| \leq G_v, \quad \left|\frac{\partial^2 V_{\pi_\phi}^{(i)}(s;\theta)}{\partial\phi_i^2}\right| \leq G_v \quad \text{and} \quad \left|\frac{\partial^2 V_{\pi_\phi}^{(i)}(s;\theta)}{\partial\phi_i\partial\phi_{-i}}\right| \leq G_v.$$

(5) **(三阶梯度有界)** $\exists L_v > 0$，使得对于 $\forall(\theta, \phi) \in \Theta \times \Phi \subset \mathbb{R}^3$ 和 $i = 1, 2$,

$$\left|\frac{\partial^3 V_{\pi_\phi}^{(i)}(s;\theta)}{\partial\theta^2\partial\phi_i}\right| \leq L_v, \quad \left|\frac{\partial^3 V_{\pi_\phi}^{(i)}(s;\theta)}{\partial\theta\partial\phi_i^2}\right| \leq L_v \quad \text{and} \quad \left|\frac{\partial^3 V_{\pi_\phi}^{(i)}(s;\theta)}{\partial\theta\partial\phi_i\partial\phi_{-i}}\right| \leq L_v.$$

(6) **(正定矩阵)** $\exists \rho_v > 0$，使得对于 $\forall(\theta,\phi) \in \Theta \times \Phi \subset \mathbb{R}^3$,

$$\begin{bmatrix} \frac{\partial^2 V_{\pi_\phi}^{(1)}(s;\theta)}{\partial \phi_1^2} & \frac{\partial^2 V_{\pi_\phi}^{(1)}(s;\theta)}{\partial \phi_1 \partial \phi_2} \\ \frac{\partial^2 V_{\pi_\phi}^{(2)}(s;\theta)}{\partial \phi_2 \phi_1} & \frac{\partial^2 V_{\pi_\phi}^{(2)}(s;\theta)}{\partial \phi_2^2} \end{bmatrix} \succ \rho_v I,$$

其中 I 是单位矩阵。

引理 E.3. 假设假设 *E.2* 成立。$\phi_*(\theta)$： $\Theta \to \Phi$ 是第 *3.4* 节中定义的映射，它将激励参数 *theta* θ 映射到正则化马尔可夫博弈$\mathcal{G}_r\theta$ 的纳什策略对应的策略参数。则有

(i) 对于 $\forall\theta \in \Theta$，有 $\|\nabla_\theta \phi_*(\theta)\|_F \le \frac{2G_v}{\rho_v}$

(ii) 对于 $\forall\theta_1, \theta_2 \in \Theta$，有 $\|\nabla_\theta \phi_*(\theta_1) - \nabla_\theta \phi_*(\theta_2)\|_F \le \frac{2L_v}{\rho_v}\left(1 + 4G_v + \frac{16G_v^2}{\rho_v}\right)\|\theta_1 - \theta_2\|_2$,

其中 $\|\cdot\|_F$ 是矩阵的 *Frobenius* 范数。

证明。见附录 E.1。 □

命题 E.4。 如果假设 *E.1* 和 *E.2* 成立，则可以得出，在优化问题 *(3.8)* 中定义的 $f_*(\theta)$ 满足以下条件：

(i) （有界） $|f_*(\theta)| \le M_f$ 对于 $\forall\theta \in \Theta$。

(ii) （光滑） $f_*(\theta)$ 是 *L*-光滑的（$\nabla_\theta f_*(\theta)$ 是 *L-Lipschitz* 连续的），其中

$$L = L_f\left(1 + \frac{2G_v}{\rho_v}\right)\left(1 + \frac{4G_v}{\rho_v}\right) + \frac{2L_v G_f}{\rho_v}\left(1 + 4G_v + \frac{16G_v^2}{\rho_v}\right).$$

证明。首先，根据条件 (1)，$|f_*(\theta)| \le M_f$很容易由 $\phi^*(\theta) \in \Phi$推出。下面的证明将重点说明 $f_*(\theta)$ 是 *L*-光滑的。

对于 $\forall\theta_1, \theta_2 \in \Theta$,

$$\begin{aligned}
\|\nabla_\theta f_*(\theta_1) - \nabla_\theta f_*(\theta_2)\|_2 &= \left\|\nabla_\theta f(\theta_1, \phi_*(\theta_1)) + \nabla_\theta \phi_*(\theta_1)^\intercal \nabla_\phi f(\theta_1, \phi_*(\theta_1)) - \nabla_\theta f(\theta_2, \phi_*(\theta_2)) - \nabla_\theta \phi_*(\theta_2)^\intercal \nabla_\phi f(\theta_2, \phi_*(\theta_2))\right\|_2 \\
&\le \|\nabla_\theta f(\theta_1, \phi_*(\theta_1)) - \nabla_\theta f(\theta_2, \phi_*(\theta_2))\|_2 + \left\|\nabla_\theta \phi_*(\theta_1)^\intercal \nabla_\phi f(\theta_1, \phi_*(\theta_1)) - \nabla_\theta \phi_*(\theta_2)^\intercal \nabla_\phi f(\theta_2, \phi_*(\theta_2))\right\|_2 \\
&\le \|\nabla_\theta f(\theta_1, \phi_*(\theta_1)) - \nabla_\theta f(\theta_2, \phi_*(\theta_2))\|_2 + \|\nabla_\theta \phi_*(\theta_1)\|_F \left\|\nabla_\phi f(\theta_1, \phi_*(\theta_1)) - \nabla_\phi f(\theta_3, \phi_*(\theta_2))\right\|_2 \\
&\quad + \left\|\nabla_\phi f(\theta_2, \phi_*(\theta_2))\right\|_2 \|\nabla_\theta \phi_*(\theta_1) - \nabla_\theta \phi_*(\theta_2)\|_F
\end{aligned} \tag{E.1}$$

我们将逐一对公式 (E.1) 中的三个项进行有界估计。

i) 根据假设 E.1 中的条件 (3) 和引理 E.3 中的性质 (i)，我们有

$$\begin{aligned}
\|\nabla_\theta f(\theta_1, \phi_*(\theta_1)) - \nabla_\theta f(\theta_2, \phi_*(\theta_2))\|_2 &\le \|\nabla_\theta f(\theta_1, \phi_*(\theta_1)) - \nabla_\theta f(\theta_2, \phi_*(\theta_2))\|_2 \\
&\le L_f \|(\theta_1, \phi_*(\theta_1)) - (\theta_2, \phi_*(\theta_2))\|_2 && \text{(Condition (3) in Assumption E.1)} \\
&\le L_f(\|\theta_1 - \theta_2\|_2 + \|\phi_*(\theta_1) - \phi_*(\theta_2)\|_2) \\
&\le L_f(\|\theta_1 - \theta_2\|_2 + \max_{\theta\in\Theta} \|\nabla_\theta \phi_*(\theta)\|_F \|\theta_1 - \theta_2\|_2) && \text{(Mean value Theorem)} \\
&\le L_f\left(1 + \frac{2G_v}{\rho_v}\right)\|\theta_1 - \theta_2\|_2 && \text{(Property (i) in Lemma E.3)}
\end{aligned} \tag{E.2}$$

ii) 根据引理 E.3 中的性质 (i) 和假设 E.1 中的条件 (3)，我们有

$$\begin{aligned}
&\|\nabla_\theta \phi_*(\theta_1)\|_F \left\|\nabla_\phi f(\theta_1, \phi_*(\theta_1)) - \nabla_\phi f(\theta_3, \phi_*(\theta_2))\right\|_2 \\
\le &\frac{2G_v}{\rho_v} \left\|\nabla_\phi f(\theta_1, \phi_*(\theta_1)) - \nabla_\phi f(\theta_3, \phi_*(\theta_2))\right\|_2 && \text{（引理 E.3 中的性质 (i)）} \\
\le &\frac{2G_v L_f}{\rho_v} \|(\theta_1, \phi_*(\theta_1)) - (\theta_2, \phi_*(\theta_2))\|_2 && \text{（假设 E.1 中的条件 (3)）} \\
\le &\frac{2G_v L_f}{\rho_v}\left(1 + \frac{2G_v}{\rho_v}\right)\|\theta_1 - \theta_2\|_2 && \text{（}
\end{aligned} \tag{E.3)}$$

iii) 根据假设 E.1 中的条件 (2) 和引理 E.3 中的性质 (ii)，我们有

$$\begin{aligned}
&\left\|\nabla_\phi f(\theta_2,\phi_*(\theta_2))\right\|_2 \left\|\nabla_\theta\phi_*(\theta_1)-\nabla_\theta\phi_*(\theta_2)\right\|_F \\
\le& G_f \left\|\nabla_\theta\phi_*(\theta_1)-\nabla_\theta\phi_*(\theta_2)\right\|_F && \text{（假设 E.1 中的条件（2））} \\
\le& \frac{2L_vG_f}{\rho_v}\left(1+4G_v+\frac{16G_v^2}{\rho_v}\right)\|\theta_1-\theta_2\|_2 && \text{(Property (ii) in Lemma E.3)}
\end{aligned} \tag{E.4}$$

结合不等式（E.1-E.4），我们可以得到

$$\|\nabla_\theta f_*(\theta_1)-\nabla_\theta f_*(\theta_2)\|_2 \le L_f(1+\frac{2G_v}{\rho_v})(1+\frac{4G_v}{\rho_v})+\frac{2L_vG_f}{\rho_v}(1+4G_v+\frac{16G_v^2}{\rho_v})\|\theta_1-\theta_2\|_2 \tag{E.5}$$

我们已经证明，$f_*(\theta)$ 是 L-光滑的，且 $L = L_f(1+\frac{2G_v}{\rho_v})(1+\frac{4G_v}{\rho_v})+\frac{2L_vG_f}{\rho_v}(1+4G_v+\frac{16G_v^2}{\rho_v})$。 □

E.1 引理 E.3 的证明

引理 E.3 的证明。当对于每个状态 $s\in\mathrm{S}$，假设 E.2 中的条件（4-6）对 $V(i)\ \pi_\phi(s;\theta)$ 成立时，显然，条件（4-6）对 $\mathbb{E}_{v^{(*)}}V(i)\ \pi_\phi(s;\theta)$ 也成立。为简化符号，本节中记 $\mathbb{E}_{v^{(*)}}V(i)\ \pi_\phi(s;\theta) := V^i(\theta,\ \phi)$。接下来我们将逐一证明两个性质。

i) 首先，我们将对 $\|\nabla_\theta\phi_*(\theta)\|_F$ 进行界定。对于任意的 $\theta\in\Theta$ 和 $\phi\in\Phi$，由于 $\phi_1\in\mathbb{R}$、$\phi_2\in\mathbb{R}$ 和 $\phi=(\phi_1,\ \phi_2)\in\mathbb{R}^2$，第 4 节中的公式 (4.3-4.4) 可写为

$$\nabla_\theta u_\theta(\phi)=\begin{bmatrix}\frac{\partial^2V^1(\theta,\phi)}{\partial\theta\partial\phi_1}\\ \frac{\partial^2V^2(\theta,\phi)}{\partial\theta\partial\phi_2}\end{bmatrix}\in\mathbb{R}^{2\times1},\qquad \nabla_\phi u_\theta(\phi)=\begin{bmatrix}\frac{\partial^2V^1(\theta,\phi)}{\partial\phi_1^2} & \frac{\partial^2V^1(\theta,\phi)}{\partial\phi_1\partial\phi_2}\\ \frac{\partial^2V^2(\theta,\phi)}{\partial\phi_1\partial\phi_2} & \frac{\partial^2V^1(\theta,\phi)}{\partial\phi_2^2}\end{bmatrix}\in\mathbb{R}^{2\times2} \tag{E.6}$$

根据引理 4.2 中的公式 (4.5) 以及假设 E.2 中的条件 (4, 6)，我们有

$$\begin{aligned}
\left\|\nabla_\theta\phi^*(\theta)\right\|_F \le& \left\|\begin{bmatrix}\frac{\partial^2V^1(\theta,\phi)}{\partial\phi_1^2} & \frac{\partial^2V^1(\theta,\phi)}{\partial\phi_1\partial\phi_2}\\ \frac{\partial^2V^2(\theta,\phi)}{\partial\phi_1\partial\phi_2} & \frac{\partial^2V^1(\theta,\phi)}{\partial\phi_2^2}\end{bmatrix}^{-1}\begin{bmatrix}\frac{\partial^2V^1(\theta,\phi)}{\partial\theta\phi_1}\\ \frac{\partial^2V^2(\theta,\phi)}{\partial\theta\phi_2}\end{bmatrix}\right\|_F \\
\le& \left\|\begin{bmatrix}\frac{\partial^2V^1(\theta,\phi)}{\partial\phi_1^2} & \frac{\partial^2V^1(\theta,\phi)}{\partial\phi_1\partial\phi_2}\\ \frac{\partial^2V^2(\theta,\phi)}{\partial\phi_1\partial\phi_2} & \frac{\partial^2V^1(\theta,\phi)}{\partial\phi_2^2}\end{bmatrix}^{-1}\right\|_F\left\|\begin{bmatrix}\frac{\partial^2V^1(\theta,\phi)}{\partial\theta\phi_1}\\ \frac{\partial^2V^2(\theta,\phi)}{\partial\theta\phi_2}\end{bmatrix}\right\|_F && \text{（假设 E.2 中的条件 (6)）} \\
\le& \frac{1}{\rho_v}\left(\left|\frac{\partial^2V^1(\theta,\phi)}{\partial\theta\phi_1}\right|+\left|\frac{\partial^2V^2(\theta,\phi)}{\partial\theta\phi_2}\right|\right) \\
\le& \frac{2G_v}{\rho_v} && \text{（假设 E.2 中的条件 (4)）}
\end{aligned} \tag{E.7}$$

ii) 其次，我们将对 $\|\nabla_\theta\phi_*(\theta_1)-\nabla_\theta\phi_*(\theta_2)\|_F$ 进行界定。对于 $\forall\theta_1,\ \theta_2\in\Theta$，

$$\begin{aligned}
\|\nabla_\theta\phi_*(\theta_1)-\nabla_\theta\phi_*(\theta_2)\|_F =& \left\|\left[\nabla_\phi u_{\theta_1}(\phi_*(\theta_1))\right]^{-1}\nabla_\theta u_{\theta_1}(\phi_*(\theta_1))-\left[\nabla_\phi u_{\theta_2}(\phi_*(\theta_2))\right]^{-1}\nabla_\theta u_{\theta_2}(\phi_*(\theta_2))\right\|_F \\
\le& \left\|\left[\nabla_\phi u_{\theta_1}(\phi_*(\theta_1))\right]^{-1}\right\|_F\left\|\nabla_\theta u_{\theta_1}(\phi_*(\theta_1))-\nabla_\theta u_{\theta_2}(\phi_*(\theta_2))\right\|_F \\
&+\left\|\nabla_\theta u_{\theta_2}(\phi_*(\theta_2))\right\|_F\left\|\left[\nabla_\phi u_{\theta_1}(\phi_*(\theta_1))\right]^{-1}-\left[\nabla_\phi u_{\theta_2}(\phi_*(\theta_2))\right]^{-1}\right\|_F
\end{aligned} \tag{E.8}$$

根据条件 (6)，我们有

$$\left\|\left[\nabla_\phi u_{\theta_1}(\phi_*(\theta_1))\right]^{-1}\right\|_F\le\frac{1}{\rho}. \tag{E.9}$$

根据条件 (4)，我们有

$$\left\|\nabla_\theta u_{\theta_2}(\phi_*(\theta_2))\right\|_F\le\left|\frac{\partial^2V^1(\theta_2,\phi_*(\theta_2))}{\partial\theta\phi_1}\right|+\left|\frac{\partial^2V^2(\theta_2,\phi_*(\theta_2))}{\partial\theta\phi_2}\right|\le2G_v. \tag{E.10}$$

根据条件 (5)，我们有

$$\begin{aligned}\left\|\nabla_\theta u_{\theta_1}(\phi_*(\theta_1)) - \nabla_\theta u_{\theta_2}(\phi_*(\theta_2))\right\|_F \leq& \left|\frac{\partial^2 V^1(\theta_1,\phi_*(\theta_1))}{\partial\theta\phi_1} - \frac{\partial^2 V^1(\theta_2,\phi_*(\theta_2))}{\partial\theta\phi_1}\right| + \left|\frac{\partial^2 V^2(\theta_1,\phi_*(\theta_1))}{\partial\theta\phi_2} - \frac{\partial^2 V^2(\theta_2,\phi_*(\theta_2))}{\partial\theta\phi_2}\right| \\ \leq& \max_{\theta,\phi}\left|\frac{\partial^3 V^1(\theta_1,\phi_*(\theta_1))}{\partial\theta^2\phi_1}\right| \|\theta_1-\theta_2\|_2 + \max_{\theta,\phi}\left|\frac{\partial^3 V^2(\theta_1,\phi_*(\theta_1))}{\partial\theta^2\phi_2}\right| \|\theta_1-\theta_2\|_2 \\ \leq& 2L_v \|\theta_1-\theta_2\|. \end{aligned} \tag{E.11}$$

最后但同样重要的是，

$$\left\|\left[\nabla_\phi u_{\theta_1}(\phi_*(\theta_1))\right]^{-1} - \left[\nabla_\phi u_{\theta_2}(\phi_*(\theta_2))\right]^{-1}\right\|_F \tag{E.12}$$

$$= \left\|\begin{bmatrix} \frac{\partial^2 V^1(\theta_1,\phi_*(\theta_1))}{\partial\phi_1^2} & \frac{\partial^2 V^1(\theta_1,\phi_*(\theta_1))}{\partial\phi_1\partial\phi_2} \\ \frac{\partial^2 V^2(\theta_1,\phi_*(\theta_1))}{\partial\phi_1\partial\phi_2} & \frac{\partial^2 V^2(\theta_1,\phi_*(\theta_1))}{\partial\partial\phi_2^2} \end{bmatrix}^{-1} - \begin{bmatrix} \frac{\partial^2 V^1(\theta_2,\phi_*(\theta_2))}{\partial\phi_1^2} & \frac{\partial^2 V^1(\theta_2,\phi_*(\theta_2))}{\partial\phi_1\partial\phi_2} \\ \frac{\partial^2 V^2(\theta_2,\phi_*(\theta_2))}{\partial\phi_1\partial\phi_2} & \frac{\partial^2 V^2(\theta_2,\phi_*(\theta_2))}{\partial\partial\phi_2^2} \end{bmatrix}^{-1}\right\|_F \tag{E.13}$$

为简化记号，我们记

$$\begin{aligned} a_1 &= \frac{\partial^2 V^1(\theta_1,\phi_*(\theta_1))}{\partial\phi_1^2}, b_1 = \frac{\partial^2 V^1(\theta_1,\phi_*(\theta_1))}{\partial\phi_1\partial\phi_2}, c_1 = \frac{\partial^2 V^2(\theta_1,\phi_*(\theta_1))}{\partial\phi_1\partial\phi_2}, d_1 = \frac{\partial^2 V^2(\theta_1,\phi_*(\theta_1))}{\partial\partial\phi_2^2} \\ a_2 &= \frac{\partial^2 V^1(\theta_2,\phi_*(\theta_2))}{\partial\phi_1^2}, b_2 = \frac{\partial^2 V^1(\theta_2,\phi_*(\theta_2))}{\partial\phi_1\partial\phi_2}, c_2 = \frac{\partial^2 V^2(\theta_2,\phi_*(\theta_2))}{\partial\phi_1\partial\phi_2}, d_2 = \frac{\partial^2 V^2(\theta_2,\phi_*(\theta_2))}{\partial\partial\phi_2^2} \\ M_1 &= det\left(\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix}\right), M_2 = det\left(\begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix}\right) \end{aligned}$$

因此，我们有

$$\begin{aligned} &\left\|\left[\nabla_\phi u_{\theta_1}(\phi_*(\theta_1))\right]^{-1} - \left[\nabla_\phi u_{\theta_2}(\phi_*(\theta_2))\right]^{-1}\right\|_F \\ =& \left\|\frac{1}{M_1}\begin{bmatrix} d_1 & -b_1 \\ -c_1 & a_1 \end{bmatrix} - \frac{1}{M_2}\begin{bmatrix} d_2 & -b_2 \\ -c_2 & a_2 \end{bmatrix}\right\|_F \\ \leq& \left|\frac{a_1}{M_1} - \frac{a_2}{M_2}\right| + \left|\frac{b_1}{M_1} - \frac{b_2}{M_2}\right| + \left|\frac{c_1}{M_1} - \frac{c_2}{M_2}\right| + \left|\frac{d_1}{M_1} - \frac{d_2}{M_2}\right| \end{aligned} \tag{E.14}$$

我们首先将对 $\left|\frac{a_1}{M_1} - \frac{a_2}{M_2}\right|$ 进行界定。

$$\left|\frac{a_1}{M_1} - \frac{a_2}{M_2}\right| \leq \frac{1}{|M_1|}|a_1 - a_1| + \frac{|a_2|}{|M_1 M_2|}|M_1 - M_2|.$$

根据条件（6），我们有

$$|M_1| \geq \rho_v \quad \text{and} \quad |M_2| \geq \rho_v. \tag{E.15}$$

根据条件（5），我们有

$$|a_1 - a_1| \leq L_v \|\theta_1 - \theta_2\|_2. \tag{E.16}$$

根据条件 (4)，我们有

$$|a_2| \leq G_v. \tag{E.17}$$

同理，根据条件 (4,5)，我们有

$$\begin{aligned} |M_1 - M_2| =& |a_1 d_1 - b_1 c_1 - a_2 d_2 + b_2 c_2| \\ \leq& |a_1||d_1 - d_2| + |d_2||a_1 - a_2| + |b_1||c_1 - c_2| + |c_2||b_1 - b_2| \\ \leq& 4G_v L_v \|\theta_1 - \theta_2\|_2. \end{aligned} \tag{E.18}$$

结合公式 (E.15-E.18)，我们有

$$\left|\frac{a_1}{M_1} - \frac{a_2}{M_2}\right| \leq \left(\frac{L_v}{\rho_v} + \frac{G_v}{\rho_v^2} \times 4G_v L_v\right) \|\theta_1 - \theta_2\|_2 = \frac{L_v}{\rho_v}\left(1 + \frac{4G_v^2}{\rho_v}\right) \|\theta_1 - \theta_2\|_2. \tag{E.19}$$

同样，我们可以计算 $\left|\frac{b_1}{M_1} - \frac{b_2}{M_2}\right|$、$\left|\frac{c_1}{M_1} - \frac{c_2}{M_2}\right|$ 和 $\left|\frac{d_1}{M_1} - \frac{d_2}{M_2}\right|$ 的界。因此，我们有

$$\left\|\left[\nabla_\phi u_{\theta_1}(\phi_*(\theta_1))\right]^{-1} - \left[\nabla_\phi u_{\theta_2}(\phi_*(\theta_2))\right]^{-1}\right\|_F \leq \frac{4L_v}{\rho_v}\left(1 + \frac{4G_v^2}{\rho_v}\right) \|\theta_1 - \theta_2\|_2 \tag{E.20}$$

结合公式 (E.9、E.10、E.11、E.20)，我们有

$$\begin{aligned}\|\nabla_\theta\phi_*(\theta_1)-\nabla_\theta\phi_*(\theta_2)\|_F \leq & \frac{1}{\rho_v}\times 2L_v\|\theta_1-\theta_2\|_2+2G_v\times\frac{4L_v}{\rho_v}(1+\frac{4G_v^2}{\rho_v})\|\theta_1-\theta_2\|_2\\ \leq & \frac{2L_v}{\rho_v}(1+4G_v+\frac{16G_v^2}{\rho_v})\|\theta_1-\theta_2\|_2 .\end{aligned} \tag{E.21}$$

□

F 定理 5.4 的证明

定理 5.4 的证明。根据假设 5.3，梯度 $\nabla_\theta f_*(\theta)$ 是 L-Lipschitz 连续的，这意味着对于 $\forall k$

$$f_*(\theta_{k+1})\leq f_*(\theta_k)+\nabla_\theta f_*(\theta_k)^\intercal(\theta_{k+1}-\theta_k)+\frac{L}{2}\|\theta_{k+1}-\theta_k\|^2 . \tag{F.1}$$

由于 $\theta_{k+1}=\theta_k-\beta\nabla_\theta f_*(\theta_k)$，我们有 $\theta_{k+1}-\theta_k=-\beta\nabla_\theta f_*(\theta_k)$。将 $\theta_{k+1}-\theta_k$代入公式 F.1，得到

$$\begin{aligned}f_*(\theta_{k+1})-f_*(\theta_k)\leq & -\beta\nabla_\theta f_*(\theta_k)^\intercal\nabla_\theta f_*(\theta_k)+\frac{L\beta}{2}\|\nabla_\theta f_*(\theta_k)\|^2\\ = & \beta(\frac{L\beta}{2}-1)\|\nabla_\theta f_*(\theta_k)\|^2 .\end{aligned} \tag{F.2}$$

由于学习率 $\beta=\frac{1}{L}$，将 β 代入公式 F.2，得到

$$f_*(\theta_{k+1})-f_*(\theta_k)\leq -\frac{1}{2L}\|\nabla_\theta f_*(\theta_k)\|^2 . \tag{F.3}$$

因此，对不等式 F 3 两边的左侧和右侧分别对 $k=0$，……，T求和，. 得到

$$\sum_{k=0}^{T}[f_*(\theta_{k+1})-f_*(\theta_k)]\leq -\frac{1}{2L}\sum_{k=0}^{T}\|\nabla_\theta f_*(\theta_k)\|^2\Longrightarrow\sum_{k=0}^{T}\|\nabla_\theta f_*(\theta_k)\|^2\leq 2L(f_*(\theta_0)-f_*(\theta_T))$$

由于 $\theta^*=\arg\min_\theta f_*(\theta)$，我们有 $f_*(\theta_T)\geq f_*(\theta^*)$。于是

$$\sum_{k=0}^{T}\|\nabla_\theta f_*(\theta_k)\|^2\leq 2L(f_*(\theta_0)-f_*(\theta^*)) \tag{F.4}$$

因此，由于范数的平方均为正，范数平方的最小值必定小于范数平方的平均值。我们有

$$\min_{k=0,\ldots,T}\|\nabla_\theta f_*(\theta_k)\|^2\leq\frac{1}{T+1}\sum_{k=0}^{T}\|\nabla_\theta f_*(\theta_k)\|^2\leq\frac{2L(f_*(\theta_0)-f_*(\theta^*))}{T+1}. \tag{F.5}$$

根据假设 5.3，$|f_*(\theta)|\leq M$ 对所有 $\theta\in\Theta$ 以及 $f_*(\theta_0)\geq f_*(\theta^*)$ 均成立，我们可以得出结论

$$f_*(\theta_0)-f_*(\theta^*)=|f_*(\theta_0)-f_*(\theta^*)|\leq|f_*(\theta_0)|+|f_*(\theta^*)|\leq 2M.$$

因此，

$$\min_{k=0,\ldots,T}\|\nabla_\theta f_*(\theta_k)\|^2\leq\frac{1}{T+1}\sum_{k=0}^{T}\|\nabla_\theta f_*(\theta_k)\|^2\leq\frac{2L(f_*(\theta_0)-f_*(\theta^*))}{T+1}\leq\frac{4LM}{T+1}. \tag{F.6}$$

□

G 实验细节 G.1 DASAC

算法 6DASAC：多智能体强化学习中的基于 SAC 的可微仲裁。

输入： 步长序列 $\left(\{\alpha(i)t\}t\in[0,T-1],\ i\in\{1,\ 2\}\ \beta_k\right)$，初始正则化参数 $\lambda_0^i,\ \ i\in\{1,\ 2\}$，截断参数 $\left\{Q(i)\max,\mathcal{E}(i)\max\right\}i\in$

{1，2}，SAC 算法的目标熵 $\bar{}\mathrm{H}$。

初始化演员参数 $\theta_0\in\Theta$、边缘化 Q 估计器参数 $\psi_0\in\Psi$以及集中式联合 Q 估计器参数（即集中式评论家） $\xi_0\in\Xi$。

对于 k=0,1,… **执行**

$Q_{\psi_0^i}^{(i)}(s,\ a^i;\ \theta_k)\ \leftarrow 0(i\in\{1,\ 2\})$

对于t=0,1… **执行**

从经验回放缓冲区中采样转移样本（$s_t,\ \ u_t^i,\ \ u_t^{-i},\ \ r_t^i,\ \ s_{t+1}$）

// 更新演员

通过最小化 KL 散度来更新策略

$$\pi_{\phi_t^i}^i(\cdot|s_t;\theta_k)\leftarrow\arg\min_{\phi\in\Phi}\ D_{KL}\left(\pi_\phi||exp\left(\frac{1}{\lambda_t^i}Q_{\psi_t^i}(s_t,a_t^i)\right)\right)$$

// 优化集中式联合 Q 估计器

使用集中式联合Q函数计算下一状态的价值函数

$$V_{t+1}^{cent,(i)}=Q_{\xi_t^{(i)}}^{cent}(s_{t+1},a_t^i,a_t^{-i};\theta_k)-\lambda_t^i\log(\pi_{\phi_t^i}(a_t^i|s_{t+1};\theta_k))+\lambda_t^{-i}\log(\pi_{\phi_t^{-i}}(a_t^{-i}|s_{t+1};\theta_k))$$

通过最小化来更新参数 ξ^i

$$\left(r_t^i+(1-is_done)\gamma V_{t+1}^{cent,(i)}-Q_{\xi_t^{(i)}}^{cent}(s_t,u_t^{(i)},u_t^{(-i)};\theta_k)\right)^2$$

//优化边缘化 Q 估计器

从学习得到的集中式联合 Q 函数计算边缘化 $Q(i)$函数

$$\widehat{Q}_{\pi_{\phi_t}}^{(i)}(s,a_t^i;\theta_k)=\mathbb{E}_{\hat{a}_t^{-i}\sim\pi_{\phi_t}^{-i}}\left[Q_{\xi_t^{(i)}}^{cent}(s_t,a_t^i,\hat{a}_t^{-i};\theta_k)\right]$$

更新估计的理想能量函数

$$\widehat{Q}_{t+1}^{(i)}(s,a_t^i;\theta_k)=(1-\bar{\alpha}_t^{(i)})\cdot Q_{\phi_t^i}^{(i)}(s,a_t^i;\theta_k)+\bar{\alpha}_t^{(i)}\cdot\widehat{Q}_{\pi_{\phi_t}}^{(i)}(s,a_t^i;\theta_k).$$

更新边缘化 Q 估计器的参数 ψ，以获得 $\mathrm{Q}_{(i)}\psi^i(s,a^i;\ \theta)\ \in\mathrm{F}Q_{\max}^{(i)}$:

$$\psi_{t+1}\leftarrow\arg\min_{\psi}\ \mathbb{E}_{\sigma_t}\left[\sum_{i\in\{1,2\}}\left(Q_{\psi^i}^{(i)}(s,a_t^i;\theta)-\widehat{\mathcal{E}}_{t+1}^{(i)}(s,a_t^i;\theta)\right)^2\right].$$

// 自动调整正则化参数

更新正则化参数 λ^i

$$\lambda_{t+1}^i\leftarrow\arg\min_{\lambda}\ \mathbb{E}_{a_t^i\sim\pi_{\phi_t}^i}\left[\lambda_t^i\log\pi_{\phi_t}^i(a_t^i|s_t)-\lambda_t^i\bar{\mathcal{H}}\right]$$

$\psi_{(*)}(\theta_k)\ =\psi_{t+1}.$

结束循环

更新激励参数

$$\theta_{k+1}\leftarrow\theta_k-\beta_k\nabla f_*(\theta_k),$$

其中 $\nabla f_*(\theta_k)$ 定义见 (4.6)。

循环结束

G.2 GridSearch(M) 和贝叶斯优化基线

作为对比，我们对激励参数进行网格搜索，并将其作为一种零阶算法。首先，我们使用相同的改进的 MASAC 算法，对激励参数从 0.0 到 0.5 进行一轮粗粒度网格搜索，步长为 0.05；随后，分别针对两个博弈进行一轮细粒度搜索。对于捕食者-猎物博弈，我们在 0.3 到 0.35 的范围内进行网格搜索，步长为 0.005；对于剪刀石头布博弈，我们则以相同的步长搜索 0.35 到 0.4 的范围。

对于改进的零阶基线，贝叶斯优化，我们采用最常用的期望改进（EI）作为采集函数，并使用随机初始点，通过*L-BFGS-B* 方法最大化采集函数，以计算下一个采样点。

G.3 实现细节

G.3.1 计算资源。我们使用一块配备 24 GB GDDR6X 显存的 NVIDIA GeForce RTX 3090 图形处理器开展实验。我们使用带 CUDA 加速的 PyTorch 框架实现代码，可在一张 GPU 卡上并行运行。

G.3.2 超参数。

超参数	值
初始化方法	正交
GRU 层数	1
Actor MLP 隐状态维度	32
Actor RNN 隐状态维度	16
评论家多层感知机和循环神经网络隐藏状态维度	64
后续全连接层数量	1
优化器	Adam
优化器 epsilon	1e-5
权重衰减	0
激活函数	ReLU
使用奖励归一化	true
回合长度	25
最后的动作层增益	0.01
批次大小	64
缓冲区大小	5000
伽马	0.99
埃普西隆	从 1.0 到 0.05
埃普西隆退火时间	50000
Q 函数损失	均方误差损失
初始正则化参数	1
目标熵系数	0.3

表 1：DASAC 实验中使用的超参数。

Meichen Song*„ Feng GaoâĂą、Boyi Liu、Zhaoran Wang、Yi Wu„