arXiv:2304.01447v1 [cs.MA] 4 Apr 2023

多智能体强化学习中的离策略动作预期

Ariyan Bighashdel A.BIGHASHDEL@TUE.NL
Daan de Geus D.C.D.GEUS@TUE.NL
Pavol Jancura P.JANCURA@TUE.NL
Gijs Dubbelman G.DUBBELMAN@TUE.NL
电气工程系 埃因霍温理工大学 荷兰埃因霍温，*5612 AZ*

摘要

多智能体强化学习（MARL）中的学习预期是一种推理范式，智能体通过预期其他智能体的学习步骤来提升彼此之间的合作。由于 MARL 使用基于梯度的优化，学习预期需要使用高阶梯度（HOG），以及所谓的 HOG 方法。现有 HOG 方法基于策略参数预期，即智能体预期其他智能体策略参数的变化。然而，目前这些现有的 HOG 方法仅应用于可微博弈或状态空间较小的博弈。在这项工作中，我们展示了在具有大状态空间的不可微博弈中，现有 HOG 方法表现不佳且效率低下，这是由其与策略参数预期及多个采样阶段相关的固有限制所导致的。为解决这些问题，我们提出了离策略动作预期（OffPA2），这是一种新颖的框架，通过离策略采样，以动作预期的方式实现学习预期，即智能体预期其他智能体动作的变化。我们从理论上分析了所提出的 OffPA2，并据此开发了多种适用于具有大状态空间的不可微博弈的 HOG 方法。我们进行了大量实验，结果表明，与现有方法相比，我们提出的 HOG 方法在效率和性能方面更优。

关键词： 多智能体强化学习、推理、学习预期、高阶梯度、动作预期

1. 引言

在多智能体系统中，智能体对其他智能体进行推理 这一范式已得到广泛探索和研究（Goodie 等人，2012年；Liu 和 Lakemeyer，2021年）。近年来，这一范式也在多智能体强化学习（MARL）这一子领域中受到研究（Wen 等人，2019年、2020年；Konan 等人，2022年）。一般而言，多智能体强化学习涉及多个智能体在环境中同时学习并相互作用。在多智能体强化学习的背景下，一种推理策略是预判其他智能体的学习步骤（Zhang 和 Lesser，2010年），即学习预期。由于多智能体强化学习采用基于梯度的优化，学习预期自然会引出高阶梯度（HOG）的使用，以及所谓的 HOG 方法（Letcher 等人，2019年）。文献中经常展示学习预期的重要性

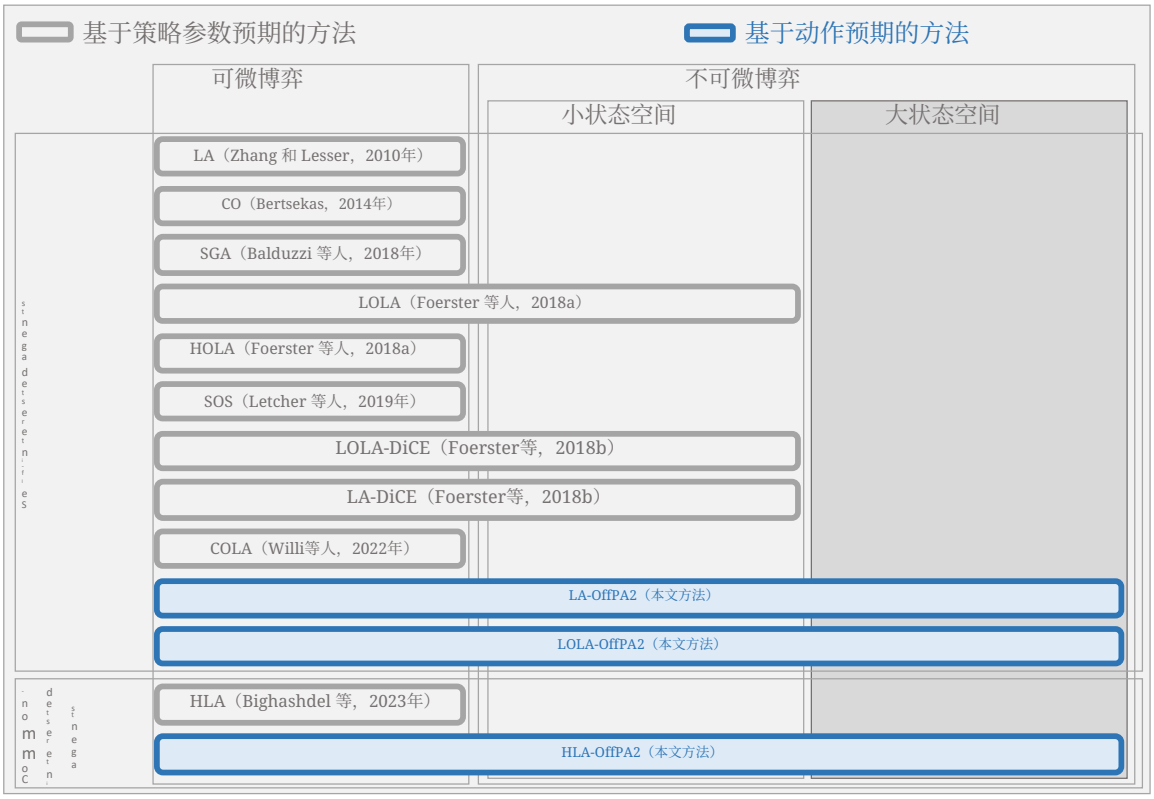


图1：HOG方法概览及其在各种博弈设定中的适用性。现有的HOG方法（灰色矩形）基于策略参数方法。这些方法仅应用于可微博弈，或状态空间较小的不可微博弈。我们提出的HOG方法（蓝色矩形）基于我们新颖的离策略动作预判（OffPA2）框架开发，可应用于状态空间较大的不可微博弈。

在 HOG 方法中的重要性已在文献中得到反复展示。例如，前瞻（LA）（Zhang 和 Lesser，2010年；Letcher 等人，2019年）利用学习预期来保证匹配硬币博弈等循环博弈中的收敛；具备对手学习意识的学习（LOLA）（Foerster 等人，2018a）利用学习预期来确保一般和博弈（如重复囚徒困境（IPD））中的合作；分层学习预判（HLA）（Bighashdel 等，2023年）则利用学习预期来提升完全合作博弈中共同利益智能体之间的协调。在本研究中，我们探讨当前 HOG 方法的局限性，并提出新的解决方案，使学习预期能够应用于更广泛的多智能体强化学习问题。图 1 概述了现有 HOG 方法和我们提出的方法的适用范围。

当前的 HOG 方法中的学习预期是基于策略参数预期 方法发展的，即智能体预期其他智能体的策略参数发生变化

（Zhang 和 Lesser，2010年；Foerster 等人，2018a,b）（见图1）。在这种方法中，首先，智能体必须能够获取其他智能体的精确参数，或根据状态-动作轨迹推断其他智能体的参数（Foerster 等人，2018a）。在许多博弈设定中，这些参数是不可见的。这会带来问题，因为随着状态空间规模增大，参数空间的维度也会随之增加，使参数推断在计算上代价高昂。此外，无论参数是推断得出的还是精确已知的，预期高维策略参数的变化都效率低下。最后，策略参数预期需要针对策略参数计算高阶梯度，而这在多智能体强化学习中已被证明颇具挑战性（Foerster等，2018b；Lu 等，2022）。当前的 HOG 方法主要假设博弈是可微的，即智能体能够获取梯度和海森矩阵（Willi等人，2022年；Letcher 等人，2019年）（图1a）。当博弈不可微时，现有 HOG 方法会结合同策略采样，采用随机策略梯度（SPG）定理（Sutton 和 Barto，2018年）来计算相对于策略参数的梯度（Foerster 等人，2018a,b）。然而，在 SPG 中估计高阶梯度，需要进行解析近似——因为标准 SPG 定理中一个智能体的学习步骤独立于其他智能体的参数——或采用多阶段采样；后者效率低下，且通常伴有较高方差，导致学习不稳定（Foerster 等人，2018a,b）。在本研究中，我们旨在提出新颖的 HOG 方法，以克服现有 HOG 方法的上述局限，使学习预期能够应用于具有大状态空间的不可微博弈。

为实现这一目标，我们提出了离策略动作预期（OffPA2），这是一种通过动作预期来学习预期的新型框架（见图1）。具体而言，OffPA2 中的智能体会在学习过程中预期其他智能体的动作变化。与策略参数预期不同，动作预期是在动作空间中进行的；在具有大状态空间的多智能体强化学习博弈中，动作空间的维度通常低于策略参数空间（Lowe 等，2017年；Peng 等，2021年）。此外，我们采用带离策略采样的确定性策略梯度（DPG）定理来估计可微目标函数。因此，能够高效计算高阶梯度，同时仍遵循标准的集中训练与分散执行（CTDE）设置，在该设置下，智能体在训练期间可以观察其他智能体的动作（Lowe 等，2017年）。我们从性能和时间复杂度两个方面对 OffPA2 进行了理论分析。所提出的 OffPA2 框架使我们能够开发高阶梯度方法；与现有的高阶梯度方法不同，这些方法适用于具有大状态空间的不可微博弈。为此，我们将 LA、LOLA 和 HLA 的原理应用于 OffPA2 框架，分别开发了 LA-OffPA2、LOLA-OffPA2 和 HLA-OffPA2 方法。我们在严格控制的研究中，将这些方法与现有高阶梯度方法进行了比较。结果表明，与现有高阶梯度方法的性能和效率会随着状态空间大小增加而急剧下降不同，我们提出的方法整体性能和效率不会因此下降。最后，我们将这些方法与标准的基于确定性策略梯度的多智能体强化学习算法进行比较，并强调了在多智能体强化学习中学习预期的重要性。下文将总结我们的贡献。

- 我们提出了 OffPA2，这是一种新颖的框架，通过动作预期来实现学习预期，从而使 HOGmethods 适用于不可微博弈

以及具有大状态空间的博弈。我们从理论上分析了所提出的动作预判方法对性能和时间复杂度的影响。

- 在我们的 OffPA2 框架中，我们提出了三种新方法，即 LA-OffPA2、LOLA-OffPA2 和 HLA-OffPA2。我们表明，我们的方法优于现有的 HOG 方法和最先进的基于确定性策略梯度的方法。

2. 相关工作

在许多现实世界的多智能体强化学习任务中，执行期间的通信约束要求采用分散式策略。在这些情况下，一种推理工具是智能体建模智能体（AMA）（Albrecht 和 Stone，2018年），即智能体显式地对其他智能体进行建模，以预测其行为。尽管 AMA 传统上假设对手是没有推理能力的天真型对手（He 等人，2016年；Hong 等人，2018年），近期研究已扩展 AMA，进一步考虑多个推理层级，让每个智能体考虑其他智能体的推理过程，从而做出更好的决策（Wen 等人，2019年，2020年）。例如，Wen 等人（2019年）提出了概率递归推理（PR2）更新规则，使多智能体强化学习智能体能够递归地推理其他智能体的信念。然而，这些方法并未考虑其他智能体的学习步骤；事实证明，在自利型智能体之间的交互否则会导致最坏情况结果的博弈中，这一点十分重要（Foerster 等人，2018a）。在第 5 节中，我们进行了若干实验，并将所提方法与这些方法进行了比较。

另一方面，HOG方法是一类使用高阶梯度来考虑其他智能体预期学习步骤的方法。这些方法包括：1）Foerster 等人（2018a）提出的 LOLA 和高阶LOLA（HOLA），旨在改善重复囚徒困境（IPD）中的合作；2）Zhang 和 Lesser（2010年）提出的前瞻（LA），旨在保证循环博弈中的收敛；3）Letcher 等人（2019年）提出的稳定对手塑形（SOS），作为 LOLA 和 LA 之间的插值，以兼得两者的优势；4）Willi等人（2022年）提出的一致性LOLA（COLA），旨在提高对手塑形的一致性；5）Bighashdel 等人（2023）提出的层次化学习预判（HLA），旨在改善完全合作型智能体之间的协调；6）Bertsekas（2014年）提出的一致性优化（CO），旨在提高零和博弈中的训练稳定性和收敛性；以及 7）Balduzzi 等人（2018年）提出的辛梯度调整（SGA），旨在提高 CO 的参数灵活性。

尽管大多数现有的 HOG 方法提出了新颖的理念，但它们仅应用于可微博弈，其中智能体可以获得精确的梯度或海森矩阵，而只有 LOLA 在不可微博弈上进行了评估。具体而言，Foerster 等人（2018a）采用 SPG 框架来估计 LOLA 中的梯度。由于标准 SPG 与其他智能体的参数无关，作者依赖于对期望回报进行泰勒展开，并结合二阶梯度的解析推导。Foerster 等人（2018b）指出，这种方法在学习过程中并不稳定。为了解决这一问题，Foerster 等人（2018b）提出了一种无限可微的蒙特卡洛估计量，称为 DiCE，以便正确优化具有任意阶梯度的随机目标函数。与元学习类似，DiCE 框架中的智能体利用内层学习循环推理并预测对手的学习步骤，并在外层学习循环中更新其参数。然而，每个智能体的每个学习循环都需要一个采样阶段

这对于高阶推理以及具有大状态空间的博弈而言效率极低，即超出矩阵博弈范围的博弈。在第 5 节中，我们开展了一系列实验，以便将我们提出的 OffPA2 框架与 DiCE 进行深入比较。

3. 问题表述与背景

我们将多智能体强化学习（MARL）设定表述为马尔可夫博弈（MG）（Littman，1994年）。MG 是一个元组 $(\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \{\mathcal{R}_i\}_{i\in\mathcal{N}}, \mathcal{T}, \rho, \gamma)$，其中 $\mathcal{N}$是智能体集合（$|\mathcal{N}| = n$），$\mathcal{S}$ 是状态集合，$\mathcal{A}_i$ 是智能体 $i \in \mathcal{N}$可能采取的动作集合。智能体 i 通过随机策略网络 π_{θ_i} 选择动作 $a_i \in \mathcal{A}_i$：π_{θ_i} 由 θ_i 参数化，并以给定状态 $s \in \mathcal{S}$为条件。给定所有智能体的动作，每个智能体 i 根据其奖励函数 $\mathcal{R}_i$：$\mathcal{S} \times \mathcal{A}_1 \times ... \times \mathcal{A}_n \to \mathbb{R}$获得奖励 r_i。给定初始状态，下一个状态按照状态转移函数 $\mathcal{T}$：$\mathcal{S} \times \mathcal{A}_1 \times ... \times \mathcal{A}_n \times \mathcal{S} \to [0, 1]$生成。我们将时域为 T 的回合记为 $\tau = (\{s^0, a_1^0, ..., a_n^0, r_1^0, ..., r_n^0\}, ..., \{s^T, a_1^T, ..., a_n^T, r_1^T, ..., r_n^T\})$，智能体 i在时间步 $t \leq T$的折扣回报定义为 $G_i^t(\tau) = \sum_{l=t}^{T} \gamma^{l-t} r_i$，其中 γ是预先设定的折扣因子。给定智能体的策略参数，期望回报可近似表示每个智能体的状态价值函数 $V_i(s, \theta_1, ..., \theta_n) = \mathbb{E}[G_i^t(\tau | s^t = s)]$。每个智能体 i 的目标是找到策略参数 θ_i，使得在初始状态分布 $\rho(s)$ 下的期望回报最大化，该目标称为性能目标 $J_i = \mathbb{E}_{\rho(s)} V_i(s, \theta_1, ..., \theta_n)$。

朴素¨梯度上升。在朴素更新规则中，智能体不会通过学习预期来更新其策略参数。具体而言，每个朴素智能体通过沿着目标函数梯度的方向更新策略参数，来最大化其性能目标

$$\nabla_{\theta_i} J_i = \mathbb{E}_{\rho(s)} \nabla_{\theta_i} V_i(s, \theta_1, ..., \theta_n). \tag{1}$$

具备对手学习意识的学习（LOLA）。与朴素智能体不同，LOLA 智能体通过对对手预期的学习步骤求导来修改其学习目标（Foerster 等人，2018a）。为简单起见，给定 $n = 2$，一阶 LOLA 智能体（智能体一）假设对手是朴素的，并使用策略参数预期来优化 $V_{\text{LOLA1}}(s, \theta_1, \theta_2 + \Delta\theta_2)$，其中 $\Delta\theta_2 = \mathbb{E}_{\rho(s)} \eta \nabla_{\theta_2} V_2(s, \theta_1, \theta_2)$，且 $\eta \in \mathbb{R}^+$ 为预测长度。利用一阶泰勒展开并对 θ_1 求导，首个 LOLA 智能体的梯度调整（Foerster 等人，2018a）表示为

$$\nabla_{\theta_1} V_1^{\text{LOLA}}(s, \theta_1, \theta_2 + \Delta\theta_2) \approx \nabla_{\theta_1} V_1 + (\nabla_{\theta_2\theta_1} V_1)^\intercal \Delta\theta_2 + \underbrace{(\nabla_{\theta_1} \Delta\theta_2)^\intercal \nabla_{\theta_2} V_1}_{\text{shaping}}, \tag{2}$$

其中 $V_1 = V_1(s, \theta_1, \theta_2)$。LOLA更新中最右侧的项可以主动塑造对手的学习过程。事实证明，该项能够有效促进各种博弈中的合作，包括 IPD（Foerster 等人，2018a,b）。LOLA更新还可以进一步扩展到非na¨ıve对手，从而得到 HOLA 智能体（Foerster 等人，2018a；Willi等人，2022年）。

前瞻（LA）。LA 智能体假设无法影响对手的学习步骤，即无法对其进行塑形（Zhang 和 Lesser，2010年；Letcher 等人，2019年）。换言之，智能体一

假设预测步骤 $\Delta\theta_2$ 与当前优化无关，即 $\nabla_{\theta_1}\Delta\theta_2 = 0$。因此，塑形项消失，第一个 LA 智能体的梯度调整将为

$$\nabla_{\theta_1} V_1^{\mathrm{LA}}(s, \theta_1, \theta_2 + \perp \Delta\theta_2) \approx \nabla_{\theta_1} V_1 + (\nabla_{\theta_2\theta_1} V_1)^{\intercal} \Delta\theta_2, \tag{3}$$

其中，$\perp$ 会阻止求导时梯度从 $\Delta\theta_2$ 流入。

层次化学习预判（HLA）。与 LOLA 和 LA 不同，HLA 旨在提升共同利益智能体在完全合作博弈中的协调能力（Bighashdel 等，2023年），即 $\mathcal{R}_i = \mathcal{R}_j = \mathcal{R}\ \forall i, j \in \mathcal{N}$，进而 $V_i = V_j = V\ \forall i, j \in \mathcal{N}$。HLA 会将智能体随机分配到不同层级，以确定其推理顺序。在每个层级中，被分配的智能体是较低层级的领导者，同时是较高层级的跟随者，并遵循两条推理规则：1）领导者了解跟随者的推理层级，并且比跟随者高一个层级；2）跟随者无法塑造领导者，只能遵循领导者的塑形计划。具体而言，若 $n = 2$，并假设智能体二是领导者（HLA-L），智能体一是跟随者（HLA-F），则领导者的梯度调整为：

$$\nabla_{\theta_2} V^{\text{HLA-L}}(s, \theta_1 + \Delta\theta_1, \theta_2) \approx \nabla_{\theta_2} V + (\nabla_{\theta_1\theta_2} V)^{\intercal} \Delta\theta_1 + (\nabla_{\theta_2}\Delta\theta_1)^{\intercal} \nabla_{\theta_1} V, \tag{4}$$

其中，$V = V(s,\ \theta_1,\ \theta_2)$ 是共同价值函数，且 $\Delta\theta_1 = \eta\nabla_{\theta_1} V$。领导者的计划是改变其参数 $\bar{\theta}_2 = \theta_2 + \eta\nabla_{\theta_2} V^{\text{HLA-L}}(s,\ \theta_1 + \Delta\theta_1,\ \theta_2)$，使得在跟随者将领导者的新参数纳入考量后，共同价值能够获得最优提升。因此，跟随者必须遵循该计划，并通过

$$\nabla_{\theta_1} V^{\text{HLA-F}}(s, \theta_1, \bar{\theta}_2) \approx \nabla_{\theta_1} V + (\nabla_{\theta_2\theta_1} V)^{\intercal} \eta \nabla_{\theta_2} V^{\text{HLA-L}}(s, \theta_1 + \Delta\theta_1, \theta_2). \tag{5}$$

4. 方法

本节提出 OffPA2，这是一种旨在将 HOG 方法应用于具有大状态空间的不可微博弈的框架。为解决策略参数预期相关的问题，我们提出了一种新颖的动作预期方法，即智能体在学习过程中预期其他智能体的动作变化。此外，我们采用离策略采样的确定性策略梯度定理来估计可微目标函数。因此，可以高效地计算高阶梯度。我们提出的 OffPA2 符合确定性策略梯度中的标准集中训练与分散执行（CTDE）设置，在训练期间，智能体可以访问其他智能体的动作（Lowe 等，2017年）。

4.1 OffPA2：离策略动作预期

我们为每个智能体 $i \in \mathcal{N}$定义一个由 θ_i 参数化的确定性策略 $\mu_{\theta_i} : \mathcal{S} \to \mathcal{A}_i$。令 $Q_i(s,\ a_1,\ \ldots\ldots,\ a_n) = \mathbb{E}[G_i^t(\tau | s^t = s,\ a_i^t = a_i \forall i \in \mathcal{N})]$ 表示状态-动作价值函数，则有 $V_i(s,\ \theta_1,\ \ldots\ldots,\ \theta_n) = Q_i(s, \mu_{\theta_1}(s)\ \ldots\ldots,\ \mu_{\theta_n}(s))$。此外，我们为每个智能体 i 定义一个随机行为策略 $\pi_{\theta_i}^b = \mu_{\theta_i} + SG$，其中 SG 服从标准高斯分布

分布。给定行为策略后，可以利用行为策略生成的轨迹离策略地学习策略参数，即 $\rho^b(s,\ \{a_i\}_{i\in\mathcal{N}},\ \{r_i\}_{i\in\mathcal{N}},\ s')$，其中 s和 s' 是连续状态。根据确定性策略梯度定理（Silver 等，2014年；Lowe 等，2017年），可以得到每个朴素智能体 i的性能目标梯度 $\nabla_{\theta_i}J_i=\mathbb{E}_{\rho^\beta(s)}\nabla_{\theta_i}Q_i(s,\ \mu_{\theta_1}(s)\ .....,\ \mu_{\theta_n}(s))$。

不失一般性，我们考虑两个智能体，即 $n=2$，并假设智能体一希望预期智能体二的学习步骤；智能体二是一名朴素学习者。在每个状态 $s\sim\rho^b(s)$ 下，智能体一将智能体二的策略参数变化预期为 $\Delta\theta_2(s)=\eta\nabla_{\theta_2}Q_2(s,\ \mu_{\theta_1}(s)\ \mu_{\theta_2}(s))$，即策略参数预期。因此，智能体一沿着以下方向更新其策略参数：

$$\nabla_{\theta_1}J_1=\mathbb{E}_{\rho^\beta(s)}\nabla_{\theta_1}Q_1(s,\mu_{\theta_1},\mu_{\theta_2+\Delta\theta_2(s)}(s)),\tag{6}$$

定理 1（动作预期）利用一阶泰勒展开，可以近似表示智能体一的性能目标梯度（式（6））：

$$\nabla_{\theta_1}J_1\approx\mathbb{E}_{\rho^\beta(s)}\nabla_{\theta_1}\mu_{\theta_1}(s)\nabla_{a_1}Q_1(s,a_1,a_2+\Delta a_2)|_{a_1=\mu_{\theta_1}(s),a_2=\mu_{\theta_2}(s)},\tag{7}$$

其中

$$\Delta a_2=\hat{\eta}_{1st}\nabla_{a_2}Q_2(s,a_1,a_2),\tag{8}$$

表示预期动作变化，其中 $\hat{}\eta_{1st}=\eta\left\|\nabla_{\theta_2}\mu_{\theta_2}(s)\right\|^2\in\mathbb{R}^+$ 是投影预测长度。

证明。见附录 A.1。

定理 1 表明，智能体一可以在动作空间中预期智能体二的学习步骤，而非在策略参数空间中进行预期。这种推理方式有两个优点。首先，在具有大状态空间的多智能体强化学习博弈中，动作空间的维度显著低于策略参数空间（Lowe 等，2017年；Peng 等，2021年）。这是因为，大状态空间需要更复杂的策略网络和更多参数，才能恰当地表示所有可能的状态。其次，与策略参数预期不同，动作预期符合确定性策略梯度中的标准集中式训练与分散式执行（CTDE）设置。在标准 CTDE 设置中，智能体在训练期间可以访问集中式状态-动作价值函数，以训练分散式策略。因此，智能体能够获知其他智能体的动作，并在训练期间进行动作预期。而在策略参数预期中，智能体还需要访问其他智能体的策略参数。

4.1.1 Influence of action anticipation on performance

我们提出的动作预判方法采用一阶泰勒近似，将对策略参数空间中学习情况的预期映射到动作空间。换言之：

$$\mu_{\theta_2+\Delta\theta_2(s)}(s)\approx a_2+\Delta a_2,\tag{9}$$

其中 $a_2 = \mu_{\theta_2}(s)$ 和 $\Delta a_2 = \hat{}\eta_{1\text{st}} \nabla_{a_2} Q_2(s, a_1, a_2)$。在下文的定理中，我们展示了这种近似如何影响 HOG 方法的原理；这些原理已在相关文献中得到理论和实验研究（Zhang 和 Lesser，2010年；Foerster 等人，2018a；Letcher 等人，2019年；Willi等人，2022年）。

定理 2 对于足够小的 $\hat{}\eta_{1st}$，存在一个 $\eta' \in \mathbb{R}^+$，使得

$$\mu_{\theta_2 + \Delta\theta'_2(s)}(s) = a_2 + \Delta a_2, \tag{10}$$

其中

$$\begin{aligned} &\Delta\theta'_2(s) = \eta' \nabla_{\theta_2} \mu_{\theta_2}(s) \nabla_{a_2} Q_2(s, a_1, a_2) \\ &a_2 = \mu_{\theta_2}(s) \\ &\Delta a_2 = \hat{\eta}_{1st} \nabla_{a_2} Q_2(s, a_1, a_2) \qquad \hat{\eta}_{1st} = \eta \left\| \nabla_{\theta_2} \mu_{\theta_2(s)} \right\|^2 \qquad \eta \in \mathbb{R}^+ \end{aligned} \tag{11}$$

证明。见附录 A.2。

根据定理 2，通过一阶泰勒展开进行动作预期，且 $\hat{}\eta_{1\text{st}}$ 足够小时，只会缩放预测长度，因为 η 和 η' 均为非负数。关于可微博弈的一般理论分析表明，缩放预测长度会影响 HOG 方法的收敛行为（Letcher 等人，2019年；Zhang 和 Lesser，2010年）。然而，在我们的 OffPA2 框架中，我们直接设定投影后的 预测长度，即 $\hat{}\eta_{1\text{st}}$，而非预测长度，即 η。因此，在策略参数空间中得到的预测长度，即 η'，能够对应令人满意的收敛行为（见第 5.2.3 节）。

4.1.2 Computing higher-order gradients

式 (7) 中的状态-动作价值函数通常是未知且不可微的。与基于确定性策略梯度的算法（Silver 等，2014年；Lowe 等，2017年）类似，我们用一个由 ω_i 参数化的可微状态-动作价值函数 $Q_i(s, a_1, ..., a_n; \omega_i)$，替代真实的状态-动作价值函数，即 $Q_i(s, a_1, ..., a_n; \omega_i) \approx Q_i(s, a_1, ..., a_n)$。状态 - 动作价值函数的参数可通过最小化由行为策略生成的回合中的离策略时间差分（TD）误差来获得（Lowe 等，2017年）：

$$\mathcal{L}(\omega_i) = \mathbb{E}_{\rho^b(s, \{a_i\}_{i \in \mathcal{N}}, \{r_i\}_{i \in \mathcal{N}}, s')}[(Q_i(s, a_1, ..., a_n; \omega_i) - y_i)^2], \tag{12}$$

其中 y_i 是TD目标值：

$$y_i = r_i + \gamma Q'_i(s, a'_1, ..., a'_n)|_{a'_i = \mu'_i(s') \; \forall i \in \mathcal{N}}, \tag{13}$$

其中，Q'_i 和 μ'_i 分别是目标状态-动作价值和策略函数。OffPA2 中目标函数的可微性对 HOG 方法尤其有益，因为这类方法需要频繁计算高阶梯度，以预判智能体的学习过程。

4.1.3 Influence of action anticipation on time complexity

除了目标函数的可微性外，动作预判方法还能降低梯度计算的复杂度，因为它需要的是预期动作变化，即 Δa_i，而非预期策略参数变化，即 $\Delta\theta_i$。假设策略网络和状态-动作价值网络都是多层感知机（与大多数实验中的做法相同），则：

定理 3 与策略参数预期相比，动作预期降低了对天真对手学习步骤进行预期的时间复杂度，降低幅度为 $O(LN^2)$，其中 L是策略网络和状态-动作价值网络中的全连接层数，N 是每层的神经元数量。

证明。参见附录 A.3。

4.2 基于 OffPA2 的 HOG方法

有了 OffPA2 框架，我们现在可以开发适用于具有大状态空间的不可微博弈的 HOG方法。在以下各节中，我们分别将 LOLA、LA 和 HLA 原则应用于 OffPA2 框架，以开发 LOLA-OffPA2、LA-OffPA2 和 HLA-OffPA2。

4.2.1 LOLA-OffPA2

如第 3 节所述，LOLA 智能体会预测并塑造其他智能体的学习步骤，以提升非团队博弈中的合作。为简单起见，设有两个智能体（$n = 2$）。LOLA-OffPA2 智能体（智能体一）预测并塑造对手（智能体二）的动作，并假定对手是天真的。利用一阶泰勒展开，第一个 LOLA-OffPA2 智能体的梯度调整为

$$\begin{aligned}\nabla_{\theta_1} J_1^{\text{LOLA-OffPA2}} &= \mathbb{E}_{\rho^\beta(s)} \nabla_{\theta_1}\mu_{\theta_1}(s) \nabla_{a_1} Q_1(s, a_1, a_2 + \Delta a_2)|_{a_1=\mu_{\theta_1}(s), a_2=\mu_{\theta_2}(s)} \\ &\approx \mathbb{E}_{\rho^\beta(s)} \nabla_{\theta_1}\mu_{\theta_1}(s) \left(\nabla_{a_1} Q_1 + (\nabla_{a_2 a_1} Q_1)^\intercal \Delta a_2 + \underbrace{(\nabla_{a_1} \Delta a_2)^\intercal \nabla_{a_2} Q_1}_{\text{action shaping}}, \right)\end{aligned} \tag{14}$$

其中

$$\begin{aligned}\Delta a_2 &= \hat{\eta}_{1\text{st}} \nabla_{a_2} Q_2(s, a_1, a_2)|_{a_1=\mu_{\theta_1}(s), a_2=\mu_{\theta_2}(s)} \\ Q_1 &= Q_1(s, a_1, a_2)|_{a_1=\mu_{\theta_1}(s), a_2=\mu_{\theta_2}(s)}.\end{aligned} \tag{15}$$

LOLA-OffPA2 更新中的最右项，即公式（14），允许主动对对手进行 动作塑造 。

在实践中，我们无需依赖泰勒展开来计算 LOLA-OffPA2 中的更新规则，因为可以使用自动微分引擎，例如 PyTorch 自动微分（Paszke 等，2019），直接计算梯度。附录中的算法 1说明了

适用于 n 智能体的 LOLA-OffPA2 优化框架。在每个状态 $s \sim \rho^b$（s）下，智能体 $i \in \mathcal{N}$ 首先预测所有智能体 $j \in \{\mathcal{N} - \{i\}\}$的动作变化：

$$\Delta a_j = \hat{\eta}_{1\text{st}} \nabla_{a_2} Q_2(s, a_1, ..., a_n)|_{a_i = \mu_{\theta_i}(s)\ \forall i \in \mathcal{N}}. \tag{16}$$

然后，智能体 i 通过以下梯度调整来更新其参数 θ_i：

$$\nabla_{\theta_i} J_i^{\text{LOLA-OffPA2}} = \mathbb{E}_{\rho^\beta(s)} \nabla_{\theta_i} \mu_{\theta_i}(s) \nabla_{a_i} Q_i(s, a_1 + \Delta a_1, ..., a_i, ..., a_n + \Delta a_n)|_{a_i = \mu_{\theta_i}(s)\ \forall i \in \mathcal{N}}. \tag{17}$$

公式（17）表示假设对手是天真的一阶 LOLA-OffPA2 智能体的更新规则。不过，我们也可以考虑二阶 LOLA-OffPA2 智能体，它会沿着一阶 LOLA-OffPA2 对手的学习步骤进行求导。同样，我们可以像 HOLA（Foerster 等人，2016年）那样，将更新规则扩展为包含高阶推理。

4.2.2 LA-OffPA2

与 LA 原则（Zhang 和 Lesser，2010年；Letcher 等人，2019年）类似，LA-OffPA2 智能体无法影响对手的学习步骤，也就是说，它们无法影响对手的动作。因此，在双智能体情形下，我们有$\nabla_{a_1} \Delta a_2 = 0$。利用一阶泰勒展开，第一个 LA-OffPA2 智能体（Foerster 等人，2018a）的梯度调整如下：

$$\begin{aligned} \nabla_{\theta_1} J_1^{\text{LA-OffPA2}} &= \mathbb{E}_{\rho^\beta(s)} \nabla_{\theta_1} \mu_{\theta_1}(s) \nabla_{a_1} Q_1(s, a_1, a_2 + \perp \Delta a_2)|_{a_1 = \mu_{\theta_1}(s), a_2 = \mu_{\theta_2}(s)} \\ &\approx \mathbb{E}_{\rho^\beta(s, \hat{a}_2)} \nabla_{\theta_1} \mu_{\theta_1}(s) \left(\nabla_{a_1} Q_1 + (\nabla_{a_2 a_1} Q_1)^\intercal \Delta a_2 \right), \end{aligned} \tag{18}$$

其中，$\perp$ 可防止在求导时梯度从 Δa_2 流入，且 Δa_2 和 Q_1 在式（15）中定义。

与 LOLA-OffPA2 的情况一样，我们可以使用自动微分引擎直接计算梯度。附录中的算法2展示了针对 n 智能体的 LA-OffPA2 优化框架。在每个状态 $s \sim \rho^b$（s）下，智能体 $i \in \mathcal{N}$ 首先使用式（16）预测所有智能体 $j \in \{\mathcal{N} - \{i\}\}$ 的动作变化。然后，智能体i 通过以下梯度调整来更新其参数 θ_i：

$$\nabla_{\theta_i} J_i^{\text{LA-OffPA2}} = \mathbb{E}_{\rho^\beta(s)} \nabla_{\theta_i} \mu_{\theta_i}(s) \nabla_{a_i} Q_i(s, a_1 + \perp \Delta a_1, ..., a_i, ..., a_n + \perp \Delta a_n)|_{a_i = \mu_{\theta_i}(s)\ \forall i \in \mathcal{N}}. \tag{19}$$

4.2.3 HLA-OffPA2

如前所述，HLA 旨在改进具有共同利益智能体的完全合作博弈中的协调。为构建 HLA-OffPA2，我们首先将 $\mathcal{M} \subseteq \mathcal{N}$定义为一个由 $m = |\mathcal{M}|$ 个智能体组成的集合，这些智能体具有共同奖励函数 $\mathcal{R} = \mathcal{R}_i = \mathcal{R}_j\ \forall i,\ j \in \mathcal{M}$。不失一般性，我们考虑 $\mathcal{M} = \mathcal{N}$，即具有共同状态-动作价值函数 Q（s，a_1，.....，a_m）的团队博弈。

层级分配。与 HLA 类似，HLA-OffPA2 中的每个智能体首先会被分配到 m 个层级中的一个，其中第一关是最低层级，而层级 m 是

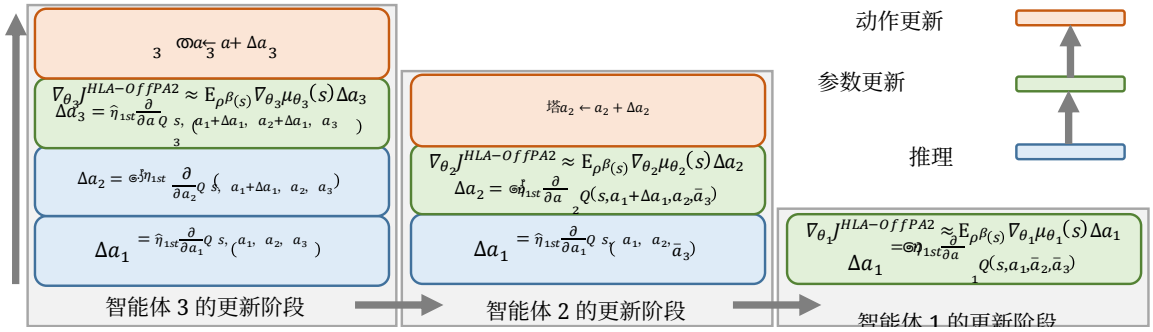


图2：HLA-OffPA2 中参数更新阶段的示例，适用于一场有三个共同利益智能体的游戏，其中智能体 1、智能体2 和智能体 3 分别被分配到层级 1、层级 2 和层级 3。

最高层级。尽管层级分配可以像 Bighashdel 等人（2023）提出的那样随机进行，但我们利用智能体对其他智能体产生的影响程度，即其塑造能力，作为衡量智能体层级的指标（实验比较见第 5.3.3 节）。我们将 i^{th} 智能体 $\mathcal{SC}_i$ 的塑造能力定义为其相对于所有其他智能体的动作塑造值之和 j:

$$\mathcal{SC}_i = \sum_{j \in \{\mathcal{M}-\{i\}\}} \left\| (\nabla_{a_i} \Delta a_j)^\intercal \nabla_{a_j} Q(a_1, ..., a_m) \right\|, \tag{20}$$

其中 $\Delta a_j = \nabla_{a_j} Q(a_1, \ldots, a_m)$。塑造能力最高的智能体被分配到最高层级，依此类推。由于 HLA-OffPA2 得益于集中式学习，HLA推理规则唯一仍需满足的约束是集中式状态-动作价值函数。

更新规则。完成层级分配后，智能体以自上而下的方式分 m 个更新阶段更新其策略参数，即每个智能体对应一个阶段：层级最高的智能体首先更新其策略参数。在每个更新阶段，相应的智能体会：1）以自下而上的方式推理跟随者（如有）的动作，即先对层级最低的智能体进行推理；2）更新其策略参数；3）更新其在下一更新阶段（如有）的动作。

如果我们设定 $m = 2$，并假设智能体二是领导者（HLA-OffPA2-L），智能体一是跟随者（HLA-OffPA2-F），则领导者的梯度调整为

$$\begin{aligned} \nabla_{\theta_2} J^{\text{HLA-OffPA2-L}} &= \mathbb{E}_{\rho^\beta(s)} \nabla_{\theta_2} \mu_{\theta_2}(s) \nabla_{a_2} Q(s, a_1 + \Delta a_1, a_2)|_{a_1 = \mu_{\theta_1}(s), a_2 = \mu_{\theta_2}(s)} \\ &\approx \mathbb{E}_{\rho^\beta(s)} \nabla_{\theta_2} \mu_{\theta_2}(s) \left(\nabla_{a_2} Q + (\nabla_{a_1 a_2} Q)^\intercal \Delta a_1 + (\nabla_{a_2} \Delta a_1)^\intercal \nabla_{a_1} Q \right), \end{aligned} \tag{21}$$

其中

$$\begin{aligned} \Delta a_1 &= \hat{\eta}_{1\text{st}} \nabla_{a_1} Q(s, a_1, a_2)|_{a_1 = \mu_{\theta_1}(s), a_2 = \mu_{\theta_2}(s)} \\ Q &= Q(s, a_1, a_2)|_{a_1 = \mu_{\theta_1}(s), a_2 = \mu_{\theta_2}(s)}. \end{aligned} \tag{22}$$

领导者的塑造计划是将其动作改变为

$$\bar{a}_2 = a_2 + \hat{\eta}_{\text{1st}} \nabla_{a_2} Q(s, a_1 + \Delta a_1, a_2)|_{a_1=\mu_{\theta_1}(s), a_2=\mu_{\theta_2}(s)}, \tag{23}$$

从而在跟随者将其新动作纳入考量后，实现共同状态-动作价值的最优提升。因此，跟随者通过以下方式调整其参数:

$$\begin{aligned} \nabla_{\theta_1} J^{\text{HLA-OffPA2-F}} &= \mathbb{E}_{\rho^{\beta}(s)} \nabla_{\theta_1} \mu_{\theta_1}(s) \nabla_{a_1} Q(s, a_1, \bar{a}_2)|_{a_1=\mu_{\theta_1}(s)} \\ &\approx \mathbb{E}_{\rho^{\beta}(s)} \nabla_{\theta_1} \mu_{\theta_1}(s) \left(\nabla_{a_1} Q + (\nabla_{a_1 a_2} Q)^{\intercal} \hat{\eta}_{\text{1st}} \nabla_{a_2} Q(s, a_1 + \Delta a_1, a_2) \right). \end{aligned} \tag{24}$$

与 LOLA-OffPA2 和 LA-OffPA2 的情况相同，我们可以使用自动微分引擎直接计算梯度。为阐明 HLA-OffPA2 中针对$m > 2$的更新规则，我们以图2中的三个共同利益智能体为例展示更新阶段，其中智能体一、智能体二和智能体三分别被分配到层级1、层级2和层级3。对于 m 智能体的情况，请参见附录中的算法3所示的 HLA-OffPA2优化框架。

5. 实验

本节开展一系列实验，旨在实现两个主要目标：1）展示学习预期在更广泛的多智能体强化学习问题中的优势，包括具有大状态空间的不可微博弈；2）展示我们提出的动作预判方法相较于策略参数预期的优势。

为实现第一个目标，我们将所提出的方法与多智能体深度确定性策略梯度（MADDPG）（Lowe等，2017年）进行比较，并为其配置了三种最先进的更新规则：1）标准更新规则（Lowe等，2017年），称为MADDPG；2）集中式策略梯度（CPG）更新规则（Peng 等，2021年），称为CPG-MADDPG；3）概率递归推理（PR2）更新规则（Wen等，2019年），称为PR2-MADDPG。为实现第二个目标，我们将基于OffPA2的所提方法与能够求解不可微博弈的现有HOG方法进行比较（见图1）。具体而言，我们分别将LOLA-OffPA2和LA-OffPA2与LOLA-DiCE和LA-DiCE进行比较。先前的研究（Foerster等，2018b）表明，LOLA-DiCE在IPD中的表现显著优于LOLA，因此我们不将LOLA-OffPA2与LOLA进行比较。由于LOLA-DiCE和LA-DiCE均基于策略参数预期，通过这些实验，我们可以凸显新颖的动作预判方法所带来的优势。与Foerster等人（2018b）的实现类似，DiCE中的智能体可以访问其他智能体的策略参数。由于原始HLA方法只能应用于可微博弈，因此我们不将HLA-OffPA2与HLA进行比较。

除非另有说明，我们根据（归一化的）平均回合奖励（AER）评估方法的性能，数值越高表示性能越好。此外，我们根据学习预期时间复杂度（LATC）评估 HOG 方法的效率，对于 HOG 方法 $\mathcal{H}$ ，其计算方式如下:

$$\text{LATC}(\mathcal{H}) = \frac{\text{per iteration training time of } \mathcal{H}}{\text{per iteration training time of the naïve version of } \mathcal{H}} - 1 \geq 0 \tag{25}$$

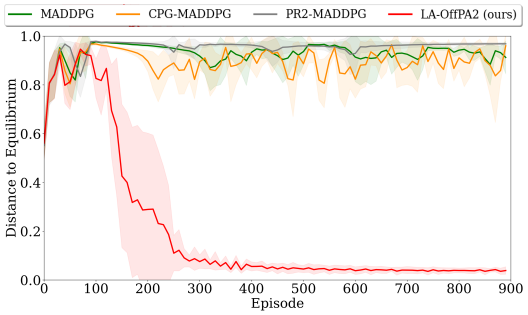


图 3：迭代旋转博弈中关于到均衡点的距离的学习曲线（↓）。

方法	到均衡的距离 ↓
LA-DiCE	0.09±0.07
LA-OffPA2（我们的方法）	**0.03±0.02**

方法	学习预期时间复杂度 ↓
LA-DiCE	1.06
LA-OffPA2（我们的方法）	**0.13**

表 1：迭代旋转博弈中 DiCE 框架与我们提出的 OffPA2 框架的 LA 原则比较。

其中， $\mathcal{H}$ 的朴素版本不执行学习预期。LATC 值越低，表示效率越高；LATC $= 0$ 意味着学习预期不会为算法增加任何时间复杂度。在接下来的章节中，我们将分别评估所提出的方法并讨论结果（实现细节请参见附录B）。

5.1 LA-OffPA2 的评估

我们在 Zhang 和 Lesser（2010）提出的不可微版本旋转博弈上评估这些方法，并将其称为迭代旋转博弈（IRG）。IRG 是一个单状态、双智能体、单动作（连续）矩阵博弈，其奖励如表 2 所示（针对两个离散动作）。然而，智能体无法访问奖励表，并且可以

	离散的动作 1	离散的动作 2
离散的动作 1	$(0, 3)$	$(3, 2)$
离散动作 2	$(1, 0)$	$(2, 1)$

表 2：迭代旋转博弈中的奖励。

它们的联合动作才能获得奖励。每个智能体 $i \in \{1, 2\}$ 都必须选择一个一维连续动作（$0 \leq a_i \leq 1$ 表示采取两种离散动作的概率。该博弈在 $a_1 = a_2 = 0.5$处有唯一的均衡点，同时也是该博弈的不动点。旋转博弈最初旨在展示：如果智能体遵循朴素的梯度更新，可能会出现循环行为。另一方面，LA 智能体可以通过考虑对手的参数调整，迅速收敛到均衡点。我们根据到均衡的距离（DtE）评估各方法的性能，该距离是当前动作与均衡点之间的欧几里得距离。

图 3 展示了 LA-OffPA2 与其他先进的基于 MADDPG 的算法的学习曲线。从图中可以看出，LA-OffPA2 是唯一能够收敛到均衡动作的方法。这些结果凸显了在 IRG 中学习预期的重要性。为了进一步展示 LA-OffPA2 的有效性，我们将 LA-OffPA2 方法与 LA-DiCE（Foerster等，2018b）进行比较，并在表 1 中报告结果。

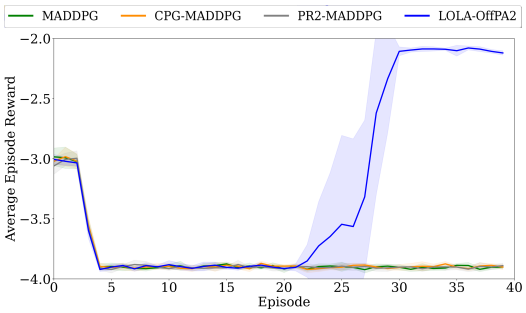


图 4：迭代囚徒困境中平均回合奖励（↑）的学习曲线。

方法	平均回合奖励 ↑
LOLA-DiCE	-2.16±0.12
LOLA-OffPA2（我们的方法）	**-2.08±0.02**

方法	学习预期时间复杂度 ↓
LOLA-DiCE	1.18
LOLA-OffPA2（我们的方法）	**0.16**

表 3：在重复囚徒困境中，比较 DiCE 框架与我们提出的 OffPA2 框架中的 LOLA 原则。

从表 1 中的 DtE 结果可以看出，这两种方法都能解决这些博弈，而 LA-OffPA2 的结果略好一些。不过，若比较这两种方法的学习预期时间复杂度（LATC），可以明显看出，LA-OffPA2 比 LA-DiCE 高效得多，这表明我们的 OffPA2 框架在结合 DiCE 时具有优势。

5.2 LOLA-OffPA2 的评估

5.2.1 Iterated prisoner's dilemma

重复囚徒困境（IPD）（Foerster 等人，2018a）是一个五状态、双智能体、双动作博弈，其奖励矩阵如表 4 所示。每个智能体必须在两种离散动作（合作或背叛）之间做出选择。该博弈进行 150 个时间步（$T = 150$）。在单次博弈版本中，智能体只有一个纳什均衡（背叛，背叛）。在重复博弈中，（背叛，背叛）也是一个纳什均衡。然而，一个更好的均

	合作	缺陷
合作	$(-1, -1)$	$(-3, 0)$
背叛	$(0, -3)$	$(-3, -3)$

表 4：重复囚徒困境中的奖励。

均衡策略是以牙还牙（TFT）：玩家先合作，之后重复对手上一轮的动作。LOLA 智能体能够影响对手的学习，以促进合作，因此最终收敛到 TFT（Letcher 等人，2019年）。我们根据平均回合奖励（AER）评估这些方法的性能。

在图 4 中，我们展示了 LOLA-OffPA2 和基于 MADDPG 的方法的学习曲线。从该图可以看出，只有 LOLA-OffPA2 能够解决该游戏，这再次凸显了学习预期的重要性。此外，我们将 LOLA-OffPA2 与专门为该游戏设计的 LOLA-DiCE（Foerster等，2018b）的性能进行了比较，并在表 3 中报告了结果。尽管两种方法都表现出较高的 AER 值，但我们的 LOLA-OffPA2 明显更加高效，因为其 LATC 值远低于 LOLA-DiCE。

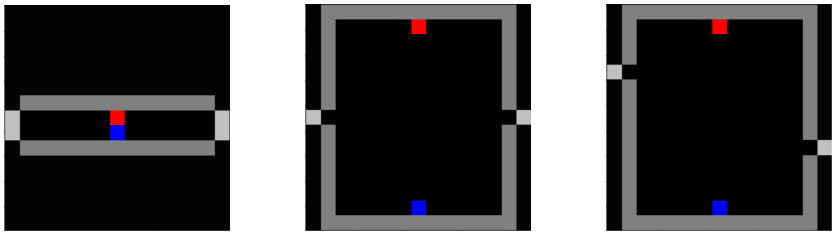


图5：Exit-Room 游戏第一关（左）、第二关（中）和第三关（右）的状态观测。

	归一化平均回合奖励 ↑			学习预期时间复杂度 ↓				
方法	$l=1$	$l=2$	$l=3$	朴素	一阶	二阶	三阶	四阶
LOLA-DiCE	0.91±0.04	0.68±0.06	0.56±0.12	0.00	1.39	2.74	4.12	5.41
LOLA-OffPA2（我们的方法）	**1.00±0.00**	**0.99±0.01**	**0.93±0.03**	0.00	**0.24**	**0.47**	**0.69**	**0.94**

表 5：LOLA-DiCE 与我们提出的 LOLA-OffPA2 在 Exit-Room 游戏中的比较，比较指标包括性能（不同游戏等级下的归一化平均回报）和效率（不同推理层级下的学习预期时间复杂度）。

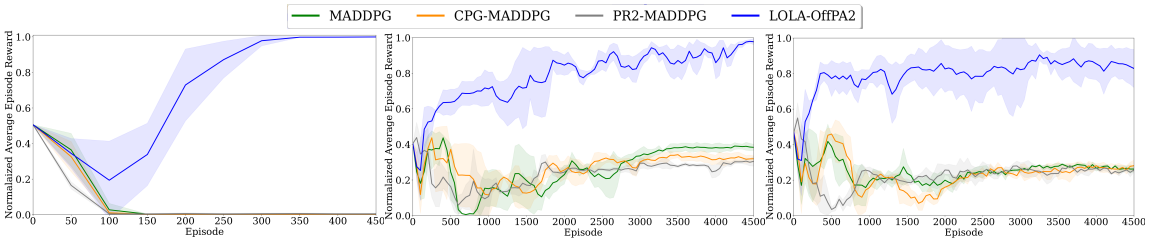
5.2.2 Multi-level Exit-Room game

在第二个实验中，我们评估 LOLA-OffPA2 在大状态空间博弈中的能力，这正是我们设想的框架应用场景。受 Vinitsky等（2019）启发，我们提出了一个包含三个复杂度等级的 Exit-Room 游戏（见图5）。Exit-Room 游戏是 IPD 的一种网格世界变体，包含两个智能体（蓝色和红色）以及 15^{2l} 个状态，其中 $l \in \{1,2,3\}$ 表示博弈的复杂度等级。智能体应当协作并向右侧的出口门移动。然而，它们也可能受到诱惑，选择从左侧出口离开，某些情况下甚至完全不离开。在等级 1 中，智能体有三种可能的动作（向左移动、向右移动或什么也不做），奖励的计算方式与 Vinitsky等（2019）相同：

$$\begin{aligned} \text{reward}_C &= \lambda_C(\text{cooperation}_{self} + \text{cooperation}_{opponent}) \\ \text{reward}_D &= \lambda_D(1 - \text{cooperation}_{self}) \\ \text{reward} &= \text{reward}_C + \text{reward}_D, \end{aligned} \tag{26}$$

其中， λ_C 和 λ_D 是一些常数，cooperation^{self} 和cooperation^{opponent} 分别是智能体及其对手到右侧门的归一化距离。在第2级和第3级中，智能体还可以执行向上移动 和向下移动动作。在第 3 级中，门的位置是随机设定的，因此智能体之间的交互更加复杂。除公式（26）中的奖励外，在第2级和第3级中，智能体还会因接近门而获得额外奖励。每个智能体会收到四张 90×90 RGB图像，代表最近四个时间步的状态观测。

图 6 对比了 LOLA-OffPA2 和基于 MADDPG 的方法在归一化平均回合奖励（NAER）方面的学习曲线。NAER 是指将每个游戏关卡中的 AER 值按最高和最低回合奖励进行归一化后的结果。从图 6 中，我们可以



(a) Exit-Room 第一关。 (b) Exit-Room 第二关。 (c) Exit-Room 第三关。

图6：Exit-Room 游戏在不同复杂度等级下的归一化平均回合奖励学习曲线（↑）。

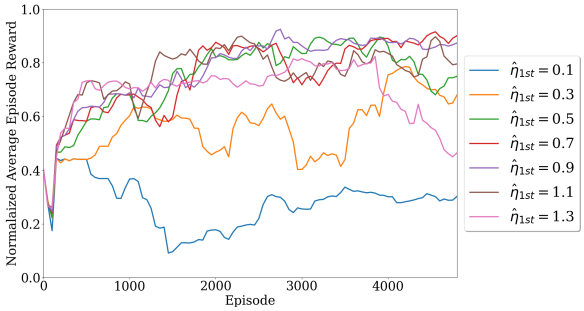


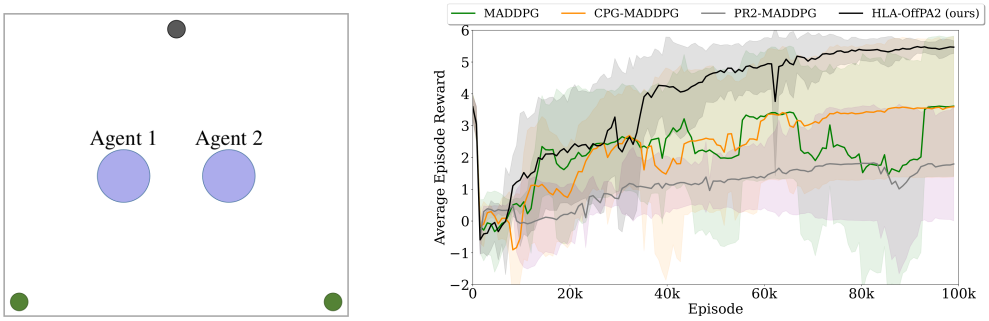
图7：投影预测长度（$\hat{\eta}_{1st}$）对 LOLA-OffPA2 在第三关 Exit-Room 游戏中收敛行为的影响。

清楚地看到，我们的 LOLA-OffPA2 显著优于其他方法，这与 IPD 矩阵博弈中的情况类似。

为突出我们提出的方法相较于现有 HOG 方法的优势，我们在表5中从性能（比较 NAER）和训练效率（比较 LATC）两方面，将 LOLA-OffPA2 与 LOLA-DiCE 进行了比较。观察表5可知，LOLA-DiCE 未能获得最高奖励，在游戏第二关和第三关中尤为明显，因为这两关的状态空间大小有所增加。这是因为状态空间规模增大时，LOLA-DiCE 智能体无法通过采样准确近似高阶梯度，因此无法进行学习预期以获得更高奖励。相比之下，我们提出的 LOLA-OffPA2 在游戏所有关卡中的 NAER 表现都更好，而且从朴素推理提升到高阶推理的效率显著高于 LOLA-DiCE。这表明我们克服了第1节所述 HOG 方法的局限性。

5.2.3 Influence of the projected prediction length

根据定理 2 中的研究结果，通过一阶泰勒展开进行动作预期会缩放预测长度，并影响 HOG方法的收敛行为。在此，我们通过直接改变投影预测长度，即



(a) 游戏示意图 (b) 学习曲线

图8：粒子协同博弈。(a) 双智能体（即紫色圆圈）与三个地标（即灰色和绿色圆圈）的博弈示意图。(b) 平均回合奖励方面的学习曲线（↑）。建议以彩色查看。

$\hat{}\eta_{1\text{st}}$，来调节状态空间中的最终预测长度，即 η'，从而改善收敛行为。我们开展了一项实验研究，以分析 $\hat{}\eta_{1\text{st}}$ 对三级出口房间博弈中 LOLA-OffPA2 收敛行为的影响（见图7）。实验重复进行四次，并以图7中的 NEAR 均值结果进行报告。图7清楚地表明，通过调节 $\hat{}\eta_{1\text{st}}$，我们可以改变 LOLA-OffPA2 的收敛行为。此外，图7表明，较低的投影预测长度值（$\hat{}\eta_{1\text{st}} = 0.1$）会抵消 OffPA2 中学习预期的效果，而较高的投影预测长度值（$\hat{}\eta_{1\text{st}} = 1.3$）会导致 LOLA-OffPA2 出现不稳定性，这可以归因于我们在定理 2 中的研究结果。

5.3 HLA-OffPA2 的评估

5.3.1 Particle-coordination game

为展示 HLA-OffPA2 的协调能力，我们在粒子环境中提出了粒子协同博弈（PCG）（Lowe 等，2017）。如图8a所示，两个智能体（紫色圆圈）中的每一个都应选择并靠近三个地标（一个灰色圆圈和两个绿色圆圈）中的一个。根据智能体与地标之间的最近距离来选择地标。假设智能体选择并靠近同一个地标。在这种情况下，它们会获得

	绿色（L）	灰色	绿方（右侧）
绿色（L）	2	0	-20
Gray	0	0.4	0
绿方（右侧）	-20	0	2

表6：粒子协同博弈中的奖励。

全局（选择绿色地标）或局部（选择灰色地标）最优奖励。如果它们选择并接近不同的地标，就会受到指定的不协调惩罚（见表 6）。每个智能体都会接收一个 10 维状态观测向量（智能体的速度和位置信息，即 4 维，以及地标的位置坐标信息，即 6 维），并选择一个 5 维独热向量，表示五种离散动作之一：向右移动、向左移动、向上移动、向下移动和停留。时域设为 25。

	粒子环境			MuJoCo 环境		
	协作导航	物理欺骗	捕食者-猎物	半猎豹	行走者	伸手者
观测	18-D	10-D（8-D）	14-D (12-D)	11-D	11-D	8-D
动作	5-D	5-D (5-D)	5-D (5-D)	3-D	3-D	1-D
动作类型	离散的	离散的	离散的	连续的	连续的	连续
时域（步）	25	25	25	100	300	50

表7：标准多智能体博弈中的规格。在混合环境中，维度以“d_1 (d_2)”的形式报告，其中 d_1 是共同利益智能体的维度， d_2 是自身利益智能体的维度。

	↑粒子环境中的 NAER			↑MuJoCo 环境中的 NAER		
方法	协作导航	物理欺骗	捕食者-猎物	半猎豹	行走者	伸手者
深度确定性策略梯度（LB）	0.00	0.00	0.00	0.00	0.00	0.00
C-MADDPG（UB）	1.00	1.00	1.00	1.00	1.00	1.00
MADDPG	0.77	0.61	0.21	0.86	0.45	0.02
CPG-MADDPG	0.78	0.67	0.18	0.88	0.46	0.05
PR2-MADDPG	0.78	0.54	0.08	0.85	0.45	0.01
HLA-OffPA2（我们的方法）	**0.88**	**0.83**	**0.44**	**0.94**	**0.67**	**0.42**

表 8：共同利益智能体的归一化平均回合奖励（NAER）方法比较。LB：下界。UB：上界。

由于智能体无法看到彼此的位置，因此该博弈颇具挑战性，而且它们可能发生协调失误。

在图8b中，我们展示了所提出的 HLA-OffPA2 与其他基于 MADDPG 的方法在平均回合奖励（AER）方面的学习曲线。如图所示，基于 MADDPG 的方法在收敛点上的方差相对较大。例如，MADDPG 智能体在策略参数更新过程中会大量受益于探索和随机性，因此偶尔能够收敛到 AER 最高的全局最优点。然而，较高的不协调惩罚会迫使智能体选择最安全的选项（灰色地标），导致最坏情况下奖励为零。从图中可以清楚看出，只有我们提出的 HLA-OffPA2 能够始终收敛到该博弈的全局最优解，这与已报道的 HLA 在完全合作型可微博弈中的结果一致（Bighashdel 等，2023年）。

5.3.2 Standard multi-agent games

在最终实验中，我们在三个粒子环境博弈中评估 HLA-OffPA2 和基于 MADDPG 的方法（Lowe 等，2017年）：1）由三个共同利益智能体参与的协作导航；2）由两个共同利益智能体和一个自利型智能体参与的物理欺骗；3）由两个共同利益（捕食者）智能体和一个自利型（猎物）智能体参与的捕食者-猎物。此外，我们还在多智能体 MuJoCo 环境中的三个博弈中比较这些方法（Peng 等，2021年）：1）双智能体半猎豹；2）双智能体行走者；3）双智能体伸手者。在混合环境（物理欺骗和捕食者-猎物）中，所有实验均使用 MADDPG 方法训练自利型智能体。博弈的具体设置见表 7。

	↑粒子环境中的 NAER			↑NAER 在 MuJoCo 环境中		
方法	协作导航	物理欺骗	捕食者-猎物	半猎豹	行走者	伸手者
HLA-OffPA2 (F)	0.85	0.80	0.38	0.92	0.63	0.38
HLA-OffPA2	**0.88**	**0.83**	**0.44**	**0.94**	**0.67**	**0.42**

表9：对我们提出的 HLA-OffPA2 方法中的层级分配进行消融研究。

我们为每个博弈分别创建了验证集和测试集，其中分别包含 100 个和 300 个随机生成的场景。在每个博弈中，我们保存验证集性能最佳的模型，并在测试集上进行测试以报告结果。所有实验均重复五次，并以归一化 AER（NAER）为指标，将均值结果报告于表 8。归一化基准设定为 MADDPG 的单智能体变体，即 DDPG（Lillicrap 等人，2016年），以及完全集中式（在学习和执行阶段均如此）的 MADDPG 变体，称为 C-MADDPG。由于所有博弈均不可微，原始 HLA 方法不再适用。

在表 8 中，我们观察到，所提出的 HLA-OffPA2 始终显著优于所有最先进的基于 MADDPG 的方法。这些结果再次证实，学习预期，尤其是我们提出的 HLA-OffPA2，能够改善共同利益智能体之间的协调，从而取得更好的结果。

5.3.3 Ablation study on hierarchy-level assignments

我们还对 HLA-OffPA2 中的层级分配进行了消融研究。我们没有通过公式 (20) 迭代地依据智能体的塑造能力对其进行排序，而是在开始时随机将智能体分配到各个层级，并在整个优化过程中固定层级。我们对这一变体进行了评估，并在表 9 中与其他方法进行了比较；该变体称为 HLA-OffPA2 (F)。结果表明，像 HLA-OffPA2 中所采用的那样，使用基于智能体塑造能力的排序策略，能够持续提升性能。

6. 结论

本文提出了 OffPA2 框架，使 HOG 方法能够应用于具有大状态空间的不可微博弈。为展示该框架的优势，我们开发了三种新颖的 HOG 方法：LA-OffPA2、LOLA-OffPA2 和 HLA-OffPA2。通过开展多项实验，我们证明，所提出的方法在性能和效率方面优于现有的 HOG 方法。此外，我们还将这些方法与多种不采用学习预期的基于 DPG 的方法进行了广泛比较，并表明学习预期能够改善智能体之间的协调并带来更高的奖励。因此，得益于我们的框架，学习预期的优势如今可以应用于更多的多智能体强化学习问题。

附录

A. 证明

A.1 定理 1 的证明

在每个状态 $s \sim \rho^b(s)$ 下，智能体一会预期智能体二的策略参数发生变化，即 $\Delta\theta_2(s) = \eta\nabla_{\theta_2}Q_2(s, \mu_{\theta_1}(s)\ \mu_{\theta_2}(s))$，并沿以下方向更新策略参数：

$$\nabla_{\theta_1}J_1 = \mathbb{E}_{\rho^{\beta}(s)}\nabla_{\theta_1}Q_1(s, \mu_{\theta_1}, \mu_{\theta_2+\Delta\theta_2(s)}(s)), \tag{27}$$

为了证明命题 1，我们需要两个假设：1）忽略状态-动作价值函数对策略参数的直接依赖；2）采用一阶泰勒展开。第一个假设是离策略强化学习中的标准做法，适用于确定性和随机性离策略演员-评论家算法（Silver 等，2014年；Degris 等人，2012年），Degris 等人（2012年）提供了支持这一假设的论证。关于第二个假设，请参阅定理 2，了解该假设如何影响性能。

根据第一个假设，我们可以将式（27）改写为

$$\nabla_{\theta_1}J_1 = \mathbb{E}_{\rho^{\beta}(s)}\nabla_{\theta_1}\mu_{\theta_1}(s)Q_1(s, a_1, \tilde{a}_2)|_{a_1=\mu_{\theta_1}(s), \tilde{a}_2=\mu_{\theta_2+\Delta\theta_2(s)}(s)}, \tag{28}$$

类似地，我们可以设定 $\Delta\theta_2(s) = \eta\nabla_{\theta_2}\mu_{\theta_2}(s)\nabla_{a_2}Q_2(s, a_1, a_2)|_{a_2=\mu_{\theta_2}(s)}$。现在，我们使用一阶泰勒展开，将预期梯度信息映射到动作空间：

$$\begin{aligned}\tilde{a}_2 &= \mu_{\theta_2+\Delta\theta_2(s)}(s) \\ &\approx \mu_{\theta_2}(s) + (\Delta\theta_2(s))^{\intercal}\nabla_{\theta_2}\mu_{\theta_2}(s),\end{aligned} \tag{29}$$

根据 $\Delta\theta_2(s)$的定义，我们有：

$$\begin{aligned}\tilde{a}_2 &\approx \mu_{\theta_2}(s) + (\eta\nabla_{\theta_2}\mu_{\theta_2}(s)\left(\nabla_{d_2}Q_2(s, a_1, a_2)\right)^{\intercal})^{\intercal}\nabla_{\theta_2}\mu_{\theta_2}(s) \\ &= \mu_{\theta_2}(s) + \nabla_{a_2}Q_2(s, a_1, a_2)\left(\eta\nabla_{\theta_2}\mu_{\theta_2}(s)\right)^{\intercal}\nabla_{\theta_2}\mu_{\theta_2}(s) \\ &= \mu_{\theta_2}(s) + \nabla_{a_2}Q_2(s, a_1, a_2)\eta\left\|\nabla_{\theta_2}\mu_{\theta_2(s)}\right\|^2 \\ &= \mu_{\theta_2}(s) + \nabla_{a_2}Q_2(s, a_1, a_2)\hat{\eta}_{1\mathrm{st}},\end{aligned} \tag{30}$$

其中，$\|.\|$ 是 l^2-范数，并且我们定义了投影预测长度$\hat{}\eta_{1\mathrm{st}} = \eta\left\|\nabla_{\theta_2}\mu_{\theta_2(s)}\right\|^2$，因为$\left\|\nabla_{\theta_2}\mu_{\theta_2(s)}\right\|^2$ 是一个正数，且与 θ_1无关。因此：

$$\tilde{a}_2 \approx a_2 + \Delta a_2. \tag{31}$$

其中，我们定义了 $\Delta a_2 = \hat{}\eta_{1\mathrm{st}}\nabla_{a_2}Q_2(s, a_1, a_2)$。将式31代入式27，得到：

$$\nabla_{\theta_1}J_1^{\mathrm{LA}} \approx \mathbb{E}_{\rho^{\beta}(s)}\nabla_{\theta_1}\mu_{\theta_1}(s)\nabla_{a_1}Q_1(s, a_1, a_2 + \Delta a_2)|_{a_1=\mu_{\theta_1}(s), a_2=\mu_{\theta_2}(s)}, \tag{32}$$

由此，定理 1 得证。

A.2 定理 2 的证明

为了证明定理 2，我们首先需要证明：

引理 4 如果使用全阶泰勒展开将预期变化从策略参数空间映射到动作空间，则存在 $\hat{}$ $\eta_{full} \in \mathbb{R}$，使得

$$\mu_{\theta_2+\Delta\theta_2}(s) = \mu_{\theta_2}(s) + \hat{\eta}_{full}\nabla_{a_2}Q_2(s,a_1,a_2), \tag{33}$$

其中

$$\begin{aligned} &\Delta\theta_2 = \eta\nabla_{\theta_2}\mu_{\theta_2}(s)\nabla_{a_2}Q_2(s,a_1,a_2) \\ &\eta \in \mathbb{R}^+ \end{aligned} \tag{34}$$

证明。对预期梯度进行全阶泰勒展开，得到：

$$\mu_{\theta_2+\Delta\theta_2}(s) = \mu_{\theta_2}(s) + (\Delta\theta_2)^\intercal\nabla_{\theta_2}\mu_{\theta_2}(s) + \frac{1}{2}(\Delta\theta_2)^\intercal H_{\mu_{\theta_2}}(s)\Delta\theta_2 + O(\|\Delta\theta_2\|^3), \tag{35}$$

其中，$H_{\mu_{\theta_2}}(s)$ 表示 μ_{θ_2} 在 s 处的海森矩阵。已知：

$$\Delta\theta_2 = (\eta\nabla_{\theta_2}\mu_{\theta_2}(s)\nabla_{a_2}Q_2(s,a_1,a_2))^\intercal, \tag{36}$$

我们有

$$\begin{aligned} \mu_{\theta_2+\Delta\theta_2}(s) =& \mu_{\theta_2}(s) + \nabla_{a_2}Q_2(s,a_1,a_2)\eta\left\|\nabla_{\theta_2}\mu_{\theta_2(s)}\right\|^2 \\ &+ \frac{1}{2}\nabla_{a_2}Q_2(s,a_1,a_2)\eta^2\left(\nabla_{\theta_2}\mu_{\theta_2}(s)\right)^\intercal H_{\mu_{\theta_2}}(s)\nabla_{\theta_2}\mu_{\theta_2}(s)\left(\nabla_{a_2}Q_2(s,a_1,a_2)\right)^\intercal \\ &+ O(\eta^3). \end{aligned} \tag{37}$$

通过定义

$$\begin{aligned} &C1(s) = \left\|\nabla_{\theta_2}\mu_{\theta_2(s)}\right\|^2 \\ &C_2(s) = \frac{1}{2}\left(\nabla_{\theta_2}\mu_{\theta_2}(s)\right)^\intercal H_{\mu_{\theta_2}}(s)\nabla_{\theta_2}\mu_{\theta_2}(s)\left(\nabla_{a_2}Q_2(s,a_1,a_2)\right)^\intercal, \end{aligned} \tag{38}$$

我们有：

$$\mu_{\theta_2+\Delta\theta_2}(s) = \mu_{\theta_2}(s) + \nabla_{a_2}Q_2(s,a_1,a_2)(\eta C_1(s) + \eta^2 C_2(s) + O(\eta^3)), \tag{39}$$

根据 $C_2(s)$ 的定义以及 $\nabla_{a_2}Q_2(s,a_1,a_2)$ 所隐含的维度约束，可以得出 $C_1(s) \in \mathbb{R}^+$ 和 $C_2(s) \in \mathbb{R}$。因此：

$$\mu_{\theta_2+\Delta\theta_2}(s) = \mu_{\theta_2}(s) + \hat{\eta}_{\text{full}}\nabla_{a_2}Q_2(s,a_1,a_2), \tag{40}$$

其中 $\hat{}\eta_{\text{full}} = \eta C_1(s) + \eta^2 C_2(s) + O(\eta^3) \in \mathbb{R}$。因此，我们证明了引理 4

现在，若我们使用全阶泰勒展开将预测长度为$\eta' \in \mathbb{R}^+$的策略参数预期变化映射到动作空间，则有：

$$\mu_{\theta_2+\Delta\theta'_2}(s) = \mu_{\theta_2}(s) + \hat{\eta'}_{\text{full}}\nabla_{a_2}Q_2(s,a_1,a_2), \tag{41}$$

其中 $\hat{\eta'}_{\text{full}} = \eta' C_1(s) + \eta'^2 C_2(s) + O(\eta'^3)$。为了证明定理 2，我们需要求出满足以下条件的 $\hat{\eta}_{1\text{st}}$ 的取值：

$$\begin{aligned} \hat{\eta}_{1\text{st}} &= \hat{\eta'}_{\text{full}} \\ &= \eta' C_1(s) + \eta'^2 C_2(s) + O(\eta'^3), \end{aligned} \tag{42}$$

同时也有 $\eta' \in \mathbb{R}^+$。忽略 $O(\eta'^3)$，并且已知 $\hat{\eta}_{1\text{st}} \in \mathbb{R}^+$，需要考虑两种情况：

- 若 $C_2(s)$ 非负，则对于 $\hat{\eta}_{1\text{st}} \in \mathbb{R}^+$ 的任意取值，都存在 $\eta \in \mathbb{R}^+$。
- 若 $C_2(s)$ 为负，则对于 $\hat{\eta}_{1\text{st}} < \frac{C_1(\quad s^2\ )()}{4\lvert C_2\ s\ \rvert}$，存在 $\eta \in \mathbb{R}^+$。

因此，当 $\hat{\eta}_{1\text{st}}$ 足够小时，即 $\hat{\eta}_{1\text{st}} < \frac{C_1(\quad s^2\ )()}{4\lvert C_2\ s\ \rvert}$，总存在 $\eta \in \mathbb{R}^+$，从而定理 2 得证。

A.3 定理 3 的证明

由于策略和状态-动作价值函数均采用神经网络进行近似，梯度预判的时间复杂度与神经网络中反向传播的时间复杂度相同。假设策略网络和状态-动作价值网络具有相同数量的隐藏层 H，且每个隐藏层中的神经元数量均为 N。因此，对于输入状态大小为 N_s 、动作大小为 N_a 的输入，网络反向传播的时间复杂度为（Lister and Stone，1995）：

- 策略网络中的反向传播时间复杂度： $O(N_s N + (H-1)N^2 + N N_a)$
- 状态-动作价值网络中的反向传播时间复杂度： $O((N_s + N_a)N + (H-1)N^2 + N)$

给定 $N > N_s + N_a$，两个网络的时间复杂度都可以上界为 $O(LN^2)$，其中我们定义了 $L = H + 1$。现在假设智能体 $i \in \mathcal{N}$ 想要预期另一个智能体 $j \in \{\mathcal{N} - \{i\}\}$的学习步骤。在策略参数预期的情况下，智能体 i 预期 $\Delta\theta_j(s)$如下：

$$\Delta\theta_j(s) = \eta \nabla_{\theta_j} \mu_{\theta_j}(s) \nabla_{a_j} Q_j(s, a_1, ..., a_n)|_{a_j = \mu_{\theta_j}(s)}. \tag{43}$$

因此，时间复杂度为 $O(LN^2) \times O(LN^2)$，换言之，即为 $O(L^2N^4)$。另一方面，在动作预期的情况下，智能体 i 预期 $\Delta a_j(s)$如下：

$$\Delta a_j = \hat{\eta}_{1st} \nabla_{a_j} Q_2(s, a_1, ..., a_n), \tag{44}$$

其复杂度为 $O(LN^2)$。因此，时间复杂度降低了$O(LN^2)$，定理 3 得证。

算法 1： 适用于一组 n 自利型智能体（$\mathcal{N}$）的 LOLA-OffPA2。

初始化 μ_{θ_i}、 Q_i、 μ_i' 和 Q_i' $\forall i \in \mathcal{N}$，并设置 $\hat{}\eta_{1\text{st}}$
适用于回合 $= 1$ 直到 max-num-episodes **为止**
接收初始状态 s
对于 $t = 1$ 直到最大回合长度 **执行**
选择动作 a_i，来自 $\pi^b_{\theta_i}(s)$ $\forall i \in \mathcal{N}$
执行动作 $a = \{a_i\}_{\forall i \in \mathcal{N}}$ 并观察奖励 $r = \{r_i\}_{\forall i \in \mathcal{N}}$ 和新状态 s'
将元组（s， a， r， s'）存储在经验回放缓冲区 $\mathcal{D}$ 中
设置 $s = s'$

从 $\mathcal{D}$ 中采样一个随机元组 K $\left(s^k,\ a^k,\ r^k,\ s'^k\right)_{k\in\{1,\ldots,K\}}$
对于 智能体 $i = 1$ 到 n **执行**
集合 $y_i^k = r_i^k + \gamma Q_i'\left(s'^k, a_1', ..., a_n'\right)\big|_{a_h'=\mu_h'(s'^k)}$，对于 $k \in \{1, ..., K\}$
通过最小化以下内容来更新状态-动作价值函数 Q_i：

$$\mathcal{L}(\omega_i) = \frac{1}{K}\sum_{k\in\{1,\ldots,K\}}\left[\left(Q_i(s^k, a_1^k, ..., a_n^k;\ \omega_i) - y_i^k\right)^2\right]$$

结束循环。
令 $a_i^k = \mu_{\theta_i}(s^k)$，对于 $k \in \{1, ..., K\}$ 和 $i \in \mathcal{N}$
对于 智能体 $i = 1$ 到 n **执行**
对 智能体 $j = 1$ 至 n **执行**
如果 $j = i$，**则** 继续
集合 $\Delta a_j^k = \hat{}\eta_{1\text{st}} \frac{\partial}{\partial a_j^k} Q_j(s^k,\ a_1^k,\ .....\ a_n^k)$，对于 $k \in \{1, ..., K\}$
结束循环
更新策略参数 θ_i 的方式：$\nabla_{\theta_i} J_i^{\text{LOLA-OffPA2}} =$

$$\frac{1}{K}\sum_{k\in\{1,\ldots,K\}} \nabla_{\theta_i}\mu_{\theta_i}(s^k)\ \frac{\partial}{\partial a_i^k} Q_i(s^k,\ a_1^k + \Delta a_1^k,\ .....,\ a_i^k,\ .....,\ a_n^k + \Delta a_n^k)$$

结束循环
更新 Q_i' 和 μ_i' $\forall i \in \mathcal{N}$
结束循环
结束循环

算法2： 针对一组 n 自利型智能体（$\mathcal{N}$）的 LA-OffPA2。

初始化 μ_{θ_i}、 Q_i、 μ_i' 和 Q_i' $\forall i \in \mathcal{N}$，并设置 $\hat{}\eta_{1\text{st}}$
对于回合 $= 1$ 到最大回合数 **执行**
接收初始状态 s
对于 $t = 1$，执行至最大回合长度 **时**
从 $\pi^b_{\theta_i}(s)$ $\forall i \in \mathcal{N}$ 中选择动作 a_i
执行动作 $a = \{a_i\}_{\forall i \in \mathcal{N}}$，并观察奖励 $r = \{r_i\}_{\forall i \in \mathcal{N}}$ 和新状态 s'
将元组 $(s,\ a,\ r,\ s')$ 存入经验回放缓冲区 $\mathcal{D}$
设置 $s = s'$

从 $\mathcal{D}$ 中随机采样 K 元组 $\{(s^k,\ a^k,\ r^k,\ s'^k)\}_{k\in\{1,\ldots,K\}}$
对于智能体 $i = 1$ 到 n **执行**
设置 $y_i^k = r_i^k + \gamma Q_i'\left(s'^k,\ a_1',\ ...,\ a_n'\right)\big|_{a_h'=\mu_h'(s'^k)}$，对于 $k \in \{1, ..., K\}$
通过最小化以下内容来更新状态-动作价值函数 Q_i：

B. 实现细节

在本节中，我们将详细介绍各方法的实现。为了对这些方法进行公平比较，我们在所有方法中使用了具有相同神经网络架构的策略和价值函数。算法 1、2 和 3 分别展示了 LOLA-OffPA2、LA-OffPA2 和 HLA-OffPA2 的优化框架。

关于部分可观测性的说明。到目前为止，我们将多智能体强化学习设定表述为一个马尔可夫博弈，并假设智能体能够访问状态空间。然而，在许多博弈中，智能体只能接收当前状态的私有状态观测。在这种情况下，多智能体强化学习设定可以表述为一个部分可观测马尔可夫博弈（PO-MG）（Littman，1994）。一个 PO-MG 是一个元组 $(\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \{\mathcal{O}_i\}_{i\in\mathcal{N}}, \{\mathcal{R}_i\}_{i\in\mathcal{N}}, \mathcal{T}, \{\Omega_i\}_{i\in\mathcal{N}}, \rho, \gamma)$，其中 $\mathcal{O}_i$ 是智能体 $i \in \mathcal{N}$的状态观测集合。每个智能体 i 通过策略 $\mu_{\theta_i} : \mathcal{O}_i \to \mathcal{A}_i$ 选择其动作$a_i \in \mathcal{A}_i$ ，该策略由 θ_i 参数化，并以给定的状态观测 $o_i \in \mathcal{O}_i$为条件。转移到新状态后，每个智能体 i 通过其观测函数 Ω_i : $\mathcal{S} \to \mathcal{O}_i$接收私有状态观测。在这种情况下，每个智能体 i的集中式状态-动作价值函数定义为 $Q_i(o_1, ..., o_n, a_1, ..., a_n) = \mathbb{E}[G_i^t(\tau | s^t = s, o_i = \Omega_i(s) \ \& \ a_i^t = a_i \ \forall i \in \mathcal{N})]$。因此，可以相应地修改所提出的 OffPA2 框架。

B.1 迭代旋转博弈与重复囚徒困境

我们为策略和价值函数采用了具有两个隐藏层（维度为 64）的多层感知机（MLP）网络。为了使状态-动作价值函数可进行任意阶微分，我们在隐藏层之间使用了 SiLU 非线性函数（Elfwing等人，2018年）。对于 IRG，我们在策略中使用 Sigmoid 函数输出 1 维连续动作；对于 IPD，则使用 Gumbel-Softmax函数（Jang等人，2017年）输出两个离散动作。通过运行 Adam 优化器（Kingma和Ba，2015年），算法以固定学习率 0。01 进行训练，在 IRG 中训练 900 个回合，在 IPD 中训练 50 个回合。OffPA2 和 DiCE 框架中的（投影）预测长度经过调优，分别设为 0.8 和 0.3。所有实验均重复五次，结果以均值和标准差表示。

B.2 Exit-Room 游戏

策略网络和价值网络均由两部分组成：编码器和解码器。编码器是包含三个卷积层（ $12 \times 90 \times 90 \to 32 \times 21 \times 21 \to 64 \times 9 \times 9 \to 64 \times 7 \times 7$ ）和两个全连接层（ $3136 \to 512 \to 128$ ）的卷积神经网络，层间使用 SiLU 非线性函数（Elfwing等人，2018年）。解码器是多层感知机网络，包含两个隐藏层，维度为 64，用于策略和价值函数。我们在策略中使用 Gumbel-Softmax函数（Jang等人，2017年）输出离散动作。通过运行 Adam 优化器（Kingma和Ba，2015年），以固定学习率 0。01 训练算法 450 个回合（第一关）和 4500 个回合（第二关和第三关）。OffPA2 和 DiCE 框架中的（投影）预测长度经过调优，分别设为 1 和 0.4。所有实验重复五次，结果以均值和标准差表示。

	$\hat{\eta}_{1\text{st}}$ 在粒子环境中			$\hat{\eta}_{1\text{st}}$ 在 MuJoCo 环境中		
	协作导航	物理欺骗	捕食者-猎物	半猎豹	行走者	伸手者
HLA-OffPA2	0.003	0.01	0.04	0.004	0.004	0.007

表10：基于 OffPA2 的方法所优化的投影预测长度。

B.3 粒子协同博弈

我们采用包含两个隐藏层、维度为 64 的多层感知机网络，用于策略和状态-动作价值函数，并在层间使用 SiLU 非线性函数（Elfwing等人，2018年）。我们在策略中使用 Gumbel-Softmax函数（Jang等人，2017年）输出离散动作。通过运行 Adam 优化器（Kingma和Ba，2015年），以固定学习率 0。01 训练算法 100k 个回合。我们将 HLA-OffPA2 智能体的投影预测长度设为 0。1。所有实验重复五次，结果以均值和标准差表示。

B.4 标准多智能体博弈

与之前一样，所有方法均使用具有相同神经网络架构的策略和状态-动作价值函数。我们采用具有两个隐藏层的多层感知机网络（粒子环境中的维度为 64，MuJoCo 环境中的维度为 256 ），并在策略和状态 - 动作价值函数中使用 SiLU 非线性函数（ Elfwing 等人，2018 年）。在粒子环境中，我们在策略中使用 Gumbel-Softmax函数（Jang等人，2017年）输出离散动作，并通过运行 Adam 优化器（Kingma和Ba，2015年），以固定学习率 0。01 训练算法 100k 回合。在 MuJoCo 环境中，我们在策略中使用双曲正切函数输出连续动作，并通过运行 Adam 优化器（Kingma和Ba，2015年），以固定学习率 0。001 训练算法 10k 回合。HLA-OffPA2智能体在所有博弈中的投影预测长度均在 $0.001 - 0.1$ 之间进行优化。优化后的投影预测长度列于表10。

参考文献

Stefano V Albrecht 和 Peter Stone。自主智能体对其他智能体的建模：一项全面综述及开放问题。人工智能，258:66–95，2018年。

David Balduzzi、Sebastien Racaniere、James Martens、Jakob Foerster、Karl Tuyls 和 Thore Graepel。n 人可微博弈的机制。载于 国际机器学习会议，第 354–363 页。PMLR，2018年。

Dimitri P Bertsekas。约束优化和拉格朗日乘子法。学术出版社，2014年。

Ariyan Bighashdel、Daan de Geus、Pavol Jancura 和 Gijs Dubbelman。使用分层学习预期协调完全合作型智能体。*arXiv* 预印本 *arXiv:2303.08307*，2023年。

Thomas Degris、Martha White 和 Richard Sutton。离策略演员-评论家。载于国际机器学习会议，2012年。Stefan Elfwing、Eiji Uchibe 和 Kenji Doya。用于强化学习中神经网络函数逼近的 Sigmoid 加权线性单元。《神经网络》，107：3–11，2018年。Jakob Foerster、Richard Y Chen、Maruan Al-Shedivat、Shimon Whiteson、Pieter Abbeel 和 Igor Mordatch。具有对手学习意识的学习。载于第*17*届自主智能体与多智能体系统国际会议论文集，第122–130页，2018a。Jakob Foerster、Gregory Farquhar、Maruan Al-Shedivat、Tim Rockt¨aschel、Eric Xing 和 Shimon Whiteson。DiCE：无限可微的蒙特卡洛估计量。载于国际机器学习会议，第1529–1538页。PMLR，2018b。Jakob N Foerster、Yannis M Assael、Nando De Freitas 和 Shimon Whiteson。利用深度多智能体强化学习学习通信。*arXiv* 预印本 *arXiv:1605.06676*，2016年。Adam S Goodie、Prashant Doshi 和 Diana L Young。竞争性博弈中的心智理论推理层级。《行为决策研究》，25(1)：95–108，2012年。He He、Jordan Boyd-Graber、Kevin Kwok 和 Hal Daum´e III。深度强化学习中的对手建模。载于国际机器学习会议，第1804–1813页。PMLR，2016年。Zhang-Wei Hong、Shih-Yang Su、Tzu-Yun Shann、Yi-Hsiang Chang 和 Chun-Yi Lee。用于多智能体系统的深度策略推断Q网络。载于第*17*届自主智能体与多智能体系统国际会议论文集，第1388–1396页，2018年。Eric Jang、Shixiang Gu 和 Ben Poole。采用 Gumble-Softmax 的分类重参数化。载于国际学习表征会议（*ICLR 2017*）。OpenReview.net，2017年。Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。载于*ICLR*，2015年。Sachin Konan、Esmaeil Seraj 和 Matthew Gombolay。在合作式与拜占庭式去中心化组队中运用互信息进行迭代推理。*arXiv* 预印本 *arXiv:2201.08484*，2022年。Alistair Letcher、Jakob Foerster、David Balduzzi、Tim Rockt¨aschel 和 Shimon Whiteson。在可微博弈中的稳定对手塑形。载于国际学习表征会议，2019年。Timothy P Lillicrap、Jonathan J Hunt、Alexander Pritzel、Nicolas Heess、Tom Erez、Yuval Tassa、David Silver 和 Daan Wierstra。利用深度强化学习进行连续控制。载于*ICLR*（海报），2016年。

Raymond Lister 和 James V Stone。《共轭梯度反向传播中各种误差函数时间复杂度的实证研究》。载于*ICNN'95*——国际神经网络会议论文集，第1卷，第237–241页。IEEE，1995年。Michael L Littman。《作为多智能体强化学习框架的马尔可夫博弈》。载于《机器学习论文集*1994*》，第157–163页。爱思唯尔，1994年。Daxin Liu 和 Gerhard Lakemeyer。《通过回归在表达性概率行动逻辑中推理信念与元信念》。载于国际人工智能联合会议。IJCAI，2021年。Ryan Lowe、Yi Wu、Aviv Tamar、Jean Harb、Pieter Abbeel 和 Igor Mordatch。《混合合作竞争环境中的多智能体演员-评论家》。载于*NIPS*，2017年。Christopher Lu、Timon Willi、Christian A Schroeder De Witt 和 Jakob Foerster。《无模型对手塑形》。载于国际机器学习会议，第14398–14411页。PMLR，2022年。Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga 等。《PyTorch：一种命令式风格的高性能深度学习库》。《神经信息处理系统进展》，第32卷，2019年。Bei Peng、Tabish Rashid、Christian Schroeder de Witt、Pierre-Alexandre Kamienny、Philip Torr、Wendelin Boehmer 和 Shimon Whiteson。《FACMAC：分解式多智能体集中式策略梯度》。*NeurIPS*，2021年。David Silver、Guy Lever、Nicolas Heess、Thomas Degris、Daan Wierstra 和 Martin Riedmiller。《确定性策略梯度算法》。载于国际机器学习会议，第387–395页。PMLR，2014年。Richard S Sutton 和 Andrew G Barto。《强化学习导论》。麻省理工学院出版社，2018年。Eugene Vinitsky、Natasha Jaques、Joel Leibo、Antonio Castenada 和 Edward Hughes。《序贯社会困境博弈的开源实现》，2019年。GitHub代码库。Ying Wen、Yaodong Yang、Rui Luo、Jun Wang 和 Wei Pan。《面向多智能体强化学习的概率递归推理》。载于第七届国际学习表征会议，*ICLR 2019*，2019年。Ying Wen、Yaodong Yang 和 Jun Wang。《通过广义递归推理对多智能体交互中的有限理性进行建模》。载于*IJCAI*，2020年。Timon Willi、Alistair Hp Letcher、Johannes Treutlein 和 Jakob Foerster。《COLA：具有对手学习意识的一致性学习》。载于国际机器学习会议，第23804–23831页。PMLR，2022年。Chongjie Zhang 和 Victor Lesser。《基于策略预测的多智能体学习》。载于*AAAI*人工智能会议论文集，第24卷，2010年。