arXiv:2602.06599v1 [cs.MA] 6 Feb 2026

使用联合经验最佳响应的样本高效策略空间响应预言机

Ariyan Bighashdel 乌得勒支大学 荷兰乌得勒支 代尔夫特理工大学 荷兰代尔夫特 a.bighashdel@uu.nl

Thiago D. Simão 埃因霍温理工大学 荷兰埃因霍温 t.simao@tue.nl

Frans A. Oliehoek 代尔夫特理工大学 荷兰代尔夫特 f.a.oliehoek@tudelft.nl

摘要

多智能体强化学习（MARL）为精确的博弈论分析提供了一种可扩展的替代方案，但面临非平稳性问题，还需要维持多样化的策略群体，以捕捉非传递交互。策略空间响应预言机（PSRO）通过迭代扩展受限博弈，并加入近似最佳响应（BR）来应对这些问题；然而，逐智能体训练最佳响应在智能体众多或模拟器计算成本高的场景中开销过高。我们提出联合经验最佳响应（*JBR*），这是一种可直接用于 PSRO 的改进方法：它在当前元策略剖面下仅收集一次轨迹，并重复利用这一联合数据集，同时为所有智能体计算最佳响应。这种方法分摊了环境交互成本，并提高了最佳响应计算的样本效率。由于 JBR 将最佳响应计算转化为离线强化学习问题，我们提出三种应对分布偏移偏差的方法：（i）保守*JBR*，采用安全策略改进；（ii）探索增强型联合经验最佳响应，通过扰动数据收集过程实现，并具备理论保证；（iii）混合最佳响应，将 JBR 与定期独立最佳响应更新交替进行。在多个多智能体基准环境中，探索增强型联合经验最佳响应实现了最佳的准确性—效率权衡，而混合最佳响应以低得多的样本成本达到了接近 PSRO 的性能。总而言之，JBR 让 PSRO 更适用于大规模策略学习，同时保持了均衡鲁棒性。

关键词

多智能体强化学习、策略空间响应预言机、博弈论

ACM 引用格式：

Ariyan Bighashdel、Thiago D. Simão 和 Frans A. Oliehoek。2026年。《样本高效的策略空间响应预言机与联合经验最佳响应》。载于*第25届自主智能体与多智能体系统国际会议论文集（AAMAS 2026）*，帕福斯，塞浦路斯，*2026年5月25日至29日*，国际自主智能体与多智能体系统基金会，10页。
https://doi.org/10.65109/

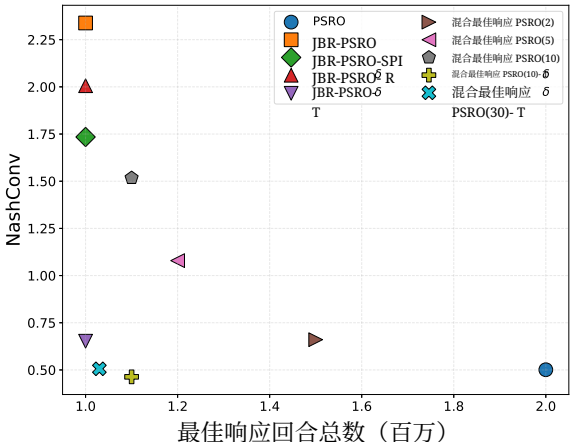

图1：Leduc扑克中的样本效率—准确性权衡。图中展示了最佳响应总回合数（百万，x轴）与100次迭代后的最小NashConv（y轴）。标准PSRO准确性高，但最佳响应样本成本最高。联合经验最佳响应（JBR-PSRO）及其增强变体——保守型（JBR-PSRO-SPI）、探索增强型（JBR-PSRO-δR、JBR-PSRO-δT）和混合型（HBR-PSRO(10/30)-δT）——大幅减少了收敛所需的最佳响应回合数；其中 JBR-PSRO-δT 实现了最佳权衡，而混合型版本的准确性接近 PSRO 水平。

1 引言

多智能体系统（MAS）是许多现实领域的基础，从自动驾驶交通管控到电子市场，智能体必须不断适应彼此的策略。博弈论为分析此类互动提供了坚实的基础，能够借助均衡概念进行研究；但由于策略空间呈组合式增长，在大型或序贯博弈中计算均衡在实践中不可行 [24, 31]。因此，需要可扩展的基于学习的方法。

多智能体强化学习（MARL）提供了这样一种替代方案，使智能体能够通过反复交互来改进其策略 [1, 26]。然而，多智能体强化学习面临两项关键挑战。首先，独立学习会导致非平稳性：随着每个智能体不断更新，其他智能体所面对的环境也在持续变化 [30]。集中训练、分散执行（CTDE） [8, 18]通过允许智能体在训练期间访问联合观测或梯度，同时在测试时独立行动，在一定程度上缓解了这一问题。然而，残余的不稳定性往往仍然存在。其次，许多博弈具有非传递性——不存在一种策略能够在全局上

本作品采用知识共享署名—国际 4.0 许可协议授权。

《第*25*届自主智能体与多智能体系统国际会议论文集》（*AAMAS 2026 29, 2026, 25* 2026），*C. Amato*、*L. Dennis*、*V. Mascardi*、*J. Thangarajah*（编），*5*月—塞浦路斯帕福斯。©自主智能体与多智能体系统国际基金会 (www.ifaamas.org)。https://doi.org/10.65109/

占据优势——而鲁棒性要求维持多样化的策略群体 [2, 3, 5, 23]。

为应对这些挑战，策略空间响应预言机（*PSRO*）框架 [12] 应运而生，它综合了多个研究领域的思想：借鉴规划领域，采用 Double Oracle 算法 [21]；借鉴协同进化领域，采用纳什记忆 [6] 和并行纳什记忆 [22]等机制；借鉴经验博弈论分析（*EGTA*）领域，纳入基于模拟的受限博弈 [25]；借鉴强化学习领域，利用基于强化学习的预言机，在高维序贯环境中训练最佳响应。PSRO 的核心是维护一个由部分策略构成的受限博弈，使用选定的求解器计算元策略（即受限博弈中各策略上的概率分布），并通过加入新的（近似）最佳响应来扩展受限博弈；这些响应是指针对当前元策略剖面（其他智能体策略的混合分布）最大化某个智能体收益的策略。通过重复这一过程，PSRO 逐步逼近均衡，同时保留多样化的策略群体，以捕捉非传递动态。整体流程总结于算法1。

尽管 PSRO 优势显著，但它存在一个主要局限：计算成本。在每次迭代中，每个智能体都必须独立计算最佳响应，这意味着需要解决与智能体数量相同的强化学习问题，而且在每种情况下都必须模拟其余智能体。在玩家众多或模拟器成本高昂的领域，这一过程很快就会变得不切实际。现有研究尝试通过并行化 [12, 20], 样本高效仿真 [28, 32], 或迁移学习 [16, 17, 27]来提高效率。虽然这些方法能缩短训练的实际耗时，但并未改变 PSRO 的基本结构：每次迭代都要为每个智能体分别进行最佳响应计算。

我们提出 联合经验最佳响应（*JBR*），这是一种可直接应用于 PSRO 的改进方法，能够在智能体之间复用联合经验。JBR 不再独立训练每个最佳响应，而是在当前元策略剖面下仅收集一次轨迹，并利用这一共享数据集为所有智能体计算最佳响应。这大幅提升了最佳响应计算的样本效率，使 PSRO 在多智能体或模拟器成本高昂的领域中也切实可行。

由于 JBR 复用的是在当前元策略下生成的数据，而非在线与环境交互，因此最佳响应计算转变为离线强化学习（离线 *RL*）问题：必须根据固定数据集优化策略，而不是通过新的环境滚动模拟进行优化。这种离线设定可能因分布偏移而引入偏差。我们提出三种互补的变体：*(i)* 保守*JBR*，通过实施安全策略改进 [13]，保证性能不会低于基线元策略；*(ii)* 探索增强型联合经验最佳响应，通过调整数据收集方式提升覆盖度，同时保留理论保证；以及*(iii)* 混合最佳响应，在 JBR 与定期执行的独立最佳响应之间交替，以兼顾效率和准确性（图 1）。

我们的研究结果表明，采用定向扰动的探索增强型JBR（**JBR-PSRO-δT**）仅需少量样本成本，即可达到接近PSRO均衡的准确性；而混合变体（**HBR-PSRO(k)-δT**）只需少量额外训练，便能达到PSRO级别的性能。在离散环境和连续环境中，PSRO式方法（包括我们的JBR变体）始终优于独立学习等标准多智能体强化学习范式以及集中训练、分散执行，这证实了在提高样本效率后，基于战略的训练仍然更加稳健。

我们的贡献主要有三点：(1) 我们提出联合经验最佳响应（*JBR*），这是对 PSRO 的一种高效改进，能够**通过在智能体之间复用共享轨迹，提高最佳响应计算的样本效率**。(2) 我们指出并解决了 JBR 的离线强化学习局限，并提出和评估了三种解决方案：保守JBR、探索增强型联合经验最佳响应（具有理论保证）以及混合最佳响应。(3) 我们开展了全面的实证研究，将 JBR 变体不仅与独立最佳响应（标准 PSRO）进行比较，还与主流多智能体强化学习基线（如独立学习和集中训练、分散执行）进行比较，结果表明，基于 PSRO 的方法始终表现出更强的战略鲁棒性。

2 相关工作

人们提出了多种 PSRO 扩展方法，以改进该框架的不同方面，包括可扩展性、探索和均衡计算。近期的一篇综述 [4] 回顾了这些研究方向。本文特别关注旨在提升 PSRO 效率 的方法。

并行化。 早期研究探索了通过并行训练最佳响应来缩短实际运行时间。深度认知层级（DCH） [12] 引入了一种层级调度，其中每个玩家都针对较低层级的均衡进行训练，从而有效地复用先前的计算结果。Pipeline PSRO（P2SRO） [20] 进一步完善了这一思路，通过协调多个并行训练的策略，并动态优化层级结构。尽管这些方法提高了吞吐量，但它们需要大量计算资源，且不会减少最佳响应的总体数量。

样本效率。 另一项研究方向旨在减少对高成本模拟的依赖。Smith 和 Wellman [28] 提出在运行 PSRO 的同时联合学习博弈的生成模型，从而能够查询所学模型，而非环境。Zhou 等人 [32]开发了 EPSRO，通过利用无需完整模拟即可推导最佳响应目标策略组合的情形，避免冗余的受限博弈评估。尽管这些方法减少了模拟器调用次数，但其效果在很大程度上取决于所学模型的质量，而重新训练最佳响应的成本并未改变。

迁移学习。第三种方法是通过在最佳响应之间复用知识来加速适应。NeuPL [16] 在单个网络中引入了跨策略的参数共享，而 Liu 等人 [15] 则针对采样混合策略进行训练，从而增强了对多样化对手策略组合的鲁棒性。随后，Liu 等人 [17] 将 NeuPL 与 JPSRO 相结合，将迁移收益扩展到相关均衡计算中。Smith 等人 [27]提出了一种混合预言机方法，通过混合针对纯策略预训练的价值函数来构建新的最佳响应。这些方法通过复用知识加速了学习，但其有效性取决于信息在不同对手之间迁移的效果，而且它们仍然一次只实例化一个新的最佳响应。

摘要。先前研究表明，通过并行化、高效模拟以及在最佳响应计算之间进行迁移，可以让 PSRO 更具实用性。尽管这些方法缩短了

训练的实际耗时，但它们基本上仍属于实现层面的优化：核心 PSRO 循环在每次迭代中仍会为每个智能体分别生成一个最佳响应。相比之下，我们提出的 JBR框架直接改变了这一结构，通过复用共享的交互经验，同时为所有智能体计算最佳响应，为提升可扩展性提供了另一条途径。

3 背景

PSRO 的核心思想是通过迭代构建一个较小的受限博弈，来近似原本难以处理的博弈；在这个受限博弈中，每个玩家只能使用部分策略。通过添加新的最佳响应（使用强化学习计算得出），逐步扩展受限博弈。这样，PSRO 在保持对完整博弈的可处理近似的同时，逐步改进策略覆盖范围。形式化地，考虑一个由以下要素定义的 n 人马尔可夫（随机）博弈：

- 一个状态空间 S，
- 每个玩家 $i \in N = \{1, \dots, n\}$ 各自的动作空间 A_i，联合动作空间 $\mathrm{A} = \times_{i=1}^{n} \mathrm{A}_i$，
- s 一个转移核 $P(\cdot \mid s, a_1, \dots, a_n)$，
- 玩家特定的奖励函数 $r_i(s, a_1, \dots, a_n)$，
- 一个折扣因子 $\gamma \in [0, 1)$，
- 以及一个初始状态分布 ρ_0。

玩家 i 的策略是一个策略 $\pi_i \in \Pi_i$，即将状态映射到 A_i 上分布的函数。给定一个策略组合 $\pi = (\pi_1, \dots, \pi_n)$，玩家 i 的效用是每步奖励的期望折扣回报：

$$u_i(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \, r_i(s_t, a_{1,t}, \dots, a_{n,t}) \;\middle|\; \pi\right].$$

因此，PSRO 可以理解为在由马尔可夫博弈诱导的标准形博弈 $\mathsf{G} = (N, (\Pi_i), (u_i))$ 上运行，其中纯策略是完整策略，收益则是它们的期望回报。实际上，这个诱导博弈难以处理，因为每个 Π_i 都极其庞大。因此，PSRO 会维护受限策略集 $X_i \subseteq \Pi_i$，其中包含迄今生成的策略，并定义受限经验博弈 $\mathsf{G} = (N, (X_i), (\hat{u}_i))$，其中 $\hat{u}_i$ 是通过对剖面 $\pi \in X = \times_i X_i$ 进行滚动模拟获得的收益估计。受限经验博弈是策略更新的基础，它是对完整博弈规模较小、基于模拟的近似，能够捕捉当前已知策略之间的相互作用。PSRO 算法在两个关键组成部分之间交替运行：

- **元策略求解器（MSS）：** 给定 G，M̂SS 会计算一个元策略剖面（混合策略）$\sigma = (\sigma_i)_{i \in N}$，其中每个 σ_i 都是一个关于 X_i 的概率分布，用于指定智能体当前策略集中各策略被采用的频率。常见选择包括纳什均衡求解器或遗憾最小化动态。
- **响应预言机（RO）：** 对于每个玩家 i，RO 会计算一个近似的最佳响应策略 $\pi_{\text{最佳响应}i}$ 以应对 σ_{-i}。形式化地，最佳响应算子为

$$br_i(\sigma_{-i}) = \arg\max_{\pi_i \in \Pi_i} u_i(\pi_i, \sigma_{-i}).$$

在实践中，响应预言机通过深度强化学习实现：智能体在诱导马尔可夫决策过程中接受训练，其中对手遵循 σ_{-i}，由此得到一个策略 π_i^{BR}，用于近似 $br_i(\sigma_{-i})$。随后将新策略添加到 X_i，从而扩展受限博弈。

算法1策略空间响应预言机（PSRO）

要求： 初始受限策略集（X_i）

```
估计 G = (N, (X_i), (û_i))
初始化 σ ← MSS(G)
当未终止时执行
    对于玩家 i ∈ N 执行
        计算 π_i^BR ← 响应预言机(σ_{-i})
        更新 X_i ← X_i ∪ {π_i^BR}
    结束循环
    重新估计 Ĝ
    更新 σ ← MSS(Ĝ)
结束循环
返回 σ
```

完整流程如算法1所示。该过程会持续进行，直到满足停止准则，通常依据遗憾值来判定。玩家 i 在元策略剖面 σ 下的遗憾值为

$$\rho_i^{\mathcal{G}}(\sigma) = \max_{\pi_i \in \Pi_i} u_i(\pi_i, \sigma_{-i}) - u_i(\sigma_i, \sigma_{-i}), \tag{1}$$

它衡量单方面偏离所能带来的潜在收益。在纳什均衡下，所有遗憾值均为零。遗憾值总和

$$\rho^{\mathcal{G}}(\sigma) = \sum_{i \in N} \rho_i^{\mathcal{G}}(\sigma), \tag{2}$$

称为*NashConv* [12]，在双人博弈中对应可剥削性。该量广泛用于跟踪PSRO实验中的收敛情况。实际上，只有在可以进行穷举最佳响应搜索的小型离散博弈中，才能精确计算该量；在较大规模或连续环境中，则必须使用学习得到的最佳响应进行近似（见第5节）。

4 方法

在本节中，我们采用统一的视角：通过基于模型的强化学习，具体而言是价值迭代，来求得最佳响应。此选择纯粹是为了清晰地阐释不同方法。在实践中，当无法进行精确的基于模型的计算时，通常会通过深度强化学习方法来实现响应预言机，这些方法近似求解相同的目标。因此，为保持一致性，所有公式均在这一基于模型的框架下给出；而在实验中，我们会根据环境采用基于模型或深度强化学习的实现。

4.1 诱导马尔可夫决策过程

在 PSRO 中，玩家 i 的最佳响应通过将多智能体随机博弈简化为单智能体马尔可夫决策过程来求得：我们固定对手的元策略剖面，将其行为纳入环境，并求解由此得到的诱导马尔可夫决策过程。

PSRO 假设博弈具有完美回忆，这意味着每个智能体都会记住迄今为止观察到和采取的完整状态与动作序列。这一假设确保在随时间做出决策时，不会遗忘任何相关信息。在完美回忆的条件下，Kuhn定理 [9] 指出，任何混合策略（即完整策略上的一个分布）都与某个实现等价

的行为策略等价，该行为策略会在每个状态指定一个可用动作上的分布。换言之，虽然与混合策略对弈在形式上会诱导出一个部分可观测马尔可夫决策过程（因为无法观察对手的策略索引），但完美回忆的假设意味着状态 s 已经包含了相关历史 h。因此，这两种表示是等价的，而且不失一般性，对手的元策略 σ_{-i} 可以表示为行为策略 $\sigma_{-i}(a_{-i} \mid s)$。[1]

形式上，固定 σ_{-i} 后，通过对对手的动作进行边缘化，可得到玩家 i 的有效转移函数与奖励函数：

$$P^{\sigma_{-i}}(s' \mid s, a_i) = \sum_{a_{-i} \in \mathcal{A}_{-i}} \sigma_{-i}(a_{-i} \mid s)\ P(s' \mid s, a_i, a_{-i}), \tag{3}$$

$$r_i^{\sigma_{-i}}(s, a_i) = \sum_{a_{-i} \in \mathcal{A}_{-i}} \sigma_{-i}(a_{-i} \mid s)\ r_i(s, a_i, a_{-i}). \tag{4}$$

（对于连续的A_{-i}，将求和替换为积分。）这定义了玩家 i的诱导单智能体马尔可夫决策过程：

$$\mathcal{M}_i^{\sigma_{-i}} \ = \ \big(\mathcal{S},\ \mathcal{A}_i,\ P^{\sigma_{-i}},\ r_i^{\sigma_{-i}},\ \gamma,\ \rho_0\big).$$

4.2 独立最佳响应

令 $Q_i(s, a_i)$表示玩家 i在诱导马尔可夫决策过程$\mathrm{M}_i^{\sigma_{-i}}$中的动作价值函数。价值迭代通过反复应用贝尔曼最优性更新来计算 Q_i：

$$Q_i^{k+1}(s, a_i) \ = \ r_i^{\sigma_{-i}}(s, a_i) \ + \ \gamma \sum_{s' \in \mathcal{S}} P^{\sigma_{-i}}(s' \mid s, a_i) \ \max_{a_i' \in \mathcal{A}_i} Q_i^k(s', a_i').$$

随着 $k \to \infty$，该序列收敛到最优动作价值函数 $Q_i^\star$。相对于 $Q_i^\star$ 的贪心策略，

$$\pi_i^{\mathrm{BR}}(s) \ \in \ \arg\max_{a_i \in \mathcal{A}_i} Q_i^\star(s, a_i),$$

即为对 σ_{-i}的最佳响应。

4.3 联合经验最佳响应

虽然独立最佳响应（§4.2）是标准方法，但它有一个关键缺点：每个智能体的最佳响应都是单独训练的。这会造成大量冗余：训练一个智能体时产生的交互经验无法被其他智能体重复利用，导致环境模拟反复进行且成本高昂。在智能体众多或动态过程成本高昂的领域，按智能体分别训练很快就会变得不切实际。

为解决这种低效问题，我们提出了**联合经验最佳响应**（*JBR*）。其核心思想是，在当前元策略剖面 $\sigma = (\sigma_j)_{j \in N}$ 下仅收集一次轨迹，并重复利用这一联合数据集，同时为所有智能体计算最佳响应。在这种设置下，数据收集期间所有玩家都按照各自的元策略行动，因此数据集 D^σ 由以下轨迹构成

$$\mathcal{D}^\sigma = \{(s_t, a_{1,t}, \ldots, a_{n,t}, r_{1,t}, \ldots, r_{n,t}, s_{t+1})\}_{t=0}^{T}$$

在执行联合策略剖面 σ时由马尔可夫博弈生成。整体流程总结于算法2中。

[1]备注：如果将对手的元策略解释为完整策略上的分布，并在每个回合开始时采样一个策略，那么从玩家 i的角度来看，对手策略索引就像一个隐变量，从而诱导出一个部分可观测马尔可夫决策过程。对此有两种标准处理方式：（i）将状态空间 $\mathcal{S}$ 表示为完整历史的集合，使该过程成为马尔可夫过程；或（ii）假设具有完美回忆，根据Kuhn定理，这可以保证每个混合策略都与某个行为策略实现等价。在这两种情况下，我们都可以将元策略 σ_{-i} 视为行为策略 $\sigma_{-i}(a_{-i} \mid s)$，从而得到诱导马尔可夫决策过程。

算法2采用联合经验最佳响应的 PSRO

要求： 初始受限策略集 (X_i)

ˆ 1: 估计 G $= (N,\ (X_i),\ (\hat{u}_i))$

ˆ 2: 初始化 $\sigma \leftarrow$ MSS(G)

3: **当**未终止 **时执行**

4: 在元策略 σ下收集联合数据集 D^σ

5: **对**玩家 $i \in N$ **执行**

6: 计算 $\pi_i^{\mathrm{BR}} \leftarrow$ 响应预言机($\mathrm{D}^{\sigma_{-i}}$)

7: 更新 $X_i \leftarrow X_i \cup \{\pi_i^{\mathrm{BR}}\}$

8: **结束循环**

9: ˆ重新估计 G

10: 更新 $\sigma \leftarrow$ MSS($\hat{\mathcal{G}}$)

11: **结束循环**

12: **返回** σ

4.4 朴素 JBR

最直接的方法是将 D^σ 视为固定的离线数据集，然后针对每个智能体 i，按照 §4.1 构造其诱导马尔可夫决策过程 $\mathrm{M}_i^{\sigma_{-i}}$。随后，在 $\mathrm{M}_i^{\sigma_{-i}}$上应用基于模型的价值迭代来求取最佳响应，并使用 D^σ 估计状态转移动态 $P^{\sigma_{-i}}$ 和奖励$r_i^{\sigma_{-i}}$。形式上， Q迭代更新为

$$Q_i^{k+1}(s, a_i) \ = \ \hat{r}_i^{\sigma_{-i}}(s, a_i) \ + \ \gamma \sum_{s' \in \mathcal{S}} \hat{P}^{\sigma_{-i}}(s' \mid s, a_i) \ \max_{a_i'} Q_i^k(s', a_i'),$$

其中，$\hat{r}_i^{\sigma_{-i}}$和$\hat{P}^{\sigma_{-i}}$分别是根据D^σ估计的奖励模型和转移模型。随后，将得到的贪心策略 π_i^{BR}加入受限博弈，与标准PSRO完全相同。

尽管概念上简单，这种朴素 *JBR* 仍存在一个关键局限：由于数据集D^σ 是在当前元策略下生成的，因此可能无法覆盖计算真实最佳响应所需的状态–动作区域。这种分布偏移使该问题成为离线强化学习问题，可能导致对最佳响应的近似效果不佳。为解决这些局限，我们提出三种互补的补救方法。第一种是基于安全策略改进（§4.5）的保守型保障机制，可确保策略的表现不会低于当前元策略。第二种是基于探索的方法（§4.6），通过在数据收集期间扰动元策略来提高覆盖度，并具有理论保证。最后，混合变体（§4.7）在联合与独立最佳响应之间交替，以平衡效率和准确性。这三种方法共同缓解了朴素 JBR 的不足。

4.5 保守JBR

为解决朴素JBR的离线强化学习偏差，我们采用安全策略改进 (SPI) [13]的原则。其核心思想是：仅允许学习到的策略在状态–动作空间中得到D^σ充分支持的区域偏离其基线 σ_i。在数据集覆盖度不足的区域，策略必须回退到 σ_i，从而保证其性能不会低于基线。

形式上，令 $N_{\mathcal{D}}(s,\ a_i)$ 表示（$s,\ a_i$）在 D^σ 中出现的次数，并固定一个阈值 $N_\wedge$。我们将不确定的状态–动作对集合定义为

$$B_i \ = \ \{(s, a_i) \ \mid \ N_{\mathcal{D}}(s, a_i) < N_\wedge\}.$$

SPI约束要求，对于所有（s， a_i） $\in B_i$，

$$\pi_i(a_i \mid s) = \sigma_i(a_i \mid s).$$

也就是说，在覆盖不足的区域，新策略会完全复现基线元策略 σ_i。在 B_i 之外，该策略通过对 $\mathrm{M}_i^{\sigma_{-i}}$ 执行离线价值迭代进行优化，并使用 D^{σ} 估计状态转移和奖励。

所得策略 π_i^{SPI}不一定是对σ_{-i}的最佳响应，但可以保证其表现不差于 σ_i（估计误差范围内），并且在 D^{σ} 提供充分覆盖的所有区域都能严格提升。因此，可以安全地将 π_i^{SPI} 加入 PSRO 的受限策略集。然而，正如我们在实验中所展示的，这种保守方法虽然优于朴素 JBR，但与采用独立最佳响应的原始 PSRO 相比，表现仍然不佳。这促使我们寻求更有效的解决方案。

4.6 探索增强型联合经验最佳响应

尽管保守JBR能够防止性能下降，但它仍然过于谨慎：改进仅限于数据集中已有充分体现的区域，因此难以达到独立最佳响应的表现。为克服这一问题，我们提出在数据收集期间，通过对元策略加入探索扰动，有意扩大覆盖度。

令 $\sigma = (\sigma_j)_{j \in N}$为当前元策略剖面。数据收集期间，我们使用一种 δ-扰动元策略，使智能体在相对于 σ的意义上大多遵循同策略，同时以概率 δ注入探索。对于每个智能体 j，定义逐状态混合行为

$$\tilde{\sigma}_j^{(\delta, \nu_j)}(a_j \mid s) \;=\; (1-\delta)\,\sigma_j(a_j \mid s) \;+\; \delta\,\nu_j(a_j \mid s), \qquad (5)$$

其中 ν_j是一种探索策略。智能体按照 $\tilde{\sigma}_j^{(\delta,\ \nu_j)}$ $= (\tilde{\sigma}^{(\delta,\ \nu_j)})_{j \in N}$独立行动，从而生成联合数据集 $\mathrm{D}^{\tilde{\sigma}}$。

我们考虑两种具体选择，由此得到两种探索策略：

4.6.1 δ-随机探索。令 $\nu_j^{\mathrm{rand}}(a_j \mid s) = \mathrm{Unif}(\mathcal{A}i)$，即

$$\tilde{\sigma}_j^{(\delta,\mathrm{rand})}(a_j \mid s) = (1-\delta)\,\sigma_j(a_j \mid s) \;+\; \delta\,\tfrac{1}{|\mathcal{A}_j|}.$$

这会在相对于 σ大多遵循同策略的同时，提供与状态无关的覆盖。

4.6.2 δ-定向探索。令 $\pi^{\mathrm{BR},\ j\ \mathrm{cur}}$ 表示智能体 j当前的最佳响应候选（即本轮 JBR 中正在训练的策略）。令 $\nu_j^{\mathrm{tgt}}(a_j \mid s)$ $= \pi^{\mathrm{BR},\ j\ \mathrm{cur}}(a_j \mid s)$，即

$$\tilde{\sigma}_j^{(\delta,\mathrm{tgt})}(a_j \mid s) = (1-\delta)\,\sigma_j(a_j \mid s) \;+\; \delta\,\pi_j^{\mathrm{BR,cur}}(a_j \mid s).$$

这会将额外样本集中在当前有望取得良好表现的最佳响应行为附近，从而在最佳响应计算所需之处改进价值估计。

4.6.3 理论影响。虽然探索扩大了覆盖范围并缓解了离线强化学习的偏差，但也改变了学习目标：智能体不再针对精确的元策略 σ进行最佳响应，而是针对其经过 δ 扰动后的变体 $\tilde{\sigma}$。这引出了一个问题：采用探索的 PSRO 是否仍能保留其理论保证?

我们现在分析这一影响。具体而言，我们证明，尽管探索改变了响应预言机的目标分布，但采用探索增强型 JBR 的 PSRO 仍能在有界误差范围内保留其收敛保证。我们关注在 $[u, u]$ 中且取值范围为 $R = u - u$ 的、收益有界的有限双人零和马尔可夫博弈。

定理4.1（探索增强型 JBR）。设 σ 为当前的元策略剖面，$\tilde{\sigma}$ 为其用于数据收集的 δ扰动变体。如果每个智能体都针对 $\tilde{\sigma}$ 计算一个 ε最佳响应，那么终止时得到的元策略将是原博弈的（$\varepsilon + 2R\delta$）-纳什均衡。

证明概要。收益关于混合策略是线性的，因此，对对手的元策略施加 δ扰动，任意策略的收益变化至多为 $2R\delta$。因此，针对 $\tilde{\sigma}$ 的最佳响应也是针对 σ的近似最佳响应，误差至多为该值。由于 PSRO 的终止检查是在 σ下进行的，因此不会遗漏任何有利可图的偏离，从而得到所述的（$\varepsilon + 2R\delta$）界。完整证明见附录A。

推论。 该结果表明，探索增强型 JBR 在受控误差范围内保留了 PSRO 的收敛保证：与均衡的偏差仅随 δ线性增长，而额外的探索则提高了覆盖度，并减少了离线强化学习偏差。不过需要注意的是，该保证仅在能够找到针对 $\tilde{\sigma}$ 的 ε-最佳响应这一假设下成立。在实践中，不保证能够从固定数据集中计算出这样的响应，因为覆盖度有限可能会妨碍可靠的策略改进。这促使我们采用多种探索策略，以丰富数据集并提高在实践中学到足够优良响应的可能性。

4.7 混合最优响应

虽然 JBR 大幅降低了仿真成本，但对离线数据的依赖会引入近似误差。保守型 JBR 和探索增强型 JBR 等变体能够缓解这些误差，但其准确性仍可能不及独立最优响应（IBR）；IBR 会在每次迭代中计算精确的最佳响应。这引出了一个问题：我们能否结合 JBR 的效率与 IBR 的准确性?

因此，我们考虑一种在两者之间交替切换的混合最佳响应方法。具体而言，在大多数迭代中使用 JBR（可能采用其某种变体），而每隔 k 次迭代便以标准的独立最优响应更新取而代之。形式上，令 t 表示迭代计数器。若 t 对 $k = 0$ 取模，则每个智能体 i通过针对 σ_{-i} 进行独立训练来计算最佳响应π_i^{IBR}；否则，使用 JBR（例如，朴素、保守型或探索增强型方法）获得响应。

这种混合型策略在大多数迭代中保留了 JBR 的样本效率优势，同时定期注入来自 IBR 的精确最佳响应，以纠正近似误差。在我们的实验中，我们评估这种定期纠正是否足以弥合基于 JBR 的变体与原始 PSRO 之间的性能差距。

5 实验

我们的实验旨在评估所提出的 *JBR*框架的可行性和可靠性 。按照设计，JBR 会计算所有智能体的最佳

响应，并且所有智能体都基于单个共享数据集同时计算，这显然能通过减少环境交互来提高样本效率。然而，核心问题在于，当数据在智能体之间重复使用时，JBR 是否仍能生成准确的最佳响应，从而保持 PSRO 的收敛行为。

5.1 实验设置

我们考虑两种经典的双人零和扑克游戏：库恩扑克 [10] 和*Leduc*扑克 [29]，以及随后介绍的三种连续多智能体粒子环境 [18]：*Simple Tag*、*Simple Adversary*和*Simple Push*。这些环境依次提高了学习可靠最佳响应的难度：库恩扑克规模较小，且被元策略完全覆盖；Leduc扑克则引入了不完全信息和更大的状态空间，从而暴露出离线学习偏差。粒子环境则进一步在连续控制和部分可观测性条件下对该方法进行压力测试；在这些环境中，联合数据复用可能会降低策略质量。

方法。 我们评估以下算法：

- **PSRO**：采用独立最佳响应的标准PSRO（基线）；
- **JBR-PSRO**：朴素的联合经验变体，利用共享数据计算所有最佳响应；
- **JBR-PSRO-SPI**：使用安全策略改进（SPI）的保守型 JBR。SPI 变体需要一个覆盖度阈值 $N_\wedge$，用于确定何时恢复为基线元策略。为隔离该方法的内在潜力，而非其对这一超参数的敏感性，我们采用经神谕调优的设置，在每次迭代时遍历 $N_\wedge \in [0, 50]$，并选择性能最佳的值。尽管这种做法并不实用，但它使我们能够评估 SPI 性能的理想化上界；
- **JBR-PSRO-δR**：采用 δ随机探索的探索增强型 JBR；
- **JBR-PSRO-δT**：采用 δ定向探索的探索增强型 JBR；
- **HBR-PSRO(k)**：每 k 次迭代在 JBR 和 IBR 之间交替的混合变体，也可以与探索结合使用（例如，**HBR-PSRO(k)-δT**）。

为了与标准多智能体强化学习范式进行比较，我们还纳入了**独立学习**（使用 DDPG [14] 的独立学习者）和**集中训练、分散执行**（使用 MADDPG [18] 的集中训练、分散执行）。

实现细节。 所有 PSRO 变体都使用相同的元策略求解器，即投影复制动态 [12]，，在每次迭代中计算经验收益矩阵的纳什均衡。在扑克领域中，最佳响应（BR）通过

通过基于模型的价值迭代，而在粒子环境中，我们采用以 DDPG 或 MADDPG 实现的深度强化学习预言机，具体取决于训练是独立进行还是集中进行。每次迭代都会为每个智能体添加一个新策略，并相应地更新元策略。尽管某些粒子环境涉及两个以上的独立智能体——例如*Simple Tag*（一名逃跑者对阵两名对手）和*Simple Adversary*（两名合作智能体对阵一名对手）——我们仍将它们视为双人博弈，把智能体分成两个对立的队伍：智能体 和对手。这种基于队伍的形式遵循*Team PSRO*框架 [19]，，确保

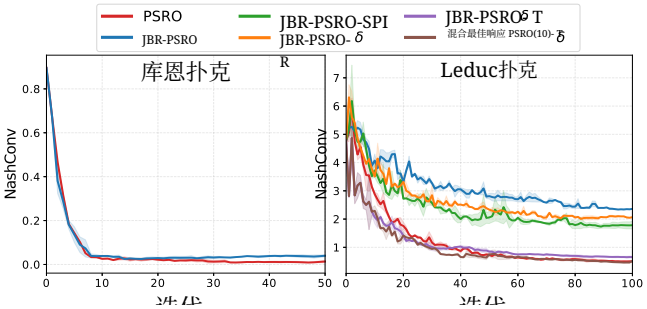

图2：PSRO 和 JBR 在两个复杂度逐渐增加的扑克游戏中的收敛情况。左：在 库恩扑克中，JBR 的表现与 PSRO 接近，表明当状态空间较小且覆盖充分时，JBR 是可行的。右：在 *Leduc*扑克中，随着复杂度增加，朴素 JBR 逐渐偏离 PSRO，揭示了促使我们提出 JBR 变体的离线学习偏差。

均衡分析在队伍层面仍然定义良好。在进行最佳响应计算时，如果队伍仅包含一名智能体，我们使用 DDPG；如果队伍包含多名合作智能体，则使用 MADDPG。

遵循先前工作 [12, 32]，，我们在扑克游戏中运行 PSRO 100 次迭代，并分配 10，000 个回合作为计算最佳响应的预算。在粒子环境中，我们运行 50 次迭代，每次迭代的最佳响应预算为 50，000 个回合。重要的是，在标准 PSRO 中，每个智能体分别获得这全部预算；而在 JBR 中，所有智能体共同分摊 同一预算，以联合计算其最佳响应，从而体现出训练效率的提升。除非另有说明，探索率 δ 设为 0。1（适用于**JBR- PSRO-δR**），以及 0。5（适用于**JBR-PSRO-δT**）；第 5.6 节报告了这些取值的敏感性分析。所有报告结果均为三个随机种子的平均值。

为评估收敛情况，我们跟踪两项指标：（i）*NashConv*，用于衡量均衡精度；（ii）最佳响应总回合数，用于量化整体样本效率。在库恩扑克和 Leduc扑克等小型离散博弈中，我们使用 OpenSpiel 的实现 [11]，精确计算 NashConv，其中通过完整树搜索获得真实的最佳响应。在连续粒子环境中，由于无法精确计算最佳响应，我们改为估计近似 *NashConv*：针对每个智能体，我们使用不同的随机初始化和扩大的最佳响应预算，训练三个最佳响应候选项，预算为 100,000 回合，并选择表现最佳的候选项来近似真实的最佳响应。

5.2 联合经验最佳响应是否可行?

我们首先研究 JBR 能否作为独立最佳响应的可靠替代方案。我们在**PSRO** 和**JBR-PSRO** 中评估库恩扑克。这是一种最简不完全信息博弈，其状态空间较小且完全覆盖。

如图2左侧面板所示，JBR-PSRO 与 PSRO 之间的性能差距在各次迭代中始终较小，这表明 JBR 紧跟标准 PSRO 的收敛行为。由于库恩扑克的状态空间较为紧凑，且在联合元策略下每个状态–动作对都会被访问，因此基于共享

数据计算最佳响应时不存在分布偏移偏差。这证实，在诱导马尔可夫决策过程能够被同策略轨迹充分覆盖的低复杂度领域中，JBR 是可行的。

5.3 朴素 JBR 何时失效?

接下来，我们考察随着最佳响应计算变得更加复杂，JBR 的表现如何。为此，我们比较**PSRO** 和**JBR-PSRO** 在*Leduc*扑克中的表现。该游戏的状态空间显著更大，决策结构也更深。与库恩扑克不同，在元策略下收集的联合数据集仅覆盖部分可达轨迹，因此，当根据共享经验计算最佳响应时，会产生离线学习偏差。

如图2右侧面板所示，JBR-PSRO 与 PSRO 之间的性能差距显著扩大：PSRO 持续降低 NashConv，而 JBR-PSRO 则停滞在较高的可利用度水平。这表明，朴素 JBR 虽然高效，但当诱导马尔可夫决策过程变为部分可观测或高维时，无法保持均衡精度。这些结果促使我们开发了下文讨论的*JBR*变体。

5.4 JBR变体的效果如何?

现在，我们评估增强版*JBR*变体能否缓解朴素 JBR 中观察到的离线学习偏差。我们重点考察**JBR-PSRO-SPI**、**JBR-PSRO-δR**和**JBR-PSRO-δT** 相较于基线**JBR-PSRO**的表现。

如图2右侧面板所示，所有增强型 JBR 变体的表现都优于朴素基线，这证实了用于提升覆盖度和稳定性的显式机制能够显著改善最佳响应的质量。不过，各方法的相对效果有所不同：**JBR-PSRO-SPI**较为保守，因为其安全策略改进约束会迫使策略在采样不足的区域回退到基线元策略，从而限制了进展。采用随机探索的**JBR-PSRO-δR**虽然提升了数据多样性，但缺乏方向性，因此收益有限。相比之下，**JBR-PSRO-δT**通过引导探索转向有希望的轨迹，取得了接近 PSRO 的表现。其δ定向探索在数据收集期间显式地以当前最佳响应为条件，从而预判下一次更新后对手策略将如何变化。这种前瞻性行为类似于对手建模，有助于生成更准确、更稳定的最佳响应。有趣的是，这种预测特性也解释了为什么在早期迭代中，**JBR-PSRO-δT** 比标准 PSRO 收敛更快，因为智能体会主动适应

朝着即将进行的元策略更新方向。

基于这些观察，我们确定 **JBR-PSRO-δT** 是最有效的变体。为了进一步缩小与 PSRO 之间剩余的微小差距，我们测试了一种混合型方法，定期用独立最佳响应替换 JBR 更新。如图2所示，这种配置 **HBR-PSRO(10)-δT**在保持 JBR 显著样本效率提升的同时，实现了 PSRO 级别的收敛。

5.5 JBR 变体的样本效率如何?

现在，我们分析不同 PSRO 方法中最佳响应计算的样本效率。图 1 展示了*Leduc*扑克中的结果，显示用于最佳响应训练的环境回合总数（百万计，x 轴）与最低

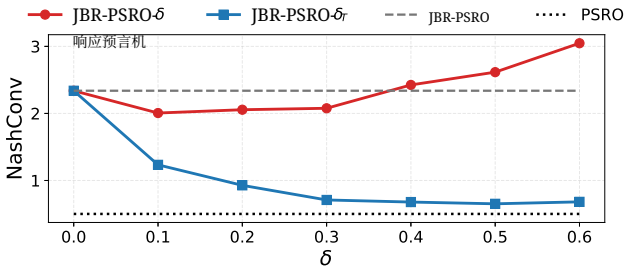

图 3：探索率 δ 对 Leduc扑克的影响。随机探索（JBR-PSRO-δR）和定向探索（JBR-PSRO-δT）在 100 次迭代后的最低 NashConv。随机探索在 δ=0.1 时达到峰值，超过 0。4 后则会退化；而定向探索的效果会持续提升至 δ=0.5，并且始终优于朴素 JBR。

在 100 次迭代后达到的 NashConv（y 轴）。这直接量化了各方法在达到相应均衡质量的同时，计算最佳响应的效率。

标准 **PSRO** 取得最低的 NashConv，但样本效率很低，因为每个智能体都会在每次迭代中独立训练一个新的最佳响应。相比之下，**JBR-PSRO** 及其增强变体会重复利用单个共享数据集，同时为所有智能体计算最佳响应，从而有效地将仿真成本分摊到各个智能体上。这大幅减少了最佳响应计算所需的环境交互次数，同时保持了有竞争力的均衡精度。

在增强变体中，**JBR-PSRO-δT** 在均衡精度与最佳响应样本效率之间实现了最佳权衡，其 NashConv 接近 PSRO，而所需的最佳响应回合仅为 PSRO 总数的一半。其混合型对应方法 **HBR-PSRO(10)-δT**和 **HBR-PSRO(30)-δT**则通过偶尔执行独立最佳响应更新，进一步缩小与 PSRO 的差距，而最佳响应回合仅略有增加。**JBR-PSRO-SPI** 更稳定但也更保守，而 **JBR-PSRO-δR** 则因探索缺乏针对性而带来适度提升。总体而言，这些结果证实，基于 JBR 的训练显著提高了最佳响应计算的样本效率，使 PSRO 能够实际应用于环境交互成本主导运行时长的领域。

5.6 对探索率的敏感性 δ

接下来，我们研究 δ 对*Leduc*扑克性能的影响。图 3 报告了不同 δ 值下，**JBR- PSRO-δR** 和**JBR-PSRO-δT** 经过 100 次迭代后的最终 NashConv。

对于 **JBR-PSRO-δR**，随着 δ 增加，性能起初有所提升，并在 $\delta = 0.1$时达到最佳。超过这一点后，随机混合程度越高，性能便逐渐下降；在 $\delta = 0.4$之后，结果甚至不如朴素的 **JBR-PSRO** 基线。这一现象反映了过度随机化带来的不稳定性，它会扭曲价值估计并削弱最佳响应信号。相比之下，**JBR-PSRO-δT** 随着 δ 增加，性能稳步提升，在 $\delta = 0.4$ 左右趋于平稳，并在 $\delta = 0.5$附近达到峰值。在 $\delta = 0.6$时性能略有下降，但在所有测试值下仍远高于 **JBR-PSRO**。这种

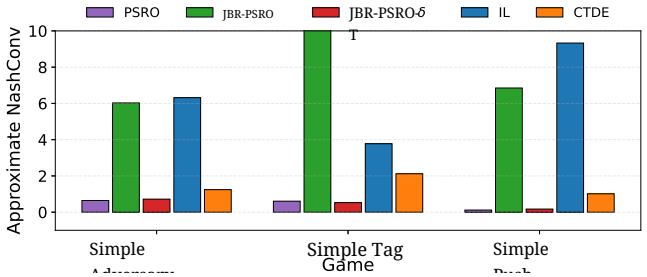

图4：连续多智能体环境中的近似 NashConv。比较 PSRO、JBR-PSRO、JBR-PSRO-δT 以及两个多智能体强化学习基线（IL/DDPG、集中训练、分散执行/MADDPG）在*SimpleTag*、*SimpleAdversary*和*SimplePush*中的表现。PSRO 和 JBR-PSRO-δT 在各博弈中取得了相当的近似 NashConv；两种 PSRO 方法均优于 IL 和 CTDE。

鲁棒性表明，在广泛的 δ取值范围内，定向探索仍能持续提升覆盖度和对手预测能力。

总体而言，这些结果证实，随机探索与定向探索的表现不同：随机探索需要进行幅度较小且谨慎的扰动，以避免学习过程失稳；而定向探索则受益于更强的混合，从而凸显元策略中的预测结构。

5.7 超越扑克：连续多智能体环境

现在，我们评估扑克中的结论能否推广至连续多智能体控制。在*Simple Tag*、*SimpleAdversary*和*Simple Push* 粒子环境中，精确的最佳响应难以计算；因此，我们报告近似 *NashConv*。对于每个智能体，我们在扩展的最佳响应预算（100,000 个回合）下，使用不同的随机初始化训练三个最佳响应候选项，并选出表现最佳的候选项，以近似真实的最佳响应（对独立学习和集中训练、分散执行进行评分时，也采用相同的流程）。

图4比较了**PSRO**、**JBR-PSRO**和**JBR-PSRO-δT**（表现最佳的增强变体），并列展示了两个标准多智能体强化学习框架：**独立学习（DDPG）**和**集中训练、分散执行（MADDPG）**。在全部三个环境中，**PSRO** 和**JBR-PSRO-δT**取得了相近的近似 NashConv，而**JBR-PSRO**则表现较差——这与 Leduc 中的发现一致，也凸显了在重复利用数据时进行定向探索的必要性。此外，PSRO 和 JBR-PSRO-δT 始终优于**独立学习**和**集中训练、分散执行**，表明它们对策略非平稳性具有更强的鲁棒性。这些结果进一步证明了 PSRO 式博弈论训练在连续域中的实用价值，并凸显了提升最佳响应计算样本效率的现实意义。

5.8 实证见解总结

综合所有实验结果，我们证明，结合定向探索时，JBR 能显著提高 PSRO 中最佳响应计算的样本效率，同时保持强劲的均衡表现。在小型且已被充分覆盖的博弈（库恩扑克）中，JBR 的收敛行为与 PSRO 相当，证明了其在采样充分条件下的可行性。在更大的

博弈（*Leduc*扑克）中，朴素 JBR 因离线偏差而表现较差，但增强变体——尤其是**JBR-PSRO-δT**——通过预判对手更新恢复了准确性；**混合最佳响应 PSRO(k)-δT**只需适度增加最佳响应回合，便弥补了与 PSRO 之间的剩余差距。消融实验表明，随机探索需要较小的混合比例，而定向探索则受益于更强的扰动（最优值约为 δ=0.5）。在连续多智能体控制中，**JBR-PSRO-δT**同样取得了与 PSRO 相当的表现，同时样本效率高得多。此外，PSRO 和 JBR-PSRO-δT 均比 IL 和 CTDE 等标准 MARL 框架表现出更强的鲁棒性，凸显了 PSRO 式博弈论训练对于稳定协作的重要性，也表明提高最佳响应计算效率能够使此类训练在更大规模的领域中切实可行。

6 讨论

实际意义。 PSRO仍然是学习具有战略稳健性的策略的最可靠方法之一，但其应用通常受到最佳响应计算成本高昂的限制。所提出的JBR框架通过在智能体之间复用经验来缓解这一限制，其中**JBR-PSRO-δT** 在均衡精度与最佳响应样本效率之间实现了最佳权衡。当需要接近精确的PSRO准确性时，混合更新（**HBR-PSRO(k)-δT**）能够以最小的额外成本恢复完整性能。

解读。 定向探索充当了一种隐式的对手建模形式：通过根据当前最佳响应来决定数据收集，智能体能够预测可能的对手适应行为，这解释了**JBR-PSRO-δT**为何能够在早期更快收敛。这一联系凸显了预测性探索如何提升元策略学习的效率与稳定性。

局限性。 尽管取得了提升，PSRO和JBR变体仍然需要许多迭代，并且所需样本量显著高于典型的多智能体强化学习方法。扩展到更大的策略群体和高维连续域仍然具有挑战性，因为元策略更新和收益矩阵增长会带来额外的计算成本。为 δ 和 k制定自适应调度，或整合学习得到的先验以引导探索，都是降低这一开销的有希望方向。

7 结论

我们提出了联合经验最佳响应（JBR）框架，通过在智能体之间复用共享交互数据，降低了 PSRO 中最佳响应计算的成本。尽管朴素 JBR 在复杂博弈中可能会受到离线学习偏差的影响，但增强型变体**JBR-PSRO-δT** 及其混合扩展 **HBR-PSRO(k)-δT**以远高于 PSRO 的样本效率，达到了与 PSRO 相当的均衡精度。在离散和连续多智能体域开展的实验表明，基于 JBR 的训练在保持 PSRO 鲁棒性的同时，也让其更适用于大规模应用。未来工作将探索自适应探索与混合调度、可扩展的元策略更新，以及将定向探索扩展至深度连续环境。

定理4.1的证明

我们给出了完整证明，表明采用探索增强型 JBR 的 PSRO 能够在有界误差范围内保持收敛保证。我们研究的是收益有界且取值范围为 $R = u - \underline{u}$ 的有限双人零和马尔可夫博弈，其中收益位于 $[\underline{u}, u]$。

引理 A.1（混合策略的线性性）。对于固定的 π_i，$u_i(\pi_i, \cdot)$ 关于对手混合策略是线性的，而 $u_i(\cdot, \cdot)$ 关于 (σ_i, σ_{-i}) 是双线性的。

证明。直接由

$$u_i(\sigma_i, \sigma_{-i}) = \sum_{\pi_i, \pi_{-i}} \sigma_i(\pi_i)\sigma_{-i}(\pi_{-i})u_i(\pi_i, \pi_{-i}).$$

□

引理 A.2（扰动下的稳定性）。对于任意 π_i 和混合策略 σ_{-i}, $\tilde{\sigma}_{-i}$,

$$\left|u_i(\pi_i, \sigma_{-i}) - u_i(\pi_i, \tilde{\sigma}_{-i})\right| \le \frac{R}{2} \left\|\sigma_{-i} - \tilde{\sigma}_{-i}\right\|_1.$$

证明。根据引理 A.1，$u_i(\pi_i, \cdot)$ 是一个有界函数的期望。期望之差的上界为 $(R/2)\|\cdot\|_1$（全变差与有界函数的对偶性）。□

引理 A.3（对扰动后的 $\Rightarrow$ 近似最佳响应到真实值的最佳响应）。令 $\pi_i^\star \in$ 最佳响应$_i(\sigma_{-i})$ 以及 $\hat{}\pi_i$ 为 ε-最佳响应于 $\tilde{\sigma}_{-i}$。则

$$u_i(\pi_i^\star, \sigma_{-i}) - u_i(\hat{\pi}_i, \sigma_{-i}) \le \varepsilon + R\left\|\sigma_{-i} - \tilde{\sigma}_{-i}\right\|_1.$$

证明。在 $\tilde{\sigma}_{-i}$ 下对期望进行加减，可得

$$\begin{aligned} &u_i(\pi_i^\star, \sigma_{-i}) - u_i(\hat{\pi}_i, \sigma_{-i}) \\ &= \left[u_i(\pi_i^\star, \sigma_{-i}) - u_i(\pi_i^\star, \tilde{\sigma}_{-i})\right] + \left[u_i(\pi_i^\star, \tilde{\sigma}_{-i}) - u_i(\hat{\pi}_i, \tilde{\sigma}_{-i})\right] \\ &\qquad + \left[u_i(\hat{\pi}_i, \tilde{\sigma}_{-i}) - u_i(\hat{\pi}_i, \sigma_{-i})\right]. \end{aligned}$$

中间项为 $\le \varepsilon$，这是由于 $\hat{}\pi_i$ 在 $\tilde{\sigma}_{-i}$ 下的最优性。另外两项各自为 $\le (R/2)\|\sigma_{-i} - \tilde{\sigma}_{-i}\|_1$，依据引理 A.2。□

定理 A.4（带扰动目标的 PSRO）。假设在终止时，每个神谕计算一个针对 ε $\tilde{\sigma}_{-i}$ $\|\tilde{\sigma}_{-i} - \sigma_{-i}\|_1 \le$-最佳响应。如果终止检查在 Δ 下执行，则 σ 是一个 (σ $\varepsilon +$ $2R\Delta$)-纳什均衡。

证明。固定玩家 i。令 $\pi_i^\star \in$ 最佳响应$_i(\sigma_{-i})$，且 $\hat{}\pi_i$ 为神谕对 $\tilde{\sigma}_{-i}$ 的 ε-最佳响应。根据引理 A.3,

$$u_i(\pi_i^\star, \sigma_{-i}) - u_i(\hat{\pi}_i, \sigma_{-i}) \le \varepsilon + R\Delta.$$

根据终止条件，$u_i(\sigma_i, \sigma_{-i}) \ge u_i(\hat{}\pi_i, \sigma_{-i})$。因此

$$u_i(\sigma_i, \sigma_{-i}) \ge \max_{\pi_i} u_i(\pi_i, \sigma_{-i}) - (\varepsilon + R\Delta),$$

这即是所需的均衡条件。对另一名玩家应用相同论证，即可得到该结论。□

在数据收集期间，我们使用按状态 δ混合：

$$\tilde{\sigma}_j(a_j \mid s) = (1 - \delta)\sigma_j(a_j \mid s) + \delta\nu_j(a_j \mid s),$$

这意味着 $\|\tilde{\sigma}_j(\cdot \mid s) - \sigma_j(\cdot \mid s)\|_1 \le 2\delta$。因此，所诱导的偏差参数满足 $\Delta \le 2\delta$。

推论 A.5（探索增强型联合经验最佳响应）。若每个神谕都对 δ 混合行为计算一个 ε-最佳响应，则 *PSRO* 的终端元策略是一个 ($\varepsilon + 2R\delta$)-纳什均衡。

该推论确立了正文中的定理4.1。

注。每个智能体都能计算针对 $\tilde{\sigma}_{-i}$ 的 ε-最佳响应这一假设，应当按照与标准PSRO相同的神谕意义来理解：它抽象掉了获得最佳响应所采用的机制。在实践中，响应是在 $\tilde{\sigma}$条件下收集的数据上训练得到的。在没有扰动（$\delta = 0$）的情况下，这会退化为一个可能存在支撑不匹配的离线强化学习问题。有了 $\delta > 0$后，探索可确保所有动作都获得非零概率，因此数据集具有更广的覆盖度，离线问题也得到了更好的适定性。因此，扰动并不会使该设置等同于在线学习，但它保证了 ε-最佳响应目标定义良好，并且可以从数据中近似得到。近期关于如何从固定的多智能体数据集——无需在线交互——施加均衡条件的研究，也从逆向学习和拟像多智能体学习的角度探讨了相关观点 [7]。

参考文献

[1] Stefano V. Albrecht、Filippos Christianos 和 Lukas Schäfer。2024。《多智能体强化学习：基础与现代方法》。麻省理工学院出版社。https: // www.marl-book.com[2] Ariyan Bighashdel、Daan de Geus、Pavol Jancura 和 Gijs Dubbelman。2023。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。[3] Ariyan Bighashdel、Daan De Geus、Pavol Jancura 和 Gijs Dubbelman。2024。多智能体强化学习中的离策略动作预期。《机器学习研究杂志》25, 67 (2024), 1–31。[4] Ariyan Bighashdel、Yongzhao Wang、Stephen McAleer、Rahul Savani 和 Frans A. Oliehoek。2024。策略空间响应预言机：综述。载于*IJCAI*。ijcai.org，7951–7961。[5] Wojciech M. Czarnecki、Gauthier Gidel、Brendan D. Tracey、Karl Tuyls、Shayegan Omidshafiei、David Balduzzi 和 Max Jaderberg。2020。现实世界中的博弈看起来像旋转的陀螺。载于*NeurIPS*。17443–17454。[6] Sevan G. Ficici 和 Jordan B. Pollack。2003。协同进化的博弈论记忆机制。载于*GECCO*。286–297。[7] Denizalp Goktas、Amy Greenwald、Sadie Zhao、Alec Koppel 和 Sumitra Ganesh。2024。高效的逆向多智能体学习。载于*ICLR*。[8] Landon Kraemer 和 Bikramjit Banerjee。2016。将多智能体强化学习作为去中心化规划的演练。*Neurocomputing* 190 (2016), 82–94。[9] Harold W Kuhn。1953。扩展式博弈与信息问题。博弈论贡献2, 28 (1953), 193–216。[10] Harold W Kuhn。2016。一种简化的双人扑克。博弈论贡献1 (2016), 97–103。[11] Marc Lanctot、Edward Lockhart、Jean-Baptiste Lespiau、Vinicius Zambaldi、Satyaki Upadhyay、Julien Pérolat、Sriram Srinivasan、Finbarr Timbers、Karl Tuyls、Shayegan Omidshafiei 等。2019。OpenSpiel：用于博弈中强化学习的框架。arXiv 预印本 arXiv:1908.09453。[12] Marc Lanctot、Vinícius Flores Zambaldi、Audrunas Gruslys、Angeliki Lazaridou、Karl Tuyls、Julien Pérolat、David Silver 和 Thore Graepel。2017。多智能体强化学习的统一博弈论方法。载于*NIPS*。4190–4203。[13] Romain Laroche、Paul Trichelair 和 Remi Tachet Des Combes。2019。利用基线自助法实现安全策略改进。载于*ICML*。3652–3661。[14] Timothy P. Lillicrap、Jonathan J. Hunt、Alexander Pritzel、Nicolas Heess、Tom Erez、Yuval Tassa、David Silver 和 Daan Wierstra。2016。基于深度强化学习的连续控制。载于*ICLR (*海报*)*。[15] Siqi Liu、Marc Lanctot、Luke Marris 和 Nicolas Heess。2022。单纯形神经种群学习：对称零和博弈中的任意混合贝叶斯最优性。载于*ICML*。13793–13806。[16] Siqi Liu、Luke Marris、Daniel Hennes、Josh Merel、Nicolas Heess 和 Thore Graepel。2022。NeuPL：神经种群学习。载于*ICLR*。[17] Siqi Liu、Luke Marris、Marc Lanctot、Georgios Piliouras、Joel Z Leibo 和 Nicolas Heess。2024。超越对称零和博弈的神经种群学习。载于*AAMAS*。1247–1255。[18] Ryan Lowe、Yi Wu、Aviv Tamar、Jean Harb、Pieter Abbeel 和 Igor Mordatch。2017。面向混合合作-竞争环境的多智能体演员-评论家。载于*NIPS*。6379–6390。[19] Stephen McAleer、Gabriele Farina、Gaoyue Zhou、Mingzhi Wang、Yaodong Yang 和 Tuomas Sandholm。2023。通过合作强化学习，在大型团队博弈中学习近似 TMECor 的 Team-PSRO。(2023), 45402–45418。[20] Stephen McAleer、John B. Lanier、Roy Fox 和 Pierre Baldi。2020。Pipeline PSRO：一种在大型博弈中求解近似纳什均衡的可扩展方法。(2020), 20238–20248。

[21] H. Brendan McMahan、Geoffrey J. Gordon 和 Avrim Blum。2003。对手控制成本函数情形下的规划。载于*ICML*。536–543。 [22] Frans A. Oliehoek、Edwin D. de Jong 和 Nikos Vlassis。2006。用于非对称博弈的并行纳什记忆。载于*GECCO*。337–344。 [23] Ricky Sanjaya、王军和 Yaodong Yang。2022。衡量国际象棋中的非传递性。《*Algorithms*》15, 5 (2022), 152。 [24] Rahul Savani 和 Theodore L. Turocy。2023。Gambit：用于博弈论计算的软件包，16.1.0 版。http://www.gambit-project.org [25] L. Julian Schvartzman 和 Michael P. Wellman。2009。在经验博弈建模中探索大型策略空间。载于*AMEC*。139–152。 [26] Yoav Shoham、Rob Powers 和 Trond Grenager。2007。如果多智能体学习是答案，那么问题是什么？《*Artificial Intelligence*》171, 7 (2007), 365–377。 [27] Max Olan Smith、Thomas Anthony 和 Michael P Wellman。2023。战略知识迁移。《机器学习研究杂志》24, 233 (2023), 1–96。

[28] Max OlanSmith和Michael PWellman。2023。共同学习经验博弈和世界模型。arXiv 预印本 arXiv:2305.14223。 [29] Finnegan Southey、Michael H. Bowling、Bryce Larson、Carmelo Piccione、Neil Burch、Darse Billings 和 D. Chris Rayner。2005。Bayes? Bluff：扑克中的对手建模。发表于*UAI*。550–558。[30] Karl Tuyls 和 Gerhard Weiss。2012。多智能体学习：基础、挑战与展望。《*AI Magazine*》33, 3 (2012), 41–41。[31] Bernhard von Stengel。2002。计算双人博弈的均衡。发表于《博弈论手册：经济学应用》。第 3 卷。爱思唯尔, 1723–1759。[32] 周明、陈景晓、文颖、张卫南、Yaodong Yang、俞勇和王军。2022年。高效策略空间响应预言机。arXiv 预印本 arXiv:2202.00633。