策略探索的　　经验博弈论分析中的正则化

Yongzhao Wang，**Michael P. Wellman**密歇根大学{wangyzh, wellman}@umich.com

arXiv:2302.04928v1 [cs.GT] 9 Feb 2023

摘要

在经验博弈论分析（EGTA）的迭代方法中，策略空间会根据对中间游戏模型的分析逐步扩展。以双重预言机算法为代表的一种常见策略探索方法，是添加对当前均衡做出最佳响应的策略。这种方法可能会遭遇过拟合及其他局限，因此，迭代式 EGTA 框架策略空间响应预言机（PSRO）的开发者对最佳响应目标进行了推广，并将其称为元策略求解器（MSS）。鉴于许多 MSS 可被视为纳什均衡的扰动版本或近似版本，我们采用明确的正则化视角来规范和分析 MSS。我们提出一种新颖的 MSS，即正则化复制动态（RRD），它只需依据遗憾准则截断这一过程。我们证明，RRD 在各种博弈中的表现优于现有 MSS。我们将研究扩展到三人博弈；在这类博弈中，收益矩阵的规模随策略数量呈立方增长，因此穷举评估策略组合可能并不可行。我们采用一种能够从不完整模型中识别解的策略组合搜索方法，并将其与使用正则化 MSS 的迭代模型构建相结合。最后，我们通过实验发现，最佳响应目标的遗憾值是策略探索表现的有力指标，这也解释了正则化在 PSRO 中为何有效。

1 引言

经验博弈论分析（***EGTA***）的方法论 [Tuyls 等，2020年; Wellman， 2016] 为基于仿真数据的模型提供了丰富的博弈推理工具。[1] 由于许多我们关注的多智能体系统难以用解析方法表达或处理，EGTA 提供了一种替代方法：通过仿真考察一系列策略，并结合博弈模型的构建与推断。能够明确纳入博弈模型的策略数量受到计算限制的显著制约，因此，选择纳入哪些策略至关重要。为获得准确的分析结果，我们要求纳入的策略具有较高的表现，并涵盖关键战略问题 [Balduzzi 等，2019]。高效地为 EGTA 组建有效策略组合所面临的挑战称为策略探索问题 [Jordan 等，2010]。

在经验博弈论分析中，策略探索最清晰的表述方式是将其置于一个迭代过程中：新策略的生成与博弈模型的估计和分析交替进行。Lanctot等人 [2017] 提出的策略空间响应预言机（PSRO）算法为迭代式经验博弈论分析提供了灵活的框架，其中每次迭代都会通过强化学习（RL）生成新策略。学习玩家在一个环境中进行训练，在该环境中，其他玩家固定采用由先前迭代得到的策略构成的策略组合（纯策略或混合策略）。关键的设计问题是，如何设定其他玩家的策略组合，使其作为训练目标。在 PSRO 中，负责推导这一目标的组件称为元策略求解器 （MSS），它以经验博弈模型为输入，并通过“求解”该模型生成目标策略组合。随后，学习智能体利用强化学习搜索一个能够对 MSS 目标作出最佳响应的策略。实际上，指定 MSS 就定义了 PSRO 的策略探索方法。

MSS 的一个显而易见的选择，是用作目标博弈分析的解概念，通常为纳什均衡（NE）。逐步加入针对当前策略集合的 NE 的最佳响应策略，这一算法称为双重预言机（DO）算法 [McMahan 等， 2003],而将 NE 用作 MSS 的 PSRO，本质上就是用强化学习计算（近似）最佳响应的 DO。虽然 DO 往往有效，但大量证据表明，针对 NE 的最佳响应并不总是策略探索的最佳方法。Schvartzman 和 Wellman [2009a] 观察到一些案例，在这些案例中，该方法趋近真实均衡的速度极慢，以至于即便加入随机策略，也能显著加快速度。更普遍地说，Lanctot 等 [2017]指出，针对纳什均衡进行最佳响应会过拟合当前的均衡策略，因此往往会产生无法泛化到整体策略空间的结果。这正是他们定义广义 MSS 概念以进行策略探索的主要动机。例如，作为一种替代性的 MSS，Lanctot等

[1] 附录 A 中提供了缩略语表。

[2017] 提出了投影复制子动力学 （PRD），该方法采用复制子动力学（RD）搜索均衡，并截断复制子更新，以确保每种纯策略的选择概率都有一个下界。

我们在这一方向上更进一步，采用明确的正则化视角来规范和分析 MSS。我们提出了一种新颖的 MSS，称为正则化复制子动力学 （RRD）。该方法基于遗憾准则，在中间经验博弈模型中截断纳什均衡搜索过程。具体而言，在 PSRO 的每次迭代中，通过运行 RD 来更新最佳响应目标策略组合；当当前策略组合相对于经验博弈的遗憾值达到指定的遗憾阈值时，停止运行。遗憾阈值是一个超参数，可以根据特定博弈类别进行调整，也可以通过退火来控制不同迭代过程中的正则化程度。我们在各种博弈中评估了 RRD 的性能，并表明，与现有的几种 MSS 相比，RRD 在收敛速度和中间经验博弈模型的质量方面表现更优。

由于收益矩阵的规模随着玩家数量呈指数增长，在 PSRO 的迭代过程中维护完全指定的模型，其成本在超过两名玩家后可能变得难以承受。为缓解这一问题，我们采用了一种与 PSRO 兼容的策略组合搜索方法，称为反向策略组合搜索 （BPS）。该方法无需模拟完整的收益矩阵即可寻找解概念。我们将 RRD 与 BPS 相结合，并在一个三人博弈中展示了这一组合的有效性。

最后，我们的实验揭示了策略探索正则化带来益处的原因。在各种设置下，我们发现，与经验博弈的精确纳什均衡相比，RRD 得出的近似经验博弈纳什均衡在完整博弈中的遗憾值更低。这不仅解释了正则化的益处，也可能为评估 PSRO 相关方法中新型 MSS 设计的潜力提供一种途径。

本研究的贡献包括：

1. RRD：一种新型 MSS，它依据遗憾值标准截断中间游戏模型中的纳什均衡搜索过程。与以往的 MSS 相比，我们的 MSS 具备策略探索所需的特性（例如，在不同博弈间具有更强的适应性）。我们证明，在多种博弈中，RRD 的表现优于文献中的相关替代方法。

2. 对采用 RRD 进行学习的全面分析，包括性能稳定性与遗憾阈值之间的关系。

3. 将选择性策略组合评估方法与 PSRO 相结合，从而无需模拟整个收益矩阵即可求解博弈。我们将 RRD 与该方法结合，并展示其在三人博弈中的学习效果。

4. 展示最佳响应目标的遗憾值与 PSRO 中 MSS 表现之间的关键关系。

2 策略探索相关工作

EGTA 中首次自动生成策略，是通过对参数化策略空间进行遗传搜索，优化相对于经验博弈均衡的表现 [Phelps 等人， 2006]。Schvartzman 和 Well- man [2009b] 在 EGTA 中将表格型强化学习用作策略生成的最佳响应预言机。这些作者还阐述了 EGTA 中策略探索 这一一般性问题，并研究了除对均衡进行最佳响应之外是否存在更好的选择 [Schvartzman和Wellman，2009a年]。Jordan 等人 [2010] 进一步拓展了这项研究，通过加入能够最大化相对于经验理性闭包的偏离收益的策略。

PSRO框架的引入显著推动了策略探索研究 [Lanctot 等， 2017]。PSRO将深度强化学习用作近似最佳响应预言机，以响应MSS选定的特定其他智能体策略组合。采用NE作为MSS时，PSRO便退化为DO算法 [McMahan 等， 2003]。为有效生成策略，Lanctot 等 [2017] 在针对NE过拟合与泛化至经验博弈之外的策略空间之间取得平衡，并提出了投影复制子动力学（PRD）。PRD采用复制子动态搜索来求解均衡 [Taylor和Jonker，1978年；Smith和 Price， 1973] 并确保每种纯策略的选择概率都有下界。为简便起见，只要上下文明确，我们通常用MSS简称采用该MSS的PSRO。例如，我们可能说“PRD”来指“采用PRD的PSRO”。

PSRO 也可以被视为对某些经典博弈学习动态的推广。例如，将当前策略的均匀分布选作 MSS，本质上会重新实现经典的虚拟博弈（FP）算法 [Brown,1951]。此外，只提取最新策略的 MSS 复现了迭代最佳响应算法。需要注意的是，产生这些动态的 MSS 仅在一种极其宽泛的意义上算作求解器；它们并未实质性地使用经验博弈模型，而是直接根据策略集得出结果。

沿着 PSRO 的思路，一些研究提出了 MSS，通过有效地正则化目标策略组合，避免对精确均衡作出最佳响应。具体而言，Wang等人 [2019] 采用了纳什均衡与均匀分布的混合，本质上是在每次 PSRO 迭代中随机决定采用双重预言机还是虚拟博弈，从而展示了结合多种 MSS 的可能性。Wright等人 [2019] 为双重预言机加入了一个调整步骤，使生成的策略网络针对先前均衡策略的混合进行微调。Balduzzi等人 [2019] 提出了一种新的 MSS，称为修正纳什均衡，旨在提高经验策略空间的多样性。Dinh等人[2022] 提出了一种用于双人零和博弈的 MSS，它对经验博弈应用在线学习，并将在线策略组合作为最佳响应目标。除了选择纳什均衡作为解概念之外，Muller等人 [2020] 还提出了一种基于演化概念的新 MSS——α-rank [Omid- shafiei等人， 2019], 以及 Marris等人 [2021] 提出了最大福利粗相关均衡（MWCCE）和最大基尼粗相关均衡（MGCCE），用于

在 PSRO 框架内计算相关均衡。

3 预备知识

正规形博弈 $\mathcal{G}=(N,(S_i),(u_i))$ 由有限玩家集合 N（以 i 为索引）、玩家 $i\in N$ 的非空策略集合 S_i，以及玩家 $i\in N$ 的效用函数 $u_i:\prod_{j\in N}S_j\to\mathbb{R}$ 组成。

混合策略 σ_i 是 S_i 中策略上的概率分布，其中 $\sigma_i(s_i)$ 表示玩家 i 选择策略 s_i 的概率。我们采用其他玩家策略组合的常规记法：$\sigma_{-i}=\prod_{j\neq i}\sigma_j$。令 $\Delta(\cdot)$ 表示某个集合上的概率单纯形。玩家 i 的混合策略空间为 $\Delta(S_i)$。类似地，$\Delta(S)=\prod_{i\in N}\Delta(S_i)$ 是混合策略剖面空间。

玩家 i 对策略组合 σ 的最佳响应由在固定其他玩家策略时使其收益最大化的策略组成：

$$br_i(\sigma_{-i})\equiv\underset{\sigma_i'\in\Delta(S_i)}{\operatorname{argmax}}\,u_i(\sigma_i',\sigma_{-i}).$$

令 $br(\sigma)\equiv\prod_{i\in N}br_i(\sigma_{-i})$ 表示策略组合 σ 的整体最佳响应对应。纳什均衡 (NE) 是满足 $\sigma^*\in br(\sigma^*)$ 的策略组合 σ^*。

博弈 $\mathcal{G}$ 中玩家 i 对策略组合 σ 的遗憾值 定义为

$$\rho_i^{\mathcal{G}}(\sigma)\equiv\max_{s_i'\in S_i}u_i(s_i',\sigma_{-i})-u_i(\sigma_i,\sigma_{-i}).$$

遗憾值刻画了玩家 i 通过单方面偏离其在 σ 中的混合策略，转而采用 S_i中的另一种策略，预期最多能获得的收益。纳什均衡中每个玩家的遗憾值均为零。如果不存在玩家能够通过单方面偏离获得超过 ϵ 的收益，则称策略组合为 ϵ-纳什均衡 (ϵ-NE)。我们将策略组合 σ 的遗憾值定义为各玩家遗憾值之和：

$$\rho^{\mathcal{G}}(\sigma)\equiv\sum_{i\in N}\rho_i^{\mathcal{G}}(\sigma).$$

受限博弈 $\mathcal{G}_{S\downarrow X}$ 是完整博弈 $\mathcal{G}$的一个投影，在该博弈中，玩家从受限策略集 $X_i\subseteq S_i$中进行选择。$\hat{}$ 经验博弈 $\mathcal{G}$ 是真实博弈 $\mathcal{G}$ 的一个模型，其中$\hat{}$收益通过仿真估算。因此，$\mathcal{G}_{S\downarrow X}=(N,(X_i),(\hat{}u_i))$ 表示一个经验博弈模型，其中$\hat{}u$ 是 u 在策略空间 X上的估计投影。

复制动态 (RD)描述了混合策略组合随时间演化的轨迹，其灵感来自自然选择 [Taylor和Jonker，1978年；Smith和Price，1973]。RD通常用作启发式均衡搜索算法。我们考虑RD的一种离散形式，其中，玩家 i选择某一策略的概率会根据其从当前混合策略转向该策略所获得的收益按比例更新。数学上，玩家 i在当前策略组合 σ 中采用策略 s_i 的复制方程为

$$\frac{d\sigma_i(s_i)}{dt}=\sigma_i(s_i)[u_i(s_i,\sigma_{-i})-u_i(\sigma_i,\sigma_{-i})].$$

在 RD 的每次迭代中，玩家 i 的混合策略 σ_i 通过 $\sigma_i\leftarrow P(\sigma_i+\alpha\frac{d\sigma_i}{dt})$ 更新，其中 α 是RD 的步长，P 是到策略单纯形的投影算子，即 $P(\sigma_i)=\operatorname{argmin}_{\sigma_i'\in\Delta}\|\sigma_i'-\sigma_i\|_2$。

PSRO如下所示，以算法1的形式呈现。MSS的选择决定了策略探索方法。

算法1PSRO，由求解器MSS参数化

输入： 初始策略集 X
$\hat{}$**1**：通过模拟 $\sigma\in X$估计 $\mathcal{G}_{S\downarrow X}$
2：将目标初始化为 $\sigma_i\leftarrow\text{Uniform}(X_i)$
3: **对于** PSRO迭代 $\tau=1,2,\ldots,\mathcal{T}$ **do**
4: **对于**玩家 $i\in N$ **执行**
5: **进行**多轮强化学习训练回合**执行**
6: 采样一个策略组合 $s_{-i}\in\sigma_{-i}$
7: 针对 s_{-i}训练最佳响应预言机 s_i'
8: **结束循环**
9: $X_i\leftarrow X_i\cup\{s_i'\}$
10: **结束循环**
11: 更新 $\hat{\mathcal{G}}_{S\downarrow X}$ 通过模拟 X上的缺失策略组合
12: 计算最佳响应目标 $\sigma\leftarrow\text{MSS}(\hat{\mathcal{G}}_{S\downarrow X})$
13: **结束循环**
14: **返回** $\hat{\mathcal{G}}_{S\downarrow X}$

4 策略探索正则化

4.1 正则化复制子动力学

为避免对 NE 的响应过拟合，我们采用显式的正则化视角来进行策略探索。具体而言，我们提出一种通过截断基于 RD 的搜索来推导近似 NE 的方法。我们提出的新型 MSS 称为正则化 ***RD***（RRD），它只需在经验博弈上运行 RD，并在当前策略组合（相对于经验博弈）的遗憾值达到指定的遗憾阈值 λ 时停止，或在达到最大迭代次数时停止。在 RRD 过程中（算法 2），每位玩家的策略都初始化为经验博弈中各策略上的均匀分布。随后反复应用复制方程，直到当前策略组合（相对于经验博弈）的遗憾值小于遗憾阈值 λ。由于 RD 通常不会收敛到精确均衡，因此无法保证有限的遗憾阈值 λ 一定会达到。因此，我们设定最大迭代次数 M；若达到该上限，则返回截至当时找到的遗憾值最低的策略组合。

请注意，RRD 支持通过一个显式参数——遗憾值阈值——直接控制正则化程度。只要衡量遗憾值所依据的效用尺度具有可比性，该参数在具有不同策略集的博弈中都具有意义。

算法 2RRD

参数：遗憾阈值 λ，RD 步长 α
$\hat{}$**输入**：一场经验博弈 $\mathcal{G}_{S\downarrow X}$
1：使用 $\sigma_i\leftarrow$ 均匀分布(X_i)初始化复制子动力学
2：**当** $\rho^{\hat{\mathcal{G}}_{S\downarrow X}}(\sigma)>\lambda$ **双重预言机**
3: **对于**玩家 $i\in N$ **执行**
4: $\sigma_i\leftarrow P(\sigma_i+\alpha\frac{d\sigma_i}{dt})$
5: **结束 for 循环**
6: **结束 while 循环**
7: **返回** σ

将 RRD 用作算法1中的 MSS，即可得到采用 RRD 的 PSRO 流程。只需稍作修改，即可支持退火方案，使 RRD 的遗憾值阈值随 PSRO迭代而变化。

4.2 使用 BPS 选择性评估策略组合

PSRO 扩展面临的一项障碍是，经验博弈的规模会随着玩家数量的增加而呈指数级增长。即使在玩家数量较少（例如三到四名玩家）的博弈中，随着策略空间不断扩大，穷举模拟收益矩阵也可能变得不可行。幸运的是，通常可以从部分收益矩阵中推导出解 [Fearnley 等人， 2015]，而且此前的经验博弈论分析研究者已经开发出一些方法，通过选择性评估策略组合来搜索均衡 [Jordan 等人，2008；Sureka和Wurman， 2005]。例如，Brinkman和 Wellman [2016] 提出的算法会维护经验博弈的一组完整子博弈 （即已对其中策略组合进行穷举评估的极大策略集）。这些子博弈的纳什均衡会被视为经验博弈的候选均衡。对于每个候选均衡，算法会评估所有偏离到子博弈外策略的单玩家偏离。如果不存在对该候选均衡有利的偏离，那么该候选均衡就确认为 经验博弈的纳什均衡。否则，算法会将有利偏离策略纳入子博弈，并继续搜索。

我们为 PSRO 开发了一种简单的策略组合搜索方法，称为反向策略组合搜索 （BPS，算法 3）。我们的方法与 Brinkman和Wellman [2016],的方法相似，但考虑了策略生成的先后顺序。在每次迭代中，BPS 都从 PSRO 最近添加到经验博弈中的策略开始搜索，然后沿着先前的 PSRO 迭代反向搜索潜在偏离。我们的动机是，最新策略最有可能参与均衡。一旦 BPS 确认了经验博弈的一个解，我们就会对该解支撑集上的子博弈应用 RRD。根据构造，该子博弈已被完全评估（这是 RD 的要求），而整个经验博弈的收益矩阵则只被部分评估。在实验中，我们表明，BPS 能够在三人博弈中成功找到最佳响应目标，而无需穷尽评估整个经验博弈。

图 1 展示了 PSRO 第三次迭代时 BPS 的机制，此时每个玩家都有四种策略。图 1 中的$4 \times 4 \times 4$立方体表示当前经验博弈的收益矩阵。绿色单元格表示在先前迭代中已评估的策略组合的收益，白色单元格表示当前子博弈均衡的潜在偏离收益。缺失的单元格表示尚未评估的策略组合的收益。请注意，当前收益矩阵并不完整。为找到经验博弈的纳什均衡，BPS 从评估由每个玩家最近添加的策略构成的子博弈开始搜索，该子博弈由红色单元格表示。由于红色单元格是一个纯策略组合，因此它也是当前子博弈的纳什均衡。接下来，BPS 评估从红色单元格出发的所有潜在偏离（白色单元格）的收益。假设蓝色单元格是某个玩家偏离收益最大的策略组合

算法 3反向策略组合搜索

ˆ 输入：具有部分收益矩阵的经验博弈 $\mathcal{G}_{S\downarrow X}$ **。1：使用策略集** $Z =$**(**Z_i**)、** $i \in N$ **初始化子博弈，其中** $Z_i = \{s_\tau^i\}$**，且** s_τ^i **是最近一次 PSRO 迭代中添加的玩家** i **策略** τ**。2：当 True 时执行 ˆ 3：**$\sigma \leftarrow$**纳什均衡搜索 (**$G_{S\downarrow Z}$**) 4：偏离存在** $\leftarrow$**False 5：对玩家** $i \in N$ **执行 6：**$s_i \leftarrow$ **argmax**$_{s' \in X_i}$ u_i**(**s', σ_{-i}**)** $- u_i$**(**σ_i, σ_{-i}**) 7：如果** $s_i \notin Z_i$ **则 8：**$Z_i \leftarrow Z_i \cup \{s_i\}$**9：偏离存在** $\leftarrow$**True 10：结束 if 11：结束 for 12：如果** $\neg$**存在偏离 则 13：返回** σ**14：结束 if 15：通过仿真评估** Z **的缺失策略组合。16：结束 while**

i。然后玩家 i 将相应的偏离策略添加到当前子博弈的策略集中。此时，当前子博弈的策略剖面空间包含红色和蓝色策略组合。接着，BPS 重复以下过程：评估当前子博弈中的所有策略组合，计算当前子博弈的纳什均衡，并评估该纳什均衡的潜在偏离。同样，如果紫色单元格是一个偏离策略组合，那么相应的偏离策略就会被添加到该子博弈的策略集中。BPS 持续重复这一过程，直到确认该子博弈的纳什均衡，也就是说，在经验博弈中找不到有利偏离。

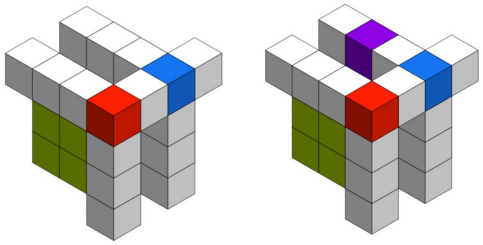


图 1 三人经验博弈的部分收益矩阵及 BPS 工作流程示意图。绿色：先前 PSRO 迭代中已评估的策略组合；白色：当前子博弈纳什均衡的偏离策略组合；红色：包含最近添加策略的策略组合；蓝色和紫色：偏离收益最大的策略组合。

5 性能与分析

5.1 实验结果

双人 Leduc 扑克

在图 2 中，我们在双人 Leduc 扑克上测试算法，并绘制了双重预言机、

PSRO 与 PRD，以及 PSRO 与 RRD（采用两种停止标准）所得到的遗憾曲线（相对于完整博弈）。我们首先观察到，与其他 MSS 相比，RRD 能迅速收敛到较低的遗憾值。值得注意的是，RRD 大幅超越了 PRD（此前该博弈中已知的最佳方法）。

为了展示将遗憾阈值作为停止准则相较于固定数量的复制子动力学更新所带来的优势，我们绘制了采用固定数量复制子动力学更新时 RRD 的最佳遗憾曲线。我们观察到，采用遗憾阈值时 RRD 的表现更好。这是因为，在不同经验博弈中，产生适当正则化程度所需的复制子动力学更新次数各不相同。

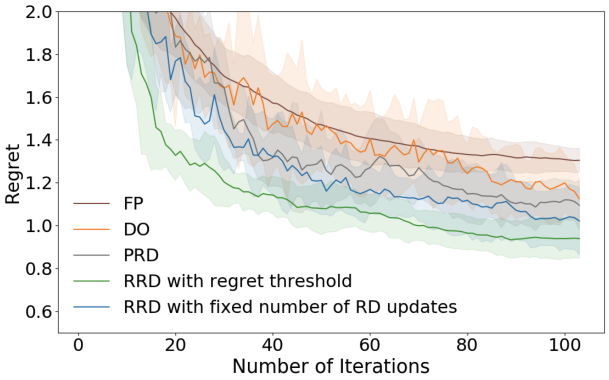


图 2 RRD 在双人 Leduc 扑克中的表现。

真实世界博弈

我们进一步在 Czarnecki等人 [2020]研究的四种“真实世界博弈”中评估了我们的算法：六角棋、四子棋、反向井字棋和围棋。我们观察到，在这四种博弈中，RRD 的收敛速度都快于虚拟博弈和双重预言机。更多实验细节以及完整博弈的说明见附录B.2。

多人博弈

我们将 BPS 与 RRD 结合应用于三人Leduc扑克。如图 3 所示，尽管 RRD 仅应用于包含精确均衡支撑集的子博弈，学习仍然受益于正则化。表1列出了在不同 PSRO迭代中，采用和不采用 BPS 时，在经验博弈中评估的策略组合平均数量。为评估一个策略组合，我们对 1000 个样本求平均来估算收益。在此示例中，与穷举估算相比，使用 BPS 节省了约 10% 的仿真工作量。我们还可以看到，随着迭代次数增加，评估所节省的工作量略有增加。例如，从第 41 次到第 50 次迭代，我们评估了 53890 个策略组合，而策略剖面空间增加了 61000 个，节省了 11.7%。以往经验博弈论分析研究中的经验表明，对于玩家超过三人的博弈，节省的比例会显著更高 [Brinkman和Wellman,2016]。

攻击图博弈

攻击图 [Miehling 等，2015] 是网络安全分析中的一种工具，用于建模对手可能借以入侵系统的路径。

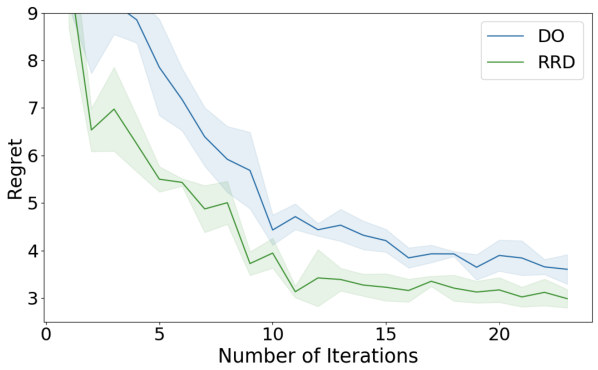


图 3　三人Leduc扑克中的 RRD 性能。

迭代次数#	使用 BPS 的策略组合数#	策略组合数#	节省百分比
10	880	1000	12.0%
20	7100	8000	11.1%
30	24667	27000	7.5%
40	58400	64000	8.8%
50	112290	125000	11.7%

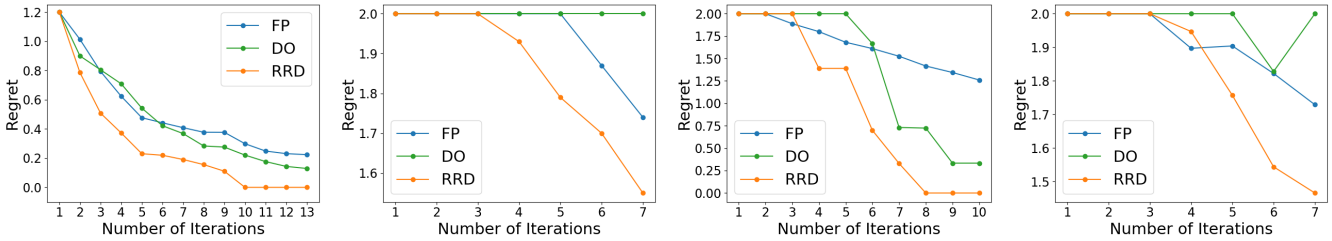
表1 三人博弈中因 BPS 而节省的配置文件模拟开销。

攻击图博弈是在攻击图上定义的双人一般和博弈：攻击者试图通过入侵一系列节点到达目标节点，而防御者则努力保护任意节点（例如，拒绝访问）。在有限时域内到达目标节点，会给攻击者带来巨大收益，并使防御者遭受重大损失。攻击和防御行动都需要付出成本。攻击者可以选择任意可行节点子集，防御者也可以防御任意节点子集，因此行动空间的组合规模很大。

在图5中，我们展示了 RRD 在一个包含100个节点、因此具有 2^{100} 种可能组合动作的攻击图博弈实例上的性能。由于该博弈规模过大，无法进行穷举分析，我们首先按照 Czarnecki 等人 [2020]提出的策略采样方法，构建一组特定的深度Q网络（DQN） [Mnih 等人。，总计包含125种具有策略多样性的策略。随后，我们对这组策略应用博弈论分析（即 FP、双重预言机和 RRD）。每条遗憾曲线均为5个随机选取的初始策略的平均值。从图5中可以观察到，尽管所研究的博弈规模较大且超出了双人零和博弈的范畴，RRD 仍能比双重预言机和 FP 更快地实现收敛，并具有更小的方差。

讨价还价博弈

我们研究一种非零和不完全信息讨价还价游戏，其中两名玩家通过序贯过程，就物品向量 [Lewis 等人，2017]展开谈判以达成协议。在我们的设定中，共有三件物品，最多进行 $T = 10$轮报价。每件物品的价值均从双方共同知晓的分布中抽取，而每名玩家只知道自己的价值实现。如果在时间 [达成协议，



(a) 六角棋。 (b) 四子棋。 (c) 反向井字棋。 (d) 围棋（尺寸=4）。

图4 RRD 与 FP 和双重预言机在 Czarnecki等人 [2020] 研究的四种博弈中的性能比较。

玩家将获得所指定物品的相应值，并按因子 $\gamma^t(\gamma = 0.9)$进行折扣。如果未能达成协议，双方玩家均获得零效用。我们使用神经网络表示谈判策略，并采用深度Q网络计算近似最佳响应。

在图6中，我们将社会福利（SW）作为性能指标，并比较由不同 MSS（即 RRD、双重预言机、虚拟博弈、MWCCE 和 MGCCE）在经验博弈中得到的相同解概念的平均社会福利。每个 SW 值均为 5 次运行的平均值，标准差见附录 B.3。我们选取五种解概念进行性能比较，包括最大社会福利纯策略组合（最大社会福利）、纳什均衡、策略上的均匀分布、MWCCE 和 MGCCE。从图6中可以观察到，RRD 生成的全部五种解的社会福利都高于其他方法。这一观察结果还表明，用于逼近某个解（例如纳什均衡、MWCCE 和 MGCCE）的最佳 MSS，可能并不是直接将该解概念用作最佳响应目标的 MSS。

5.2 遗憾阈值变化时的稳定性

为了研究学习性能相对于遗憾阈值 λ 的稳定性，我们为 RRD 选取了广泛的 λ，并将 PSRO 在这些λ下最后一次迭代时的遗憾值，与双人 Leduc 扑克中双重预言机的遗憾值进行比较。我们在图7a中绘出了这些遗憾值。从图7a可以看出，范围内所有λ的学习性能都优于双重预言机，这说明 RRD 的性能相对于遗憾阈值 λ 具有稳定性。此外，我们观察到，随着遗憾阈值 λ 的值增大，学习性能先提升，随后变差。这意味着，过度或不足的正则化都会损害整体学习性能。

6 解释 MSS 的性能

早期关于策略探索的研究认识到，针对精确纳什均衡进行最佳响应并不理想，并证明了通过采用其他 MSS 或其他方法，可以取得优于双重预言机的效果。然而，迄今为止，尚无令人满意的解释说明什么样的最佳响应目标才有效。我们先简要讨论一些早期研究，再根据实验观测提出一个新的见解，以帮助解释为什么正则化在这一情境下会有所帮助。

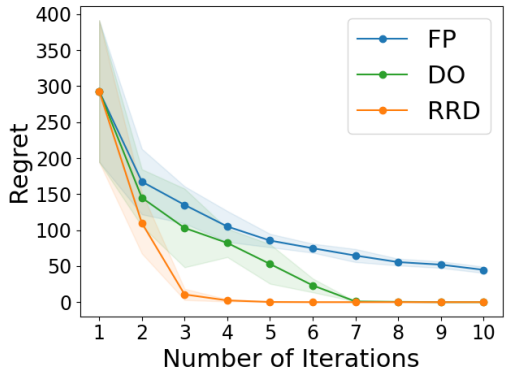


图5 RRD 在攻击图博弈中优于虚拟博弈和双重预言机。

6.1 与先前正则化相关的方法

在一项早期的策略探索研究中，Schvartzman 和 Wellman [2009a] 发现，向纳什均衡中加入噪声会改变均衡搜索的路径，并加快整体学习速度。但这种噪声为何能加快学习，尚未得到解释。如上所述，PSRO 的提出者专门引入了 PRD，以解决对纳什均衡过拟合的问题 [Lanctot等人， 2017]。PRD 通过确保经验博弈中的每种策略在最佳响应目标中都具有固定的最低概率，来促进探索。出于类似的考虑，Wang 等人 [2019] 提出随机交替使用虚拟博弈和双重预言机，而 Wright 等人 [2019] 则通过针对先前的对手进行调优，调整了对纳什均衡的最佳响应。

与这一研究方向相关，Balduzzi等人 [2019] 引入了博弈景观这一术语，用以描述经验博弈所涵盖的概念性策略空间。他们提出了一种名为修正纳什均衡的 MSS，旨在从定性上扩展博弈景观。然而，博弈景观的概念主要通过一个特定的简单博弈示例来说明，虽然颇具启发性，却没有提供操作性定义或度量。

6.2 一种新颖的解释

我们的关键洞见是，策略探索的表现与最佳响应目标相对于完整博弈的遗憾值密切相关。为展示这一现象，图 7b 给出了双人 Leduc扑克中采用 RRD 的 PSRO 的遗憾曲线。这两条曲线分别展示了纳什均衡和 RRD 的真实博弈遗憾

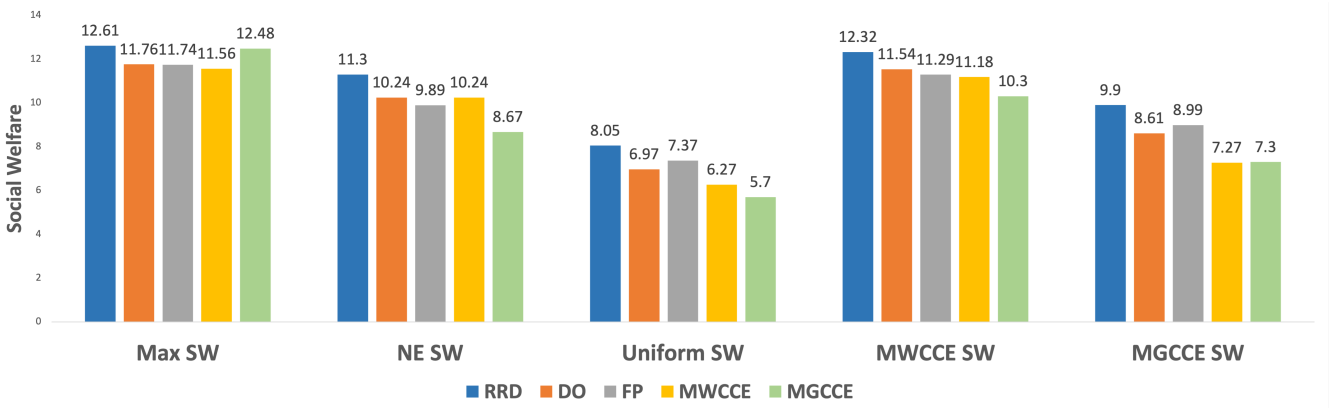
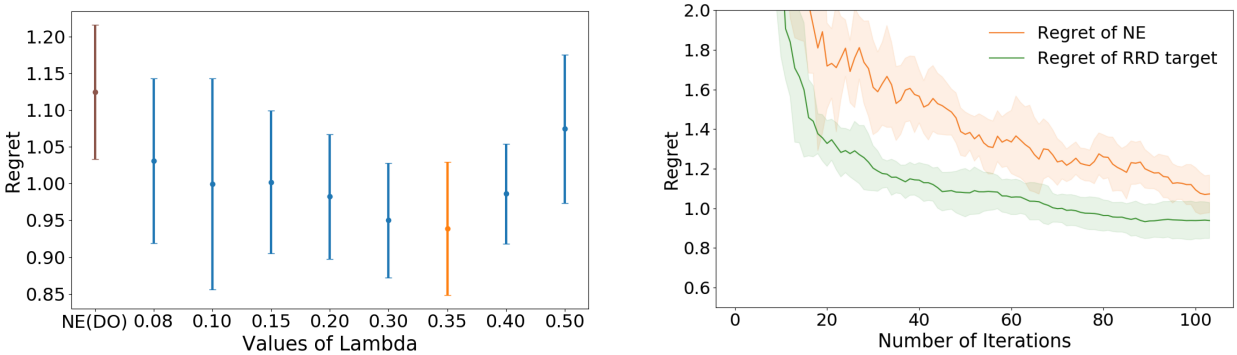


图6 讨价还价博弈中的 RRD 性能。每种颜色代表一个 MSS，每组颜色表示相应经验博弈中某一解概念的 SW。



（a）遗憾阈值的范围。　　（b）应用正则化后遗憾值降低。

图7 双人 Leduc 扑克中使用 RRD 学习的性质。

，这些遗憾值均在每次 PSRO 迭代时计算。请注意，在整个运行过程中，RRD 解的遗憾值远低于经验纳什均衡。我们在由其他 MSS 生成的 PSRO 运行中也观察到了同样的现象（见附录 B.4）。换言之，尽管 RRD 在经验博弈中的遗憾值高于纳什均衡（λ 与零相比），但它在完整博弈中的遗憾值始终更低。由于我们的最终目标是在完整博弈中获得低遗憾值解，这有助于解释为何 RRD 所施加的正则化显然能稳定地提升策略探索的表现。

请注意，这一观察的适用范围有限；最小化完整博弈遗憾值并不总能为策略探索提供最优的最佳响应目标。这是因为，完整博弈中遗憾值最低的策略组合在相邻的 PSRO 迭代之间可能变化不大，因此据此选择目标可能会损害所构建经验博弈的多样性。我们通过明确的实验对此进行了检验：将经验博弈中相对于完整博弈遗憾值最低的策略组合用作 MSS。结果证实，这种极端的目标选择确实不利于策略探索（详情见附录 C.2）。

7 结论

我们提出 RRD，作为一种基于正则化、适用于 PSRO 的新型 MSS。通过控制遗憾阈值，可以调整正则化程度，以适应特定的策略探索情境。在实验中，我们展示了 RRD 在多种博弈中的表现优于若干现有 MSS，并研究了使用 RRD 进行学习的诸多特性。为扩展至双人以上博弈，我们提出了 BPS——一种与 PSRO 兼容的策略组合搜索方法，可避免对博弈矩阵进行穷举仿真。我们展示了在三人 Leduc扑克中将 BPS 与 RRD 结合时正则化带来的益处。最后，我们证明，策略探索的表现与最佳响应目标的遗憾值密切相关，而正则化可以显著降低最佳响应目标的遗憾值，从而促进学习效果的提升。

参考文献

[Balduzzi等人， 2019] David Balduzzi、Marta Garnelo、Yoram Bachrach、Wojciech M Czarnecki、Julien Perolat、Max Jaderberg 和 Thore Graepel。《开放式学习》

对称零和博弈。载于第36届国际机器学习大会，页码 434–443，2019年。[Brinkman和Wellman， 2016] Erik Brinkman 和 Michael P. Wellman。《限价订单市场中的着色与效率》。载于 IJCAI-16算法博弈论研讨会，2016年。[Brown， 1951] George W. Brown。《通过虚拟博弈迭代求解博弈》。载于 T. C. Koopmans 编，《生产与分配的活动分析》，页码 374–376。Wiley，1951年。[Czarnecki 等人， 2020] Wojciech Marian Czarnecki、Gauthier Gidel、Brendan Tracey、Karl Tuyls、Shayegan Omidshafiei、David Balduzzi 和 Max Jaderberg。《现实世界博弈看起来像陀螺》。载于第34届神经信息处理系统大会，2020年。[Dinh等人， 2022] Le Cong Dinh、Stephen Marcus McAleer、Zheng Tian、Nicolas Perez-Nieves、Oliver Slumbers、David Henry Mguni、Jun Wang、Haitham Bou Ammar、Yaodong 和 Yang。《在线双重预言机》。机器学习研究汇刊，2022年10月。[Fearnley等人， 2015] John Fearnley、Martin Gairing、Paul Goldberg 和 Rahul Savani。《通过收益查询学习博弈均衡》。机器学习研究杂志，16:1305–1344，2015年。[Jordan 等人， 2008] Patrick R. Jordan、Yevgeniy Vorobeychik 和 Michael P. Wellman。《在经验博弈中搜索近似均衡》。载于第7届自主智能体与多智能体系统国际会议，页码 1063–1070，2008年。[Jordan 等人， 2010] Patrick R. Jordan、L. Julian Schvartzman 和 Michael P. Wellman。《经验博弈中的策略探索》。载于第9届自主智能体与多智能体系统国际会议，页码 1131–1138，2010年。[Lanctot 等人， 2017] Marc Lanctot、Vinicius Zambaldi、Audr¯unas Gruslys、Angeliki Lazaridou、Karl Tuyls、Julien P´erolat、David Silver 和 Thore Graepel。《多智能体强化学习的统一博弈论方法》。载于第31届神经信息处理系统年会，页码 4190–4203，2017年。[Lanctot 等人， 2019] Marc Lanctot、Edward Lockhart、Jean-Baptiste Lespiau、Vinicius Zambaldi、Satyaki Upadhyay、Julien P´erolat、Sriram Srinivasan、Finbarr Timbers、Karl Tuyls、Shayegan Omidshafiei 等。《OpenSpiel：用于博弈强化学习的框架》。arXiv 预印本 arXiv:1908.09453，2019年。[Lewis等人， 2017] Mike Lewis、Denis Yarats、Yann N Dauphin、Devi Parikh 和 Dhruv Batra。《成交还是不成交？谈判对话的端到端学习》。arXiv 预印本 arXiv:1706.05125，2017年。[Marris等人， 2021] Luke Marris、Paul Muller、Marc Lanctot、Karl Tuyls 和 Thore Graepel。《通过相关均衡元求解器实现超越零和博弈的多智能体训练》。载于第38届国际机器学习大会，页码 7480–7491，2021年。[McAleer 等人， 2021] Stephen McAleer、John Lanier、Pierre Baldi 和 Roy Fox。CFR-DO：一种用于扩展式博弈的双重预言机算法。载于《AAAI-21游戏强化学习研讨会》，2021年。[McKelvey and Palfrey， 1995] Richard D. McKelvey 和 Thomas R. Palfrey。标准式博弈的分位数响应均衡。《博弈与经济行为》，10(1)：6–38，1995年。[McKelvey and Palfrey， 1998] Richard D. McKelvey 和 Thomas R. Palfrey。扩展式博弈的分位数响应均衡。《实验经济学》，1(1)：9–41，1998年。[McKelvey 等人， 2006] Richard D. McKelvey、Andrew M. McLennan 和 Theodore L. Turocy。Gambit：博弈论软件工具，2006年。[McMahan 等人， 2003] H. Brendan McMahan、Geoffrey J. Gordon 和 Avrim Blum。对手控制成本函数时的规划。载于第20届国际机器学习会议，第536–543页，2003年。[Miehling等人， 2015] Erik Miehling、Mohammad Rasouli 和 Demosthenis Teneketzis。贝叶斯攻击图上部分可观测传播过程的最优防御策略。载于第2届ACM移动目标防御研讨会，第67–76页，2015年。[Mnih等人， 2013] Volodymyr Mnih、Koray Kavukcuoglu、David Silver、Alex Graves、Ioannis Antonoglou、Daan Wierstra 和 Martin Riedmiller。使用深度强化学习玩 Atari游戏。arXiv预印本 arXiv:1312.5602，2013年。[Muller等人， 2020] Paul Muller、Shayegan Omidshafiei、Mark Rowland、Karl Tuyls、Julien Perolat、Siqi Liu、Daniel Hennes、Luke Marris、Marc Lanctot、Edward Hughes 等人。多智能体学习的通用训练方法。载于第8届国际学习表征会议，2020年。[Omidshafiei等人， 2019] Shayegan Omidshafiei、Christos Papadimitriou、Georgios Piliouras、Karl Tuyls、Mark Rowland、Jean-Baptiste Lespiau、Wojciech M. Czarnecki、Marc Lanctot、Julien Perolat 和 Remi Munos。 α-排名：通过演化进行多智能体评估。《科学报告》，9(1)：1–29，2019年。[Phelps 等人， 2006] S. Phelps、M. Marcinkiewicz、S. Parsons 和 P. McBurney。一种在 n人非零和博弈中自动获取策略的新方法。载于第5届国际自主智能体与多智能体系统联合会议，第705–712页，2006年。[Schvartzman和Wellman，2009a年] L. Julian Schvartzman 和 Michael P. Wellman。在经验博弈建模中探索大型策略空间。载于AAMAS-09智能体介导电子商务研讨会，2009年。[Schvartzman和Wellman，2009b年] L. Julian Schvartzman 和 Michael P. Wellman。通过经验博弈论分析和强化学习改进CDA策略。载于第8届国际自主智能体与多智能体系统会议，第249–256页，2009年。

[Smith and Price, 1973] J. Maynard Smith 和 George R. Price。动物冲突的逻辑。《自然》，246(5427)：15–18，1973年。[Sureka和Wurman, 2005] Ashish Sureka 和 Peter R. Wurman。使用禁忌最佳响应搜索，在标准式博弈中寻找纯策略纳什均衡。载于第*4*届国际自主智能体与多智能体系统联合会议，第1023–1029页，2005年。[Taylor and Jonker, 1978] Peter D. Taylor 和 Leo B. Jonker。进化稳定策略与博弈动态。《数学生物科学》，40(1-2)：145–156，1978年。[Tuyls 等人, 2020] Karl Tuyls、Julien P´erolat、Marc Lanctot、Edward Hughes、Richard Everett、Joel Z. Leibo、Csaba Szepesv´ari 和 Thore Graepel。经验博弈论分析的界限与动态。自主智能体与多智能体系统，34：7，2020年。[Wang等人, 2019] Yufei Wang、Zheyuan Ryan Shi、Lantao Yu、Yi Wu、Rohit Singh、Lucas Joppa 和 Fei Fang。利用实时信息，在绿色安全博弈中应用深度强化学习。载于第*33*届美国人工智能协会人工智能会议，第1401–1408页，2019年。[Wang等人, 2022] Yongzhao Wang、Qiurui Ma 和 Michael P. Wellman。在经验博弈论分析中评估策略探索。载于第*23*届自主智能体与多智能体系统国际会议，第1346–1354页，2022年。[Wellman, 2016] Michael P. Wellman。将智能体纳入基于智能体的建模。自主智能体与多智能体系统，30：1175–1189，2016年。[Wright等人, 2019] Mason Wright、Yongzhao Wang 和 Michael P. Wellman。博弈中的迭代深度强化学习：通过历史感知训练提升稳定性。载于第*20*届*ACM*经济学与计算会议，第617–636页，2019年。

缩略语表

缩写	定义
BPS	反向策略组合搜索
DO	双重预言机
EGTA	经验博弈论分析
FP	虚拟博弈
MSS	元策略求解器
MRCP	最小遗憾约束剖面
NE	纳什均衡
PRD	投影复制子动态
PSRO	策略空间响应预言机
QRE	量化响应均衡
RD	复制动态
RL	强化学习
RRD	正则化复制动态

表 2 按字母顺序排列的缩略语表。

B 额外实验结果

B.1 在固定迭代次数下相较于复制子动力学的优势

在图 8 中，我们展示了达到遗憾阈值 λ 所需的遗憾匹配更新次数。我们将迭代次数的下限固定为 10000。如图 8 所示，达到遗憾阈值 λ 所需的遗憾匹配更新次数会随PSRO迭代而变化，这再次凸显了动态调整遗憾匹配更新次数的必要性，而非将其固定不变。此外，有趣的是，学习过程处于中期时，迭代次数较高，而在开始和结束时则较低。这与一种刻板印象相矛盾：在策略更多样化的经验博弈中，复制子动力学需要更多迭代才能收敛。

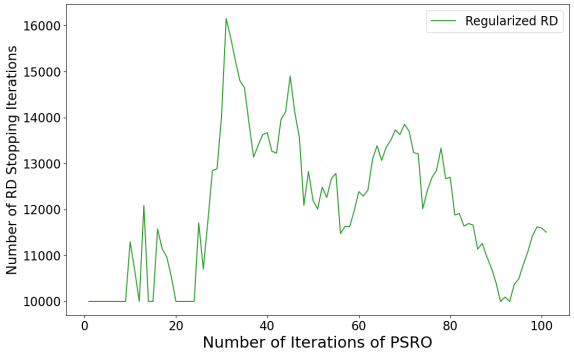


图 8 在不同 PSRO迭代下达到遗憾阈值 λ 所需的遗憾匹配更新次数。

B.2 博弈描述

现实世界博弈实例

我们实验中的现实博弈，是 Czarnecki 等人 [2020] 从完整博弈中提炼出的博弈模型。博弈描述如下。

- 六角棋是一种双人棋盘游戏，玩家试图连接六边形棋盘的两侧。
- 四子棋是一种双人连接类棋盘游戏，玩家选择一种颜色，然后轮流将彩色棋子落入棋盘。游戏目标是率先用己方棋子连成一条水平、垂直或对角线上的四子连线。
- 反向井字棋是井字棋的一种变体，只有当对手连成一线而自己没有连成一线时，玩家才算获胜。
- 4×4 围棋是一种棋盘比标准围棋更小的围棋。

现实世界游戏的实验细节

我们按照 Wang等人 [2022], 提出的一致性指标评估所有 MSS，该指标用于衡量中间博弈模型的质量。在所有实验中，每位玩家在开始时都会获得一组表现较差的策略。所有 MSS 使用相同的初始策略集。由于这些博弈均表示为矩阵博弈，最佳响应计算和遗憾计算都是确定性的，因此不报告误差条。对于 RRD，我们在所有实验中都设置了遗憾阈值$\lambda = 0.5$ 。

讨价还价博弈的实验细节

我们采用 Lewis等人 [2017]提出的讨价还价博弈模型。每位玩家都会获得一个随机生成的价值函数，该函数为每件物品赋予非负值。价值函数受到以下约束：（1）对每位玩家而言，所有物品的总价值为 10；（2）每件物品对至少一位玩家的价值非零；（3）有些物品对双方玩家的价值都非零。这些约束确保双方玩家不可能同时获得最高分，而且没有任何物品对双方都毫无价值，因此谈判将具有竞争性。经过 10 轮后，玩家可以在未达成协议的情况下结束谈判，此时双方玩家均得 0 分。对于 RRD，我们在所有实验中都设置了遗憾阈值 $\lambda = 0.3$ 。

攻击图博弈的实验细节

我们遵循 Miehling等人 [2015]的攻击图模型。如正文所述，攻击图由一个有向无环图 $\mathcal{G} = (V, E)$ 给出，其中顶点 $v \in V$ 表示安全条件，边 $e \in E$ 则标记为利用方式。攻击图博弈由一个附加了其他规范的攻击图定义。该博弈是一种双人部分可观测随机博弈，在有限个时间步 T上进行。在每个时间步 t，博弈状态由图状态给出，即节点是否处于激活状态（即是否已被攻击者攻陷），由 $s_t(v) \in \{0, 1\}$ 表示。假设攻击者能够完全观测图状态，而防御者则会收到关于该状态的噪声观测 $o_t(v) \in \{0, 1\}$ ，其依据是

每个节点 v对应的公认概率 $P_v(o = 0 \mid s = 1)$ 和$P_v(o = 1 \mid s = 0)$。正观测称为警报。

攻击者和防御者在每个时间步 t 同时行动。防御者的原子动作是防御一个节点，因此，原子动作就是 V。防御者在任意时间步的动作空间为 2^V，这意味着它可以选择防御节点的任意子集。攻击者的原子动作集合会随时间变化，并取决于当前的图状态。只有当边的起始节点处于激活状态时，该边上的利用方式才可行。没有父节点的节点称为根节点，可以在没有前提条件的情况下受到攻击。对于攻击 *AND*节点这一特殊情况，攻击者的原子动作被视为攻击一个节点，而不是选择所有入边上的利用方式。因此，攻击者的原子动作可以理解为：选择边（可行的利用方式）来攻击 *OR* 节点，或从父节点全部处于激活状态的 *AND* 节点中选择节点。攻击者在任意时间步的动作空间是可行原子动作的幂集。

防御者动作优先于攻击者动作，也就是说，任何被防御的节点 v 都会变为非活跃状态。否则，活跃节点保持活跃；任何遭受攻击的 与 节点 v 会以概率 $P(v)$变为活跃状态，而任何 或 节点则根据被攻击边的成功概率 $P(e)$变为活跃状态。

每个目标节点 v在其处于活跃状态的所有时间步中，都会给攻击者带来奖励 $R_A(v)$，并给防御者带来惩罚 $P_D(v)$。任何智能体的原子动作 a 都具有成本：对于防御者而言，防御节点的成本为 $c_{a,\ D}(v)$；对于攻击者而言，选择与 节点的成本为 $c_{a,\ A}(v)$，选择边的成本为 $c_{a,\ A}(e)$。为简单起见，我们在记号中省略动作 a的参数（v 或 e）。当上下文清楚时，我们也省略下标 D 和 A，仅使用 v_a 和 c_a 分别表示 a 的目标节点和动作 a 的成本。

防御者在任意时间步的损失（负奖励）是其所有原子动作的成本（即被防御节点的总成本），加上行动后处于活跃状态的目标节点所带来的惩罚。防御者的长期收益是随时间推移的损失的折扣期望和。类似地，攻击者的长期收益是每个时间步收益的折扣期望和，其中每个时间步的收益是活跃目标节点的奖励减去该时间步所使用原子动作的成本。

任一玩家的策略（博弈中的纯策略）都会将其在任意步骤的观测映射为一组行动。对于攻击者，该映射是从状态到行动集合的映射。防御者只能部分观测状态，因此其策略会将观测历史映射为行动集合。在我们的实现中，为了便于求解，我们将防御者的过去观测长度限制为固定的 h。求解博弈意味着找到一对混合策略（纯策略上的概率分布），每位玩家各一个，使其构成纳什均衡。

B.3 讨价还价博弈中的社会福利标准差

在表3中，我们列出了不同解的社会福利标准差。这对应于正文中的图6。

MSSs	最大社会福利	NE	均匀分布	MWCCE	MGCCE
RRD	0.60	1.14	1.40	0.73	1.05
DO	0.73	1.11	1.39	0.77	0.98
FP	0.80	1.58	1.63	1.01	1.53
MWCCE	0.67	1.11	1.66	0.70	2.49
MGCCE	1.00	1.78	1.60	1.11	1.88

讨价还价博弈标准差表。

B.4 使用纳什均衡的 PSRO 中通过正则化降低遗憾值

在图9中，我们在双人 Leduc 扑克中使用 NE 运行 PSRO，并绘制了 RRD 给出的最佳响应目标相对于真实博弈的遗憾值，以及中间经验博弈中 NE 相对于真实博弈的遗憾值。从图9可以看出，与使用 RRD 的 PSRO（图7b）一样，对于给定的中间经验博弈，应用正则化后，最佳响应目标的遗憾值再次降低。

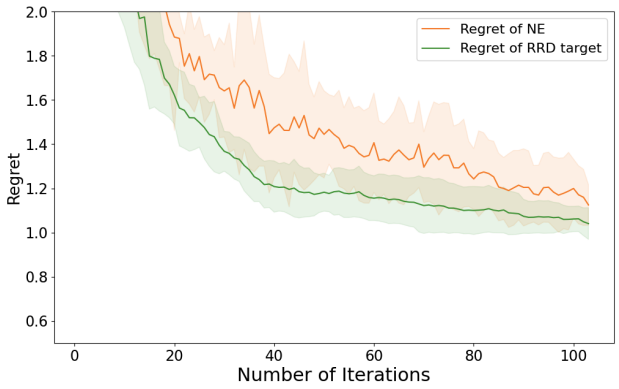


图 9 在以纳什均衡作为 MSS 的 PSRO 运行中应用正则化后遗憾值降低。

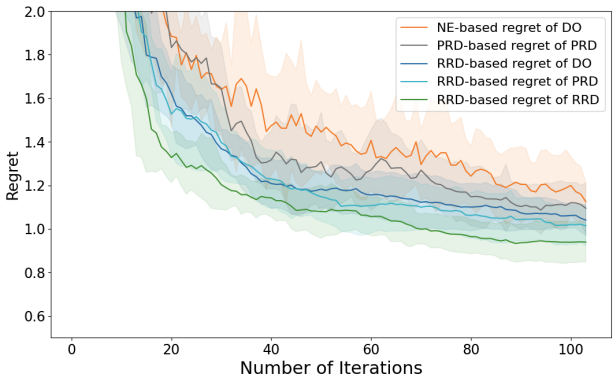
B.5 使用一致性评估求解器进行性能检查

在图 10 中，我们遵循一致性规则 [Wang 等人，2022], 这是一项衡量所生成经验博弈质量的关键指标。根据一致性准则，我们比较具有相同 RRD 遗憾值的 MSS，并证实在所生成经验博弈的质量方面，RRD 的提升速度快于双重预言机和 PRD。

B.6 实验参数

我们在 Leduc 上的实验中使用 OpenSpiel Lanctot等人 [2019] 的默认参数集：经验博弈中的每个收益条目均为 1000 次重复模拟的平均值；我们采用深度Q网络作为最佳响应预言机，其参数见表 4。扑克博弈是不对称的，因为总是有一名玩家先行动。

PRD 的实现采用策略概率下界 1e-10、最大步数 1e5 和步长 1e-3。RD 使用相同的步长，但步数可变



Wang等人 [2022]提出的一致性度量下，RRD 在双层 Leduc 扑克中的表现，图10。

参数	值
学习率	1e-2
批量大小	32
经验回放缓冲区大小	1e4
回合数	1e4
优化器	adam
层大小	256
层数	4
初始探索率	1
最终探索率	0.1
探索衰减时长	3e6
折扣因子	0.999
网络更新步数	10
目标网络更新步数	500

表 4 深度Q网络超参数

由遗憾阈值 λ控制的步数。我们测试了 λ 从 0 到 0.6 时 RRD 的学习性能。在 Leduc扑克中，当 $\lambda = 0.35$ 时，学习性能最佳。在三人Leduc扑克中，我们进行了 $\lambda = 0.6$实验。

节点数	100
成本	攻击者的均匀分布范围为 [0, 1]；防御者的均匀分布范围为 [2, 4]
奖励	均匀分布范围为 [10, 20]
惩罚	在 [7, 10]中均匀分布
#目标节点	6
激活概率	在 [0.6, 1] 中均匀分布
误报概率	[0, 0 中均匀分布。2]

表 5 攻击图博弈的超参数。

B.7 代码可用性

根据双盲政策，我们将在评审结束后公开代码。

B.8 计算资源

我们使用了一颗 2×3.0 GHz 英特尔至强金牌 6154 中央处理器，配备 16GB 内存。

C 使用其他新颖 MSS 的 PSRO

C.1 量子响应均衡作为 MSS 时的结果

博弈论分析中的一个常见假设是玩家具有理性（即玩家按照纳什均衡行事）。由于我们的正则化方法在一定程度上阻止玩家在经验博弈中采取纳什均衡，因此可以将其视为限制玩家理性的一种方式，这也使我们的方法自然地与量化响应均衡（QRE）联系起来[McKelvey和Palfrey, 1995年；McKelvey和Palfrey， 1998],一种有限理性的均衡概念。QRE 的一种常见形式是 Logit 均衡，在这种均衡中，玩家的策略形式如下

$$\sigma_i(s_i) = \frac{exp(\tau u_i(s_i, \sigma_{-i}))}{\sum_{s_i' \in S_i} exp(\tau u_i(s_i', \sigma_{-i}))}, \forall s_i \in S_i, i \in N.$$

其中， τ 是一个用于控制玩家理性的参数。

为了考察量子响应均衡作为 MSS 的表现，我们评估了采用量子响应均衡的 PSRO 的学习表现。具体而言，我们使用 Gambit [McKelvey 等人，2006] 在每次 PSRO迭代中计算经验博弈的量子响应均衡，并分析其学习表现。

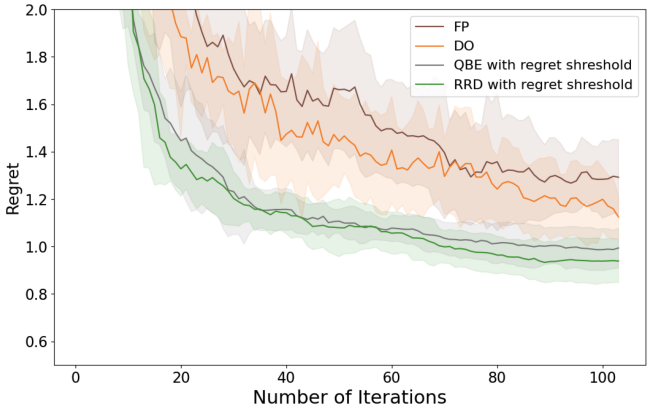


图 11 双层 Leduc 扑克中采用量子响应均衡的 PRSO 的表现。

图 11 展示了在 Leduc扑克中将量子响应均衡作为 MSS 时的学习表现。作为对比，我们还绘制了采用与量子响应均衡相同遗憾阈值的 RRD 学习曲线。尽管量子响应均衡在最后略有发散，但仍展现了在 PSRO 中将量子响应均衡用作 MSS 的潜力。

C.2 将 MRCP 用作 MRCP 的 MSS 定义的结果

经验博弈中最接近完整博弈解的策略组合称为最小遗憾约束策略组合 （MRCP） [Jordan 等人，2010]。形式上，$\bar{\sigma}$ 是 MRCP，当且仅当：

$$\bar{\sigma} = \underset{\sigma \in \Delta(X)}{\operatorname{argmin}} \sum_{i \in N} \rho_i^{\mathcal{G}}(\sigma)$$

因此，MRCP 的遗憾值为衡量 X 对策略相关空间 [的覆盖程度提供了一个自然指标 Wang 等，2022]。

使用 MRCP 进行学习的实验

我们观察到，在经验博弈中存在全局遗憾值低于纳什均衡的策略组合；正则化的实验结果表明，与这些策略组合进行训练，相较于双重预言机，能够带来更好的学习性能。一个自然的问题是：与最稳定的策略组合目标进行训练，是否能最大程度地促进策略探索（例如，使用 MRCP 作为 MSS）。

为回答这一问题，我们在矩阵形式的双人库恩扑克和一个合成双人零和博弈中，将 MRCP 作为 MSS，与 DO 和 FP 的表现进行比较。在库恩扑克中，我们随机选择 4 个起始点并实施 PSRO。图 12a–12d 显示，在 4 个起始点中的 3 个起始点上，MRCP 的收敛速度略快于 DO。对于矩阵博弈，图 12e 和 12f 展示了应用 MRCP 的好处，但其表现因起始点不同而有所差异。

从图 12 可以看出，MRCP 在启发式策略生成方面具有一定作用。然而，从收敛速度和计算复杂度来看，使用 MRCP 的优势并不理想。我们还发现，在布洛托等其他博弈中，使用 MRCP 的收敛速度可能慢于 DO 和 PRD。

实验表明，针对经验博弈中遗憾值最低的策略组合进行训练，并不一定能带来更好的整体学习表现。这是因为向经验博弈中加入新策略后，其中遗憾值最低的策略组合可能不会发生太大变化，因此在 PSRO迭代过程中会持续加入相似的策略。持续加入相似策略，在 PSRO迭代过程中只能带来有限的性能提升。

现在，我们通过表7所示的一个矩阵博弈来说明，为什么追求遗憾值极低的最佳响应目标可能导致学习速度缓慢。该矩阵博弈为每个玩家包含1000个策略。收益矩阵中所有缺失的条目均为(0，0)。让我们从第一个策略(s_1，s_1)开始PSRO。该矩阵博弈旨在使双重预言机的均衡搜索路径很长（许多现实世界的博弈也具有这一特征）。具体而言，通过对(s_1，s_1)进行最佳响应，每个玩家都会向经验博弈中添加$s2$，其新的纳什均衡为(s_2，s_2)。类似地，如果我们在每次PSRO迭代中都对均衡进行最佳响应，首先会得到一个新的纳什均衡(s_3，s_3)，随后沿着对角线经历一条很长的均衡路径，直到到达完整博弈的纳什均衡(s_{1000}，s_{1000})。

不失一般性地，假设我们处于第2次迭代（即经验博弈包含(s_1，s_2)，且(s_2，s_2)是一个经验纳什均衡）。该经验（对称）博弈的MRCP约为(1$s1$，0$s2$)，遗憾值为0.0112=0.022（对玩家求和）（精确MRCP的遗憾值甚至更低）。经验纳什均衡的遗憾值为 $0.1 * 2 = 0.2$，因为从(s_2，s_2)偏离到 s_3。当对MRCP进行最佳响应时，我们添加 s_{500}（仅考虑经验博弈之外的偏离策略），此后经验博弈的MRCP保持不变（即(1s_1，0s_2)）。因此，继续对MRCP进行最佳响应可能会再次添加一些类似于 s_{500} 的策略，并且可能无法显著改善学习性能。

假设 RRD 在 (s_1，s_2) 上给出的概率为 (0.5，0.5)，那么对 ($0.5s_1$，$0.5s_2$) 作最佳响应会直接导向均衡策略 —— 策，跳出双重预言机漫长的均衡路径。($0.5s_1$，$0.5s_2$) 的遗憾值为 $(0.005 \times 0.5 + 0.199 \times 0.5 - 0.011 \times 0.25 - 0.1 * 0.25) \times 2 = (0.102 - 0.02775) \times 2 = 0.074252 = 0.1485$ (0.02（MRCP 的遗憾值）$<$ 0.1485（RRD 的遗憾值）$<$ 0.2（纳什均衡的遗憾值）)。

在本例中，针对遗憾值相对较低的完整博弈遗憾策略组合进行最佳响应时，RRD 不会像双重预言机那样陷入漫长的对角线轨迹。与此同时，它的遗憾值并没有 MRCP 的遗憾值那么低，因此每次 PSRO 迭代都会显著更新最佳响应目标，而不会一直保持相似。

使用 MRCP 学习的其他性质

理论上，经验博弈中可能存在多个 MRCP，而且一般而言，MRCP 不一定是纯策略组合。此外，单纯将 MRCP 用作 MSS 并不能保证收敛到纳什均衡，因为最佳响应策略可能已经存在于经验博弈中。我们将 MRCP 的这一性质定义如下。

定义。 若策略空间为 $X \subseteq S$ 的经验博弈关于 ***MRCP*** 是封闭的，则

$$\forall i \in N, s_i = \underset{s_i' \in S_i}{\operatorname{argmax}}\, u_i(s_i', \sigma_{-i}) \in X_i.$$

为说明这一概念，考虑表 6 中的对称零和矩阵博弈。从每个玩家的第一个策略开始，使用 MRCP 实施 PSRO，我们得到包含 a^1 和 a^2 的经验博弈。由于 (a_1^1，a_2^1) 是一个 MRCP（考虑所有纯策略和混合策略组合），且针对该策略组合进行最佳响应会再次得到 a^2，因此该经验博弈是封闭的，永远不会扩展到真实博弈；在真实博弈中，真实纳什均衡为 (a_1^3，a_2^3)。

	a_2^1	a_2^2	a_2^3
a_1^1	(0, 0) [2]	(-1, 1) [6]	(-0.5, 0.5)
a_1^2	(1, -1) [6]	(0, 0) [10]	(-5, 5)
a_1^3	(0.5, -0.5)	(5, -5)	(0, 0)

表 6 MRCP 的对称零和博弈。策略组合的遗憾值显示在方括号中。

在我们的实验中，我们通过只引入经验博弈之外偏离收益最高的新策略来处理这一问题，从而保证收敛。另一种方法是在出现这一问题时切换 DO 和 MRCP；由于 DO 的收敛性质，这样也能保证收敛。

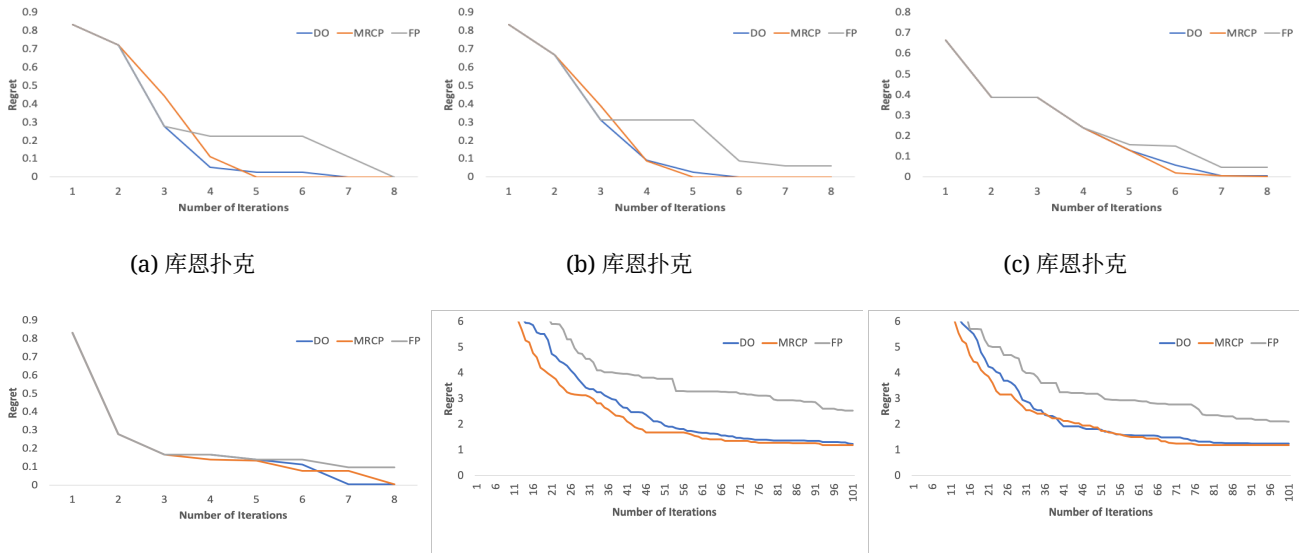
D RRD 的收敛

D.1 收敛性证明

我们首先定义 ϵ-接近 的概念，它可以视为 PSRO 的停止条件。

定义 (ϵ-接近)。 策略空间为 $X \subseteq S$ 的经验博弈，关于某个 ϵ-纳什均衡 $\sigma \in \Delta(X)$ 和算子 o，当且仅当 $o(\sigma) \in X$ 时为 ϵ-闭合。

例如，如果 o 是最佳响应算子且 $\epsilon = 0$，则此定义意味着不存在从



(a) 库恩扑克 (b) 库恩扑克 (c) 库恩扑克

(d) 库恩扑克 (e) 零和博弈 (f) 零和博弈

图12 将 MRCP 用作 MSS 的性能。纵轴表示基于 MRCP 的遗憾值。

经验博弈的纳什均衡 σ 出发的有利偏离，因此 σ 是完整博弈的纳什均衡。当 $\epsilon \neq 0$时， ϵ-接近度表示经验博弈的 ϵ-纳什均衡 σ 的偏离策略已存在于经验博弈中。请注意，对于给定的 ϵ，经验博弈中可能存在无限多个 ϵ-纳什均衡，因此 ϵ-接近度的定义与特定的 ϵ-纳什均衡相关联。

接下来我们证明，若某个经验博弈相对于某个 ϵ-纳什均衡 $\sigma \in \Delta(X)$ 和最佳响应算子o是 ϵ-闭合的，那么 σ 就是完整博弈的一个 ϵ-纳什均衡。

引理1。 若一个策略空间为 $X \subseteq S$的经验博弈相对于某个 ϵ-纳什均衡 $\sigma \in \Delta(X)$ 和最佳响应算子 o是 ϵ-闭合的，那么 σ 就是完整博弈 $\mathcal{G}$的一个 ϵ-纳什均衡。

证明。 由于 σ 是经验博弈中的一个 ϵ-纳什均衡，因此在经验博弈中不存在会导致遗憾值大于 ϵ的偏离策略。数学上，我们有 $\forall i \in N, \max_{s_i' \in X_i} u_i(s_i', \sigma_{-i}) - u_i(\sigma_i, \sigma_{-i}) \leq \epsilon$。由于最佳响应算子针对真实博弈寻找最佳偏离，且该最佳偏离落在经验博弈中，因此我们有$\forall i \in N, \max_{s_i' \in S_i} u_i(s_i', \sigma_{-i}) - u_i(\sigma_i, \sigma_{-i}) \leq \epsilon$。于是， σ 就是完整博弈 $\mathcal{G}$ 的一个 ϵ- 纳什均衡。 □

考虑有限策略空间 S。我们证明以下定理：如果在 PSRO 的每次迭代中，我们针对经验博弈的一个 ϵ-NE 进行训练，那么最终得到的经验博弈至少包含一个 ϵ-NE。通过将 ϵ 设为一个可达的 λ，我们证明了定理1。

定理1。 假设可以访问精确的最佳响应预言机，则采用与可达遗憾阈值 λ 相关联的正则化复制动态的策略空间响应预言机将收敛到一个至少包含一个 λ-*NE* 的经验博弈。

证明。 由于我们拥有有限策略空间，针对特定 σ 的 S、 ϵ 接近度始终可以通过训练达到在每次迭代中都相对于 ϵ-纳什均衡进行训练。一旦达到 ϵ 接近度，根据引理1，相应的 σ 就是完整博弈中的 ϵ-纳什均衡。由于 PSRO [的种群性质，可能还存在遗憾值低于 ϵ 的 2022]，策略组合。 □

这一结果将双重预言机及其收敛保证推广到了训练目标不局限于 NE 的情形。值得注意的是，早期研究 [Dinh等人，2022年；McAleer 等人， 2021]，也证明了类似结果。然而，这些研究忽略了经验博弈的一个基本事实：经验博弈会生成一个策略剖面空间，其中可能存在遗憾值远小于目标策略组合的策略组合。这一事实揭示了使用博弈模型促进博弈学习的一项主要优势：借助博弈模型，我们只需在策略空间中找到一个 NE，而不像在线学习那样要求一系列策略组合收敛到 NE。

D.2 性能混淆

一个常见的疑问是：既然 RRD 的理论收敛保证弱于双重预言机，为什么 RRD 仍然有用？需要注意，McMahan等人 [2003] 对双重预言机收敛性的证明，描述的是它是否收敛，而非收敛速度。在最坏情况下，双重预言机需要将所有策略（假设博弈是有限的）都加入经验博弈，才能实现收敛，这显然只是一个平凡结论。在我们关注的博弈中（例如 Leduc扑克），策略空间通常非常庞大，无法枚举。尽管双重预言机等算法能够保证收敛到纳什均衡，但这些算法可能需要探索大量策略才能实现收敛，远超计算预算。策略探索的目标是在预算范围内尽可能学到接近纳什均衡的策略。因此，策略探索更关注在计算预算内学习 MSS 的表现，而

不太关注预算之外的长期收敛。尽管 RRD 在极限情况下的收敛保证较弱，但它能够通过控制正则化程度，在预算范围内快速学到稳定的策略组合，这符合策略探索的理念。如果追求绝对收敛，可以在 RRD 快速生成有用策略后，于某次 PSRO迭代将 RRD 切换为双重预言机。

	a_2^1	a_2^2	a_2^3	...	a_2^{500}	...	a_2^{1000}
a_1^1	(0, 0)	(0, 0.011)	(0, 0)	...	(0. 0.01)	...	(0,0.005)
a_1^2	(0.011, 0)	(0.1, 0.1)	(0.1, 0.2)	...	...	...	(0,0.199)
a_1^3	(0, 0)	(0.2, 0.1)	(0.2, 0.2)	...	...	...	(0, 0)
...	...	...	...	...	...	...	...
a_1^{500}	(0.01, 0)	...	...	...	...	...	(0, 0)
...	...	...	...	...	...	...	...
a_1^{1000}	(0.005, 0)	(0.199, 0)	(0, 0)	...	(0, 0)	...	(100, 100)

表7 一个具有较长纳什均衡路径的博弈实例。