控制与决策
Control and Decision

基于可配置CFR的海上基地防护安全博弈策略求解

罗俊仁, 张万鹏, 谷学强, 陈璟

引用本文:
罗俊仁, 张万鹏, 谷学强, 等. 基于可配置CFR的海上基地防护安全博弈策略求解[J]. 控制与决策, 2025, 40(8): 2503-2512.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2024.1172

基于可配置 CFR 的海上基地防护安全博弈策略求解

罗俊仁，张万鹏，谷学强，陈 璟†

(国防科技大学 智能科学学院，长沙 410073)

摘 要: 围绕海上基地的攻防可看作一个多阶段序贯对抗过程, 通常可建模为不完美信息零和博弈. 针对海上基地防护安全博弈问题, 构建不完美信息序贯博弈模型, 分析博弈模型各要素; 围绕近似纳什均衡策略的快速求解, 提出可配置反事实遗憾最小化 (CogCFR) 算法, 利用基类 CFR 算法与元控制器可动态控制 CFR 的超参数; 以海上多个海上基地防护为试验背景, 利用 CogCFR 求解海上基地防护资源分配策略. 针对有限理性对手, 提出考虑约束的单侧信任域鲁棒对手利用策略更新方式. 实验结果表明: 可配置反事实遗憾最小化相比动态加权反事实遗憾最小化计算时效性更强、参数更少; 算法具有较好的应用可行性和领域泛化性, 可为序贯交互类博弈对抗问题策略求解提供参考.

关键词: 海上基地；安全博弈；资源分配；反事实遗憾最小化；可配置

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2024.1172

引用格式: 罗俊仁, 张万鹏, 谷学强, 等. 基于可配置 CFR 的海上基地防护安全博弈策略求解 [J]. 控制与决策, 2025, 40(8): 2503-2512.

Configurable CFR for strategy solving of security game in maritime base protectioon

LUO Jun-ren，*ZHANG Wan-peng*，*GU Xue-qiang*，*CHEN Jing*†

(College of Intelligence Science and Technology，National University of Defense Technology，Changsha 410073，China)

Abstract: The offensive and counterattack around the maritime island can be regarded as a multi-stage sequential counterattack process, which is usually modeled as a zero-sum game with imperfect information, the game model elements are analysized. Aiming at the security game problem of maritime base protection, the imperfect information sequential game model is constructed. A configurable counterfactual regret (CogCFR) minimization algorithm based on base CFR variants and meta-controller is proposed to solve the approximate Nash equilibrium strategy quickly, which can dynamically control the CFR hyperparameters. For the experimental background of multiple maritime island protection, the CogCFR minimization algorithm is used to solve the resource allocation strategy of maritime island protection. This paper presents a robust opponent exploitation strategy updating method with unilateral trust region considering constraints for bounded rational opponent. The experimental results show that the CogCFR minimization with meta-learning is more efficient and has fewer parameters than the dynamic weighted CFR minimization. The algorithm has good application feasibility and domain generalization, and can provide reference for strategy solving of sequential interactive game.

Keywords: maritime base；security game；resource allocation；counterfactual regret minimization；configurable

0 引 言

随着国际贸易的扩大, 为守护海上利益, 亟需采用定点及伴随等方式保障海上基地 (争端岛屿、油气田、潜在交战区、运输保障及补给点等利益相关区) 安全. 近年来, 美军陆续提出了远征前进基地作战[1]、分布式海上作战[2]、马赛克战[3] 等作战概念. 此外, 各国均在持续加快水下侦察预警、海基一体化防空反导等系统的研发. 其中, 水下侦察预警系统主要包括潜艇、无人潜航器、水下预置系统、反潜直升机等, 主要用于增强海上基地水下防护能力. 海基一体化

收稿日期: 2024-10-04；**录用日期:** 2025-03-17.
基金项目: 国家自然科学基金项目 (61702528).
责任编委: 刘德荣.
†通信作者. E-mail: chenjing001@vip.sina.com.

防空反导系统主要负责海上基地利益区海空目标的打击. 在考虑可能遭受威胁的基础上, 有效分配海上基地防护资源, 对于维护海上利益十分关键.

未来, 海上对抗面临高复杂环境、不完美信息、强博弈对抗、高动态响应等挑战. 现实世界中大多数交互对抗情境均可建模成不完美信息博弈(imperfect information games, IIGs): 局中人持有私有信息, 同时也可能存在任何局中人均不知晓的信息. 近年来, 一些研究聚焦采用博弈理论研究海上冲突对抗中资源分配问题. 其中, Keith 等[4] 构建了一体化网络和防空系统资源分配问题的扩展式博弈模型, 并利用反事实遗憾最小化 (counterfactual regret minimization, CFR)[5] 求解博弈的纳什均衡和鲁棒响应策略. Ganzfried 等[6] 构建了热点冲突海域策略规划随机博弈模型, 分别利用值迭代虚拟对弈和策略迭代虚拟对弈求解近似纳什均衡策略. Luo 等[7] 综合分析网络上布洛托上校博弈模型, 设计了基于策略空间响应预言机的策略求解算法. McCarthy[8] 构建了军事力量设计和作战行动一体的随机博弈模型, 利用 Shapley 值迭代算法消除劣势行动求解博弈的纯策略纳什均衡. Du 等[9] 利用区间直觉模糊 TOPSIS 法研究了海上基地占位问题. Zeng 等[10] 构建了海上基地攻防博弈模型, 利用 CFR 算法求解纳什均衡策略. 此外, 安全博弈 (security game)[11] 作为一类描述序贯交互过程的博弈模型, 常被用于毒品禁运、武器走私、非法贸易、野生动物保护、林业保护、城市犯罪和网络防护等安全领域资源分配问题的建模.

随着人工智能技术在星际争霸 AI (Alphastar)[12] 和德州扑克 AI (Pluribus)[13] 上的成功, 博弈策略求解算法也取得了长足的发展. 围绕博弈纳什均衡策略求解[14], 即没有局中人能够通过偏离该均衡来提升自身状况的策略, 不完美信息博弈[15] 的策略求解算法主要包括从线性规划、虚拟自对弈 (fictitious self play, FSP)、反事实遗憾最小化 (CFR)、在线凸优化、策略空间响应预言机 (policy space response oracle, PSRO) 和博弈强化学习等. 主要可分为基于博弈理论的 CFR 类、基于强化学习的值函数估计类和基于在线凸优化的 OMD 类. 本文主要聚焦 CFR 类算法, 该类算法借助迭代流程来降低双方的遗憾, 逐步引导每位局中人的时间平均策略或末轮迭代策略趋向于纳什均衡策略. 其中, CFR+ [16] 的提出是一项重要的里程碑, 在实践中其收敛速率较基础 CFR 快了一个数量级. 具体而言, CFR+交替更新双方的策略, 运用“遗憾匹配+ (regret matching +, RM+)”作为其遗憾最小化器, 而非“遗憾匹配(RM)”, 采用线性折扣方案, 将迭代次数t对平均策略的贡献依据t进行加权. 这一改进在求解两人有限德州扑克问题时发挥了关键作用[17]. CFR 类算法及变体构成了该领域各类算法的对比测试基准[17–19]. 当前一些研究主要聚焦利用若干创新改良技巧来获得更快的收敛速度. Brown 等[18] 提出“折扣 CFR (Discount CFR, DCFR)”算法, 与经典 CFR 对每个迭代予以统一加权有所不同, DCFR 有效性源自于采用的折扣计算方案, 这些方案为后期迭代赋予了更大的权重. DCFR 通过运用 3 个超参数α(调控正累计遗憾)、β(调控负累计遗憾) 和γ(调控累计策略) 为后期迭代赋予更大权重, 进而将折扣应用于遗憾值与平均策略. Farina 等[19]提出了“预测性 CFR+ (Predictive CFR+, PCFR+)”, 采用预测性 Blackwell 可达性理论, 引入在线凸优化技术来加速算法迭代收敛. 尽管这些 CFR 变体借助折扣实现了加速收敛, 但它们对固定折扣方案的依赖极大地限制了其潜能. 为减轻 CFR 类型算法在手动算法设计方面所面临的挑战, Xu 等[20] 提出了 AutoCFR, 引入一种新颖的方式, 采用搜索语言系统地探索 CFR 类型算法的广阔空间, 并通过进化过程学习新颖的变体. Zhang 等[21] 提出贪婪加权算法, 采用基于运行时观测动态调整迭代权重的遗憾最小化算法, 但该研究主要聚焦于正则式博弈近似均衡求解, 在扩展式博弈策略求解问题上表现欠佳. Xu 等[22]提出了动态折扣 CFR(dynamic DCFR, DDCFR) 算法, 采用强化学习框架来获取性能优良的折扣方案, 可在运行时动态调整 DCFR 的超参数. 虽然该算法在多个基准测试博弈求解问题上验证了有效性, 但进行实时折扣权重计算时需要进行特征计算、策略训练以及网络推断, 这会产生额外的计算成本. 此外, 它还假定在训练游戏中习得的具有可变超参数的策略能够良好地泛化至感兴趣的目标游戏中.

总体来看, 传统的安全博弈通常采用两阶段或三阶段 Stackelberg 博弈或多智能体随机博弈模型, 无法很好地描述攻防双方序贯交互对抗的过程. 此前的 CFR 类算法超参数不好调控, 算法泛化性差. 鉴于此, 本文首先分析海上基地防护面临的问题, 构建两方多阶段攻防对抗的不完美信息序贯博弈模型; 其次, 针对多阶段博弈策略求解问题, 设计可配置反事实遗憾最小化 (configurable CFR, CogCFR) 算法进行博弈求解, 使用更为简单的动态折扣方案生成算法控制 CFR 类算法的超参数 (α、β和γ), 并与其他 CFR 类算法变体进行对比分析; 最后, 针对有限

理性对手, 提出对手利用的鲁棒响应策略求解算法.

本文贡献主要有以下 4 点: 1) 提出了利用博弈理论来研究海上基地跨域防空问题; 2) 构建了面向海上基地跨域防空的不完美信息博弈模型; 3) 设计了可配置反事实遗憾最小化算法求解博弈策略; 4) 给出了应对有限理性对手的鲁棒响应策略求解方式.

1　海上基地防护安全博弈分析建模

1.1　多阶段攻防场景分析

本文采用多阶段交互来描述双方的攻防对抗过程, 攻击方是先手方, 防御方是后手方. 防御方的目标是采用部署反水下渗透装备设施和一体化防空系统保护海上基地, 最大化守护海上基地范围内安全; 攻击方的目标是采用侦察破坏和物理打击等措施摧毁海上基地, 最小化海上基地防护地域面积. 典型海上基地防护攻防对抗场景如图 1 所示.

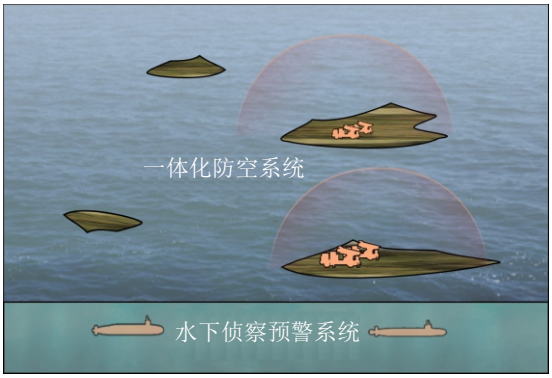

图1　典型海上基地防护攻防对抗场景

将敌我双方的对抗过程看作是有 4 个攻防双方博弈对抗阶段的攻防过程, 将双方围绕侦察破坏与反渗透的攻防阶段放置在物理打击攻防阶段之中:

1) 防御方的物理部署阶段, 即部署一体化防空系统;

2) 攻击方的侦察破坏水下攻击阶段, 即利用水下无人潜航器、潜艇、蛙人等对防御方基地实施水下渗透侦察;

3) 防御方的反渗透水下部署阶段, 即部署水下声纳、无人潜航器、调用反潜直升机等实施反潜;

4) 攻击方物理攻击阶段, 即利用陆基火炮或水面舰艇发射导弹对防御方实施攻击.

考虑由n套部署在各个海上基地上的防空导弹组成的一体化防空系统, 多阶段博弈对抗的防御方, 其目标不仅是保卫海上基地不受敌方侦察渗透破坏, 还需反制可能受到的敌物理攻击 (火炮或导弹). 攻击方在执行物理攻击前, 通常会采用侦察渗透破坏防御方部署的物理防御设施. 本文中, 防空系统按每个防空单元来分析, 相关参数定义如表 1 所示, 部分参数采用“归一化”定义. 整个博弈过程如表 2 所示.

表1　相关参数设置

符号	名称	约束
N	海上基地数量	无
D	一体化防空系统部署位置	无
r	每套一体化防空系统的覆盖半径	$(0,1]$
n_d	防空系统单元数	$n_d < N$
n_a	受物理空袭基地数	$n_a < N$
m_d	有反渗透防御能力海上基地数	$m_d < n_d$
m_a	反渗透防御能力脆弱基地数	$m_a < n_a$
n_{cd}	反侦察破坏分队数量	$n_{cd} < m_d$
n_{ca}	具备侦察破坏攻击分队数	$n_{ca} < m_a$
p	有效物理防御概率	$(0,1]$
p_s	反渗透检测概率	$(0,1]$
p_d	有效反渗透防御概率	$(0,1]$
p_a	有效侦察破坏攻击概率	$(0,1]$

表2　多阶段博弈对抗过程

阶段	形态	局中人	动作空间	行动
1	物理防御	防御方	$C_N^{n_d}$	部署防空系统
2	反渗透	自然	$C_{n_d}^{m_d}$	选定反渗透基地
3	侦察破坏	自然	$C_{n_d}^{m_a}$	选定侦察破坏基地
4	侦察破坏	攻击方	$C_{m_d}^{n_{ca}}$	侦察破坏部分基地
5	反渗透	自然	$2^{n_{ca}}$	检测侦察破坏有效性
6	反渗透	防御方	$C_{m_d}^{n_{cd}}$	为防空系统指派防护队
7	物理打击	攻击方	$C_N^{n_a}$	攻击部分海上基地

1.2　不完美信息序贯博弈

海上基地多阶段攻防对抗过程可以建模成一类典型有随机因素 (自然人) 影响的不完美信息序贯博弈模型. 攻防双方的多阶段博弈对抗过程可以采用信息不完美条件下的序贯交互来刻画.

定义 1　不完美信息序贯博弈可表示为$G = \langle P, H, A, I, \{u_i\}\rangle$.

定义 1 中: 有限个局中人P之间的博弈交互过程可以表示为树结构, 树节点表示博弈状态, 连边表示每个局中人$i \in P$的动作, 包括自然 (随机) 节点c. 基于当前节点的历史信息 (包含私有信息)$h \in H$, 局中人从可采动作集合$A(h)$中选择动作$a \in A$. 历史信息序列得到更新$h' \sqsubset h$, 然后轮到下一个局中人$P(h)$行动, 当到达叶子节点$z \in Z$时, 每个局中人获得回报$\{u_i\}$. 在二人零和博弈中, 满足$u_1(z)+u_2(z)=0$. 每个局中人的不完美信息主要体现在信息集 (information set) 上. 在局中人i拥有的信息集I_i中, 所有节点$h, h' \in I_i$对局中人i而言都是不可分辨的. 因此, 每个节点$h \in H$都可划分至具体的信息集I_i中. 由此可以将$P(h)$和$A(h)$都重新定义至信息

集I_i上得到$P(I_i)$和$A(I_i)$. 假定局中人拥有完美回忆, 不会遗忘此前任何信息.

1.3 博弈模型要素分析

博弈双方在进行攻防对抗过程中, $A_{i,j}$表示局中人i在过程j的行为集, 对应$I_{i,j}$表示局中人i在过程j的信息集. 采用布尔覆盖矩阵表示防空系统对海上基地的覆盖关系, 其中$c_{i,j}$表示在i位置的防空系统是否覆盖了j海上基地, 即在半径r以内. 使用s_i表示海上基地i的面积, 采用v_i表示海上基地的价值, 攻防双方的策略表示为$\sigma=(\sigma_1,\sigma_2)$.

假设物理域和网络域防御概率统计独立, 则海上基地i受防空单元j保护的概率$p_{i,j}$可表示为

$$p_{i,j}=\begin{cases}0,\ c_{i,j}=0;\\ 0,\ j\in a_c(\sigma),\ j\notin d_c(\sigma);\\ p,\ j\notin a_c(\sigma);\\ pp_d(1-p_a),\ j\in a_c(\sigma),\ j\in d_c(\sigma).\end{cases}\tag{1}$$

其中: $a_c(\sigma)$和$d_c(\sigma)$分别为策略σ下遭敌攻击和受网络防护队防护的防空系统, p_a和p_d分别为遭敌攻击和受网络防护队防护的概率.

零和博弈中, 攻击方的收益值为防御方收益值的相反数, 防御方的收益值可表示为

$$u_1(\sigma)=-u_2(\sigma)=-\sum_{i\in a_p(\sigma)}v_i\prod_{j\in d_p(\sigma)}(1-p_{i,j}).\tag{2}$$

其中: $p_{i,j}$为海上基地i受防空单元j保护的概率, $a_p(\sigma)$为策略σ中受攻击的海上基地集合, $d_p(\sigma)$为策略σ中有防空系统的海上基地.

防御方的平均收益值可表示为

$$E_{(\sigma_1,\sigma_2)}\Big[-\sum_{i\in a_p(\sigma_1,\sigma_2)}v_i\prod_{j\in d_p(\sigma_1,\sigma_2)}(1-\phi(\sigma_1,\sigma_2))\Big],\tag{3}$$

其中$\phi(\sigma_1,\sigma_2)$为攻防双方纯策略组合.

2 基于可配置 CFR 资源分配策略求解

2.1 纳什均衡

给定博弈模型, 局中人的目标是寻求一个最大化期望值的策略. 然而, 策略的期望值最终取决于对手的策略 (或策略集). 当对手策略未知时, “保守”的方法是计算近似纳什均衡解 (Nash equlibrium, NE).

定义 2(纳什均衡解) $\sigma^*=(\sigma_i,\sigma_{-i})$是由每个局中人的最佳响应组成的策略剖面 (strategy profile), 即

$$\forall i,u_i(\sigma_i^*,\sigma_{-i}^*)=\max_{\sigma_i'}u_i(\sigma_i',\sigma_{-i}^*).$$

在两人零和博弈中, 纳什均衡解是一个不败策略, 并且解的存在性具有理论证明, 能够在多项式时间内求解. 任何局中人偏离纳什均衡解都不会得到更大的利益, 反而会被对方利用.

在纳什均衡解概念中, 双方都采用均衡策略, 任何偏离均衡策略的一方所获得的收益将减小. 对于非完全理性的对手, 其策略与纳什均衡策略之间的“距离”, 为其他均衡策略局中人创造了可利用度 (exploitability). 在求解 (近似) 纳什均衡解时, 可利用度是均衡策略质量的衡量标准, 指该策略在预期中相对于最坏情况下的对手策略所达到的少于博弈价值的量. 通常也将这个差值称为一个策略的可利用度.

定义 3(可利用度) 两人零和博弈中, 策略σ_i的可利用度 (exploitability) 表示为

$$e(\sigma_i)=u_i(\sigma_i^*,BR(\sigma_i^*))-u_i(\sigma_i,BR(\sigma_i)).$$

当观察到对手持续偏离纳什均衡解时, 可以结合对手可利用度, 利用对手的缺点变换己方的策略, 而不是持续使用均衡策略.

2.2 反事实遗憾最小化

对于状态空间小的不完美信息博弈问题, 可以转化成优化问题 (如混合整数线性规划或混合整数非线性规划). 但随着信息集数量和平均信息集大小的增长, 优化问题难以在多项式时间内求解. 一些一阶 (first-order) 方法可以以$O(1/\epsilon)$收敛到纳什均衡, 结果比反事实遗憾最小化算法更接近纳什均衡, 但该类方法扩展性比较差, 难以直接推广应用.

在一些大型博弈中, 一阶方法的收敛速度不如一些 CFR 变体, 而且需要精细调参. CFR 是目前最先进的能够在大型不完美信息博弈中生成高效策略的技术之一, 是一种在两人零和博弈中收敛到纳什均衡的迭代算法, 保证在两人零和博弈中计算出近似纳什均衡策略.

2.2.1 遗憾最小化

遗憾是一个在线学习概念, 用来度量局中人执行除实际动作以外的动作获得的收益. 在单局中人重复博弈场景中, 局中人每个回合需在$\|A\|$个动作中选择一个. 第t次迭代, 定义每个动作分配的回报为$v^t(a)$, 该回报局中人不可观, 且每轮迭代可变. 局中人通过选择一个动作上的概率分布σ^t, 获得期望回报$v^t=\sum_{a\in A}\sigma^t(a)v^t(a)$.

定义 4(遗憾) 用于衡量局中人已选择的决策序列$(\sigma^1,\sigma^2,\dots,\sigma^T)$与过去“可行却未行”的序列$s\in S$的收益差值, 即根据对过去博弈中行为动作的遗憾程度来调整将来的动作选择策略. 遗憾定义如下:

$$R^T(s_a)=\sum_{t=1}^{T}(v^t(a)-v^t).$$

将其扩展到多人平均意义下每局局中人所选择的策略在与T轮迭代中收益最大策略的收益差值R_i^T, 有

$$R_i^T=\max_{\sigma_i'}\sum_{t=1}^{T}(v_i(\sigma_i',\sigma_{-i}^t)-v_i(\sigma^t)).$$

定义 5(平均遗憾)　σ_i^t为局中人i在第t回合的博弈中所采取的策略, 则局中人i在T次博弈中的平均遗憾为

$$R_i^T=\frac{1}{T}\max_{\sigma_i'\in\Sigma_i}\sum_{t=1}^{T}(u_i(\sigma_i',\sigma_{-i}^t)-u_i(\sigma^t)).$$

如果产生某一局中人的策略使得当$T\to\infty$时, $R_i^{T,+}/T\to 0$, 其中$R_i^{T,+}=\max\{R_i^T,0\}$表示遗憾都为非负, 则称该策略是遗憾最小的.

以遗憾匹配 (regret matching) 为代表的遗憾最小化方法, 适用于规范化博弈 (normal formal game). 遗憾匹配依据一个正比于正遗憾的动作概率分布, 用随机的方式选择动作. 对于可选动作集A_i中的每一个动作a, 存储该动作的每轮迭代计算得到的遗憾, 在接下来的第$T+1$轮迭代中, 更新策略计算如下:

$$\sigma_i^{T+1}(a)=\frac{R_i^{T,+}(a)}{\sum_{b\in A_i}R_i^{T,+}(b)},$$

其中动作a的遗憾表示在过去T轮博弈中, 局中人i没有采取该动作而产生的累加遗憾.

2.2.2　反事实遗憾最小化

对于大型序贯博弈, 要计算并最小化平均整体遗憾R_i^T是不现实的. 反事实遗憾最小化算法的基本思想是将整体遗憾分解为一组可独立信息集, 在每个独立的信息集上引入反事实遗憾的概念, 通过不断迭代最小化每个信息集上的反事实遗憾从而最小化平均整体遗憾, 此时得到的平均策略达到近似纳什均衡.

定义 6(平均策略)　对于每个信息集$I\in L_i$, 每个动作$a\in A(I)$, $\bar{\sigma}_i^t$为局中人i从第 1 次到第T次博弈的平均策略, 有

$$\bar{\sigma}_i^t(I)(a)=\frac{\sum_{t=1}^{T}\pi_i^{\sigma^t}(I)\sigma^t(I)(a)}{\sum_{t=1}^{T}\pi_i^{\sigma^t}(I)}.$$

在两人零和博弈中, 若两个局中人的平均整体遗憾小于ϵ, 则平均策略遵循2ϵ纳什均衡.

在扩展博弈中, 博弈中间点的收益未知, CFR 算法正是通过定义中间节点的遗憾, 从而可以在每个信息集上使用遗憾匹配算法. 首先考虑一个指定信息集$I\in L_i$和局中人i在该信息集上的选择, 定义$u_i(\sigma,h)$为所有局中人遵循策略σ进行博弈并到达了动作序列h时的期望效益值, 定义反事实效益值 (counterfactual utility) 为在除了局中人以外的所有局中人都遵循策略σ而局中人i的策略被改为故意到达信息集I的情况下, 进行博弈并到达信息集I的期望效益值. $\pi^{\sigma}(h,h')$是在策略组下, 从动作序列h到h'的转移概率. 则有

$$u_i(\sigma,I)_{=}\frac{\sum_{h\in I,h'\in Z}\pi_{-i}^{\sigma}(h)\pi^{\sigma}(h,h')u_i(h')}{\pi_{-i}^{\sigma}(I)}.$$

对于所有的$a\in A(I)$, 定义$\sigma|_{I\to a}$为除了在信息集I上总是选择动作a以外, 其余与σ完全相同的策略. 则即时反事实遗憾 (immediate counterfactual regret) 为

$$R_{i,\mathrm{imm}}^T(I)=\frac{1}{T}\max_{a\in A(I)}\sum_{t=1}^{T}\pi_{-i}^{\sigma^t}(I)(u_i(\sigma^t|_{I\to a},I)-u_i(\sigma^t,I)).\quad(4)$$

直观上, 这是局中人i在信息集I上所做决策的遗憾, 由动作a的反事实效益值与信息集I的反事实效益值之间的差值, 再乘上一个表示局中人i试图到达信息集I的反事实概率作为加权项.

寻找博弈中近似纳什均衡的关键点是如何最小化每个信息集上的即时反事实遗憾. 即时反事实遗憾的关键特征是可以通过仅控制$\sigma_i(I)$来最小化它, 对于所有的$I\in L_i$和$a\in A(I)$, 有

$$R_i^T(I,a)=\frac{1}{T}\sum_{t=1}^{T}\pi_{-i}^{\sigma^t}(I)(u_i(\sigma^t|_{I\to a},I)-u_i(\sigma^t,I)).$$

$T+1$轮的策略为

$$\sigma_i^{T+1}(a)=\begin{cases}\dfrac{R_i^{T,+}(I,a)}{\sum_{a\in A(I)}R_i^{T,+}(I,a)}, & \sum_{a\in A(I)}R_i^{T,+}(I,a)>0;\\ \dfrac{1}{|A(I)|}, & \text{otherwise.}\end{cases}\quad(5)$$

式 (5) 揭示了反事实遗憾与更新下一轮策略动作概率分布之间的关系, 即每种动作被选择的概率与其在过去已经进行的T轮博弈中没被选择而产生的累加的正遗憾成正比, 若不存在正遗憾, 则给每个

动作分配均等的概率.

基础 CFR 算法在现有条件下只能解决10^{12}规模的博弈问题, 局限在于局中人i根据式 (5) 选择动作, 则 $R^T_{i,\mathrm{imm}}(I)\leqslant \Delta_{u,i}\sqrt{|A_i|}/\sqrt{T}$, 从而有 $R^T_i(I)\leqslant \Delta_{u,i}|L_i|\sqrt{|A_i|}/\sqrt{T}$, 其中 $|A_i|=\max_{h:p(h)=i}|A(h)|$. 平均整体遗憾的上界与信息集的数量成线性关系, 这成为原始 CFR 算法的瓶颈之处.

2.2.3 蒙特卡洛反事实遗憾最小化

蒙特卡洛反事实遗憾最小化 (Monte Carlo CFR, MCCFR)[23] 是一种采样类 CFR 算法变体, 通过引入反事实遗憾的样本估计, 有效估计策略更新中必要部分的遗憾. 在大型博弈中能在短期快速收敛, 但长期可能因为累积方差 (variance) 增高导致收敛变慢.

MCCFR 的衍生算法中两个具有代表性的是基于结果抽样的蒙特卡洛反事实遗憾最小化算法 (OS-MCCFR) 和基于外部抽样的蒙特卡洛反事实遗憾最小化算法 (ES-MCCFR), 前者在每次迭代中只抽样一次博弈, 后者抽样机会节点和对手的行为.

2.3 CogCFR 算法及鲁棒响应

2.3.1 基类 CFR 算法与元控制器

当前各类 CFR 算法的性能对比一致性不强, 如何利用基类算法随时间自适应调整可变参数是一种可行方法. 与利用演化策略[20] 或强化学习[22] 来优化超参数选择方法不同, 本文提出采用元控制器 (meta-controller) 来调控基类 CFR 算法, 无需提前配置和预训练. 本文超参数选择主要采用区间数进行限制, 针对不可微分评价指标“可利用度”, 本文超参数优化主要采用零阶优化方式. CogCFR 算法结构如图 2 所示.

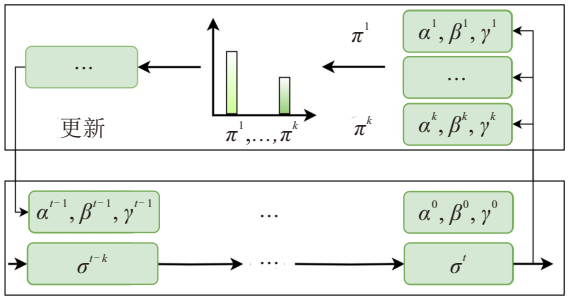

图2 CogCFR 算法结构

本文采用的多个重要的基类 CFR 算法如表 3 所示. 对于指数 CFR 算法[24], 其参数为

$$\alpha(I,a)=\exp\Big(r_i(I,a)-\frac{\sum\limits_{a'\in A(I)}r_i(I,a')}{|A(I)|}\Big). \tag{6}$$

定理 1 若$\alpha\in[1,8]$, $\beta\in[-8,8]$, $\gamma\in[8,40]$, 则对于两人零和博弈, 采用 CogCFR 算法迭代T后所得平均策略为$\dfrac{9}{\sqrt{T}}\Delta\Big(\dfrac{8}{3}\sqrt{|A|}+\dfrac{2}{\sqrt{T}}\Big)$纳什均衡.

表3 不同 CFR 基类算法的计算方式

算法	累积遗憾 R_i^t	局部策略更新 σ_i^{t+1}	累积策略 C^t
CFR	$R_i^{t-1}(I,a)+r_i^t(I,a)$	$[R_i^t(I,a)]^+\Big/\sum\limits_{a'\in A(I)}[R_i^t(I,a')]^+$	$C^{t-1}(I,a)+\pi_i^{\sigma^t}(I)\sigma_i^t(I,a)$
CFR+	$[R_i^{t-1}(I,a)+r_i^t(I,a)]^+$	$R_i^t(I,a)\Big/\sum\limits_{a'\in A(I)}R_i^t(I,a')$	$C^{t-1}(I,a)+t\pi_i^{\sigma^t}(I)\sigma_i^t(I,a)$
LCFR[18]	$R_i^{t-1}(I,a)+t*r_i^t(I,a)$	$[R_i^t(I,a)]^+\Big/\sum\limits_{a'\in A(I)}[R_i^t(I,a')]^+$	$C^{t-1}(I,a)+t\pi_i^{\sigma^t}(I)\sigma_i^t(I,a)$
ECFR[24]	$R_i^t(I,a)=R_i^{t-1}(I,a)+\alpha(I,a)*r_i^t(I,a),\text{if } r_i^t(I,a)>0$ $R_i^t(I,a)=R_i^{t-1}(I,a)+\alpha(I,a)*\beta,$其它	$[R_i^t(I,a)*\alpha(I,a)]^+\Big/\sum\limits_{a'\in A(I)}[R_i^t(I,a')\alpha(I,a)]^+$	$C^{t-1}(I,a)+\pi_i^{\sigma^t}(I)\sigma_i^t(I,a)*\alpha(I,a)$
DCFR[18]	$R_i^t(I,a)=R_i^{t-1}(I,a)\odot d_i^{t-1}+r_i^t(I,a),$ $d_i^t[a]=\begin{cases}\dfrac{t^\alpha}{t^\alpha+1}, & R_i^t[a]>0\\ \dfrac{t^\beta}{t^\beta+1}, & \text{otherwise}\end{cases}$	$[R_i^t(I,a)]^+\Big/\sum\limits_{a'\in A(I)}[R_i^t(I,a')]^+$	$C^{t-1}(I,a)\Big(\dfrac{t-1}{t}\Big)^\gamma+\pi_i^{\sigma^t}(I)\sigma_i^t(I,a)$
DCFR+[18]	$\Big[R_i^{t-1}(I,a)\dfrac{(t-1)^\alpha}{(t-1)^\alpha+1}+r_i^t(I,a)\Big]^+$	$R_i^t(I,a)\Big/\sum\limits_{a'\in A(I)}R_i^t(I,a')$	$C^{t-1}(I,a)\Big(\dfrac{t-1}{t}\Big)^\gamma+\pi_i^{\sigma^t}(I)\sigma_i^t(I,a)$
PCFR+[19]	$[R_i^{t-1}(I,a)+r_i^t(I,a)]^+$	$\tilde{R}_i^{t+1}(I,a)\Big/\sum\limits_{a'\in A(I)}\tilde{R}_i^{t+1}(I,a'),$其中 $\tilde{R}_i^{t+1}(I,a)=[R_i^t(I,a)+v_i^{t+1}(I,a)]^+$	$C^{t-1}(I,a)+t^2\pi_i^{\sigma^t}(I)\sigma_i^t(I,a)$
PDCFR+[19]	$\Big[R_i^{t-1}(I,a)\dfrac{(t-1)^\alpha}{(t-1)^\alpha+1}+r_i^t(I,a)\Big]^+$	$\tilde{R}_i^{t+1}(I,a)\Big/\sum\limits_{a'\in A(I)}\tilde{R}_i^{t+1}(I,a'),$其中 $\tilde{R}_i^{t+1}(I,a)=\Big[R_i^t(I,a)\dfrac{t^\alpha}{t^\alpha+1}+v_i^{t+1}(I,a')\Big]^+$	$C^{t-1}(I,a)\Big(\dfrac{t-1}{t}\Big)^\gamma+\pi_i^{\sigma^t}(I)\sigma_i^t(I,a)$

证明　根据文献 [22] 引理 5, 给定动作集A, 假定局中人i使用 CogCFR 迭代T次, 则该局中人的“遗憾”至多为$\Delta|\mathcal{I}_i|\sqrt{|A|}\sqrt{T}$, 平均遗憾最多为$\frac{8}{3}\Delta|\mathcal{I}_i|\sqrt{|A|}/\sqrt{T}$.

根据文献 [18] 引理 1, 对于所有i, 如果$B>0$, $C\leqslant 0, \sum_{t=1}^{i} x_t \geqslant C$且$\sum_{t=1}^{T} x_t \leqslant B$, 则称有界实数序列$x_1,\dots,x_T$是$BC$可靠的. 对于任意$BC$可靠序列和任意非减权重序列$w_t\geqslant 0, \sum_{t=1}^{T}(w_t x_t)\leqslant w_T(B-C)$.

根据定义, 任意次迭代t时的瞬时遗憾最小值为$-\Delta$. 对于t次迭代, 则有$\frac{(t-1)^{\beta_t}}{(t-1)^{\beta_t}+1}\leqslant\frac{(t-1)^0}{(t-1)^0+1}=\frac{1}{2}$. 因此, 每个决策点上每个行动的遗憾大于$-2\Delta$.

考虑策略迭代序列$\sigma'^1,\dots,\sigma'^T$, 对应加权权重取值为$w_{a,t}=\prod_{i=t+1}^{T}\left(\frac{i-1}{i}\right)^{\gamma_i}$. 第$t$次迭代信息集$I$上行动$a$的遗憾为$R'^t(I,a)$. 则有

$$R^T(I,a)\leqslant\frac{8}{3}\Delta\sqrt{|A|}\sqrt{T}.$$

由于$w_{a,t}$是递增序列, 假定$B=\frac{8}{3}\Delta\sqrt{|A|}\sqrt{T}$, $C=-2\Delta$, 有

$$R'^T(I,a)\leqslant\frac{8}{3}\Delta\sqrt{|A|}\sqrt{T}+2\Delta. \tag{7}$$

根据γ的区间上下界, “加权和”满足

$$\sum_{t=1}^{T} w_{a,t}\geqslant\sum_{t=1}^{T}\Big(\prod_{i=t+1}^{T}\left(\frac{i-1}{i}\right)^8\Big)=\sum_{t=1}^{T}\left(\frac{t}{T}\right)^8\geqslant\frac{T}{9}. \tag{8}$$

对应平均遗憾满足

$$R_i'^{w,T}=\max_{a\in\mathcal{A}}\frac{R'^T(I,a)}{\sum_{t=1}^{T} w_{a,t}}\leqslant\frac{9}{\sqrt{T}}\Delta\Big(\frac{8}{3}\sqrt{|A|}+\frac{2}{\sqrt{T}}\Big). \tag{9}$$

根据式 (9) 可得, 该方法T次迭代最终平均策略为$\frac{9}{\sqrt{T}}\Delta|\mathcal{I}|\Big(\frac{8}{3}\sqrt{|A|}+\frac{2}{\sqrt{T}}\Big)$纳什均衡. □

2.3.2　鲁棒对手利用

对手建模是博弈理论中一个重要的研究方向, 重点研究如何建立一个清晰的模型来预测对手的行为. 同时按照建模方式区分: 1) 在不完美信息博弈中, 更新博弈树节点上的对手概率分布, 推导出对手的典型模型, 得到最优的对应策略的显示建模方法; 2) 弥补在缺乏对手大量历史数据的基础上, 采取的隐式建模方法. 主流的对手利用方法有数据偏差 CFR[25]、安全对手利用[26]、行为约束 CFR [27]、贝叶斯对手利用[28] 和约束 CFR [29].

虽然, 在应对基于均衡策略高水平对手时, 基于均衡解的对手建模技术可以提高局中人的自适应能力, 但该技术同时也限制了局中人的适应性, 因为无法充分利用相对较弱的对手. 只有当局中人不能足够迅速地适应对手以利用其策略时, “安全”的概念才有意义. 具有理想对手模型的局中人应该能够调整其策略, 以采取针对所有对手的近似最佳反应. 即应针对每个 (弱的或强的) 对手, 以最大程度地利用为目标, 并依靠适应来避免被利用, 而不是将适应限制在偏离均衡策略的有限范围之内. 针对理性对手的精确和近似纳什均衡策略, 无法利用有限理性对手的非均衡策略. 当前主流方法是求解近似鲁棒最佳响应, 鲁棒最佳响应对策是最大保守纳什均衡策略与最大进取最佳对策策略的折衷.

两人零和不完美信息博弈通常可转化为双线性鞍点问题 (bilinear saddle point problem, BSPP) 模型

$$\min_{x\in X}\max_{y\in Y}\langle x,Uy\rangle=\max_{y\in Y}\min_{x\in X}\langle x,Uy\rangle. \tag{10}$$

其中: U为收益, xy分别为局中人的序贯形式策略. 围绕博弈收益约束[30], 本文采用单侧信任域的方式设计带约束的 CFR 算法, 即根据历史交互信息获取关于Y的相关信息, 分析己方预期策略, 有

$$X^t:=\{x\in X:\max_{y\in Y^{t-1}} x^TU_iy\geqslant\delta\}. \tag{11}$$

为了平衡对手利用率和己方被利用率, 本文采用探索协调因子ϕ, 即

$$x^t:=(1-\phi)\arg\max_{x\in X^t}\max_{y\in Y^{t-1}} x^TU_iy+\phi\arg\max_{x\in X^t} x^TU_iy^{t-1}. \tag{12}$$

3　海上基地防护策略实验分析

3.1　实验环境设置

本文实验平台采用 Rog Strix G634J 笔记本电脑, CPU 为 Intel(R) i9-14900HX, 2.2 GHz. GPU 型号为 GeForce RTX 4080. 程序采用 Julia 与 Python 语言混合编码.

3.2　参数配置元控制分析

为了对比各类不同算法的有效性, 本文主要基准算法有:

CFR$(\alpha=\infty,\beta=\infty,\gamma=0)$,
CFR+$(\alpha=\infty,\beta=-\infty,\gamma=2)$,
DCFR$(\alpha=1.5,\beta=0,\gamma=2)$[18],
DCFR+$(\alpha=1.5,\beta=0,\gamma=4)$[20],
PDCFR+$(\alpha=2.3,\beta=0,\gamma=5)$[31].

采用的超参数主要包括: α_t用于调整正遗憾值的折扣权重$\left(\frac{t-1}{t}\right)^{\alpha_t}$, β_t用于调整负遗憾值的折扣权重$\left(\frac{t-1}{t}\right)^{\beta_t}$, γ_t用于调整累积策略的折扣权重$\left(\frac{t-1}{t}\right)^{\gamma_t}$.

本文 CogCFR 中设置参数的取舍范围为($\alpha \in [1,8], \beta \in [-8,8], \gamma \in [8,40]$), 参数随时间的取值为

$$\alpha = 1 + (3/1\,000)t,$$
$$\beta = 8 - (5/1\,000)t,$$
$$\gamma = 8 + (5/1\,000)t.$$

相关参数取值如表 4 所示.

表4 相关参数取值

符号	取值范围
N	$9, 10, 11, 12, 13$
D	$[4, 8]$
r	$[0.2, 0.4]$
n_d	$[4, 8]$
n_a	$[2, 6]$
m_d	$[0.2, 1]$
m_a	$[0.2, 1]$
n_{cd}	$[0.2, 0.8]$
n_{ca}	$[0.2, 0.8]$
p	$[0.6, 0.95]$
p_s	$[0.6, 0.95]$
p_d	$[0.6, 0.95]$
p_a	$[0.05, 0.4]$
δ	$[-2, -1]$
ϕ	$[0.1, 0.9]$

本文主要对比 CogCFR、DCFR+、PDCFR+三类算法, 相关结果如图 3 所示. 由图 3 可见所提出方法的有效性, 其收敛速度较 DCFR+和 PDCFR 算法更快.

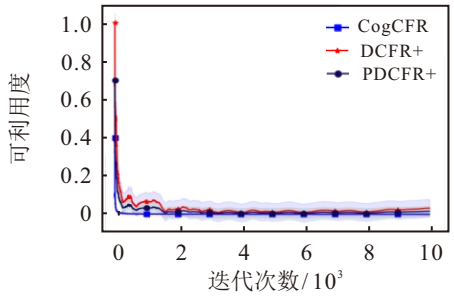

图3 参数配置元控制分析

3.3 最优资源分配策略实证分析

为了求解海上基地防护对应的资源分配策略, 本文首先采用归一化方式生成海上基地位置和对应效用值信息, 对应一体化防空系统的防御半径设置为$(0, 1]$.

当$N = 10$时, CogCFR 算法对应单个实验想定中最优资源分配策略如表 5 所示. 根据基地编号, 求解最优资源分配策略. 攻击方可能主要在 1 号和 3 号基地实施侦察渗透, 打击 2 号、3 号和 6 号基地. 防御方的最优资源分配策略为将一体化防空反导系统部署至高价值 (效用值) 的 1 号、3 号、4 号、5 号、8 号、9 号基地, 同时在 1 号和 3 号基地部署水下侦察预警系统, 做好反渗透准备.

表5 防御方最优资源分配策略 ($N = 10$)

编号	经纬度(归一化)	效用值	打击/侦察	部署/反渗透
1	(0.036, 0.717)	0.479	侦察	部署/反渗透
2	(0.700, 0.713)	0.159	打击	
3	(0.772, 0.625)	0.711	打击/侦察	部署/反渗透
4	(0.122, 0.754)	0.969		部署
5	(0.374, 0.720)	0.713		部署
6	(0.624, 0.817)	0.474	打击	
7	(0.923, 0.267)	0.506		
8	(0.813, 0.323)	0.707		部署
9	(0.288, 0.848)	0.210		部署
10	(0.977, 0.470)	0.295		

可以看出, 算法在1 600次时已经收敛, 为进一步分析算法的有效性, 本文区分迭代次数 (100, 500, 1 500, 3 000) 组织资源分配策略的可视化分析. 资源部署结果如图 4 所示. 从上至下, 每个子图分别代表算法迭代100, 500, 1 500, 3 000次时, 攻防对抗过程. 10 个海上基地中 2 号, 3 号和 6 号基地可能遭受敌物理打击.

3.4 问题规模扩展性分析

由于不完美信息博弈的动作空间太大, 随着问题规模扩大, 计算耗时较长. 本文围绕海上基地数量进行扩展性分析, 算法收敛需要的迭代次数呈线性增长、时间呈指数增长. 当基地数量超过13时, 序贯式策略组合近7.85×10^8个, 完成计算需要的内存空间更大. 相关扩展性结果如表 6 所示.

3.5 对手利用鲁棒响应策略

面对非理性对手, 如何有效评估风险, 根据对手历史信息制定反制对手的鲁棒响应策略十分关键. 这里将收益约束设置为δ, 将探索协调因子设置为ϕ. 本文根据对手利用率和己方被利用率指标, 采用网格搜索的方式, 利用校正决定系数 (adjusted R-square) 来分析最佳权重系数. 本文主要采用的对比基准算法包括基于数据偏差 CFR(DBCFR) 和基于约束的 CFR(CCFR) 两类, 相关结果如表 7 所示. 总

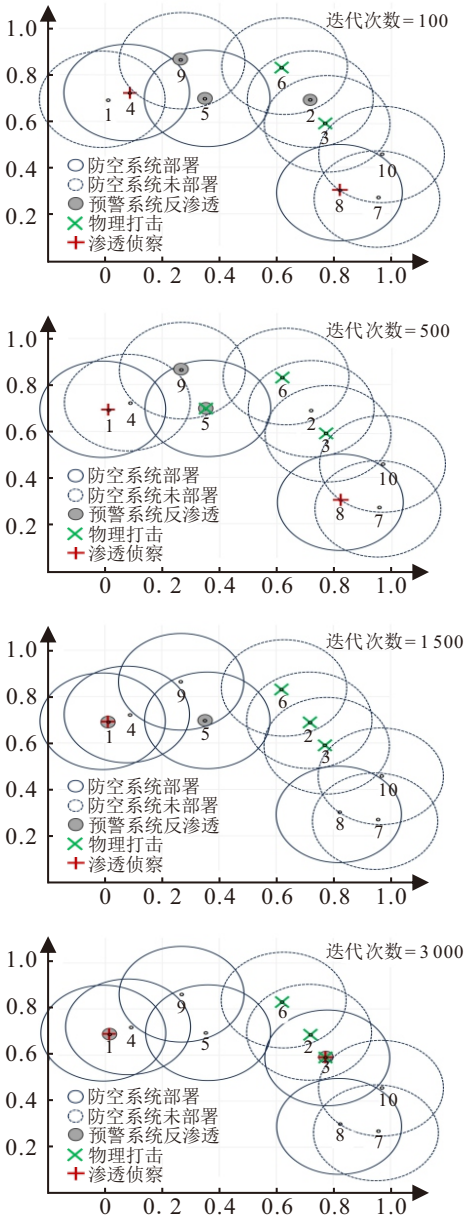

图4　资源分配策略学习过程可视化

表6　不同数量规模海上基地设置及算法性能表现

基地数量	信息集大小/10^8	计算时间/s	防护收益
9	2.25	385	−1.82
10	3.21	496	−1.34
11	4.41	636	−1.74
12	5.89	941	−1.86
13	7.65	1 105	−1.93

表7　不同数量规模海上防护收益及计算时间

基地数量	DBCFR/s	CCFR/s	CogCFR/s
9	−0.580/248	−0.584/1 599	−0.589/1 860
10	−0.248/355	−0.249/1 800	−0.252/1 990
11	−0.392/530	−0.395/1 800	−0.398/2 160
12	−0.609/720	−0.620/1 801	−0.629/2 390
13	−0.607/947	−0.612/1 800	−0.621/2 530

体来看, 基于 CogCFR 动态探索协调增加了算法的计算时间.

4　结　论

强对抗条件下的多方序贯对抗问题可以结合具象化场景组织博弈模型构建, 区分“近似理性均衡解”或“鲁棒对手响应”展开博弈策略求解算法设计, 首先求解基准“蓝图”策略, 然后根据具体情况分析对手可能的行动策略, 并制定最佳反制策略. 围绕海上基地防护安全博弈分析, 结合多阶段攻防场景, 构建不完美信息博弈模型, 分别针对理性对手和非理性对手设计基于超参数调度的反事实遗憾最小化方法及考虑约束的 CogCFR 算法. 实验结果表明了该类算法的有效性和扩展性.

考虑到多阶段交互的扩展性, 以及如何处理信息不对称情形, 未来将尝试引入信号博弈和斯塔克伯格博弈等基类模型构建信息不对称条件下的多阶段动态博弈模型. 此外, 在算法设计方面, 一是可引入正则化项, 嵌入在线凸优化相关算法加快算法收敛, 设计具有最佳、随机或末轮迭代收敛 (best、random、last-iterate convergence)的高效策略学习方法; 二是耦合利用深度强化学习方法高效学习遗憾值, 辅助决策时规划策略搜索.

参考文献 (References)

[1] Sean W I. An analysis of artificial intelligence performance and behavior within the model of expeditionary advanced base operations[D]. Monterey: Naval Postgraduate School, 2024: 5-9.

[2] Priebe M, Vick A J, Heim J L, et al. Distributed operations in a contested environment[R]. RAND Project, 2019.

[3] Bryan C, Dan P, Hassison S. Mosaic warfare exploiting artificial intelligence and autonomous systems to implement decision-centric operations[R]. Center for Strategic and Budgetary Assessments, 2020.

[4] Keith A, Ahner D. Counterfactual regret minimization for integrated cyber and air defense resource allocation[J]. European Journal of Operational Research, 2021, 292(1): 95-107.

[5] Zinkevich M, Johanson M, Bowling M, et al. Regret minimization in games with incomplete information[C]. Proceedings of the 21st International Conference on Neural Information Processing Systems. British Columbia, 2007: 1729-1736.

[6] Ganzfried S, Laughlin C, Morefield C. Parallel algorithm for Nash equilibrium in multiplayer stochastic games with application to naval strategic planning[C]. Distributed Artificial Intelligence. Cham: Springer International Publishing, 2020: 1-13.

[7] Luo J R, Zhang W P, Gu X Q, et al. GER-PSRO: Graph embedding representation based reinforcement learning in sequential colonel blotto game[C]. 2024 43rd Chinese Control Conference. Kunming, 2024: 8162-8167.

[8] McCarthy J. Unifying strategic military force design and operational warfighting: A stochastic game approach[D]. Atlanta: Georgia Institute of Technology, 2024.

[9] Du B K, Xiong W, Wang H T, et al. AG600 maritime base location decision based on the interval intuitionistic fuzzy TOPSIS method[J]. IEEE Access, 2022, 10: 82483-82492.

[10] Zeng B, Wang R, Li H P, et al. Nash equilibrium strategy and attack-defense game model for naval support base[J]. Systems Engineering and Electronics, 2022, 44(8): 2570-2580.

[11] Wang Z, Yuan Y, An B, et al. An overview of security games[J]. Journal of Command and Control, 2015, 1(2): 121-149.

[12] Vinyals O, Babuschkin I, Czarnecki W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning[J]. Nature, 2019, 575(7782): 350-354.

[13] Brown N, Sandholm T. Superhuman AI for multiplayer poker[J]. Science, 2019, 365(6456): 885-890.

[14] Nash J. Equilibrium points in *n*-person games[J]. Proceedings of the National Academy of Sciences, 1950: 48-49.

[15] 罗俊仁, 张万鹏, 苏炯铭, 等. 计算机博弈中序贯不完美信息博弈求解研究进展[J]. 控制与决策, 2023, 38(10): 2721-2748.
(Luo J R, Zhang W P, Su J. Research progress on sequential imperfect information game solving in computer games[J]. Control and Decision, 2023, 38(10): 2721-2748.)

[16] Tammelin O. Solving large imperfect information games using CFR+ [J/OL]. 2024, arXiv: 1407.5042.

[17] Bowling M, Burch N, Johanson M, et al. Heads-up limit hold'em poker is solved[J]. Science, 2015, 347(6218): 145-149.

[18] Brown N, Sandholm T. Solving imperfect-information games via discounted regret minimization[C]. Proceedings of the AAAI Conference on Artificial Intelligence, Hawaii, USA, 2019, 33(1): 1829-1836.

[19] Farina G, Kroer C, Sandholm T. Faster game solving *via* predictive Blackwell approachability: Connecting regret matching and mirror descent[C]. Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2021, 35(6): 5363-5371.

[20] Xu H, Li K, Fu H B, et al. AutoCFR: Learning to design counterfactual regret minimization algorithms[C]. Proceedings of the AAAI Conference on Artificial Intelligence, Arlington, Virginia, 2022, 36(5): 5244-5251.

[21] Zhang H, Lerer A, Brown N. Equilibrium finding in normal-form games *via* greedy regret minimization[C]. Proceedings of the AAAI Conference on Artificial Intelligence, Arlington, Virginia, 2022, 36(9): 9484-9492.

[22] Xu H, Li K, Fu H, et al. Dynamic discounted counterfactual regret minimization[C]. The 12th International Conference on Learning Representations. Vienna, 2024:1-9.

[23] Lanctot M, Waugh K, Zinkevich M, et al. Monte Carlo sampling for regret minimization in extensive games[C]. Proceedings of the 22nd International Conference on Neural Information Processing Systems. Piscataway: IEEE, 2009: 1078-1086.

[24] Li H, Wang X, Qi S, et al. Solving imperfect-information games via exponential counterfactual regret minimization[J/OL]. 2020, arXiv: 2008.02679.

[25] Johanson M, Bowling M. Data biased robust counter strategies[C]. Proceedings of the 12th International Conference on Articial Intelligence and Statistics. Florida, 2009: 264-271.

[26] Ganzfried S, Sandholm T. Safe opponent exploitation[J]. ACM Transactions on Economics and Computation, 2015, 3(2): 1-28.

[27] Farina G, Kroer C, Sandholm T. Regret minimization in behaviorally-constrained zero-sum games[C]. International Conference on Machine Learning. 2017: 1107-1116.

[28] Ganzfried S, Sun Q Y. Bayesian opponent exploitation in imperfect-information games[C]. 2018 IEEE Conference on Computational Intelligence and Games. Maastricht, 2018: 1-18.

[29] Davis T, Waugh K, Bowling M. Solving large extensive-form games with strategy constraints[C]. Proceedings of the AAAI Conference on Artificial Intelligence, Hawaii, USA, 2019, 33(1): 1861-1868.

[30] Bernasconi M, Cacciamani F, Fioravanti S, et al. Exploiting opponents subject to utility constraints in extensive-form games[C]. Proceedings of the 35th Conference on Neural Information Processing Systems. Canada, 2021: 13177-13188.

[31] Xu H, Li K, Liu B, et al. Minimizing weighted counterfactual regret with optimistic online mirror descent[J/OL]. 2024, arXiv: 2404.13891.

作者简介

罗俊仁 (1989−), 男, 博士生, 主要研究方向为非对称信息博弈、策略搜索, E-mail: luojunren17@nudt.edu.cn;

张万鹏 (1981−), 男, 研究员, 博士, 主要研究方向为大数据智能、智能演进, E-mail: wpzhang@nudt.edu.cn;

谷学强 (1983−), 男, 副研究员, 博士, 主要研究方向为智能规划与决策, 边缘智能, E-mail: xueqiang_gu@nudt.edu.cn;

陈璟 (1972−), 男, 教授, 博士, 主要研究方向为认知决策博弈、分布式智能, E-mail: chenjing001@vip.sina.com.