布洛托上校博弈模型及求解方法研究进展

罗俊仁, 邹明我, 陈少飞, 张万鹏, 陈璟

引用本文

罗俊仁, 邹明我, 陈少飞, 张万鹏, 陈璟. 布洛托上校博弈模型及求解方法研究进展[J]. 计算机科学, 2024, 51(1): 84-98.

LUO Junren, ZOU Mingwo, CHEN Shaofei, ZHANG Wanpeng, CHEN Jing. Research Progress on Colonel Blotto Game Models and Solving Methods [J]. Computer Science, 2024, 51(1): 84-98.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600011

布洛托上校博弈模型及求解方法研究进展

罗俊仁　邹明我　陈少飞　张万鹏　陈　璟
国防科技大学智能科学学院　长沙 410073
(luojunren17@nudt.edu.cn)

摘　要　对抗条件下的资源分配是大多数博弈决策问题的核心。从拟合最优解到博弈均衡解,基于博弈论的资源分配策略求解是认知决策领域的前沿课题。文中围绕对抗条件下资源分配的布洛托上校博弈模型和求解方法展开综述分析。首先,简要介绍了离线与在线策略学习的区别,策略博弈与相关解概念,在线优化与遗憾值;其次,梳理了6类布洛托上校博弈典型模型(连续布洛托上校博弈、离散布洛托上校博弈、广义布洛托上校博弈、广义乐透布洛托博弈、广义规则布洛托上校博弈与在线离散布洛托上校博弈);然后,区分2个阶段(离线与在线)3类博弈场景(单次、重复、多阶段),分析了多类布洛托上校博弈求解方法;最后,从典型应用探索、广义博弈模型、博弈求解方法、未来研究展望共4方面进行了未来研究前沿分析及展望。通过对当前布洛托上校博弈进行概述,期望能为对抗条件下资源分配与博弈论相关领域的研究带来启发。

关键词:资源分配;布洛托上校博弈;近似纳什均衡;在线凸优化;期望遗憾;高概率遗憾
中图分类号　TP181

Research Progress on Colonel Blotto Game Models and Solving Methods

LUO Junren,ZOU Mingwo,CHEN Shaofei,ZHANG Wanpeng and CHEN Jing
College of Intelligence Science and Technology,National University of Defense Technology,Changsha 410073,China

Abstract　Resource allocation under confrontation conditions is the core of most game decision problems. From fitting optimal solution to game equilibrium solution,resource allocation strategy solving based on game theory is a frontier topic in cognitive decision-making field. This paper summarizes and analyzes the Colonel Blotto game model and its solution method for adversarial resource allocation. Firstly,the differences between offline and online strategy learning,strategy game and related solution concepts,online optimization and regret value are briefly introduced. Secondly,six types of Colonel Blotto game models(continuous Blotto game,discrete Colonel Lotto game,generalized Colonel Blotto game,generalized Lotto Blotto game,generalized rule Colonel Lotto game and online discrete Colonel Lotto game). Then,this paper distinguishes 2 stages(offline and online) and 3 types of game scenarios(single,repeated,multi-stage),and analyzes the solution method of Colonel Blotto game. Finally,the future research frontiers are analyzed and prospected from four aspects:typical application exploration,generalized game model,game solving method and future research prospect. The main purpose is to give an overview of the current Colonel Blotto game,hoping to enlighten the research on resource allocation and game theory under confrontation condition.

Keywords　Resource allocation,Colonel Blotto game,Approximate Nash equilibrium,Online convex optimization,Expected regret,High-probability regret

1　引言

资源分配是决策科学领域的核心问题。资源拥有方需要将有限或可重复使用的资源分配给若干个目标(对象),根据待分配资源与目标对象之间的类型和数量关系,考虑相应限制条件和待优化目标,构建资源分配模型,得出资源分配方案。在电力分配[1]、网络服务分配[2]、安全设备布设[3]、作战兵力布势[4]、军事攻防资源分配[5]、在轨服务资源分配[6]、云计算服务分配[7]、政治选举[8]等领域,很多问题都可以建模成资源分配问题。运筹学与博弈论作为决策科学的两个分支,为资源分配问题提供了建模工具与求解方法。资源分配问题相关研究的很多情境均涉及多个决策者之间的策略交互,而博弈论特别适合建模交互式决策过程。从博弈论视角来看,资源分配可用于多种场景、多类问题的建模。根据博弈局中人之间的合作、对抗、混合关系,资源分配可分为合作条件下资源分配问题、对抗条件下资源分配问题以及混合式

到稿日期:2023-06-01　返修日期:2023-09-27
基金项目:国家自然科学基金(61806212);湖南省研究生创新项目(CX20210011)
This work was supported by the National Natural Science Foundation of China(61806212) and Hunan Postgraduate Innovation Project (CX20210011).
通信作者:陈璟(chenjing001@vip.sina.com)

条件下资源分配问题，其中对抗条件下资源分配问题是本文的研究核心。近年来，人工智能技术的相关研究从计算智能、感知智能逐步向决策智能聚焦，相关方法正从传统上以数据拟合为核心求解最优值转向以博弈论为核心求解均衡[9]。

布洛托上校博弈(Colonel Blotto Game，CBG)是一类典型的对抗性资源分配博弈，简称布洛托博弈，其概念最早由Borel于1921年提出[10]，被认为是现代博弈论的起源[11]。初始版本的布洛托上校博弈相关术语与军事相关，描述了一个两人博弈对抗场景：博弈双方均有固定的资源预算，双方同时将资源分配到多个战场。当局中人在某个战场上分配的资源比对手多时，他就赢得该战场并获得对应的战场价值，而输了的一方获得的战场价值为零，获得较多战场价值的一方即为最后的赢家。虽然博弈规则非常简单，有着著名的模型谱系，但由于策略空间规模大，且问题形式多样，布洛托上校博弈策略求解至今仍未完全解决。布洛托上校博弈与军事资源分配问题密切相关，并且由于其模型简单而具有通用性，在实际中也可以依据事实对该模型进行更改。布洛托上校博弈早期的相关应用主要聚焦在军事和后勤问题上[12]，这类问题的资源对象可以是兵力、武器装备、弹药，研究目标是对抗双方应该如何分配作战资源而获得战斗的胜利。由于其“赢者通吃”属性，美国等西方国家的政治选举问题也可以建模成布洛托上校博弈问题，即候选人如何分配时间、金钱、人力等资源至各个州，而获得比其对手更多的选票[13]。2021年，兰德公司围绕马赛克战作战概念，发布《布洛托上校博弈对马赛克战的启示》研究报告[14]，借助布洛托上校博弈模型，从作战资源分配能力层面探讨了马赛克战是否比传统集成式作战模式更具优势的问题，分析了马赛克战在未来作战模式中作战资源分配方面的优越性与局限性。图1描述了布洛托上校博弈模型下同构资源和异构资源在多战场上的分配场景，用来检验马赛克战模式下武器资源的分配是否比传统单一集成的作战模式更具优越性。马赛克式作战力量通常由数量较多、功能专一的多类新型作战单元(无人机、无人船、战列舰、巡逻机等)组成，而传统大型武器平台通常数量较少、功能综合、不能分解(如F-35战斗机、航空母舰、B-21轰炸机等)。同构布洛托上校博弈模型是指作战力量具备单一作战能力的情形，如图1左侧所示，传统上校可支配的作战资源是一支由10架航空器(每架携带5枚导弹)组成的编队，马赛克上校可支配的作战资源是一支由50架无人机(每架携带1枚导弹)组成的编队，在中间战场上分配更多导弹的上校将赢得该战场。异构布洛托上校博弈模型是指作战力量具备多种作战能力的情形，如图1右侧所示，传统上校有10艘战列舰(每艘5枚导弹、1部雷达)和15架海上巡逻机(每架2枚导弹、1部雷达和1台照像机)，马赛克上校可支配的作战资源有10艘声呐舰(每艘1部雷达)、25艘导弹驱逐舰(每艘2枚导弹)、15架海上监视无人机(每架1部雷达、1台照像机)、15架武装无人机(每架2枚导弹)。每个战场的获胜条件依据是导弹、雷达和照像机等提供的联合作战能力。

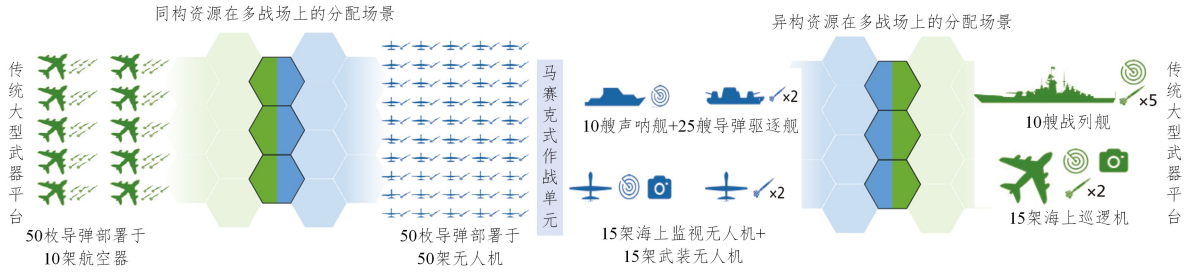

图1 布洛托上校博弈模型下同构资源和异构资源在多战场上的分配场景

Fig. 1 Allocation scenario of homogeneous and heterogeneous resources on multiple battlefields with Colonel Brotto's game model

本文主要从博弈论视角来分析资源分配问题，聚焦面向资源分配的各类布洛托上校博弈模型，区分离线与在线博弈场景，对比分析布洛托上校博弈的求解方法。本文的整体结构如图2所示，第2章介绍策略博弈与解概念、离线与在线博弈求解、在线组合优化基本原理；第3章分析资源分配博弈场景，梳理多类布洛托上校博弈模型；第4章区分场景对比分析博弈策略求解方法；第5章梳理面临的问题及挑战，并展望未来的研究方向。

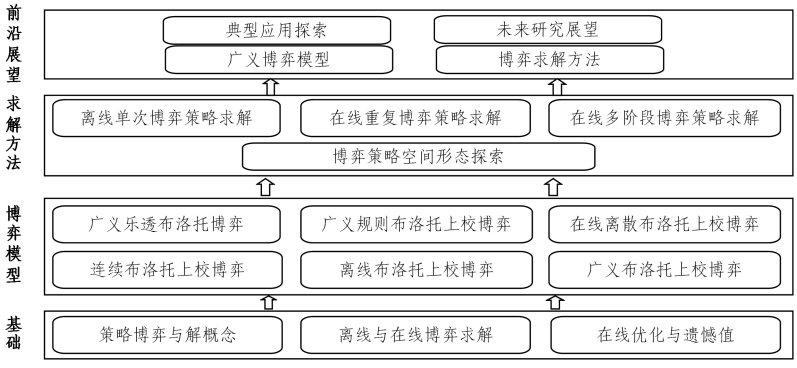

图2 本文整体结构框架

Fig. 2 Overall structure framework of this paper

2 策略博弈基础

博弈论是一种借助数学工具来建模分析交互式决策的方法论，其建立可以追溯到20世纪初[10]，而后，摩根斯坦和纳什等对策略博弈以及均衡的概念进行了界定[15]，根据局中人之间是否形成协议将其分为合作博弈和非合作博弈。在非合作博弈中，局中人彼此独立行动，每个局中人都在自己偏好的前提下尽力获得最佳结果；合作博弈中，局中人可以达成共识并签署有约束力的合约来实施协调一致的行动。在博弈的过程中，根据局中人之间的行动顺序，博弈可以分为正则式博弈、斯塔克尔伯格博弈、扩展式博弈、重复博弈、多阶段博弈、马尔可夫博弈等。正则式博弈是指在博弈过程中，局中人同时采取行动，或者虽然不是同时采取行动，但是后行动者不知道先行动者采取的具体行动；斯塔克尔伯格博弈也称“主从”博弈，常用于攻防问题建模，是指在博弈过程中区分先手方和后手方来分析各方采取的行动；扩展式博弈是指在博弈过程中，局中人的行动有先后顺序，并且后行动者能够观察到先行动者所采取的具体行动；重复博弈指的是同一个博弈进行多次的博弈过程，构成重复博弈的单次博弈(one-shot game)也称为“原博弈”或“阶段博弈”；多阶段博弈也称动态博弈，主要是指博弈时序上有先后顺序，后决策的局中人在做出决策之前可以看到之前局中人所做出的决策，甚至包括自己的决策。

2.1 策略博弈与解概念

策略博弈可以表示成一个元组 $G=(N,(S^j)_{j\in N},(u^j)_{j\in N})$，其中 N 是局中人集合，$|N|$ 是局中人集合的势，S^j 是局中人 $j\in N$ 的策略集合，$u^j:S\to R$ 是局中人 $j\in N$ 的收益函数。

如果每个局中人的策略集合都是有限的，则称该博弈为有限博弈。当 $|N|=2$ 时，称该博弈为两人博弈。定义 $\sum_{j=1}^{N}u^j=a$，如果 a 为固定常数，则称该博弈为常和博弈；当 $a=0$ 时，称为零和博弈。

2.1.1 典型对抗场景

1)单次博弈。$G=(N,(S^j)_{j\in N},(u^j)_{j\in N})$ 是单次博弈的条件为：模拟只发生单次事件，每个局中人 $j\in N$ 知道博弈的所有参数和细节 $(N,S^j,u^j,\forall j\in N)$，所有局中人同时独立地采取他们的行动。$S^j$ 的每个元素都被称为局中人 $j\in N$ 的纯策略，而集合 S 的每个元素都被称为一个策略组合(strategy profile)。

2)重复博弈。$G=(N,(S^j)_{j\in N},(u^j)_{j\in N})$ 是重复博弈的条件为：模拟发生次数为 T 的重复事件，对于其中的每一个阶段 t，每个局中人 $i\in N$ 知道博弈的部分参数 $(N,S^j_{j=i},u^j_{j=i},j\in N)$，即除了博弈的共同参数外，局中人只知道自己的策略空间和收益，而对其他局中人的策略空间和收益是未知的，所有局中人同时独立地选择他们的行动，按照此规定进行 T 次博弈直到结束，其中次数可以为有限次或无限次。

3)多阶段博弈。$G=(N,(S^j)_{j\in N},(u^j)_{j\in N})$ 是多阶段博弈的条件为：模拟发生次数为 T 的多阶段事件，对于其中的每一个阶段 t，每个局中人 $i\in N$ 知道博弈的部分参数 $(N,S^j_{j=i},u^j_{j=i},j\in N)$，即除了博弈的共同参数外，局中人只知道自己的策略空间和收益，而对其他局中人的策略空间和收益是未知的，所有局中人同时独立地选择他们的行动，按照此规定进行 T 次博弈，直到结束。

2.1.2 解概念

一般将非合作博弈分为4类：完全信息静态博弈、完全信息动态博弈、不完全信息静态博弈、不完全信息动态博弈。与之对应的有4种均衡：纳什均衡(Nash Equilibrium，NE)、子博弈精炼纳什均衡、贝叶斯纳什均衡、精炼贝叶斯纳什均衡。其中，纳什均衡常被简单地称为均衡或者均衡点，是策略博弈中最核心、最重要的解概念，它体现了博弈的稳定性。在纳什均衡下，每个局中人针对其他局中人的行为，选择对自己最有利的行动。

1)纳什均衡

博弈 $G=(N,(S^j)_{j\in N},(u^j)_{j\in N})$ 的纳什均衡 $s_*=(s^j_*)_{j\in N}\in S$ 是指对于每个局中人 $j\in N$，满足不等式约束的策略组合：

$$u^j(s_*)\geqslant u^j(s^j,s^{-j}_*),\forall s^j\in S^j$$

其中，s^{-j}_* 表示策略组合 s_* 中除局中人 j 外其他局中人的策略集合；(s^j,s^{-j}_*) 表示局中人 j 的策略用 s^j 代替，而其他局中人的策略不变。

设 $\varepsilon\geqslant 0$，博弈 $G=(N,(S^j)_{j\in N},(u^j)_{j\in N})$ 的近似纳什均衡(ε-NE)策略 $s_*\in S$，对于 $\forall j\in N$，$s^j\in S^j$ 满足：

$$u^j(s^j_*,s^{-j})+\varepsilon\geqslant u^j(s^j_*,s^{-j}_*)\geqslant u^j(s^j,s^{-j}_*)-\varepsilon$$

2)混合策略纳什均衡

有限策略博弈 $G=(N,(S^j)_{j\in N},(u^j)_{j\in N})$ 至少存在一个混合纳什均衡[16]。用 $\Delta(S^j)$ 表示在 S^j 上的概率分布集合，且 $\Delta:=\mathrm{X}_{j\in N}\Delta(S^j)$ 表示所有局中人的概率分布集合，则 $\Delta(S^j)$ 中的元素(一个纯策略及其对应的概率赋值)称为局中人 j 的混合策略，Δ 中的元素称为混合策略组合。用 $\tilde{u^j}:\Delta\to R$ 表示对于任意的混合策略组合为 $(\sigma^j)_{j\in N}\in\Delta$ 时，局中人 j 的期望收益。如果策略博弈 G 是一个有限式博弈，则当且仅当 $\tilde{u^j}$ 取得最大值时，$(\sigma^j_*)_{j\in N}\in\mathrm{X}_{j\in N}\Delta(S^j)$ 是博弈 $G=(N,(S^j)_{j\in N},(u^j)_{j\in N})$ 的混合策略纳什均衡。

2.2 离线与在线博弈求解

近年来，随着人工智能技术的发展，从感知智能、认知智能到决策智能，博弈论为决策问题建模提供了有力支撑。“大数据+算力+算法”的数据驱动范式，为博弈问题求解提供了通用解决方案。其中，基于学习(深度学习、强化学习)设计的迭代式问题求解方法是离线博弈策略学习的基础范式。然而，与离线批式(batch)利用模拟多次对抗学习博弈策略不同，在线博弈对抗过程博弈各方处于策略解耦合状态。在线博弈对抗策略的求解本质是一个流式(flow)学习过程，需要根据少量的此前交互样本来做决策。根据具体对抗场景，当前的相关研究将博弈求解区分为两大类：离线求解和在线求解。离线求解包括围绕博弈均衡查找(finding)、计算(computation)与学习(learning)等方法，如利用与模拟器交互的迭代式学习方法、基于算法博弈论的一阶优化方法、基于环境模型的预训练方法，这些方法的求解对象是离线预训练模型或博弈蓝图策略。在线求解包括围绕适应性应对对手的单次博弈策略搜索(search)方法、重复博弈策略无遗学习(no regret learning)方法和多阶段博弈策略优化(optimization)方法等，这类方法的求解对象是适应性、

鲁棒性、安全性等反制应对策略。

2.3 在线优化和遗憾值(界)

2.3.1 在线组合优化

在线组合优化[17]是在线线性优化[18]问题的一个实例，著名的多臂赌博机(Multi-Armed Bandits，MAB)问题[19]和在线最短路径(Online Shortest Path，OSP)问题[20]都属于这类可利用事前交互序列进行预测进而决策的问题。这类问题假设策略学习器和对手之间进行多阶段博弈，不同阶段的博弈之间存在联系，学习器不仅需要关心即时收益，还需要立足长远目标进行动作选择，因为当前的动作可能会影响未来的收益。此外，学习器决策时可能不了解对手的一些信息，也不知道过去的行为收益甚至博弈的一些参数。

在线线性优化问题中，学习器和对手之间进行 T 轮博弈，$S\subset R^D$ 表示学习器的动作集，$D\in N\backslash\{0\}$，在每一个时间阶段 $t\in[T]$，学习器在不知道对手动作的情况下选择动作向量 $\tilde{\boldsymbol{p}}_t\in S$，产生一个损失向量 $\boldsymbol{l}_t\in[0,1]^D$，即在时间阶段 t，学习器选择动作和对手对抗后的损失。该损失对学习器的公开程度决定了学习器能够获取的最大信息反馈量。在线线性优化问题中，学习器的信息反馈一般包括以下 3 种情形。

1)完全(full)信息反馈：在阶段 t 结束时，学习器能够观察到所有动作的损失向量 $\boldsymbol{l}_t$。

2)半赌博机(semi-bandit)信息反馈：在阶段 t 结束时，学习器能够观察到所做动作的损失向量 $\boldsymbol{l}_t$。

3)赌博机(bandit)信息反馈：在阶段 t 结束时，学习器只能够观察损失标量 $L(\tilde{\boldsymbol{p}}_t)=(\boldsymbol{l}_t)^{\mathrm{T}}\tilde{\boldsymbol{p}}_t$，即所做动作总的损失。

需要注意的是，在半赌博机信息反馈的情形中，还存在由所做动作的损失经过简单的推理获得未知动作损失的特殊情形，即带“侧面观测”(side observations)的半赌博机信息反馈[21]。其是否存在由实际问题决定，但不可否认，如果能推理获得未知动作的损失，则针对不同的动作能够获得更准确的权重估计，从而对下一阶段的动作生成有重要影响。

与不完全信息博弈中的信息补全思想类似，通过信息补全可以获得尽可能多的对手信息，是己方获得有利决策优势的直接途径。使用符号 $X\succ Y$ 表示反馈设置为 X 时所获得的信息较反馈设置为 Y 时所获得的信息更多，则在线线性优化问题中不同的信息反馈设置的关系如下：

完全信息≻侧面观测半赌博机信息≻半赌博机信息≻赌博机

2.3.2 多臂赌博机

MAB 问题模型是有限信息反馈的序列学习问题中最基本的模型之一。MAB 按照损失(或者收益)产生方式的不同，分为随机型(stochastic)赌博机和对抗型(adversarial)赌博机。随机型赌博机中每支臂的收益服从一个固定但未知的概率分布[22]，在每次试验中，学习器选择其中一支臂，而后获得对应的损失，通过重复试验，学习器有望获知赌博臂收益的概率分布参数，从而在后续的试验中最小化损失。Lai 等[23]是较早研究随机型赌博机问题的学者，并首次提出上置信界(Upper Confidence Bounds，UCB)算法。UCB 算法的核心思想是面对不确定性时保持乐观态度，其总是假设任何不确定性都将对学习器有正面影响。Auer 等[24]在 Lai 等的基础上提出 α-UCB 算法，实现了关于时间范围 T 的对数遗憾上界；Bubeck 等[25]、Garivier 等[26]针对 Auer 实现的遗憾上界的常数项进行了改进。随机型赌博机问题由于问题复杂度不高，研究理论较为完善。

面向 MAB 问题的 Exp3 算法最早由 Auer 等[27]提出。Exp3 是 Hedge 算法[28]的变体，是一种代表探索和利用的指数权重算法。在线对抗过程中的收益一般不满足独立同分布假设，因为对手的行为会随着时间或者策略的改变而变化，则 UCB 算法在在线对抗过程中会遭受线性遗憾[29]。与随机型赌博机不同，对抗型赌博机中每支臂的收益由对手决定。对抗型赌博机经常被用来建模学习器和对手之间的博弈问题，因为对抗型赌博机提供了一个良好的权衡探索和利用的模型框架，而探索和利用的权衡正是在线对抗问题普遍面临和需要解决的难点。对抗型赌博机中的对手类型又区分为健忘型(oblivious)对手和非健忘型(non-oblivious)对手[30]。在健忘型对手场景中，不同动作的损失(或者收益)由对手事先设定好(也可以看成和对手对抗产生)，并且在学习器开始行动之后不进行更改；在非健忘型对手场景中，如果对手具备学习能力，能够依据学习器过去的表现而随时更改策略，则称之为非健忘性对手，也称适变型(adaptive)对手[29,31]。Auer 等[27]提出的 Exp3 算法能够针对健忘型对手的赌博机问题实现 $O(\sqrt{DT\ln D})$ 的期望遗憾上界(D 表示赌博机的赌博臂数量，在其他问题中表示动作维度)。Exp3 被认为是求解对抗型赌博机问题的基准算法，其核心是利用指数权重更新不同动作的权重[32]。在 Exp3 算法的基础上，Auer 等又通过引入“专家建议”进行动作采样，提出 Exp4 算法并获得和最佳专家几乎相同的收益。另外，还有很多技术成功应用于对抗型赌博机算法设计，如在线镜像下降(Online Mirror Descent，OMD)[33]和正则化跟风(Follow The Regularized Leader，FTRL)[34]等工具。

2.3.3 遗憾值及界

在线学习通常会采用事后理性视角来分析[35]，利用遗憾值(界)作为策略的衡量指标。遗憾定义为：

$$r_T=\sum_{t=1}^{T}L(\tilde{p}_t)-\min_{p\in S}\sum_{t=1}^{T}L(p)$$

其中，p 表示在全局范围内产生损失最小的最佳单动作，$\tilde{p}_t$ 表示 r 方在 t 阶段选择的行动。

该定义表示遗憾是学习器在时间范围 T 内采取实际动作 $\tilde{p}_t$ 产生的累积损失与事后来看选择最佳固定动作产生的累积损失之差。Auer 等[27]称该遗憾为弱遗憾，因为该定义中事后理性采取的最佳动作在时间范围 T 内是固定的，并不是每个时间阶段 t 内针对对手策略的最佳动作。如果遗憾是通过期望整合动作采样的随机性来获得的，则使用期望遗憾来衡量算法性能[36]。在在线线性优化问题中，如果对手为健忘型[30]，则期望遗憾的定义如下：

$$R_T=\mathrm{E}[\sum_{t=1}^{T}L(\tilde{p}_t)]-\min_{p\in S}\sum_{t=1}^{T}L(p)$$

该期望遗憾也被称为策略遗憾[31]。如果对手为非健忘型，那么期望遗憾的定义如下：

$$R_T=\max_{p\in S}E[\sum_{t=1}^{T}L(\tilde{p}_t)-\sum_{t=1}^{T}L(p)]$$

当 $T\to\infty$ 时，如果算法可以确保 $R_T T\to 0$，则认为该算法是遗憾最小化算法。

当前面向 MAB 问题的算法性能衡量指标主要包括期望遗憾和高概率遗憾。现有大多数研究 MAB 问题的文献通常利用期望整合动作采样的随机性以获得遗憾界，故使用期望遗憾来衡量算法性能[36]。Auer 等[27]提出的 Exp3 算法获得的遗憾界，就是期望遗憾界。此外，一些研究采用高概率遗憾界来衡量算法的性能[20]。相对于期望遗憾界，高概率遗憾界的求解引入了统计分析，得到了概率下可确保的界。

对抗型赌博机问题中不同对手的期望遗憾定义不同。针对健忘型对手，事后遗憾是根据时间范围 T 内产生损失最小的最佳单动作来计算的；而针对适变型对手，事后遗憾根据每个时间阶段 t 产生损失最小的动作来计算[29,31]。学术界将的针对健忘型对手的期望遗憾定义称为伪遗憾(pseudo-regret)或弱遗憾(weak regret)[27,36]，其中考虑了对手的强弱水平。相对于期望遗憾界，高概率遗憾界的获得相对困难，有时需要对原有的算法进行重大修改或者在遗憾证明过程中采用大量复杂的技巧[36]。针对上述对抗型赌博机问题，Auer 等[27]提出了一种 Exp3 算法的变体 Exp3.P，其获得的高概率遗憾界基本和 Exp3 算法实现的期望遗憾界一样好。Neu 针对健忘型对手赌博机问题提出 Exp3-IX 算法[36]，其以高概率 $1-\delta$ 实现了 $R_T\leqslant O(\sqrt{(\log D+\log(1/\delta))TD})$ 的遗憾上界。Abernethy 等[37]基于高概率遗憾保证的机制，提出了一个面向赌博机问题的通用框架(包括适变型对手的赌博机问题)，以获得高概率遗憾界。表 1 总结了面向不同多臂赌博机问题的经典代表算法的性能表现。

表 1 不同多臂赌博机问题的经典算法性能表现

Table 1 Performance of classical algorithms for different multiple arm gambling machine problems

	完全信息	半信息	赌博机	随机型	对抗型	期望\高概率遗憾界
α-UCB[24]	√			√		$O(\sum(\ln T))$
Exp3[27]			√		√	$O(\sqrt{DT\ln D})$
Exp3-IX[36]			√		√	$O(\sqrt{(\log D+\log(1/\delta))TD})$
OMD[38]			√		√	$O(\sqrt{T})$
FPL[39]		√			√	$O(m\sqrt{DT\log D})$
Hedge[40]			√		√	$O(D^{3/2}\sqrt{T})$

3 布洛托上校博弈典型模型

Gross 等[12]是早期比较系统地研究布洛托上校博弈问题的学者，其面向简单的战场兵力配置问题，设置了 3 种不同的对抗条件。对于战场数量 $n\geqslant 3$ 的情形，设置对抗双方拥有数量相等的资源预算，也称之为对称预算，反之为非对称预算；若对同一个战场的价值评估相等，则称之为同质(homogeneous)战场，反之称为异质(heterogeneous)战场；不同战场之间价值相等，则称之为等价战场，反之称为非等价战场。

布洛托上校博弈与军事资源分配问题密切相关，特别是异质资源的布洛托上校博弈模型，是探索作战资源分配策略优劣的有效工具。

布洛托上校博弈中，局中人同时分配兵力至 n 个战场，如果 A 的分配方案 $(a_1,\cdots,a_m)$ 满足 $\sum_{i\in[n]}a_i\leqslant X^A$，$B$ 的分配方案 $(b_1,\cdots,b_m)$ 满足 $\sum_{i\in[n]}b_i\leqslant X^B$，那么局中人 $\phi\in\{A,B$ 在战场 i 上的收益为 $R_i^\phi(a_i,b_i)$，总收益为 $\sum_{i\in[n]}R_i^\phi(a_i,b_i)$。

布洛托上校博弈 $G=(N,(S^j)_{j\in N},(u^j)_{j\in N})$ 是一个资源分配博弈的条件是：对于任意 $j\in N$，存在 $n^j,X^j\in(0,\infty)$，使得 S^j 和 $\{(s_1^j,\cdots,s_{n^j}^j):\sum_{i\in[n_j]}s_i^j\leqslant X^j\}\subseteq R^n$ 之间是一一映射。

根据这个定义，局中人可以选择的动作以 n 元组的形式出现，X^j 是局中人 j 的有限预算，将局中人 j 的预算约束条件称为 $\sum_{i\in[n_j]}s_i^j\leqslant X^j$。

局中人 A 和 B 之间进行单次博弈，每个局中人可用资源预算分别为 X^A 与 X^B。不失一般性，假设局中人同时分配资源至 n 个战场($n\geqslant 3$)。局中人在每个战场 $i\in[n]$ 上的值对应两个参数 $w_i^A,w_i^B>0$。局中人的纯策略 $P\in\{A,B\}$ 可表示为一个向量 $\boldsymbol{x}^P=(x_i^P)_{i\in[n]}\in R_{\geqslant 0}^n$，满足预算约束 $\sum_{i=1}^n x_i^P\leqslant X_P$。

在每个战场 i，当局中人 P 分配的资源严格多于对手时，其将完全赢得相应战场上的值 w_i^P，对手无收益。对于平局情形 $x_i^A=x_i^B$，局中人 A 的收益为 αw_i^A，局中人 B 的收益为 $(1-\alpha)w_i^B$，其中 $\alpha\in[0,1]$ 为固定参数。对于纯策略组合 (x^A,x^B)，局中人 A 与 B 的收益为：

$$\Pi_A(x^A,x^B)=\sum_{i=1}^n w_i^A\cdot\beta_A(x_i^A,x_i^B)$$

$$\Pi_B(x^A,x^B)=\sum_{i=1}^n w_i^B\cdot\beta_B(x_i^A,x_i^B)$$

对于所有 $x,y\in R_{\geqslant 0}$，布洛托函数 β_A 与 β_B 定义如下：

$$\beta_A(x,y)=\begin{cases}1, & \text{if } x>y\\ \alpha, & \text{if } x=y\\ 0, & \text{if } x<y\end{cases}$$

$$\beta_B(x,y)=\begin{cases}1, & \text{if } y>x\\ 1-\alpha, & \text{if } y=x\\ 0, & \text{if } y<x\end{cases}$$

由于应用场景和条件假设不同，相应的布洛托博弈模型也有差异，如表 2 所列。

表 2 不同条件假设下的布洛托博弈模型

Table 2 Colonel Blotto game models with different assumptions

模型	主要特点
连续布洛托上校博弈	策略空间连续
离散布洛托上校博弈	策略空间离散
广义布洛托上校博弈	资源预算、战场价值评估等条件不作约束
广义乐透布洛托博弈	引入竞争成功函数，利用概率以描述不确定性
广义规则布洛托上校博弈	考虑资源预置与非对称资源效果

3.1 连续布洛托上校博弈

对于连续布洛托上校博弈，通常假设每个局中人的收益

函数是连续的，局中人的混合策略为其行为策略空间上的Borel概率测度。如对局中人A，其混合策略集表示为Δ_X，则混合策略的支撑集为：

$$Supp(p):=\bigcap\{K\subset X|K\text{ 为紧集},p(K)=1\}$$

对于两人博弈，根据支撑集的大小，混合策略$p\in\Delta_X$可分类为：纯策略p，即对于一些$x\in X$，有$Supp(p)=x$，p为Dirac测度δ_x；有限支撑集混合策略，即支撑集的大小是有限的，p可以写成凸组合$p=\sum_{x\in Supp(p)}p(x)\delta_x$；无限支撑集混合策略。

两人博弈中，对于混合策略$(p,q)\in\Delta$，其中$\Delta:=\Delta_X\times\Delta_Y$，局中人的平均收益为：

$$\Pi_A(p,q):=\int_{X\times Y}\beta_A(x,y)d(p\times q)$$

当支撑集$Supp(p)$和$Supp(q)$均有限时，则：

$$\Pi_A(p,q):=\sum_{x\in Supp(p)}\sum_{y\in Supp(q)}p(x)q(y)\beta_A(x,y)$$

Adam等[41]设计了用于连续布洛托上校博弈的均衡计算双重Oracle方法，Ganzfried[42]设计了面向连续布洛托上校博弈近似纳什均衡的冗余虚拟对弈方法。

3.2 离散布洛托上校博弈

现实世界中的很多资源不可分割，如战场对抗环境下的兵力资源[12]，以及西方政治选举竞争中的人力及物力等资源[43]。考虑资源不可分割的情形，离散布洛托上校博弈(Discrete Colonel Blotto，DCB)记作CB_n^D，局中人的预算与分配约束都是整数，即$X^A,X^B\in N\backslash\{0\}$，则局中人$\phi\in\{A,B\}$的策略集合为$(x_1^\phi:x_n^\phi):x_i^\phi\in N,\forall i\in[n]$，且$\sum_{j\in[n]}x_j^\phi\leqslant X^\phi$。

关于CB_n^D的研究主要聚焦在同质战场(即常和布洛托上校博弈)的条件下，且主要致力于降低求解方案的复杂性[44]，因为CB_n^D中博弈双方的纯策略数量随着战场数量和资源预算的增长呈指数级增长，所以，虽然布洛托上校博弈的基本规则简单，但其模型本身所涉及的条件假设可以十分复杂，故布洛托上校博弈问题的求解十分困难。目前的研究主要聚焦在CB_n^C和常和CB_n^D等部分限制性条件下，对CB_n的问题的求解至今仍充满挑战。

3.3 广义布洛托上校博弈

n个战场且资源预算为X_A和X_B的广义布洛托上校博弈(Generalized Colonel Blotto，GCB)记作$GCB_n^{X_A,X_B}$，局中人$P\in\{A,B\}$的策略集合为$x^P\in R_{\geqslant 0}^n:\sum_{i=1}^n x_i^P\leqslant X_P$，局中人分别采用纯策略$x^A$和$x^B$时，收益为$\Pi_P(x^A,x^B)$。

在GCB_n中，条件假设非常宽泛，对对抗双方的资源预算、战场价值评估等条件不作约束。对GCB_n比较完整的定义最早由Kovenock等[45]提出，Vu等[46]在Kovenock等的基础上对单个战场的胜负规则(布洛托函数)作了更一般化的假设，用因子α来表示一方的收益系数，具有更广泛的表示意义。由于放松了约束，GCB_n的问题求解(通常指其纳什均衡求解)更为复杂。

当对抗双方对同一战场的价值评估相等，其他条件和GCB_n相同时，双方对抗之后的收益之和为战场价值之和，是一个固定的值，故称该博弈为常和布洛托上校博弈，记为GCB_n^C。零和博弈可以看成是收益之和为零的特殊常和博弈。Roberson[47]提出了求解等价战场、双方预算比值满足特定范围问题的方法。Schwartz等[48]扩展了Roberson的研究，放松了等价战场的约束，考虑不等价战场情况下的GCB_n^C问题。

3.4 广义乐透布洛托博弈

针对布洛托上校博弈中赢者通吃的规则过于苛刻的问题，广义乐透布洛托博弈(Generalized Lottery Blotto，GLB)(记作$GLB_n(\zeta)$)假设每个局中人均有一定概率获得相应的收益。原始赢者通吃布洛托函数被改成基于竞争成功函数(Contest Success Functions，CSF)：$\zeta_A,\zeta_B:R_{\geqslant 0}^2\to R$。对于纯策略组合$x^A$和$x^B$，博弈各方收益分别为：

$$\Pi_\zeta^A(x^A,x^B)=\sum_{i=1}^n w_i^A\cdot\zeta_A(x_i^A,x_i^B)$$

$$\Pi_\zeta^B(x^A,x^B)=\sum_{i=1}^n w_i^B\cdot\zeta_B(x_i^A,x_i^B)$$

Vu等[49]利用单变量分布构建混合策略的方法给出了广义乐透布洛托博弈的近似均衡求解方法。Li等[50]给出了两方广义乐透布洛托博弈的纯策略纳什均衡求解方法，其中博弈局中人之间的预算和估值都是不对称的，博弈采用了Tullock竞争成功函数[51]。

3.5 广义规则布洛托上校博弈

如何泛化赢者通吃规则是广义规则布洛托上校博弈(General Rule Colonel Blotto，GRCB)的核心。其中，Vu等[52]研究了偏袒(favoritism)布洛托上校博弈，将资源预置(pre-allocation)与非对称资源效果(asymmetric resource's effectiveness)看成某种形式上的偏袒。对于n个战场的偏袒布洛托上校博弈，记作CB_n^F，局中人A和B的纯策略为$x^A=x_{i_{i\in[n]}}^A\in S^A$和$x^B=x_{i_{i\in[n]}}^B\in S^B$。对应收益函数分别为：

$$\Pi_{CB_n^F}^A(x^A,x^B)=\sum_{i\in[n]}w_i\beta(x_i^A,q_ix_i^B-p_i)$$

$$\Pi_{CB_n^F}^B(x^A,x^B)=\sum_{i\in[n]}w_i[1-\beta(x_i^A,q_ix_i^B-p_i)]$$

其中，$\beta:R_{\geqslant 0}^2\to[0,1]$满足：

如果$x>y$，则$\beta(x,y)=1$；

如果$x=y$，则$\beta(x,y)=\alpha$；

如果$x<y$，则$\beta(x,y)=0$。

Aspect等[53]研究了带热启动的布洛托上校博弈模型。Gupta等[54]建立了面向网络安全的三阶段布洛托上校博弈模型。Paarporn等[55]将信息的隐藏、混淆和泄漏建模成两阶段布洛托上校博弈模型。

3.6 在线离散布洛托上校博弈

在线离散布洛托上校博弈中，学习器与对手在n个战场上对抗T个阶段，在阶段$t\in[T]$，每个战场$i\in[n]$的价值$b_t(i)>0$，满足$\sum_{i=1}^n b_t(i)=1$，学习器的策略向量满足：$\boldsymbol{S}_{k,n}:=\{z\in N^n:\sum_{i=1}^n z(i)=k\}$。对于一个给定战场数量为$n$、部队数量为$k$的在线离散布洛托上校博弈，可以相应构建一个有向无环图，在线离散布洛托上校博弈的策略集合$S_{k,n}$与图上从起点s至终点d之间的所有路径集合$G_{k,n}$一一映射。因此，完全信息反馈对应可以观测边上的所有损失，半赌博机反馈对应仅可观测到所选路径上边的损失，赌博机反馈对应仅能观测到所选路径上边的聚合损失。此外，对于半赌博机反馈，学习器可以通过侧面观测图的结构来推导出相关信息[56]。需要注意的是，这类侧面观测与多臂赌博机的“侧面观测”[21]

不同，前者表示图中边的观测，后者表示路径（行动、路径）的观测。在线离散布洛托上校博弈信息反馈方式如表 3 所列。

表 3 在线布洛托上校博弈信息反馈设置

Table 3 Online Colonel Blotto game information feedback

信息反馈	对手策略	战场价值	战场结果	战场总损失
完全信息	√	√	√	√
侧面观测	部分可推理	√	√	√
半赌博机		√	√	√
赌博机				√

通过对 Gyorgy 等[20]定义的网络路由问题对应的图结构进行改进，可以将在线布洛托上校博弈问题转化为在线最短路径问题。在 OSP 问题中，学习器的动作集是有向无环图（Directed Acyclic Graph，DAG）上从源头到目的地的一组路径[57]。OSP 可以用一个 DAG 定义，DAG 有以下属性：有两个特殊的顶点，即源点和目的点，分别表示为 s 和 d；P 表示从 s 到 d 的所有路径集，DAG 的顶点集和边集分别用 V 和 E 表示。设置 $|V|\geqslant 2$ 以及 $|E|\geqslant 1$，且每条边 $e\in E$ 至少属于一条路径 $p\in P$。用 n 表示集合 P 中最长的路径长度，即 $\|p\|\leqslant n, \forall p\in\{P\}$。

给出时间范围 $T\in N$，在线最短路径问题（Online Shortest Path，OSP）定义如下：在阶段 $t\in[T]$，每条边 $e\in E$ 对应一个由对手确定的标量损失 $l_t(e)\in[0,1]$，学习器在未知标量损失的情况下选择一条路径 $p_t\in P$，并产生该路径所有边总的损失 $L_t(p_t)=\sum_{e\in p_t} l_t(e)$。在阶段结束时，学习器会观察到一些反馈（该路径上每条边的损失或者仅该路径上所有边损失的和）。学习器的目标是最小化期望遗憾值。这里定义的在线最短路径问题是在线组合优化问题的一个实例（具有动作向量的维度是 $|E|$），其可以看作映射每条路径 $p\in P$ 到 $\{0,1\}^{|E|}$ 的一个向量，所有与在线组合优化相关的概念都可用于在线最短路径，包括遗憾、期望遗憾、健忘型、非健忘型的对手以及学习器在每个阶段结束时观察到的反馈设置。

4 布洛托上校博弈求解方法

4.1 博弈策略空间形态探索

Czarnecki 等[58]结合实证博弈策略分析理论分析了多类博弈的策略空间形态，提出了陀螺猜想。假定策略空间为分层的博弈几何体，即策略空间 Π 可分解成 k 层，$\bigcup_i L_i=\Pi$，其中 $\forall_{i\neq j} L_i\cap L_j=\varnothing$，层之间的策略满足传递性压制关系，$\forall_{i<j,\pi_i\in L_i,\pi_j\in L_j} f(\pi_i,\pi_j)>0$，且存在 $z\in N$ 满足 $i<z$ 时有 $|L_i|\leqslant|L_{i+1}|$，$i\geqslant z$ 时有 $|L_i|\geqslant|L_{i+1}|$。

布洛托上校博弈的策略空间形态如图 3 所示，分别选定“10 资源 5 战场”博弈场景，首先获取种群策略之间的收益矩阵，通过计算收益矩阵的“石头-剪刀-布”（Rock-Paper-Scissors，RPS）环和“纳什聚类中心”（Nash Clusterings）来分析策略空间形态。对于对称零和博弈，由于决策空间（战场）具备置换不变性，布洛托上校博弈要求局中人在所有可能的排列中均匀地混合以避免被利用。面向布洛托上校博弈的学习方法生成的策略也具备非传递性。

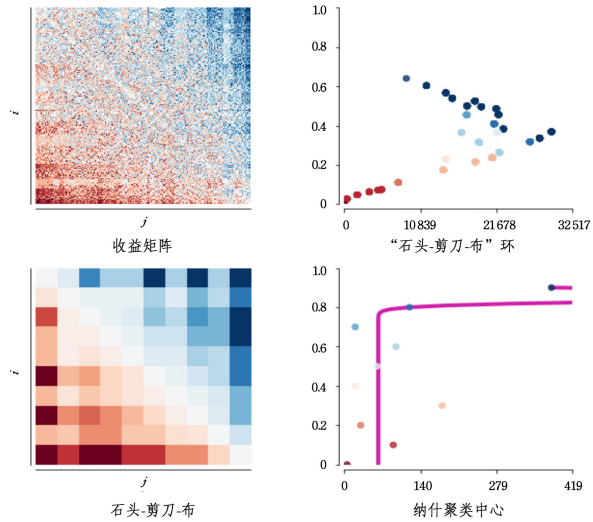

图 3 多类布洛托上校博弈的策略空间形态

Fig. 3 Strategy space form of multiple kinds of Colonel Blotto game

Omidshafiei 等[59]围绕收益张量，采用 α-rank 响应图分析探索了各类博弈策略的空间形态。其中布洛托上校博弈的相关结果如图 4 所示。对于“5 资源 3 战场”博弈场景，根据收益矩阵对各类策略进行评级，获得各策略之间的响应图，然后采用图统计分析-主成分或图谱分析-聚类-压缩等步骤，将各类博弈策略映射到参数化博弈结构中。

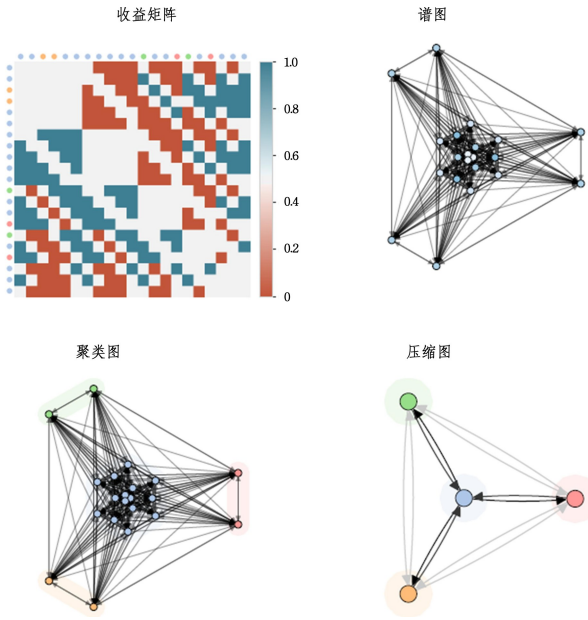

图 4 布洛托上校博弈的策略空间形态

Fig. 4 Strategy space form of Colonel Blotto game

布洛托上校博弈求解仍然面临诸多挑战。下面区分 2 个阶段（离线与在线）3 类博弈场景（单次、重复、多阶段），梳理布洛托上校博弈求解方法。

4.2 离线单次博弈策略求解

从博弈论的角度出发，将基于布洛托上校博弈模型的单次对抗条件下的资源分配问题称为离线布洛托上校博弈。相关研究大部分建立在离线布洛托上校博弈条件下，致力于寻找该条件下的纳什均衡。其中，对抗双方对战场数量 n、对方的资源预算以及布洛托函数拥有共同知识假设。

Gross 等[12]研究了常和布洛托上校博弈中最简单的一种情形，即博弈双方的资源预算相等，战场数量 $n=3$，且设置布洛托函数在双方分配策略相等时平分战场价值，即作者构造的布洛托上校博弈是对称博弈。现有的很多研究针对离散布洛托上校博弈问题进行求解。与广义布洛托上校博弈不同，离散布洛托是一个有限式博弈。因为要求资源预算以及局中人的分配策略至少是整数的某个粒度，故局中人的分配策略数量是有限的。Hortala-Vallve 等[60]研究了对抗双方拥有相同预算的离散布洛托上校博弈问题，描述了纯策略纳什均衡存在时布洛托函数的形式（当双方分配相同的资源至同一个战场时，双方平均分得该战场的价值）以及维持纳什均衡时对抗双方的代价。由于离散布洛托上校博弈的策略空间随着战场数量和预算数量的增加呈指数级增长，因此相关研究主要聚焦在同质战场（即常和布洛托上校博弈）条件，且主要致力于降低求解方法的计算复杂性。

4.2.1 线性规划求解

Ahmadinejad 等[61]研究了常和离散布洛托上校博弈问题，使用线性规划方法来寻找最优策略。Behnezhad 等[44]提出了第一个多项式时间算法，发现即使是单纯形方法（尽管它的指数运行时间），在实际应用中也比椭球体方法表现得更好。

假设 $N=\{r,m\}$，u^r 和 u^m 分别表示局中人 r 和 m 益函数；$s^r\in S^r$ 和 $s^m\in S^m$ 分别表示 r 和 m 纯策略集合中的纯策略；用 $p:S^r\to[0,1]$ 表示局中人 r 某个纯策略对应的概率，且满足 $\sum_{s^r\in S^r}p_{s^r}=1$；同理，用 $q:S^m\to[0,1]$ 表示局中人 m 某个纯策略对应的概率。由于离线布洛托上校博弈是常和博弈，因此可以利用极小极大原理找到该博弈的一个纳什均衡。根据混合纳什均衡概念，如果 p^* 和 q^* 能够同时最大化 r 和 m 的收益，则由 (p^*,q^*) 构成的混合策略是离线布洛托上校博弈的混合纳什均衡。通过求解以下线性规划方程式，可以获得 p^*；同理，通过类似的线性规划方程式可以获得 q^*。

$$\max\ u^r$$
$$\text{s.t.}\ \begin{cases}\sum_{s^r\in S^r}p_{s^r}=1\\ \sum_{s^r\in S^r}p_{s^r}h_m^r(s^r,s^m)\geqslant u^r,\ \forall\, s^m\in S^m\end{cases}$$

由于 $|S^r|$ 和 $|S^m|$ 是关于战场数量和资源预算的指数函数，单纯地采用上述线性规划的方法面临计算效率缓慢的问题，因此该形式下的线性规划方法在实际中并不适用。

4.2.2 最优单变量耦合

Roberson[47]研究了更具一般性的常和布洛托上校博弈中的纳什均衡求解问题，在未限制战场数量以及博弈双方的资源预算是否相等的情况下，提出了一个基于 Copula 理论的博弈求解方法，获得了该条件下博弈双方的最优单变量分布。但是，Copula 理论的表述十分复杂，实现起来十分困难。Kovenock 等[45]研究了常和布洛托上校博弈的最优单变量分布的求解，给出了一类由特殊方程的正数解描述的最优单变量分布，但是仍然无法从这些最优单变量分布集合中构造满足预算约束的联合分布。Schwartz 等[48]针对任意数量战场但等价的常和布洛托上校博弈问题，求解出唯一的最优单变量分布，并证明了存在由这些最优单变量分布的 n 变量联合分布，但是未能求解出该联合分布。

在常和离线布洛托上校博弈中，如果一组单变量分布 $\{F_i^r,F_i^m\}_{i\in n}$ 满足以下两个条件，则称其为最优单变量分布。

1）如果局中人 r 根据分布 F_i^r 分配资源 x_i^r 至战场 i，$i\in[n]$，则其分配资源的期望满足预算约束 X^r，即：

$$\sum_{i\in[n]}[\mathbb{E}_{x_i^r\sim F_i^r}x_i^r]\leqslant X^r$$

2）如果局中人 m 根据 F_i^m 分配资源给战场 i 的同时，局中人 r 根据 F_i^r 分配资源给战场 i，则局中人 r 的策略是最优的。换句话说，对于局中人 r 的任意纯策略 $\tilde{x}^r$，所有战场期望收益的和满足以下不等式：

$$\sum_{i\in[n]}\mathbb{E}_{x\sim F_i^r,y\sim F_i^m}[W_i^r\beta^r(x,y)]\geqslant\sum_{i\in[n]}\mathbb{E}_{y\sim F_i^m}[W_i^r\beta^r(\tilde{x}_i^r,y)]$$
$$\sum_{i\in[n]}\mathbb{E}_{x\sim F_i^r,y\sim F_i^m}[W_i^m\beta^m(x,y)]\geqslant\sum_{i\in[n]}\mathbb{E}_{x\sim F_i^r}[W_i^m\beta^m(x,\tilde{x}_i^m)]$$

其中，W_i 表示战场 i 的价值，β 为布洛托函数。

基于最优单变量求解常和离线布洛托上校博弈的方法基本上可以归结为以下两个步骤：

1）确定每个战场上局中人的最优单变量分布，即放松预算约束，只保持期望相同，并在每个战场上寻求最优分配策略；

2）构造一个在 1）中发现的单变量分布的 n 变量联合分布，使从该联合分布中提取的任何策略都满足预算约束。

上述构造最优单变量的求解方法难度较大，而且往往只能针对特定条件下（如常和＋离散＋等价战场＋相同预算等）的离散布洛托上校博弈问题进行求解，普适性的结论往往很难获得。此外，全支付拍卖（All-Pay Auction，APA）的均衡刻画经常被用作研究布洛托上校博弈均衡的工具（用于在上述两个步骤中的第 1 步中来构造最优的单变量分布）。

在一个全支付拍卖中，竞标者秘密决定他们各自的出价来竞争同一个物品，出价最高者赢得该物品并获得其价值，且所有竞标者支付他们各自的出价（包括竞得物品的获胜者）[62-63]。通过上述分析可知，对于这类构造单变量的求解方法，普适性的结论往往很难获得。很多学者通过研究发现，可以构建每个战场上局中人的最优单变量分布，但不能确定是否存在由这些分布构造的混合策略。通过构造最优单变量的方法求解常和布洛托上校博弈的纳什均衡，仍面临很大的挑战。

4.2.3 动态策略迭代

Mcmahan 等[64]提出了利用子博弈增量迭代方式求解博弈的双重预言机（Double Oracle，DO）方法。Lanctot 等[65]将 DO 方法与深度强化学习方法结合，提出了策略空间响应预言机（Policy Space Response Oracle，PSRO）。Zou 等[66]基于 DO 算法提出面向离线布洛托上校博弈求解的 ϵ-DO 算法，使其能够求解离线布洛托上校博弈的近似纳什均衡。

Gemp 等[67]提出了基于策略梯度的大型多人博弈纳什均衡近似方法 ADIDAS，其与同伦型、李雅普诺夫和迭代式多矩阵博弈求解方法类似。Anthony 等[68]采用元博弈理论，利用自对弈强化学习设计了随机虚拟对弈的最佳响应策略迭代方法 FPPI。Jacob 等[69]基于无憾学习与在线凸优化理论，提出了面向类人对抗的 KL 散度正则化搜索方法 pi-KL。Strang 等[70]利用主成分权衡分析方法，重点分析了 Blotto 博弈的

对称性，基于 Disc 博弈的“平衡”结构表示分析了策略之间的互锁循环部分。Bertrand 等[71]提出了基于技能和一致性的扩展版 Elo 策略来评估方法，在多类投影后的策略空间中分析各类策略的分布情况，如图 5 所示。

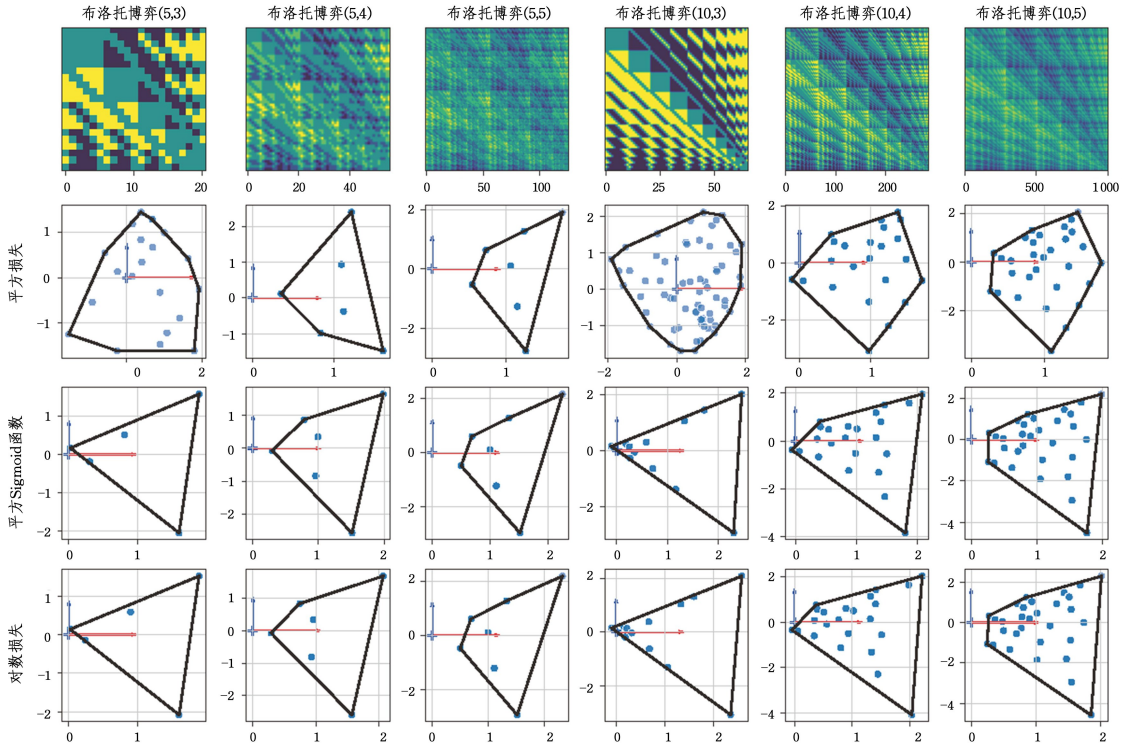

图 5 布洛托上校博弈的策略传递性与循环性分析

Fig. 5 Analysis of strategic transitivity and cyclicity in Colonel Blotto game

4.3 在线重复博弈策略求解

4.3.1 重复性对抗场景

在线布洛托上校博弈主要描述了与对手之间进行多次资源分配对抗的情形。在现实情况中，如果资源是重复可利用的，则每一次对抗双方的预算都和上一阶段相同(剩余部分资源不会为对抗结果带来好处)。假设博弈的其他共同参数在每次对抗中也是相同的，则双方之间的多轮次对抗可以看成是重复博弈。

将基于布洛托上校博弈模型的在线资源分配问题称为在线布洛托上校博弈，主要描述学习器和对手之间进行多次资源分配对抗的情形。对于其中的每一次对抗，学习器在不知道部分信息的条件下做出决策(如战场价值或者当前对手的预算等)，当一次对抗结束之后，学习器接收部分信息反馈(如分配策略的收益等)。在这种情况下，学习器通常需要连续地动态学习，并调整利用已知信息和探索获得新信息之间的权衡，生成良好策略以实现最小化累积损失(或者最大化累积收益)的目标。在线布洛托上校博弈本质上是具有组合结构的在线资源分配问题，是在线组合优化问题的一个实例。

4.3.2 Exp3 类算法

在上述经典 MAB 问题中，一次只摇一支臂；在组合多臂赌博机问题中，一次拉动的不是一支臂，而是多支臂组成的集合，称之为超臂(super arm)[72]。拉完这个超臂后，超臂所包含的每个基准臂会给一个反馈，而这个超臂整体上会给学习器带来某种复合的收益。因此，在线布洛托上校博弈可以直接建模为组合多臂赌博机(Combinatorial Multi-Armed Bandit，CMAB)模型，其中每个纯策略对应一支超臂，同时 CMAB 模型可以完全捕捉到关于博弈的信息反馈。信息反馈量的大小决定了 CMAB 问题属性，类似于博弈论中完全信息和不完全信息问题的区别，对应的求解方法迥异。由于 CMAB 问题仍然隶属于在线线性优化问题的框架，因此，类比于在线线性优化问题，针对完全信息反馈的在线资源动态分配问题，Freund 等[28]提出了 Hedge 算法，实现了 $O(\sqrt{T\log|S|})$次线性遗憾界，其中 $S\subset\{0,1\}^D$，其是使用二进制方式来表示动作的一种方法；Koolen 等[73]对全信息反馈的在线线性优化问题进行深入研究，提出了一种 Hedge 算法的变体，实现了 $O(\sqrt{Tn\log(D/n)})$遗憾界，其中 $n=\max_{p\in S}\{\|p\|_1\}$。针对赌博机反馈，Dani 等[40]提出了几何 Hedge 算法，获得了 $O(D^{3/2}\sqrt{T})$的期望遗憾界，这是首个获得关于时间范围 T 的次线性期望遗憾界的算法。另一个针对赌博机反馈的在线线性优化问题的算法是在线镜像下降算法，该算法基于凸优化中的镜像下降(Mirror Descent，MD)方法进行构造，并首次由 Cesa-Bianch 等[30]分析了镜像下降和在线学习之间的联系，之后 Abernethy 等[38]提出首个针对赌博机反馈的在线线性优化问题的镜像下降算法。由于经典的 MAB 问题中，一次只摇一支臂，因此该问题下的信息反馈符合完全信息反馈和赌博机反馈的情形；而半赌博机信息反馈则需要在组合多臂赌博机问题下才会发生，故针对半赌博机信息反馈的研究并不多。Alon 等[74]将不同程度的赌博机反馈信息类型建模为图结构，并提出强可观察图、弱可观察图和不可观察图的定义，分别对应于上述 3 种不同的信息反馈情形。

4.3.3 在线最短路径

Exp3 算法是一种面向多臂赌博机问题的遗憾最小化算法，经过改进可以实现组合多臂赌博机问题的遗憾最小化。在线学习问题通常对算法有实时性要求，借助在线最短路径问题(Online Shortest Path，OSP)良好的图结构，可以实现 Exp3 算法高效的运行效率。Gyorgy 等[20]研究了经典的路由问题，需要在路由网络上依次选择路径进行数据包的传输，实现时间范围 T 内的数据包传输累积时间的最小化；其同时考虑了完全信息反馈和半赌博机信息反馈的情形。Cesa-Bianchi 等[75]考虑了半赌博机信息反馈和赌博机反馈的在线最短路径问题，并分析了对路径进行均匀采样是不可取的。Vu 等[76]针对具有组合结构的在线布洛托资源分配问题，提出作用于有向无环图的 Edge 算法，实现了关于时间范围 T 的次线性期望遗憾上界，同时该算法的运行时间要优于经典的 COMBAND 算法。以上研究以多臂赌博机为实例，研究了不同问题类型下的求解方法，为在线布洛托上校博弈问题的建模与求解提供了良好支撑。

4.4 在线多阶段博弈策略求解

4.4.1 多阶段对抗场景

在上一节讨论的在线布洛托上校博弈问题中，学习器不需要关心如何将总预算分配至各个时间阶段 t，因为每个时间阶段 t 的预算(称为阶段预算)都是相同的，或者说没有总预算的约束，只考虑阶段预算，而且每次对抗的阶段预算都相等，而博弈的其他共同参数在每次对抗也是相同的，故双方之间在时间范围 T 内的对抗其实是重复博弈。重复博弈中的资源可以是非消耗性资源，具备重复使用的特点，然而在现实对抗性活动中，消耗性资源也是普遍存在的，此时，局中人之间的对抗形式会发生变化。假设在时间范围 T 内存在总预算约束，则学习器需要兼顾两个层面的问题。

1)在上层级，学习器面临在时间范围 T 内的阶段预算分布优化问题，即如何将总预算分配至各个时间阶段 t。学习器基于观察到的历史信息反馈进行分配，各个时间阶段的预算是相关的，某些时间阶段分配的预算增加(相对于平均数)意味着其他些阶段分配的预算会减少。

2)在下层级，学习器在每个时间阶段 t 和对手(类型为健忘型或者非健忘型对手[30])进行单次对抗条件下的布洛托上校博弈。该问题和重复博弈条件下的布洛托上校博弈问题不同的是，学习器在时间范围 T 内的每个阶段预算可能不同。而其中的单次对抗问题，仍然属于组合多臂赌博机问题框架，涉及到多个战场之间的资源分配，必须整体考虑分配的合理性，任何仅仅专注于一个战场的做法都是不可取的。这类似于组合多臂赌博机问题中，如何分配有限的阶段预算给超臂(将多个赌博臂称为超臂)，从而获得可观的收益。

4.4.2 赌博机背包

将上述包含两个层级问题的在线布洛托上校博弈称为多阶段对抗条件下的资源分配问题。目前对于在线布洛托上校博弈问题，大部分的研究如上节中所述，是基于固定的阶段预算进行，即不考虑上层级的时间范围 T 内的阶段预算分布优化问题，学习器在时间范围 T 内的每个阶段预算都是相等的，利用下层级算法在时间范围 T 内进行重复博弈实现遗憾最小化，而考虑多阶段对抗条件下资源分配问题的不多。与多阶段对抗条件下资源分配问题相近的研究为受背包容量约束的赌博机背包(Bandits with Knapsacks，BwK)问题，其是一个在供应/预算限制下的多臂赌博机问题的一般模型[77]。

BwK 问题主要包括随机型背包赌博机[77]、对抗型背包赌博机[78]、情境型背包赌博机[79]、非平稳背包赌博机[80]等。背包赌博机问题的典型应用场景[81]有赌博机背包、重复斯塔克尔伯格博弈背包、预算受控重复首价拍卖等[82]。

Ashwinkumar 等[83]提出了两种基于线性规划的随机 BwK 问题求解方法。Agarwal 等[84]将随机型 BwK 设置推广至分别包括任意凹收益和任意凸约束的情形，并分别获得 $O\left(L|\mathbf{1}_d|\sqrt{\frac{D}{T}\ln\left(\frac{DTd}{\delta}\right)}\right)$ 和 $O\left(|\mathbf{1}_d|\sqrt{\frac{D}{T}\ln\left(\frac{DTd}{\delta}\right)}\right)$ 的高概率遗憾界。其中，d 是收益或者损失的维度，L 为 Lipschitz 常数。Agarwal 等[79]研究了一种上下文 BwK 问题，上下文由未知的独立同分布线性产生，学习器据此进行动作采样，目的是在每种资源预算约束内实现收益最大化并获得高概率遗憾界。Agarwal 等[85]针对 Badanidiyuru 等[86]研究的上下文 BwK 问题，在 Agarwal 等[87]提出的算法上做出改进，获得的高概率遗憾界比 Badanidiyuru 等[86]提出的方法提高了$\sqrt{a}$因子(a 表示资源种类)，而且其算法运行时间是策略集大小的对数平方根。Immorlica 等[78]研究了包括随机型、对抗型、全信息反馈、半赌博机信息反馈以及上下文赌博机在内的多种不同问题下的 BwK 问题，并设计 LagrangeBwK 算法求解对抗型 BwK 问题，获得了高概率遗憾界。Li 等[88]采用原始-对偶的视角来研究 BwK 问题，从对偶的角度强调背包约束对遗憾的影响，并基于原始问题和对偶问题共同定义次优动作。

4.4.3 组合赌博机背包

Leon 等[89]研究了带总预算约束的动态布洛托上校博弈问题，学习器将有限的资源分配至有限的时间阶段，而在每个时间阶段，学习器和对手同时分配资源至多个战场进行对抗，每次对抗结束后，学习器仅仅能够获得总战场损失信息反馈。其目的是经过时间范围 T 次对抗后，实现期望遗憾的最小化。以上研究针对带背包容量约束的多臂赌博机问题进行建模和求解，结合在线布洛托上校博弈问题的研究，可以进一步求解在线多阶段布洛托上校博弈问题。

Sankararaman 等[90]研究了带背包容量约束的半赌博机反馈的组合赌博机背包(CombinatorialBwK，CBwK)问题，其收益服从固定的分布。文章指出，在传统的随机组合多臂赌博机问题中，立足于找到每一轮次中对应于最佳期望收益的动作即可；而在带背包容量约束的半反馈随机组合多臂赌博机中，其主要挑战在于需要求解所有轮次中最佳期望收益的动作分布，但受限于组合多臂赌博机问题中指数级的动作空间，该问题求解难度很大。

如上所述，在线组合优化是在线线性优化的一个实例，其中 $S\subset\{0,1\}^D$，即在线组合优化中每个动作都是一个 D 维的 0-1 向量。这和组合优化问题的变量形式也是契合的，因为组合优化问题本就是一类在离散状态下求极值的最优化问题[91]。相应地，在线线性优化中的有关定义可以转移到在线组合优化中来。学习器根据获取的信息反馈指导下一阶段的动作采样，目标是实现遗憾最小化。当前典型的布洛托上校博弈求解方法如表 4 所列。

表 4 典型布洛托上校博弈求解方法

Table 4 Typical solving methods for Colonel Blotto game

博弈场景	类别	基本思想及性能分析	主要参考文献
离线	线性规划求解	采用解析式方法计算精确解，但计算复杂性高，算法效率不高	文献[44,61]
离线	最优单变量耦合	利用最优单变量分布耦合得到边缘分布，但变量分布的独立性无法保证，近似解仍存在误差	文献[45,47-48]
离线	动态策略迭代	通常利用迭代式学习与分布式训练方式，但多人大型博弈均衡策略难学习，策略学习难保证收敛	文献[58-59,71]
在线	在线最短路径 Exp3	可以利用观测信息反馈优化在线策略更新，但无法很好地应对大型博弈空间	文献[66,76]
在线	赌博机背包 BwK	通常采用上下两层优化方式，同步优化两层目标，但很难同时优化策略与资源约束	文献[76,89]

5 研究前沿及展望

5.1 广义博弈模型

在广义布洛托模型的基础上，还有很多不同限制条件下的布洛托上校博弈变体模型，如斯塔克尔伯格博弈[92]、考虑资源预置的三阶段博弈[93]、连续博弈[94]、动态防御者-攻防者布洛托上校博弈[95-96]、贝叶斯广义布洛托上校博弈[97]、联盟布洛托上校博弈[98]、多维私人信息的多人布洛托上校博弈[99]、成比例资源分配布洛托上校博弈[100]、联网布洛托上校博弈[101]、布尔值型布洛托上校博弈模型[102]和私人布洛托上校博弈[103]等。当前围绕布洛托上校博弈模型的相关扩展主要聚焦以下几点：1)改造“赢者通吃”，设计新型布洛托函数，如采用兰彻斯特毁伤率[4]，建模比例资源分配的 Tullock 竞争成功函数；2)资源约束条件，如多阶段博弈中总体资源量受限[89]；3)博弈局中人数量，如针对多人(方)对抗场景的各类模型[99]；4)状态与动作空间连续，如针对多维连续状态空间[104]。

5.2 博弈求解方法

5.2.1 离线布洛托上校博弈求解

均衡点刻画仍然有待探索和解决，而构造最优单变量分布求解其纳什均衡存在以下缺点：一是理论比较复杂，求解难度大；二是不确定是否存在由最优单变量分布构造的混合策略；三是至今没有得出比较通用的结论，只在部分限制条件下得到结果。Perchet 等[104]提出了面向两人连续布洛托上校博弈的耦合与采样方法，可以求解非对称预算、异构非对称价值设置下的博弈。Beaglehole 等[105]提出的面向多人布洛托上校博弈的蒙特卡洛马尔可夫链(Monte Carlo Markov Chains，MCMC)采样方法，大大降低了计算时间复杂度。Noel 等[106]提出利用强化学习方法来求解布洛托上校博弈。此外，针对多人布洛托上校博弈的求解方法仍充满挑战。

5.2.2 在线布洛托上校博弈求解

现有对在线布洛托上校博弈的研究还很少，特别是对于半赌博机信息反馈的布洛托上校博弈问题，只能借助多臂赌博机等在线组合优化问题进行问题分析和求解。面向多臂赌博机问题的 Exp3 经典算法应用于在线布洛托上校博弈问题时需要进行改进，以获得遗憾界保证，同时兼顾算法的运行效率。将策略空间映射成有向无环图的策略求解方式，仍无法很好地扩展至策略行动空间较大的博弈模型中[89]。现有对多阶段对抗条件下在线布洛托上校博弈问题的研究还很少，而对于半赌博机信息反馈的情形处于空白。与该问题相近的赌博机背包问题研究中考虑对手设置的也很少，其收益服从固定的分布，属于随机多臂赌博机问题的范畴，不适用于对抗条件。此外，如何求解在线多阶段资源受约束条件下的布洛托上校博弈策略，如何为在线布洛托上校博弈的序贯策略学习 BwK 类问题设计有遗憾界保证、用于在线学习的元算法(Meta Algorithm)仍充满挑战[81]。

5.3 典型应用探索

基于布洛托博弈的典型应用主要包括：基于选区的政治选举资源投入[13,107]，面向互联网用户的广告投放[108]，无线电干扰对抗领域双方的功率分配[109]，空中无线传能中信令传输信道分配的隐私保护机制[110]，卫星通信智能抗干扰中多信道功率分配[111]。此外，如下领域的典型应用也值得持续关注。

1)竞争性资源分配

布洛托上校博弈模型能够在一定程度上近似模拟实际战场资源分配的对抗场景，通过求解不同对抗条件下多战场资源分配问题，充分合理利用战场资源，优化资源分配，为分布式环境下的资源分配决策提供技术和理论支撑。布洛托上校博弈分配方法结合兰彻斯特战损方程，可对“分而击之”等经典的战术战法进行验证。未来针对复杂问题(例如异构资源、异构战场)的求解，可为更复杂的战术战法的验证提供有效技术支撑。

2)通信频谱分配

相关研究将多维资源分配博弈问题建模成布洛托上校博弈[112]，多个网络服务提供商(Network Service Providers，NSPs)竞争向一组用户提供无线连接。用户可以是单个移动设备、一组本地化的物联网(IoT)设备，甚至是需要无线回程的校园网络。网络服务提供商之间相互竞争，通过分配可用带宽为用户提供无线服务，以实现总收益的最大化。NSPs 向每个用户提议使用一定带宽提供无线连接，然后用户决定连接到该 NSP，其提供的带宽使其收益最大化。

3)云服务调度

相关研究利用布洛托上校博弈模型来分析云服务系统的安全防御问题[113]，攻击者通过选择有限数量的 CPU 来攻击云服务设备，防御者需要寻找针对高级持久性威胁的防御策略。

4)复杂网络攻防

相关研究利用布洛托上校博弈描述网络上的攻防对抗问题[101,114]，将网络联通度、平均路径长度和传输能力作为网络性能评估指标，基于可行动作集设计均衡求解的协同演化算法，相关模型可用于互联网安全、车联网通信、交通系统效率、谣言传播控制、关键基础设施安全等问题的研究[115]。

5.4 未来研究展望

1)考虑适变型对手

现实中的对手一般具有学习能力，特别是未来高科技

战争背景下，智能体的学习能力、应变能力将会有很大突破。所以，更符合实际的情况是考虑适变型对手，对手越理性、智能，越能模拟现实中的对抗场景。虽然目前这方面的研究还很少，难度很大，但对其进行研究具有十分重要的实际意义。

2）考虑异构型资源分配

为了更加贴近实际对抗，应该考虑异构型资源的分配，即考虑多种不同类型的资源的分配，在这些不同类型的资源中，不同的组合会产生不同的效用，对于具体的战场，就不是简单地以数量多取胜，而是以不同的资源类型进行匹配产生不同的作战效用，基于此判定对抗双方的输赢。在这种情况下，学习者不仅仅需要考虑固定资源分配策略的组合优化问题，还需要考虑不同资源类型之间匹配的组合优化问题。虽然问题难度升级，但这更加贴近实际。

3）提高算法遗憾界性能

现有的算法大多数实现的是关于时间范围 T 的次线性遗憾界，还有提升的空间。未来努力的方向包含两个方面：一方面是致力于提升遗憾界，另一方面是尽量缩短达到遗憾界所需要的交互轮次。

4）提高算法的运行效率

针对资源分配问题，Exp3-U 以及 Exp3-G 算法相较于目前的其他算法在运行效率方面已经有较大的优势，但是离在线对抗的实际要求还存在差距，因此，致力于开发探索出高效的算法也是需要重点关注的问题之一。

结束语 对抗条件下的资源分配一直以来都是军事作战不可忽视的问题，是智能决策的核心问题。在具有环境高复杂、动态不确定、博弈强对抗等突出特征的未来竞争环境，如何实现对抗条件下的资源合理有效分配是一个值得重点关注的问题。本文围绕对抗条件下的资源分配问题，聚焦多类离线与在线布洛托博弈模型，借助博弈学习和在线组合优化理论，梳理多类博弈问题求解方法。本文为对抗条件下资源分配问题的研究提供了一种博弈论视角，期望能为资源分配与博弈论交叉领域的相关研究带来启发，激发相关研究人员的兴趣。

参 考 文 献

[1] ZHAO J, YANG C. Graph Reinforcement Learning for Predictive Power Allocation to Mobile Users[EB/OL]. (2022-03-08)(2022-11-01). https://arxiv.org/abs/2203.03906.

[2] PANIGRAHY N K, BASU P, NAIN P, et al. Resource Allocation in One-Dimensional Distributed Service Networks with Applications [J]. Performance Evaluation. 2020, 142: 102110.

[3] ABDALLAH M. Effects of Behavioral Decision-Making in Game-Theoretic Frameworks for Security Resource Allocation in Networked Systems [D]. West Lafayette: Purdue University Graduate School, 2022.

[4] JI X, ZHANG W, XIANG F, et al. A Swarm Confrontation Method Based on Lanchester Law and Nash Equilibrium [J]. Electronics, 2022, 11(6): 896.

[5] ZHANG X X, GE B F, TAN Y J. Multi-Attribute Game Theoretic Model for Resource Allocation in Military Attack Defense Application[J]. Journal of National University of Defense Technology, 2018, 40(5): 153-160.

[6] LIU B Y, YE X B, ZHOU C F, et al. Composite Mode On-Orbit Service Resource Allocation Based on Improved DQN[J]. Acta Aeronautica et Astronautica Sinica, 2020, 41(5): 9.

[7] YAN W, JINKUAN W, JINGHAO S. A Game-Theoretic Based Resource Allocation Strategy for Cloud Computing Services[J]. Scientific Programming, 2016, 2016(Pt. 2): 1629893. 1.

[8] MYERSON R B. Incentives to Cultivate Favored Minorities Under Alternative Electoral Systems [J]. American Political Science Review, 1993, 87(4): 856-869.

[9] JIANG Y J, KUANG K, WU F. Big Data Intelligence: From the Optimal Solution of Data Fitting to The Equilibrium Solution of Game Theory[J]. CAAI Transactions on Intelligent Systems, 2020, 15(1): 175-182.

[10] BOREL E. La Théorie Du Jeu Et Les Equations Intégralesa Noyau Symétrique [J]. Comptes rendus de l'Académie des Sciences, 1921, 173(1304/1305/1306/1307/1308): 58.

[11] NAKAYAMA M. The Dawn of Modern Theory of Games[J]. Advances in Mathematical Economics, 2006, 9: 73-97.

[12] GROSS O, WAGNER R. A Continuous Colonel Blotto Game [R]. Rand Project Air Force Santa Monica Ca, 1950.

[13] KOVENOCK D, ROBERSON B. Coalitional Colonel Blotto Games with Application to The Economics of Alliances [J]. Journal of Public Economic Theory, 2012, 14(4): 653-676.

[14] GRANA J, LAMB J, O'DONOUGHUE N A. Findings on Mosaic Warfare from a Colonel Blotto Game [R]. Rand National Defense Research Inst Santa Monica Ca, 2021.

[15] MORGENSTERN O, VON NEUMANN J. Theory of Games and Economic Behavior [M]. Princeton University Press, 1953.

[16] NASH J. Non-Cooperative Games [D]. Annals of mathematics, 1951: 286-295.

[17] KALAI A, VEMPALA S. Efficient Algorithms for Online Decision Problems [J]. Journal of Computer and System Sciences, 2005, 71(3): 291-307.

[18] HANNAN J. Approximation to Bayes Risk in Repeated Play [J]. Contributions to the Theory of Games, 1957, 3(2): 97-139.

[19] ROBBINS H. Some Aspects of The Sequential Design of Experiments [J]. Bulletin of the American Mathematical Society, 1952, 58(5): 527-535.

[20] GYÖRGY A, LINDER T, LUGOSI G, et al. The On-Line Shortest Path Problem Under Partial Monitoring [J]. Journal of Machine Learning Research, 2007, 8(10): 2369-2403.

[21] KOCÁK T, NEU G, VALKO M, et al. Efficient Learning by Implicit Exploration in Bandit Problems with Side Observations [C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. 2014: 613-621.

[22] BARTLETT P, DANI V, HAYES T, et al. High-Probability Regret Bounds for Bandit Online Linear Optimization [C]// Proceedings of the 21st Annual Conference on Learning Theory-COLT. 2008: 335-342.

[23] LAI T L, ROBBINS H. Asymptotically Efficient Adaptive Allocation Rules [J]. Advances in Applied Mathematics, 1985, 6(1): 4-22.

[24] AUER P, CESA-BIANCHI N, FISCHER P. Finite-Time Analysis of The Multiarmed Bandit Problem [J]. Machine Learning, 2002, 47(2): 235-256.

[25] BUBECK S, CESA-BIANCHI N. Regret Analysis of Stochastic and Non-Stochastic Multi-Armed Bandit Problems [J]. Founda-

tions and Trends © in Machine Learning, 2012, 5(1): 1-122.
[26] GARIVIER A, CAPPÉ O. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond [C]// Proceedings of the 24th Annual Conference on Learning Theory. 2011: 359-376.
[27] AUER P, CESA-BIANCHI N, FREUND Y, et al. The Non-Stochastic Multiarmed Bandit Problem [J]. SIAM Journal on Computing, 2002, 32(1): 48-77.
[28] FREUND Y, SCHAPIRE R E. A Decision-Theoretic Generalization of On-Line Learning and An Application to Boosting [J]. Journal of Computer and System Sciences, 1997, 55(1): 119-139.
[29] BARD N D C. Online Agent Modelling in Human-Scale Problems [D]. Edmonton: University of Alberta, 2016.
[30] CESA-BIANCHI N, LUGOSI G. Prediction, Learning, and Games [M]. Cambridge: Cambridge university press, 2006.
[31] ARORA R, DEKEL O, TEWARI A. Online Bandit Learning Against an Adaptive Adversary: From Regret to Policy Regret. [EB/OL]. (2012-06-27) [2022-12-01]. https://arxiv.org/abs/1206.6400.
[32] LITTLESTONE N, WARMUTH M K. The Weighted Majority Algorithm [J]. Information and Computation, 1994, 108(2): 212-261.
[33] WARMUTH M K, JAGOTA A K. Continuous and Discrete-Time Nonlinear Gradient Descent: Relative Loss Bounds and Convergence [C]// Electronic Proceedings of the 5th International Symposium on Artificial Intelligence and Mathematics. 1997.
[34] GORDON G J. Regret Bounds for Prediction Problems [C]// Proceedings of the Twelfth Annual Conference on Computational Learning Theory. 1999: 29-40.
[35] MORRILL D. Hindsight Rational Learning for Sequential Decision-Making: Foundations and Experimental Applications [D]. Alberta: University of Alberta, 2022.
[36] NEU G. Explore No More: Improved High-Probability Regret Bounds for Non-Stochastic Bandits [C]// Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2. 2015: 3168-3176.
[37] ABERNETHY J, RAKHLIN A. Beating the Adaptive Bandit with High Probability [C]// Proceedings of the Information Theory and Applications Workshop. 2009: 280-289.
[38] ABERNETHY J, HAZAN E E, RAKHLIN A. Competing in the Dark: An Efficient Algorithm for Bandit Linear Optimization [C]// Proceedings of the 21st Annual Conference on Learning Theory(COLT 2008). 2008: 263-273.
[39] NEU G, BARTÓK G. An Efficient Algorithm for Learning with Semi-Bandit Feedback [C]// Proceedings of the International Conference on Algorithmic Learning Theory. 2013: 234-248.
[40] DANI V, KAKADE S M, HAYES T. The Price of Bandit Information for Online Optimization [C]// Proceedings of the 20th International Conference on Neural Information Processing Systems. 2007: 345-352.
[41] ADAM L, HORCIK R, KASL T, et al. Double Oracle Algorithm for Computing Equilibria in Continuous Games [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 5070-5077.
[42] GANZFRIED S. Algorithm for Computing Approximate Nash Equilibrium in Continuous Games with Application to Continuous Blotto [J]. Games, 2021, 12(2): 47.
[43] BEHNEZHAD S, BLUM A, DERAKHSHAN M, et al. From Battlefields to Elections: Winning Strategies of Blotto and Auditing Games [C]// Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms. 2018: 2291-2310.
[44] BEHNEZHAD S, DEHGHANI S, DERAKHSHAN M, et al. Faster and Simpler Algorithm for Optimal Strategies of Blotto Game [C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. 2017: 369-375.
[45] KOVENOCK D, ROBERSON B. Generalizations of the General Lotto and Colonel Blotto Games [J]. Economic Theory, 2021, 71(3): 997-1032.
[46] VU D Q, LOISEAU P, SILVA A. Approximate Equilibria in Non-constant-sum Colonel Blotto and Lottery Blotto Games with Large Numbers of Battlefields [EB/OL]. (2019-10-15) [2022-12-01]. https://arxiv.org/abs/1910.06559v1.
[47] ROBERSON B. The Colonel Blotto Game [J]. Economic Theory, 2006, 29(1): 1-24.
[48] SCHWARTZ G, LOISEAU P, SASTRY S S. The Heterogeneous Colonel Blotto Game [C]// 2014 7th International Conference on Network Games, Control and Optimization (NetGCoop). 2014: 232-238.
[49] VU D Q, LOISEAU P, SILVA A. Approximate Equilibria in Generalized Colonel Blotto and Generalized Lottery Blotto Games [EB/OL]. (2019-10-15) [2022-12-01]. https://arxiv.org/abs/1910.06559v2.
[50] LI X, ZHENG J. Pure strategy Nash equilibrium in 2-Contestant Generalized Lottery Colonel Blotto Games [J]. Journal of Mathematical Economics, 2022, 103: 102771.
[51] TULLOCK G. Efficient rent-seeking revisited[M]//The Political Economy of Rent-seeking. Boston, MA: Springer US, 1988: 91-94.
[52] VU D Q, LOISEAU P. Colonel Blotto Games with Favoritism: Competitions with Pre-Allocations and Asymmetric Effectiveness [C]// Proceedings of the 22nd ACM Conference on Economics and Computation. 2021: 862-863.
[53] ASPECT L, EWERHART C. Colonel Blotto Games with A Head Start [R]. Department of Economics-University of Zurich, 2022.
[54] GUPTA A, SCHWARTZ G, LANGBORT C, et al. A Three-Stage Colonel Blotto Game with Applications to Cyber Physical Security [C]// Proceedings of the American Control Conference. 2014: 3820-3825.
[55] PAARPORN K, CHANDAN R, KOVENOCK D, et al. Strategically Revealing Intentions in General Lotto Games [EB/OL]. (2021-10-23) [2022-12-01]. https://arxiv.org/abs/2110.12099.
[56] VU D Q, LOISEAU P, SILVA A, et al. Path Planning Problems with Side Observations—When Colonels Play Hide And-Seek [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 2252-2259.
[57] TAKIMOTO E, WARMUTH M K. Path Kernels and Multiplicative Updates [J]. The Journal of Machine Learning Research, 2003, 4: 773-818.

[58] CZARNECKI W M，GIDEL G，TRACEY B，et al. Real World Games Look Like Spinning Tops [C]// In Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020：17443-17454.

[59] OMIDSHAFIEI S，TUYLS K，CZARNECKI W M，et al. Navigating the Landscape of Multiplayer Games [J]. Nature Communications. 2020，11(1)：1-17.

[60] HORTALA-VALLVE R，LLORENTE-SAGUER A. Pure Strategy Nash Equilibria in Non-Zero Sum Colonel Blotto Games [J]. International Journal of Game Theory，2012，41(2)：331-343.

[61] AHMADINEJAD A，DEHGHANI S，HAJIAGHAYI M，et al. From Duels to Battlefields：Computing Equilibria of Blotto and Other Games [J]. Mathematics of Operations Research，2019，44(4)：1304-1325.

[62] BAYE M R，KOVENOCK D，DE VRIES C G. The Solution to The Tullock Rent-Seeking Game When R＞ 2：Mixed-Strategy Equilibria and Mean Dissipation Rates [J]. Public Choice，1994，81(3)：363-380.

[63] HILLMAN A L，RILEY J G. Politically Contestable Rents and Transfers [J]. Economics & Politics，1989，1(1)：17-39.

[64] MCMAHAN H B，GORDON G J，BLUM A. Planning in The Presence of Cost Functions Controlled by An Adversary [C]// Proceedings of the 20th International Conference on Machine Learning(ICML-03). 2003：536-543.

[65] LANCTOT M，ZAMBALDI V，GRUSLYS A，et al. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning [C]// Proceedings of the 31th International Conference in Neural Information Processing Systems. 2017：4193-4206.

[66] ZOU M，CHEN J，LUO J，et al. Equilibrium Approximating and Online Learning for Anti-Jamming Game of Satellite Communication Power Allocation [J]. Electronics，2022，11(21)：3526.

[67] GEMP I，SAVANI R，LANCTOT M，et al. Sample-based Approximation of Nash in Large Many-Player Games via Gradient Descent[C]// Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems. 2022：507-515.

[68] ANTHONY T，ECCLES T，TACCHETTI A，et al. Learning To Play No-Press Diplomacy with Best Response Policy Iteration [C]// Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020：17987-18003.

[69] JACOB A P，WU D J，FARINA G，et al. Modeling Strongand Human-Like Gameplay With KL-Regularized Search[C]// International Conference on Machine Learning. PMLR，2022：9695-9728.

[70] STRANG A，SEWELL D，KIM A，et al. Principal Trade-off Analysis [EB/OL]. (2022-06-09) [2022-12-01]. https://arxiv.org/abs/2206.07520.

[71] BERTRAND Q，CZARNECKI W M，GIDEL G. On the Limitations of Elo：Real-World Games，are Transitive，not Additive [EB/OL]. (2022-06-21) [2022-12-01]. https://arxiv.org/abs/2206.12301.

[72] CHEN W，WANG Y，YUAN Y，et al. Combinatorial Multi-Armed Bandit and Its Extension to Probabilistically Triggered Arms [J]. The Journal of Machine Learning Research，2016，17(1)：1746-1778.

[73] KOOLEN W M，WARMUTH M K，KIVINEN J，et al. Hedging Structured Concepts [C]// Proceedings of the COLT. 2010：93-105.

[74] ALON N，CESA-BIANCHI N，DEKEL O，et al. Online Learning with Feedback Graphs：Beyond Bandits [C]// Proceedings of the Conference on Learning Theory. 2015：23-35.

[75] CESA-BIANCHI N，LUGOSI G. Combinatorial Bandits [J]. Journal of Computer and System Sciences，2012，78(5)：1404-1422.

[76] VU D Q，LOISEAU P，SILVA A. Combinatorial Bandits for Sequential Learning in Colonel Blotto Games [C]// Proceedings of the IEEE 58th Conference on Decision and Control (CDC). 2019：867-872.

[77] BADANIDIYURU A，KLEINBERG R，SLIVKINS A. Bandits with Knapsacks [J]. Journal of the ACM(JACM)，2018，65(3)：1-55.

[78] IMMORLICA N，SANKARARAMAN K A，SCHAPIRE R，et al. Adversarial Bandits with Knapsacks [C]// Proceedings of the IEEE 60th Annual Symposium on Foundations of Computer Science(FOCS). 2019：202-219.

[79] AGRAWAL S，DEVANUR N. Linear Contextual Bandits with Knapsacks [C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016：3458-3467.

[80] LIU S，JIANG J，LI X. Non-stationary Bandits with Knapsacks [EB/OL]. (2022-05-25) [2022-12-01]. https://arxiv.org/abs/2205.12427.

[81] CASTIGLIONI M，CELLI A，KROER C. Online Learning with Knapsacks：The Best of Both Worlds [EB/OL]. (2022-02-28). [2022-12-01] https://arxiv.org/abs/2202.13710.

[82] GAITONDE J，LI Y，LIGHT B，et al. Budget Pacing in Repeated Auctions：Regret and Efficiency without Convergence [EB/OL]. (2022-05-18) [2022-12-01]. https://arxiv.org/abs/2205.08674.

[83] ASHWINKUMAR B，ROBERT K，ALEKSANDRS S. Bandits with Knapsacks [C]// Proceedings of the 54th IEEE Symposium on Foundations of Computer Science. ACM，2013.

[84] AGRAWAL S，DEVANUR N R. Bandits with Concave Rewards and Convex Knapsacks [C]// Proceedings of the Fifteenth ACM Conference on Economics and Computation. 2014：989-1006.

[85] AGRAWAL S，DEVANUR N R，LI L. An Efficient Algorithm for Contextual Bandits with Knapsacks，and An Extension to Concave Objectives [C] // Proceedings of the Conference on Learning Theory. 2016：4-18.

[86] BADANIDIYURU A，LANGFORD J，SLIVKINS A. Resourceful Contextual Bandits [C]// Proceedings of the Conference on Learning Theory. 2014：1109-1134.

[87] AGARWAL A，HSU D，KALE S，et al. Taming the Monster：A Fast and Simple Algorithm for Contextual Bandits [C]// Proceedings of the International Conference on Machine Learning. 2014：1638-1646.

[88] LI X，SUN C，YE Y. The Symmetry between Arms and Knapsacks：A Primal-Dual Approach for Bandits with Knapsacks [C]// Proceedings of the International Conference on Machine Learning. 2021：6483-6492.

[89] LEON V，ETESAMI S R. Bandit Learning for Dynamic Colonel Blotto Game with A Budget Constraint [C]// Proceedings of the

60th IEEE Conference on Decision and Control (CDC). 2021: 3818-3823.
[90] SANKARARAMAN K A, SLIVKINS A. Combinatorial Semi-Bandits with Knapsacks [C]// Proceedings of the International Conference on Artificial Intelligence and Statistics. 2018: 1760-1770.
[91] WONG R T. Combinatorial Optimization: Algorithms and Complexity [J]. SIAM Review, 1983, 25(3): 424.
[92] CLEMPNER J B. Reveling Misleading Information for Defenders and Attackers in Repeated Stackelberg Security Games [J]. Engineering Applications of Artificial Intelligence, 2022, 110: 104703.
[93] CHANDAN R, PAARPORN K, MARDEN J R. When Showing Your Hand Pays Off: Announcing Strategic Intentions in Colonel Blotto Games [C]// Proceedings of the American Control Conference (ACC). 2020: 4632-4637.
[94] SEDDIGHIN S. Campaigning via LP's: Solving Blotto and Beyond [D]. Baltimore: University of Maryland, 2019.
[95] FERGUSON B L, SHISHIKA D, MARDEN J R. Ensuring the Defense of Paths and Perimeters in Dynamic Defender-Attacker Blotto Games (dDAB) on Graphs [C]// Proceedings of the 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton). 2022: 1-7.
[96] CHEN A K, FERGUSON B L, SHISHIKA D, et al. Path Defense in Dynamic Defender-Attacker Blotto Games (dDAB) with Limited Information [EB/OL]. (2022-04-08) [2022-12-01]. https://arxiv.org/abs/2204.04176.
[97] PAARPORN K, CHANDAN R, ALIZADEH M, et al. A General Lotto Game with Asymmetric Budget Uncertainty [EB/OL]. (2021-06-23) [2022-12-01]. https://arxiv.org/abs/2106.12133.
[98] SHAH V, MARDEN J R. Battlefield transfers in coalitional Blotto games [EB/OL]. (2023-04-04) [2023-06-30]. https://arxiv.org/abs/2304.02068.
[99] EWERHART C, KOVENOCK D. A Class of N-Player Colonel Blotto Games with Multidimensional Private Information [J]. Operations Research Letters, 2021, 49(3): 418-425.
[100] ANBARCI N, CINGIZ K, ISMAIL M S. Proportional Resource Allocation in Dynamic N-Player Blotto Games [J]. Mathematical Social Sciences, 2023, 125: 94-100.
[101] GUAN S, WANG J, JIANG C, et al. Colonel Blotto Game Aided Attack-Defense Analysis in Real-World Networks [C]// 2018 IEEE Global Communications Conference (GLOBECOM). IEEE, 2018: 1-6.
[102] BOIX-ADSERÀ E, EDELMAN B L, JAYANTI S. The Multi-player Colonel Blotto Game [C]// Proceedings of the 21st ACM Conference on Economics and Computation. 2020: 47-48.
[103] DONAHUE K, KLEINBERG J. Private Blotto: Viewpoint Competition with Polarized Agents [EB/OL]. (2023-02-27) [2023-06-30]. https://arxiv.org/abs/2302.14123.
[104] PERCHET V, RIGOLLET P, LE GOUIC T. An Algorithmic Solution to The Blotto Game Using Multi-Marginal Couplings [C]// Proceedings of the 23rd ACM Conference on Economics and Computation. 2022: 208-209.
[105] BEAGLEHOLE D, HOPKINS M, KANE D, et al. Sampling Equilibria: Fast No-Regret Learning in Structured Games [EB/OL]. (2022-01-26) [2022-12-01]. https://arxiv.org/abs/2201.10758.
[106] NOEL J C G. Reinforcement Learning Agents in Colonel Blotto [EB/OL]. (2022-04-04) [2022-12-01]. https://arxiv.org/abs/2204.02785.
[107] THOMAS C. N-Dimensional Blotto Game with Heterogeneous Battlefield Values [J]. Economic Theory, 2018, 65(3): 509-544.
[108] MASUCCI A M, SILVA A. DefensiveResource Allocation in Social Networks [C]// 2015 54th IEEE Conference on Decision and Control (CDC). IEEE, 2015: 2927-2932.
[109] ZOU M, CHEN J, LUO J, et al. Equilibrium Approximating and Online Learning for Anti-Jamming Game of Satellite Communication Power Allocation [J]. Electronics, 2022, 11(21): 3526.
[110] ZHANG L, WANG Y, HAN Z. Safeguarding UAV-Enabled Wireless Power Transfer Against Aerial Eavesdropper: A Colonel Blotto Game [J]. IEEE Wireless Communications Letters, 2021, 11: 503-507.
[111] WEI P, WANG S D, LU R M, et al. Multi-Channel Power Distribution for Anti-Jamming Based on Asymmetric Colonel Blotto Game [J]. Journal of National University of Defense Technology, 2023, 45: 35-48.
[112] HAJIMIRSADEGHI M, SRIDHARAN G, SAAD W, et al. Inter-Network Dynamic Spectrum Allocation Via a Colonel Blotto Game [C]// Proceedings of the Annual Conference on Information Science and Systems (CISS). 2016: 252-257.
[113] MIN M, XIAO L, XIE C, et al. Defense Against Advanced Persistent Threats: A Colonel Blotto Game Approach [C]// Proceedings of the IEEE International Conference on Communications (ICC). 2017: 1-6.
[114] GUAN S, WANG J, YAO H, et al. Colonel blotto games in network systems: Models, strategies, and applications [J]. IEEE Transactions on Network Science and Engineering, 2019, 7(2): 637-649.
[115] FERDOWSI A, SAAD W, MANDAYAM N B. Colonel Blotto Game for Sensor Protection in Interdependent Critical Infrastructure [J]. IEEE Internet of Things Journal, 2020, 8(4): 2857-2874.

LUO Junren, born in 1989, Ph.D candidate. His main research interests include imperfect information game and adversarial team game.

CHEN Jing, born in 1972, Ph.D, professor, Ph.D supervisor. His main research interests include cognitive decision-making gaming and distributed intelligence.

(责任编辑:柯颖)