

面向对抗条件下资源分配的在线多阶段布洛托博弈求解方法

陈少飞, 邹明我, 苏小龙, 罗俊仁, 冯俊侨

(国防科技大学智能科学学院, 湖南 长沙 410073)

摘要: 未来战场上的作战资源分配是一个存在总资源预算约束的多阶段对抗问题, 具有环境高复杂性、动态不确定性、博弈强对抗性。基于布洛托博弈模型, 首先把多阶段对抗场景下的资源分配问题建模为双层在线布洛托博弈, 然后将原资源分配问题转化为有向无环图上的在线最短路径问题, 并借鉴拉格朗日博弈对资源分配问题进行分析求解。此外, 提出 LagrangeBwK-Exp3-G 算法以实现多阶段对抗条件下资源分配问题的高概率遗憾最小化, 进一步通过数学推导获得关于时间范围 T 的高概率遗憾界。最后, 设计一个多阶段对抗条件下的卫星通信多信道功率分配实验, 从而验证 LagrangeBwK-Exp3-G 算法具有良好性能。

关键词: 多阶段对抗; 布洛托博弈; 资源分配; 高概率遗憾

中图分类号: TP39

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202341

Online multi-stage Colonel Blotto game solving method for resource allocation under contested condition

CHEN Shaofei, ZOU Mingwo, SU Xiaolong, LUO Junren, FENG Junqiao

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

Abstract: The allocation of combat resources on the future battlefield is a multi-stage confrontation problem with total resource budget constraints, which is characterized by high complexity of environment, dynamic uncertainty, and strong game confrontation. Based on the Blotto game model, the research firstly modelled the resource allocation problem in the multi-stage confrontation scenario as a two-level online Blotto game, then transformed the original problem into an online shortest path problem on a directed acyclic graph to realize the intuitive formulation of the resource allocation problem. The resource allocation problem was analyzed and solved by referring to the Lagrange game. In addition, the LagrangeBwK-Exp3-G algorithm was proposed to minimize the high probability regret of the resource allocation problem under the condition of multi-stage antagonism, and the high-probability regret bound of the algorithm on the time range T was obtained by mathematical derivation. Finally, a multi-channel power allocation experiment of satellite communication under the condition of multi-stage confrontation was designed to verify the good performance of LagrangeBwK-Exp3-G algorithm.

Key words: multi-stage adversarial, Blotto game, resource allocation, high probability regret

0 引言

美军近年来提出的“分布式作战”“马赛克战”等新型作战概念都强调“将传统的大型武器平台分解成分散、独立的作战单元, 再根据任务和环境将

作战单元重组”的作战能力^[1], 力图通过降低作战系统成本、提高作战系统可扩展性和灵活性来谋求战略主动和绝对优势。作战资源分配建立了行动方案 and 作战资源之间的连接关系, 为作战单元的分散、重组提供了重要支撑。在环境高复杂、动态不

确定、博弈强对抗的未来战场^[2],作战资源分配,特别是对抗条件下的作战资源分配,是实现新型作战概念下良好作战效能的关键因素。兰德公司2021年发布《布洛托上校博弈对马赛克战的启示》^[3],报告借助布洛托上校博弈(以下简称布洛托博弈)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单一集成的作战模式更具优势的问题,展示了在未来作战模式中马赛克战在作战资源分配上的优越性与局限性。

布洛托博弈是一种两人参与的竞争性资源分配零和博弈,主要用来建模对抗条件下的有限资源分配问题,其概念最早由Borel于1921年提出^[4]。该博弈模型综合考虑了参与者之间的对抗性和参与者与环境之间的互动性,从而最大化己方的收益。虽然布洛托博弈的基本规则十分简单,但该模型本身涉及的条件假设可以十分复杂,导致布洛托博弈问题的求解变得十分困难^[5]。近些年,由于机器学习等计算技术的快速发展和应用需求(如通信信道功率分配^[6]、云计算资源分配^[7]、边缘计算功率分配^[8])的增加,布洛托博弈问题的许多模型变体和求解算法相继被提出,包括攻防布洛托博弈^[9]、有限信息下的动态攻防布洛托博弈^[10]、随机布洛陀博弈^[11]、动态多人布洛陀博弈^[12]等。然而,这些研究大多假设博弈模型已知,然后寻求最优或者均衡求解策略。

作战资源分配经常存在多个阶段对抗的情况。博弈双方可以根据已经发生的对抗历史进行后续行动的在线决策,但需要做好利用已知信息和探索新信息之间的权衡。这种在线布洛托博弈问题的主要性能衡量指标是遗憾,即学习者分别采取实际动作和最佳动作产生的累积损失之差。为了降低决策的遗憾,近些年提出了Exp3-IX^[13]、Hedge^[14]、几何Hedge^[15]等求解算法,在理论上降低了遗憾上界。然而,这些模型和算法并不适用于描述和求解存在总预算约束的多阶段作战资源分配的博弈问题。

基于此,本文以作战资源分配为背景,围绕多阶段对抗条件下的资源分配问题展开研究。由于现实对抗性活动中双方的资源存在总预算约束,将该问题建模为一种新的双层级在线布洛托博弈资源分配问题。在上层级,学习者面临在时间范围 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 ;在下层级,学习者在每个时间阶段 t 和对手进行一次对抗条件下的布洛托博弈。此外,

本文提出LagrangeBwK-Exp3-G算法来求解问题,在后文中通过数学证明获得了关于时间范围 T 的 $O\left(\sqrt{nT|E|\ln\frac{T\cdot|E|}{\delta}}+2\sqrt{nT|P|\ln|P|}\right)$ 高概率遗憾界,并分析得到关于时间范围 T 的 $O(|E|T)$ 高效运行效率。

本文的结构如下:第1节介绍研究现状;第2节对多阶段对抗条件下的资源分配问题进行描述和建模;第3节介绍LagrangeBwK-Exp3-G算法并用数学推导出问题的高概率遗憾界;第4节通过卫星多信道功率分配实验验证算法的有效性;第5节进行论文总结。

1 对抗条件下的资源分配概述

1.1 对抗条件下的资源分配问题

资源分配是决策科学领域的核心问题。在资源分配问题中,根据待分配资源与目标对象之间的类型和数量,并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。博弈论作为决策科学的一个重要分支,为资源分配问题提供了建模工具与求解方法。电力分配^[15]、网络资源分配^[16]、安全设备布设^[17]、军事兵力分配^[18]、军事资源分配^[19]、在轨资源分析^[20]等领域中的很多问题可以被建模成资源分配问题。近年来,随着人工智能技术的发展,基于“大数据+算力+算法”的数据驱动范式为博弈问题的求解提供了通用解决方案。其中,基于深度学习和强化学习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,在线博弈对抗与离线博弈对抗的批处理式策略学习方法不同,在线博弈对抗策略的求解本质是一个流式学习过程,需要根据此前的少量交互样本来做决策。对抗条件下资源分配问题可以建模成非合作博弈模型,基于神经网络表示的迭代式学习方法和基于在线组合优化的无悔学习方法是对抗条件下资源分配在线策略求解的首选方案。

1.2 在线布洛托博弈

智能博弈技术在游戏领域接连取得重要突破^[21],并在军事指挥控制系统领域展现出广泛的应用前景^[22]。从博弈论的角度出发,离线布洛托博弈可以看作是一次性完全信息对抗条件下的资源分配问题,关键在于寻找离线布洛托博弈条件下的纳什均衡。而双方进行多次对抗,是一种在线布洛

托博弈问题,在该问题中学习者通常需要连续动态学习博弈,不断根据历史经验知识动态调整资源分配策略,需要做好利用已知信息和探索新信息之间的权衡。经分析,在线布洛托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,其中包括著名的多臂赌博机^[23]和在线最短路径(online shortest path, OSP)问题^[24]。在多臂赌博机问题中,假设赌博机有 k 个摇臂,学习者投一个游戏币后可以按下任意一个摇臂,每个摇臂以一定的概率吐出硬币作为回报,且每个摇臂的中奖概率不同,学习者的目标是通过寻找良好的策略以获得最大化的累积回报。多臂赌博机问题本质上是一个在线资源分配问题,即学习者学习如何分配有限的游戏币给不同的摇臂以获得最大化的累积回报。因此,在线布洛托博弈问题的求解可以从多臂赌博机问题的求解中获得启发。

对于随机型赌博机问题,Lai等人^[25]首次提出上置信界(upper confidence bounds, UCB)算法,Auer等人^[26]在Lai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。对于对抗性赌博机问题,Auer等人^[27]提出Exp3算法并针对遗忘型对手的赌博机问题实现 $O(\sqrt{DT \ln D})$ 的期望遗憾上界,其中 D 表示赌博臂数量。

在线布洛托博弈问题的主要性能衡量指标是遗憾,针对上述对抗性赌博机问题,Neu^[13]针对Exp3算法提出Exp3-IX算法,以高概率 $1-\delta$ 实现 $R_T \leq O(\sqrt{(\log D + \log(1/\delta))TD})$ 的遗憾上界。针对完全信息反馈的在线资源动态分配问题,Freund^[14]提出了Hedge算法实现了 $O(\sqrt{T \log |S|})$ 次线性遗憾界,其中 $S \subset \{0,1\}^D$ 是一种使用二进制方式来表示动作的方法,针对赌博机反馈,Dani等人^[15]提出几何Hedge算法,获得 $O(D^{3/2} \sqrt{T})$ 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。

在线布洛托博弈问题中,由于学习者的策略数量是博弈参数(战场数量和资源预算)的指数函数,经典的Exp3算法的运行时间是博弈参数的指数级大小,而在线学习问题通常对算法有实时性要求,而借助OSP问题良好的图结构,可以将在线布洛托博弈问题转化为在线最短路径问题,提高算

法运行效率。

2 在线多阶段布洛托博弈建模

本文以作战资源分配为背景,围绕基于布洛托博弈的多阶段对抗条件下资源分配问题展开研究,首先,本文考虑了对手因素,所获得的求解方法更适用于实际对抗中对手不确定、博弈强对抗的场景,支撑在对抗条件下进行作战资源分配的有效实施。其次,本文选择的在线布洛托博弈问题考虑到对抗双方之间进行多次对抗,更符合实际中的情况。最后,本文考虑的多阶段对抗条件下资源分配问题,需要综合考虑双层级的资源分配问题,是重复博弈条件下资源分配问题的进一步延伸,增加了问题求解的难度,也更符合实际的对抗场景。在多阶段对抗条件下的资源分配问题中,下层级的每一次对抗仍然是基于一次性对抗条件下布洛托博弈模型进行构建的。

2.1 一次性对抗条件下布洛托博弈模型

考虑两人博弈问题,对抗双方简称为 r 和 m ,有 T 轮博弈,其中每一轮 $t \in T$,对抗双方进行一次静态布洛托博弈,假设 t 时刻共有 n 个战场(假设战场权重相同), r 方资源预算为 S_r , $S_r \in \mathbf{R}$, r 方分配至战场 k 的资源数量为 R_k ,需满足约束 $\sum_{k=1}^n R_k \leq S_r, k=1,2,3,\dots,n$,资源分配的向量表示为 $\tilde{\mathbf{R}}=[R_k], 1 \leq k \leq n$;同理, m 方预算为 S_m , $S_m \in \mathbf{R}$, m 方分配至战场 k 的资源数量为 M_k ,需满足约束 $\sum_{k=1}^n M_k \leq S_m, k=1,2,3,\dots,n$,资源分配的向量表示为 $\tilde{\mathbf{M}}=[M_k], 1 \leq k \leq n$ 。对于 r 方而言,将单个战场输赢规则定义为效用函数 l 。

$$l(R_k - M_k) = \begin{cases} 1, R_k - M_k > 0 \\ 0.5, R_k - M_k = 0 \\ -1, R_k - M_k < 0 \end{cases} \quad (1)$$

即对该战场分配资源多的一方获胜。假设该问题为零和博弈,则双方效用函数的关系表示为:

$$\begin{aligned} U_r(\tilde{\mathbf{R}}, \tilde{\mathbf{M}}) &= -U_m(\tilde{\mathbf{R}}, \tilde{\mathbf{M}}) \\ &= \sum_{k=1}^n W_k l(R_k - M_k) \end{aligned} \quad (2)$$

其中 W_k 表示战场 k 的价值。

在线线性优化问题中,学习者和对手之间进行 T 轮次博弈($T \in \mathbf{N}^*$),学习者的动作集表示为 $\mathbf{S} \subset \mathbf{R}^D, D \in \mathbf{N}^*$ 表示动作维度。在每一个时间阶段 $t \in [T]$,学习者在不知道对手动作的情况下选择

动作向量 $\tilde{p}_t \in S$ ，由对手产生一个损失向量 $l_t \in [0, 1]^D$ ，即学习者在时间阶段 t 对手对抗后的损失。该损失对学习者的公开程度决定了学习者能够获取的最大信息反馈量。在线线性优化问题中学习者的信息反馈一般包括以下3种情形。

(1) 完全信息反馈：在阶段 t 结束时，学习者能够观察到所有动作的损失向量 l_t 。

(2) 半信息反馈：在阶段 t 结束时，学习者能够观察到所做动作的损失向量 l_t 。

(3) 赌博机反馈：在阶段 t 结束时，学习者只能观察到损失标量 $L(\tilde{p}_t) = (l_t)^T \tilde{p}_t$ ，即所做动作的总损失。

需要注意的是，在半信息反馈的情形中，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，带侧面观测的半信息反馈是否存在由实际问题决定。但是不可否认，如果能推理获得未知动作的损失，那么针对不同的动作能够获得更准确的权重估计，从而对下一阶段动作的采样产生重要的影响。这个思想和不完全信息博弈中的信息补全思想是相同的，在不完全信息博弈领域，通过信息补全手段获得尽可能多的对手信息，是使己方获得有利决策优势的最直接途径。多阶段对抗条件下的资源分配问题可以获得带侧面观测的半信息反馈。

由于在线线性优化问题中的算法衡量指标是遗憾最小化，因此定义问题的优化目标为最小化如下遗憾：

$$R_T = \sum_{t=1}^T L(\tilde{p}_t) - \min_{p \in S} \sum_{t=1}^T L(p) \quad (3)$$

其中 p 为阶段 t 的最佳策略。该定义表示的遗憾是学习者在时间范围 T 内采取实际动作 $\tilde{p}_t$ 和选择最佳动作 p 产生的累积损失之差，采用该定义计算出来的遗憾界称为高概率遗憾界。

2.2 在线最短路径问题

多阶段对抗条件下资源分配问题中，对抗双方的策略集为有限离散向量，因此可以借助在线最短路径问题良好的图结构对问题进行求解。策略集可以用一个有向无环图 (directed acyclic graph, DAG) 定义，DAG有以下属性：有两个特殊的顶点，源点 s 和终止点 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 $|V| \geq 2$ 以及 $|E| \geq 1$ ，且每条边 $e \in E$ 至少包含一条路径 $p \in P$ 。用 n 表示路径集 P 中最长的路径长度，即

$\|p\|_1 \leq n, \forall p \in P$ 。给出时间范围 $T \in \mathbf{N}$ ，OSP问题定义如下：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 $l_t(e) \in [0, 1]$ ，学习者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的总损失 $L_t(p_t) = \sum_{e \in p_t} l_t(e)$ 。在阶段结束时，学习者获得带侧面观测的半信息反馈并对下一阶段的分配策略进行指导。学习者的目标是最小化 T 个阶段对抗的累积遗憾值。

图1展示了在 $S_r = S_m = 3, n = 3$ 情况下的 DAG。从初始点 s 到终止点 d 共有10条路径，代表10种不同的资源分配策略，其中每条边代表分配至当前阶段的资源数量。例如，图中红色的路径代表第1阶段分配2个资源，第2阶段分配0个资源，第3阶段分配1个资源，即分配策略为 $(2, 0, 1)$ 。

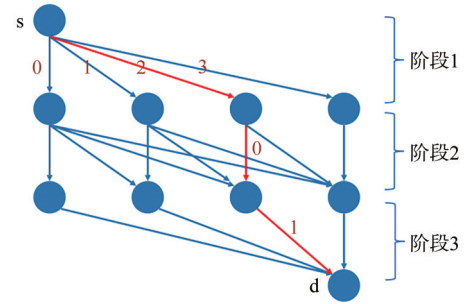


图1 在 $S_r = S_m = 3, n = 3$ 情况下的 DAG，红色路径表示的分配策略为 $(2, 0, 1)$

借助 DAG 的图结构，接下来阐述布洛托博弈模型在半信息反馈设置中能够获取额外信息的原理。在图1中，假设红色路径代表 r 方在时间阶段 t 实际选择的分配策略。以阶段1举例，如果 r 方在阶段1失败了，则边2能够揭示边0和边1的损失（边2称为揭示边，边0和边1称为被揭示边），因为边0和边1分配的资源比实际选择的少，且边0和边1在阶段1的损失不小于边2对应的损失。如果 r 方在阶段1赢了，其在阶段1的损失为0，则边2能够揭示边3的损失，即边3在阶段1的损失为0。通过这种侧面观察所得到的信息（简称为带侧面观测的半信息反馈），可以更加准确地估计部分边的损失，进一步合理地更新对应边的权重，对下一次博弈对抗产生正确的指导作用。

2.3 拉格朗日博弈问题

多阶段对抗条件下资源分配问题的求解可以从“受背包容量约束的赌博机” (bandits with knap-

sacks, BwK) 问题的求解中获得启发。Immorlica 等人^[28]提出一个受背包容量约束的随机多臂赌博机 (stochastic bandits with knapsacks, SBwK) 问题, 其中 S 表示有限赌博臂的集合, $X = (X(a))_{a \in S}$ 是关于赌博臂的概率分布。设 R 是一个有限的资源集, B_0 表示资源的总预算, 如果学习者选择拉动赌博臂 $a \in S$, 将消耗 $\omega_i(a)$ 大小的资源, 同时返回赌博臂的期望奖励 $r(a)$, 其中 $i \in R$, 且假设不同种类的资源 i 的预算都是相等的。该 SBwK 问题由以下线性规划的形式给出:

$$\max \sum_{a \in S} X(a)r(a) \quad \text{s.t.} \begin{cases} \sum_{a \in S} X(a) = 1 \\ \sum_{a \in S} X(a)\omega_i(a) \leq \frac{B_0}{T}, \forall i \in R \\ 0 \leq X(a) \leq 1, \forall a \in S \end{cases} \quad (4)$$

利用对偶变量 λ 得到上述线性规划的拉格朗日函数, 如式 (5) 所示:

$$\Gamma(X, \lambda) = \sum_{a \in S} X(a)r(a) + \sum_{i \in R} \lambda_i \left(1 - \frac{T}{B_0} \sum_{a \in S} X(a)\omega_i(a) \right) \quad (5)$$

Immorlica 等人^[28]根据上述拉格朗日函数定义了以下拉格朗日博弈: 存在原始玩家 1 和对偶玩家 2, 玩家 1 在每轮游戏中负责选择一支赌博臂 a , 玩家 2 负责选择一种资源 i , 二者动作对的收益如下所示。

$$\Gamma(a, i) = r(a) + 1 - \frac{T}{B_0} \omega_i(a), \forall a \in S, i \in R \quad (6)$$

其中, $\Gamma(a, i)$ 代表玩家 1 的最大化奖励目标, 而对于玩家 2 而言, $\Gamma(a, i)$ 则表示消耗, 玩家 2 的目标是最小化它。Immorlica 等人^[28]随后设计了面向重复拉格朗日博弈的 LagrangeBwK 算法, 其中两个玩家分别采用算法 ALG1 和 ALG2。

基于上述工作, 本文将在第 3 节以 LagrangeBwK 算法作为主要框架求解多阶段对抗条件下的资源分配问题, 其中 ALG1 的目标是最大化 $\Gamma(a, i)$, 用于下层级资源分配, 而 ALG2 的目标是最小化 $\Gamma(a, i)$, 用于上层级将总预算分配至各个时间阶段 t 。对于每一轮对抗, 对抗双方基于各自的阶段预算在 n 个战场上分配资源, 二者之间的效益通过 $\Gamma(a, i)$ 计算出来, 并反馈给 ALG1 和 ALG2。学习者需要权衡资源预算和消耗, 进而采取全局较优的动作。问题的关键在于如何将下层级算法的效用 (损失) 传递到上层级, 为下一阶段的预算分配

提供依据。从直观上来说, 衡量多阶段对抗条件下资源分配问题的算法性能的指标是遗憾, 下面给出有关高概率遗憾的重要引理。

引理 1 根据文献[28]中的定理 4.4, 对 $\delta \in (0, 1)$, 分别用 $R_{1, \delta/T}(T)$ 和 $R_{2, \delta/T}(T)$ 表示算法 ALG1 和 ALG2 的 $1 - \delta$ 高概率遗憾上界, 则算法 LagrangeBwK 的遗憾上界至少以 $1 - O(\delta T)$ 概率维持在 $O\left(\frac{T}{B_0}\right)\left(R_{1, \delta/T}(T) + R_{2, \delta/T}(T) + \sqrt{T \log(|R| \frac{T}{\delta})}\right)$ 。

3 在线多阶段布洛托博弈求解

3.1 在线多阶段布洛托博弈

本文研究的多阶段对抗条件下的分配问题是指对抗双方各自受己方总资源预算约束的双层级资源分配问题。在上层级, 学习者面临在时间范围 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 ; 在下层级, 学习者在每个时间阶段 t 和对手进行一次对抗条件下的布洛托博弈。与该问题相近的研究为 BwK 问题, BwK 最早由 Badanidiyuru 等人^[29]提出, Agrawal 等人^[30]研究了一种上下文 BwK 问题。Immorlica 等人^[28]研究了包括随机型、对抗型、全信息反馈、半信息反馈以及上下文赌博机在内的多种 BwK 问题, 并设计 LagrangeBwK 算法求解对抗型 BwK 问题, 获得高概率遗憾界。Li 等人^[31]采用原始-对偶 (primal-dual) 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。Leon 等人^[32]研究了带总预算约束的动态布洛托博弈问题, 提出 LagrangeBwK-Edge 在线学习算法, 并实现了高概率的遗憾界。以上研究对总资源预算约束的双层级资源分配问题进行求解, 并获得了特定条件下的高概率遗憾界, 由于布洛托博弈模型假设的特殊性, 多阶段对抗条件下的资源分配问题属于带侧面观测的半信息反馈问题。接下来, 设计 LagrangeBwK-Exp3-G 算法, 利用这一事实降低信息不确定性, 提高算法求解性能。

3.2 LagrangeBwK-Exp3-G 算法设计

LagrangeBwK-Exp3-G 沿用 LagrangeBwK (算法 1) 的双层求解框架, 分别针对下层级 ALG1 和上层级 ALG2 求解其高概率遗憾界。ALG2 的作用是根据每次对抗的反馈结果, 将总预算分配至各个

时间阶段 t 。在每个对抗阶段 t ，ALG1 基于 ALG2 的资源分配结果最大化对抗收益。接下来，分别设计 ALG1 算法和 ALG2 算法并给出对应的高概率遗憾界，整合形成 LagrangeBwK-Exp3-G 算法并分析算法运行效率。

算法1 LagrangeBwK

Input: 资源预算 B_0, T , ALG1 算法, ALG2 算法;

```

1: for  $t = 1, \dots, T$  do
2:   ALG1 算法选择一支赌博臂  $a_t \in S$ ;
3:   ALG2 算法选择其中一种资源  $i_t \in R$ ;
4:   观察奖励  $r_t(a_t)$  和消耗  $\omega_{t,i}(a_t), \forall i \in R$ ;
5:    $\Gamma(a_t, i_t)$  作为奖励反馈给 ALG1;
6:    $\Gamma(a_t, i_t)$  作为消耗反馈给 ALG2;
7: end for

```

3.2.1 ALG1 算法设计

为获得下层级 ALG1 算法的高概率遗憾界，本文基于 Exp3-OE 算法^[33]，设计基于 DAG 图结构的 Exp3-G 算法。

算法2 Exp3-G

Input: $\beta_2 \in (0, 1)$, T , DAG;

```

1: 初始化, 对于 DAG 中的每一条边  $e \in E, \omega_e^0 = 1, Q = \{0\}, u_0 = s, k = 0, d$  表示 DAG 的终止点;
2: for  $t = 1, 2, \dots, T$  do
3:   for  $k \leq n$  do
4:     以概率  $q_{e_{[u_k, u_{k+1}]}}^t = (1 - \gamma) \frac{\omega_{e_{[u_k, u_{k+1}]}}^t H^t(u_{k+1}, d)}{H^t(u_k, d)} + \gamma \frac{|\{p \in P: e \in E\}|}{|P|}$ 
     从  $u_k$  的继承点集合  $C(u_k)$  中采样一个继承点  $u_{k+1}$ ;
5:     增加顶点  $u_{k+1}$  到集合  $Q$  中;
6:   end for
7:   连接集合  $Q$  中所有的顶点得到采样路径  $p^t \in P$ , 即分配策略;
8:   和对手对抗产生增益  $g_e^t$ .

```

```

9:   计算增益估计  $\tilde{g}_e^t = \begin{cases} \frac{g_e^t + \beta_2}{q_e^t}, & \text{if } e \in p^t \\ \frac{\beta_2}{q_e^t}, & \text{otherwise} \end{cases}$ ;

```

```

10: 更新 DAG 中各个边的权重:  $\omega_e^{t+1} = \omega_e^t e^{\eta \tilde{g}_e^t}$ ;
11: end for

```

为衡量 Exp3-G 算法的性能，下面给出算法的高概率遗憾上界，所求取的遗憾上界是关于时间范围 T 的次线性函数。

定理1 对于任意 $\delta \in (0, 1)$, $0 < \beta_2 \leq 1$, $0 \leq \gamma < 0.5$, 且 $\eta > 0$, $2\eta n|P| \leq \gamma$, 则 Exp3-G 算法以高概率 $1 - \delta$ 实现式 (7) 所示的遗憾上界:

$$\hat{L}^T - \min_{p \in P} L^T(p) \leq T\gamma + 2\eta n|P| + \frac{n}{\beta_2} \ln \frac{E}{\delta} + \frac{\ln |P|}{\eta} + TE\beta_2 \quad (7)$$

若给定参数 $\beta_2 = \sqrt{\frac{n}{T|E|} \ln \frac{|E|}{\delta}}$, $\gamma = 2\eta n|P|$ 且 $\eta = \sqrt{\frac{\ln |P|}{4Tn|P|}}$, 则当 $T \geq \max \left\{ \frac{n}{|E|} \ln \frac{|E|}{\delta}, 4|P| \ln |P| \right\}$,

Exp3-G 算法以高概率 $1 - \delta$ 实现的遗憾上界为:

$$R_\delta(T) = \hat{L}^T - \min_{p \in P} L^T(p) \leq 2 \left(\sqrt{nT|E| \ln \frac{|E|}{\delta}} + 2\sqrt{nTP \ln |P|} \right) \quad (8)$$

证明: 设对于每条路径 $p \in P$, 有 $G_p^T = \sum_{t=1}^T g_p^t$ 且 $\tilde{G}_p^T = \sum_{t=1}^T \tilde{g}_p^t$. 对于边 $e \in E$, 有 $G_e^T = \sum_{t=1}^T g_e^t$, $\tilde{G}_e^T = \sum_{t=1}^T \tilde{g}_e^t$, $\hat{G}^T = \sum_{t=1}^T g_p^t$, 其中 p^t 为 t 时刻实际选择的路径 (由实际选择的边组成)。

算法2中的步骤4为路径采样方法, 从 u_k 的继承点集合 $C(u_k)$ 中采样一个继承点 u_{k+1} 的概率为:

$$q_{e_{[u_k, u_{k+1}]}}^t = (1 - \gamma) \frac{\omega_{e_{[u_k, u_{k+1}]}}^t H^t(u_{k+1}, d)}{H^t(u_k, d)} + \gamma \frac{|\{p \in P: e \in E\}|}{|P|} \quad (9)$$

其中 $H^t(u, v) = \sum_{p \in P} \prod_{e \in p} \omega_e^t(e)$ 表示经过顶点 (u, v) 的所有路径 (不一定是 DAG 中的完整路径) 的权重之和, 根据该路径采样方法, 不局限于从源 s 或 u_k 开始选择边, 在整个 DAG 中对任意边进行采样的概率为:

$$q_{e_{[u, v]}}^t = (1 - \gamma) \frac{H^t(s, u) \omega_{e_{[u, v]}}^t H^t(v, d)}{H^t(s, d)} + \gamma \frac{|\{p \in P: e \in E\}|}{|P|} \quad (10)$$

由于 $H^t(u, v) = \sum_{p \in P_{u, v}} \prod_{e \in p} \omega_e^t(e)$, $\omega^t(p) = \prod_{e \in p} \omega_e^t(e)$,

故式 (10) 可以写成:

$$q_e^t = (1 - \gamma) \frac{\sum_{p: e \in p} \omega^t(p)}{W^t} + \gamma \frac{|\{p \in P: e \in E\}|}{|P|} \quad (11)$$

即边被选择的概率为经过该边的所有完整路径被选择的概率之和。完整路径被选择的概率为:

$$q_p^t = (1 - \gamma) \frac{\omega^t(p)}{W^t} + \frac{\gamma}{|P|} \quad (12)$$

由于 $W^t = \sum_{p \in P} \omega^t(p), \forall t \in [T]$ 且

$$\begin{aligned} \omega^{t+1}(p) &= \omega^t(p) \cdot e^{\eta \tilde{g}^t(p)}, \forall p \in P, t \in [T-1] \\ \omega_e^t &= \omega_e^{t-1} e^{\eta \tilde{g}_e^t} \end{aligned} \quad (13)$$

首先获得 $\ln\left(\frac{W^T}{W^0}\right)$ 的下界:

$$\ln\left(\frac{W^T}{W^0}\right) = \ln \sum_{p \in P} e^{\eta \tilde{G}_p^T} - \ln |P| \geq \eta \max_{p \in P} \tilde{G}_p^T - \ln |P| \quad (14)$$

且

$$\begin{aligned} \ln \frac{W^t}{W^{t-1}} &= \ln \sum_{p \in P} \frac{\omega^{t-1}(p)}{W^{t-1}} e^{\eta \tilde{g}^t(p)} \\ &= \ln \left(\sum_{p \in P} \frac{q^t(p) - \frac{\gamma}{|P|}}{1 - \gamma} e^{\eta \tilde{g}^t(p)} \right) \\ &\leq \ln \left(\sum_{p \in P} \frac{q^t(p) - \frac{\gamma}{|P|}}{1 - \gamma} \left(1 + \eta \tilde{g}^t(p) + \eta^2 (\tilde{g}^t(p))^2 \right) \right) \\ &\leq \ln \left(1 + \sum_{p \in P} \frac{q^t(p)}{1 - \gamma} \left(\eta \tilde{g}^t(p) + \eta^2 (\tilde{g}^t(p))^2 \right) \right) \\ &\leq \frac{\eta}{1 - \gamma} \sum_{p \in P} q^t(p) \tilde{g}^t(p) + \frac{\eta^2}{1 - \gamma} \sum_{p \in P} \left(q^t(p) (\tilde{g}^t(p))^2 \right) \end{aligned} \quad (15)$$

其中, 第二个等式使用了完整路径被选择的概率定义式 (12), 最后一个不等式对于任意 $x > -1$, 有 $\ln(1+x) \leq x$. 由于:

$$\begin{aligned} \eta \tilde{g}^t(p) &= \eta \sum_{e \in p} \tilde{g}_e^t(e) \\ &\leq \eta \sum_{e \in p} \frac{1 + \beta_2}{\gamma} \\ &\leq \frac{\eta n(1 + \beta_2) |P|}{\gamma} \leq 1 \end{aligned} \quad (16)$$

其中 $q_e^t \geq \frac{\gamma}{|P|}$, 故第一个不等号成立, 由于:

$$\begin{aligned} \sum_{p \in P} q^t(p) \tilde{g}^t(p) &= \sum_{p \in P} q^t(p) \sum_{e \in p} \tilde{g}_e^t(e) \\ &= \sum_{e \in E} \tilde{g}_e^t(e) \sum_{p \in P: e \in p} q^t(p) \\ &= \sum_{e \in E} \tilde{g}_e^t(e) q_e^t \\ &= g^t(p') + E\beta_2 \end{aligned} \quad (17)$$

其中第三个等号中的 q_e^t 是对 DAG 图中任意边进行采样的概率, 定义为式 (11). 第四个等号中 p' 为

实际中采样的路径。

$$\begin{aligned} \sum_{p \in P} q^t(p) (\tilde{g}^t(p))^2 &= \sum_{p \in P} q^t(p) \left(\sum_{e \in p} \tilde{g}_e^t(e) \right)^2 \\ &\leq n \sum_{p \in P} q^t(p) \sum_{e \in p} (\tilde{g}_e^t(e))^2 \\ &= n \sum_{e \in E} (\tilde{g}_e^t(e))^2 \sum_{p \in P: e \in p} q^t(p) \\ &= n \sum_{e \in E} (\tilde{g}_e^t(e))^2 q_e^t \\ &= n \sum_{e \in E} \tilde{g}_e^t(e) q_e^t \frac{\beta_2 + 1_{\{e \in p'\}} g^t(e)}{q_e^t} \\ &\leq n(1 + \beta_2) \sum_{e \in E} \tilde{g}_e^t(e) \end{aligned} \quad (18)$$

其中 n 为战场数量, 第一个不等式运用算术均值和二次均值, 由于 $g^t(e) \leq 1$, 故第二个不等式成立。则:

$$\ln \frac{W^t}{W^{t-1}} \leq \frac{\eta}{1 - \gamma} (g^t(p') + E\beta_2) + \frac{\eta^2 n(1 + \beta_2)}{1 - \gamma} \sum_{e \in E} \tilde{g}_e^t(e) \quad (19)$$

由式 (19) 可得到:

$$\begin{aligned} \ln \frac{W^T}{W^{T-1}} &\leq \frac{\eta}{1 - \gamma} (\hat{G}^T + TE\beta_2) + \frac{\eta^2 n(1 + \beta_2)}{1 - \gamma} \sum_{e \in E} \hat{G}^T(e) \\ &\leq \frac{\eta}{1 - \gamma} (\hat{G}^T + TE\beta_2) + \frac{\eta^2 n(1 + \beta_2)}{1 - \gamma} |P| \max_{p \in P} \hat{G}^T(p) \end{aligned} \quad (20)$$

由于 $\sum_{e \in E} \tilde{G}^T(e) \leq \sum_{p \in P} \tilde{G}^T(p)$, 故第二个不等号成

立。结合式 (14), 有:

$$\hat{G}^T \geq (1 - \gamma - \eta n(1 + \beta_2) |P|) \max_{p \in P} \tilde{G}^T(p) - \frac{1 - \gamma}{\eta} \ln |P| - TE\beta_2 \quad (21)$$

由文献 [34] 中的引理 6.7 可知, 对于任意 $\delta \in (0, 1)$, $0 < \beta_2 \leq 1$ 且 $e \in E$, 有:

$$\mathbb{P} \left[G_e^T > \tilde{G}_e^T + \frac{1}{\beta_2} \ln \frac{|E|}{\delta} \right] \leq \frac{\delta}{|E|} \quad (22)$$

使用式 (22) 得: 对于任意 $\delta \in (0, 1)$, 以下不等式成立的概率至少为 $1 - \delta$, 由不等式转化为概率不等式, 只需要考虑 $\max$ 项。

$$\begin{aligned} \hat{G}^T &\geq (1 - \gamma - \eta n(1 + \beta_2) |P|) \left(\max_{p \in P} \tilde{G}^T(p) - \frac{n}{\beta_2} \ln \frac{E}{\delta} \right) - \\ &\quad \frac{1 - \gamma}{\eta} \ln |P| - TE\beta_2 \end{aligned} \quad (23)$$

由于 $\hat{G}^T = T - \hat{L}^T$, 且对于任意 $p \in P$, 有 $G^T(p) =$

$T-L^T(p)$, 将式 (23) 转化为以下不等式, 且成立的概率至少为 $1-\delta$.

$$\begin{aligned} \hat{L}^T &\leq T - (1-\gamma-\eta n(1+\beta_2)|P|) \cdot \\ &\left(\max_{p \in P} \tilde{G}^T(p) - \frac{n}{\beta_2} \ln \frac{E}{\delta} \right) + \frac{1-\gamma}{\eta} \ln |P| + TE\beta_2 = \\ &T - (1-\gamma-\eta n(1+\beta_2)|P|) \cdot \\ &\left(T - \min_{p \in P} L^T(p) - \frac{n}{\beta_2} \ln \frac{E}{\delta} \right) + \frac{1-\gamma}{\eta} \ln |P| + TE\beta_2 = \\ &T(\gamma+\eta n(1+\beta_2)|P|) + (1-\gamma-\eta n(1+\beta_2)|P|) \cdot \\ &\min_{p \in P} L^T(p) + (1-\gamma-\eta n(1+\beta_2)|P|) \cdot \\ &\frac{n}{\beta_2} \ln \frac{E}{\delta} + \frac{1-\gamma}{\eta} \ln |P| + TE\beta_2 \end{aligned} \quad (24)$$

则:

$$\begin{aligned} \hat{L}^T - \min_{p \in P} L^T(p) &\leq T(\gamma+\eta n(1+\beta_2)|P|) + \\ &\frac{n}{\beta_2} \ln \frac{E}{\delta} + \frac{1-\gamma}{\eta} \ln |P| + TE\beta_2 \\ &\leq T\gamma + 2\eta nT|P| + \frac{n}{\beta_2} \ln \frac{E}{\delta} + \frac{\ln |P|}{\eta} + TE\beta_2 \quad (25) \\ \text{令 } \beta_2 &= \sqrt{\frac{n}{T|E|} \ln \frac{|E|}{\delta}}, \gamma = 2\eta n|P|, \eta = \sqrt{\frac{\ln |P|}{4Tn|P|}}, \end{aligned}$$

则:

$$\begin{aligned} \hat{L}^T - \min_{p \in P} L^T(p) &\leq 2 \left(\sqrt{nT|E| \ln \frac{|E|}{\delta}} + \right. \\ &\left. 2\sqrt{nT|P| \ln |P|} \right) \end{aligned} \quad (26)$$

3.2.2 ALG2 算法设计

本文采用 Hedge^[14] 中的 Hedge 算法作为 ALG2 算法, Immorlica 等人^[28] 证明了 Hedge 算法在全信息反馈条件下能够实现以下高概率遗憾上界。

$$R_\delta(T) = O\left(\sqrt{T \ln \left(\frac{|A|}{\delta}\right)}\right) \quad (27)$$

Hedge 算法的伪代码如算法 3 所示。

3.2.3 LagrangeBwK-Exp3-G 算法

以上介绍了上层级 Hedge 算法和下层级 Exp3-G 算法, 并通过数学推导证明了 Exp3-G 实现的 $1-\delta$ 高概率遗憾上界, 其中 $\delta \in (0, 1)$ 。将上层级 Hedge 算法 (ALG2) 和下层级 Exp3-G 算法 (ALG1) 进行整合, 设计求解多阶段对抗条件下资源分配问题的 Lagrange-BwK-Exp3-G 算法, 实现多阶段对抗条

算法 3 Hedge

Input: $\beta_1 \in (0, 1), T$;

```

1: 初始化: 对于任意  $i \in A, \omega_i^t = \frac{1}{|A|}$ ,  $A$  表示动作集;
2: for  $t = 1, \dots, T$  do
3:   以概率  $p_i^t = \frac{\omega_i^t}{\sum_{i \in A} \omega_i^t}$  采样一个动作  $i_t \in A$ ;
4:   获得损失  $\{l_i^t\}, \forall i \in A$ ;
5:   更新动作权重  $\omega_i^{t+1} = \omega_i^t \cdot \beta_1^{l_i^t}, \forall i \in A$ ;
6: end for

```

件下资源分配问题的高概率遗憾最小化。设 τ 为算法终止的时间阶段, 在此之前的时间阶段 t , 学习者将按照式 (28) 在固定的预算约束 m 内进行策略选择。

$$S = \{p \in Z_+^n: 1^T p \leq m\} \quad (28)$$

在时间阶段 τ , 设置学习者将剩余的资源 $x(\tau)$ 作为 τ 轮次的预算进行策略选择。在时间阶段 τ 之后, 学习者的阶段预算都为零, 此时是否继续和对手进行对抗, 需要根据具体的布洛托战场胜负规则 (式 (1)) 确定, 如果布洛托规则规定在双方分配资源相等的情况下, 学习者获胜, 则即使学习者分配的资源数量为零, 在对手也分配资源为零的战场上, 学习者就能够获得该战场的价值收益。

$$\tilde{S}_\tau = \{p \in Z_+^n: 1^T p = x(\tau)\} \quad (29)$$

在上层级, 使用 Hedge 算法维持每一轮的资源消耗和总预算在时间范围 T 内的资源分配之间的平衡, 二者产生的回报 ($\Gamma_{\text{resource}}^t$ 和 Γ_{time}^t) 如式 (30) 所示, 在时间阶段 t 对抗结束后, 通过 Hedge 算法从 $\Gamma_{\text{resource}}^t$ 和 Γ_{time}^t 中进行选择, 作为奖励传递给 Exp3-G 算法, 进行边的权重更新。每一轮的资源消耗和阶段预算的回报设计为:

$$\begin{aligned} \Gamma_{\text{resource}}^t &= r^t(p^t) + 1 - \frac{T}{B} \cdot \omega^t(p^t) \in [1-c, 2] \quad (30) \\ \Gamma_{\text{time}}^t &= r^t(p^t) \in [0, 1] \end{aligned}$$

其中 $\omega^t(p^t)$ 表示学习者在时间阶段 t 的预算消耗, 且 $\omega^t(p^t) \leq m = \frac{cB}{T}$, B 表示资源总量, c 表示学习者在每一轮对抗过程中最多分配总预算在时间范围 T 内的平均值的 c 倍。下面给出 LagrangeBwK-Exp3-G 算法过程。

根据引理 1, 求解 LagrangeBwK-Exp3-G 的高概率遗憾上界。由于 Exp3-G 算法的 $1-\delta$ 高概率遗憾上界为:

算法4 LagrangeBwK-Exp3-G

Input: $B, T, n, m = c \frac{B}{T}$, ALG1, ALG2;

```

1: for  $t = 1, \dots, T$  do
2:   对手选择分配策略;
3:   利用 ALG1 算法采样一条路径  $p^t \in S$ ;
4:   if  $x(t) \leq m$ 
5:     利用 ALG1 算法采样一条路径  $p^t \in \tilde{S}$ ;
6:   算法终止;
7: end if
8: 计算  $\Gamma_{\text{resource}}^t$  和  $\Gamma_{\text{time}}^t$ ;
9: 利用 ALG2 选择  $\Gamma_{\text{resource}}^t$  或者  $\Gamma_{\text{time}}^t$ ;
10: if ALG2 选择  $\Gamma_{\text{resource}}^t$ 
11:   将  $\Gamma_{\text{resource}}^t$  作为奖励传递给 ALG1;
12: if ALG2 选择  $\Gamma_{\text{time}}^t$ 
13:   将  $\Gamma_{\text{time}}^t$  作为奖励传递给 ALG1;
14: end if
15: 使用 ALG1 更新 DAG 中边的权重;
16: 将  $\Gamma_{\text{resource}}^t$  和  $\Gamma_{\text{time}}^t$  作为损失传递给 ALG2 更新二者的权重;
17: end if

```

$$R_\delta(T) = \hat{L}^T - \min_{p \in P} L^T(p) \leq 2 \left(\sqrt{nT|E| \ln \frac{|E|}{\delta}} + 2\sqrt{nT|P| \ln |P|} \right) \quad (31)$$

Hedge 算法在全信息反馈条件下的 $1-\delta$ 高概率遗憾上界为:

$$R_\delta(T) = O \left(\sqrt{T \ln \left(\frac{|A|}{\delta} \right)} \right) \quad (32)$$

其中, 引理1中的 B_0 在本问题中扩展为 $\frac{B}{cB/T} = \frac{T}{c}$, 满足 $O(\frac{T}{B_0}) = O(1)$, 则 LagrangeBwK-Exp3-G 算法高概率遗憾上界可以由定理2给出。

定理2 LagrangeBwK-Exp3-G 算法以高概率实现如下所示的遗憾上界。

$$\begin{aligned} R(T) &\leq O(1) \cdot \left(O \left(2 \left(\sqrt{nT|E| \ln \frac{T \cdot |E|}{\delta}} + 2\sqrt{nT|P| \ln |P|} \right) \right) + O \left(\sqrt{T \ln \left(\frac{T}{\delta} \right)} \right) \right) \\ &= O \left(\sqrt{nT|E| \ln \frac{T \cdot |E|}{\delta}} + 2\sqrt{nT|P| \ln |P|} \right) \end{aligned} \quad (33)$$

其中 $|E| = O(nm^2)$, $m = O(\frac{B}{T})$ 且 $|P| = O(2^{\min\{n-1, m\}})$, 满足 $\ln |P| \leq O(m) = O(\frac{B}{T})$ 。

3.2.4 算法运行效率分析

下面对 LagrangeBwK-Exp3-G 算法的运行效率

进行简单分析。LagrangeBwKExp3-G 算法的运行效率取决于 Exp3-G 算法, 由于 Exp3-G 算法主要面向下层级的资源分配策略的选择, 表现为 DAG 中的路径采样, Exp3-G 算法的作用对象仍然是 DAG 中的边, 如上文所述, $H^t(u, v) = \sum_{p \in P_{u,v}} \prod_{e \in p} \omega^t(e)$ 表示经过顶点 (u, v) 的所有路径的权重之和, 则 $H^t(s, d)$ 的运行时间为 $O(|E|)$ (s 和 d 分别为 DAG 图中的源点和终点), 则 Exp3-G 算法的运行时间为 $O(|E|T)$ 。针对在线学习问题的遗憾最小化算法, 一般考虑 T 的取值明显大于其他参数的设置, 因此本文设计的 Exp3-G 算法在提高算法运行效率方面的作用是显著的。

4 对抗条件下卫星多信道功率分配实验

4.1 卫星多信道功率分配问题描述

随着通信系统软硬件技术的不断发展, 卫星跳频通信系统每次使用多个信道成为了可能。例如, 多用户同步正交跳频的卫星通信系统中, 通信方可以同时申请占用多个信道, 以增大通信系统总容量。然而, 随着侦查技术的不断进步, 高空长航时侦查无人机以及具有抵近侦查能力的小卫星等对抗设备逐渐实用化, 干扰方对通信方实施跟踪干扰变得越来越容易。因此, 卫星通信中每次占用多信道进行信息传输面临的挑战和问题仍然十分严峻, 对多信道功率分配抗干扰问题的研究具有重大意义。

在具有总电量约束的卫星通信多信道功率分配抗干扰场景中, 假设电力资源无法支持卫星通信抗干扰持续作战, 通信方需要在固定燃料储备的条件下进行必要的通信信息传输。假设发电机燃料和功率的转化比为 1:1, 此时可以直接将燃料的总量当作功率的总预算。考虑在功率总预算约束下通信方和干扰方之间的多阶段对抗条件下功率资源分配问题, 分别设置对抗双方功率总预算相等和不相等的情形, 验证算法是否能够实现累积遗憾最小化的目标。

在多阶段对抗条件下的每一次对抗问题中, 通信方使用多个信道进行信息传输, 而干扰方通过侦查技术手段, 能够对通信方实施跟踪干扰或阻塞式干扰, 但是受限于总干扰功率的限制, 不能同时阻塞用户所有的信道。对抗双方同时分配功率资源至所选信道, 各个信道之间相互独立, 对于通信信道中的每一个信道 k , 如果接收机接收到的信噪比满

足信噪比计算式 $\frac{R_k}{M_k + N_k} \geq \beta$ ，则可以成功传输信息，否则不能进行传输。其中 β 由服务类型决定， N_k 是通道 k 对应的白噪声的方差。结合多信道功率分配应用背景，将式 (1) 表示的经典布洛陀博弈效用函数 l 重新设计为：

$$l(R_k - (M_k + N_k)\beta) = \begin{cases} 1, R_k - (M_k + N_k)\beta \geq 0 \\ -1, R_k - (M_k + N_k)\beta < 0 \end{cases} \quad (34)$$

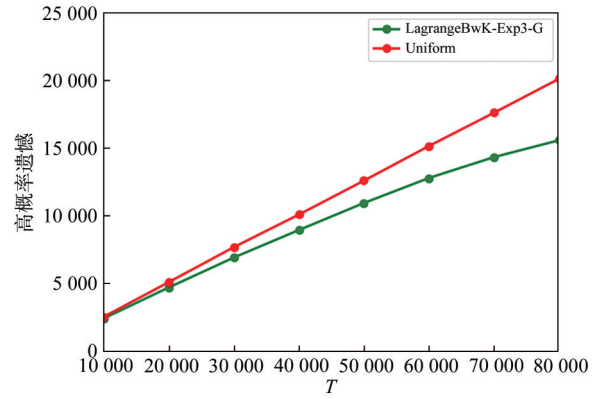
4.2 对抗双方功率总预算相等

考虑对抗双方功率资源总预算相等的情形，其中干扰方在时间范围 T 内平均分配总预算至各个时间阶段 t ，即每个时间阶段的资源数为 B/T 。在每一个时间阶段 t ，干扰方随机选择一个信道，并将 $1W$ 的功率分配至该信道，直至 B/T 的功率分配完，即干扰方采取的策略服从随机均匀分布。而对于通信方，每一次对抗的功率预算满足 $\omega^t(p^t) \leq m = c \frac{B}{T}$ ，其中 $c=2$ ，即通信方在每个时间阶段的资源数是资源总数在时间范围 T 内的平均数的 2 倍。对于每一次对抗，根据 LagrangeBwK-Exp3-G 对 DAG 图中各个边的权重进行调整，从而达到控制下一阶段预算的目的。设置信道数量 $n=3$ 以及 $c=2$ ，且各个信道价值为 $[1/3, 1/3, 1/3]$ ，分别验证 $m=4$ 和 $m=6$ 的情形。为进一步说明 LagrangeBwK-Exp3-G 算法的有效性，实验采用 Uniform 算法作为 LagrangeBwK-Exp3-G 算法的对比算法，Uniform 算法设置的每个时间阶段的资源数量和 LagrangeBwK-Exp3-G 算法相同，但是 Uniform 算法不考虑上一阶段的信息反馈，在各个时间阶段以及各个信道上的分配策略服从均匀分配。实验结果如图 2 所示。

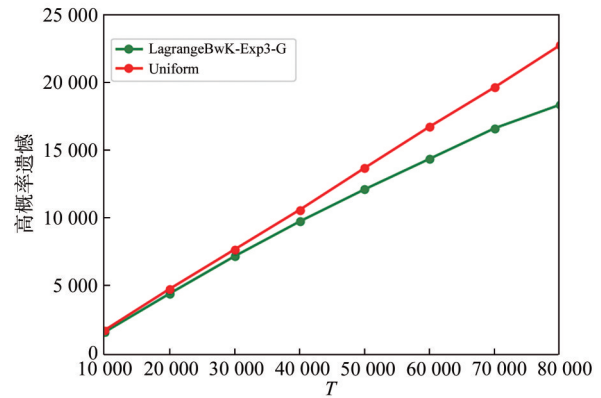
从图 2 中可以看出，对于不同的功率总预算约束 B 和功率阶段预算 m ，LagrangeBwK-Exp3-G 算法都能够实现关于时间范围 T 的次线性遗憾，而 Uniform 算法只能实现关于时间范围 T 的线性遗憾（遗憾曲线关于 T 成线性关系），通过对比显示了 LagrangeBwK-Exp3-G 算法的有效性。由于 LagrangeBwK-Exp3-G 算法使得 $T \rightarrow \infty$ 时， $R_T/T \rightarrow 0$ ，验证了 LagrangeBwK-Exp3-G 算法是遗憾最小化算法。

4.3 对抗双方功率总预算不相等

考虑对抗双方功率资源总预算不相等的情形，根据信噪比计算式，通信方进行抗干扰的成本要比干扰方高，设置对手每次对抗时分配 $m+1$ 至 $n-1$ 个信道，剩下一个信道分配 B/T 的功率，而通信方不



(a) $n=3, m=4, B=2T$



(b) $n=3, m=6, B=3T$

图2 对抗双方预算相等时，累积遗憾对比结果

知道干扰方在哪个信道分配了 B/T 功率，通信方每次对抗的功率预算满足 $\omega^t(p^t) \leq m = c \frac{B}{T}$ ，即通信方每次对抗的预算最多为 m ，而总资源在时间范围 T 内的平均值为 B/T 。根据信噪比计算式，通信方和干扰方在同一个信道上分配的功率相等时，通信方仍然能够成功传输信息，因此通信方最优策略为将 B/T 功率分配至干扰方也分配了 B/T 功率的信道上。通信方需要在消耗较少资源的情况下，快速分辨出对手分配 B/T 功率所在的信道，从而最大化在时间范围 T 内的总增益（即最小化遗憾）。设置 $n=3$ 和 $c=2$ ，且各个信道价值为 $[1/3, 1/3, 1/3]$ ，分别验证 $m=4$ 和 $m=6$ 的情形。实验仍然采用 Uniform 算法作为 LagrangeBwK-Exp3-G 算法的对比算法，进一步说明 LagrangeBwK-Exp3-G 算法的有效性，Uniform 算法设置的每个时间阶段的资源数量和 LagrangeBwK-Exp3-G 算法相同，且在各个时间阶段以及各个战场上的分配策略服从均匀分布。实验结果如图 3 所示。

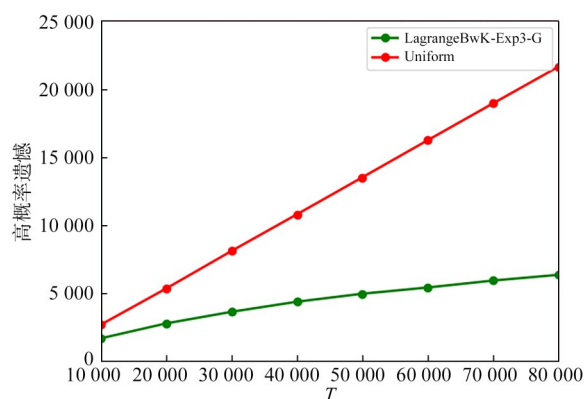
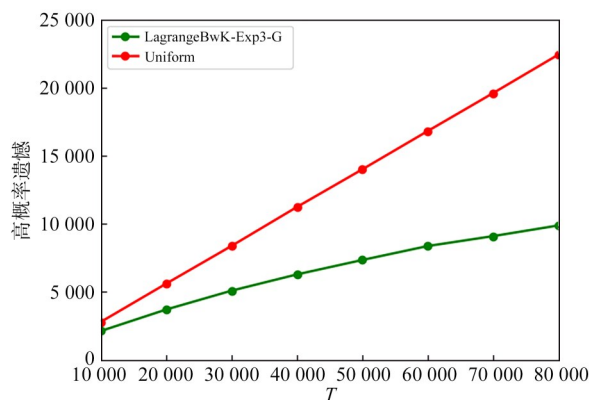
(a) $n=3, m=4, B=2T$ (b) $n=3, m=6, B=3T$

图3 对抗双方预算不相等时, 累积遗憾对比结果

对于不同的功率资源总预算 B 和阶段预算 m , LagrangeBwK-Exp3-G 算法都能够实现关于时间范围 T 的次线性遗憾。这说明随着对抗活动的进行, 算法能够找到对手分配 B/T 功率所在的信道, 且总功率在时间范围 T 内的分配策略也为 B/T , 从而实现通信方在时间范围 T 内的总增益最大化, 而Uniform 算法实现的累积遗憾曲线关于 T 成线性关系, 因为实验设置干扰方的资源预算远远大于通信方的资源预算, 每次对抗时干扰方在 $n-1$ 个信道中的每个信道上分配 $m+1$ 功率, 剩余一个信道分配 B/T 的功率, 通信方只能在剩余的信道上分配足够的资源才能赢得该信道价值。Uniform 算法无法针对干扰方在上一时间阶段的分配策略做出策略调整, 因此通信方在绝大多数的对抗中失去信道价值。通过LagrangeBwK-Exp3-G 算法实现的累积遗憾曲线以及与Uniform 算法的对比分析, 显示了LagrangeBwK-Exp3-G 算法的有效性, 证明了LagrangeBwK-Exp3-G 算法是遗憾最小化算法。

5 结论

本文设计LagrangeBwK-Exp3-G 算法求解了多阶段对抗条件下的资源分配问题, 实现了关于时间范围 T 的高概率次线性遗憾界。多阶段对抗条件下的资源分配问题是一种在线多阶段布洛托博弈问题, 学习者每次获得的信息反馈属于半信息反馈。学习者不但要考虑在上层级如何在时间范围 T 内将总预算分配至各个时间阶段 t 的问题, 还要考虑在下层级的每个时间阶段 t 和对手进行一次性对抗条件下的布洛托博弈问题。上下层级的两个算法相辅相成, 它们的目标是最小化时间周期 T 内的损失。该问题的难点在于如何将下层级的损失合理地反馈至上层级, 进而对上层级的分配做出指导, 这需要解决奖励和消耗的平衡问题。奖励是由对手决定的, 这增加了问题的不确定性。本文设计的LagrangeBwK-Exp3-G 算法考虑了消耗和奖励之间的平衡, 实现了关于时间范围 T 的高概率次线性遗憾界, 并结合卫星多信道功率分配实验对LagrangeBwK-Exp3-G 算法的运行效率进行分析, 结果表明该算法具有高效的运行效率。

本文考虑的功率资源是单一类型的资源, 为了更加贴近实际, 未来的工作应该考虑异构资源的分配。不同的资源组合产生不同的作战效用, 单一战场的输赢不是简单地以数量取胜, 而是以不同类型的资源组合产生的作战效用为标准。学习者不仅需要考虑指定资源分配策略的组合优化问题, 还需要考虑不同类型的资源之间的组合优化问题, 虽然问题难度升级, 但是更加贴近实际, 为指挥员的战场决策提供了良好支撑。

参考文献:

- [1] 郭行, 符文星, 闫杰. 浅析美军马赛克作战概念及启示[J]. 无人系统技术, 2020, 3(6): 92-106.
GUO H, FU W X, YAN J. Analysis and inspiration of the U. S. force's concept of mosaic warfare[J]. Unmanned Systems Technology, 2020, 3(6): 92-106.
- [2] 吴明曦. 智能化战争: AI军事畅想[M]. 北京: 国防工业出版社, 2020.
WU M X. Intelligent warfare: the imagination of the AI military[M]. Beijing: National Defense Industry Press, 2020.
- [3] GRANA J, LAMB J, O'DONOUGHUE N A. Findings on mosaic warfare from a colonel blotto game [M]. Santa Monica: RAND Corporation, 2021.
- [4] BOREL E. La théorie du jeu et les équations intégrales à noyau symétrique[J]. Comptes Rendus De l'Académie Des Sciences, 1921, 173: 1304-1308.

- [5] KOVENOCK D, ROBERSON B. Generalizations of the general lotto and colonel blotto games[J]. *Economic Theory*, 2021, 71(3): 997-1032.
- [6] ZOU M W, CHEN S F, LUO J R, et al. An evolutionary learning approach for anti-jamming game in cognitive radio confrontation[C]// *Proceedings of 2022 IEEE International Conference on Systems, Man, and Cybernetics*. Piscataway: IEEE Press, 2022: 3210-3215.
- [7] ZHANG L, WANG Y, MIN M H, et al. Privacy-aware laser wireless power transfer for aerial multi-access edge computing: a colonel blotto game approach[J]. *IEEE Internet of Things Journal*, 2023, 10(7): 5923-5939.
- [8] ZHOU X B, DONG X D, ZHAO L P, et al. Learning-driven cloud resource provision policy for content providers with competitor[J]. *IEEE Transactions on Cloud Computing*, 2022, 10(3): 1913-1924.
- [9] CHEN A K, FERGUSON B L, SHISHIKA D, et al. Path defense in dynamic defender-attacker blotto games (dDAB) with limited information [C]// *Proceedings of 2023 American Control Conference (ACC)*. Piscataway: IEEE Press, 2023: 447-453.
- [10] SHISHIKA D, GUAN Y, DOROTHY M, et al. Dynamic defender-attacker blotto game[C]// *Proceedings of 2022 American Control Conference (ACC)*. Piscataway: IEEE Press, 2022: 4422-4428.
- [11] KHARLAMOV V V. On asymptotic strategies in the stochastic colonel blotto game[J]. *Theory of Probability & Its Applications*, 2022, 67(2): 318-326.
- [12] ANBARCI N, CINGIZ K, ISMAIL M S. Proportional resource allocation in dynamic n-player Blotto games[J]. *Mathematical Social Sciences*, 2023, 125: 94-100.
- [13] NEU G. Explore no more: improved high-probability regret bounds for non-stochastic bandits[J]. *Advances in Neural Information Processing Systems*, 2015, 2015-January: 3168-3176.
- [14] 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.
- [15] 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*. New York: ACM, 2007: 345-352.
- [16] 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.
- [17] 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.
- [18] 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*. New York: ACM, 2021: 862-863.
- [19] 张骁雄, 葛冰峰, 谭跃进. 军事攻防中的多属性资源分配对策模型[J]. *国防科技大学学报*, 2018, 40(5): 153-160.
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.
- [20] 刘冰雁, 叶雄兵, 周赤非, 等. 基于改进DQN的复合模式在轨服务资源分配[J]. *航空学报*, 2020, 41(5): 256-264.
LIU B Y, YE X B, ZHOU C F, et al. Allocation of composite mode on-orbit service resource based on improved DQN[J]. *Acta Aeronautica et Astronautica Sinica*, 2020, 41(5): 256-264.
- [21] 孙宇祥, 彭益辉, 李斌, 等. 智能博弈综述: 游戏AI对作战推演的启示[J]. *智能科学与技术学报*, 2022, 4(2): 157-173.
SUN Y X, PENG Y H, LI B, et al. Overview of intelligent game: enlightenment of game AI to combat deduction[J]. *Chinese Journal of Intelligent Science and Technology*, 2022, 4(2): 157-173.
- [22] 李宪港, 李强. 典型智能博弈系统技术分析及指控系统智能化发展展望[J]. *智能科学与技术学报*, 2020, 2(1): 36-42.
LI X G, LI Q. Technical analysis of typical intelligent game system and development prospect of intelligent command and control system[J]. *Chinese Journal of Intelligent Science and Technology*, 2020, 2(1): 36-42.
- [23] ROBBINS H. Some aspects of the sequential design of experiments[J]. *Bulletin of the American Mathematical Society*, 1952, 58(5): 527-535.
- [24] 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.
- [25] LAI T L, ROBBINS H, et al. Asymptotically efficient adaptive allocation rules[J]. *Advances in Applied Mathematics*, 1985, 6(1): 4-22.
- [26] AUER P, CESA-BIANCHI N, FISCHER P. Finite-time analysis of the multiarmed bandit problem[J]. *Machine Learning*, 2002, 47(2): 235-256.
- [27] AUER P, CESA-BIANCHI N, FREUND Y, et al. The nonstochastic multiarmed bandit problem[J]. *SIAM Journal on Computing*, 2002, 32(1): 48-77.[LinkOut]
- [28] IMMORLICA N, SANKARARAMAN K A, SCHAPIRE R, et al. Adversarial bandits with knapsacks[C]// *Proceedings of 2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. Piscataway: IEEE Press, 2019: 202-219.
- [29] BADANIDIYURU A, KLEINBERG R, SLIVKINS A. Bandits with knapsacks[C]// *Proceedings of 2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. Piscataway: IEEE Press, 2013: 207-216.
- [30] AGRAWAL S, DEVANUR N R. Linear contextual bandits with knapsacks[EB]. *arXiv preprint*, 2015, arXiv: 1507.06738.
- [31] LI X, SUN C, YE Y. The symmetry between arms and knapsacks: a primal-dual approach for bandits with knapsacks[EB]. *arXiv preprint*, 2021, arXiv: 2102.06385.
- [32] LEON V, ETESAMI S R. Bandit learning for dynamic colonel blotto game with a budget constraint[C]// *Proceedings of 2021 60th IEEE Conference on Decision and Control (CDC)*. Piscataway: IEEE Press, 2021: 3818-3823.
- [33] VU D Q, LOISEAU P, SILVA A, et al. Path planning problems with side observations—when colonels play hide-and-seek[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(2): 2252-2259.
- [34] CESA-BIANCHI N, LUGOSI G. *Prediction, learning, and games*[M]. Cambridge: Cambridge University Press, 2006.

[作者简介]

陈少飞 (1987—), 男, 博士, 国防科技大学智能科学学院副教授, 主要研究方向为多智能体系统、机器学习和博弈。





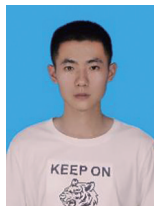
邹明我（1992- ），男，国防科技大学智能科学学院硕士生，主要研究方向为智能决策、博弈和组合优化。



罗俊仁（1989- ），男，国防科技大学智能科学学院博士生，主要研究方向为不完美信息博弈和多智能体学习。



苏小龙（2000- ），男，国防科技大学智能科学学院硕士生，主要研究方向为强化学习。



冯俊侨（2001- ），男，国防科技大学军政基础教育学院本科生，主要研究方向为任务规划和博弈。