软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2022,33(9):3223−3235 [doi: 10.13328/j.cnki.jos.006386]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种满足差分隐私的图赌博机算法*

卢世银, 王广辉, 邱梓豪, 张利军

(计算机软件新技术国家重点实验室 (南京大学), 江苏 南京 210023)

通信作者: 张利军, E-mail: zhanglj@lamda.nju.edu.cn

摘　要: 图赌博机是一种重要的不确定性环境下的序列决策模型, 在社交网络、电子商务和推荐系统等领域都得到了广泛的应用. 目前, 针对图赌博机的工作都只关注如何快速识别最优摇臂从而最小化累积遗憾, 而忽略了在很多应用场景中存在的隐私保护问题. 为了克服现有图赌博机算法的缺陷, 提出了一种满足差分隐私的图赌博机算法 GAP (图反馈下的差分隐私摇臂消除策略). 一方面, GAP 算法阶段性地根据摇臂的经验平均奖赏更新摇臂选取策略, 并在计算摇臂的经验平均奖赏时引入拉普拉斯噪声, 从而确保恶意攻击者难以根据算法输出推算摇臂奖赏数据, 保护了隐私. 另一方面, GAP 算法在每个阶段根据精心构造的反馈图的独立集探索摇臂集合, 有效地利用了图形式的反馈信息. 证明了 GAP 算法满足差分隐私性质, 具有与理论下界相匹配的遗憾界. 在仿真数据集上的实验结果表明: GAP 算法在有效保护隐私的同时取得了与现有无隐私保护的图赌博机算法相当的累积遗憾.

关键词: 图赌博机; 差分隐私; 不确定性环境下的序列决策; 独立集; 拉普拉斯噪声

中图法分类号: TP181

中文引用格式: 卢世银, 王广辉, 邱梓豪, 张利军. 一种满足差分隐私的图赌博机算法. 软件学报, 2022, 33(9): 3223–3235. http://www.jos.org.cn/1000-9825/6386.htm

英文引用格式: Lu SY, Wang GH, Qiu ZH, Zhang LJ. Differentially Private Algorithm for Graphical Bandits. Ruan Jian Xue Bao/Journal of Software, 2022, 33(9): 3223–3235 (in Chinese). http://www.jos.org.cn/1000-9825/6386.htm

Differentially Private Algorithm for Graphical Bandits

LU Shi-Yin, WANG Guang-Hui, QIU Zi-Hao, ZHANG Li-Jun

(State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210023, China)

Abstract: Graphical bandit is an important model for sequential decision making under uncertainty and has been applied in various real-world scenarios such as social network, electronic commerce, and recommendation system. Existing work on graphical bandits only investigates how to identify the best arm rapidly so as to minimize the cumulative regret while ignoring the privacy protection issue arising in many real-world applications. To overcome this deficiency, a differentially private algorithm is proposed, termed as graph-based arm elimination with differential privacy (GAP), for graphical bandits. On the one hand, GAP updates the arm selection strategy based on empirical mean rewards of arms in an epoch manner. The empirical mean rewards are perturbed by Laplace noise, which makes it hard for malicious attackers to infer rewards of arms from the output of the algorithm, and thus protects the privacy. On the other hand, in each epoch, GAP carefully constructs an independent set of the feedback graph and only explores arms in the independent set, which effectively utilize the information in the graph feedback. It is proved that GAP is differential private and its regret bound matches the theoretical lower bound. Experimental results on synthetic datasets demonstrate that GAP can effectively protect the privacy and achieve cumulative regret comparable to that of existing non-private graphical bandits algorithm.

Key words: graphical bandits; differential privacy; sequential decision making under uncertainty; independent set; Laplace noise

赌博机问题最早于 1933 年由 Thompson 在一篇研究医学实验的论文中提出[1]. 假设针对一种病情存在多种疗法, 它们的治愈率未知. 每当一个新病人到来时, 医生需要选择一种疗法对病人进行治疗. 之后, 医生会观测到治疗

* 基金项目: 国家自然科学基金 (61976112); 江苏省自然科学基金 (BK20200064)
收稿时间: 2021-03-31; 修改时间: 2021-05-04; 采用时间: 2021-05-26; jos 在线出版时间: 2022-06-15

效果, 即治愈或未治愈, 并据此改进对每种疗法治愈率的估计. 随着所诊治病人数目的增多, 医生对每种疗法治愈率的估计也越来越准确. 医生的目标是最大化治愈病人数. 上述问题之所以称为赌博机问题, 是因为其内在机制与具有多个摇臂 (arm) 的赌博机很相似. 具体来说, 考虑一个与多臂赌博机迭代交互的学习者 (learner). 在每个回合, 学习者首先选取一个摇臂, 之后获得采样自所选摇臂奖赏分布的随机奖赏 (stochastic reward), 最后根据获得的奖赏更新下一回合的摇臂选取策略. 学习者的表现可以用累积遗憾来衡量, 其定义为学习者实际获得的累积奖赏与始终选择最优摇臂的累积奖赏的差值. 为了最小化累积遗憾, 学习者需要平衡好探索 (选取先前较少选取的摇臂来获得更多信息) 和利用 (选取经验平均奖赏最大的摇臂来累计更多奖赏). 除了医学实验, 多臂赌博机也被广泛用于建模很多现实场景中的序列决策问题, 如电子商务[2]、新闻推荐[3]和社交网络[4,5].

然而, 理论研究表明多臂赌博机的最小最大化遗憾界 (minimax regret bound) 为 $\Theta(K\log T)$[6,7], 其中 K 为摇臂数目, T 为回合数. 该遗憾界随着摇臂数目线性增长, 因此无法应用于摇臂太多的情形. 另一方面, 多臂赌博机假设学习者只能观测到所选摇臂的奖赏, 无法建模很多现实场景如推荐系统和社交网络中广泛存在的额外奖赏信息[8]. 例如, 在推荐系统中, 我们可以利用商品之间的相似性, 根据用户对单个商品的反馈来近似判断其对相似商品的喜好程度[9]. 为了克服多臂赌博机的局限性, 文献 [10] 提出了一种多臂赌博机的变体——图赌博机 (graphical bandits), 用一个无向图 G 来刻画学习者获得的反馈信息的结构. 图中的每个顶点对应一个摇臂, 相似摇臂之间通过无向边连接. 在每个回合, 学习者除了获得所选摇臂的奖赏, 还能额外观测到与之相邻摇臂的奖赏信息. 对于图赌博机, 文献 [10] 设计了一种遗憾界为 $\mathrm{O}(\theta\log T)$ 的算法, 其中 θ 为反馈图 G 的团覆盖数 (clique covering number), 满足 $\theta\leqslant K$, 并且对于良性图 (benign graphs), 有 $\theta<<K$. 因此, 与多臂赌博机相比, 图赌博机受摇臂数目的影响更小.

自从文献 [10] 提出图赌博机以来, 机器学习领域已经涌现出了大量关于图赌博机的工作[11−17]. 然而, 现有工作都只注重最小化累计遗憾, 而没有考虑到选择摇臂时对隐私的保护. 以推荐系统为例, 每个商品对应一个摇臂, 根据商品的相似性建立反馈图. 在推荐系统中, 图赌博机算法需要不断地选择摇臂 (推荐商品给用户) 并根据用户反馈 (对所推荐商品的点击或购买情况) 更新摇臂选取策略. 现有图赌博机算法的摇臂选取策略对用户反馈的变动十分敏感. 通过跟踪算法输出随用户反馈的变化情况, 攻击者能够学习到用户的兴趣偏好[18]. 因此, 现有的图赌博机算法有泄露用户隐私的风险, 无法适用于隐私敏感的场景.

差分隐私 (differential privacy) 是一种被广泛认可的隐私保护模型[19,20], 要求更改数据集中的一条记录后, 算法的输出几乎保持不变, 从而确保攻击者难以根据算法的输出推算数据集中的隐私信息. 与传统的隐私保护模型相比, 差分隐私模型的主要优点在于: (1) 不限制恶意攻击者的背景知识; (2) 给出了衡量算法隐私保护效果的量化指标. 近年来, 设计满足差分隐私的赌博机算法已经成为了机器学习领域的研究热点[21−25]. 然而, 现有满足差分隐私的赌博机算法都是为多臂赌博机设计的, 无法利用图形式的反馈信息. 因此, 对于图赌博机问题, 直接应用现有差分隐私赌博机算法只能取得次优的累积遗憾.

在本文中, 我们提出一种既满足差分隐私又能有效利用图形式反馈信息的赌博机算法 GAP. 通过将 T 回合划分成若干个阶段并隔离不同阶段的奖赏数据, GAP 算法将用户反馈序列的单个改变对摇臂选取策略的影响限制到了某个阶段内. GAP 算法只根据每个阶段结束时摇臂的经验平均奖赏 (empirical mean reward) 来选取摇臂, 并通过添加噪声的方式进一步减小了对单个用户反馈的依赖, 从而保护了用户隐私. 我们证明了给定差分隐私预算 ϵ, GAP 算法满足 ϵ-差分隐私. 此外, 为了在保护用户隐私的同时减少累计遗憾, GAP 算法在每个阶段开始时构造反馈图的独立集 (independent set), 并在每个回合只选取独立集中的摇臂, 从而有效地挖掘了图形式的反馈信息, 降低了探索摇臂集的成本. 我们通过理论分析为 GAP 算法建立了 $\mathrm{O}(\alpha\ln T/\epsilon)$ 的遗憾界. 其中, α 表示反馈图 G 的独立数, 满足 $\alpha\leqslant\theta\leqslant K$. 该遗憾界与无隐私保护的图赌博机问题和满足差分隐私的多臂赌博机问题的理论下界相匹配, 表明 GAP 算法在理论上是最优的. 最后, 我们在仿真数据集上测试了 GAP 算法的运行效果. 测试结果表明: GAP 算法能够有效的保护用户隐私, 并且其累积遗憾与现有无隐私保护的图赌博机算法相当.

本文的主要贡献如下:

(1) 提出了一种新颖的图赌博机算法 GAP, 通过阶段划分和随机噪声来保护隐私, 通过构造反馈图的独立集

来有效地利用图形式的反馈信息.

(2) 证明了 GAP 算法满足差分隐私, 并能够取得与理论下界相匹配的最优遗憾界.

(3) 构造了一个图赌博机仿真数据集, 并在仿真数据集上对 GAP 算法进行实验. 实验结果表明: GAP 算法能够在保护隐私的同时取得与无隐私保护的图赌博机算法相当的累积遗憾.

1 相关工作

在本节中, 我们简要回顾图赌博机和差分隐私方面的相关工作.

1.1 图赌博机

文献 [10] 最早研究图赌博机, 提出了一种基于置信上界 (upper confidence bound, UCB) 的图赌博机算法, 并证明了该算法能够取得 $\mathrm{O}(\theta \ln T + K)$ 的遗憾界. 算法的主要思想是根据每个摇臂的平均奖赏和观测次数来选取摇臂, 联合估计所选摇臂及其邻居的期望奖赏. 文献 [11] 改进了文献 [10] 中所提出的算法, 通过引入额外的探索项将遗憾界提升至 $\mathrm{O}(\theta \ln T)$.

文献 [12] 提出了一种基于连续消除 (successive elimination, SE) 的图赌博机算法, 根据反馈图的结构来调整每个摇臂的探索率, 取得了优于 UCB 类算法的遗憾界. 文献 [13] 将文献 [12] 中所提出的算法扩展到了反馈图是有向图、随时间变化且对学习者不完全可见的复杂场景, 并且建立了 $\mathrm{O}(\alpha_{\max} \ln T)$ 的遗憾界, 其中 $\alpha_{\max}$ 是反馈图在 T 回合中的最大独立数 (maximum independence number).

文献 [14−16] 设计了一系列基于汤普森采样 (Thompson sampling, TS) 的图赌博机算法, 但是只推导出了贝叶斯遗憾界, 理论保障弱于具有普通遗憾界的 UCB 类和 SE 类图赌博机算法. 文献 [11] 为 TS 类图赌博机算法建立了首个 $\mathrm{O}(\theta \ln T)$ 的普通遗憾界. 文献 [17] 通过引入分层技术将 UCB 类和 TS 类图赌博机算法的普通遗憾界进一步提升至 $\mathrm{O}(\alpha \ln^2 T)$.

文献 [26] 研究了摇臂奖赏分布随时间变化的场景, 设计了一系列具有次线性动态遗憾界的图赌博机算法, 包括能够自主感知分布变化的自适应算法. 文献 [27] 考虑了奖赏数据可能受到攻击污染的场景, 提出了一种对攻击鲁棒的图赌博机算法, 能够高效利用存在污染的奖赏数据, 取得接近对数级的遗憾界.

1.2 差分隐私

差分隐私由 Dwork 等人于 2006 年提出[20], 已经成为了机器学习和数据挖掘中被广泛使用的隐私保护模型[28−30]. 文献 [21] 最早研究随机赌博机中的隐私保护问题, 提出了两种满足差分隐私的多臂赌博机算法. 虽然这两种算法使用不同的技术 (UCB 和 TS) 来平衡探索和利用, 但都是通过向奖赏数据中添加噪声的方式保护隐私. 为了降低添加的噪声对奖赏期望估计造成的误差, 文献 [21] 构造了一个二叉树, 每个节点对应一个回合的噪声. 在计算前 t 回合的累积奖赏时, 只计入二叉树中从根到第 t 个叶子节点的路径中的噪声. 文献 [21] 证明了所提出的算法能够取得 $\mathrm{O}(K \ln^3 T/\epsilon)$ 的遗憾界.

文献 [22] 通过改进置信上界的构造方式和使用更加精细的分析技巧将文献 [21] 中的遗憾界提高到了 $\mathrm{O}(K \ln^{2.5} T/\epsilon)$. 然而, 该遗憾界对 T 的依赖仍然比无隐私保护的多臂赌博机算法的最优遗憾界 $\mathrm{O}(K \ln T)$ 多了 $\mathrm{O}(\ln^{1.5} T)$. 为了得到满足差分隐私且遗憾界最优的多臂赌博机算法, 文献 [23] 摒弃了文献 [21,22] 中使用的二叉树机制, 提出了一种基于稀疏向量机制[31]的算法, 并且证明了所提出算法取得了最优的 $\mathrm{O}(K \ln T/\epsilon)$ 遗憾界.

除了多臂赌博机外, 也有一些文献研究其他类型赌博机中的隐私保护问题[24,25]. 然而, 关于图赌博机的隐私保护问题从未被研究过, 本文提出的算法是首个满足差分隐私的图赌博机算法.

2 理论基础

在本节中, 我们给出图赌博机和差分隐私的形式化定义, 并介绍相关理论基础.

2.1 图赌博机

图赌博机可以看成是在学习者和环境之间重复进行的游戏. 记摇臂集 $[K] = \{1, \ldots, K\}$, 反馈图 $G = ([K], E)$. 对

于每个摇臂 $a\in[K]$, 我们用 $\mathcal{N}(a)$ 来表示该摇臂和其邻居所构成的集合:

$$\mathcal{N}(a)=\{a\}\cup\{a'\in[K]|(a,a')\in E\}.$$

记每个摇臂 $a\in[K]$ 的奖赏概率分布和期望奖赏分别为 $\mathbb{P}_a$ 和 $\mu_a=\mathbb{E}[r], r\sim\mathbb{P}_a$. 在每个回合 $t=1,\dots,T$, 首先学习者选择一个摇臂 $a_t\in[K]$. 然后, 环境从每个摇臂的奖赏概率分布中独立采样产生该回合所有摇臂的奖赏 $r_{t,1}\sim\mathbb{P}_1,\dots,r_{t,K}\sim\mathbb{P}_K$. 本文研究奖赏有界的情况. 不失一般性, 我们假设所有摇臂的奖赏都属于 $[0,1]$. 最后, 学习者观测到所选摇臂及其邻居的奖赏 $\{r_{t,a}|a\in\mathcal{N}(a_t)\}$, 并据此更新下一回合的摇臂选取策略. 学习者的目标是最大化所选摇臂的累积期望奖赏 $\sum_{t=1}^{T}\mu_{a_t}$. 记 a^* 为最优摇臂, 即期望奖赏最大的摇臂:

$$a^*=\arg\max_{a\in[K]}\mu_a.$$

我们用累积遗憾 (cumulative regret) 来衡量学习者的表现:

$$\mathrm{Reg}(T)=\sum_{t=1}^{T}(\mu_{a^*}-\mu_{a_t}).$$

可以看到, 最大化所选摇臂的累积期望奖赏等价于最小化累积遗憾.

图赌博机问题的难度与反馈图 G 的结构有关. 我们引入独立集和独立数来刻画图 G 的结构.

定义 1. 独立集[32]. 给定一个图 $G=(V,E)$. V 表示顶点集, E 表示边集. 对于图 G 的一个顶点子集 $\Omega\subseteq V$, 如果 Ω 中的任意两个顶点都不相邻, 即:

$$\forall u,v\in\Omega,\ (u,v)\notin E,$$

则称 Ω 为图 G 的独立集. 如果 Ω 还满足以下条件: Ω 之外的任意一个顶点都与 Ω 之内的某个顶点相邻, 则称 Ω 为图 G 的极大独立集.

定义 2. 独立数[32]. 我们将图 G 的所有独立集构成的集合记为 $\mathcal{I}(G)$. 图 G 的独立数 α 是 $\mathcal{I}(G)$ 中最大的独立集所包含的顶点数目:

$$\alpha=\max_{\Omega\in\mathcal{I}(G)}|\Omega|.$$

直观上来看, 反馈图 G 的独立数越小, 说明图中的顶点总体上的连接程度越高, 算法选择单个摇臂后能观测到的摇臂奖赏数据越多, 问题越容易.

2.2 差分隐私

差分隐私是密码学中的一项隐私保护技术, 旨在通过向算法的输出插入噪声来防止攻击者根据算法的输出逆向推算用户数据, 从而达到保护用户隐私的目的. 记第 t 回合所有摇臂的奖赏所构成的奖赏向量为 $r_t=[r_{t,1},\dots,r_{t,K}]$, 我们先定义奖赏序列的相似性.

定义 3. 相似奖赏序列[20]. 给定两个按照时间顺序构成的奖赏序列 $R=r_1,\dots,r_T$ 和 $R'=r'_1,\dots,r'_T$, 如果它们至多相差一个奖赏向量中的一项, 即存在 $s\in[T], k\in[K]$, 使得对于 $t=1,\dots,s-1,s+1,\dots,T$, 有 $r_t=r'_t$, 且对于 $i=1,\dots,k-1,k+1,\dots,K$, 有 $r_{s,i}=r'_{s,i}$, 则称 R 和 R'相似.

基于相似奖赏序列, 我们如下定义赌博机算法的差分隐私性质.

定义 4. 差分隐私[20]. 给定赌博机算法 $\mathcal{M}$, 用 $\mathcal{M}(R)$ 表示 $\mathcal{M}$ 作用于奖赏序列 R 上所输出的摇臂序列, 如果对于任意两个相似的奖赏序列 R 和 R'以及任意摇臂序列集 $S\subset[K]^T$, 有:

$$\Pr[\mathcal{M}(R)\in S]\leqslant\exp(\epsilon)\Pr[\mathcal{M}(R')\in S] \tag{1}$$

则称 $\mathcal{M}$ 满足 ϵ-差分隐私.

直观上来看, 公式 (1) 表明改变某个回合某个摇臂的奖赏对算法输出的摇臂序列的影响很小. 因此, 恶意攻击者难以根据算法输出的摇臂序列推算出摇臂的奖赏. 在很多现实场景如推荐系统中, 摇臂的奖赏 (如某个商品的点击情况或购买情况) 与用户隐私相关, 应避免被泄露. 满足 ϵ-差分隐私的赌博机算法能够有效地保护用户隐私. 参数 ϵ 刻画了隐私保护的程度, 值越小说明对隐私保护得越好.

3 算法描述

本文所提出的算法基于活跃摇臂消除算法 (active arm elimination, AAE)[33]. 为了便于理解, 在描述本文所提出的算法之前, 我们首先简要介绍 AAE 算法. AAE 算法的基本思想是维持一个活跃摇臂集 $\mathcal{A}\subseteq[K]$, 并重复迭代以下两步直到 T 回合结束.

(1) 依次选择 $\mathcal{A}$ 中的每个摇臂, 然后基于观测到的奖赏数据更新 $\mathcal{A}$ 中每个摇臂的经验平均奖赏.

(2) 找出 $\mathcal{A}$ 中经验平均奖赏值最大的摇臂作为标杆摇臂, 然后从 $\mathcal{A}$ 中删除经验平均奖赏显著小于标杆摇臂的经验平均奖赏的摇臂.

对于多臂赌博机问题, 理论分析可以证明以至少 $1-1/T$ 的概率, 经过 $\mathrm{O}(\ln T)$ 次迭代后活跃摇臂集 $\mathcal{A}$ 中只包含最优摇臂. 一方面, 由于每次迭代最多选择次优摇臂 $|\mathcal{A}|\leqslant K$ 次, 并且每次选取次优摇臂所造成的遗憾不超过 1. 另一方面, $\mathrm{O}(\ln T)$ 次迭代后只有最优摇臂会被选取. 因此, AAE 算法的期望遗憾不超过:

$$\mathrm{O}(K\ln T)(1-1/T)+\mathrm{O}(T)(1/T)=\mathrm{O}(K\ln T).$$

该遗憾界对于多臂赌博机问题是最优的.

然而, AAE 算法不满足差分隐私, 并且对于图赌博机问题是次优的. 针对图赌博机, 为了设计出满足差分隐私且遗憾界最优的算法, 我们对 AAE 算法做了 3 点创造性的改动. 首先, 为了保护隐私, 我们在计算摇臂的经验平均奖赏时引入拉普拉斯噪声, 使得攻击者难以根据摇臂的经验平均奖赏推算摇臂的单次奖赏. 其次, 为了最小化引入的拉普拉斯噪声对算法识别最优摇臂的影响, 我们将 T 回合分成若干阶段, 只在每个阶段的最后一个回合进行经验平均奖赏的计算操作和次优摇臂的删除操作. 最后, 为了得到对于图赌博机最优的遗憾界, 我们在每个阶段开始时构造一个由活跃摇臂集 $\mathcal{A}$ 所诱导的反馈子图 $G(\mathcal{A})$ 的独立集 Ω, 然后只选择 Ω 中的摇臂, 从而将次优摇臂的单次探索成本从 $\mathrm{O}(K)$ 降到了 $\mathrm{O}(\alpha)$.

我们将所提出的算法命名为 GAP (graph based arm elimination with differential privacy), 其伪代码如算法 1 所示.

算法 1. GAP 算法.

输入: 回合数 T, 隐私预算 ϵ, 置信度 $1-\delta$;

输出: 摇臂序列 $a_1,\ldots,a_T$.

1. $\tau=1,\ s_1=1,\ \mathcal{A}_1=[K],\ \bar{\mu}_{0,1}=\ldots=\bar{\mu}_{0,K}=0$;
2. **while** $t\leqslant T$ **do**
3. **for** 每个摇臂 $a\in\mathcal{A}_\tau$ **do**
4. $o_{\tau,a}=0,\ c_{\tau,a}=0$;
5. **end for**
6. $\Omega_\tau=GetIndSet(\mathcal{A}_\tau,\{\bar{\mu}_{\tau-1,a}|a\in\mathcal{A}_\tau\})$;
7. **for** 每个摇臂 $a\in\Omega_\tau$ **do**
8. $n_{\tau,a}=0$;
9. **end for**
10. $L_{\tau,1}=2^{5+2\tau}\ln(8|\mathcal{A}_\tau|\tau^2/\delta)$;
11. $L_{\tau,2}=2^{3+\tau}\epsilon^{-1}\ln(4|\mathcal{A}_\tau|\tau^2/\delta)$;
12. $L_\tau=\lceil\max\{L_{\tau,1},L_{\tau,2}\}\rceil,\ s_{\tau+1}=s_\tau+L_\tau|\Omega_\tau|$;
13. **for** $t=s_\tau,s_\tau+1,\ldots,\min\{T,s_{\tau+1}-1\}$ **do**
14. 选择摇臂 $a_t=\arg\min_{k\in\Omega_\tau}n_{\tau,k}$;
15. $n_{\tau,a_t}=n_{\tau,a_t}+1$, 观测奖赏 $\{r_{t,a}|a\in\mathcal{N}(a_t)\}$
16. **for** 每个摇臂 $a\in\mathcal{N}(a_t)\cap\mathcal{A}_\tau$ **do**

17. $o_{\tau,a}=o_{\tau,a}+1,\ c_{\tau,a}=c_{\tau,a}+r_{\tau,a}$;
18. **end for**
19. **end for**
20. **for** 每个摇臂 $a\in\mathcal{A}_\tau$ **do**
21. $\bar{\mu}_{\tau,a}=c_{\tau,a}/o_{\tau,a}+\eta_a,\ \eta_a\sim\text{Laplace}(0,1/(\epsilon L_\tau))$
22. **end for**
23. $\mathcal{A}_{\tau+1}=\mathcal{A}_\tau-\{a|\bar{\mu}_{\tau,a}<\max_{i\in\mathcal{A}_\tau}\bar{\mu}_{\tau,i}-\sqrt{2\ln(8|\mathcal{A}_\tau|\tau^2/\delta)/L_\tau}-2\ln(4|\mathcal{A}_\tau|\tau^2/\delta)/(\epsilon L_\tau)\}$;
24. $\tau=\tau+1$;
25. **end while**
26. **function** *GetIndSet*$(\mathcal{A},\{\bar{\mu}_a|a\in\mathcal{A}\})$
27. $\Omega=\varnothing$;
28. **while** $\mathcal{A}\neq\varnothing$ **do**
29. $a'=\arg\max_{a\in\mathcal{A}}\bar{\mu}_a$;
30. $\Omega=\Omega\cup\{a'\},\ \mathcal{A}=\mathcal{A}-\mathcal{N}(a')$;
31. **end while**
32. **return** Ω;

根据算法 1, 步骤 1 初始化阶段序号 $\tau=1$, 第 1 阶段的起始回合 s_1, 第 1 阶段的活跃摇臂集 $\mathcal{A}_1$, 每个摇臂的经验平均奖赏 $\bar{\mu}_{0,1}=\ldots=\bar{\mu}_{0,K}$. 在每个阶段 τ (步骤 2–25), 首先对活跃摇臂集 $\mathcal{A}_\tau$ 中每个摇臂 a 初始化其奖赏被观测到的次数 $o_{\tau,a}$ 和累积奖赏 $c_{\tau,a}$ (步骤 3–5). 然后, 通过调用函数 GetIndSet 来获取一个由活跃摇臂集 $\mathcal{A}_\tau$ 所诱导的反馈子图 $G(\mathcal{A}_\tau)$ 的独立集 Ω_τ (步骤 6), 初始化 Ω_τ 中每个摇臂被选择的次数 $n_{\tau,a}$ (步骤 7–9). 接着, 步骤 10–12 计算 $\mathcal{A}_\tau$ 中每个摇臂的奖赏所需要被观测到的最少次数 L_τ, 并据此确定下一阶段的起始回合 $s_{\tau+1}$. 之后, 在第 τ 阶段的每个回合 (步骤 13–19), 依次执行以下操作: 从独立集 Ω_τ 中选择到目前为止被选择次数最少的摇臂 (步骤 14), 观测所选摇臂及其邻居的奖赏 (步骤 15), 对这些摇臂更新其奖赏被观测到的次数 $o_{\tau,a}$ 和累积奖赏 $c_{\tau,a}$ (步骤 16–18). 最后, 步骤 20–22 根据第 τ 阶段内 $\mathcal{A}_\tau$ 中每个摇臂被观测到的奖赏数据和随机产生的拉普拉斯噪声来计算每个摇臂的经验平均奖赏 $\bar{\mu}_{\tau,a}$. 步骤 23 从 $\mathcal{A}_\tau$ 中删除经验平均奖赏显著低于最大经验平均奖赏的摇臂来得到下一阶段的活跃摇臂集 $\mathcal{A}_{\tau+1}$. 步骤 24 更新阶段序号.

算法 1 通过调用函数 GetIndSet 来得到独立集. 该函数接受一个摇臂集 $\mathcal{A}$ 以及 $\mathcal{A}$ 中每个摇臂的经验平均奖赏 $\{\bar{\mu}_a|a\in\mathcal{A}\}$ 作为输入, 输出一个 $\mathcal{A}$ 所诱导的反馈子图 $G(\mathcal{A})$ 的极大独立集 Ω. 虽然构造一个反馈子图的极大独立集是一个平凡问题[13], 但是函数 GetIndSet 的新颖之处在于其构造极大独立集时贪心地选择经验平均奖赏最大的摇臂. 具体来说, 首先初始化独立集 Ω 为空集 (步骤 27). 然后, 重复步骤 29–30 直到摇臂集 $\mathcal{A}$ 变成空集: 选择 $\mathcal{A}$ 中经验平均奖赏最大的摇臂, 将该摇臂加入 Ω, 从 $\mathcal{A}$ 中删除该摇臂及其邻居. 最后, 步骤 32 返回最终生成的极大独立集 Ω. 注意到函数 GetIndSet 在每个阶段只被调用一次, 每次调用的时间复杂度为 $\text{O}(\alpha)$. 因为每个阶段的长度相比于前一阶段都至少增加一倍, 所以 T 回合内总的阶段数为 $\text{O}(\ln T)$. 因此, 引入函数 GetIndSet 所带来的额外时间成本为 $\text{O}(\alpha\ln T)$.

4 理论分析

我们首先证明 GAP 算法的差分隐私性质. 为此, 引入以下定义和引理.

定义 5. 全局敏感度[34]. 给定一个以序列 R 为输入、值域为 $\mathbb{R}^K$ 的函数 f, 我们定义 f 的全局敏感度 $\phi(f)$ 为:

$$\varphi(f)=\sup_{R\sim R'}\|f(R)-f(R')\|_1,$$

其中, $R\sim R'$ 表示序列 R 与序列 R' 相似.

引理 1. 拉普拉斯机制[34]. 给定一个全局敏感度不超过 B 的函数 f, 通过给 f 添加服从拉普拉斯分布的噪声

$\eta_1 \sim \text{Laplace}(0, B/\epsilon), \ldots, \eta_K \sim \text{Laplace}(0, B/\epsilon)$ 得到的新函数:

$$f'(R) = f(R) + [\eta_1, \ldots, \eta_K],$$

满足 ϵ -差分隐私.

由引理 1, 我们可以证明以下定理.

定理 1. GAP 算法满足 ϵ -差分隐私.

证明: 考虑两个相似的奖赏序列 R 和 R', 它们的第 t 个奖赏向量的第 k 维分量不同, 即 $r_{t,k} \neq r'_{t,k}$. 设回合 t 所在的阶段为 τ, 即 $s_\tau \leqslant t < s_{\tau+1}$. 若 τ 是最后一个阶段, 由于 GAP 算法在每个阶段的摇臂选取策略由前一阶段中的最后一个回合所决定, 所以 $r_{t,k} \neq r'_{t,k}$ 不影响 GAP 算法的运行. 否则, 定义函数:

$$u(R) = \left[\frac{c_{\tau,1}}{o_{\tau,1}}, \ldots, \frac{c_{\tau,K}}{o_{\tau,K}}\right].$$

因为摇臂奖赏在 $[0,1]$ 区间内, 所以有:

$$u(R) - u(R') \leqslant \frac{|r_{t,k} - r'_{t,k}|}{o_{\tau,k}} \leqslant \frac{1}{o_{\tau,k}}.$$

接下来, 我们分析 $o_{\tau,k}$ 的下界. 根据 GAP 算法的步骤 14, 每隔 $|\Omega_\tau|$ 回合, Ω_τ 中的每个摇臂恰好被选择一次. 因为 Ω_τ 是 $\mathcal{A}_\tau$ 所诱导的反馈子图 $G(\mathcal{A}_\tau)$ 的极大独立集, 所以每隔 $|\Omega_\tau|$ 回合, 对于 Ω_τ 中的每个摇臂, 其奖赏被观测至少一次. 进而, 由 GAP 算法的步骤 12 可得 $o_{\tau,k} \geqslant (s_{\tau+1} - s_\tau)/|\Omega_\tau| = L_\tau$. 因此, 函数 u 的全局敏感度满足 $\phi(u) \leqslant 1/L_\tau$. 注意到 GAP 算法在步骤 21 给函数 u 的每个维度都添加了拉普拉斯噪声 $\eta_a \sim \text{Laplace}(0, 1/(\epsilon L_\tau))$, 由引理 1 立即可得 GAP 算法满足 ϵ -差分隐私. 证毕.

接着, 我们分析 GAP 算法的累积遗憾, 首先引入以下两个聚集不等式 (concentration inequality).

引理 2[35]. 设 $x_1, \ldots, x_n$ 为 n 个独立同分布的支撑集 (support set) 为 $[0,1]$ 的随机变量, 记其共同期望为 μ, 对任意常数 $\rho > 0$, 有:

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right| \geqslant \rho\right] \leqslant 2\exp(-2n\rho^2).$$

引理 3[34]. 设 x 服从参数为 $(0, \lambda)$ 的拉普拉斯分布, 对任意常数 $\rho > 0$, 有:

$$\Pr[|x| \geqslant \rho] = \exp(-\rho/\lambda).$$

根据引理 2 和引理 3, 我们可以分析 GAP 算法在步骤 21 所计算的经验平均奖赏 $\bar{\mu}_{\tau,a}$ 与真实期望奖赏 μ_a 之间的偏离情况.

引理 4. 以至少 $1-\delta$ 的概率, 对除了最后一个阶段之外的任意阶段 τ 和任意摇臂 $a \in \mathcal{A}_\tau$, 下式成立:

$$|\bar{\mu}_{\tau,a} - \mu_a| < \sqrt{\frac{\ln(8|\mathcal{A}_\tau|\tau^2/\delta)}{2L_\tau}} + \frac{\ln(4|\mathcal{A}_\tau|\tau^2/\delta)}{\epsilon L_\tau}.$$

证明: 对除了最后一个阶段之外的任意阶段 τ 和任意摇臂 $a \in \mathcal{A}_\tau$, 根据引理 2, 有:

$$\Pr\left[\left|\frac{c_{\tau,a}}{o_{\tau,a}} - \mu_a\right| \geqslant \sqrt{\frac{\ln(8|\mathcal{A}_\tau|\tau^2/\delta)}{2L_\tau}}\right] \leqslant 2\exp\left(-2o_{\tau,a}\cdot\frac{\ln(8|\mathcal{A}_\tau|\tau^2/\delta)}{2L_\tau}\right) \leqslant \frac{\delta}{4|\mathcal{A}_\tau|\tau^2} \tag{2}$$

其中, 第 2 个不等式用到了 $o_{\tau,a} \geqslant L_\tau$.

另一方面, 由引理 3 可得:

$$\Pr\left[|\eta_a| \geqslant \frac{\ln(4|\mathcal{A}_\tau|\tau^2/\delta)}{\epsilon L_\tau}\right] \leqslant \frac{\delta}{4|\mathcal{A}_\tau|\tau^2} \tag{3}$$

根据式 (2)、式 (3) 和 GAP 算法的步骤 21, 有:

$$\Pr\left[|\bar{\mu}_{\tau,a} - \mu_a| \geqslant \sqrt{\frac{\ln(8|\mathcal{A}_\tau|\tau^2/\delta)}{2L_\tau}} + \frac{\ln(4|\mathcal{A}_\tau|\tau^2/\delta)}{\epsilon L_\tau}\right] \leqslant \frac{\delta}{2|\mathcal{A}_\tau|\tau^2}.$$

最后, 对 $a \in \mathcal{A}_\tau$ 和 $\tau = 1, 2, \ldots$ 应用联合界 (union bound), 并注意到:

$$\sum_{\tau=1}^{+\infty}\frac{1}{\tau^2}=\frac{\pi^2}{6}\leqslant 2,$$

即可证明引理 4. 证毕.

接下来, 我们给出并证明 GAP 算法的大概率遗憾界 (high probability regret bound).

定理 2. 以至少 $1-\delta$ 的概率, GAP 算法的累积遗憾满足:

$$\mathrm{Reg}(T)\leqslant\left(\frac{176\alpha}{\Delta_{\min}^2}+\frac{32\alpha}{\epsilon\Delta_{\min}}\right)\ln\left(\frac{8K\lceil\log_2\Delta_{\min}^{-1}\rceil^2}{\delta}\right)\tag{4}$$

证明: 我们首先通过数学归纳法证明最优摇臂 a^* 始终在活跃摇臂集 $\mathcal{A}_\tau$ 中. 对于第一个阶段 $\tau=1$, $a^*\in\mathcal{A}_1=[K]$ 是平凡的结论. 假设 $a^*\in\mathcal{A}_\tau$, 下面我们证明 $a^*\in\mathcal{A}_{\tau+1}$. 用 $\bar{a}^*$ 表示阶段 τ 中经验平均奖赏最大的摇臂. 根据引理 4, 有:

$$\overline{\mu}_{\tau,\bar{a}^*}-\overline{\mu}_{\tau,a^*}\leqslant\overline{\mu}_{\tau,\bar{a}^*}-\mu_{\bar{a}^*}+\mu_{a^*}-\overline{\mu}_{\tau,a^*}\leqslant 2\left(\sqrt{\frac{\ln(8|\mathcal{A}_\tau|\tau^2/\delta)}{2L_\tau}}+\frac{\ln(4|\mathcal{A}_\tau|\tau^2/\delta)}{\epsilon L_\tau}\right).$$

因此, 由 GAP 算法的步骤 23 可得 $a^*\in\mathcal{A}_{\tau+1}$.

下面, 我们证明经过有限个阶段后, 所有次优摇臂 (即满足 $\mu_a<\mu_{a^*}$ 的摇臂) 都被从活跃摇臂集 $\mathcal{A}_\tau$ 中移除. 我们考察 GAP 算法在第 τ 阶段执行步骤 23 的情况. 对任意次优摇臂 $a\in\mathcal{A}_\tau-\{a'|\mu_{a'}=\mu_{a^*}\}$, 记其次优间隔为 $\Delta_a=\mu_{a^*}-\mu_a$, 根据引理 4 和 GAP 算法的步骤 10–12, 有:

$$\overline{\mu}_{\tau,a^*}-\overline{\mu}_{\tau,a}=\overline{\mu}_{\tau,a^*}-\mu_{a^*}+\mu_a-\overline{\mu}_{\tau,a}+\Delta_a>\Delta_a-2\left(\sqrt{\frac{\ln(8|\mathcal{A}_\tau|\tau^2/\delta)}{2L_\tau}}+\frac{\ln(4|\mathcal{A}_\tau|\tau^2/\delta)}{\epsilon L_\tau}\right)\geqslant\Delta_a-2^{-(1+\tau)}\geqslant\Delta_{\min}-2^{-(1+\tau)}.$$

令 $\tau_*=\lceil\log_2\Delta_{\min}^{-1}\rceil$, 则有:

$$\overline{\mu}_{\tau_*,a^*}-\overline{\mu}_{\tau_*,a}>2^{-\tau_*}-2^{-(1+\tau_*)}=2^{-(1+\tau_*)}\geqslant\sqrt{\frac{2\ln(8|\mathcal{A}_{\tau_*}|\tau_*^2/\delta)}{2L_{\tau_*}}}+\frac{2\ln(4|\mathcal{A}_{\tau_*}|\tau_*^2/\delta)}{\epsilon L_{\tau_*}}.$$

根据 GAP 算法的步骤 23 可得 $a\notin\mathcal{A}_{\tau_*+1}$, 即经过 τ_* 个阶段后, 活跃摇臂集 $\mathcal{A}_\tau$ 只包含最优摇臂. 注意到由于摇臂奖赏在 $[0,1]$ 区间内, 每个回合的遗憾不超过 1, 因此 GAP 算法的遗憾满足:

$$\mathrm{Reg}(T)\leqslant\sum_{\tau=1}^{\tau_*}L_\tau|\Omega_\tau|\leqslant\alpha\sum_{\tau=1}^{\tau_*}L_\tau\leqslant\alpha\ln\left(\frac{8K\tau_*^2}{\delta}\right)\sum_{\tau=1}^{\tau_*}\left(2^{5+2\tau}+2^{3+\tau}\epsilon^{-1}\right)\leqslant\left(\frac{176\alpha}{\Delta_{\min}^2}+\frac{32\alpha}{\epsilon\Delta_{\min}}\right)\ln\left(\frac{8K\lceil\log_2\Delta_{\min}^{-1}\rceil^2}{\delta}\right).$$

证毕.

最后, 由定理 2 可以直接推出 GAP 算法的期望遗憾界 (expected regret bound):

定理 3. GAP 算法的累积遗憾的期望满足:

$$\mathbb{E}[\mathrm{Reg}(T)]\leqslant\Delta_{\min}^{-1}\left(\frac{176}{\Delta_{\min}}+\frac{32}{\epsilon}\right)\alpha\ln T+\left(\frac{176\alpha}{\Delta_{\min}^2}+\frac{32\alpha}{\epsilon\Delta_{\min}}\right)\ln\left(8K\lceil\log_2\Delta_{\min}^{-1}\rceil^2\right)+1.$$

证明: 在公式 (4) 中取 $\delta=1/T$, 根据全期望公式有:

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}(T)]=&\mathbb{E}[\mathrm{Reg}(T)|\text{式 (4)成立}]\cdot\Pr[\text{式 (4)成立}]+\mathbb{E}[\mathrm{Reg}(T)|\text{式 (4)不成立}]\cdot\Pr[\text{式 (4)不成立}]\\ \leqslant&\mathbb{E}[\mathrm{Reg}(T)|\text{式 (4)成立}]+T\cdot\Pr[\text{式 (4)不成立}]\leqslant\left(\frac{176\alpha}{\Delta_{\min}^2}+\frac{32\alpha}{\epsilon\Delta_{\min}}\right)\ln\left(\frac{8K\lceil\log_2\Delta_{\min}^{-1}\rceil^2}{1/T}\right)+1.\end{aligned}$$

证毕.

根据定理 3, GAP 算法的期望遗憾界为 $\mathrm{O}((\alpha/\Delta_{\min}+\alpha/\epsilon)\ln T)$. 该遗憾界依赖反馈图的独立数 α、最小次优间隔 $\Delta_{\min}$ 和隐私预算 ϵ. 对于图赌博机问题, 反馈图的独立数越小、最小次优间隔越大, 问题越容易. 这是因为反馈图的独立数越小说明每回合能够观测到的摇臂奖赏数据越多, 最小次优间隔越大越容易根据经验平均奖赏将最优摇臂与次优摇臂分割开. 另一方面, ϵ 越小说明隐私保护得越好, 相应地所需要付出的代价也越大. 可以看到, GAP 算法的遗憾界很好地反映了差分隐私下的图赌博机问题的难易程度与反馈图的独立数、最小次优间隔和隐私预

算之间的关系. 最后, 我们指出 GAP 算法的遗憾界与图赌博机问题的理论下界[13]和差分隐私下的赌博机问题的理论下界[24]相匹配, 因此是最优的.

5 实 验

在本节, 我们测试 GAP 算法的隐私保护效果和累积遗憾. 我们使用仿真数据集, 其构造方式如下. 首先, 设定摇臂数 $K=10$. 然后, 使用 Erdos-Renyi 模型[36]来生成反馈图 G. 具体地, 对任意两个摇臂 $u\neq v$, Erdos-Renyi 模型以概率 p 连接 u 和 v, 其中 p 是实验参数. 直观上来看, 随着 p 增大反馈图 G 变得稠密, 相应地图 G 的独立数 α 变小. 最后, 分两步生成每个摇臂的奖赏数据. 第 1 步确定每个摇臂的期望奖赏. 不失一般性, 我们取前两个摇臂为最优摇臂, 后 8 个摇臂为次优摇臂. 最优摇臂的期望奖赏被设置为 $\mu_1=\mu_2=0.9$. 对于次优摇臂 $k=3,\ldots,10$, 其期望奖赏为 $\mu_k=0.9-\Delta_{\min}-0.05(k-3)$. 其中, $\Delta_{\min}$ 为实验参数, 表示最小次优间隔. 第 2 步根据摇臂的期望奖赏随机生成奖赏数据. 对于期望奖赏为 μ 的摇臂, 其奖赏数据采样自均值为 μ、方差为 0.01、支撑集为 $[0,1]$ 的截断正态分布 (truncated normal distribution).

5.1 隐私保护

本小节探究 GAP 算法的隐私保护效果. 设定回合数 $T=100\,000$, 置信度 $1-\delta=1-1/T$. 首先, 随机生成 T 回合的奖赏向量序列 $\mathbf{r}_1,\ldots,\mathbf{r}_T$. 然后, 在该奖赏向量序列上运行 GAP 算法, 记录 GAP 算法输出的摇臂序列 $a_1,\ldots,a_T$. 接着, 随机选取某个回合的某个摇臂的奖赏数据, 将其改为 0 后再次运行 GAP 算法, 记录 GAP 算法输出的新的摇臂序列 $a'_1,\ldots,a'_T$. 最后, 比对 GAP 算法先后输出的摇臂序列是否完全相同 $a_1=a'_1,\ldots,a_T=a'_T$. 重复以上步骤 100 次, 计算 GAP 算法先后输出的摇臂序列相同的次数所占的比例. 比例越大说明越难通过 GAP 算法输出的摇臂序列逆向推算出摇臂的奖赏数据, 即隐私保护得越好. 表 1 列出了不同实验参数 (最小次优间隔 $\Delta_{\min}$、连接概率 p 和隐私预算 ϵ) 下的实验结果.

表 1 GAP 算法在不同实验参数下的隐私保护效果

$\Delta_{\min}$	p	ϵ	Ratio of same-sequence (%)	$\Delta_{\min}$	p	ϵ	Ratio of same-sequence (%)	$\Delta_{\min}$	p	ϵ	Ratio of same-sequence (%)
0.05	0.1	0.05	94	0.1	0.1	0.05	97	0.2	0.1	0.05	98
0.05	0.1	0.1	91	0.1	0.1	0.1	95	0.2	0.1	0.1	96
0.05	0.1	0.2	89	0.1	0.1	0.2	90	0.2	0.1	0.2	92
0.05	0.2	0.05	93	0.1	0.2	0.05	94	0.2	0.2	0.05	94
0.05	0.2	0.1	90	0.1	0.2	0.1	91	0.2	0.2	0.1	91
0.05	0.2	0.2	88	0.1	0.2	0.2	88	0.2	0.2	0.2	91
0.05	0.3	0.05	91	0.1	0.3	0.05	93	0.2	0.3	0.05	92
0.05	0.3	0.1	89	0.1	0.3	0.1	92	0.2	0.3	0.1	90
0.05	0.3	0.2	86	0.1	0.3	0.2	89	0.2	0.3	0.2	90

从表 1 可以看出, 在所有实验参数下, GAP 算法先后输出相同摇臂序列的占比都超过了 85%, 表明 GAP 算法可以很好地保护隐私. 通过对比不同试验参数下的实验结果可以发现, 总体上最小次优间隔 $\Delta_{\min}$ 越大, GAP 算法先后输出相同摇臂序列的可能性越大. 这是因为 $\Delta_{\min}$ 越大, GAP 算法越能更早地识别出最优摇臂并把次优摇臂移除活跃摇臂集 $\mathcal{A}_\tau$. 根据 GAP 算法的运行逻辑, 当 $\mathcal{A}_\tau$ 只包含最优摇臂时, GAP 算法会始终选择最优摇臂, 不受奖赏数据影响. 连接概率 p 越小, 反馈图的独立数越大, 平均意义上 GAP 算法每回合观测到的奖赏数据越少, 因此改变某个回合的某个摇臂的奖赏数据影响到 GAP 算法的可能性也越小. 最后, 隐私预算 ϵ 越小, GAP 算法在计算经验平均奖赏时引入的拉普拉斯噪声越大, 越不容易受到摇臂奖赏数据变化的影响, 从而越有可能先后输出相同的摇臂序列.

5.2 累计遗憾

本小节测试 GAP 算法的累积遗憾. 为了对比, 我们也同时测试了不满足差分隐私的图赌博机算法 AlphaSample[13]、满足差分隐私的多臂赌博机算法 DPSE[23]和 GAP 算法所基于的活跃摇臂消除算法 AAE[33]. 为了展示 GAP 算法中

使用的用来构造独立集的函数 GetIndSet 的优越性, 我们也测试了不使用 GetIndSet 的 GAP 算法变种 GAPU. 在 GAPU 中, 每个阶段的独立集 Ω_τ 从活跃摇臂集 $\mathcal{A}_\tau$ 所诱导的反馈子图 $G(\mathcal{A}_\tau)$ 的所有独立集中均匀采样得到.

与第 5.1 节相同, 设定回合数 $T=100\,000$, 置信度 $1-\delta=1-1/T$. 分别固定最小次优间隔 $\Delta_{\min}$、连接概率 p 和隐私预算 ϵ 这 3 个实验参数中的两个, 变化另外一个. 图 1–图 3 给出了不同算法在各种实验参数下的遗憾增长曲线. 从中可以看出, GAP 算法的遗憾始终显著小于 DPSE 和 AAE. 这是因为 DPSE 和 AAE 都是为多臂赌博机问题设计的, 无法利用图赌博机问题中额外的反馈信息加速识别最优摇臂. 与之相反, GAP 算法通过构造独立集并只选择独立集中的摇臂的方式降低了探索活跃摇臂集所需要的回合数, 进而能更早地识别最优摇臂. 此外, GAPU 算法的遗憾在几乎所有实验参数下均明显大于 GAP 算法. 虽然 GAPU 算法识别最优摇臂所需要的回合数与 GAP 算法几乎一样, 但是由于使用的是均匀采样的独立集, 在识别最优摇臂前选择次优摇臂所造成的遗憾较大. 与此不同, GAP 算法通过精心设计的 GetIndSet 函数构造独立集, 贪心地选择经验平均奖赏最高的摇臂来完成探索, 降低了总的探索成本. 最后, 我们比较 GAP 算法与不满足差分隐私的图赌博机算法 AlphaSample. 虽然 GAP 算法的遗憾大于 AlphaSample, 但是两者之间的差距很小. 这说明 GAP 算法能够在保护隐私的同时取得与现有无隐私保护算法相当的累积遗憾, 具有很高的可用性.

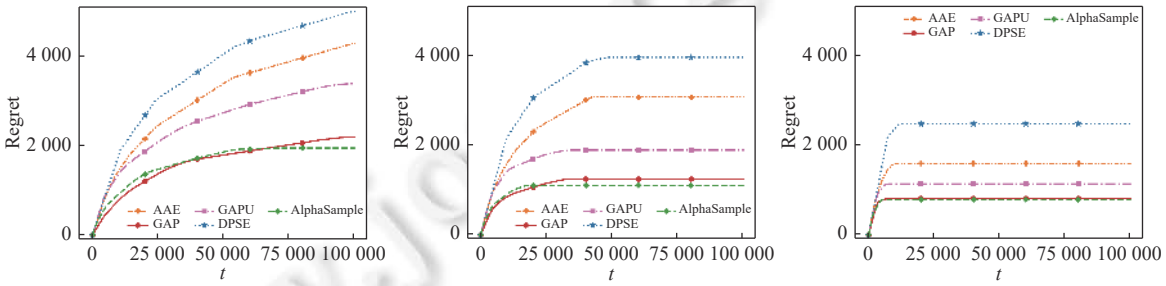

图 1 固定 $\epsilon=0.05, p=0.2$, 变化 $\Delta_{\min}$, 比较 4 种算法的性能

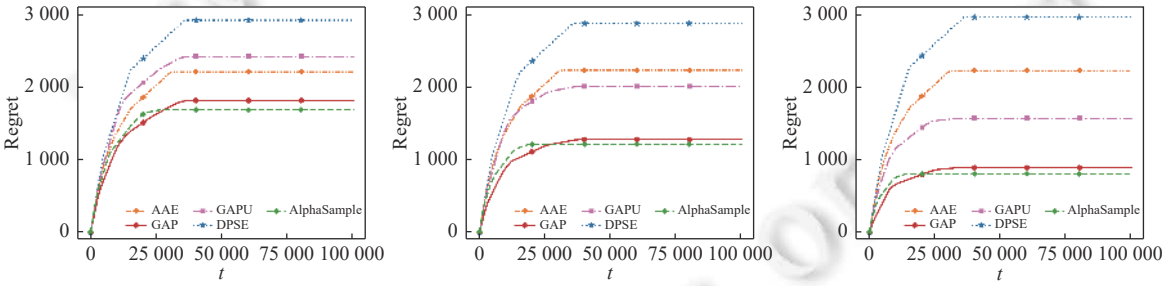

图 2 固定 $\Delta_{\min}=0.1$, $\epsilon=0.1$, 变化 p, 比较 4 种算法的性能

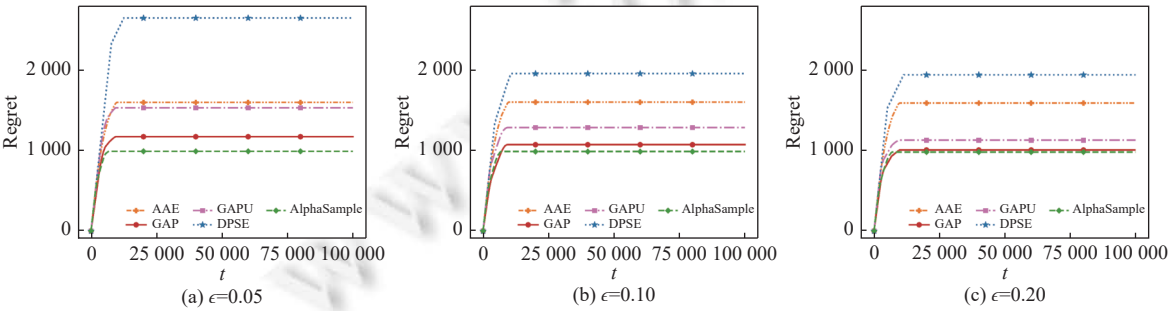

图 3 固定 $\Delta_{\min}=0.2, p=0.1$, 变化 ϵ, 比较 4 种算法的性能

通过观察 GAP 算法在不同实验参数下的累积遗憾可以发现, GAP 算法的累计遗憾及其与 AlphaSample 算法之间的差距都随着最小次优间隔 $\Delta_{\min}$、连接概率 p 和隐私预算 ϵ 的增大而减小. 这与定理 2 和定理 3 中给出的遗憾界相符. 结合前一小节中所观测到的 GAP 算法的隐私保护效果随隐私预算 ϵ 的变化情况, 我们发现隐私预算 ϵ 的设置对 GAP 算法的影响具有两面性. 一方面, 减小隐私预算 ϵ 可以更好的保护隐私. 另一方面, 减小隐私预算 ϵ 也会增加累计遗憾. 因此在实际应用中, 需要根据具体问题设定隐私预算 ϵ 的取值, 权衡好隐私保护和累计遗憾.

6 总 结

本文提出了一种满足差分隐私性质的图赌博机算法 GAP. 一方面, GAP 算法通过对回合进行阶段分割和添加拉普拉斯噪声的方式确保恶意攻击者难以根据算法的输出逆向推算摇臂的奖赏数据, 从而保护了隐私. 另一方面, 为了有效利用图形式的反馈信息, GAP 算法使用了精心设计的独立集构造函数, 贪心地选取经验平均奖赏最高的摇臂来探索活跃摇臂集, 从而降低了探索成本和累计遗憾. 我们证明了 GAP 算法能够在满足 ϵ -差分隐私的同时取得 $\mathrm{O}(\alpha \ln T/\epsilon)$ 的遗憾界. 我们也在仿真数据集上测试了 GAP 算法的隐私保护效果和累积遗憾. 实验结果显示: GAP 算法输出的摇臂序列受单个摇臂奖赏变化的影响很小, 表明 GAP 算法能够很好保护隐私, 并且 GAP 算法的累积遗憾与现有无隐私保护的图赌博机算法相当, 具有很高的可用性.

除图赌博机外, 另一种建模摇臂之间相似关系的赌博机模型是利普希茨赌博机[37]. 这两种赌博机模型分别适用于不同的问题场景, 用到的算法技术也大不相同. 未来工作尝试研究利普希茨赌博机中的隐私保护问题, 探索满足差分隐私的利普希茨赌博机算法.

References:

[1] Thompson WR. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 1933, 25(3–4): 285–294. [doi: 10.2307/2332286]

[2] Wang LM, Huang HK, Chai YM. Choosing multi-issue negotiating object based on trust and K-armed bandit problem. Ruan Jian Xue Bao/Journal of Software, 2006, 17(12): 2537–2546 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/2537.htm [doi: 10.1360/jos172537]

[3] Li LH, Chu W, Langford J, Schapire RE. A contextual-bandit approach to personalized news article recommendation. In: Proc. of the 19th Int'l Conf. on World Wide Web. Raleigh: ACM, 2010. 661–670. [doi: 10.1145/1772690.1772758]

[4] Bnaya Z, Puzis R, Stern R, Felner A. Bandit algorithms for social network queries. In: Proc. of the 2013 Int'l Conf. on Social Computing. Alexandria: IEEE, 2013. 148–153. [doi: 10.1109/SocialCom.2013.29]

[5] Chen W, Wang YJ, Yuan Y. Combinatorial multi-armed bandit: General framework, results and applications. In: Proc. of the 30th Int'l Conf. on Machine Learning. Atlanta: JMLR.org, 2013. 151–159.

[6] Lai TL, Robbins H. Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 1985, 6(1): 4–22. [doi: 10.1016/0196-8858(85)90002-8]

[7] Auer P, Cesa-Bianchi N, Fischer P. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 2002, 47(2–3): 235–256. [doi: 10.1023/A:1013689704352]

[8] Mannor S, Shamir O. From bandits to experts: On the value of side-observations. In: Proc. of the 24th Int'l Conf. on Neural Information Processing Systems. Granada: Curran Associates Inc., 2011. 684–692.

[9] Alon N, Cesa-Bianchi N, Gentile C, Mannor S, Mansour Y, Shamir O. Nonstochastic multi-armed bandits with graph-structured feedback. SIAM Journal on Computing, 2017, 46(6): 1785–1826. [doi: 10.1137/140989455]

[10] Caron S, Kveton B, Lelarge M, Bhagat S. Leveraging side observations in stochastic bandits. In: Proc. of the 28th Conf. on Uncertainty in Artificial Intelligence. Catalina Island: AUAI Press, 2012. 142–151.

[11] Hu BS, Mehta NA, Pan JP. Problem-dependent regret bounds for online learning with feedback graphs. In: Proc. of the 36th Conf. on Uncertainty in Artificial Intelligence. Tel Aviv: UAI Press, 2019. 852–861.

[12] Buccapatnam S, Eryilmaz A, Shroff NB. Stochastic bandits with side observations on networks. In: Proc. of the 2014 ACM Int'l Conf. on Measurement and Modeling of Computer Systems. Austin: ACM, 2014. 289–300. [doi: 10.1145/2591971.2591989]

[13] Cohen A, Hazan T, Koren T. Online learning with feedback graphs without the graphs. In: Proc. of the 33rd Int'l Conf. on Machine

Learning. New York: JMLR.org, 2016. 811–819.

[14] Tossou ACY, Dimitrakakis C, Dubhashi D. Thompson sampling for stochastic bandits with graph feedback. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. San Francisco: AAAI Press, 2017. 2660–2666.

[15] Liu F, Buccapatnam S, Shroff N. Information directed sampling for stochastic bandits with graph feedback. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2018. 3643–3650.

[16] Liu F, Zheng ZZ, Shroff NB. Analysis of thompson sampling for graphical bandits without the graphs. In: Proc. of the 34th Conf. on Uncertainty in Artificial Intelligence. Monterey: AUAI Press, 2018. 13–22.

[17] Lykouris T, Tardos É, Wali D. Feedback graph regret bounds for Thompson Sampling and UCB. In: Proc. of the 31st Int'l Conf. on Algorithmic Learning Theory. San Diego: PMLR, 2020. 592–614.

[18] Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. In: Proc. of the 2008 IEEE Symp. on Security and Privacy. Oakland: IEEE, 2008. 111–125. [doi: 10.1109/SP.2008.33]

[19] Ye QQ, Meng XF, Zhu MJ, Huo Z. Survey on local differential privacy. Ruan Jian Xue Bao/Journal of Software, 2018, 29(7): 1981–2005 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5364.htm [doi: 10.13328/j.cnki.jos.005364]

[20] Dwork C. Differential privacy. In: Proc. of the 33rd Int'l Colloquium on Automata, Languages, and Programming. Venice: Springer, 2006. 1–12. [doi: 10.1007/11787006_1]

[21] Smith AD, Thakurta AG. (Nearly) optimal algorithms for private online learning in full-information and bandit settings. In: Proc. of the 27th Advances in Neural Information Processing Systems. Lake Tahoe, 2013. 2733–2741.

[22] Tossou ACY, Dimitrakakis C. Algorithms for differentially private multi-armed bandits. In: Proc. of the 30th AAAI Conf. on Artificial Intelligence. Phoenix: AAAI Press, 2016. 2087–2093.

[23] Sajed T, Sheffet O. An optimal private stochastic-MAB algorithm based on optimal private stopping rule. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 5579–5588.

[24] Shariff R, Sheffet O. Differentially private contextual linear bandits. In: Proc. of the Advances in Neural Information Processing Systems 2018. Montréal, 2018. 4301–4311.

[25] Tossou ACY, Dimitrakakis C. Achieving privacy in the adversarial multi-armed bandit. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. San Francisco: AAAI Press, 2017. 2653–2659.

[26] Lu SY, Hu Y, Zhang LJ. Stochastic bandits with graph feedback in non-stationary environments. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI Press, 2021. 8758–8766.

[27] Lu SY, Wang GH, Zhang LJ. Stochastic graphical bandits with adversarial corruptions. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2021. 8749–8757.

[28] Zhang XJ, Wang M, Meng XF. An accurate method for mining top-*k* frequent pattern under differential privacy. Journal of Computer Research and Development, 2014, 51(1): 104–114 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2014.20130685]

[29] Wang JY, Liu C, Fu XC, Luo XD, Li XX. Crucial patterns mining with differential privacy over data streams. Ruan Jian Xue Bao/Journal of Software, 2019, 30(3): 648–666 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5686.htm [doi: 10.13328/j.cnki.jos.005686]

[30] Wang YX, Lei J, Fienberg SE. Learning with differential privacy: Stability, learnability and the sufficiency and necessity of ERM principle. The Journal of Machine Learning Research, 2016, 17(1): 6353–6392.

[31] Dwork C, Roth A. The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 2014, 9(3-4): 211–407. [doi: 10.1561/0400000042]

[32] West DB. Introduction to Graph Theory. 2nd ed., Prentice Hall: Upper Saddle River, 2001.

[33] Even-Dar E, Mannor S, Mansour Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. The Journal of Machine Learning Research, 2006, 7: 1079–1105.

[34] Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. Journal of Privacy and Confidentiality, 2017, 7(3): 17–51. [doi: 10.29012/jpc.v7i3.405]

[35] Hoeffding W. Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association, 1963, 58(301): 13–30. [doi: 10.1080/01621459.1963.10500830]

[36] Erdös P, Rényi A. On the evolution of random graphs. Publication of the Mathematical Institute of the Hungarian Academy of Sciences, 1960, 5(1): 17–61.

[37] Kleinberg R, Slivkins A, Upfal E. Multi-armed bandits in metric spaces. In: Proc. of the 40th Annual ACM Symp. on Theory of Computing. Victoria British: ACM, 2008. 681–690. [doi: 10.1145/1374376.1374475]

附中文参考文献:

[2] 王黎明, 黄厚宽, 柴玉梅. 基于信任和K臂赌博机问题选择多问题协商对象. 软件学报, 2006, 17(12): 2537–2546. http://www.jos.org.cn/1000-9825/17/2537.htm [doi: 10.1360/jos172537]

[19] 叶青青, 孟小峰, 朱敏杰, 霍峥. 本地化差分隐私研究综述. 软件学报, 2018, 29(7): 1981–2005. http://www.jos.org.cn/1000-9825/5364.htm [doi: 10.13328/j.cnki.jos.005364]

[28] 张啸剑, 王淼, 孟小峰. 差分隐私保护下一种精确挖掘top-*k*频繁模式方法. 计算机研究与发展, 2014, 51(1): 104–114. [doi: 10.7544/issn1000-1239.2014.20130685]

[29] 王金艳, 刘陈, 傅星珵, 罗旭东, 李先贤. 差分隐私的数据流关键模式挖掘方法. 软件学报, 2019, 30(3): 648–666. http://www.jos.org.cn/1000-9825/5686.htm [doi: 10.13328/j.cnki.jos.005686]

卢世银(1996－), 男, 博士生, 主要研究领域为机器学习与优化.

邱梓豪(1996－), 男, 硕士生, CCF 学生会员, 主要研究领域为机器学习与优化.

王广辉(1995－), 男, 硕士, 主要研究领域为机器学习与优化.

张利军(1986－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为机器学习与优化.