arXiv:2302.01324v1 [cs.LG] 2 Feb 2023

# 全老虎机反馈下非单调随机子模最大化的随机贪心学习

**Fares Fourati**
阿卜杜拉国王科技大学
fares.fourat@kaust.edu.sa

**Vaneet Aggarwal**
普渡大学 & 阿卜杜拉国王科技大学
vaneet.aggarwal@kaust.edu.sa

**Christopher John Quinn**
爱荷华州立大学
cjquinn@iastate.edu

**Mohamed-Slim Alouini**
阿卜杜拉国王科技大学
slim.alouini@kaust.edu.sa

## 摘要

我们研究了具有全臂反馈和随机奖励、用于子模最大化的无约束组合多臂老虎机问题。以往研究在奖励函数为子模且单调的假设下，探讨了同一问题。本文研究一个更一般的问题，即奖励函数不一定单调，且仅假设其期望具有子模性。我们提出随机贪心学习（RGL）算法，并从理论上证明，对于时域 $T$ 和臂数 $n$，该算法可达到 $\tilde{\mathcal{O}}(nT^{\frac{2}{3}})$ 的 $\frac{1}{2}$- 遗憾上界。实验还表明，在子模和非子模设置下，RGL 的实际表现均优于其他全臂反馈变体。

## 1 引言

随机多臂老虎机由 Robbins（1952年）首次提出，用于形式化描述若干具有挑战性的决策问题，例如临床决策、投资、定价、影响力最大化和产品推荐。决策者的目标可以建模为最大化一个特定的奖励函数，该函数取决于其随时间做出的决策。决策者需要在探索（尝试次优臂）与利用（选择已选定的臂）之间权衡，而关于遗憾的有效保证已得到广泛研究（Thompson，1933年；Auer，2002年；Auer 等人，2002a；Auer 和 Ortner，2010年；Agrawal 和 Goyal，2012年；Gopalan 等人，2014年）。

多臂老虎机问题的一种自然扩展是组合多臂老虎机问题。在每一轮中，智能体不再只选择一个基臂，而是选择一组臂，并获得该集合的联合奖励。如果智能体只能获得所选臂集合的奖励，则称为全老虎机反馈。反之，如果智能体还能获得有关其选择的更多信息，例如该集合中每个单个臂的奖励，则称为半老虎机反馈。前一种情形更具挑战性，因为决策者能够用于决策的信息远少于后一种情形。本文聚焦于前一种情形。

近年来，具有子模奖励函数的组合多臂老虎机问题受到了广泛关注（Nie 等人，2022年；Niazadeh 等，2020）。形式上，设集合函数 $f: 2^{\Omega} \to \mathbb{R}$ 定义在有限底集 Ω 上；如果其满足边际收益递减性质，则称其为子模函数：对于所有 $A \subseteq B \subseteq \Omega$，以及 $x \in \Omega \backslash B$，都有 $f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B)$。子模性假设的动机来自多个现实场景。例如，在某个区域开设更多超市，由于需求饱和，将导致边际收益递减。因此，子模函数被广泛用作经济学和算法博弈论中的效用函数。此外，在组合优化的许多重要场景中也会出现子模性，例如图割（Goemans 和 Williamson，1995；Iwata 等人，2001）、拟阵秩函数（Edmonds，2003）以及集合覆盖问题（Feige，1998）。

多臂老虎机的研究涉及两种不同的设定：对抗性设置 中，对手生成奖励序列，并且可以根据智能体先前的决策来生成该序列（Auer 等人，2002b）；而随机设定 中，每个动作的奖励都独立地从某个（未知）分布中抽取（Auer 等人，2002a）。对于标准多臂老虎机而言，对抗性设置更难应对，而其研究结果可直接用于随机设定，作为一种可行策略（Lattimore 和 Szepesv´ari，2020）。然而，子模多臂老虎机的研究并非如此。在该领域的既有研究中（Roughgarden 和 Wang，2018；Niazadeh 等，2020），对抗性老虎机环境会选择一系列子模函数 $\{f_1, ..., f_T\}$。在本文中，我们聚焦于

26th人工智能与统计国际会议（AISTATS）2023年论文集，西班牙瓦伦西亚。PMLR：第 XXX 卷。版权所有 © 2023 作者。

随机奖励函数。因此，我们假设一种更宽松的子模性条件，即期望意义下的子模性（如定义 1 所定义）。也就是说，我们所考虑问题中的随机函数 $f_t$ 的实现不必是子模的，因此对抗性算法不再适用于这种设定。

**定义 1.** 定义在有限底集 $\Omega$ 上的随机集合函数 $f: 2^\Omega \to \mathbb{R}$，若满足期望意义下的边际收益递减性质，则称其为期望意义下子模：对于所有 $A \subseteq B \subseteq \Omega$ 以及 $x \in \Omega \backslash B$，有：

$$\mathbb{E}f(A \cup \{x\}) - \mathbb{E}f(A) \geq \mathbb{E}f(B \cup \{x\}) - \mathbb{E}f(B). \quad (1)$$

文献中的一些研究假设函数是单调子模函数，因为这类函数更易处理，且可以获得更强的保证（Nie 等人，2022年；Chen 等人，2020年）。若对任意 $A \subseteq B \subseteq \Omega$ 都有 $f(A) \leq f(B)$，则称子模集合函数 $f: 2^\Omega \to \mathbb{R}$ 是单调的。本文考虑一个更一般的问题，即函数不一定是单调的。

非单调子模最大化有若干实际应用场景，包括优化特征选择（Das 和 Kempe，2008；Khanna 等人，2017；Elenberg 等人，2018）和数据摘要（Mirzasoleiman 等人，2016）。由于训练数据可能出现过拟合，特征选择优化可以建模为非单调子模最大化（Fahrbach 等人，2018）。数据摘要旨在选出具有代表性的数据点子集；典型的效用函数是子模的，但并非单调的，从而对较大的解施加惩罚（Tschiatschek 等人，2014；Dasgupta 等人，2013）。更多相关示例见附录 A。

**贡献：** 本文的主要贡献总结如下：

**i.** 我们提出了随机贪心学习（RGL），这是首个针对具有非单调随机子模奖励函数和全老虎机反馈的随机组合多臂老虎机问题而设计的算法。该算法具有较低的存储和计算复杂度。

**ii.** 我们证明，RGL 可实现遗憾上界保证：对于时域 $T$ 和臂数 $n$，上界为 $\mathcal{O}(n T^{\frac{2}{3}} \log(T)^{\frac{1}{3}})$。

**iii.** 实验结果表明，在期望奖励和累积遗憾方面，RGL 优于其他全老虎机反馈方法。

**相关工作：** 次模最大化是 NP-困难。Feige 等人（2011）证明，对于任意常数 $\varepsilon > 0$，任何能够达到 $(\frac{1}{2} + \varepsilon)$ 近似比的算法，都需要对非单调子模函数进行指数数量的预言机查询。此外，他们提出了若干贪心算法，例如确定性自适应算法和随机自适应算法，其近似比分别为 1/3 和 2/5。

算法。近期，Buchbinder 等，2015；Buchbinder 和 Feldman，2018 提出了线性时间 1/2-近似算法。我们的工作将 Buchbinder 等（2015）中的贪心算法从非随机离线设置扩展到随机在线设置，并为随机在线场景证明了遗憾保证。在实际场景中，奖励是随机的；因此，智能体必须在噪声奖励下在线优化探索与利用。在随机函数的样本支持下，在线设置需要权衡探索与利用，才能有效保证遗憾，这使问题更加复杂。

近期，非单调子模最大化问题已在对抗性设置下得到研究（Roughgarden 和 Wang，2018），其中提出了一种全信息条件下的贪心算法，该算法可实现 1/2 遗憾上界 $\tilde{\mathcal{O}}(nT^{\frac{1}{2}})$。除了随机设置与对抗性设置之间的差异外，我们的工作在反馈方式上也有所不同。他们研究的是全信息条件下的问题，即执行动作 $S_t$ 后，不仅能获得奖励 $f_t(S_t)$，还能获得整个函数 $f_t(\cdot)$。我们研究的是全老虎机反馈条件下的问题，也就是说，我们的智能体可用于决策的信息要少得多。对抗性设置下的全老虎机反馈近期也得到研究（Niazadeh 等，2020），其中提出的算法可实现 1/2 遗憾上界 $\tilde{\mathcal{O}}(nT^{\frac{2}{3}})$。

## 2 问题陈述

本节将正式定义本文研究的问题。设 $\Omega$ 为所有基臂的集合，$u_i$ 为索引为 $i$ 的臂，且 $n = |\Omega|$ 为臂数。我们考虑一个具有固定时域 $T$ 的序贯决策问题，其中在每个时间步 $t$，智能体选择一个臂子集（动作）$S_t \subseteq \Omega$。在每个时间步 $t$，智能体使用随机函数 $f(S_t)$ 对所选子集进行采样，并获得样本奖励。

我们假设奖励函数 $f(\cdot)$ 是随机的，且在期望意义下子模（见定义 1），但不一定单调，并且在给定子集的条件下独立同分布。不失一般性，我们假设 $f(\cdot)$ 在 $[0,1]$ [1] 上有界。智能体的目标是在时间范围 $T$ 结束之前最大化累积奖励。

衡量在线学习器随时间推移的表现的一种标准指标，是将其表现与能够访问奖励函数 $f(\cdot)$ 的期望的最优最大化器 $OPT$ 的智能体进行比较，

$$OPT = \arg\max_{S \subseteq \Omega} \mathbb{E}f(S). \quad (2)$$

[1] 通过考虑归一化子模函数 $g(S) = (f(S) - f_{min})/(f_{max} - f_{min})$，这些结果可以直接推广到具有最小值 $f_{min}$ 和最大值 $f_{max}$ 的一般期望意义下子模函数 $f(\cdot)$。

最大化一般的非单调子模函数是一个 NP-困难问题。Feige 等人（2011）研究了非单调子模最大化的困难性，并假设函数 $f(\cdot)$ 通过值预言机获得。他们证明，对于任意常数 $\varepsilon > 0$，任何达到 $(\frac{1}{2} + \varepsilon)$ 近似的算法都需要指数数量的预言机查询。随后，Buchbinder 等人（2015）在离线非随机设定下在线性时间内实现了 $\frac{1}{2}$-近似。因此，我们将智能体的累积奖励与 $\frac{1}{2}T\mathbb{E}f(OPT)$ 进行比较，并将累积 $\frac{1}{2}$-遗憾记为 $\mathcal{R}_{\frac{1}{2}}(T)$，其中，

$$\mathcal{R}_{\frac{1}{2}} = \sum_{t=1}^{T}(\frac{1}{2}f(OPT) - f(S_t)) \tag{3}$$

请注意，$\frac{1}{2}$-遗憾是随机的，其随机性源于奖励函数 $f(\cdot)$ 的随机性，以及随时间推移所选择的动作（子集）的随机性。因此，我们主要关注于最小化智能体的期望 $\frac{1}{2}$-遗憾，其定义如下：

$$\mathbb{E}[\mathcal{R}_{\frac{1}{2}}] = \frac{1}{2}T\mathbb{E}[f(OPT)] - \sum_{t=1}^{T}\mathbb{E}[f(S_t)], \tag{4}$$

其中，期望取自 $f(\cdot)$ 的随机性和所选动作序列的随机性。为便于记号表示，在本文余下部分中，我们使用 $\mathcal{R}(T)$ 代替 $\mathcal{R}_{\frac{1}{2}}(T)$。

## 3 所提出的 RGL 算法

本节介绍我们所提出的 RGL 算法，该算法改编自 Buchbinder 等，2015 针对非随机 $f(\cdot)$ 所提出的离线算法。RGL 的伪代码见算法 1。

对于我们所考虑的问题（无约束动作空间），如果奖励函数是单调的，那么最佳集合就是包含所有基臂 Ω 的集合。一个简单的算法（无需探索）即可达到 $\alpha = 1$ 的近似比。然而，当奖励函数是非单调的时，增加臂不再必然是明智的选择。因此，跟踪两个集合 $X_i$（初始为 $\emptyset$）和 $Y_i$（初始为 Ω）是一种有效策略。RGL 依次遍历所有单个臂，并决定将其添加到基臂集合 $X_i$ 中，或将其从基臂集合 $Y_i$ 中移除。添加或移除任意臂的决策均采用随机贪心方式，并利用边际增益的经验估计，直至对所有单个臂都作出决策；随后，算法利用所确定的最佳臂集合。

设 $X_i$ 和 $Y_i$ 是两个臂集合。初始时，$X_0 = \emptyset$ 和 $Y_0 = \Omega$。该算法包含 $n$ 个阶段，其中 $n$ 是臂的数量；每个阶段包含 $m$ 个子阶段，其中 $m$ 是用于估计给定臂集合质量的重复次数。在第 $i$ 个阶段（共 $n$ 个阶段）中，智能体估计以下两个随机变量的期望，$a_i$ 和 $b_i$，定义如下：

$$\begin{aligned} a_i &= f(X_{i-1} \cup \{u_i\}) - f(X_{i-1}) \\ b_i &= f(Y_{i-1} \setminus \{u_i\}) - f(Y_{i-1}). \end{aligned} \tag{5}$$

**Algorithm 1** RGL

**Require:** Set of base arms $\Omega$, horizon $T$

```
X_0 ← ∅, Y_0 ← Ω, n ← |Ω|
m ← ⌈(T sqrt(25/32 log(T)))^{2/3}⌉
for arm index i ∈ {1, ..., n} do
    ā_i ← 0 and b̄_i ← 0
    for sample j ∈ {1, ..., m} do
        Play X_{i-1} ∪ {u_i}, X_{i-1}, Y_{i-1}, and Y_{i-1} \ {u_i}
        ā_i ← ā_i + (f_j(X_{i-1} ∪ {u_i}) − f_j(X_{i-1}))/m
        b̄_i ← b̄_i + (f_j(Y_{i-1} \ {u_i}) − f_j(Y_{i-1}))/m
    end for
    a'_i ← max(ā_i, 0) and b'_i ← max(b̄_i, 0)
    with probability (a'_i / (a'_i + b'_i)) do
        X_i ← X_{i-1} ∪ {u_i} and Y_i ← Y_{i-1}
    else
        Y_i ← Y_{i-1} \ {u_i} and X_i ← X_{i-1}
end for
for remaining time do
    Play X_n
end for
```

由于 $f(\cdot)$ 是随机的，因此 $a_i$ 和 $b_i$ 也是随机的，即使在给定 $X_{i-1}$、$Y_{i-1}$ 和臂 $u_i$ 的条件下也是如此。为了估计给定集合 $X_i$ 和 $Y_i$ 以及臂 $u_i$ 时它们的期望，智能体对式 (5) 中的四个随机集合值各采样 $m$ 次。将已选集合 $S$ 的第 $j$ 个样本记为 $f_j(S)$，将执行动作 $S$ 的经验均值定义如下：

$$\bar{f}(S) := \frac{1}{m}\sum_{j=1}^{m} f_j(S). \tag{6}$$

因此，智能体经过 $m$ 次重复后，计算它们的经验均值 $\bar{a}_i$ 和 $\bar{b}_i$，即：

$$\begin{aligned} \bar{a}_i &= \bar{f}(X_{i-1} \cup \{u_i\}) - \bar{f}(X_{i-1}) \\ \bar{b}_i &= \bar{f}(Y_{i-1} \setminus \{u_i\}) - \bar{f}(Y_{i-1}). \end{aligned} \tag{7}$$

这两个估计值对于决策过程至关重要。$\bar{a}_i$ 衡量将臂 $u_i$ 添加到 $X_{i-1}$ 所产生的预期影响，而 $\bar{b}_i$ 衡量从 $Y_{i-1}$ 中移除臂 $u_i$ 所产生的预期影响。通过计算一个取决于这两个估计值 $\bar{a}_i$ 和 $\bar{b}_i$ 的特定概率，以贪心且概率性的方式做出决策，定义如下：

$$p = \frac{a'_i}{a'_i + b'_i}, \tag{8}$$

其中 $a'_i = \max(\bar{a}_i, 0)$ 和 $b'_i = \max(\bar{b}_i, 0)$，这解释了该算法为何称为随机贪心。在特殊情况 $a'_i = b'_i = 0$ 下，我们令 $p = 1$。

以概率 $p$，智能体将单个臂 $i$ 加入臂集合 $X_i$，并将其保留在臂集合 $Y_i$ 中；以概率 $1-p$，智能体从臂集合 $Y_i$ 中移除臂 $i$，并保留与 $X_{i-1}$ 中相同的臂。因此，对于所有 $i = 1, \ldots, n$，都有 $X_i \subseteq Y_i$。检查完所有 $n$ 个单个臂后，很容易看出，根据该算法的构造，集合 $X_n$ 和 $Y_n$ 都恰好包含相同的臂，即 $X_n = Y_n$。因此，在决定了每个 $n$ 个基臂之后，智能体会在剩余时间内利用 $X_n$。

**注 1.** 请注意，我们的随机贪心算法中的随机性对于实现 $1/2$ 近似保证至关重要。采用确定性决策的同一算法，即当 $a_i \geq b_i$时添加索引为 $i$ 的臂，只能达到 $1/3$ 近似保证（*Buchbinder* 等，*2015*）。

RGL 的存储复杂度和每轮时间复杂度都较低。在利用阶段，RGL 只需存储所选基臂集合 $X_n$ 的索引，最多为 $n$，且无需进一步计算。在探索阶段的第 $i$ 阶段，RGL 需要更新 $\bar{a}_i$ 和 $b_i$ 的经验均值，并更新 $X_i$ 和 $Y_i$。因此，RGL 的存储复杂度为 $\mathcal{O}(n)$，每轮时间复杂度为 $\mathcal{O}(1)$。

## 4 遗憾分析

本节将给出本文的主要结果，即所提出算法的期望累积1 2-遗憾上界。在介绍主要结果之前，我们先给出证明主要结果时会用到的引理。

**引理 1.** 对于每个 $i \in \{1, \cdots, n\}$，都有 $\mathbb{E}[a_i + b_i] \geq 0$，其中 $a_i$、$b_i$ 如(5)所定义。

证明。根据构造，$X_{i-1} \subseteq Y_{i-1} \backslash \{u_i\}$ 和 $u_i \in Y_{i-1}$。因此，根据子模性的定义 1，将 $u_i$ 添加到 $Y_{i-1} \backslash \{u_i\}$ 的期望边际增益小于或等于将 $u_i$ 添加到 $X_{i-1}$ 的边际增益，

$$\begin{aligned} &\mathbb{E}\left[f\left(Y_{i-1}\right) - f\left(Y_{i-1} \backslash \{u_i\}\right)\right] \\ &\quad \leq \mathbb{E}[f\left(X_{i-1} \cup \{u_i\}\right) - f\left(X_{i-1}\right)]. \end{aligned} \tag{9}$$

将 (5) 代入 (9) 得到 $\mathbb{E}[-b_i] \leq \mathbb{E}[a_i]$，整理后即完成证明。

□

对于每个臂 $u_i$，智能体会将以下动作列表 $\mathcal{S}_i = [X_{i-1}, X_{i-1} \cup \{u_i\}, Y_{i-1}, Y_{i-1} \backslash \{u_i\}]$ 恰好执行$m$次，然后计算边际增益估计。为确定 $m$，我们为所有动作 $S_i$ 的经验估计~~采用相同大~~小的置信半径 $\text{rad} := \sqrt{2\log(T)/m}$。增大 $m$ 可以提高经验估计围绕其均值的集中性，从而提升使用这些经验估计所作决策的质量。然而，增大 $m$ 会导致花费更多时间执行其值可能远离$\frac{1}{2}f(OPT)$的动作，从而造成较高的累积遗憾。

将测试臂 $u_i$ 时所采取动作的经验均值集中在其统计均值附近这一事件记为，

$$\mathcal{E}_i := \bigcap_{S \in \mathcal{S}_i} \{|\bar{f}(S) - \mathbb{E}[\bar{f}(S)]| < \text{rad}\} \tag{10}$$

然后，我们将干净事件 $\mathcal{E}$ 定义为：截至并包括臂 $u_n$，所有已执行动作的经验均值都在其对应统计均值的半径范围内：

$$\mathcal{E} := \mathcal{E}_1 \cap \cdots \cap \mathcal{E}_n. \tag{11}$$

所执行的具体动作序列取决于先前动作及其奖励的经验估计。然而，给定任意时刻 $t$ 执行的当前动作 $S_t$，随机奖励 $f(S_t)$ 与过去的动作及其奖励相互独立。

利用霍夫丁界，我们证明 $\mathcal{E}$ 以高概率发生。接下来，我们利用经验均值的集中性 (10) 以及期望意义下子模性的性质（定义 1），推导后续结论。

**备注 2.** 在干净事件 $\mathcal{E}_i$ *(10)* 下，对所有 $S \in \mathcal{S}_i$，

$$|\bar{f}(S) - \mathbb{E}[\bar{f}(S)]| < \text{rad}\,.$$

因此，由于 $X_{i-1}$ 属于 $\mathcal{S}_i$，

$$\mathbb{E}[\bar{f}(X_{i-1})] - rad \leq \bar{f}(X_{i-1}) \leq \mathbb{E}[\bar{f}(X_{i-1})] + \text{rad}\,.$$

$X_{i-1} \cup \{u_i\}$、$Y_{i-1}$ 和 $Y_{i-1} \backslash \{u_i\}$ 之间也有类似的关系。

因此，

$$\begin{aligned} &\mathbb{E}[\bar{f}(X_{i-1} \cup \{u_i\})] - \mathbb{E}[\bar{f}(X_{i-1})] - 2rad \\ &\quad \leq \bar{f}(X_{i-1} \cup \{u_i\}) - \bar{f}(X_{i-1}) \\ &\quad = \frac{1}{m}\sum_{j=1}^{m}(f_j(X_{i-1} \cup \{u_i\}) - f_j(X_{i-1})) \\ &\quad = \bar{a}_i. \qquad \text{(by (7))} \end{aligned}$$

因此，

$$\mathbb{E}[a_i] - 2rad \leq \bar{a}_i$$

使用类似步骤，可以很容易地验证，

$$\begin{aligned} \mathbb{E}[a_i] - 2rad \leq \bar{a}_i \leq \mathbb{E}[a_i] + 2rad \\ \mathbb{E}[b_i] - 2rad \leq \bar{b}_i \leq \mathbb{E}[b_i] + 2rad. \end{aligned} \tag{12}$$

**推论 1.** 在干净事件 $\mathcal{E}$下，对每个$1 \leq i \leq n$，$\bar{a}_i + b_i \geq -4rad$。

证明。在干净事件 $\mathcal{E}$, $\bar{a}_i \geq \mathbb{E}[a_i] - 2rad$ and $b_i \geq \mathbb{E}[b_i] - 2rad$下。由于 $\mathbb{E}[a_i + b_i] \geq 0$（根据引理 1），因此，$\bar{a}_i + b_i \geq \mathbb{E}[a_i + b_i] - 4rad \geq -4rad$。 □

**引理 2**。 定义 $OPT_i := (OPT \cup X_i) \cap Y_i$。在干净事件 $\mathcal{E}$下，对于每个 $1 \leq i \leq n$，我们有

$$\begin{aligned}&\mathbb{E}[f(OPT_{i-1}) - f(OPT_i))] \\ &\leq \frac{1}{2}\mathbb{E}[f(X_i) - f(X_{i-1}) + f(Y_i) - f(Y_{i-1})] + 5rad.\end{aligned} \tag{13}$$

证明。只需证明在任何形如 $X_{i-1} = S_{i-1}$ where $S_{i-1} \subseteq \{u_1, ..., u_{i-1}\}$ 的事件条件下，该不等式成立，且该事件的概率$X_{i-1} = S_{i-1}$ 非零。证明的其余部分均假设条件为此事件。我们通过考虑以下四种可能情形来证明引理 2，其中 $\bar{a}_i$ 和 $b_i$:

**情形 1 ($\bar{a}_i \geq 0$ 和 $b_i \leq 0$)**：在这种情况下 $b_i \leq 0 \Rightarrow b'_i = 0 \Rightarrow \frac{a'_i}{a'_i + b'_i} = 1$。因此，$Y_i = Y_{i-1}$ 和 $X_i = X_{i-1} \cup \{u_i\}$。由于 $X_i = X_{i-1} \cup \{u_i\}$，我们有

$$\begin{aligned}\mathbb{E}[a_i] &= \mathbb{E}[f(X_{i-1} \cup \{u_i\}) - f(X_{i-1})] \\ &= \mathbb{E}[f(X_i) - f(X_{i-1})].\end{aligned} \tag{14}$$

由于 $Y_i = Y_{i-1}$，我们要证明的关系 (13) 可化简为,

$$\begin{aligned}&\mathbb{E}[f(OPT_{i-1}) - f(OPT_i)] \\ &\leq \frac{1}{2}\mathbb{E}[f(X_i) - f(X_{i-1})] + 5rad.\end{aligned}$$

Notice, $OPT_i = (OPT \cup X_i) \cap Y_i = OPT_{i-1} \cup \{u_i\}$. 如果 $u_i \in OPT \Rightarrow OPT_i = OPT_{i-1}$。因此,

$$\begin{aligned}&\mathbb{E}[f(OPT_i) - f(OPT_{i-1})] \\ &= 0 \\ &\leq \frac{\bar{a}_i}{2} && \text{(by case 1 condition)} \\ &\leq \frac{\mathbb{E}[a_i]}{2} + rad && \text{(using (12))} \\ &= \frac{1}{2}\mathbb{E}[f(X_i) - f(X_{i-1})] + rad && \text{(by (14))} \\ &\leq \frac{1}{2}\mathbb{E}[f(X_i) - f(X_{i-1})] + 5rad.\end{aligned}$$

现在考虑 $u_i \notin OPT$。根据 $OPT_i$的定义， $OPT_{i-1} \subseteq Y_{i-1}$ 成立，因此这里有 $OPT_{i-1} \subseteq Y_{i-1} \backslash \{u_i\}$，且 $(Y_{i-1} \backslash \{u_i\}) \cup \{u_i\} = Y_{i-1}$，所以根据期望意义下的子模性的定义 1,

$$\begin{aligned}&\mathbb{E}[f(Y_{i-1})] - \mathbb{E}[f(Y_{i-1} \backslash \{u_i\})] \\ &\leq \mathbb{E}[f(OPT_{i-1} \cup \{u_i\})] - \mathbb{E}[f(OPT_{i-1})].\end{aligned} \tag{15}$$

对 (15) 取否定，我们得到

$$\begin{aligned}&\mathbb{E}[f(OPT_{i-1})] - \mathbb{E}[f(OPT_{i-1} \cup \{u_i\})] \\ &\leq \mathbb{E}[f(Y_{i-1} \backslash \{u_i\})] - \mathbb{E}[f(Y_{i-1})] && \text{(negating (15))} \\ &= \mathbb{E}[b_i] && \text{(by def. of } b_i \text{ (5))} \\ &\leq \bar{b}_i + 2rad && \text{(using (12))} \\ &\leq \frac{\bar{a}_i}{2} + 2rad && \text{(condition for case 1)} \\ &\leq \frac{1}{2}\mathbb{E}[a_i] + 3rad && \text{(using (12))} \\ &= \frac{1}{2}\mathbb{E}[f(X_i) - f(X_{i-1})] + 3rad && \text{(by (14))} \\ &\leq \frac{1}{2}\mathbb{E}[f(X_i) - f(X_{i-1})] + 5rad.\end{aligned}$$

**情形2** （$\bar{a}_i < 0$ 和 $b_i \geq 0$）：此情形与情形1类似，其证明请参阅附录 B.2。

**情形 3 ($\bar{a}_i < 0$ 和 $b_i < 0$)**：对于此情形，根据 $a'_i$ 和 $b'_i$ 的定义，我们有 $a'_i = b'_i = 0$。因此，臂 $u_i$ 的选择概率将设为 $\frac{a'_i}{a'_i + b'_i} = 1$，这意味着 $X_i = X_{i-1} \cup \{u_i\}$ 和 $Y_i = Y_{i-1}$。因此，我们有

$$\begin{aligned}\mathbb{E}[a_i] &= \mathbb{E}[f(X_{i-1} \cup \{u_i\}) - f(X_{i-1})] \\ &= \mathbb{E}[f(X_i) - f(X_{i-1})].\end{aligned} \tag{16}$$

因此，只需证明

$$\mathbb{E}(f(OPT_{i-1}) - f(OPT_i))) \leq \frac{1}{2}\mathbb{E}[a_i] + 5rad. \tag{17}$$

注意， $OPT_i = (OPT \cup X_i) \cap Y_i = OPT_{i-1} \cup \{u_i\}$。此外，根据推论 1，在干净事件下，对于每个 $1 \leq i \leq n$，都有 $\bar{a}_i + b_i \geq -4rad$。由于 $b_i < 0$，于是 $\bar{a}_i \geq -4rad$，我们有

$$\frac{\bar{a}_i}{2} + 2rad \geq 0 > \bar{b}_i. \tag{18}$$

如果 $u_i \in OPT$，那么 $OPT_i = OPT_{i-1}$。因此，我们有

$$\begin{aligned}&\mathbb{E}[f(OPT_i) - f(OPT_{i-1})] \\ &= 0 \\ &\leq \frac{\bar{a}_i}{2} + 2rad && \text{(by (18))} \\ &\leq \frac{1}{2}\mathbb{E}[a_i] + 3rad && \text{(using (12))} \\ &\leq \frac{1}{2}\mathbb{E}[a_i] + 5rad.\end{aligned}$$

If $u_i \notin OPT$, then $OPT_{i-1} \subseteq Y_{i-1}$ and $(Y_{i-1} \backslash \{u_i\}) \cup \{u_i\} = Y_{i-1}$ Thus, like in case 1, (15)

成立。对（15）取否定，可得，

$$\begin{aligned}
&\mathbb{E}f(OPT_{i-1}) - \mathbb{E}f(OPT_{i-1}\cup\{u_i\}) \\
&\quad\le \mathbb{E}f\left(Y_{i-1}\backslash\{u_i\}\right) - \mathbb{E}f\left(Y_{i-1}\right) && \text{(from (15))} \\
&\quad= \mathbb{E}[b_i] && \text{(from def. (5))} \\
&\quad\le \bar{b}_i + 2rad && \text{(using (12))} \\
&\quad\le \frac{\bar{a}_i}{2} + 4rad && \text{(by (18))} \\
&\quad\le \frac{1}{2}\mathbb{E}[\bar{a}_i] + 5rad.
\end{aligned}$$

**情况4** ($\bar{a}_i \ge 0$ 和 $b_i > 0$)：在这种情况下，$a_i' = \bar{a}_i$ 和 $b_i' = b_i$。因此，根据算法，以概率 $\frac{a_i'}{a_i'+b_i'}$，$X_i \leftarrow X_{i-1}\cup\{u_i\}$　$Y_i \leftarrow Y_{i-1}$ 且，以概率 $\frac{b_i'}{a_i'+b_i'}$，$Y_i \leftarrow Y_{i-1}\setminus\{u_i\}$ 且 $X_i \leftarrow X_{i-1}$。我们有，

$$\begin{aligned}
&\mathbb{E}\left[f\left(X_i\right) - f\left(X_{i-1}\right) + f\left(Y_i\right) - f\left(Y_{i-1}\right)\right] \\
&= \mathbb{E}[\mathbb{E}\left[f\left(X_i\right) - f\left(X_{i-1}\right) + f\left(Y_i\right) - f\left(Y_{i-1}\right)|\,\bar{a}_i, \bar{b}_i\right]] \\
&= \mathbb{E}[\frac{\bar{a}_i}{\bar{a}_i+\bar{b}_i}\mathbb{E}\left[f\left(X_{i-1}\cup\{u_i\}\right) - f\left(X_{i-1}\right)\right] \\
&\quad + \frac{\bar{b}_i}{\bar{a}_i+\bar{b}_i}\mathbb{E}\left[f\left(Y_{i-1}\backslash\{u_i\}\right) - f\left(Y_{i-1}\right)\right]] \\
&= \mathbb{E}\left[\frac{\bar{a}_i\mathbb{E}[a_i]}{\bar{a}_i+\bar{b}_i} + \frac{\bar{b}_i\mathbb{E}[b_i]}{\bar{a}_i+\bar{b}_i}\right] && \text{(def. of } a_i \text{ and } b_i\text{)} \\
&\ge \mathbb{E}\left[\frac{\bar{a}_i(\bar{a}_i-2rad)}{\bar{a}_i+\bar{b}_i} + \frac{\bar{b}_i(\bar{b}_i-2rad)}{\bar{a}_i+\bar{b}_i}\right] && \text{(using (12))} \\
&= \mathbb{E}\left[\frac{\bar{a}_i^2}{\bar{a}_i+\bar{b}_i} + \frac{\bar{b}_i^2}{\bar{a}_i+\bar{b}_i} - \frac{2rad(\bar{a}_i+\bar{b}_i)}{\bar{a}_i+\bar{b}_i}\right] \\
&= \mathbb{E}\left[\frac{\bar{a}_i^2+\bar{b}_i^2}{\bar{a}_i+\bar{b}_i}\right] - 2rad.
\end{aligned}$$

因此，

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}\left[\frac{\bar{a}_i^2+\bar{b}_i^2}{\bar{a}_i+\bar{b}_i}\right] - rad \\
&\le \frac{1}{2}\mathbb{E}\left[f\left(X_i\right) - f\left(X_{i-1}\right) + f\left(Y_i\right) - f\left(Y_{i-1}\right)\right].
\end{aligned} \tag{19}$$

此外，

$$\begin{aligned}
&\mathbb{E}[f(OPT_{i-1}) - f(OPT_i)] \\
&= \mathbb{E}[\frac{\bar{a}_i}{\bar{a}_i+\bar{b}_i}\mathbb{E}[f(OPT_{i-1}) - f(OPT_{i-1}\cup\{u_i\})] \\
&\quad + \frac{\bar{b}_i}{\bar{a}_i+\bar{b}_i}\mathbb{E}[f(OPT_{i-1}) - f(OPT_{i-1}\setminus\{u_i\})]]
\end{aligned}$$

如果 $u_i \notin OPT \Rightarrow$第二项为零且 $OPT_{i-1}\subseteq Y_{i-1}\backslash\{u_i\}$。因此，根据期望意义下的子模性，

$$\begin{aligned}
&\mathbb{E}[f(OPT_{i-1}) - f(OPT_{i-1}\cup\{u_i\})] \\
&\qquad\le \mathbb{E}[f(Y_{i-1}\setminus\{u_i\}) - f(Y_i)] \\
&\qquad= \mathbb{E}[b_i] \\
&\qquad\le \bar{b}_i + 2rad,
\end{aligned}$$

因此，如果 $u_i \notin OPT$，那么

$$\begin{aligned}
&\mathbb{E}[f(OPT_{i-1}) - f(OPT_i)] \\
&\qquad\le \mathbb{E}[\frac{\bar{a}_i}{\bar{a}_i+\bar{b}_i}(\bar{b}_i+2rad) + \frac{\bar{b}_i}{\bar{a}_i+\bar{b}_i}0] \\
&\qquad= \mathbb{E}[\frac{\bar{a}_i\bar{b}_i}{\bar{a}_i+\bar{b}_i}] + 2rad\mathbb{E}[\frac{\bar{a}_i}{\bar{a}_i+\bar{b}_i}].
\end{aligned} \tag{20}$$

如果 $u_i \in OPT \Rightarrow$第一项为零且 $X_{i-1}\subseteq(OPT\cup X_{i-1})\cap Y_{i-1}\setminus\{u_i\}$。因此，根据期望意义下的子模性，我们有

$$\begin{aligned}
&\mathbb{E}[f(OPT_{i-1}) - f(OPT_{i-1}\setminus\{u_i\})] \\
&\qquad\le \mathbb{E}[f(X_{i-1}\setminus\{u_i\}) - f(X_i)] \\
&\qquad= \mathbb{E}[a_i] \\
&\qquad\le \bar{a}_i + 2rad.
\end{aligned}$$

因此，如果 $u_i \in OPT$那么

$$\begin{aligned}
&\mathbb{E}[f(OPT_{i-1}) - f(OPT_i)] \\
&\qquad\le \mathbb{E}[\frac{\bar{a}_i}{\bar{a}_i+\bar{b}_i}0 + \frac{\bar{b}_i}{\bar{a}_i+\bar{b}_i}(\bar{a}_i+2rad)] \\
&\qquad= \mathbb{E}[\frac{\bar{a}_i\bar{b}_i}{\bar{a}_i+\bar{b}_i}] + 2rad\mathbb{E}[\frac{\bar{b}_i}{\bar{a}_i+\bar{b}_i}].
\end{aligned} \tag{21}$$

由于在此情形下我们对 ($\bar{a}_i \ge 0$ 和 $b_i > 0$) 进行条件化，因此有

$$\mathbb{E}[\frac{\bar{a}_i}{\bar{a}_i+\bar{b}_i}] \ge 0 \qquad \text{and} \qquad \mathbb{E}[\frac{\bar{b}_i}{\bar{a}_i+\bar{b}_i}] \ge 0$$

结合界 (20) 和 (21)，我们得到

$$\begin{aligned}
&\mathbb{E}[f(OPT_{i-1}) - f(OPT_i)] \\
&\le \mathbb{E}[\frac{\bar{a}_i\bar{b}_i}{\bar{a}_i+\bar{b}_i}] + 2rad\mathbb{E}[\frac{\bar{b}_i}{\bar{a}_i+\bar{b}_i}] + 2rad\mathbb{E}[\frac{\bar{b}_i}{\bar{a}_i+\bar{b}_i}] \\
&= \mathbb{E}[\frac{\bar{a}_i\bar{b}_i}{\bar{a}_i+\bar{b}_i}] + 2rad,
\end{aligned} \tag{22}$$

无论 $u_i$ 是否属于 $OPT$，该结论均成立。

对于 $x+y>0$，根据柯西-施瓦茨不等式，

$$\frac{xy}{x+y} \le \frac{1}{2}\frac{x^2+y^2}{x+y}. \tag{23}$$

结合上述观察可知，

$$\begin{aligned}
&\mathbb{E}[f(OPT_{i-1}) - f(OPT_i)] \\
&\overset{(22)}{\le} \mathbb{E}[\frac{\bar{a}_i\bar{b}_i}{\bar{a}_i+\bar{b}_i}] + 2rad \\
&\overset{(23)}{\le} \frac{1}{2}\mathbb{E}[\frac{\bar{a}_i^2+\bar{b}_i^2}{\bar{a}_i+\bar{b}_i}] + 2rad \\
&\overset{(19)}{\le} \frac{1}{2}\mathbb{E}\left[f\left(X_i\right) - f\left(X_{i-1}\right) + f\left(Y_i\right) - f\left(Y_{i-1}\right)\right] + 3rad \\
&\le \frac{1}{2}\mathbb{E}\left[f\left(X_i\right) - f\left(X_{i-1}\right) + f\left(Y_i\right) - f\left(Y_{i-1}\right)\right] + 5rad.
\end{aligned}$$

□

**推论 2.** 在干净事件 $\mathcal{E}$下，

$$\mathbb{E}[f(X_n)] + \frac{5}{2} n\, rad \geq \frac{1}{2}\mathbb{E}[f(OPT)]. \tag{24}$$

证明。 对 $1 \leq i \leq n$，将引理 2 中的 (13) 求和可得，

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}\left[f\left(OPT_{i-1}\right) - f\left(OPT_i\right)\right] \\
&\quad \leq 5n\, rad + \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\left[f\left(X_i\right) - f\left(X_{i-1}\right)\right] \\
&\qquad + \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\left[f\left(Y_i\right) - f\left(Y_{i-1}\right)\right].
\end{aligned}$$

注意到上面的求和具有望远镜性质。化简后，

$$\begin{aligned}
&\mathbb{E}\left[f\left(OPT_0\right) - f\left(OPT_n\right)\right] \\
&\leq 5nrad + \frac{1}{2} \cdot \mathbb{E}\left[f\left(X_n\right) - f\left(X_0\right) + f\left(Y_n\right) - f\left(Y_0\right)\right] \\
&\leq 5nrad + \frac{\mathbb{E}\left[f\left(X_n\right) + f\left(Y_n\right)\right]}{2}.
\end{aligned}$$

注意到 $OPT_0 = OPT$ 和 $OPT_n = X_n = Y_n$，即可得到结果。 □

讨论完关键引理后，下面的结果给出了 RGL 的期望累积 $_{1\,2}$-遗憾的界。

**定理 1**。 对于第 *2* 节中定义的、满足 $T \geq 2$ 的序贯决策问题，*RGL* 的期望累积 $_{1\,2}$-遗憾至多为 $\mathcal{O}(nT^{\frac{2}{3}}\log(T)^{\frac{1}{3}})$。

证明。 我们首先在干净事件的条件下考察期望累积遗憾。

$$\begin{aligned}
&\mathbb{E}(R(T)|\mathcal{E}) \\
&= \frac{1}{2}T\mathbb{E}[f\left(OPT\right)] - \sum_{t=1}^{T}\mathbb{E}[f\left(S_t\right)] \\
&= \sum_{t=1}^{T}\left(\frac{1}{2}\mathbb{E}[f(OPT)] - \mathbb{E}[f\left(S_t\right)]\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\Bigg[\left(\frac{1}{2}\mathbb{E}[f(OPT)] - \mathbb{E}[f\left(X_{i-1}\right)]\right) \\
&+ \left(\frac{1}{2}\mathbb{E}[f(OPT)] - \mathbb{E}[f\left(X_{i-1} \cup \{u_i\}\right)]\right) \\
&+ \left(\frac{1}{2}\mathbb{E}[f(OPT)] - \mathbb{E}[f\left(Y_{i-1}\right)]\right) \\
&+ \left(\frac{1}{2}\mathbb{E}[f(OPT)] - \mathbb{E}[f\left(Y_{i-1} \setminus \{u_i\}\right)]\right)\Bigg] \\
&+ \sum_{t=4nm+1}^{T}\left(\frac{1}{2}\mathbb{E}[f(OPT)] - \mathbb{E}[f\left(S_t\right)]\right).
\end{aligned} \tag{25}$$

我们将总和分为两部分：第一部分计算探索阶段产生的累积遗憾，第二部分计算利用阶段产生的累积遗憾。在探索阶段，对于每个臂 $u_i$，智能体会分别将四个子集 $X_i$、 $Y_i$、$X_{i-1} \cup \{u_i\}$ 和 $Y_{i-1} \setminus \{u_i\}$ 各选择 $m$ 次。因此，智能体共探索 $4mn$ 个时间步。由于智能体在时间步 $t$ 选择的任意子集 $S_t$ 所对应的 $f(\cdot)$ 均有界于 $[0,1],$，

$$\frac{1}{2}\mathbb{E}[f\left(OPT\right)] - \mathbb{E}[f\left(S_t\right)] \leq \frac{1}{2}. \tag{26}$$

将 (26) 代入 (25)，可得

$$\begin{aligned}
&\mathbb{E}[R(T) \mid \mathcal{E}] \\
&\quad \leq 4nm\frac{1}{2} + \sum_{t=T_{n+1}}^{T}\left(\frac{1}{2}\mathbb{E}[f(OPT)] - \mathbb{E}[f\left(S_t\right)]\right) \\
&\quad = 2nm + \sum_{t=T_{n+1}}^{T}\left(\frac{1}{2}\mathbb{E}[f(OPT)] - \mathbb{E}[f\left(X_n\right)]\right).
\end{aligned}$$

根据推论 2, $\frac{1}{2}\mathbb{E}[f(OPT) - f(X_n)] \leq \frac{5}{2} n\, rad$。因此，

$$\begin{aligned}
\mathbb{E}[R(T) \mid \mathcal{E}] &\leq 2nm + \sum_{t=T_{n+1}}^{T}\left(\frac{5}{2} n\, rad\right) \\
&\leq 2nm + \frac{5}{2}Tn\, rad.
\end{aligned}$$

由于 $rad = \sqrt{2\log(T)/m},$ 我们有

$$\mathbb{E}(R(T) \mid \mathcal{E}) \leq 2nm + \frac{5}{2}Tn\sqrt{2\frac{\log(T)}{m}}.$$

上述不等式对所有严格大于零的 $m$ 均成立。因此，为了得到更紧的界，我们求使左侧最小的 $m^*$。精确的最小化解为：

$$m^\star = \left(T\sqrt{\frac{25}{32}\log(T)}\right)^{2/3}.$$

因此，我们选择 $m = \lceil m^\star \rceil$。

$$\begin{aligned}
\mathbb{E}(R(T) \mid \mathcal{E}) &\leq 2n\lceil m^\star \rceil + \frac{5}{2}nT\sqrt{\frac{2\log(T)}{\lceil m^\star \rceil}} \\
&\leq 2n\lceil m^\star \rceil + \frac{5}{2}nT\sqrt{\frac{2\log(T)}{m^*}}
\end{aligned}$$

对于 $T \geq 2, m^\star \geq \frac{1}{2}$ 因此 $\lceil m^\star \rceil \leq 2m^*$。因此，我们有

$$\begin{aligned}
\mathbb{E}(R(T) \mid \mathcal{E}) &\leq 4nm^\star + \frac{5}{2}nT\sqrt{\frac{2\log(T)}{m^*}} \\
&\leq \mathcal{O}(nT^{\frac{2}{3}}\log(T)^{\frac{1}{3}})
\end{aligned}$$

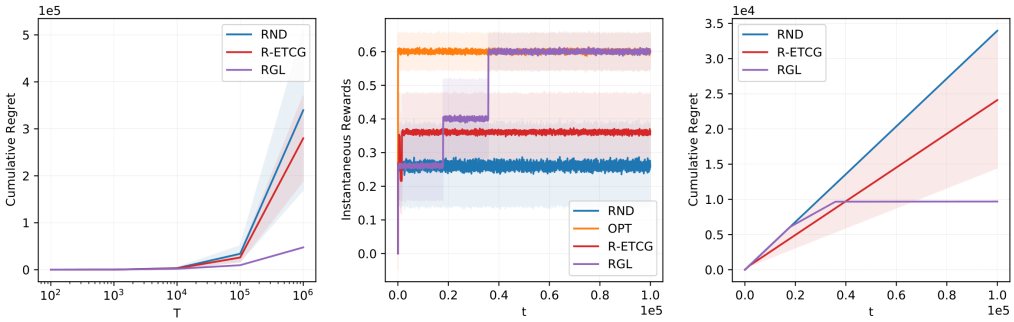


图1：非单调随机子模奖励函数的比较结果。从左到右，图中分别展示了累积遗憾随时间步 $T$ 的变化、即时奖励随时间步 $t$ 的变化，以及累积遗憾随时域 $t$ 的变化。

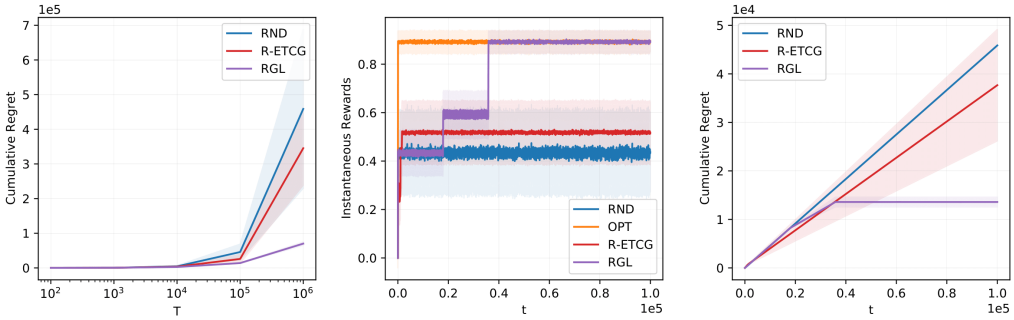


图2：非单调随机非子模奖励函数的比较结果。从左到右，图中分别展示了累积遗憾随时间步 $T$ 的变化、即时奖励随时间步 $t$ 的变化，以及累积遗憾随时域 $t$ 的变化。

¯在坏事件下，即好事件 $\mathcal{E}$的补集 $\bar{\mathcal{E}}$，由于奖励在 $[0,1]$, it¯中有界，很容易看出 $\mathbb{E}(R(T) \mid \bar{\mathcal{E}}) \leq T$。此外，通过使用附录B中的引理 3以及 Hoeffding 不等式（Hoeffding，1994），我们有 $\mathbb{P}(\bar{\mathcal{E}}) \leq \frac{8n}{T^4}$，参见附录B中的引理 4。因此，我们得到 $\mathbb{E}(R(T)) \leq \mathcal{O}(\bar{n}T^{\frac{2}{3}}\log(T)^{\frac{1}{3}})$。 □

**注3.** 当时间范围 $T$ 未知时，我们可以使用几何倍增技巧，将结果扩展为任意时刻算法。具体而言，我们选取一个几何序列 $T_i = T_0 2^i$，用于 $i \in \{1,2,\cdots\}$，其中 $T_0$ 是一个足够大的数，使算法能够完成初始化；随后在时间区间 $T_{i+1} - T_i$ 内运行 *RGL*，并在每个区间开始时完全重启（*Besson*和*Kaufmann*，*2018*）。根据 *Besson*和 *Kaufmann*（*2018*）中的定理*4*可知，遗憾上界保留了原有的 $T^{2/3}\log(T)^{1/3}$ 依赖关系，仅常数因子有所变化。

## 5 实验

本节中，我们在非单调、子模和非子模设置下对 RGL 算法进行实证评估。关于更多实验，请参阅附录 D.1中的线性奖励减成本实验，以及附录D.2中的社交网络上的收入最大化实验。

我们将方法与精确最优解进行比较，并计算不同重复次数下累积全反馈遗憾的经验均值，而非累积1 2-遗憾，¯其定义如下：

$$\bar{\mathcal{R}}(T) = \frac{1}{rep}\sum_{n=1}^{rep}\sum_{t=1}^{T}(f(OPT) - f(S_t)).$$

我们在所选集合 $S$ 的一个非单调随机子模函数上测试这些算法，该函数定义为 $f(S)$ =min(max($g(S)$ + $\varepsilon$, 0), 1)，其中 $\varepsilon \sim \mathcal{N}(\mu, \sigma)$。在实验中，我们选择了一个非单调子模示例 $g(S)$，其中 $g(\{\}) = 0.2, g(\{1\}) = 0$, $g(\{2\})$ =0。6, $g(\{1,2\}) = 0.2$。注意，$\mathbb{E}[f(S)] = g(S)$, 且 $g(S)$ 是子模的。

在第二个实验中，我们选择了一个非单调非次模示例 $g(S)$, 其中 $g(\{\}) = 0.3, g(\{1\}) = 0, g(\{2\}) = 0.5, g(\{1,2\})$ $= 0.9$。请注意，$\mathbb{E}[f(S)] = g(S)$，且 $g(S)$ 不是子模的。

我们在 $T \in \{10^2, 10^3, 10^4, 10^5, 10^6\}$ 个时间范围内运行该方法。我们假设 $\varepsilon \sim \mathcal{N}(0, 0.1)$。我们对实验进行 $rep = 20$ 次重复并取平均值。我们对大小为 50 的窗口内的瞬时奖励取平均值。

我们使用最优解，该解在第一次实验中为 $\{2\}$，在第二次实验中为 $\{1,2\}$，并在在线设置中运行它，其中最优智能体（OPT）仅利用

直到 $T$始终选择最佳臂集合，见算法 2。此外，我们将其与随机多臂老虎机（RND）进行比较，见算法 3；在每个时间步，该算法从 Ω 中随机选择一个子集，其中每个臂以$_1$ $_2$的概率独立采样。离线设置中的随机算法具有$_1$ $_4$近似保证（Feige 等人，2011）。此外，我们还将其与一种在线单调子模最大化算法 ETCG 进行比较（Nie 等人，2022）。遗憾的是，单调子模最大化的在线算法需要额外的输入，即基数 $k$。因此，我们定义了 R-ETCG，见算法 4：该算法首先生成一个随机 $k \sim \mathcal{U}(0, n)$，然后找到最佳 $k$ 臂。

从子模函数情形的图 1 可以看出，RGL 达到了最优值。从图 2 可以看出，在非子模情形下，RGL 仍然达到了最优值。在这两项实验中，RGL 的表现均优于上述所有基准。尽管理论尚未针对非子模情形展开，但该方法在这类情形下仍然表现良好。此外，所提出的算法优于 R-ETCG，这表明针对单调函数的算法不能直接应用于非单调情形。

**注 4.** $O(T^{2/3})$ 的累积遗憾上界取决于时域 $T$（而非时间步 $t$）（参见所有图中左侧的子图，其中累积遗憾曲线随 $T$ 增长）。对于固定的时间范围 *T*，*RGL* 找到了最优臂集合，因此在固定时间范围 *T* 下，其累积遗憾相对于时间为常数 $t$（图 *1* 和图 *2* 右侧的子图）。此外，理论保证针对的是最坏情形，即理论给出了遗憾的上界，而对于某些实例，遗憾会更低。

## 6 结论

本文提出了 RGL，这是首个在全臂反馈下实现在线随机非单调子模最大化的算法；也就是说，智能体只能获得所选臂集合的奖励，而无法获得关于单个臂的额外信息。理论上，该算法在时域为 $T$、臂数为 $n$ 时，可达到 $\tilde{\mathcal{O}}(nT^{\frac{2}{3}})$ 的$_1$ $_2$-遗憾上界。此外，实验结果表明，该算法在全臂反馈下优于所考察的基线方法。

值得注意的是，现有关于全老虎机反馈下子模老虎机问题的研究，在单调函数设置中也达到了 $\tilde{\mathcal{O}}(T^{\frac{2}{3}})$ 遗憾上界（Nie 等人，2022年；Niazadeh 等，2020年）。此外，在对抗性设置下针对非单调函数的研究结果也是 $\tilde{\mathcal{O}}(T^{\frac{2}{3}})$（Niazadeh 等，2020年）。尽管尚未针对这一设置研究正式的下界，但证明这样的下界或改进所有这些设置下的遗憾界仍是一个开放问题。

## 7 致谢

本研究部分得到了美国国家科学基金会资助，资助项目编号为 2149588 和 2149617。

### 参考文献

Shipra Agrawal 和 Navin Goyal。多臂老虎机问题中汤普森采样的分析。载于学习理论会议论文集，第 39–1 页。《机器学习研究杂志》研讨会与会议论文集，2012年。Georgios Amanatidis、Federico Fusco、Philip Lazos、Stefano Leonardi 和 Rebecca Reiffenh¨auser。受背包约束的快速自适应非单调子模最大化。《神经信息处理系统进展》，33：16903–16915，2020年。

Peter Auer。利用置信界权衡利用与探索。《机器学习研究期刊》，3(Nov):397–422，2002年。

Peter Auer 和 Ronald Ortner。重新审视置信上界：随机多臂老虎机问题的改进遗憾界。《匈牙利数学期刊》，61(1-2):55–65，2010年。

Peter Auer、Nicolo Cesa-Bianchi 和 Paul Fischer。多臂老虎机问题的有限时间分析。《机器学习》，47(2):235–256，2002a。

Peter Auer、Nicolo Cesa-Bianchi、Yoav Freund 和 Robert E Schapire。非随机多臂老虎机问题。*SIAM*计算杂志，32(1):48–77，2002b。

Lilian Besson 和 Emilie Kaufmann。倍增技巧对多臂老虎机问题能做什么、不能做什么。*arXiv*预印本 *arXiv:1803.06971*，2018年。

Niv Buchbinder 和 Moran Feldman。子模最大化问题的确定性算法。《*ACM* 算法汇刊》（*TALG*），14(3):1–20，2018年。

Niv Buchbinder、Moran Feldman、Joseph Seffi 和 Roy Schwartz。无约束子模最大化的紧致线性时间 (1/2)-近似。*SIAM* 计算期刊，44(5):1384–1402，2015年。

Lin Chen、Mingrui Zhang、Hamed Hassani 和 Amin Karbasi。黑盒子模最大化：离散和连续情形。载于国际人工智能与统计会议，第 1058–1070 页。PMLR，2020年。

Abhimanyu Das 和 David Kempe。线性回归中的子集选择算法。载于第四十届 *ACM* 计算理论研讨会论文集，第 45–54 页，2008年。

Anirban Dasgupta、Ravi Kumar 和 Sujith Ravi。通过子模性和离散度进行摘要。载于协会第*51*届年会论文集

计算语言学（第*1*卷：长篇论文），第1014–1022页，保加利亚索非亚，2013年8月。计算语言学协会。

Jack Edmonds。子模函数、拟阵与某些多面体。载于组合优化——尤里卡，你缩小了！，第11–26页。施普林格，2003。Ethan R Elenberg、Rajiv Khanna、Alexandros G Dimakis和Sahand Negahban。受限强凸性蕴含弱子模性。统计年鉴，46 (6B)：3539–3568，2018年。Matthew Fahrbach、Vahab Mirrokni和Morteza Zadi-moghaddam。具有近乎最优自适应性和查询复杂度的非单调子模最大化。*arXiv*预印本 *arXiv:1808.06932*，2018年。Uriel Feige。用于近似集合覆盖的ln n阈值。*ACM*学报（*JACM*），45(4)：634–652，1998。Uriel Feige、Vahab S Mirrokni和Jan Vondr´ak。最大化非单调子模函数。*SIAM*计算杂志，40(4)：1133–1153，2011。Michel X Goemans和David P Williamson。使用半正定规划改进最大割和可满足性问题的近似算法。*ACM*学报（*JACM*），42(6)：1115–1145，1995。Aditya Gopalan、Shie Mannor和Yishay Mansour。复杂在线问题的汤普森采样。载于第*31*届国际机器学习会议论文集，第32卷机器学习研究论文集，第100–108页，中国北京，2014年6月22–24日。PMLR。Wassily Hoeffding。有界随机变量之和的概率不等式。载于*Wassily Hoeffding*文集，第409–426页。施普林格，1994。Satoru Iwata、Lisa Fleischer和Satoru Fujishige。用于子模函数最小化的组合强多项式算法。*ACM*学报（*JACM*），48 (4)：761–777，2001。Rajiv Khanna、Ethan Elenberg、Alex Dimakis、Sahand Ne-gahban和Joydeep Ghosh。通过弱子模性实现可扩展的贪心特征选择。载于人工智能与统计，第1560–1568页。PMLR，2017。Tor Lattimore和Csaba Szepesv´ari。老虎机算法。剑桥大学出版社，2020年。Siyu Lei、Silviu Maniu、Luyi Mo、Reynold Cheng和Pierre Senellart。在线影响力最大化。载于第*21*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第645–654页，2015。Shuai Li、Fang Kong、Kejie Tang、Qizhi Li和Wei Chen。线性阈值模型下的在线影响力最大化。神经信息处理系统进展，33：1192–1204，2020年。Wei Lu和Laks VS Lakshmanan。社交网络上的利润最大化。载于*2012 IEEE*第*12*届国际数据挖掘会议，第479–488页。IEEE，2012年。

Baharan Mirzasoleiman、Ashwinkumar Badanidiyuru 和 Amin Karbasi。《快速约束子模最大化：个性化数据摘要》。载于国际机器学习会议，第 1358–1367 页。PMLR，2016年。

Rad Niazadeh、Negin Golrezaei、Joshua Wang、Fransisca Susan 和 Ashwinkumar Badanidiyuru。通过离线贪心进行在线学习：在市场设计和优化中的应用。*EC 2021*，《管理科学》期刊，2020年。Guanyu Nie、Mridul Agarwal、Abhishek Kumar Umrawal、Vaneet Aggarwal 和 Christopher John Quinn。一种用于全老虎机反馈下子模最大化的先探索后承诺算法。载于第三十八届人工智能不确定性会议，2022年。Pierre Perrault、Jennifer Healey、Zheng Wen 和 Michal Valko。预算约束在线影响力最大化。载于国际机器学习会议，第7620–7631页。PMLR，2020年。Sharon Qian 和 Yaron Singer。统计子集选择问题的快速并行算法。神经信息处理系统进展，第32卷，2019年。Lijing Qin 和 Xiaoyan Zhu。通过熵正则项提升推荐系统的多样性。载于第二十三届国际人工智能联合会议。Citeseer，2013年。Herbert Robbins。序贯实验设计的若干方面。《美国数学学会通报》，58(5):527–535，1952年。Tim Roughgarden 和 Joshua R. Wang。一种用于在线无约束子模最大化的最优学习算法。载于第*31*届学习理论会议论文集，第1307–1325页，2018年。Sho Takemori、Masahiro Sato、Takashi Sonoda、Janmajay Singh 和 Tomoko Ohkuma。多重约束下的子模老虎机问题。载于人工智能不确定性会议，第191–200页。PMLR，2020年。William R Thompson。根据两组样本的证据，一个未知概率超过另一个未知概率的可能性。《生物计量学》，25(3-4):285–294，1933年。Sebastian Tschiatschek、Rishabh K Iyer、Haochen Wei 和 Jeff A Bilmes。学习用于图像集合摘要的子模函数混合模型。神经信息处理系统进展，第27卷，2014年。Sharan Vaswani、Branislav Kveton、Zheng Wen、Mohammad Ghavamzadeh、Laks VS Lakshmanan 和 Mark

Schmidt。用于影响力最大化的模型无关在线学习。载于国际机器学习会议，第3530–3539页。PMLR，2017。

Zheng Wen、Branislav Kveton、Michal Valko 和 Sharan Vaswani。具有半老虎机反馈的独立级联模型下的在线影响力最大化。神经信息处理系统进展，30，2017。

# A. 非单调子模最大化的动机示例

## A.1 数据摘要

随着每天产生海量数据，如何从数据点中选出优质的代表性子集仍是一项挑战。通常，用于衡量整个数据集的子集覆盖度或多样性的效用函数具有子模性（Mirzasoleiman 等人，2016）。然而，能够兼顾多样性的效用函数不一定是单调的，因为它们会对较大的解集施加惩罚（Tschiatschek 等人，2014；Dasgupta 等人，2013）。

## A.2 特征选择

非单调子模最大化算法的一个重要应用，是对特征选择等学习问题进行建模（Das 和 Kempe，2008；Khanna 等人，2017年；Elenberg 等人，2018年；Qian 和 Singer，2019年）。由于可能会对训练数据过拟合，特征选择优化可以建模为非单调子模最大化问题（Fahrbach 等人，2018年）。

## A.3 推荐系统

推荐包含冗余信息的项目会导致效用的边际收益递减。通过上下文子模组合赌博机框架，研究人员对向用户依次推荐项目集合这一问题进行了研究（Qin 和 Zhu，2013年；Takemori 等人，2020年）。这种优化不一定是单调的，因为添加更多推荐内容可能会产生反作用（Amanatidis 等人，2020年）。

## A.4 影响力最大化

营销新开发产品的一种可行方式，是选出一组极具影响力的人，并希望他们将该产品推荐给各自的社区。近期一系列研究将这一问题视为多臂老虎机问题（带有额外反馈），且不要求了解网络和扩散模型（Lei 等人，2015年；Wen 等人，2017；Vaswani 等人，2017；Li 等人，2020年；Perrault 等人，2020年）。大多数研究都假设基数或预算存在固定约束。然而，从影响力中获得的收入减去成本，以实现收入最大化的模型，通常是一个非单调的无约束子模最大化问题（Lu 和 Lakshmanan，2012年）。

# B 附加引理与证明

## B.1 CleanEvent 的概率

Hoeffding 不等式（Hoeffding，1994）是一种用于界定有界随机变量概率的强大方法。我们陈述该不等式，然后利用它证明 $\mathcal{E}$ 以高概率发生。

**引理 3.**（***Hoeffding*** 不等式）。设 $X_1$、$X_2$、.....、$X_n$ 是取值有界于 $[0,1]$ 的独立随机变量，且 $X$ 是它们的经验均值。那么，对于任意 $\varepsilon > 0$，有

$$\mathbb{P}(|\bar{X} - \mathbb{E}(X)| \geq \varepsilon) \leq 2\exp\left(-2n\varepsilon^2\right).$$

**引理 4.** 干净事件 $\mathcal{E}$ 的概率满足

$$\mathbb{P}(\mathcal{E}) \geq 1 - \frac{8n}{T^4}. \tag{27}$$

证明。 将引理 3 应用于动作 $S$ 的 $m$ 次奖励的经验均值 $\bar{f}(S)$，并令 $\epsilon = \text{rad} = \sqrt{2\log(T)/m}$，我们得到

$$\begin{aligned}
\mathbb{P}\left[\left|\bar{f}(S) - f(S)\right| \geq \text{rad}\right] &\leq 2\exp\left(-2m\,\text{rad}^2\right) && \text{(by Lemma 3)} \\
&= 2\exp(-2m(2\log(T)/m)) \\
&= 2\exp(-4\log(T)) \\
&= \frac{2}{T^4}.
\end{aligned} \tag{28}$$

对于每个臂 $u_i$，智能体恰好执行以下动作列表 $\mathcal{S}_i = [X_{i-1}, X_{i-1} \cup \{u_i\}, Y_{i-1}, Y_{i-1} \setminus \{u_i\}]$ $m$ 次，然后计算边际增益估计。因此，对于任意单个动作 $S \in \mathcal{S}_i$，我们可以将其样本均值 $\bar{f}(S)$ 落在指定置信半径内（即上述事件的补事件）的概率界定为

$$\begin{aligned}\forall S \in \mathcal{S}_i \quad \mathbb{P}\left[\left|\bar{f}(S) - f(S)\right| < \mathrm{rad}\right] &= 1 - \mathbb{P}\left[\left|\bar{f}(S) - f(S)\right| \geq \mathrm{rad}\right] \\ &\geq 1 - \frac{2}{T^4}. \end{aligned} \tag{29}$$

现在我们着重对 $\mathbb{P}(\mathcal{E}_i \mid X_{i-1} = X,\ Y_{i-1} = Y)$ 进行界定。通过对上一阶段确定的集合进行条件化，$X_{i-1} = X$，$Y_{i-1} = Y$，我们知道当前阶段 $i$ 将执行的所有动作，即 $\mathcal{S}_i$。所有动作的奖励均有界于 $[0,1]$，且在给定相应动作的条件下相互独立。

$$\begin{aligned}\mathbb{P}(\mathcal{E}_i \mid X_{i-1} = X, Y_{i-1} = Y) &= \mathbb{P}\left(\bigcap_{S \in \mathcal{S}_i} \left\{\left|\bar{f}(S) - \mathbb{E}\left[f(S)\right]\right| < \mathrm{rad}\right\} \mid X_{i-1} = X, Y_{i-1} = Y\right) && \text{(by (10))} \\ &= \prod_{S \in \mathcal{S}_i} \mathbb{P}\left(\left\{\left|\bar{f}(S) - f(S)\right| < \mathrm{rad}\right\} \mid X_{i-1} = X, Y_{i-1} = Y\right) \\ & && \text{(rewards are independent when conditioned on actions)} \\ &\geq \left(1 - \frac{2}{T^4}\right)^{|\mathcal{S}_i|} && \text{(by (29))} \\ &= \left(1 - \frac{2}{T^4}\right)^4 && (30)\end{aligned}$$

有了这一点，我们便可以给出干净事件 $\mathcal{E}$的概率下界，

$$\begin{aligned}\mathbb{P}(\mathcal{E}) &= \mathbb{P}(\mathcal{E}_1 \cap \cdots \cap \mathcal{E}_n) && \text{(by (11))} \\ &= \prod_{i=1}^{n} \mathbb{P}(\mathcal{E}_i \mid \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}) \\ &= \prod_{i=1}^{n} \sum_{X,Y} \mathbb{P}(X_{i-1} = X, Y_{i-1} = Y, \mathcal{E}_i \mid \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}) && \text{(law of total probability)} \\ &= \prod_{i=1}^{n} \sum_{X,Y} \mathbb{P}(X_{i-1} = X, Y_{i-1} = Y \mid \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}) \times \mathbb{P}(\mathcal{E}_i \mid X_{i-1} = X, Y_{i-1} = Y, \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}) \\ &= \prod_{i=1}^{n} \sum_{X,Y} \mathbb{P}(X_{i-1} = X, Y_{i-1} = Y \mid \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}) \times \mathbb{P}(\mathcal{E}_i \mid X_{i-1} = X, Y_{i-1} = Y) \\ &\geq \prod_{i=1}^{n} \sum_{X,Y} \mathbb{P}(X_{i-1} = X, Y_{i-1} = Y \mid \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}) \times \left(1 - \frac{2}{T^4}\right)^4 && \text{(by (30))} \\ &= \prod_{i=1}^{n} \left(1 - \frac{2}{T^4}\right)^4 \sum_{X,Y} \mathbb{P}(X_{i-1} = X, Y_{i-1} = Y \mid \mathcal{E}_1, \ldots, \mathcal{E}_{i-1}) \\ &= \prod_{i=1}^{n} \left(1 - \frac{2}{T^4}\right)^4 \\ &= \left(1 - \frac{2}{T^4}\right)^{4n} \\ &\geq \left(1 - \frac{8n}{T^4}\right). && \text{(Bernoulli's inequality)}\end{aligned}$$

□

**B.2 引理2中情形2的证明。**

只需证明在任意形如 $X_{i-1} = S_{i-1}$ where $S_{i-1} \subseteq \{u_1, ..., u_{i-1}\}$ 的事件条件下，该不等式成立，其中概率 $X_{i-1} = S_{i-1}$ 非零。以下证明均假设是在这一事件条件下进行的。正文将引理 2 的证明分为 4 种情形，并在正文中给出了其中三种情形的详细证明。这里给出情形2的证明，以求完整。

证明。 **情形2 (**$\bar{a}_i < 0$ 和 $\bar{b}_i \geq 0$**)**：在此情形下，$\bar{a}_i \leq 0 \Rightarrow a'_i = 0 \Rightarrow \frac{b'_i}{a'_i + b'_i} = 1$。因此，$X_i = X_{i-1}$ 和 $Y_i = Y_{i-1} \setminus \{u_i\}$。由于 $Y_i = Y_{i-1} \setminus \{u_i\}$，我们有

$$\begin{aligned} \mathbb{E}[b_i] &= \mathbb{E}[f(Y_{i-1} \setminus \{u_i\}) - f(Y_{i-1})] \\ &= \mathbb{E}[f(Y_i) - f(Y_{i-1})]. \end{aligned} \tag{31}$$

由于 $X_i = X_{i-1}$，我们要证明的关系（13）可化简为：

$$\mathbb{E}[f(OPT_{i-1}) - f(OPT_i)] \leq \frac{1}{2}\mathbb{E}[f(Y_i) - f(Y_{i-1})] + 5rad.$$

注意，

$$OPT_i = (OPT \cup X_i) \cap Y_i = OPT_{i-1} \setminus \{u_i\} \tag{32}$$

若 $u_i \notin OPT \Rightarrow OPT_i = OPT_{i-1}$。因此，

$$\begin{aligned} \mathbb{E}[f(OPT_{i-1}) - f(OPT_i)] &= 0 \\ &\leq \frac{\bar{b}_i}{2} && \text{（根据情形2的条件）} \\ &\leq \frac{\mathbb{E}[b_i]}{2} + rad && \text{(using concentration)} \\ &= \frac{1}{2}\mathbb{E}[f(Y_i) - f(Y_{i-1})] + rad && \text{(by (31))} \\ &\leq \frac{1}{2}\mathbb{E}[f(Y_i) - f(Y_{i-1})] + 5rad. \end{aligned}$$

现在考虑 $u_i \in OPT$。根据 $OPT_{i-1}$ 的定义，$X_{i-1} \subseteq OPT_{i-1}$。由于 $u_i \notin X_{i-1}$。于是，$X_{i-1} \subseteq OPT_{i-1} \setminus \{u_i\}$。因此，根据期望意义下的子模性，

$$\mathbb{E}[f(X_{i-1} \cup \{u_i\})] - \mathbb{E}[f(X_{i-1})] \geq \mathbb{E}[f(OPT_{i-1} \setminus \{u_i\}) \cup \{u_i\}))] - \mathbb{E}[f(OPT_{i-1} \setminus \{u_i\})]. \tag{33}$$

这样我们就可以完成该界的证明：

$$\begin{aligned} \mathbb{E}[f(OPT_{i-1})] - \mathbb{E}[f(OPT_i)] &= \mathbb{E}[f(OPT_{i-1})] - \mathbb{E}[f(OPT_{i-1} \setminus \{u_i\})] && \text{(by (32))} \\ &= \mathbb{E}[f(OPT_{i-1} \setminus \{u_i\}) \cup \{u_i\}))] - \mathbb{E}[f(OPT_{i-1} \setminus \{u_i\})] \\ &\leq \mathbb{E}[f(X_{i-1} \cup \{u_i\})] - \mathbb{E}[f(X_{i-1})] && \text{(by (33))} \\ &= \mathbb{E}[a_i] && \text{(by def. of } a_i\text{)} \\ &\leq \bar{a}_i + 2rad && \text{(using concentration)} \\ &\leq \frac{\bar{b}_i}{2} + 2rad && \text{(condition for case 2)} \\ &\leq \frac{1}{2}\mathbb{E}[b_i] + 3rad && \text{(using concentration)} \\ &= \frac{1}{2}\mathbb{E}[f(Y_i) - f(Y_{i-1})] + 3rad && \text{(by (31))} \\ &\leq \frac{1}{2}\mathbb{E}[f(Y_i) - f(Y_{i-1})] + 5rad. \end{aligned}$$

□

# C 基准

现在，我们将讨论用于评估所提出算法性能的基准。

### C.1 最优老虎机

最优老虎机（OPT）需要将最优臂集合作为输入，并且在直到$T$的整个时间内仅利用该集合，见算法 2。最优臂集合应预先已知，或通过某种离线算法找到。

**算法 2**OPT

**要求：** 时域 $T$，解 $S^\star$
  **对**时间步 $t \in \{1, \dots, T\}$ **执行**
    选择 $S^\star$
  **结束循环**

### C.2 随机老虎机

随机多臂老虎机（RND）在每个时间步随机选择 Ω 的一个子集，其中每个臂都以概率 $\frac{1}{2}$ 独立采样，见算法 3。离线设置中的随机算法具有 $\frac{1}{4}$-近似保证（Feige 等人，2011）。

**算法 3**RND

**要求：** 基础臂集合 Ω，时域 $T$
  $n \leftarrow |\Omega|$
  **对于**时间步 $t \in \{1, \dots, T\}$ **执行**
    $S^{(t)} \leftarrow \emptyset$
    **对于** $i \in \{1, \dots, n\}$ **执行**
      **以概率** $\frac{1}{2}$ **执行**
        $S^{(t)} \leftarrow S^{(t)} \cup \{u_i\}$
    **结束循环**
    选择 $S^{(t)}$
  **循环结束**

#### C.3 R-ETCG 多臂老虎机

探索后承诺贪心（ETCG）（Nie 等人，2022年）是一种针对全老虎机反馈下单调子模最大化的在线算法，并且在单调设置下具有已证明的保证。只有在存在约束时，子模单调最大化才有意义；否则，只要添加臂总是有益的，智能体就会选择所有臂。因此，单调子模最大化的在线算法至少需要一个额外输入，例如基数约束 $k$。为了使其适用于我们的无约束非单调设置，我们定义了随机 ETCG（R-ETCG）：它首先随机生成一个基数预算 $k \sim \mathcal{U}\{0, n\}$，然后找到最佳的 $k$ 个臂，见算法 4。

# D 更多实验评估

本节中，我们在另一种非单调设置下对 RGL 算法进行实证评估。我们将该方法与精确最优解进行比较，并计算不同重复次数下累积全反馈遗憾的经验均值，而非累积 $\frac{1}{2}$-遗憾，定义如下，

$$\bar{\mathcal{R}}(T) = \frac{1}{rep} \sum_{n=1}^{rep} \sum_{t=1}^{T} (f(OPT) - f(S_t)).$$

我们测试 $n = 8$ 个基臂，时域为 $T \in \left\{10^2, 10^3, 10^4, 10^5, 10^6\right\}$。我们对 $rep = 9$ 次重复实验取平均。我们对窗口大小为 50 的窗口内的即时奖励取平均。

**Alg算法 4**R-ETCG

**要求：**基臂集合 Ω，时域 $T$初始化 $S^{(0)} \leftarrow \emptyset$，$n \leftarrow |\Omega|$，$k \leftarrow \mathcal{U}\{0, n\}$初始化 $m \leftarrow \left\lceil \left( \frac{T\sqrt{2}}{n+2nk\sqrt{2\log(T)}} \right)^{2/3} \log(T) \right\rceil$ **对于**阶段 $i \in \{1, \ldots, k\}$ **执行以下操作：对于**臂 $a \in \Omega \backslash S^{(i-1)}$ **执行**游玩 $S^{(i-1)} \cup \{a\}$ $m$ 次，计算经验均值 $\bar{f}\left(S^{(i-1)} \cup \{a\}\right)$ **结束循环**$a_i \leftarrow \arg\max_{a \in \Omega \backslash S^{(i-1)}} \bar{f}\left(S^{(i-1)} \cup \{a\}\right)$ $S^{(i)} \leftarrow S^{(i-1)} \cup \{a_i\}$**结束循环；对于**剩余时间 **执行**游玩 $S^{(k)}$**结束循环**

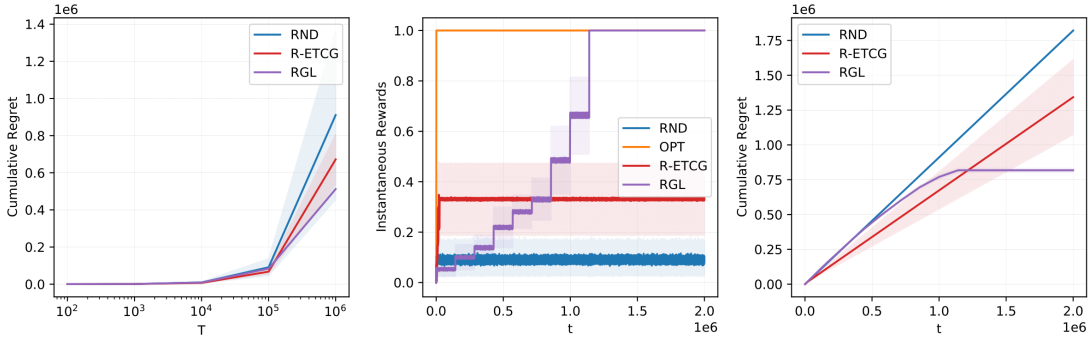


图 3：非单调随机奖励函数的比较结果。从左到右，图中分别展示了累积遗憾随时间步 $T$ 的变化、即时奖励随时间步 $t$ 的变化，以及累积遗憾随时域 $t$ 的变化。

### D.1 线性奖励减成本

我们测在所选集合 $X$上的非单调随机函数上测试算法，其定义如下　　s,

$$f(X) = \begin{cases} 1, & \text{if } X = \{5, 6, 7, 8\} \\ min(max(\sum_{a \in X} r(a) - \frac{|X|}{k^\star}, 0), 1), & \text{otherwise} \end{cases}$$

其中，$r(a)$ 是单个臂的随机奖励函数，其中 $\forall X$，$\forall a \in X$，$r(a) \in [0, 1]$。事实上，我们选择$r(a) \sim min(max(\mathcal{N}(\mu_a, \sigma), 0), 1)$，其中 $\forall X$，$\forall a \in X$，$\mu_a \in [0, 1]$。
我们固定子模函数 $k^\star = 6$ 的预言机常数。我们选择 $\sigma = 0.02$，以及由所有 $\mu_a$ 构成的向量 $\mu$，例如 $\mu$ 的取值从 0 到 0.35，步长为 0.05。可以很容易验证，集合 $\{5, 6, 7, 8\}$ 是最优臂子集。

从图 3 可以看出，在上述定义的基准中，我们提出的算法 RGL 是唯一达到最优值的算法（中间图），并且就时域 T 而言，其累积遗憾最小（左图）。就时间步 t 而言（右图），与 RND 类似，RGL 起初的累积奖励高于 R-ETCG，这可以用 RGL 较长的早期探索阶段来解释；然而，在后续时间步中，RGL 通过利用最优臂集合，取得了比 R-ETCG 更低的累积遗憾，表现优于 R-ETCG。

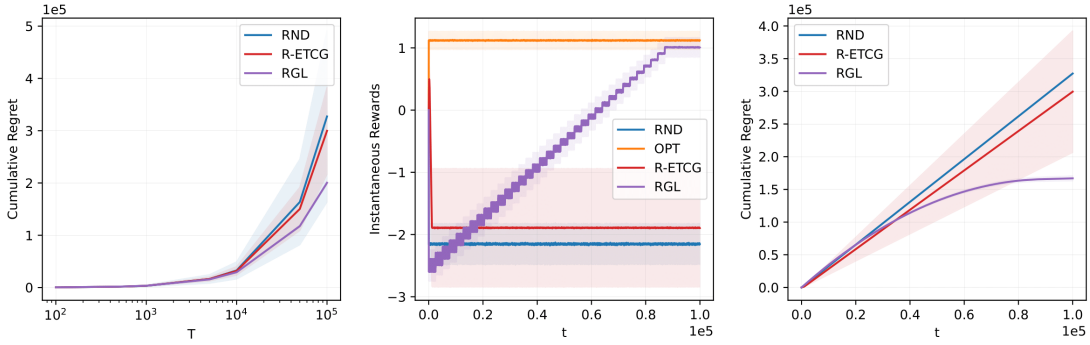


图 4：社交网络上的收入最大化。从左到右，图中分别显示累积遗憾随时间步 $T$ 的变化、即时奖励随时间步 $t$ 的变化，以及累积遗憾随时域 $t$ 的变化。

### D.2 社交网络上的收入最大化

在若干现实场景中，非单调目标函数更有意义。例如，在社交网络上的收入最大化问题中，相较于仅优化影响力（单调）并将预算作为约束，优化总收入（影响力减去成本；非单调）更有意义。前一种方案会用完全部预算，而收入最大化方案可能只会使用其中一部分。

我们通过影响力最大化减去成本，在社交网络上测试 RGL 的非单调收入最大化表现。影响力最大化实际上是一个次模最大化问题；减去添加节点的成本后，该问题变为非单调（附录 A.4）。

我们使用 Kar ate 网络，其中包含 34 个节点，并使用预言机函数 $f$；对于节点子集 $S$ ，

$$f(S) = \mathcal{N}(\sum_{c \in \mathcal{C}} \max_{a \in S \cap c} d(a), \sigma) - \alpha |S|,$$

其中， $\mathcal{C}$ 表示社区集合， $d(a)$表示节点 a 的度， $\mathcal{N}$ 表示正态分布， $\alpha$ 是取决于成本的正数常量。如图 4 所示，RGL 的表现优于所有其他算法，因为它的累积遗憾最低，并且几乎达到了最优即时奖励。