多臂 老虎机问题与广义时间分区奖励 s

Ronald C. van den Broek[1] , Rik Litjens[1] , Tobias Sagis[1] , Luc Siecker[1] , Nina Verbeeke[1] 和 Pratik Gajane[1]

[1]埃因霍温理工大学 r.c.v.d.broek@student.tue.nl, r.litjens@student.tue.nl,t.g.m.sagis@student.tue.nl, l.r.siecker@student.tue.nl, n.c.verbeeke@student.tue.nl, p.gajane@tue.nl

arXiv:2303.00620v1 [cs.LG] 1 Mar 2023

摘要

具有序贯性质的决策问题，即过去所做的决策可能会影响未来，常用于为许多实际重要的应用建模。在一些现实世界应用中，关于某项决策的反馈会延迟到达，并可能以部分奖励的形式，通过不同的延迟时间被观测到。受此类场景启发，我们提出了一种新的问题表述，称为具有广义时间分区奖励的多臂老虎机。为形式化描述一项决策的反馈如何分布于多个时间步，我们引入 β-离散度性质。我们推导出所考虑问题中任何一致高效算法的性能下界。此外，我们提出了一种名为 TP-UCB-FR-G 的算法，并证明了其性能度量的上界。在某些场景下，我们的上界优于最先进水平。我们还提供实验结果，以验证所提出的算法和理论结果。

1 引言

经典多臂老虎机（MAB，或简称老虎机问题）是用于建模序贯决策的框架 [Bubeck 和 Cesa-Bianchi，2012a]。在多臂老虎机问题中，学习智能体面对一组有限的 K 臂，智能体所做的决策以拉动某个臂表示。智能体通过数值奖励获得有关所做决策的反馈。多臂老虎机文献通常关注这样一种场景：假设拉动某个臂后会立即获得奖励。相比之下，延迟反馈多臂老虎机的相关研究（例如，[Joulani 等，2013；Mandel 等人， 2015]）假设拉动某个臂与观察到相应奖励之间存在延迟。在这些研究中，假设奖励集中在单个延迟轮次中。通过允许将奖励划分为以不同延迟观察到的部分奖励，可以扩展这一设置。这类老虎机问题称为具有时间分区奖励的多臂老虎机（TP-MAB），由 [Romano 等人， 2022] 首次提出。

在 TP-MAB 设置中，智能体会在多个轮次中收到奖励的子集。某个臂的累积奖励是拉动该臂所获得的部分奖励之和。 [Romano 等人， 2022] 提出了 α-平滑性来描述奖励结构。α-平滑性指出，一组连续部分奖励中的最大奖励不能超过最大奖励的一定比例（精确定义见定义 2）。然而，如果累积奖励并非均匀分布， α-平滑性的假设就不太适用。本文提出了一种更通用的方式，用于描述某个臂的延迟累计奖励如何分布在多个轮次中。

作为一个应用示例，我们可以考虑提供大规模开放在线课程（MOOC）的网站（例如 Coursera、Khan Academy 和 edX）。这类网站旨在为用户推荐有用的课程。这个问题可以建模为 TP-MAP问题。一门课程由一系列视频讲座组成，可以视作一条臂。智能体可以向用户推荐一门课程，这相当于拉动一条臂。当学生学习一门课程时，智能体可以观察到部分奖励（例如，通过查看观看时长留存率）。在这种设置下， α-平滑性很少能反映实际的累积奖励分布。许多学生会观看课程开头的视频讲座，却从不看完最后几讲，因此部分奖励的离散度并不均匀。因此，关于延迟反馈多臂老虎机的现有研究及 [Romano 等人提出的算法 2022] 可能无法为用户推荐相关课程。受这类场景启发，我们探讨了一种更通用的奖励结构表述方式。

我们的贡献

1. 我们提出了一种新颖的多臂老虎机表述，以一种通用方式描述某个臂的延迟累计奖励如何分布在各轮次中。2. 我们证明了所考虑问题中任意一致高效算法的性能度量下界。3. 我们设计了算法 TP-UCB-FR-G，并证明了其性能度量的上界。在某些场景下，所证明的上界比最先进水平更紧。

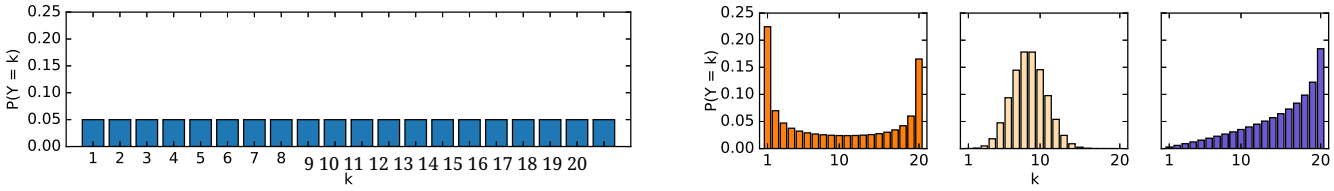

图 1：概率分　　布 α-平滑性（左）以及使用我们的方法得到的可能概率分布 β-离散度（右）

4. 我们提供了实验结果，验证了理论结果的正确性以及所提算法的有效性。

2 背景与相关工作

延迟反馈在线学习是文献中一个经过充分研究的问题。受篇幅限制，关于这一主题的工作必然无法一一列举，其中包括[Weinberger 和 Ordentlich, 2002; Agarwal 和 Duchi, 2011; Mesterharm, 2005; Mesterharm, 2007; Desautels 等人, 2014; Zinkevich 等人，2009]。本节其余部分将重点讨论具有延迟反馈的多臂老虎机。

[Joulani 等人，2013] 研究了非匿名延迟反馈老虎机问题，并提出了一种 UCB 算法的变体 [Auer 等人, 2002] 作为解决方案。在 [Joulani 等人，2013], 的研究中，假设能够获知哪个动作产生了特定的延迟奖励。[Wang 等人，2021] 将该问题扩展到包含匿名反馈的情形，并且无需准确掌握奖励区间的先验知识。近期，研究者在不同于本文的多臂老虎机设置中探讨了多种延迟反馈场景，例如线性老虎机和上下文老虎机 [Arya 和 Yang, 2020; Zhou 等人, 2019; Vernade 等人 , 2020a]，非平稳老虎机[Vernade 等人, 2020b]。此外，[Pike-Burke 等人，2018] 和[Cesa-Bianchi 等人，2022]研究了延迟、聚合且匿名的反馈情形。

过去大多数关于延迟多臂老虎机设置的研究都假设，某个臂的全部奖励会在某个有界延迟之后一次性观测到[Joulani 等，2013；Mandel 等人， 2015]；或者在来自具有有限期望值的无界分布的随机延迟之后观测到[Gael 等人，2020；Vernade 等人， 2017]。本文研究的是这样一种设置：某个臂的奖励会在一个区间内逐步发放，且最大延迟值有限。这与我们旨在建模的应用场景相符，例如第 1 节提到的大规模开放在线课程提供商。据我们所知， [Romano 等人， 2022], 最早分析了这种设置。他们提出了具有时间分区奖励的多臂老虎机（TP-MAB）设置。在 TP-MAB 设置中，拉动某个臂所获得的随机奖励会被分为若干部分奖励，并在拉动该臂后的有限轮次内陆续观测到。[Romano 等人， 2022] 假设臂奖励遵循α-平滑性性质（具体定义见定义 2）。如图 1 最左侧的直方图所示，如果某个臂的奖励是 α-平滑的，那么在一组连续轮次中观测到部分奖励的概率是均匀分布的。

尽管 [Romano 等人， 2022] 的研究在 TP-MAB 设置中取得了令人鼓舞的结果，但其依据的是 α-平滑性成立这一强假设。因此，他们提出的解决方案不适用于更广泛的场景，例如奖励以非均匀方式分区的情况。为此，我们提出使用一般分布，以更准确地刻画所获得的奖励如何分区。设想这样一种场景：我们能够获得有关累积奖励如何分布到各轮次的信息。第 1 节所述的大规模开放在线课程平台向用户推荐课程，就是这类场景的一个例子。通过泛化奖励结构，我们的方法能够处理每轮最大奖励并非在各轮次间均匀分区的情况，例如图 1 右侧所示的情况。

[Romano 等人， 2022] 提出了两种基于 UCB 算法的新算法，利用了 α-平滑性：TP-UCB-EW 和 TP-UCB-FR。这两种算法都将一个假设的 α 值作为输入，并用它计算类似于 UCB [Auer 等人， 2002]中的置信项。[Romano 等人， 2022] 指出，在时间范围较短时 T，TP-UCB-EW 算法表现更好；而从长远来看，TP-UCB-FR 的表现更出色。此外，他们还表明，TP-UCB-FR 算法的累积遗憾会受到假设 α的显著影响，而对于不同的 α 输入值，TP-UCB-EW 的累积遗憾变化相对较小。因此，我们认为，在泛化设置中利用假设分布时，TP-UCB-FR 的设置也最为合适。随后，我们将 TP-UCB-FR 作为所提算法的基线。

3 问题表述

考虑一个具有 K 个臂、时间范围为 T 轮次的多臂老虎机问题，其中 K、$T \in \mathbb{N}$。在每一轮 $t \in \{1, 2, ..., T\}$，从臂集合 $\{1, 2, ..., K\}$ 中选择一个臂。

对于所考虑的问题，算法 $\mathfrak{A}$ 在经过 T 个时间步后的性能可以使用期望遗憾（或简称遗憾）来衡量，记为 $\mathcal{R}_T(\mathfrak{A})$。

定义1。（遗憾）算法 $\mathfrak{A}$ 在经过 T 个时间步后的遗憾为 $\mathcal{R}_T(\mathfrak{A})$: $= \mu^* T - \sum_{i=1}^{K} \mu_i \cdot \mathbb{E}[N_i(T)]$,，其中 μ^* : $= \max_{1 \le i \le K} \mu_i$，且 $N_i(T)$ 表示截至时间 t 臂 i 被选择的次数。

总奖励在一组轮次中按时间分区

$T' = \{t, t+1, ..., t+\tau_{\max}-1\}$。令 $x^i_{t,m}(m \in T')$ 表示学习器在第m轮收到的分区奖励，该奖励是在第 t轮拉取臂 i 后产生的。智能体知道是哪次臂拉取产生了这份奖励。学习器至迟会在经过$\tau_{\max}$延迟后收齐累积奖励。每轮奖励 $x^i_{t,m}$ 是一个随机变量的实现，

随机变量 $X^i_{t,m}$，其支持集为 $[0, \overline{X}^i_m]$。学习器在第 t 轮拉动臂 i 所获得的累积奖励记为 r^i_t，它是随机变量 R^i_t 的一个实现，满足 R^i_t：$= \sum_{n=t}^{t+\tau_{\max}-1} X^i_{t,\,n}$ 其支持集为 $[0, \overline{R}^i]$。显然，我们观察到 $\overline{R}^i := \sum_{n=t}^{\tau_{\max}} \overline{X}^i_n$。

[Romano 等人， 2022] 已表明，在实践中，某个臂的每轮奖励能够提供该臂累积奖励的信息。 [Romano 等人， 2022] 引入了一个 α-平滑性属性（定义见定义 2），用于划分时间间隔奖励，使得每个分区都对应一组连续每轮奖励之和。形式化地，设 $\alpha \in T'$ 满足 α 是 $\tau_{\max}$ 的因子。每个分区的基数（下文称为“z-群”）记为 $\phi := \frac{\tau_{\max}}{\alpha}$，其中 $\phi \in \mathbb{N}$。与上文类似，现在可以将每个 z-群 $z^i_{t,\,k}, k \in \{1, 2, ..., \alpha\}$ 定义为随机变量 $Z^i_{t,\,k}$ 的一个实现，使得对于每个 k:

$$Z^i_{t,k} := \sum_{n=t+(k-1)\phi}^{t+k\phi-1} X^i_{t,n} \tag{1}$$

使得 $Z^i_{t,\,k}$ 的支持集为 $[0, \overline{Z}^i_{\alpha,k}]$。

定义 2（α-平滑性）。对于 $\alpha \in \{1, ..., \tau_{max}\}$，奖励 α是 $\frac{\tau_{max}}{\alpha} \in \mathbb{N}$ 平滑的，当且仅当 $i \in \{1, ..., K\}$并且对于每个 $k \in \{1, 2, ..., \alpha\}$ 和 Z^i_t，随机变量 $_{,\,k}$ 相互独立，且满足 $\overline{Z}^i_{\alpha,\,k} = \overline{Z}^i_\alpha = \frac{\overline{R}^i}{\alpha}$。

α-平滑性确保所有时间分区奖励都有助于约束同一窗口内未来奖励的值。如果满足 α-平滑性，那么 z-群 $\overline{Z}^i_{\alpha,\,k}$ 中的最大累积奖励对于所有 z-群 $k \in \{1, 2, ..., \alpha\}$ 都是相等的。因此，我们可以说， $\forall_k \in \{1, 2, ..., \alpha\}$、$\overline{Z}^i_{\alpha,\,k} = \overline{Z}^i_\alpha$。

然而，对于累积奖励未在各轮次之间均匀分区的场景， α-平滑性的假设并不适用。为适应此类场景，本文旨在推广奖励在z-群之间的离散度。为此，需要消除每个 z-群获得部分奖励的概率都相等这一假设。为实现这一点，我们用 β-离散度性质替代 α-平滑性，从而能够对累积奖励未在各轮次之间均匀分布的场景进行建模；也就是说，该性质允许 $\overline{Z}^i_{\alpha,\,k}$ 在不同z-群之间有所差异。

3.1 OurSolution 方法： β-离散度性质

我们不妨设，对于每个部分奖励，都有一个离散随机变量 Y，表示观察到该部分奖励时所属的 z-群的索引。因此， Y 可能取值的集合为 $\{1, 2, ...\alpha\}$。每个 Y 的可能取值都对应一个特定概率，该概率表示在 z-群 k 中观察到部分奖励的概率。这意味着奖励在 z-群上的分布可以是非均匀的。形式上，我们将这种 β-离散度定义如下:

定义 3（β-离散度）。设 Y为一个离散随机变量，表示观测到部分奖励的 z-群的索引。 Y具有离散概率分布 $\mathbb{F}$，其定义域大小有限，为 α；该分布由概率质量函数 $P_{\mathbb{F}}(k)$（$k \in \{1, , \ldots, \alpha\}$）定义。若臂 i的奖励满足以下条件，则称其为 β-离散：臂 i的部分奖励按照任意离散分布 $\mathbb{F}$进行分配。对于每个 k、 i， z-群的上界定义如下:

$$\overline{Z}^i_{\alpha,k} = P_{\mathbb{F}}(k) \cdot \overline{R}^i \tag{2}$$

根据关于累积奖励如何分布于各轮次的先验信息，只要符合 β-离散的定义，就可以使用任何合适的概率分布 $\mathbb{F}$ 来定义奖励离散度。例如，可以使用取值范围从 1 到 α 的贝塔 - 二项分布来描述奖励离散度。图 1 右侧的直方图展示了其中一些具有不同形状参数的贝塔-二项分布，但选择其他参数或分布来匹配我们希望建模的场景，其可能性几乎是无穷无尽的。

4 遗憾下界

利用 β-离散度性质，我们可以为一致高效策略推导出以下下界，即任何遗憾为 $o(T^x)$ 且满足 $x < 1$ 的策略。

定理 1. 对于具有 β-离散度性质的 TP-MAB 问题，经过 T时间步后，应用于该问题的任意一致高效策略 $\mathfrak{U}$ 的遗憾下界为

$$\lim_{T\to+\infty} \inf \frac{\mathcal{R}_T(\mathfrak{U})}{\ln T} \geq \sum_{i:\mu_i<\mu^*} \frac{2}{(\alpha+1)} \sum_{\mathrm{k}=1}^{\alpha} \mathrm{kB(k)} \cdot \alpha \sum_{\mathrm{k}=1}^{\alpha} (\mathrm{B(k)})^2 \frac{\Delta_i}{\alpha \mathcal{KL}\left(\frac{\mu_i}{R_{\max}}, \frac{\mu^*}{R_{\max}}\right)}$$

其中 $\Delta_i :=\mu^*-\mu_i$， $\mathcal{KL}(p, q)$: $=$均值为 p 和 q 的伯努利随机变量之间的 Kullback-Leibler 散度。

与 [Romano 等人， **2022**] 给出的下界比较

在假设 α-平滑性的情况下， [Romano 等人， 2022] 为 TP-MAB 证明了以下下界:

$$\lim_{T\to+\infty} \inf \frac{\mathcal{R}_T(\mathfrak{U})}{\ln T} \geq \sum_{i:\mu_i<\mu^*} \frac{\Delta_i}{\alpha \mathcal{KL}\left(\frac{\mu_i}{R_{\max}}, \frac{\mu^*}{R_{\max}}\right)}. \tag{3}$$

请注意，在 [Romanoet 等人于 2022] 式 (3) 中给出的下界适用于 [平滑性的情况时，我们在定理 1 中给出的下界与之相同。

然而，对于所考虑的问题设置，当

$$\frac{2}{(\alpha+1)}\sum_{\mathrm{k=1}}^{\alpha}\mathrm{kB(k)}\cdot\alpha\sum_{\mathrm{k=1}}^{\alpha}\mathrm{(B(k))}^2>1.$$

定理 1 的证明思路
我们首先构造两个多臂老虎机问题实例，它们要求试图解决问题的算法采取不同的行为。然后，我们利用分布变化论证来证明，任何一致高效算法都无法高效地区分这两个实例。

定理 1 的完整证明见附录 A。

5 所提算法与遗憾上界

本节提出一种利用 TP-MAB 设置中 β-离散度性质的算法，并证明其遗憾上界。

5.1 所提算法：**TP−UCB−FR−G**

我们提出的算法 TP-UCB-FR-G 是对 Romano 等人[提出的算法 TP-UCB-FR 的扩展 2022]。在 TP-UCB-FR-G 中，最重要的改动是置信区间 c_{t-1}^i，该区间经过严格构造，以适应 β-离散度性质。

该算法的输入包括平滑常数 $\alpha\in[\tau_{max}]$、最大延迟 τ_{max} 以及概率质量函数 B。算法利用 B，以便在所有延迟部分奖励均被观测到之前，对某个臂作出适当判断。具体做法是用虚构实现值（换言之，即期望估计奖励）替代尚未收到的部分奖励。

在第 t轮，与区间 $H:=\{t-\tau_{max}+1,...,t-1\}$内拉取的每个臂相关联的虚构奖励向量分别为 $\tilde{\boldsymbol{x}}_h^i=\left[\tilde{x}_{h,\ 1}^i,\ldots,\tilde{x}_{h,\ \tau_{\max}}^i\right]$，其中 $h\in H$，若$h+j\le t$（奖励已被观测到），则 $\tilde{x}_{h,\ j}^i:=x_{h,\ j}^i$；若 $h+j>t$（奖励将在未来被观测到），则 $\tilde{x}_{h,\ j}^i=0$。相应的虚构累积奖励为 $\tilde{r}_h^i:=\sum_{j=1}^{\tau_{\max}}\tilde{x}_{h,\ j}^i$。

在算法的初始化阶段（第 2–4 行），每个臂都会被拉取一次。之后，在每个时间步 t，通过分别使用公式 (4) 和 (5) 计算估计期望奖励 $\hat{}\ R_{t-1}^i$ 和置信区间 c_{t-1}^i，从而确定每个臂 i 的上置信界 u_{t-1}^i。

$$\hat{R}_{t-1}^i:=\frac{1}{N_i(t-1)}\left(\sum_{h=1}^{t-\tau_{\max}}r_h^i\mathbb{I}_{\{i_h=i\}}+\sum_{h\in H}\tilde{r}_h^i\mathbb{I}_{\{i_h=i\}}\right).\tag{4}$$

其中，$N_i(t-1):=\sum_{h=1}^{t-1}\mathbb{I}_{\{i_h=i\}}$ 是截至第 $t-1$轮，臂i 被拉动的次数。

$$c_{t-1}^i=\frac{\phi\bar{R}^i}{N_i(t-1)}\sum_{k=1}^{\alpha}kB(k)$$

算法 1 TP-UCB-FR-G

1: 输入：$\alpha\in[\tau_{max}],\tau_{max}\in\mathbb{N}^*,B\sim Y$
2: 对于 $t\in\{1,...,K\}$，执行
3: 拉取一个臂 $i_t=t$
4: 结束循环
5: 对于 $t\in\{K+1,...,T\}$ 执行
6: 对于 $t\in\{1,...,K\}$ 执行
7: 按照（4）和（5）计算 $\hat{}\ R_{t-1}^i$ 和 c_{t-1}^i
8: $u_{t-1}^i\leftarrow\hat{}\ R_{t-1}^i+c_{t-1}^i$
9: 结束循环
10: 选择臂 $i_t=z=\mathrm{argmax}_{i\in[K]}\,u_{t-1}^i$
11: 观察 $x_{h,\ t-h+1}^{i_h}$，对于 $h\in\{t-\tau_{max}+1,...,t\}$
12: 结束循环

$$+\bar{R}^i\sqrt{\frac{2\ln(t-1)\sum_{k=1}^{\alpha}\left(B(k)\right)^2}{N_i(t-1)}}.\tag{5}$$

然后，算法拉动上置信界 u_{t-1}^i 最高的臂 i，并观察其奖励。

5.2 TP-UCB-FR-G 的遗憾上界

定理 2。在具有 β-离散度奖励的 TP-MAB 设置中，若 B(k) 与随机变量 Y 的概率质量函数 $P_{\mathbb{F}}(k)$ 匹配，则 TP-UCB-FR-G 在经过 T 个时间步后的遗憾上界为

$$\begin{aligned}\mathcal{R}_T(\mathit{TP-UCB-FR-G})\le&\sum_{i:\mu_i<\mu^*}\frac{4\ln T\,(\bar{\mathrm{R}}^{\mathrm{i}})^2\sum_{\mathrm{k=1}}^{\alpha}\mathrm{(B(k))}^2}{\Delta_i}\\&\cdot\left(1+\sqrt{1+\frac{\Delta_i\phi\sum_{\mathrm{k=1}}^{\alpha}\mathrm{kB(k)}}{\bar{R}^i\ln T\sum_{k=1}^{\alpha}\left(B(k)\right)^2}}\right)\\&+2\phi\sum_{\mathrm{k=1}}^{\alpha}\mathrm{kB(k)}\sum_{i:\mu_i<\mu^*}\bar{R}^i+\left(1+\frac{\pi^2}{3}\right)\sum_{i:\mu_i<\mu^*}\Delta_i\end{aligned}$$

注意，$\sum_{\mathrm{k=1}}^{\alpha}\mathrm{kB(k)}=\mathbb{E}[Y]$，这意味着随机扩散变量 Y 的期望值会影响算法的上界。另一个有趣的因素是 $\sum_{k=1}^{\alpha}(B(k))^2$。这个因素可以看作奖励之间 Friedman 重合指数 [的近似值。重合指数决定了两个奖励点在同一个 z-群中被观测到的概率。其最小值为 $\frac{1}{\alpha}$，并在满足 α-平滑性属性（均匀分布）时达到。如果所有奖励都落入同一个 z-群，其值则达到最大值 1。

与 [Romano 等人 **2022]** 给出的**TP−UCB−FR**上界进行比较

让我们将定理 2 中给出的上界与 [Romano 等人， 2022] 中给出的上界进行比较。要使后一个上界成立，作为输入提供给其算法的 α 估计值也必须与真实奖励分布的 α 相匹配。请注意，在 α-平滑性情况下，$\mathbb{E}[Y]=\frac{\alpha+1}{2}$。对于具有 $\mathbb{E}[Y]<\frac{\alpha+1}{2}$ 的 $Y\sim\mathbb{F}$，我们给出的遗憾上界更低。此外，选择一个具有低均值的 β-离散度分布 $\mathbb{F}$

以及低重合指数，相较于选择 $\mathbb{F}$ 奖励集中在末尾附近（均值较高）且分布不够分散（重合指数较高）的情况，使用定理 2 将得到更好的上界。

定理 2 的证明概要
这里我们简要证明定理 2。完整证明见附录 B。

该方法分为三个步骤。首先，我们证明最优臂的估计值显著低于其均值的概率有界于 t^{-4}。其次，我们证明次优臂的估计值显著高于其均值的概率有界于 t^{-4}。最后，我们评估该算法区分最优臂与次优臂之间差异的表现。

6 实验结果

本节将所提算法 TP-UCB-FR-G 与 TP-UCB-FR [Romano 等人、 2022],UCB1 [Auer 等人、 2002], 以及 Delayed-UCB1[Joulani 等人 2013]进行比较。我们考察 TP-UCB-FR-G 在奖励分布各不相同的设置中的表现。我们采用 [Romano 等人提出的实验设置 2022]。具体而言，包括两个合成生成环境和一个真实世界的播放列表推荐场景。在这些设置中，我们沿用 [Romano 等人 2022], 所提供实验中使用的学习器，并使用 TP-UCB-FR-G 创建新的学习器配置。对于新学习器的输入分布，我们采用贝塔-二项分布，并为每个学习器设置独特的参数值。贝塔-二项分布使我们能够对极端场景进行建模，从而有望获得更有见地的实验结果。正如附录 C.4 中延后进行的实验所示，其他分布并不具备贝塔-二项分布的灵活性。本节图表中，我们使用记号 TP-UCB-FR-G(α, dist name) 表示采用我们算法的学习器，其中 dist name 是贝塔-二项分布的名称，具体参数见表 1。有关所用贝塔-二项分布和实验设置的更多细节，请参见附录 C.4。

设置 1：均匀奖励分布

在此设置下，我们评估 α 对 TP-UCB-FR-G 的影响。我们设置 $K = 10$、 $\tau_{\max} = 100$ 轮次，并且

分布名称	α	β
极端前置	1	100
非常前置	1	16
开始	2	8
开始中间	2	4
中间	5	5
中间后置	4	2
end	8	2
极端后置	16	1

表 1：贝塔-二项分布的参数值

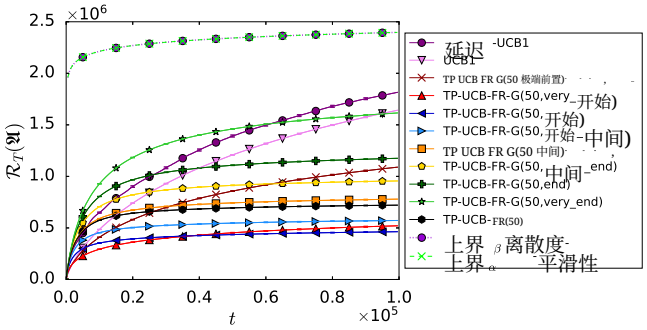

图 2：设置 1 中遗憾随时间的变化（ $\alpha_{est} = 50$）

通过令 $\overline{R}^i = 100\zeta^i$，使最大奖励满足 $\zeta \in \{1, 3, 6, 9, 12, 15, 18, 21, 22, 23\}$，从而使学习器更难收敛到最优臂。累计奖励满足 $Z^i_{t,k} \sim \frac{\overline{R}^i}{\alpha} U[0, 1]$，并且我们采用平滑常数为 $\alpha = 20$ 的 α设置。我们在时间范围$T = 10^5$ 内运行该算法，并对 100 次独立运行的结果取平均值。我们运行该设置 $\alpha_{est} \in \{5, 10, 20, 25, 50\}$，其中 α_{est}是 α平滑常数的估计值。为模拟底层数据生成分布和 α 值可能未知的现实世界应用，我们使用 α_{est}来估计该常数。

结果

我们来关注 $\alpha_{est} = 50$[1]的结果。图 2 展示了其性能随时间的变化，并给出了对应设置下的理论上界。请注意，β-离散度和 α-平滑性的上界相同，这符合该均匀扩散设置下的预期。 t 取值较小时上界较高，这是由上界方程中的常数项导致的，这些常数取决于实验设置。各臂的最大奖励累积值是影响最大的因素。首先，已有研究表明， [Romano 等人 2022]发现，实践中 α_{est} 取较大（乐观）值会带来更好的性能。然而，α_{est} 取过度乐观的值并不一定会带来更好的性能。因此，TP-UCB-FR 在很大程度上受到 α_{est}错误设定的影响。请注意，学习器 TP-UCB-FR-G(50, begin) 和 TP-UCB-FR-G(50, begin middle) 的表现显著优于[Romano 等人 2022]提出的 TP-UCB-FR 学习器。事实上，TP-UCB-FR-G(50, begin) 和 TP-UCB-FR-G(50, begin middle) 与 TP-UCB-FR 渐近地近似平行，因此在较大的时间范围 T 下也能带来显著的性能增益。这表明，我们的贡献改进了 [Romano 等人 2022]所研究设置下的性能上界。

进一步的结果表明，对于较低的 α_{est} 值，采用偏向开头分布的 TP-UCB-FR-G 学习器表现略好。在 α_{est} =5 这一特定情况下，对附录 C 中详述的精确遗憾结果进行数值分析后发现，我们提出的 TP-UCB-FR-G(5, begin middle) 学习器的表现优于

[1] α_{est} 其他取值的完整结果将在附录 C 中给出。

TP-UCB-FR 学习器 [Romano 等人， 2022]，优势达 α_{est} %。此外，随着 α_{est} 增大，偏向开头的 TP-UCB-FR-G 学习器的性能提升速度快于 TP-UCB-FR，因此在 $\alpha_{est} = 20$ 时性能提升 $\approx$ 22.1%，在 α_{est} =50 时则提升至 $\approx$ 36.1 %。我们观察到，TP-UCB-FR-G(α_{est}, begin) 本质上是“理想”的学习器，因为在测试的设置中，它的表现始终优于 [Romano 等人， 2022] 提出的学习器。这些结果表明，过度乐观或不够乐观的问题，本质上是以另一种形式承袭自 [Romano 等人， 2022] 的设置。过度乐观的 [2]
TP-UCB-FR-G 学习器通常表现较差。这可以归因于过度乐观的学习器通常具有重合指数较高的假设分布，因为奖励被假设为更集中在开头或末尾。在设置 1 中， $\alpha_{est} = 50$ 条件下“极端前置”和“极端后置”学习器的重合指数分别为 ≈ 0.51 和 0.14。这两个数值显著高于同一设置中“前置”和“后置”学习器的重合指数 ≈ 0.05。类似地，采用中间导向分布的不够乐观的学习器表现也较差。这可以用其假设分布的期望值来解释：正如名称所示，该期望值高于偏向开头的分布。

设置 2：非均匀奖励分布

第二种设置旨在测试 TP-UCB-FR-G 算法在这样的场景中的表现：拉取某个臂后，后续时间步内累计奖励的分布是非均匀的。此设置中的奖励分布满足 $Z^i_{t,\ k} \sim \frac{\overline{R}^i}{\alpha} Beta[a^i_k, b^i_k]$ 其中 $Beta$ 是贝塔分布，具有参数 a、b，使奖励按照相应设置的离散度进行分布。同样，我们对 $K = 10$ 个臂进行建模，平滑常数为 α，最大奖励为 $\alpha = 20$，使得收敛到最优臂所需的时间更长。也就是说， $\overline{R}^i = 100\zeta^i$，其中 $\zeta \in \{1, 3, 6, 9, 12, 15, 18, 21, 22, 23\}$。不过， $\tau_{\max}$、α_{est} 以及学习器所假设的贝塔分布参数有所不同。具体配置见附录 C。总体而言，共有 12 种组合，由 4 种配置和每种配置下的 3 个场景组成。配置之间的差异在于 $\tau_{\max}$ 和 α_{est}，而场景之间的差异则在于分布参数。通常有一个均匀场景（与设置 1 相同）、一个在拉取后较晚观测到奖励的场景（设置 2.1），以及一个在拉取后立即观测到结果的场景（设置 2.2）。我们使用与设置 1 中相同的估计分布的学习器（见表 1）。

结果

对 $\tau_{\max} = 100$ 和$\alpha_{est} = 50$ 运行所提出的设置 2.1 和 2.2，得到的结果与图 2 中设置 1 的结果在视觉上完全相同。然而，对数值结果的分析表明，两者之间确实存在差异。累积遗憾通常会从设置 2.2 开始略微增加

[2]如果 TP-UCB-FR-Glearner 采用「尾部导向」的贝塔-二项分布，并因此预期大部分奖励分布在前 z个群中，我们就认为它是「乐观的」。

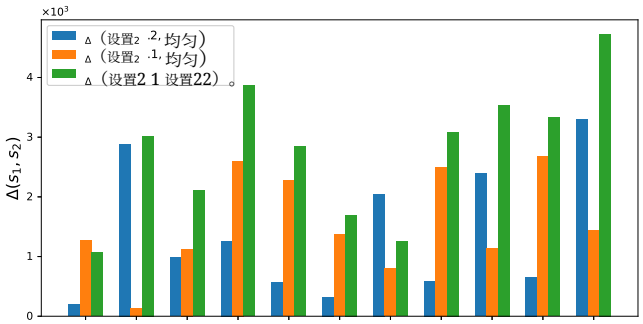

图 3：设置 2 中不同场景下学习器的累积遗憾差异， $\tau_{\max} = 100$、$\alpha_{est} = 50$

接着是均匀设置，随后是设置 2.1。

图 3 展示了聚合奖励以非均匀方式分布时，各设置下累积遗憾的差异。我们将 $\Delta(s_1, s_2)$ 定义为设置 2.1、设置 2.2、均匀}下 $s_1, s_2 \in$ {的累积遗憾绝对差值，即设置 s_1 与设置 s_2 之间的差异。请注意，图 3 所示的差异仅为边际差异。例如，Δ(设置 2.1，设置 2.2) $\approx 4.8 \times 10^3$ 对学习器 TP-UCB-FR-G(50, 极端后置) 而言，这是所有比较设置中观察到的平均遗憾最大差异。由于 TP-UCB-FR-G(50, 极端后置) 在 T 上的遗憾平均值为 $\approx 1.61 \times 10^6$，观察到的变化 $\approx 0.3\%$ 可以忽略不计。此外，使用不同的 $\tau_{\max}$ 和 α_{est} 值进行相同实验，似乎也证实了这种边际变化。例如， $\tau_{\max} = 200$ 和 $\alpha_{est} = 20$ 下的设置 2 中，平均遗憾的最大变化仅为 $\approx 0.5\%$。

这些结果表明，在累计奖励均匀分布的设置中，TP-UCB-FR-G 学习器的性能与累计奖励非均匀分布的设置中的性能无法区分。因此，我们可以与设置 1 的结论保持一致；TP-UCB-FR-G(α,begin) 相较于[Romano 等人提出的学习器，能显著提升性能 2022]。在上述设置中，我们观察到的增益高达 $\approx 48.2\%$。详细的性能总结见附录 C。再次说明，由于分布的选择具有灵活性，因此性能还有可能进一步提升。

图 4 在设置 2.2 的结果上方绘制了 TP-UCB-FR-G 的理论上界以及 TP-UCB-FR 算法的上界。该图表明，在此设置下，本文提出的上界更紧。请注意，TP-UCB-FR-G 和 TP-UCB-FR 的理论上界仅适用于能够精确假设数据生成分布的特定学习器

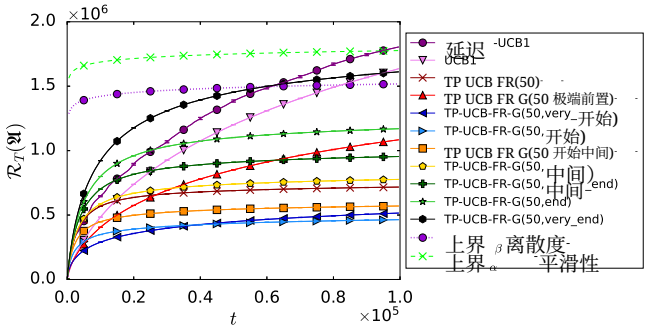

图 4：设置 2.2 中 $\tau_{\max} = 100$ 和$\alpha_{est} = 50$的遗憾随时间变化

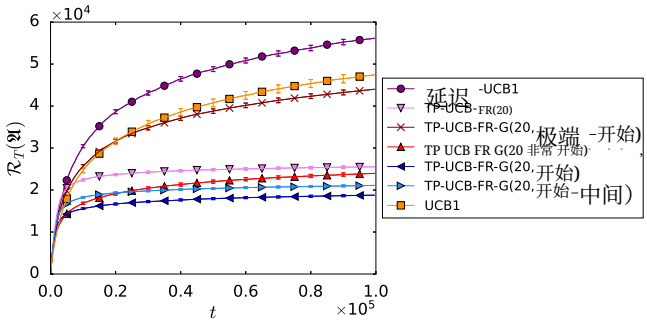

图 5：Spotify 设置中面向开始的学习器的遗憾随时间的变化

，而“极端后置”学习器超过了 β-离散度上界。这说明了乐观地估计假设分布的另一个原因。

用户推荐：Spotify 播放列表

为在真实世界数据集上验证我们的算法，我们使用 [Brost 等人、 2019]介绍的 Spotify 数据集，解决由 [Romano 等人、 2022], 提出的用户推荐问题。我们选择播放次数最多的 $K = 6$ 个播放列表作为待推荐的臂。每次选中一个播放列表i 时，前 $N = 20$ 首歌曲对应的奖励实现 x_t^i 均从数据集中进行采样。在此设置下， α-平滑性为 $\alpha = 20$，最大延迟$\tau_{\max} = 4N = 80$，结果取 100 次独立运行的平均值。

结果

在图 5 中，我们观察到，乐观学习器的表现显著优于 [Romano 等人 2022]提出的学习器 TP-UCB-FR(20)。我们重点关注学习器 TP-UCB-FR-G(20, begin)，因为它的表现迄今为止最佳。与 TP-UCB-FR(20) 相比，我们观察到，该学习器在时间范围 T 内的平均遗憾降低了 $\approx 26.3\%$。表 2 完整总结了 TP-UCB-FR-G 学习器在 Spotify 场景中的性能增益。

我们的第二个观察是，与设置 1 的结论一致，过于乐观的学习器，例如

学习器 TP-UCB-FR-G($\alpha_{est} = 20$)	遗憾 ($\times 10^4$)	降低 (%)
极端前置	4.40	≈ -72.5
非常开始	2.40	≈ 5.9
开始	**1.88**	$\approx$ **26.3**
开始 中间	2.11	≈ 17.2

表 2：TP-UCB-FR-G 学习器及其相较于 TP-UCB-FR(20) 遗憾的降低幅度 $= 2.55 \times 10^4$

TP-UCB-FR-G(20,极端前置)，其表现明显不如 TP-UCB-FR(20)。如表 2 所示，平均遗憾增加了 ≈ 72.5 %。尽管如此，由于当 t 较大时，TP-UCB-FR-G(20,开始) 的表现明显优于 TP-UCB-FR(20)，因此它更适合用于播放列表推荐。

7 总结与未来工作

本文使用一种称为“具有广义时间分区奖励的多臂老虎机”的新型建模方法，对存在延迟反馈的序贯决策问题进行建模。为了推广延迟奖励分布的情形，我们引入了 β-离散度性质。针对具有 β-离散度性质的 TP-MAB 设置，我们给出了一个下界，该下界可能比具有 α-平滑性性质的 TP-MAB 设置的下界更紧。具体而言，我们证明，即使两个设置中的 α 相等，假设分布的均值较高或重合指数较高，也会使本文引入的设置具有更紧的界。我们还提出了利用 β-离散度性质的 TP-UCB-FR-G 算法，并证明在某些场景下，该算法的上界可以低于 TP-UCB-FR 算法的上界，因此也低于经典 UCB1 和 Delayed-UCB1 算法的上界。最后，我们通过合成数据和真实世界数据开展了大量实验，实证表明，我们的算法优于 TP-UCB-FR 及其他 UCB 算法。在使用真实世界数据的实验中，与最先进水平的 TP-UCB-FR 算法相比，累积遗憾最多降低了 26.3%。

未来的研究方向包括：

1. 目前， β-离散度属性仅适用于定义域大小为 α 的有限定义域上的离散概率分布。一项有趣的未来研究课题是探讨如何去除这些限制，因为这可能使我们的算法更具灵活性，并拓宽其实际应用范围。2. 另一种可能的扩展是考虑将臂视为子集的设置，使每个臂子集都被赋予不同的 α-值和概率分布。在将臂视为簇的设置中，这可能会很有优势，例如 [Pandey 等人提出的设置， 2007]3. 此外，另一个有趣的课题是研究奖励划分所依据的时间跨度（即 $\tau_{\max}$）可变的情形。在目前的工作中，我们

假设该时间跨度固定且对所有臂保持一致。如果去除这一假设，可能有助于我们考虑实际应用中 $\tau_{\max}$ 并非固定的情况，例如，在在线广告的生命周期（$\tau_{\max}$）内观察其奖励（广告的点击量），而该生命周期是可变的。

参考文献

[Agarwaland Duchi, 2011] Alekh Agarwal 和 John C Duchi。《分布式延迟随机优化》。载于《神经信息处理系统进展》，第 24 卷，2011 年。

[Arya 和 Yang, 2020] Sakshi Arya 和 Yuhong Yang。用于具有延迟奖励的上下文多臂老虎机问题的随机分配与非参数估计。《统计与概率快报》，164：108818，2020 年。

[Auer 等人, 2002] Peter Auer、Nicol`o Cesa-Bianchi 和 Paul Fischer。多臂老虎机问题的有限时间分析。《机器学习》2002 年第 47 卷第 2 期，47：235–256，2002 年 5 月。

[Brost 等人, 2019] Brian Brost、Rishabh Mehrotra 和 Tristan Jehan。《音乐流媒体会话数据集》。CoRR, abs/1901.09851，2019 年。

[Bubeck 和 Cesa-Bianchi，2012a] S´ebastien Bubeck 和 Nicol`ò Cesa-Bianchi。《随机与非随机多臂老虎机问题的遗憾分析》。《机器学习基础与趋势》，5(1):1–122，2012。

[Bubeck 和 Cesa-Bianchi，2012b] S´ebastien Bubeck 和 Nicol`ò Cesa-Bianchi。《随机与非随机多臂老虎机问题的遗憾分析》。CoRR，abs/1204.5721，2012。

[Cesa-Bianchi 等人, 2022] Nicol`ò Cesa-Bianchi、Tommaso Cesari、Roberto Colomboni、Claudio Gentile 和 Yishay Mansour。《具有复合匿名反馈的非随机老虎机》。《机器学习研究杂志》，23(277):1–24，2022。

[Desautels 等人, 2014] Thomas Desautels、Andreas Krause 和 Joel W. Burdick。《高斯过程老虎机优化中探索与利用权衡的并行化》。《机器学习研究杂志》，15(1):3873–3923，2014年1月。

[Friedman, 1987] William Frederick Friedman。《重合指数及其在密码分析中的应用》，第49卷。加利福尼亚爱琴海公园出版社，1987年。

[Gael 等人, 2020] Manegueu Anne Gael、Claire Vernade、Alexandra Carpentier 和 Michal Valko。《具有臂相关延迟的随机多臂老虎机》。载于《第37届国际机器学习会议论文集》，《机器学习研究论文集》第119卷，第3348–3356页，2020年7月13日至18日。

[Hoeffding, 1963] Wassily Hoeffding。有界随机变量之和的概率不等式。《美国统计学会杂志》，58(301):13–30，1963 年。

[Joulani 等人, 2013] Pooria Joulani、Andr´as Gy¨orgy 和 Csaba Szepesv´ari。延迟反馈下的在线学习。载于《第 30 届国际机器学习会议论文集——国际机器学习会议》第 28 卷，ICML'13，2013 年。[Lai 和 Robbins, 1985] T.L Lai 和 Herbert Robbins。渐近高效的自适应分配规则。《应用数学进展》，6(1):4–22，1985 年。[Lee, 2012] Peter M. Lee。3.1 二项分布。Wiley，第 4 版，2012 年。

[Mandel 等人, 2015] Travis Mandel、Yun-En Liu、Emma Brunskill 和 Zoran Popovi´c。《队列方法：处理延迟、启发式方法、先验数据及多臂老虎机评估》。载于第二十九届 AAAI 人工智能会议，2015年。

[Mesterharm, 2005] Chris Mesterharm。《具有延迟标签反馈的在线学习》。载于第16届算法学习理论国际会议论文集，第399–413页，2005年。

[Mesterharm, 2007] Chris Mesterharm. 在线学习的改进。博士论文，罗格斯大学，2007。[Pandey 等人, 2007] Sandeep Pandey、Deepayan Chakrabarti 和 Deepak Agarwal。具有相关臂的多臂老虎机问题。载于第24届《机器学习》国际会议论文集，第721–728页，2007。

[Pike-Burke 等人, 2018] Ciara Pike-Burke、Shipra Agrawal、Csaba Szepesvari 和 Steffen Grunewalder。具有延迟、聚合匿名反馈的老虎机问题。载于国际机器学习会议，第4105–4113页。PMLR，2018。

[Romano 等人, 2022] Giulia Romano、Andrea Agostini、Francesco Trov`o、Nicola Gatti 和 Marcello Restelli。具有时间分区奖励的多臂老虎机问题：部分反馈何时发挥作用。第三十一届国际人工智能联合会议论文集，2022。

[Vernade等人, 2017] Claire Vernade、Olivier Capp´e 和 Vianney Perchet。用于延迟转化的随机老虎机模型。发表于人工智能不确定性会议，澳大利亚悉尼，2017年8月。

[Vernade 等人，2020a] Claire Vernade、Alexandra Carpen- tier、Tor Lattimore、Giovanni Zappella、Beyza Ermis 和 Michael Br¨uckner。具有随机延迟反馈的线性老虎机。发表于第37届国际机器学习会议论文集，《机器学习研究论文集》第119卷，第9712–9721页，2020年7月13日至18日。

[Vernade 等人，2020b] Claire Vernade、Andr´as Gy¨orgy 和 Timothy A. Mann。具有中间观测的非平稳延迟老虎机。发表于第37届国际机器学习会议论文集，ICML'20，2020年。

[Wang 等人, 2021] Siwei Wang、HaoyunWang 和 Longbo Huang。《具有复合匿名反馈的多臂老虎机自适应算法》

《AAAI 人工智能会议论文集》，35(11):10210–10217，2021。[Weinberger 和 Ordentlich，2002] M.J. Weinberger 和 E. Ordentlich。《个体序列的延迟预测》。《IEEE 信息论汇刊》，48(7):1959–1976，2002。[Zhou 等人，2019] Zhengyuan Zhou、Renyuan Xu 和 Jose Blanchet4。《随机延迟下广义线性上下文老虎机中的学习》。（Curran Associates Inc.），2019。[Zinkevich 等人，2009] Martin Zinkevich、John Langford 和 Alex Smola。《慢速学习者速度快》。《神经信息处理系统进展》，第 22 卷，2009。

定理 1 的证明

证明。证明思路遵循 [Bubeck 和 Cesa-Bianchi, 2012b], 中定理 2.2 的证明，该证明基于 [Lai 和 Robbins, 1985]。由于 β 离散度性质不会影响 ϕ 群的基数，因此我们可以推广到单次臂拉取可获得多个奖励的设置。我们定义一个辅助的 TP-MAB 设置，其中：

- 只存在两条臂，其期望值为 μ_1 和 μ_2，满足 $\mu_2 < \mu_1 < 1$。
- 每条臂的奖励上界等于最大上界，即 $\overline{R}_t^i = \overline{R}_{\max}$
- 每个 z 群中的总奖励 $Z_{t,\ k}^i$ 相互独立，并且每个 z 群中奖励的期望值为 $\mathrm{B}(k) \cdot \mu_i$。
- -群中的总奖励，$Z_{t,\ k}^i$是一个缩放伯努利随机变量，满足 $Z_{t,k}^i \in \{0, \mathrm{B}(k) \cdot \overline{R}_{\max}\}$
- 在时间 t 拉取一个臂$\{Z_{t,1}^i, ..., Z_{t,\alpha}^i\}$ 会获得奖励，这些奖励都能在拉取时立即被观测到。

在这个下界证明中，我们可以直接观察到：在所有部分奖励都能同时被观测到的设置中，找到最优臂不可能比在奖励分散到一组轮次 $\{t, t+1, ..., \tau_{max}\}$ 中的设置更困难。因此，在这一设定下得到的下界，也对应于我们的 β-离散设置中的下界。

为了说明一个臂相对于其最大值的优劣，我们为每个臂推导出一个新的替代均值，如下所示：$\mu_{A_i} = \frac{\mu_i}{\overline{R}_{\max}}$。请注意，正如之前所述，$\mu_{A_i} < 1$。

令 $\mathbb{E}[\mathrm{N}_i(T)]$ 表示在一组轮次 T 中拉取某个臂i的预期次数。为了计算 $\mathbb{E}[\mathrm{N}_1(\{t, t+1, ..., \tau_{max}\})]$ 和 $\mathbb{E}[\mathrm{N}_2(\{t, t+1, ..., \tau_{max}\})]$,，我们可以使用缩放后的奖励值，而不会损失有效性。现在考虑上述 TP-MAB 设置的第二个修改后的实例，唯一的区别是臂 2 现在是最优臂，且满足 $\mu_{A_1} < \mu'_{A_2} < 1$。我们可以证明，选择各个臂的学习智能体无法区分这些不同的实例。这一推理意味着对次优臂被选择次数的下界。我们知道，$x \mapsto \mathcal{KL}(\mu_{A_1}, x)$ 是一个连续函数，并且可以为每个 $\epsilon > 0$ 找到一个 μ'_{A_2}，使得：

$$\mathcal{KL}(\mu_{A_2}, \mu'_{A_2}) \leq (1+\epsilon)\mathcal{KL}(\mu_{A_2}, \mu_{A_1}) \tag{6}$$

证明遵循[Bubeck 和 Cesa-Bianchi, 2012b] 的研究中给出的步骤，以推导任何一致策略 $\mathfrak{U}$ 的下界。

步骤 1: $\mathbb{P}(C_t) = o(1)$

在此证明中，我们稍微调整了奖励的记号，使序列 $Z_{1,1}^i$, ..., $Z_{n,\alpha}^i$ 中的每个变量都表示臂 i 在时间步 $k \in \{1, 2, ..., \alpha\}$ 被拉动 n 次时获得的累积奖励。$Z_{s,k}^i$ 表示臂 i 在第 s 次拉动后、时间步 $s \in \{1, 2, ..., n\}$ 的累积奖励，即拉动之后的时间步 k。使用此记号，我们可以将 $\mathcal{KL}(\mu_{Z_2}, \mu'_{Z_2})$ 的经验估计定义为：

$$\widehat{\mathcal{KL}}_{\alpha\beta} := \sum_{n=1}^{s}\sum_{k=1}^{\alpha} \ln \frac{\mu_{A_2} Z_{n,k}^2 + (1-\mu_{A_2})(1-Z_{n,k}^2)}{\mu'_{A_2} Z_{n,k}^2 + (1-\mu'_{A_2})(1-Z_{n,k}^2)}$$

据此，我们定义一个事件，将原始智能体的行为与修改版本联系起来

$$C_t := \left\{ \alpha N_2(t) < f_t \quad \text{and} \quad \widehat{\mathcal{KL}}_{\alpha N_2(t)} \leq (1-\varepsilon/2)\ln t \right\} \tag{7}$$

with

$$f_t = \left(\frac{2}{\alpha+1} \sum_{\mathrm{k}=1}^{\alpha} \mathrm{kB(k)} \cdot \alpha \sum_{\mathrm{k}=1}^{\alpha} (\mathrm{B(k)})^2 \right) \frac{1-\epsilon}{\mathcal{KL}(\mu_{A_2}, \mu'_{A_2})} \ln t$$

利用[Bubeck 和 Cesa-Bianchi, 2012b] 中定义的测度变换恒等式，以及式 (7) 中 C_t 定义的第二个不等式：

$$\mathbb{P}'(C_t) = \mathbb{E}\left[1_{C_t} \exp\left(-\widehat{\mathcal{KL}}_{\alpha N_2(t)} \right) \right] \geq e^{-(1-\varepsilon/2)\ln t} \mathbb{P}(C_t)$$

然后，我们首先重新排列上述不等式中的各项，得到

$$\begin{aligned} \mathbb{P}(C_t) &\leq t^{(1-\varepsilon/2)} \mathbb{P}'(C_t) \\ &\leq t^{(1-\varepsilon/2)} \mathbb{P}'(\alpha N_2(t) < f_t) \\ &\leq t^{(1-\varepsilon/2)} \frac{\mathbb{E}'[t - N_2(t)]}{t - f_t/\alpha} \\ &= o(1) \end{aligned}$$

在上面的方程中，我们使用了 $\mathbb{P}'(C_t) \leq \mathbb{P}'(\alpha N_2(t) < f_t)$、马尔可夫不等式，以及策略 $\mathfrak{U}$ 一致高效这一事实（即 $\mathbb{E}[N_2(t)] = o(t^\gamma)$，其中 $\gamma < 1$）。

Step 2: $\mathbb{P}(\alpha N_2(t) \leq f_t) = o(1)$

根据 [Bubeck 和 Cesa-Bianchi, 2012b] 中的定理 2.2，并注意到我们始终有

1. $\sum_{\mathrm{k}=1}^{\alpha} \mathrm{kB(k)} \geq 1 \implies \frac{2}{\alpha+1} \sum_{\mathrm{k}=1}^{\alpha} \mathrm{kB(k)} \geq \frac{2}{\alpha+1}$
2. $\sum_{\mathrm{k}=1}^{\alpha} (\mathrm{B(k)})^2 \in [\frac{1}{\alpha}, 1] \implies \alpha \sum_{\mathrm{k}=1}^{\alpha} (\mathrm{B(k)})^2 \in [1, \alpha]$
3. $\frac{2}{\alpha+1} \sum_{\mathrm{k}=1}^{\alpha} \mathrm{kB(k)} \cdot \alpha \sum_{\mathrm{k}=1}^{\alpha} (\mathrm{B(k)})^2 \geq \frac{2}{(\alpha+1)} > 0$

接下来，我们定义两个事件

$$E_1 = \alpha N_2(t) < f_t$$

and

$$E_2 = \left(\frac{\alpha+1}{2\alpha} \cdot \frac{1}{\sum_{\mathrm{k}=1}^{\alpha} \mathrm{kB(k)} \cdot \sum_{\mathrm{k}=1}^{\alpha} (\mathrm{B(k)})^2} \cdot \frac{\mathcal{KL}(\mu_{Z_2}, \mu'_{Z_2})}{(1-\varepsilon)\ln t} \cdot \max_{\beta < f_t/\alpha} \widehat{\mathcal{KL}}_{\alpha\beta} \right.$$

$$\leq \frac{1-\varepsilon/2}{1-\varepsilon} \cdot \frac{\alpha+1}{2\alpha} \cdot \frac{\mathcal{KL}(\mu_{Z_2}, \mu'_{Z_2})}{\sum_{\mathrm{k}=1}^{\alpha} \mathrm{kB(k)} \cdot \sum_{\mathrm{k}=1}^{\alpha} (\mathrm{B(k)})^2} \Bigg)$$

从而得到:

$$o(1) = \mathbb{P}(C_t) \leq \mathbb{P}(E_1 \wedge E_2)$$

利用事件 E_2 的大数强律，s.t. $\lim_{t\to+\infty} \mathbb{P}(E_2) = 1$, 我们可以得出 $\mathbb{P}(E_1) = \mathbb{P}(\alpha N_2(t) < f_t) = o(1)$，并且对于 $t \to +\infty$，我们有$\mathbb{E}[N_2(t)] > f_t/\alpha$。

最后一步

根据公式 (6)，我们知道，对于 $t \to +\infty$：

$$\begin{aligned}
\mathbb{E}[N_2(t)] &> f_t/\alpha \\
&= \frac{2}{\alpha+1} \sum_{\mathrm{k}=1}^{\alpha} \mathrm{kB(k)} \cdot \alpha \sum_{\mathrm{k}=1}^{\alpha} (\mathrm{B(k)})^2 \\
&\quad \frac{1-\varepsilon}{\alpha \mathcal{KL}\left(\mu_{A_2}, \mu'_{A_2}\right)} \ln t \\
&\geq \frac{2}{\alpha+1} \sum_{\mathrm{k}=1}^{\alpha} \mathrm{kB(k)} \cdot \alpha \sum_{\mathrm{k}=1}^{\alpha} (\mathrm{B(k)})^2 \\
&\quad \frac{1-\varepsilon}{\alpha(1+\varepsilon) \mathcal{KL}\left(\mu_{A_2}, \mu_{A_1}\right)} \ln t
\end{aligned}$$

其中，定理陈述是通过利用 ε 的任意性，将 μ_{A_1} 替换为 $\frac{\mu^*}{R_{\max}}$、将 μ_{A_2} 替换为$\frac{\mu_2}{R_{\max}}$，并对所有次优臂求和而得出的。□

B 定理 2 的证明

B.1 预备知识

次优臂拉取次数的期望与算法遗憾之间的关系如下:

$$\mathcal{R}_T(\texttt{TP-UCB-FR-G}) = \sum_{i:\mu_i<\mu^*} \Delta_i \mathbb{E}[N_i(T)]$$

定义臂 i 在 $N_i(t)$ 次臂拉取中累积奖励的真实经验均值:

$$\hat{R}_t^{i,\,\text{true}} := \frac{1}{N_i(t)} \sum_{h=1}^{t} r_h^i \mathbb{1}_{\{i_h=i\}}$$

上述数值假设某次臂拉取的累积奖励已知，即使部分奖励仍将在未来陆续到来。我们对真实经验均值与观测到的经验均值之间的差异作如下界定:

$$\begin{aligned}
&\hat{R}_t^{i,\,\text{true}} - \hat{R}_t^i \\
&= \frac{1}{N_i(t)} \sum_{h=1}^{t} \sum_{j=1}^{\tau_{\max}} \left(x_{h,j}^i - \tilde{x}_{h,j}^i\right) \mathbb{1}_{\{i_h=i\}} \\
&\leq \frac{1}{N_i(t)} \sum_{h=1}^{t} \sum_{j=1}^{\tau_{\max}} (x_{h,j}^i - \tilde{x}_{h,j}^i) \\
&= \frac{1}{N_i(t)} \sum_{h=\max\{1,t-\tau_{\max}+2\}}^{t} \sum_{j=t-h+2}^{\tau_{\max}} x_{h,j}^i \qquad (8) \\
&\leq \frac{1}{N_i(t)} \sum_{k=1}^{\alpha} k\phi \bar{R}^i B(k) \qquad (9) \\
&= \frac{\phi \bar{R}^i}{N_i(t)} \sum_{k=1}^{\alpha} kB(k) \qquad (10)
\end{aligned}$$

式（8）表明，真实均值与观测均值之差，等于最多已被拉取的 $\tau_{\max}-1$ 个臂所对应的、尚未观测到的所有未来奖励之和。索引 h 越接近当前时间 t，存在未观测奖励的已拉取臂就越多。因此，可以通过遍历式（9）中的所有 z 群，计算仍待观测的最大奖励，并通过索引 k 对较晚的 z 群赋予更高权重，从而界定未观测奖励的总量。此外，式（9）成立是因为具有 β 离散度性质。

事实 1（Hoeffding不等式[Hoeffding, 1963]）。设 $X_1, \ldots, X_n$ 是 $[0,1]$ 中的随机变量，且满足 $\mathbb{E}[X_t \mid X_1, \ldots, X_{t-1}] = \mu$。令 $S_n = X_1 + \cdots + X_n$。那么，对于所有 $a \geq 0$

$$\mathbb{P}\{S_n \leq n\mu - a\} \leq e^{-2a^2/n}.$$

推导上界

根据算法 TP-UCB-FR-G 的构造，次优臂 i 被拉取的预期次数上界可表示如下:

$$\mathbb{E}[N_i(t)] \leq \ell + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=\ell}^{t-1} \mathbb{P}\left\{\left(\hat{R}_{t,s}^* + c_{t,s}^*\right) \leq \left(\hat{R}_{t,s_i}^i + c_{t,s_i}^i\right)\right\} \qquad (11)$$

其中$\hat{R}_{t,s}^*$和 $c_{t,s}^*$ 分别是最优臂的经验均值和置信项，而 $\hat{R}_{t,s_i}^i$ 和 c_{t,s_i}^i 表示臂 i 的经验均值和置信项。

要使（11）成立，以下三个不等式之一也必须成立:

$$\hat{R}_{t,s}^* \leq \mu^* - c_{t,s}^* \qquad (12)$$

$$\hat{R}_{t,s_i}^i \geq \mu_i + c_{t,s_i}^i \qquad (13)$$

$$\mu^* < \mu_i + 2c_{t,s_i}^i \qquad (14)$$

我们先关注（12），并得到以下结果:

$$\begin{aligned}
&\mathbb{P}\left(\hat{R}_{t,s}^* - \mu^* \leq -c_{t,s}^*\right) \\
&= \mathbb{P}\left(\hat{R}_{t,s}^{*\,\text{true}} - \mu^* \leq -c_{t,s}^* + \hat{R}_{t,s}^{*\,\text{true}} - \hat{R}_{t,s}^*\right) \\
&\leq \mathbb{P}\left(\hat{R}_{t,s}^{*\,\text{true}} - \mu^* \leq -c_{t,s}^* + \frac{\phi \bar{R}^i}{s} \sum_{k=1}^{\alpha} kB(k)\right) \\
&= \mathbb{P}\left(s\hat{R}_{t,s}^{*\,\text{true}} \leq s\mu^* - s\sqrt{\frac{2\ln t \sum_{k=1}^{\alpha} \left(\bar{R}^* B(k)\right)^2}{s}}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \exp\left\{-\frac{\left(2\sqrt{\frac{2\ln t\sum_{k=1}^{\alpha}\left(\bar{R}^{*}B(k)\right)^{2}}{s}}\right)^{2}s^{2}}{\sum_{l=1}^{s}\sum_{k=1}^{\alpha}\left(\bar{R}^{*}B(k)\right)^{2}}\right\}\\
&\leq e^{-4\ln t}\\
&= t^{-4}
\end{aligned}
$$

其中我们使用霍夫丁不等式（定义见事实 1）、倒数第二步和 $c_{t,s_i}^{i} := \frac{\phi\bar{R}^{i}}{s}\sum_{k=1}^{\alpha}kB(k) + \bar{R}^{i}\sqrt{\frac{2\ln t\sum_{k=1}^{\alpha}(}{B\ k^{-2}())}}$

^{s} 同样，可以从式（13）推导出该上界：

$$
\begin{aligned}
\mathbb{P}\left(\hat{R}_{t,s_i}^{i} - \mu_i \geq c_{t,s_i}^{i}\right) &\leq \mathbb{P}\left(\hat{R}_{t,s}^{i,\,\text{true}} - \mu_i \geq \bar{R}^{i}\sqrt{\frac{2\ln t}{\alpha s_i}}\right)\\
&\leq e^{-4\ln t}\\
&= t^{-4}
\end{aligned}
$$

其中，我们使用了霍夫丁不等式（定义见事实 1），以及根据定义，$\hat{}R_{t,\ s_i}^{i} \leq \hat{}_{R^{i},\ t}^{\ true}{}_{,\ s_i}$ 这一事实。剩下的只需考虑式（14）。假设式（14）不成立，即 $\mu^{*} \geq \mu^{i} + 2c_{t,\ s_i}^{i}$。这等价于

$$
\Delta_i \geq 2\left(\frac{\phi\bar{R}^{i}}{s_i}\sum_{k=1}^{\alpha}kB(k) + \sqrt{\frac{2\ln t\sum_{k=1}^{\alpha}\left(\bar{R}^{*}B(k)\right)^{2}}{s_i}}\right)
$$

重新排列各项。

$$
\begin{aligned}
&\frac{\Delta_i^{2}}{4} + \frac{\phi^{2}(\ \bar{R}^{i})^{2}}{s_i^{2}} \qquad \sum_{k=1}^{\alpha}kB(k)\Big)^{2} - 2\left(\frac{\Delta_i\phi(\quad \bar{R}^{i}}{2s_i}\left(\sum_{k=1}^{\alpha}kB\right)(k)\right)\\
&\geq \frac{2\ln t\sum_{k=1}^{\alpha}\left(\bar{R}^{*}B(k)\right)^{2}}{s_i^{2}\dfrac{\Delta_i^{2}}{4} + \phi^{2}(\ \bar{R}^{i})^{2}\left(\sum_{k=1}^{\alpha}kB(k)\right)^{2} - 2s_i\left(\dfrac{\Delta_i\phi(\bar{R}^{i})}{2}\right)\sum_{k=1}^{\alpha}kB(k)}\\
&\qquad \mathrm{n}\,t\sum^{\alpha+1}{}_{k=1}\left(\bar{R}^{*}B(k)\right)^{2}\Big)\\
&\geq 0
\end{aligned}
$$

通过求解 s_i，可以得出以下结论：

$$
\begin{aligned}
s_i \geq\ & \frac{2\phi\bar{R}^{i}\sum_{\mathrm{k}=1}^{\alpha}\mathrm{kB(k)}}{\Delta_i} + \frac{4\ln t\,(\bar{\mathrm{R}}^{\mathrm{i}})^{2}\sum_{\mathrm{k}=1}^{\alpha}(\mathrm{B(k)})^{2}}{\Delta_i^{2}}\\
&+ 4\frac{\sqrt{\left(\ln t\,(\bar{\mathrm{R}}^{\mathrm{i}})^{2}\sum_{\mathrm{k}=1}^{\alpha}(\mathrm{B(k)})^{2}\right)\left(1 + \frac{\sum_{\mathrm{k}=1}^{\alpha}\mathrm{kB(k)}\,\Delta\phi\bar{R}^{i}}{\ln t\,(\bar{\mathrm{R}}^{\mathrm{i}})^{2}\sum_{\mathrm{k}=1}^{\alpha}(\mathrm{B(k)})^{2}}\right)}}{\Delta_i^{2}}\\
s_i \geq\ & \frac{2\phi\bar{R}^{i}\sum_{\mathrm{k}=1}^{\alpha}\mathrm{kB(k)}}{\Delta_i} + \left(\frac{4\ln t\,(\bar{\mathrm{R}}^{\mathrm{i}})^{2}\sum_{\mathrm{k}=1}^{\alpha}(\mathrm{B(k)})^{2}}{\Delta_i^{2}}\right)\\
&\cdot\left(1 + \sqrt{1 + \frac{\Delta_i\phi\bar{R}^{i}\sum_{\mathrm{k}=1}^{\alpha}\mathrm{kB(k)}}{\ln t\,(\bar{\mathrm{R}}^{\mathrm{i}})^{2}\sum_{\mathrm{k}=1}^{\alpha}(\mathrm{B(k)})^{2}}}\right)
\end{aligned}
$$

因此，我们选择

$$
\begin{aligned}
l = \Bigg\lceil & \frac{2\phi\bar{R}^{i}\sum_{\mathrm{k}=1}^{\alpha}\mathrm{kB(k)}}{\Delta_i} + \left(\frac{4\ln t\,(\bar{\mathrm{R}}^{\mathrm{i}})^{2}\sum_{\mathrm{k}=1}^{\alpha}(\mathrm{B(k)})^{2}}{\Delta_i^{2}}\right)\\
&\cdot\left(1 + \sqrt{1 + \frac{\Delta_i\phi\bar{R}^{i}\sum_{\mathrm{k}=1}^{\alpha}\mathrm{kB(k)}}{\ln t\,(\bar{\mathrm{R}}^{\mathrm{i}})^{2}\sum_{\mathrm{k}=1}^{\alpha}(\mathrm{B(k)})^{2}}}\right)\Bigg\rceil
\end{aligned}
$$

以确保对于 $s_i \geq l$，式（14）中的不等式始终不成立。

$$
\begin{aligned}
&\mathbb{E}\left[N_i(t)\right]\\
&\leq \Bigg\lceil \frac{2\phi\bar{R}^{i}\sum^{\alpha}{}_{\mathrm{k}=1}\mathrm{kB(k)}}{\Delta_i} + \frac{4\ln\ t(\bar{\mathrm{R}}^{\mathrm{i}})^{2}\sum^{\alpha}{}_{\mathrm{k}=1}(\mathrm{B(k)})^{2}}{\Delta_i^{2}}\Bigg)\\
&\quad\cdot\left(1 + \sqrt{1 + \frac{\Delta_i\phi\bar{R}^{i}\sum^{\alpha}{}_{\mathrm{k}=1}\mathrm{kB(k)}\ln(\mathrm{R_i})_{\mathrm{k}=1}}{t^{-2}\sum^{\alpha}}}\right)\Bigg\rceil\\
&\quad(\mathrm{B(k)})^{2}\Bigg)\\
&\quad + \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_i=\ell}^{t-1}\left[\mathbb{P}\left(\hat{}R_{t,s}^{*} - \mu^{*} \leq -c_{t,s}^{*}\right)\right]\\
&\quad + \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_i=\ell}^{t-1}\left[\mathbb{P}\left(\hat{}R_{t,s_i}^{i} - \mu_i \geq c_{t,s_i}^{i}\right)\right]\\
&\leq \frac{2\phi\bar{R}^{i}\sum^{\alpha}{}_{\mathrm{k}=1}\mathrm{kB(k)}}{\Delta_i} + \Bigg(\frac{4\ln\ t(\bar{\mathrm{R}}^{\mathrm{i}})^{2}\sum^{\alpha}{}_{\mathrm{k}}}{}^{2}(\mathrm{B(k)})\\
&\quad\cdot\left(1 + \sqrt{1 + \frac{\Delta_i\phi\bar{R}^{i}\sum^{=1}{}^{\alpha}{}_{\mathrm{k}=1}\mathrm{kB(k)}\ln}{t^{-2}\sum^{\alpha}}}\right)\\
&\quad(\mathrm{R_i})_{\mathrm{k}=1}(\mathrm{B(k)})^{2}\Bigg)\\
&\quad + 1 + \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{s_i=\ell}^{t-1}2t^{-4}\\
&\leq \frac{2\phi\bar{R}^{i}\sum^{\alpha}{}_{\mathrm{k}=1}\mathrm{kB(k)}}{\Delta_i} + \frac{4\ln\ t(\bar{\mathrm{R}}^{\mathrm{i}})^{2}\sum^{\alpha}{}_{\mathrm{k}=1}(\mathrm{B(k)})^{2}}{\Delta_i^{2}}\Bigg)\\
&\quad\cdot\left(1 + \sqrt{1 + \frac{\Delta_i\phi\bar{R}^{i}\sum^{\alpha}{}_{\mathrm{k}-1}\mathrm{kB(k)}\ln}{t^{-2}\sum^{\alpha}}}\right)\\
&\quad(\mathrm{R_i})_{\mathrm{k}=1}(\mathrm{B(k)})^{2}\Bigg)\\
&\quad + 1 + \frac{\pi^{2}}{3}
\end{aligned}
$$

定理陈述可由以下事实得出：$\mathcal{R}_T(\mathfrak{U}_{\mathrm{FR}}) = \sum_{i:\ \mu_i<\mu^{*}}\Delta_i\mathbb{E}\left[N_i(T)\right]$。

C 实验环境细节

C.1 技术细节

代码在一台服务器上运行，该服务器配备 2 个 Intel Xeon 4110 2.1Ghz（32 个超线程）中央处理器和 384GB 内存。仿真过程中未使用 GPU 加速。操作系统为 Ubuntu 16.04.7 LTS。仿真代码使用 Python 3.9.12 编写。此外，我们使用 Conda 管理库，实验中使用了以下库：

- NumPy 1.23.4
- pandas 1.5.1
- tqdm 4.64.1
- SciPy 1.9.3
- Matplotlib 3.5.3

由于本研究部分基于[Romano 等人 2022], 的研究成果，我们以他们的代码为基础进行了调整，以便为我们的学习器运行实验[3]。总体而言，在上述硬件上运行所有设置约需 96 小时。

C.2 TP-UCB-FR-G学习器配置

我们将采用贝塔-二项分布变体的新学习器引入实验设置中。首先简要介绍贝塔-二项分布本身，随后介绍我们提出的变体。有关贝塔-二项分布的更完整概述，请参阅 [Lee 2012]。

设 $N \in \mathbb{N}$、 $\alpha \in \mathbb{R}^+$ 和 $\beta \in \mathbb{R}^+$，并注意，在此设置中， α 是一个任意常数，而非本文其余部分中的离散度参数。对于 $x \in \{0, .., N\}$，概率质量函数定义如下：

$$BetaBinom(N, \alpha, \beta)(x) = \binom{N}{x} \frac{B(x+\alpha, N-x+\beta)}{B(\alpha, \beta)} \tag{15}$$

其中， $B(u, v)$ 是某些 u、 $v \in \mathbb{R}$ 对应的贝塔函数。可以看出，该分布的定义域并不等于满足 β-离散度性质所需的 $\{1, 2, \ldots, \alpha\}$ 的 α 大小定义域。因此，我们在公式 16 中创建了一个定义域属于该范围的 Beta-二项分布变体。

$$BetaBinomVariant(x) = BetaBinom(\alpha - 1, a, b)(x - 1) \tag{16}$$

其中， α 表示 z-群的数量，a 和 b 为分布参数。该函数的定义域是所需的整数区间 $\{1, 2, \ldots, \alpha\}$。请注意，这种变体不再适用计算均值或方差的公式。对于每种实验设置，我们都会添加 8 个按照这种变体分布的合成学习器。其具体参数见表 1，相应的概率质量函数也绘制在图 6 中。请注意，分布名称是根据其在 x轴上的“中心”选取的。也就是说，该位置是 x轴上观测到的概率最高的位置。

[3]代码随补充材料一并提供。

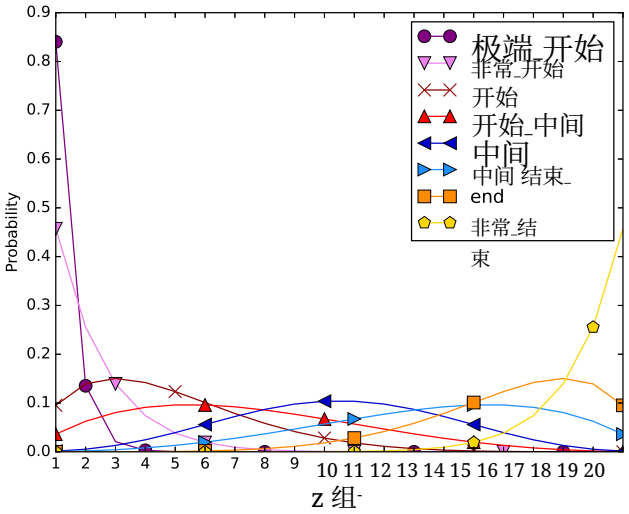

图 6：不同贝塔-二项配置的概率质量函数

C.3 实验设置

本节详细介绍第 6 节中所述的实验设置，以及为验证正文中的结果而开展的一些后续实验。

设置 1

此设置的主要目标是探究调整 α 对 TP-UCB-FR-G 学习器的影响。我们针对 α、 $\alpha_{est} \in [5, 10, 20, 25, 50]$ 的不同取值运行此设置。在此设置中，我们对 $k = 10$ 个臂进行建模，在 $\tau_{\max} = 100$ 轮次中收集奖励，并将最大奖励设为 $\overline{R}^i = 100i$。[Romano 等人 2022] 提出的学习器的累计奖励满足 $Z^i_{t,\ k} \sim \frac{\overline{R}^i}{\alpha} U[0, 1]$。此设置中的 TP-UCB-FR-G 学习器假设采用不同的累计奖励结构，即 $Z^i_{t,\ k} \sim \frac{\overline{R}^i}{\alpha} Beta(a^i_k, b^i_k)$。实验在时间范围 $T = 10^5$ 内进行，α-平滑常数为 $\alpha = 20$。结果取 100 次独立运行的平均值。

第 6 节给出了 $\alpha_{est} = 50$ 的结果。在本节中，我们重点关注 TP-UCB-FR-G 在$\alpha_{est} = [5, 10, 20, 25]$时的结果。这些结果分别绘制在图 7、图 8、图 9 和图 10 中。这些结果与第 6 节的结论一致。也就是说，随着 α_{est} 增加，以开始为中心的学习器表现得越来越好，优于其他学习器。对于 $\alpha_{est} = 5$，我们可以看到，对于较大的 t，以开始为中心的学习器表现更好。这表明，在 α_{est} 值较低的情况下，开始过于乐观。

表 3 总结了设置 1 中使用 begin 分布时的完整性能增益 δ^+ 。请注意， δ^+ 表示相对于 TP-UCB-FR 的遗憾，平均遗憾的降幅百分比。

[4]设置 1 中其他分布的数据见补充材料

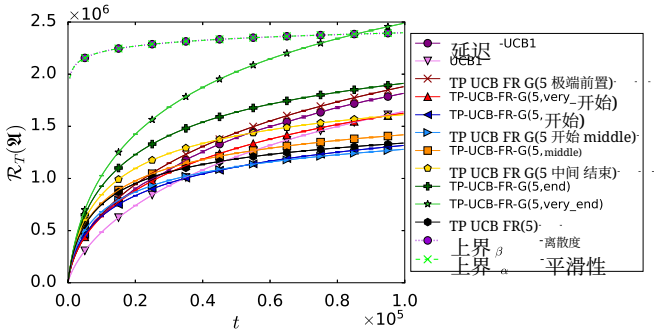

图7：设置 1 中遗憾随时间的变化（ $\alpha_{est} = 5$）

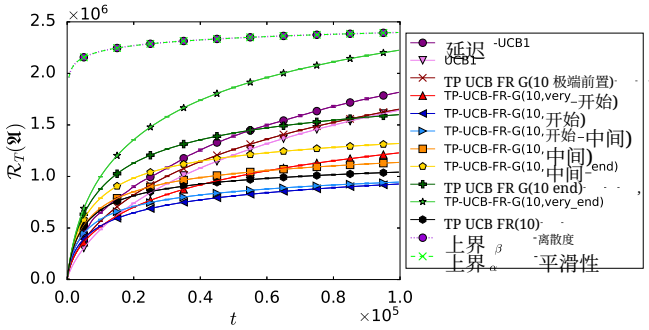

图 8：设置 1 中遗憾随时间的变化（ $\alpha_{est} = 10$）

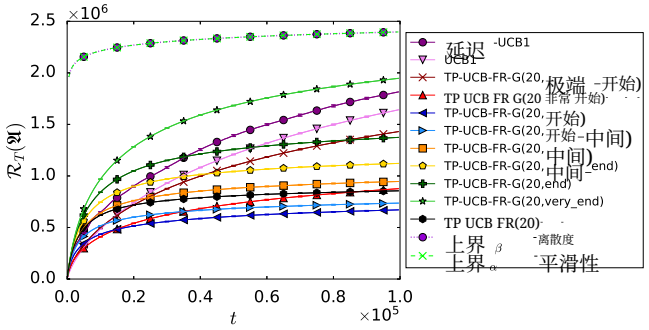

图 9：设置 1 中 $\alpha_{est} = 20$ 随时间变化的遗憾

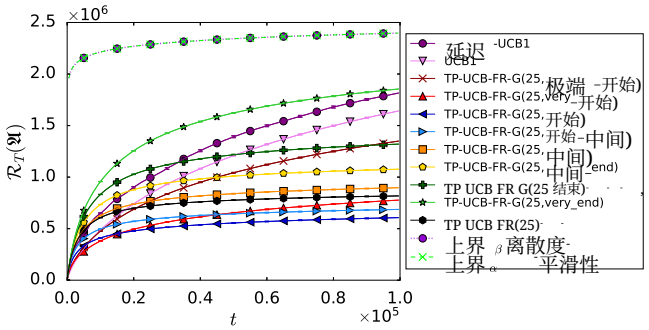

图 10：设置 1 中遗憾随时间的变化，其中 $\alpha_{est} = 25$

α	遗憾 TP-UCB-FR	遗憾 TP-UCB-FR-G	δ^+ (%)
5	1.34×10^6	1.28×10^6	4.6*
10	1.04×10^6	9.27×10^5	10.9
20	8.56×10^5	6.71×10^5	21.6
25	8.18×10^5	6.07×10^5	25.8
50	7.21×10^5	4.64×10^5	35.6

表 3：TP-UCB-FR-G 在设置 1 中带来的性能增益总结。*对于 $\alpha_{est} = 5$，我们展示起始中间分布，因为这是唯一一种性能优于其他地方隐含假设的 begin 分布的配置。

设置	参数向量
	a^i
均匀	$\mathbf{1}_\alpha$
设置 2.1	[2,4,6,8,10,10,10,10,10,10]
设置 2.2	[10,10,10,10,10,10,8,6,4,2]
	b^i
均匀	$\mathbf{1}_\alpha$
设置 2.1	[10,10,10,10,10,10,8,6,4,2]
设置 2.2	[2,4,6,8,10,10,10,10,10,10]

表 4：配置 1 下不同场景的分布参数

设置 2

在第二种设置中，我们评估了不同（非均匀）数据生成分布对延迟部分奖励的影响。此外，还测试了具有更高 $\tau_{\max}$ 的配置。我们再次对 $K = 10$ 个臂进行建模，假设每次拉取臂后都会获得累计奖励，且 $Z^i_{t,\,k} \sim \frac{R^i}{\alpha} Beta(a^i_k, b^i_k)$。时间范围设为 $T = 10^5$。我们评估了四种配置和三个场景。这四种配置与 $\tau_{\max}$ 和 α 相关。三个场景的区别在于拉取臂后部分奖励的分布。第一个场景采用均匀聚合奖励分布，与设置 1 相同。其余两个场景的奖励分别在 $\tau_{\max}$ 区间末尾（设置 2.1）较高，以及在拉取臂后不久（设置 2.2）较高。向量 a^i 和 b^i 表示 $k \in \{1, \ldots, \alpha\}$ 中的所有 a^i_k 和 b^i_k 值，可在引用表格 4、5、6 和 7 中找到。

- 配置 1：$\tau_{\text{最大}} = 100$，$\alpha = 10$，见表 4
- 配置 2 $\tau_{\text{最大}} = 100$，$\alpha = 50$，见表 5
- 配置 3 $\tau_{\text{最大}} = 200$，$\alpha = 20$，见表 6
- 配置 4 $\tau_{\text{最大}} = 200$，$\alpha = 100$，见表7

结果

如主论文第 6 节所述，同一配置下不同场景的结果在视觉上无法区分。同一节中对图 3 的分析表明，学习器数量增加时并无显著差异。分析还表明，所有学习器的表现

设置	参数向量
	a^i
均匀	$\mathbf{1}_\alpha$
设置 2.1	[2,4,. . ., 48、50、. . ., 50]
设置 2.2	[50,. . ., 50,48, . . ., 4,2]
	b^i
均匀	$\mathbf{1}_\alpha$
设置 2.1	[50,. . ., 50,48, . . ., 4,2]
设置 2.2	[2,4,。。。, 48,50, 。。。, 50]

表 5：配置 2 下不同场景的分布参数

设置	参数向量
	a^i
均匀	$\mathbf{1}_\alpha$
设置 2.1	[2,4,. . . , 18,20, . . . , 20]
设置 2.2	[20,. . . , 20,18, . . . , 4,2]
	b^i
均匀	$\mathbf{1}_\alpha$
设置 2.1	[20,. . . , 20,18, . . . , 4,2]
设置 2.2	[2,4,. . . , 18,20, . . . , 20]

表 6：配置 3 下不同场景的分布参数

设置	参数向量
	a^i
均匀	$\mathbf{1}_\alpha$
设置 2.1	[2,4,. . . , 98, 100, . . . , 100]
设置 2.2	[100,. . . , 100,98, . . . , 4,2]
	b^i
均匀	$\mathbf{1}_\alpha$
设置 2.1	[100,. . . , 100,98, . . . , 4,2]
设置 2.2	[2,4,. . . , 98,100, . . . , 100]

表7：配置4下不同场景的分布参数

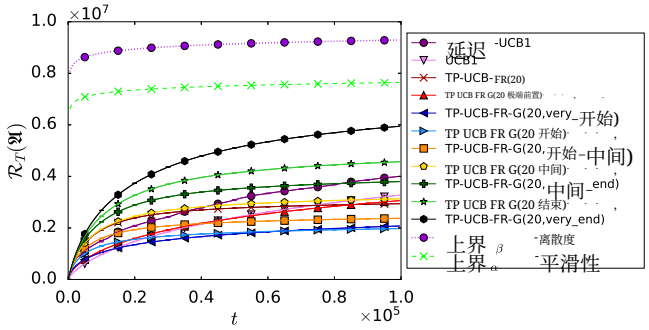

图 11：设置 2.1 中遗憾随时间的变化， $\tau_{\max}$ =200， $\alpha_{est} = 20$

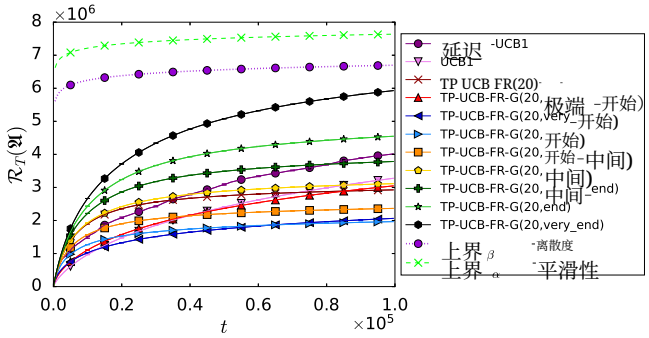

图 12：设置 2.2 下遗憾随时间的变化， $\tau_{\max}$ =200, $\alpha_{est} = 20$

在设置 2.2 中最佳，在均匀设置中较差，在设置 2.1 中最差。由于无法通过视觉加以区分，配置 1 和配置 2 的结果在视觉上与设置 1 中使用 $\alpha_{est} = 10$和 $\alpha_{est} = 50$的结果完全相同。同样，使用均匀分布进行设置 2 测试的结果，在视觉上与设置 2.1 和 2.2 的结果完全相同。为了展示我们的上界相较于 TP-UCB-FR 上界的变化，我们给出了在设置 2.1 和 2.2 下运行配置 3 和 4 的结果。所得图表见图 11、12、13 和 14。此外，表 8、9 和 10 完整总结了使用 begin 分布[5]时，设置 2 中不同配置的性能增益 δ^+ 。

[5]设置 2 中其他分布的数据见补充材料

$\tau_{\max}$	α	遗憾 TP-UCB-FR	遗憾 TP-UCB-FR-G	δ^+ (%)
100	10	1.04×10^6	9.27×10^5	10.9
100	50	7.19×10^5	4.66×10^5	35.2
200	20	2.94×10^6	1.98×10^6	32.7
200	100	2.45×10^6	1.27×10^6	48.2

表8：设置 2.1 中 TP-UCB-FR-G 带来的性能增益 δ^+

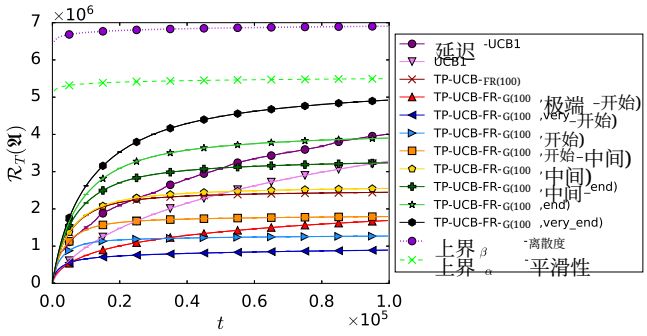

图 13：设置 2.1 中遗憾随时间的变化，$\tau_{\max}$ =200, $\alpha_{est} = 100$

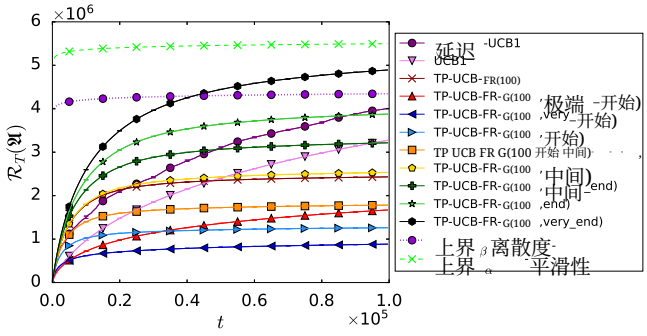

图 14：设置 2.2 中遗憾随时间的变化，$\tau_{\max}$ =200, $\alpha_{est} = 100$

$\tau_{\max}$	α	遗憾 TP-UCB-FR	遗憾 TP-UCB-FR-G	δ^+ (%)
100	10	1.04×10^6	9.25×10^5	11.0
100	50	7.17×10^5	4.64×10^5	35.3
200	20	2.93×10^6	1.97×10^6	32.8
200	100	2.43×10^6	1.26×10^6	48.1

表 9：TP-UCB-FR-G 在设置 2.2 中的性能增益总结

$\tau_{\max}$	α	遗憾 TP-UCB-FR	遗憾 TP-UCB-FR-G	δ^+ (%)
100	10	1.04×10^6	9.28×10^5	10.8
100	50	7.18×10^5	4.65×10^5	35.2
200	20	2.94×10^6	1.98×10^6	32.7
200	100	2.44×10^6	1.27×10^6	48.0

表10：TP-UCB-FR-G 在设置 2 均匀条件下的性能增益 δ^+

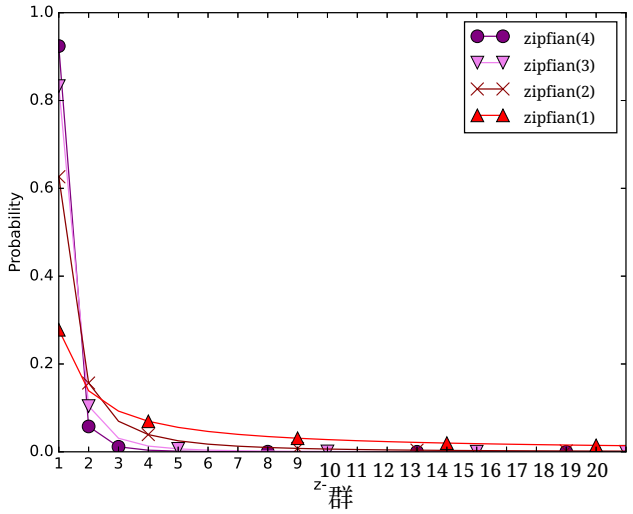

图15：不同参数下齐夫分布的概率质量函数

Spotify 场景

Spotify 场景采用与 [Romano 等人 2022] 中详细说明的相同预处理和配置，以准确比较其 TP-UCB-FR 算法与我们提出的 TP-UCB-FR-G 算法的性能。我们对 100 次独立运行的结果取平均值。该场景的结果见图 5。

C.4 补充实验

作为**TP-UCB-FR-G**输入的替代分布

在第 6 节中，我们描述了使用参数取值不同的贝塔-二项分布时，各种设置所得到的结果。通常，分布的定义域必须限定在区间 $[1, \alpha]$ 内，才能将其用作输入。下面，我们介绍几种可以限制为具有此类定义域的有限离散分布，以及实验结果。

齐夫分布

The Zipfian distribution 可以让我们用不同的参数描述“以开始为导向”的分布，如图15所示。鉴于我们观察到以开始为导向的分布与我们的算法配合良好，齐夫分布在这种设置下也应表现出色。

我们来考虑设置 1。我们针对α_{est} = 20开展实验，这是对平滑因子的完美估计。此外，我们只考虑以开始为导向的学习器。我们添加了 4 个采用齐夫分布的学习器，参数为 $s \in [1, 2, 3, 4]$，名称分别为 zipfian(1)、zipfian(2)、zipfian(3) 和 zipfian(4)。我们观察到，TP-UCB-FR-G zipfian(1) 学习器的表现优于 [Romano 等人 2022], 提出的学习器，但不如以开始为导向的贝塔-二项学习器。这些结果表明，对于 $\zeta \geq 2$，zipfian(ζ) 学习器的表现不如大多数贝塔-二项学习器，也无法提升结果。

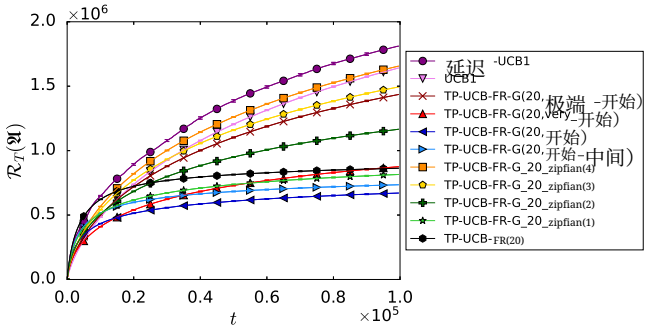

图 16：设置 1 下齐夫分布与贝塔-二项分布的比较（$\alpha_{est} = 20$）

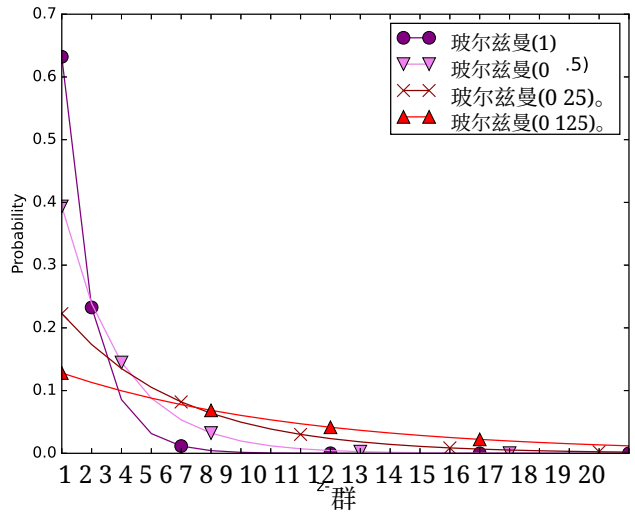

图 17：不同参数的玻尔兹曼分布概率质量函数

玻尔兹曼分布

同样，我们也可以使用玻尔兹曼分布来描述“以开始为导向”的分布。我们将输入分布记为 boltzmann(λ)。我们注意到，随着 $\lambda \to 0$，该分布呈现出更平滑的开始导向性。这个实验尤其有趣，因为对于 $\lambda = 0.25$ 和 $\lambda = 0.125$，该分布以开始为中心进行了精心设计，因此可能带来性能提升。我们再次考虑设置 1。我们针对 $\alpha_{est} = 20$ 开展实验，这代表了对平滑因子的完美估计。我们考虑以开始为导向的贝塔-二项学习器，并为 $\lambda \in [1, 0.5, 0.25, 0.125]$添加玻尔兹曼分布学习器。该实验的结果见图 18。我们观察到，玻尔兹曼分布，尤其是 Boltzmann(0.5)，对于 $t = [0, 15000]$的表现略好，而且其曲线与非常靠前的学习器几乎完全相同。这是合乎逻辑的，因为图 6 和图 17 中的曲线几乎完全相同。还值得注意的是，对于 $t = [0, 38000]$，Boltzmann(0.25) 的表现略好于开始。对于 $t > 38000$，开始仍然是遗憾最低的学习器。由于

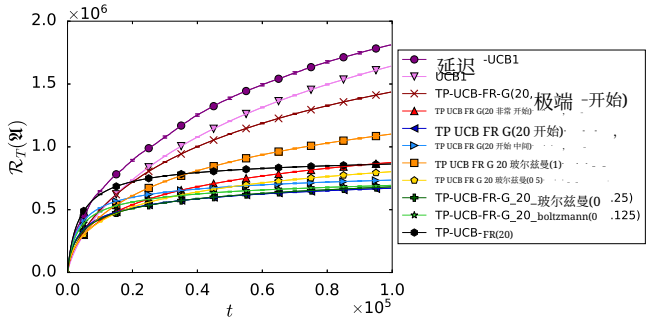

图 18：设置 1 下 Boltzmann 与贝塔-二项分布的比较，$\alpha_{est} = 20$

$\lambda = 0.125$ 开始低估，而 $\lambda = 1$、$\lambda = 0.5$ 明显高估，因此我们有充分证据表明，这种分布缺乏所需的灵活性，无法在性能增益方面胜过贝塔-二项分布。该实验的精确遗憾值可在补充材料中找到。我们排除极端前置和 Boltzmann(1)，因为它们明显高估；同时排除 Delayed-UCB 和 UCB，因为它们与此比较无关。

根据上述结果，我们得出结论：贝塔-二项分布仍然具有最佳性能和灵活性，而玻尔兹曼分布在这两个方面都表现不足。

超几何分布

对于 mathematical description of

超几何分布，我们参考 [Lee, 2012]。简而言之，若随机变量X 的概率质量函数定义如下，则称其服从超几何分布

$$p_X(k) = P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{17}$$

其中，N为总体规模，K为总体中的成功次数，n为抽取次数，k为概率质量函数所求概率对应的成功数。我们将n和K设为$\alpha - 1$，即比 z 群的数量少一，并将输入的k的每个值都加 1（与我们的贝塔-二项分布实现类似），从而将 $p_X(k)$限制在定义域 $[1, \alpha]$内。随后，我们可以通过选择 $N \geq 2\alpha$来调整分布的形状。本节中，我们将使用超几何分布的学习器记作 Hypergeom(N)，其中塑形参数 $N \in \mathbb{N}$ 和 $N \geq 2\alpha = 40$。我们将介绍几种面向开始的超几何分布，用于 $N \in [50, 100, 200, 300, 400, 500]$。图 19 展示了 N对应的概率质量函数。

我们将纳入以开始为导向的贝塔-二项式学习器进行比较，并使用 $\alpha_{est} = 20$ 运行设置 1，此时平滑因子估计完美。该实验的结果如图 20 所示

再次观察到，在整个时间范围 T 内取平均时，TP-UCB-FR-G(20, begin) 是表现最佳的学习器。然而，对于 $t \in [0, 15000]$，采用超几何分布的 hypergeom(N)（$N \in [200, 300, 400, 500]$）表现略好。由于超几何分布所对应曲线的斜率，这些分布的表现会逐渐下降，

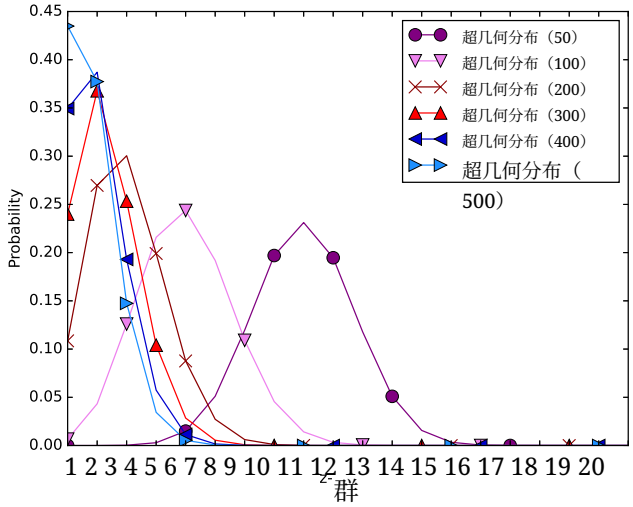

图19：不同参数的超几何分布的概率质量函数

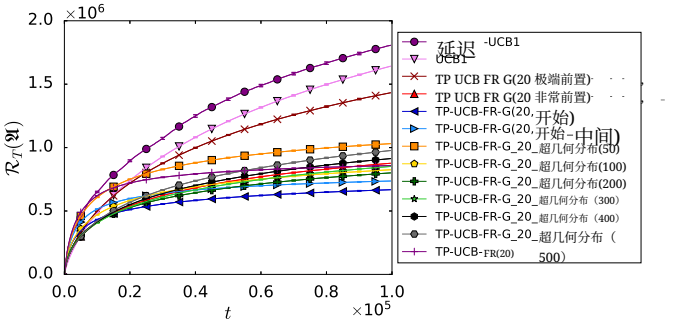

图 20：设置 1 中超几何分布与贝塔-二项分布的比较，使用 $\alpha_{est} = 20$

随着 t 变大。由于 N 之间存在较大差距，可以搜索 N 的最合适取值，或许能找到一种表现优于 begin 贝塔-二项学习器的分布，但这需要针对 $200 \leq N \leq 500$ 运行设置 1，代价非常高。此外，我们也可以用类似方式微调贝塔-二项学习器：在 begin 附近针对不同的 α、β 运行设置 1，这样所需的运行次数会少得多。再次说明，贝塔 - 二项分布在灵活性和性能方面仍然更胜一筹。

假设分布与上界之间的关系如定理 2 所述，只有当作为输入提供给算法的假设分布与数据生成分布相匹配时，TP-UCB-FR-G 学习器的上界才成立。本节将深入探讨这一点，并讨论在假设分布不匹配的情况下，哪类学习器仍能保持在该上界之下。

我们注意到，尽管 [Romano 等人 2022], 并未明确提及这一现象，但它同样存在于 TP-UCB-FR 算法的上界中，而且取决于数据生成分布的 α，而非作为输入给出的估计值。由于该算法的假设分布是单变量的（仅取决于 α），因此更容易检验哪些学习器不会超过上界：在所有测试的设置中，提高 α_{est} 总能带来更好的结果（见第 6 节）。这意味着，当 $\alpha_{est} \in [\alpha, \tau_{\max}]$ 时，学习器通常会保持在上界以下。

然而，TP-UCB-FR-G 的上界取决于数据生成分布的三个性质，即其期望值、重合指数和 α_o。从结果中（见第 6 节）可以看出，假设期望值较低的学习器通常表现更好。但假设期望值也不能过低，因为第二个因素——重合指数——在非常偏向 begin 的分布中会相对较高，从而增大遗憾。这使得难以通过经验估计判断哪类学习器能够保持在理论上界以下。这也表明，与 TP-UCB-FR 的情况类似，α 较高的学习器表现更好。推导这三个参数之间的精确相关性非常复杂，可作为未来的研究方向。