多臂老虎机：模型、理论与应用

清华大学经济管理学院　王天宇　郭紫辰

2020 年 6 月 19 日

摘要

多臂老虎机问题是强化学习中常见的一种问题设置。我们总结了几种经典多臂老虎机算法（贪婪算法、$\varepsilon-$贪婪算法、UCB、汤普森采样）与上下文老虎机算法(LinUCB、CTS）的应用场景和理论基础，并讨论了它们的优缺点。在此基础上，我们主要聚焦于CMAB在学界的应用，使用Mathematica实现了算法仿真，得到了各种算法的相对性能比较, 结果与目前学术界真实数据集结论基本一致。文末展望了多臂老虎机算法的应用前景和未来发展。在交叉学科的融合中, 提出未来发展方向是拓展到多维臂定义、约束和信息反馈等。

关键词: 多臂老虎机, 汤普森采样, 上下文信息, 推荐系统

注：本文正文(不含附录和摘要) 共**4984**字。

1 问题介绍

与已知样本数据的统计学习相比，强化学习侧重于在与环境的交互中训练一个能自动选择动作、并根据动作形成的反馈动态调整自己的策略的算法。由于强化学习模式更接近人的思维方式，因此正逐渐成为学界研究的主流，如自动驾驶、多智能体决策等。(Sutton and Barto (2018))

而在线学习则可以看作是强化学习的一个特例。在线学习的核心便是"**探索-利用**"的权衡。探索是为了更好了解陌生环境的性质，而利用便是对该环境性质去优化。倘若对环境探索次数过多，固然可以有效认识所在环境，但是其往往前期损失过大；倘若探索不足，则会盲目选择自己认为的最优解，有时会与最优结果相差甚远。而如何在给定的时间段内尽可能最大化收益减少后悔，便是在线学习关注的问题。

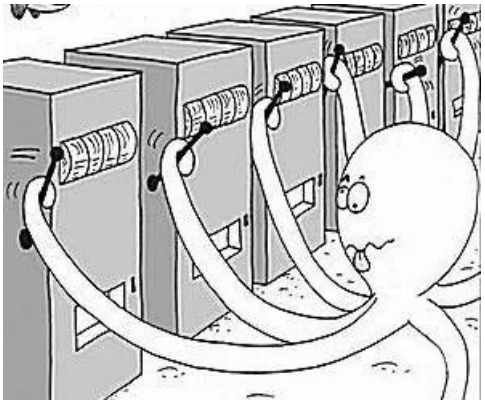


图 1: 多臂老虎机问题图示

而作为在线学习的代表, **多臂老虎机(multi-armed bandits**, 以下简称MAB)则介绍了这样一个模型。假设我们在游戏厅，手上有T个游戏币，有K个老虎机(臂)排放在我们面前。我们给老虎机按顺序标号，老虎机i以p_i概率吐出一元钱(p_i对于我们是未知的)，而每摇一次老虎机需要花费1游戏币，也就是说我们最多能够摇T次，那么如何使得我们最终收益最大呢？

在这里衡量算法的指标是**后悔(regret)**。考虑前t次实验，我们可以通过算法得到的累计收益与可能的预期最优收益进行比较，因此累积t时间的后悔期望值可记为：

$$R(T) := \max_{a \in A} \mu(a)T - \sum_{s=1}^{T} \mu(a_s).$$

如下不同算法便会对$R(T)$不同的后悔阶数，如$O(logT), O(T)$等。而我们理论上希望算法能够使得R(T)的阶数更小。这样从长期来看($T \to \infty$)算法便是更好的。

2 算法分析

在本节对算法的证明中，我们仅考虑$K = 2$的情形。$K \geq 2$的情形类似。

2.1 贪婪算法和$\varepsilon-$贪婪算法

首先我们考虑最简单的情况，即对每个臂均先探索若干次。达到某一时间点的时候便选择之前平均收益最好的臂。也即把整个时间段自然地拆成探索和利用两段，算法如下[1]： 在这里，更多的

Algorithm 1 简单贪婪算法示例

1.探索阶段：对每个臂均尝试N次;
2.选择其中具有最高平均收益的臂$\hat{a}$;
3.利用阶段：在之后每轮均采用臂$\hat{a}$.

探索使得在利用阶段即使我们选择不为最优的臂，从统计意义看仍然能够得到不太差的结果。为了计算算法的后悔阶，我们首先给出如下的**Hoeffding不等式**[1]来分析，对某个臂a来说在探索N次之后，$\widehat{\mu(a)}$为N次之后的平均收益，而$\mu(a)$为其真实收益。

$$Pr\left\{|\widehat{\mu(a)}-\mu(a)|\leq\sqrt{\frac{2logT}{N}}\right\}\geq 1-\frac{1}{T^4} \tag{1}$$

以K = 2为例，假设[1]左边成立，易见探索阶段的后悔值最多为N, 而利用阶段每次后悔值可分析得为$O(\sqrt{\frac{logT}{N}})$(见附录)。我们有因此总共T阶段的后悔值：

$$R(T)\leq N+O(\sqrt{\frac{logT}{N}}\times(T-2N))\leq N+O(\sqrt{\frac{logT}{N}}\times T).$$

倘若选择 $N=T^{\frac{2}{3}}(logT)^{\frac{1}{3}}$, 使得不等式右边最小，则有$R(T)\leq O(T^{\frac{2}{3}}(logT)^{\frac{1}{3}})$。结合1不成立的情况分析可知$\mathbb{E}[R(T)]\leq O(T^{\frac{2}{3}}(logT)^{\frac{1}{3}})$.而对$K>2$的情形，我们有如下定理(见Slivkins (2019))：

定理 1. *简单贪婪算法得到的后悔*$\mathbb{E}[R(T)]\leq T^{\frac{2}{3}}\times O(KlogT)^{\frac{1}{3}}$

然而，贪婪算法在探索阶段的表现非常差，因此我们尝试对算法引入一定的随机因素，便得到了如下的$\varepsilon-$贪婪算法和相应的后悔界:

Algorithm 2 $\varepsilon-$贪婪算法示例

for 每阶段 t = 1, 2, ..., T **do**
  选取[0, 1]的随机数ε
  if $\varepsilon\leq\varepsilon_t$ **then**
    探索：任意选臂;
  else
    利用：选取目前最高平均收益的臂.
  end if
end for

定理 2. *如果选取*$\varepsilon_t=t^{-\frac{1}{3}}(Klogt)^{\frac{1}{3}}$, $\epsilon-$*贪婪算法得到的后悔*$\mathbb{E}[R(t)]\leq t^{\frac{2}{3}}\times O(Klogt)^{\frac{1}{3}}$

尽管两种贪婪算法的界看似一致，$\varepsilon-$贪婪算法的t是任意的，即任意时刻均有这样的后悔上界，因此相对贪婪算法会更好。然而上述的贪婪算法在探索阶段都较为盲目，并不能根据实时的探索反馈进行策略的调整。基于这样的思想, Auer (2002)便提出了如下的适应性算法:

[1]这其中N是与T有关的预先确定的固定值

2.2 基于置信区间(UCB)的算法

在某个臂a采样$n_t(a)$次之后，使用Hoeffdling不等式得到的置信区间为:

$$UCB_t(a) = \widehat{\mu_t(a)} + \sqrt{\frac{2logT}{n_t(a)}}^2$$

$$LCB_t(a) = \widehat{\mu_t(a)} - \sqrt{\frac{2logT}{n_t(a)}}$$

随着$n_t(a)$的增加, $[LCB_t(a), UCB_t(a)]$的长度会渐渐变小。我们便会发现某些臂会比其他臂表现更显著, 也即该臂的下界大于其他臂的上界： 对K = 2的后悔界证明见附录。对一般的K，我们有如下

Algorithm 3 消除算法示例

保留每个臂
for 每个大阶段直到结束 **do**
  尝试所有可能的臂;
  删除所有的臂a, s.t. $\exists$ arm a' with $UCB_t(a) < LCB_t(a')$;
end for

的结论(见Slivkins (2019)):

定理 3. *UCB类算法得到的后悔*$\mathbb{E}[R(t)] \leq O(\sqrt{KtlogT}), \forall t \leq T$

而总体上，UCB类算法能够得到的后悔还有:

$$\mathbb{E}[R(T)] \leq O(logT) \left[\sum_{arms\ a\ with\ \mu(a)<\mu(a^*)} \frac{1}{\mu(a^*) - \mu(a)} \right]$$

可以看到, UCB算法$O(logT)$的后悔界是其相比贪婪算法的最大优势, 主要原因便在于其利用阶段的步骤都根据反馈动态调整。因此在后续的研究中得到了很多推广，如Bayes UCB等(Kaufmann et al. (2012))。

2.3 汤普森采样(TS)

贪婪算法的一大劣势便是主动探索不足。汤普森采样(Thompson Sampling，下文简称TS)也即后验采样(posterior sampling), 由Thompson (1933)提出。假设收益分别为服从Bernouli分布, 则收益整体的先验分布为Beta分布，那么其后验分布仍为Beta分布。而TS相比贪婪算法的核心优势便是根据已知数据从模型中随机采样而不是简单地比较平均值。

根据贝叶斯更新，我们有下面的算法规则：

[2]UCB, 即upper confidential bound

Algorithm 4 TS算法示例

for t = 1, 2, ..., T **do**
  for k = 1, 2, ..., K **do**
    对每个臂值采样$\hat{\mu_k} \sim Beta(\alpha_k, \beta_k)$;
  end for选取$\hat{\mu_k}$最大的作为本次的选择臂a_t;
  选择a_t并观察到收益r_t. 更新分布$(\alpha_{a_t}, \beta_{a_t}) \Leftarrow (\alpha_{a_t} + r_t, \beta_{a_t} + 1 - r_t)$
end for

TS的好处便是其在采样同时完成探索-利用两个任务，由Beta分布的性质(见附录)可知，倘若某一臂之前选择较小，其采样概率便较为分散；否则对采样的参数估计较准, 好的臂也更容易被选中利用。而其复杂度如下(Slivkins (2019))：

定理 4. *TS算法得到的后悔*$\mathbb{E}[R(t)] \leq O(\sqrt{KtlogT}), \forall t \leq T$

TS算法取得了和UCB算法同样的后悔界，这是由于其在证明分析中Russo et al. (2017)利用UCB的值进行后悔分解。也可看作是UCB和TS算法的联系。

而在一般的设置下，TS算法也被推广至MCMC, Gibbs采样，自助采样等估计方法Russo et al. (2017)。由于其采样的可操作性得到了广泛应用，相比UCB这种需要了解精确分布的算法更具有现实意义(Osband and Van Roy (2017))，并且其在连续空间的探索表现更好(Russo and Van Roy (2014))。如在动态定价(Ferreira et al. (2018))、推荐系统等领域。

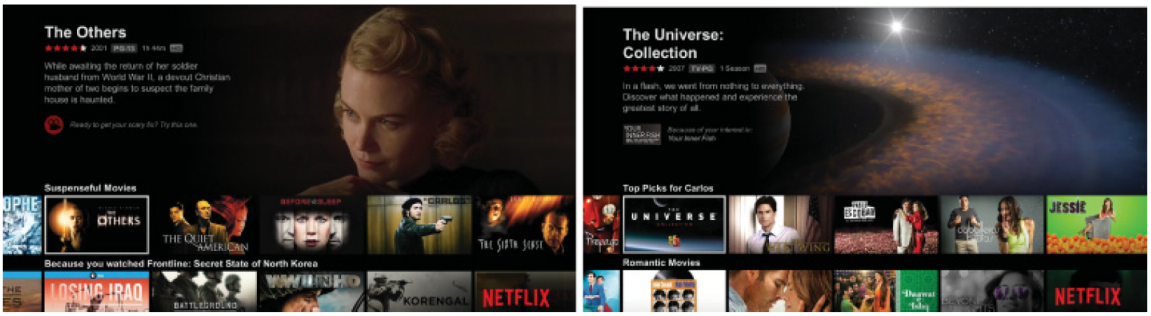
3 从CMAB到推荐系统

上下文老虎机问题(Contextual MAB, 下文简称CMAB)即基于上下文信息有效选择行动。相比简单的MAB设置，CMAB核心要求用户的异质性，表示不同用户特征的特征向量如身份、地理信息等均存在较大差异。因此不同商品（臂$a \in A_t$）在不同用户的效用差别较大，会导致经典MAB模型表现不佳。

CMAB考虑这样一个问题: 在时刻t，观察到当前用户信息$x_{t,a}$ ，随后根据已有的知识选择商品a_t展示给用户并得到。这样一阶段后系统便得到了$(x_{t,a}, a_t, r_{t,a_t})$这样的信息。我们的目标仍然是选择使得使得下式最小：

$$R_\pi(T) = \mathbb{E}\left[\sum_{t=1}^{T} r_{t,a_t^*}\right] - \mathbb{E}\left[\sum_{t=1}^{T} r_{t,a_t}\right].$$

CMAB作为MAB个性化的推广，目前主要被应用在推荐系统中。虽然传统推荐系统本身有成熟的协同过滤、基于内容等算法(Ricci et al. (2011))，比如通过用户历史行为、内容描述等特征给用户选择, 但其在新用户时会遇到冷启动并且随时间用户会兴趣漂移等问题。这种问题可以用MAB模型的思路去解决。



基于探索-利用的思想和已有的算法, Li et al. (2010)便提出了LinUCB算法。他们首先对用户信息和收益(用户点击率, click-through rate, CTR)的关系给出如下的线性假设：

$$E[r_{t,a}|x_{t,a}] = x_{t,a}^T \theta_a$$

这里θ_a也就是LinUCB算法的参数，然后他们通过对训练集(D_a, c_a)进行岭回归估计用户对商品的效用:

$$\min_{\theta_{a,j}} \sum_{i=1}^{m} \left(c_{a,i} - \sum_{j=1}^{d} \theta_{a,j} x_{ij} \right)^2 + \sum_{i=1}^{d} \theta_{a,j}^2$$

这样求解得到系数的估计：

$$\hat{\theta}_a = (D_a^T D_a + I_d)^{-1} D_a^T c_a.$$

而由根据UCB算法对θ_a置信区间的估计，他们在论文中提出的算法如下：

Algorithm 5 LinUCB算法示例

for t = 1, 2, ..., T **do**
  for k = 1, 2, ..., K **do**
    if a 之前未设置参数 **then**
      设置A_a为单位阵;
      设置b_a为d维零向量。
    end if
    $\hat{\theta_a} \Leftarrow A_a^{-1} b_a$;
    $p_{t,a} \Leftarrow \hat{\theta_a}^T x_{t,a} + \alpha \sqrt{x_{t,a}^T A_a^{-1} x_{t,a}}$.
  end for
  选取$\hat{\mu_k}$最大的作为本次的选择臂a_t;
  选择a_t并观察到收益r_t.
  更新参数$A_{a,t} \Leftarrow A_{a,t} + x_{t,a_t} x_{t,a_t}^T, b_{a_t} \Leftarrow b_{a_t} + r_t x_{t,a_t}$
end for

这样的算法可以被推广至一般意义上的线性收益模型，也可以使用基于上下文的汤普森采样去解决(CTS, Agrawal and Goyal (2013))。整体来说这样突破了传统推荐系统贪婪方法的桎梏，为算法提供了处理兴趣漂移问题的可能。并且能够达到本问题的后悔上界$O(\sqrt{KdT})$.

而倘若问题的目标函数结构不清晰, 在近期关于CMAB的研究中，也有其他思路如在给定策略集中进行选择可能最好的策略，首先对选取前N次数据探索收集并评价不同臂的收益，然后利用贪婪或UCB算法选择进行利用Slivkins (2019)。

更为广泛地，CMAB象征了在实际决策考虑辅助信息的情形。在实际问题中考虑辅助信息与决

策变量的相关，是人工智能与运筹学结合的重要前景。不过CMAB算法在真实推荐系统的工业界得到应用，如下图仍然特征工程数据提取等的参与，也需要算法能够实时收集用户反馈，需要较高要求的流计算框架等。如下图则是CMAB对系统从学习到数据收集探索进行的循环，因此算法落地仍然需要工业界框架思维的改进。

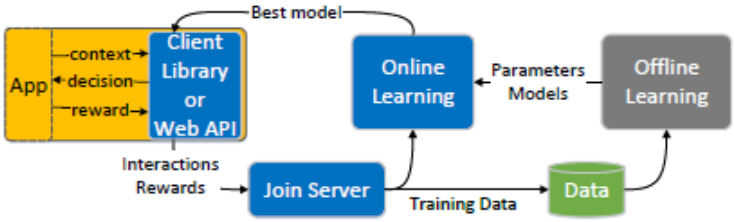


图 2: CMAB问题系统设计(https://zhuanlan.zhihu.com/p/35753281)

4 案例研究

多臂老虎机的各种算法注重信息的渐进获取、信息获取成本的慎重计算、已知信息的最大利用，因此对低信息获取量、高信息获取成本或信息变动迅速的实际应用问题具有重要的现实意义。鉴于多臂老虎机算法的以上特征，该算法在不同场景下得到了广泛运用。一些常见的应用场景包括：

(1)动态推荐网站对访客的界面设计展示和广告, 如最大化CTR;

(2)流量大的电子商务平台获取客户偏好、设计动态定价(最大化销售利润);

(3)推荐系统如大众点评, 如最大化推荐成功率与好评率;

(4)高频交易中金融资产的筛选, 如最大化投资收益;

(5)医疗系统采集信息为患者选择体检计划, 如最大化检测效率。

本文用Mathematica对MAB算法进行模拟仿真，测试多种不同的算法在普通MAB和CMAB问题中的实际表现。因此，当没有上下文信息时，第i个臂的预期回报是i，每次选择的后悔值为$5-i$; 有上下文信息时，后悔值就是预期回报最高的臂收益减去该臂对应的收益值。有两个上下文变量a和b，第i个臂的回报是一个随机变量，收益服从$N(i+ai+bi^2,1)$, 其中$a,b\sim N(0,1)$。

我们采用不同的MAB算法选择回报最高的臂，从而最小化后悔。在没有上下文的情形下，我们对比了(1) 随机选择 (2) 简单贪婪算法(3) ε-贪婪算法（4） UCB算法(5)汤普森采样这五种算法。有上下文时，我们对比了(1)随机选择 (2)LinUCB算法 (3)上下文汤普森采样法。测试结果如下图所示：

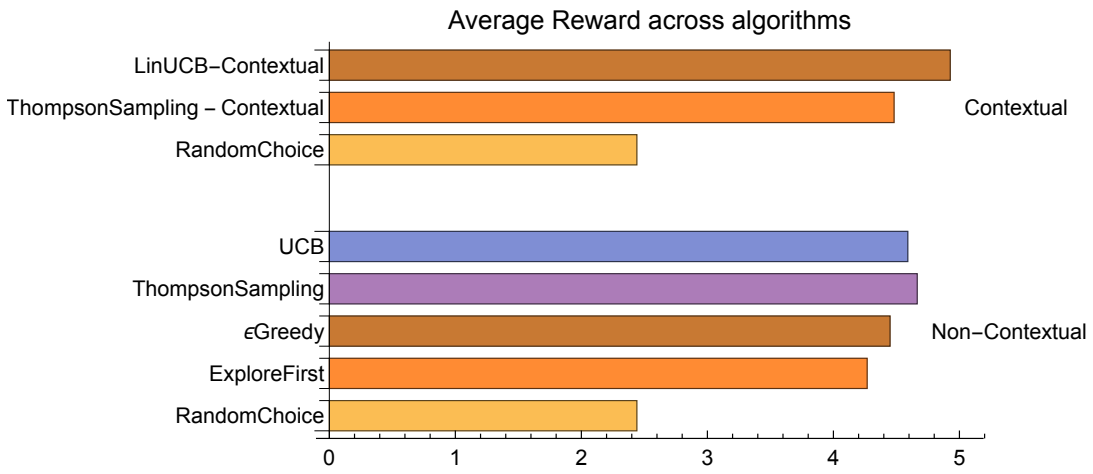


图 3: Mathematica 实验仿真结果 (T = 1000, K = 5)

图中可以看出，汤普森算法在两种情形下表现都很好，这充分说明其背后的贝叶斯概率逻辑稳健。而LinUCB虽然在无上下文时表现稍差于汤普森算法，但有上下文时，借助LinUCB的最大化行为，能够更频繁地提取回报最高的臂，从而达到比汤普森算法更好的结果。

这与雅虎公司Chapelle and Li (2011)下图[3]得出的结论相似。在实际问题中，每个用户都被表示为一个高维的原始特征向量(即背景信息)，该特征向量表示用户的年龄、性别、地理位置、行为等信息。为了让这些稀疏的信息更加精炼，雅虎采取主成分分析法提炼其中的核心特征(缩减至20维)，将其作为CMAB的参数。他们也指出UCB的最大化行为能让它的回报更大, 每阶段加入一定概率的探索能够有效处理兴趣漂移等情况。

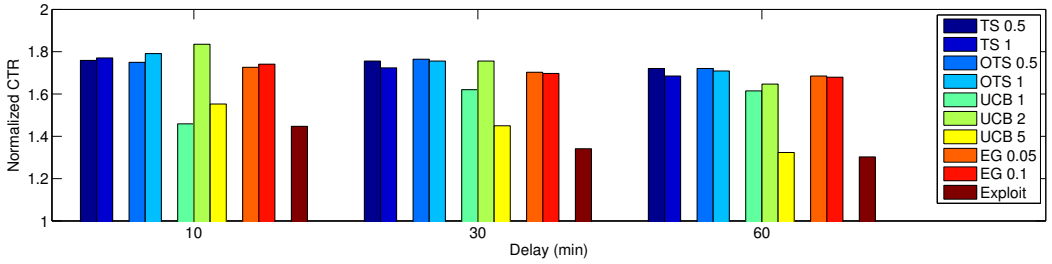


图 4: Chapelle and Li (2011)实验结果及对应方法参数设定

5 展望与结论

而在现实环境中，MAB问题也产生了很多变体。比如在不同臂的探索时转移有成本或探索次数有上限等约束(Simchi-Levi and Xu (2019))，抑或随着环境的推移臂的参数会发生变换，抑或是。而这些理论问题以动态定价为例，便对应价格变换之间产生的成本和库存约束，以及顾客行为的兴趣漂移等，也在学界和业界近年来得到了广泛应用。下图(Bouneffouf and Rish (2019))便描述了其目前被应用到的场景:

[3]Delay表示受到框架和计算成本等限制，奖励反馈给算法的延迟时间

	MAB	Non-stationary MAB	CMAB	Non-stationary CMAB
Healthcare	√		√	
Finance	√			
Dynamic pricing		√		
Recommendr system	√	√	√	√
Maximization	√			
Dialogue system			√	
Telecomunication	√			
Anomaly detection	√			

图 5: MAB的应用领域

从上图可以看出, MAB及CMAB算法在（1）数据获取次数少（2）不同臂选择的机会成本高、潜在利润大（3）短保质期、快节奏的臂选择（4）数据需运行时、实时、连续优化时可以展现出较大的优化能力，作为一种经典、稳定的强化学习算法，在电商、媒体、搜索、销售、医疗等行业中都有巨大的应用空间。

从理论上看, 现阶段MAB问题吸引了计算机科学、运筹学、经济学等多领域的学者。在学科交叉中，MAB研究已经在传统MAB问题上作出了很大的扩充：

一是考虑无穷多个臂的老虎机问题(Slivkins (2019))，这对应连续的老虎机问题(continuum-armed bandit)。动态定价便是此例。经典的框架分析便是考虑Lip连续并使用离散化转化为离散的问题。而在此基础上对臂的选择范围和约束条件也有进一步的研究，如`BwK`, `BwSC`(Simchi-Levi and Xu (2019))等, 从而推广到更为广阔的在线学习问题, 都是动态定价实际问题约束下的MAB模型抽象。

二, 对于传统的MAB和CMAB问题, 在联系实际落地地应用场景形成了较多有趣的思路。在**深度强化学习**的框架下，比如有运用主成分分析和神经网络的优势最大化利用上下文信息(Collier and Llorens (2018));在**传统推荐算法**和**社会网络**的交流中，对MAB的信息反馈也从全局臂选择向具有相关性的臂发展，如引入不同臂之间条件期望(Gupta et al. (2019))或考虑网络中其他用户信息等(Wu et al. (2016)), 从而更好地将协同过滤框架引入进去;在与**博弈论**的融合中, 用户只能进行两两臂之间的选择，如决斗老虎机(dueling bandit, Sui et al. (2018))等。

展望未来，本文认为多臂老虎机算法还有更大的应用空间——不仅是在算法（后悔上界、权重更新模式、选择策略）上可以优化，而且对老虎机问题的性质约束定义也可进行改良（如对臂的定义从一维选择到多维选择，从而实现更深入的定制化内容）。若能如此优化，则该类强化学习算法在实际应用中必然会创造更大的价值。

参考文献

Agrawal, Shipra, Navin Goyal. 2013. Thompson sampling for contextual bandits with linear payoffs. *International Conference on Machine Learning*. 127–135.

Auer, Peter. 2002. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* **3**(Nov) 397–422.

Bouneffouf, Djallel, Irina Rish. 2019. A survey on practical applications of multi-armed and contextual bandits. *arXiv preprint arXiv:1904.10040* .

Chapelle, Olivier, Lihong Li. 2011. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*. 2249–2257.

Collier, Mark, Hector Urdiales Llorens. 2018. Deep contextual multi-armed bandits. *arXiv preprint arXiv:1807.09809* .

Ferreira, Kris Johnson, David Simchi-Levi, He Wang. 2018. Online network revenue management using thompson sampling. *Operations research* **66**(6) 1586–1602.

Gupta, Samarth, Shreyas Chaudhari, Gauri Joshi, Osman Yağan. 2019. Multi-armed bandits with correlated arms. *arXiv preprint arXiv:1911.03959* .

Kaufmann, Emilie, Olivier Cappé, Aurélien Garivier. 2012. On bayesian upper confidence bounds for bandit problems. *Artificial intelligence and statistics*. 592–600.

Li, Lihong, Wei Chu, John Langford, Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th international conference on World wide web*. 661–670.

Osband, Ian, Benjamin Van Roy. 2017. Why is posterior sampling better than optimism for reinforcement learning? *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2701–2710.

Ricci, Francesco, Lior Rokach, Bracha Shapira. 2011. Introduction to recommender systems handbook. *Recommender systems handbook*. Springer, 1–35.

Russo, Daniel, Benjamin Van Roy. 2014. Learning to optimize via posterior sampling. *Mathematics of Operations Research* **39**(4) 1221–1243.

Russo, Daniel, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen. 2017. A tutorial on thompson sampling. *arXiv preprint arXiv:1707.02038* .

Simchi-Levi, David, Yunzong Xu. 2019. Phase transitions and cyclic phenomena in bandits with switching constraints. *Advances in Neural Information Processing Systems*. 7521–7530.

Slivkins, Aleksandrs. 2019. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272* .

Sui, Yanan, Masrour Zoghi, Katja Hofmann, Yisong Yue. 2018. Advancements in dueling bandits. *IJCAI*. 5502–5510.

Sutton, Richard S, Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Thompson, William R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**(3/4) 285–294.

Wu, Qingyun, Huazheng Wang, Quanquan Gu, Hongning Wang. 2016. Contextual bandits in a collaborative environment. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 529–538.

6 附录

6.1 关于贪婪算法的利用阶段的后悔界

假设最优臂为a^*, 并且该算法选择其他臂$a \neq a^*$.则其平均收益一定高于最优臂即$\widehat{\mu(a)} > \widehat{\mu(a^*)}$，在满足Hoeffding不等式的条件下，我们有如下的不等式:

$$\mu(a) + \sqrt{\frac{2logT}{N}} \geq \widehat{\mu(a)} > \widehat{\mu(a^*)} \geq \widehat{\mu(a^*)} - \sqrt{\frac{2logT}{N}}$$

因此每阶段的后悔最大值为:

$$\mu(a^*) - \mu(a) \leq \sqrt{\frac{2logT}{N}} = O(\sqrt{\frac{logT}{N}})$$

.

6.2 UCB算法 K =2的后悔界分析

我们记t为删除较差臂的停时，也就是a和a'臂的置信区间仍然由重叠，这时候由Hoeffding不等式有：

$$\Delta := |\mu(a) - \mu(a')| \leq 2(\sqrt{\frac{2logT}{n_t(a)}} + \sqrt{\frac{2logT}{n_t(a')}})$$

带入每个$n_t(a) = n_t(a') = \frac{t}{2}$,有:

$$\Delta \leq 4\sqrt{\frac{2logT}{t/2}} = O(\sqrt{logT}t)$$

因此，直到阶段t累计的后悔值为:

$$R(t) \leq \Delta \times t \leq O(t\sqrt{\frac{logT}{t}}) = O(\sqrt{tlogT}).$$

之后，同样的考虑选错臂的情况(即Hoeffding不等式左端不成立)，类似地最终仍然有： $\mathbb{E}[R(t)] \leq O(\sqrt{tlogT})$.

6.3 Beta分布的图示性质

在2.3节中, 对每个臂未知的收益$\theta = (\theta_1, \theta_2, ..., \theta_k)$事先位置。我们首先假设$\theta$的先验分布为Beta分布, 参数分别为$\alpha = (\alpha_1, \alpha_2, ..., \alpha_k)$, $\beta = (\beta_1, \beta_2, ..., \beta_k)$。对应到每个臂来说，也即:

$$p(\theta_k) = \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)}\theta_k^{\alpha_k - 1}(1 - \theta_k)^{\beta_k - 1}$$

对于这个$Beta$分布，下图中的action 1, action 2, action 3则分别表示$Beta(601, 401), Beta(401, 601), Beta(2, 3)$等。从中可以看出随着$Beta(\alpha_k, \beta_k)$中参数的增大，其采样产生可能的值范围越来越窄，直到收敛至$\mathbb{E}_p(\theta_k) =$

$\frac{\alpha_k}{\alpha_k+\beta_k}$。而action 3由于仅采样几次，其分布仍然接近于$Beta(1,1) = U(0,1)$, 提供了继续探索的空间。

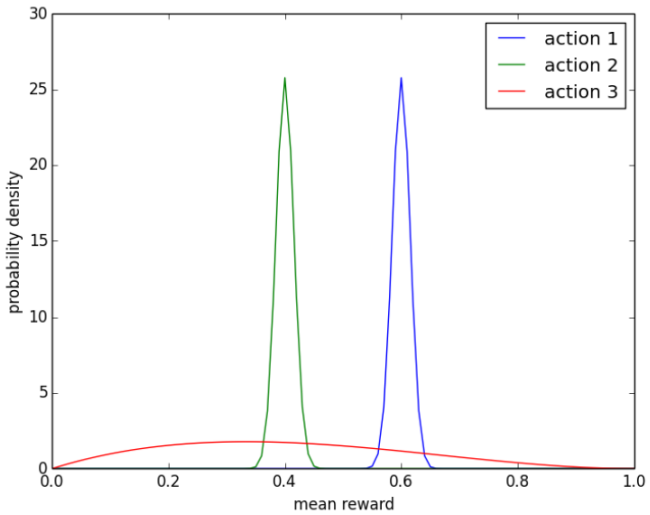


图 6: 不同参数下Beta分布示意图

6.4 其他文件说明

MABexample.nb和**MABexample.pdf**里附有Mathematica实验运行的代码, **graph.pdf**为不同算法设置下实验regret/reward的结果图。