随机多臂老虎机

Stochastic Multi-Armed Bandit Problem

傅奕诚

fuycc@zju.edu.cn

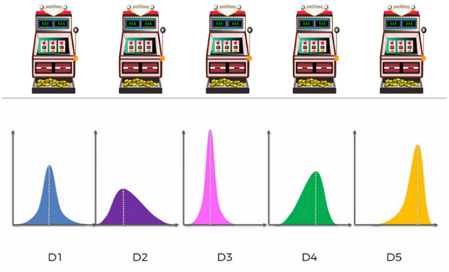
浙江大学计算机科学与技术学院

2025 年 4 月 18 日

目录

多臂老虎机问题

- 玩家希望最大化自己的奖励
- 玩家通过获得反馈调整策略
- 如何进行“探索-利用”权衡（exploration-exploitation balance）

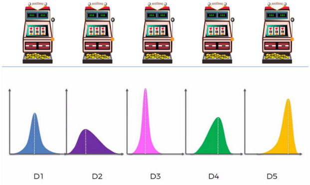


- **Scenario:** Pull machine k → sample from ***unknown*** reward distribution D_k → observe reward.
- **Problem:** Given a finite number of pulls T, how can I optimize my winnings?
- How much should I **explore**? How much should I **exploit**?

图: 多臂老虎机

多臂老虎机问题

- 经典场景：一个赌鬼要玩多臂老虎机，摆在他面前有 K 个臂（Arms）或动作选择（Actions），每一轮游戏中，他要选择拉动一个臂并会获得一个**奖励**．如果总共玩 T 轮，他该如何**最大化**奖励？



- **Scenario:** Pull machine k → sample from ***unknown*** reward distribution D_k → observe reward.
- **Problem:** Given a finite number of pulls T, how can I optimize my winnings?
- How much should I **explore**? How much should I **exploit**?

- 其它场景：新闻网站、动态定价、投资组合...

例子	动作	奖励
投资组合	选择一个股票买入	股票的涨跌
动态定价	选择一个价格交易	商品销售的收益
新闻网站	展示一则新闻	是否被用户点击

多臂老虎机问题

反馈（Feedback）

- 完全反馈（full feedback）：所有股票的涨跌
- 部分反馈（partial feedback）：任何更低（高）价格都被接受（拒绝）
- 老虎机反馈（bandit feedback）：该新闻是否被用户点击

奖励（Reward）

- 随机奖励/IID 奖励：玩家或者的奖励随机取自一个未知概率分布
- 对抗性奖励：奖励可以任意，能由一个“对手”有针对性的选择
- ...

随机多臂老虎机问题

过程:

- 在每轮 $t \in [T]$，玩家选择一个臂 $a_t \in \mathcal{A} = \{a_1, \cdots, a_K\}$
- 玩家获得该臂对应的随机奖励 $r_t \sim \mathcal{R}(a_t)$ $\left(r_t \in [0,1]\right)$
- 玩家依据过往轮次的奖励情况调整选择策略，实现奖励最大

说明:

- 奖励分布的均值记为 $\mu(a_k) = \mathbb{E}[\mathcal{R}(a_k)]$，$k \in [K]$
- 最优臂 a^* 的奖励均值 $\mu^* = \max_{a \in \mathcal{A}} \mu(a)$
- 奖励均值差异 $\Delta(a) = \mu^* - \mu(a)$

遗憾分析

我们需要设计 MAB 算法实现最大化奖励，实际上就是找最优臂．那么，分析 MAB 算法的性能就是在分析算法能否找到最优臂．我们用遗憾（regret）来度量实际选择和最优选择的差异．

遗憾（regret）或累计遗憾（cumulative regret）

- 伪遗憾（pseudo-regret）：

$$R(T) = \sum_{t=1}^{T}(\mu^* - \mu(a_t)) = \mu^* \cdot T - \sum_{t=1}^{T}\mu(a_t)$$

- 期望遗憾（expected regret）：$\mathbb{E}[R(T)]$

在 MAB 问题中，我们常常关注算法遗憾界（regret bound）．一个好的遗憾界是次线性的（sub-linear），即

$$\frac{regret\ bound}{T} \to 0, \quad T \to \infty$$

这意味着算法能够逐渐学到最优臂．

重要定理：Hoeffding 不等式

定理

(***Hoeffding* 不等式*-1***) 假设 $X_1, X_2, \cdots, X_n$ 是 $[0,1]$ 上的独立随机变量，样本均值为 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$，$\mu = \mathbb{E}[\bar{X}_n]$. 对于任意 $\varepsilon > 0$ 有：

$$P(|\mu - \bar{X}_n| \geq \varepsilon) \leqslant 2\exp(-2n\varepsilon^2)$$

该不等式是集中不等式（concentration inequalities）之一，直观上来说 $P(|\mu - \bar{X}_n| \leqslant small) \geqslant 1 - small$.

这里，$[\mu - \varepsilon, \mu + \varepsilon]$ 是置信区间 (confidence interval)，ε 是置信半径 (confidence radius). 若令 $\varepsilon = \sqrt{\frac{\alpha \log T}{n}}$，则有

$$P(|\mu - \bar{X}_n| \geq \varepsilon) \leqslant 2T^{-2\alpha}, \ \forall \alpha > 0$$

一般取 $\alpha = 2$.

重要定理：Hoeffding 不等式

定理

(***Hoeffding* 不等式-2**) 假设 $X_1, X_2, \cdots, X_n$ 是 $\{0,1\}$ 上的独立随机变量且 $\mathbb{E}[X_i] = p_i$，样本均值为 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$，$\mu = \frac{1}{n}\sum_{i}^{n} p_i$. 对于任意 $0 < \lambda < 1 - \mu$ 有：

$$P(\bar{X}_n \geqslant \mu + \lambda) \leqslant \exp(-nd(\mu + \lambda, \mu))$$

对于任意 $0 < \lambda < \mu$，

$$P(\bar{X}_n \leqslant \mu - \lambda) \leqslant \exp(-nd(\mu - \lambda, \mu))$$

注：$d(a, b) = a \ln \frac{a}{b} + (1 - a) \ln \frac{1-a}{1-b}$

定理

对于正数数 α, β，$F^{beta}_{\alpha,\beta}(y) = 1 - F^{B}_{(\alpha+\beta-1),y}(\alpha - 1)$.

目录

贪心算法

探索阶段：将每个臂各尝试 N 次
利用阶段：
for $t > KN$ **do**
  选择平均奖励最高的臂 $\hat{a} = \arg\max_a Q_t(a)$
  观察奖励 r_t，$N_{t+1}(\hat{a}) = N_t(\hat{a}) + 1$，$Q_{t+1}(\hat{a}) = Q_t(\hat{a}) + \frac{r_t - Q_t(\hat{a})}{N_{t+1}(\hat{a})}$
end for

贪心算法的遗憾分析

定理

贪心算法的遗憾界为 $O(T^{2/3}(K\log T)^{1/3})$.

证明：考虑 $K=2$ 的情况，遗憾产生当且仅当选择了次优臂 $a \neq a^*$.
显然，探索阶段的遗憾为

$$R(\mathit{exploration}) \leqslant N$$

对于利用阶段，我们分为两种情况考虑：

- 事件 E：所有臂 a 均满足 $P(|\mu(a)-Q(a)| \leqslant \sqrt{2\log T/N}) \geqslant 1-2T^{-4}$
- 事件 $\bar{E}$：事件 E 的补集

则有：

$$\begin{aligned}\mathbb{E}[R(\mathit{exploitation})] &\leqslant \mathbb{E}[R(\mathit{exploitation})|E] \times P(E) \\ &\quad + \mathbb{E}[R(\mathit{exploitation})|\bar{E}] \times P(\bar{E}) \\ &\leqslant \mathbb{E}[R(\mathit{exploitation})|E] + T \times O(1/T^4)\end{aligned}$$

贪心算法的遗憾分析（续）

定理

贪心算法的遗憾界为 $O(T^{2/3}(K\log T)^{1/3})$.

证明（续）：记 $\mathit{rad} = \sqrt{2\log T/N}$. 在事件 E 下产生遗憾时，有

$$\mu(a) + \mathit{rad} \geqslant Q(a) > Q(a^*) \geqslant \mu(a^*) - \mathit{rad}$$

整理得 $\mu(a^*) - \mu(a) \leqslant 2\mathit{rad}$. 那么

$$\begin{aligned}
\mathbb{E}[R(\mathit{exploitation})] &\leqslant \mathbb{E}[R(\mathit{exploitation})|E] \times P(E) \\
&\quad + \mathbb{E}[R(\mathit{exploitation})|\bar{E}] \times P(\bar{E}) \\
&\leqslant \mathbb{E}[R(\mathit{exploitation})|E] + T \times O(1/T^4) \\
&\leqslant (T - 2N) \cdot 2\mathit{rad} + O(1/T^3)
\end{aligned}$$

综合探索和利用的遗憾可得 $\mathbb{E}[R(T)] \leqslant N + 2\mathit{rad}T + O(1/T^3)$. 若令 $N = T^{2/3}(\log T)^{1/3}$，则有 $\mathbb{E}[R(T)] \leqslant O(T^{2/3}(\log T)^{1/3})$.

贪心算法的遗憾分析（续）

定理

贪心算法的遗憾界为 $O(T^{2/3}(K\log T)^{1/3})$.

证明：考虑 $K>2$ 的情况，探索阶段的遗憾为

$$R(\mathit{exploration}) \leqslant N(K-1)$$

利用阶段的遗憾为

$$\begin{aligned}\mathbb{E}[R(\mathit{exploitation})] &\leqslant \mathbb{E}[R(\mathit{exploitation})|E] \times P(E) \\ &\quad + \mathbb{E}[R(\mathit{exploitation})|\bar{E}] \times P(\bar{E}) \\ &\leqslant (T-NK)\cdot 2\mathit{rad} + O(1/T^3)\end{aligned}$$

综合探索和利用的遗憾可得 $\mathbb{E}[R(T)] \leqslant NK + 2\mathit{rad}T + O(1/T^3)$. 若令 $N=(T/K)^{2/3}\cdot O(\log T)^{1/3}$，则有 $\mathbb{E}[R(T)] \leqslant O(T^{2/3}(K\log T)^{1/3})$.

ϵ-贪心算法

- 前期探索过多导致没必要的遗憾
- 贪心算法固定了探索阶段，容易陷入次优解

1: **for** $t = 1, 2, \cdots, T$ **do**
2: 以 ϵ_t 的概率探索：随机选择一个臂
3: 以 $(1-\epsilon_t)$ 的概率利用：选择 $a_t = \arg\max_a Q_t(a)$
4: 观察奖励 r_t，$N_{t+1}(\hat{a}) = N_t(\hat{a}) + 1$，$Q_{t+1}(\hat{a}) = Q_t(\hat{a}) + \frac{r_t - Q_t(\hat{a})}{N_{t+1}(\hat{a})}$
5: **end for**

定理

令 $\epsilon_t = t^{-1/3}(K\log t)^{1/3}$，$\epsilon$-贪心算法的遗憾界为 $O(T^{2/3}(K\log T)^{1/3})$.

目录

上置信界算法

- ϵ-贪心算法的探索过于随机

对于每个候选臂 $k=1,\cdots,K$，令 $Q_0(a_k)=0$，$N_0(a_k)=0$
for $t=1,\cdots,T$ **do**
  if $t\leqslant K$ **then**
    初始化顺序选择每个臂
  else
    选择 $a_t=\arg\max_a(Q_t(a)+\sqrt{\frac{2\ln t}{N_t(a)}})$
  end if
  观察奖励 r_t，$N_{t+1}(a_t)=N_t(a_t)+1$，$Q_{t+1}(a_t)=Q_t(a_t)+\frac{r_t-Q_t(a_t)}{N_{t+1}(a_t)}$
end for

$(Q_t(a)+\sqrt{\frac{2\ln t}{N_t(a)}})$ 较大有两种情况：1. 奖励比较大；2. 不确定性较大

UCB 的遗憾分析

定理

UCB 算法的遗憾界为 $O(\sqrt{KT\log T})$.

证明：由于遗憾的产生是因为没有选择到最优臂而获得了次优奖励，因此遗憾又可以表示成：

$$\mathbb{E}[R(T)] = T\mu^* - \sum_{i=1}^{K} \mu_i \mathbb{E}[N_{T+1}(a)]$$

其中，$\mathbb{E}[N_{T+1}](a)]$ 代表到第 T 轮结束后，臂 a 被选择到的期望次数. 由于算法初始化会将每个臂都选择一遍，因此每个臂被选择次数为

$$N_{T+1}(a) = 1 + \sum_{t=K+1}^{T} \mathbb{I}\{a_t = a\}$$

其中 $\mathbb{I}\{\cdot\}$ 代表指示函数. 不难看出，如果我们要对算法进行遗憾分析，实际上是在分析算法选择次优臂的上界，即分析 $N_{T+1}(a)$ 的上界.

UCB 的遗憾分析（续）

证明：令 $c_{t,s} = \sqrt{(2\log t)/s}$，对于任何正整数 l，有

$$
\begin{aligned}
N_{T+1}(a) &\leqslant \sum_{t=K+1}^{T} \mathbb{I}\{a_t = a, N_t(a) < l\} + \sum_{t=K+1}^{T} \mathbb{I}\{a_t = a, N_t(a) \geqslant l\} \\
&\leqslant l + \\
&\sum_{t=K+1}^{T} \mathbb{I}\{Q_t(a^*) + c_{t,N_t(a^*)} \leqslant Q_t(a) + c_{t,N_t(a)}, N_t(a) \geqslant l\} \\
&\leqslant l + \sum_{t=K+1}^{T} \mathbb{I}\{\min_{0<s<t} Q_s(a^*) + c_{t,N_s(a^*)} \leqslant \max_{l \leqslant k < t} Q_k(a) + c_{t,N_k(a)}\} \\
&\leqslant l + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{k=l}^{t-1} \mathbb{I}\{Q_s(a^*) + c_{t,N_s(a^*)} \leqslant Q_k(a) + c_{t,N_k(a)}\}
\end{aligned}
$$

UCB 的遗憾分析（续）

证明：我们以同样的方式定义事件 $E(a,a^*)$，即臂 a 与 a^* 获得的平均奖励落在各自的置信区间中．在事件 $E(a,a^*)$ 下，若 $Q_s(a^*)+c_{t,N_s(a^*)}\leqslant Q_k(a)+c_{t,N_k(a)}$，则有：

$$\mu^*\leqslant Q_s(a^*)+c_{t,N_s(a^*)}\leqslant Q_k(a)+c_{t,N_k(a)}\leqslant \mu(a)+2c_{t,N_k(a)}$$

整理下得 $\mu^*-\mu(a)\leqslant 2c_{t,N_k(a)}$．可以验证的是，该式仅在 $l\leqslant\frac{8\log t}{\Delta(a)^2}$ 下成立，因此我们在后续分析中取 $l=\left\lceil\frac{8\log T}{\Delta(a)^2}\right\rceil$．
而当事件 $E(a,a^*)$ 没有发生时，则下述不等式至少有一个成立

$$Q_s(a^*)\leqslant\mu^*-c_{t,N_s(a^*)}$$
$$\mu(a)+c_{t,N_k(a)}\leqslant Q_k(a)$$

这两个不等式分别代表最优臂被低估的情况与次优臂被高估的情况．

UCB 的遗憾分析（续）

证明：因此，基于上述分析与 Hoeffding 不等式，我们可以得到：

$$N_t(a) \leqslant \left\lceil \frac{8\log T}{\Delta(a)^2} \right\rceil + \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{k=\left\lceil \frac{8\ln T}{\Delta(a)^2}\right\rceil}^{t-1} \{Q_s(a^*) + c_{t,N_s(a^*)} \leqslant Q_k(a) + c_{t,N_k(a)}\}$$

$$\leqslant \left\lceil \frac{8\log T}{\Delta(a)^2} \right\rceil$$

$$+ \sum_{t=1}^{\infty}\sum_{s=1}^{t-1}\sum_{k=\left\lceil \frac{8\ln T}{\Delta(a)^2}\right\rceil}^{t-1} \left(P(Q_s(a^*) \leqslant \mu^* - c_{t,N_s(a^*)}) + P(Q_k(a) \geq \mu(a) + c_{t,N_k(a)})\right)$$

$$\leqslant \left\lceil \frac{8\log T}{\Delta(a)^2} \right\rceil + \sum_{t=1}^{\infty}\sum_{s=1}^{t}\sum_{k=1}^{t} 2t^{-4} = \left\lceil \frac{8\log T}{\Delta(a)^2} \right\rceil + \sum_{t=1}^{\infty} 2t^{-2}$$

$$\leqslant \frac{8\log T}{\Delta(a)^2} + 1 + \frac{\pi^2}{3}$$

UCB 的遗憾分析（续）

证明： T 轮之后的遗憾为：

$$\mathbb{E}[R(T)] = T\mu^* - \sum_{i=1}^{K} \mu(a_i)\mathbb{E}[N_{T+1}(a_i)] = \sum_{a:\Delta(a)>0} \Delta(a)\mathbb{E}[N_{T+1}(a)]$$

$$\leqslant 8 \sum_{a:\Delta(a)>0} \frac{\log T}{\Delta(a)} + (1+\frac{\pi^2}{3})(\sum_{i=1}^{K}\Delta(a_i)) = O(\log T \sum_{a:\Delta(a)>0} \frac{1}{\Delta(a)})$$

该界称为问题依赖的（instance-dependent）遗憾界.
进一步讨论，若对于所有 $\Delta(a) \leqslant \epsilon$，则累计遗憾为 $O(\epsilon T)$. 反之，若 $\Delta(a) > \epsilon$，则产生的遗憾为 $O(\frac{K}{\epsilon}\log T)$. 综合这两种情况，我们取 $\epsilon = \sqrt{\frac{K}{T}\log T}$，则可以得到问题独立的（instance-independent）遗憾界 $O(\sqrt{KT\log T})$.

UCB 的遗憾分析 - 2

证明：回顾我们的分析目标

$$N_{T+1}(a) \leqslant \sum_{t=K+1}^{T} \mathbb{I}\{a_t = a, N_t(a) < l\} + \sum_{t=K+1}^{T} \mathbb{I}\{a_t = a, N_t(a) \geqslant l\}$$

基于之前的分析可以发现关键在于当事件 E 发生的情况下有 $l \leqslant \frac{8 \log t}{\Delta(a)^2}$，也就是说，次优臂 a 被选择的次数 $N_t(a) = O(\frac{\log T}{\Delta(a)^2})$，即 $\Delta(a) = \sqrt{\frac{\log T}{N_t(a)}}$. 另一方面，基于 Hoeffding 不等式，事件 E 不发生的概率随着轮次的增加趋近于 0. 因此，次优臂产生的遗憾为

$$R(t; a) = N_t(a) \cdot \Delta(a) = O(\sqrt{N_t(a) \log T})$$

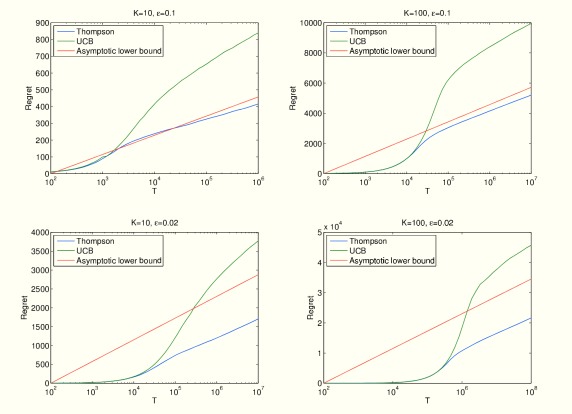
则累计遗憾为

$$R(t) = O(\sqrt{\log T}) \sum_{a \in \mathcal{A}} \sqrt{N_t(a)} \leqslant O(\sqrt{\log T}) \sqrt{K \sum_{a \in \mathcal{A}} N_t(a)} = O(\sqrt{Kt \log T})$$

目录

汤普森采样

- 汤普森采样（Thompson sampling，TS）最早于 1933 年由 William R. Thompson 提出
- 雅虎 An Empirical Evaluation of Thompson Sampling，NIPS 2011
- 2015 年前后才出现了一些理论分析的文章
- 简洁优雅



汤普森采样

```
对于每个候选臂 k = 1, ⋯ , K，令 S_t(a_k) = 0，F_t(a_k) = 0
for t = 1, ⋯ , T do
    for k = 1, ⋯ , K do
        从 Beta(S_t(a_k) + 1，F_t(a_k) + 1) 分布中采样 θ_t(a_k)
    end for
    选择 a_t = arg max_a θ_t(a)，并观察奖励 r_t
    if r_t = 1 then
        S_{t+1}(a_t) = S_t(a_t) + 1
    else
        F_{t+1}(a_t) = F_t(a_t) + 1
    end if
end for
```

- 只能用 beta 分布？共轭先验分布.
- 只能用于 $r_t \in \{0, 1\}$ 的情况？以 $r_t \in [0, 1]$ 为概率从 $\{0, 1\}$ 中抽取一个数作为反馈.

TS 的遗憾分析

定理

TS 算法的遗憾界为 $O(\sqrt{KT\ln T})$.

证明：延续 UCB 算法的证明思路，由于遗憾的产生是因为没有选择到最优臂而获得了次优奖励，因此遗憾为：

$$\mathbb{E}[R(T)] = T\mu^* - \sum_{i=1}^{K} \mu(a_k)\mathbb{E}[N_{T+1}(a_k)]$$

要分析这个遗憾，等价于分析任意次优臂被选中的次数 $\mathbb{E}[N_{T+1}(a_k)]$.

TS 的遗憾分析（续）

证明：

- $\hat{\mu}_t(a) = \frac{S_t(a)}{N_t(a)}$ 是奖励的经验期望（empirical mean），$N_t(a_k) = S_t(a_k) + F_t(a_k)$.
- 对于每个 $a \neq a^*$，令两个阈值参数 $x(a)$ 与 $y(a)$ 满足 $\mu(a) < x(a) < y(a) < \mu^*$. 对于在连续区间上，一定能找到这样的 $x(a)$ 与 $y(a)$ 存在.
- 记事件 $E_t^{\mu}(a)$ 为 $\hat{\mu}_t(a) \leqslant x(a)$，事件 $E_t^{\theta}(a)$ 为 $\theta_t(a) \leqslant y(a)$.
- 记 $p_t(a) = P(\theta_t(a^*) > y(a) | \mathcal{H}_{t-1})$，其中，$\mathcal{H}_{t-1} = \{a_\tau,\ r_\tau(a_\tau),\ \tau = 1, \cdots, t\}$ 代表时间步 t 之前的历史.

TS 的遗憾分析（续）

证明：回到要分析的目标 $\mathbb{E}[N_{T+1}(a)]$. 我们基于上面定义的两个事件对其进行分解：

$$
\begin{aligned}
\mathbb{E}[N_{T+1}(a)] &= \sum_{t=1}^{T} P(a_t = a) \\
&= \underbrace{\sum_{t=1}^{T} P(a_t = a, E_t^{\mu}(a), E_t^{\theta}(a))}_{(1)} + \underbrace{\sum_{t=1}^{T} P(a_t = a, E_t^{\mu}(a), \overline{E_t^{\theta}(a)})}_{(2)} \\
&\quad + \underbrace{\sum_{t=1}^{T} P(a_t = a, \overline{E_t^{\mu}(a)})}_{(3)}
\end{aligned}
$$

（1）代表估计与采样都较好的情况下选择了 a；（2）代表估计较好而采样不好的情况下选择了 a；（3）代表估计不好的情况下选择了 a.

TS 的遗憾分析 - (1) 式证明

证明：对于（1）式，

引理

对于所有时间步 t，$a \neq a^*$ 以及 $\mathcal{H}_{t-1}$ 的实现 H_{t-1}，有

$$P(a_t = a, E_t^{\mu}(a), E_t^{\theta}(a)|H_{t-1}) \leqslant \frac{(1-p_t(a))}{p_t(a)} P(a_t = a^*, E_t^{\mu}(a), E_t^{\theta}(a)|H_{t-1})$$

假设历史 H_{t-1} 使得事件 $E_t^{\mu}(a)$ 成立. 如果该事件不成立，则上面引理中的不等式两边都为 0，不等式恒成立. 事实上，对于事件 $E_t^{\theta}(a)$ 也能做同样的假设. 基于条件概率的性质，证明上述不等式等价于证明：

$$P(a_t = a|E_t^{\theta}(a), H_{t-1}) \leqslant \frac{(1-p_t(a))}{p_t(a)} P(a_t = a^*|E_t^{\theta}(a), H_{t-1})$$

TS 的遗憾分析 - (1) 式证明 - 引理证明

我们从不等式左侧开始证明.

$$
\begin{aligned}
&P(a_t = a|E_t^\theta(a), H_{t-1}) = P(\theta_t(a') \leqslant \theta_t(a), \forall a'|E_t^\theta(a), H_{t-1}) \\
&\leqslant P(\theta_t(a') \leqslant y(a), \forall a'|E_t^\theta(a), H_{t-1}) \\
&= P(\theta_t(a^*) \leqslant y(a)|H_{t-1}) \cdot P(\theta_t(a') \leqslant y(a), \forall a' \neq a^*|E_t^\theta(a), H_{t-1}) \\
&= (1 - p_t(a)) \cdot P(\theta_t(a') \leqslant y(a), \forall a' \neq a^*|E_t^\theta(a), H_{t-1}) \\
&= \frac{(1 - p_t(a))}{p_t(a)} \cdot P(\theta_t(a^*) > y(a)|H_{t-1}) \cdot P(\theta_t(a') \leqslant y(a), \forall a' \neq a^*|E_t^\theta(a), H_{t-1}) \\
&= \frac{(1 - p_t(a))}{p_t(a)} \cdot P(\theta_t(a') \leqslant y(a) < \theta_t(a^*), \forall a' \neq a^*|E_t^\theta(a), H_{t-1}) \\
&\leqslant \frac{(1 - p_t(a))}{p_t(a)} \cdot P(a_t = a^*|E_t^\theta(a), H_{t-1})
\end{aligned}
$$

TS 的遗憾分析 - (1) 式证明

证明：对于（1）式，利用上述引理与条件期望的性质，可以得到：

$$
\begin{aligned}
(1) &= \sum_{t=1}^{T} \mathbb{E}[P(a_t = a, E_t^{\mu}(a), E_t^{\theta}(a)|\mathcal{H}_{t-1})] \\
&\leqslant \sum_{t=1}^{T} \mathbb{E}[\frac{(1-p_t(a))}{p_t(a)} P(a_t = a, E_t^{\mu}(a), E_t^{\theta}(a)|\mathcal{H}_{t-1})] \\
&= \sum_{t=1}^{T} \mathbb{E}[\mathbb{E}[\frac{(1-p_t(a))}{p_t(a)} \mathbb{I}(a_t = a, E_t^{\mu}(a), E_t^{\theta}(a))|\mathcal{H}_{t-1}]] \\
&= \sum_{t=1}^{T} \mathbb{E}[\frac{(1-p_t(a))}{p_t(a)} \mathbb{I}(a_t = a, E_t^{\mu}(a), E_t^{\theta}(a))]
\end{aligned}
$$

TS 的遗憾分析 - (1) 式证明

证明：令 τ_k 表示最优臂 a^* 被第 k 次选到的时间步．可以观察到的是概率 $p_t(a)$ 仅在 $\theta_t(a^*)$ 的分布变化时才变化，即最优臂被选择时．因此，$p_t(a)$ 对于 $t \in \{\tau_k + 1, \cdots, \tau_{k+1}\}$（对于任意 k）不变．因此：

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}[\frac{(1-p_t(a))}{p_t(a)} \mathbb{I}(a_t = a, E_t^{\mu}(a), E_t^{\theta}(a))] \\
&= \sum_{t=1}^{T} \mathbb{E}[\frac{(1-p_{\tau_k+1}(a))}{p_{\tau_k+1}(a)} \sum_{t=\tau_k+1}^{\tau_{k+1}} \mathbb{I}(a_t = a, E_t^{\mu}(a), E_t^{\theta}(a))] \\
&\leqslant \sum_{t=0}^{T-1} \mathbb{E}[\frac{(1-p_{\tau_k+1}(a))}{p_{\tau_k+1}(a)}]
\end{aligned}
$$

这个不等式说明了，次优臂的选择次数能够被选择最优臂概率的方程所束缚住．

TS 的遗憾分析 - (1) 式证明

证明：下面的引理进一步证明了 $\mathbb{E}[\frac{1}{p_{\tau_k+1}(a)}-1]$ 的上界.

引理

令 τ_k 表示最优臂 a^* 被第 k 次选到的时间步，则对于 $i \neq 1$，有

$$\mathbb{E}[\frac{1}{p_{\tau_k+1}(a)}-1] \leqslant \begin{cases} \frac{3}{\Delta'(a)} & for\ k < \frac{8}{\Delta'(a)} \\ \Theta(e^{-\Delta'(a)^2 k/2} + \frac{e^{-D(a)k}}{(k+1)\Delta'(a)^2} + \frac{1}{e^{\Delta'(a)^2 k/4}-1}) & for\ k \geqslant \frac{8}{\Delta'(a)} \end{cases}$$

其中 $\Delta'_a = \mu^* - y(a)$，$D(a) = y(a)\ln\frac{y(a)}{\mu^*} + (1-y(a))\ln\frac{1-y(a)}{1-\mu^*}$.

引理证明略，请参考原文 [2].

TS 的遗憾分析 - (1) 式证明

证明：因此，基于上述引理，我们可以得到式（1）的上界：

$$(1) \leqslant \frac{24}{\Delta'(a)^2} + \sum_{j \geqslant 8/\Delta'(a)} \Theta\left(e^{-\Delta'(a)^2 j/2} + \frac{1}{(j+1)\Delta'(a)^2} e^{-D_a j} + \frac{1}{e^{\Delta'(a)^2 j/4} - 1}\right)$$

接下来我们证明式（2）的上界和式（3）的上界．这两个引理遵循的事实是，随着次优臂 i 被选择次数增加，违背事件 $E_t^{\mu}(a)$ 与 $E_t^{\theta}(a)$ 的概率会指数衰减．这两个引理的证明都需要用到 Hoeffding 不等式．

TS 的遗憾分析 - (2) 式证明

证明：对于（2）式，

引理

对于 $a \neq a^*$，有

$$\sum_{t=1}^{T} P(a_t = a, E_t^{\mu}(a), \overline{E_t^{\theta}(a)}) \leqslant L_T(a) + 1$$

其中，$L_T(a) = \frac{\ln T}{d(x(a),\ y(a))}$.（注：$d(a,b) = a \ln \frac{a}{b} + (1-a) \ln \frac{1-a}{1-b}$）

引理证明：我们首先对式（2）进行分解：

$$\begin{aligned}(2) = &\sum_{t=1}^{T} P(a_t = a, N_t(a) \leqslant L_T(a), \overline{E_t^{\theta}(a)}, E_t^{\mu}(a)) \\ &+ \sum_{t=1}^{T} P(a_t = a, N_t(a) > L_T(a), \overline{E_t^{\theta}(a)}, E_t^{\mu}(a))\end{aligned}$$

TS 的遗憾分析 - (2) 式证明 - 引理证明

引理证明：对于分解式的第一部分，显然有：

$$
\begin{aligned}
&\sum_{t=1}^{T} P(a_t = a, N_t(a) \leqslant L_T(a), \overline{E_t^{\theta}(a)}, E_t^{\mu}(a)) \\
&\leqslant \mathbb{E}[\sum_{t=1}^{T} \mathbb{I}(a_t = a, N_t(a) \leqslant L_T(a))] \\
&\leqslant L_T(a)
\end{aligned}
$$

TS 的遗憾分析 - (2) 式证明 - 引理证明

引理证明：对于分解式的第二部分，我们回忆事件 $E_t^{\mu}(a)$ 代表 $\hat{\mu}_t(a) \leqslant x(a)$，$E_t^{\theta}(a)$ 代表 $\theta_t(t) \leqslant y(a)$，且 $N_t(a)$ 与 $E_t^{\mu}(a)$ 由 $\mathcal{H}_{t-1}$ 决定，则有

$$
\begin{aligned}
&\sum_{t=1}^{T} P(a_t = a, N_t(a) > L_T(a), \overline{E_t^{\theta}(a)}, E_t^{\mu}(a)) \\
&= \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}(a_t = a, N_t(a) > L_T(a), \overline{E_t^{\theta}(a)}, E_t^{\mu}(a))] \\
&= \mathbb{E}[\sum_{t=1}^{T} \mathbb{E}[\mathbb{I}(a_t = a, N_t(a) > L_T(a), \overline{E_t^{\theta}(a)}, E_t^{\mu}(a)) | \mathcal{H}_{t-1}]] \\
&= \mathbb{E}[\sum_{t=1}^{T} \mathbb{I}(N_t(a) > L_T(a), E_t^{\mu}(a)) \cdot P(a_t = a, \overline{E_t^{\theta}(a)} | \mathcal{H}_{t-1})] \\
&= \mathbb{E}[\sum_{t=1}^{T} \mathbb{I}(N_t(a) > L_T(a), \hat{\mu}_t(a) \leqslant x(a)) \cdot P(\theta_t(a) > y(a) | \mathcal{H}_{t-1})]
\end{aligned}
$$

TS 的遗憾分析 - (2) 式证明 - 引理证明

引理证明：由定义可知，$\theta_t(a)$ 是取自于分布 $Beta(\hat{\mu}_t(a)(N_t(a)+1)+1, (1-\hat{\mu}_t(a))(N_t(a)+1))$. 进一步，记 $\alpha^{'}(a) = x(a)(N_t(a)+1)+1$，$\beta^{'}(a) = (1-x(a))(N_t(a)+1)$. 给定 $\mathcal{H}_{t-1} = H_{t-1}$ 使得 $N_t(a) > L_T(a)$ 和 $\hat{\mu}_t(a) \leqslant x(a)$ 成立，即 $\mathbb{I}(N_t(a) > L_T(a), \hat{\mu}_t(a) \leqslant x(a)) = 1$（否则为 0）. 利用 Heoffding 不等式可以得到：

$$
\begin{aligned}
P(\theta_t(a) > y(a) | \mathcal{H}_{t-1} = H_{t-1}) &\leqslant 1 - F^{beta}_{\alpha_i^{'},\beta_i^{'}}(y(a)) \\
&= F^{B}_{N_t(a)+1, y(a)}(x(a)(N_t(a)+1)) \\
&\leqslant e^{-(N_t(a)+1)\,d(x(a),y(a))} \\
&\leqslant e^{-L_T(a)\,d(x(a),y(a))} \leqslant \frac{1}{T}
\end{aligned} \tag{1}
$$

因此，可证得分解式的第二部分的一个上界是 1.

TS 的遗憾分析 - (3) 式证明

证明（续）：对于（3）式，

引理

对于 $i \neq 1$，有

$$\sum_{t=1}^{T} P(a_t = a, \overline{E_t^{\mu}(a)}) \leqslant \frac{1}{d(x(a),\ \mu(a))} + 1$$

TS 的遗憾分析 - (3) 式证明 - 引理证明

引理证明：这里，令 τ_k 表示次优臂 a 被第 k 次选择到的时间步，显然可以获得下述第一行的不等式．然后，我们利用事实 $\hat{\mu}_{\tau_{k+1}}(a) = \frac{S_{\tau_{k+1}}(a)}{k+1} \leqslant \frac{S_{\tau_{k+1}}(a)}{k}$ 与 Hoeffding 不等式，我们可以得到第二行不等关系．

$$
\begin{aligned}
\sum_{t=1}^{T} P(a_t = a, \overline{E_t^{\mu}(a)}) &\leqslant \sum_{k=0}^{T-1} P(\overline{E_i^{\mu}(\tau_{k+1})}) = \sum_{k=0}^{T-1} P(\hat{\mu}_{\tau_{k+1}}(a) > x(a)) \\
&\leqslant \sum_{k=0}^{T-1} P(\frac{S_{\tau_{k+1}}(a)}{k} > x(a)) \\
&\leqslant 1 + \sum_{k=1}^{T-1} \exp(-kd(x(a), \mu(a))) \\
&\leqslant 1 + \frac{1}{d(x(a), \mu(a))}
\end{aligned}
$$

TS 的遗憾分析（合）

证明：最后，通过将上述三个上界组合在一起，我们就可以获得最终遗憾的上界．通过选择 $x(a)=\mu(a)+\frac{\Delta(a))}{3}$ 与 $y(a)=\mu_1-\frac{\Delta(a)}{3}$，有
$\Delta^{'}(a)^2=(\mu_1-y(a))^2=\frac{\Delta(a)^2}{9}$，
$d(x(a),\mu(a))\geqslant 2(x(a)-\mu(a))^2=\frac{2\Delta(a)^2}{9}$ 以及
$d(x(a),y(a))\geqslant 2(y(a)-x(a))^2\geqslant\frac{2\Delta(a)^2}{9}$.

$$
\begin{aligned}
\mathbb{E}[N_t(a)] &\leqslant \frac{24}{\Delta^{'}(a)^2}+\sum_{j\geqslant 8/\Delta^{'}(a)}^{T-1}\Theta(e^{-\Delta^{'}(a)^2 j/2}+\frac{1}{(j+1)\Delta^{'}(a)^2}e^{-D(a)j}+\frac{1}{e^{\Delta^{'}(a)^2 j}} \\
&\quad + L_i(T)+1+\frac{1}{d(x(a),\mu(a))}+1 \\
&\leqslant \sum_{j\geqslant 8/\Delta^{'}(a)}^{T-1}\Theta(e^{-\Delta^{'}(a)^2 j/2}+\frac{1}{(j+1)\Delta^{'}(a)^2}+\frac{1}{j\Delta^{'}(a)^2})+O(\frac{\ln T}{\Delta(a)^2}) \\
&= \Theta(\frac{1}{\Delta^{'}(a)^2}+\frac{\ln T}{\Delta^{'}(a)^2})+O(\frac{\ln T}{\Delta(a)^2})=O(\frac{\ln T}{\Delta(a)^2})
\end{aligned}
$$

TS 的遗憾分析（合）

证明：

$$\mathbb{E}[N_t(a)] \leqslant O(\frac{\ln T}{\Delta(a)^2})$$

因此，对于所有满足 $\Delta(a) \geqslant \sqrt{\frac{K \ln T}{T}}$ 的臂，遗憾为 $O(\sqrt{\frac{T \ln T}{K}})$；对于所有满足 $\Delta(a) \leqslant \sqrt{\frac{K \ln T}{T}}$ 的臂，遗憾为 $O(\sqrt{KT \ln T})$.

综上，我们可以得到上界 $O(\sqrt{KT \ln T})$.

总结

- 多臂老虎机问题
- 随机多臂老虎机
- 贪心算法 -> UCB -> TS
- 遗憾分析（分而治之）
  - “探索”遗憾与“利用”遗憾
  - “好事件”与“坏事件”
  - 放缩 +Hoeffding 用于 bound
- contextual bandits、bayesian bandits、adversarial bandits. . .

参考文献

- Finite-time Analysis of the Multiarmed Bandit Problem. Peter Auer, et al. Machine Learning. 2002.
- Near-Optimal Regret Bounds for Thompson Sampling. Shipra Agrawal, et al. Journal of the ACM. 2017.
- Introduction to Multi-Armed Bandits. Aleksandrs Slivkins.
- Bandit Algorithms. Tor Lattimore and Csaba Szepesvari.