Thompson Sampling
Александра Харь, 774

Многорукие бандиты, постановка задачи:
Пусть у нас есть автомат с N ручками, в каждый момент времени $t = 1, 2, 3, \dots$ игрок выбирает одну из этих ручек. Для каждой ручки есть какое-то свое распределение выигрыша (игроку неизвестное). Сразу после того, как игрок сыграл конкретную ручку, он получает выигрыш.

Пусть μ_i - матожидание выигрыша на ручке i; $\mu^* := \max_i \mu_i,\ \Delta_i := \mu^* - \mu_i$. Тогда определим суммарный regret:

$$\mathbb{E}[R(T)] = \mathbb{E}[\sum_{t=1}^{T}(\mu^* - \mu_{i(t)})] = \sum_i \Delta_i \cdot \mathbb{E}[k_i(T+1)] \tag{1}$$

Мы будем рассматривать Бернуллиевских бандитов: распределения выигрышей на ручек являются бернуллиевскими. Без ограничения общности, будем считать, что первая ручка имеет максимальный средний выигрыш ($\mu^* = \mu_1$)

Algorithm 1 (Thompson Sampling using Beta prior)

For each arm $i = 1, \dots, N$ set $S_i = 0, F_i = 0$
For each $t = 1, 2, \dots$ **do**
    For each arm $i = 1, \dots, N$ sample $\theta_i(t)$ from the $Beta(S_i + 1, F_i + 1)$ distribution
    Play arm $i(t) := argmax_i \theta_i(t)$ and observe reward r_t
    If $r_t = 1$, then $S_{i(t)} = S_{i(t)} + 1$,
        else $F_{i(t)} = F_{i(t)} + 1$

Теорема. Для N-рукого стохастического бандита, используя Thompson Sampling using Beta priors, имеет место следующая оценка:

$$\mathbb{E}[R(T)] \leqslant O(\sqrt{NTlnT}) \tag{2}$$

Для начала введем некоторые обозначения:
• $i(t)$ - ручка, которую играл игрок в момент времени t
• $k_i(t)$ - количество игр ручки i до момента времени $t-1$
• $S_i(t)$ - количество успешных игр ручки i до момента времени $t-1$
• эмпирическое среднее $\hat{\mu}_i(t) = \dfrac{\sum_{\tau=1:\ i(\tau)=i}^{t-1} r_i(\tau)}{k_i(t)+1}$
• величины x_i, y_i впоследствии будем выбирать таким образом, чтобы $\forall i \hookrightarrow \mu_i < x_i < y_i < \mu_1$
• обозначим $E_i^{\mu}(t)$ событие: $\hat{\mu}_i(t) \leqslant x_i,\ E_i^{\theta}(t)$ событие: $\theta_i(t) \leqslant y_i$
• $\mathcal{F}_{t-1} = \{i(w), r_{i(w)}(w),\ w = 1, \dots, t-1\}$ - история до момента времени $t-1$
• $p_{i,t} = P(\theta_1(t) > y_i | \mathcal{F}_{t-1})$

Лемма 1. Для любого $t \in [1, T]$ и $t \neq 1$:

$$P(i(t) = i, E_i^{\mu}(t), E_i^{\theta}(t) | \mathcal{F}_{t-1}) \leqslant \frac{1 - p_{i,t}}{p_{i,t}} P(i(t) = 1, E_i^{\mu}(t), E_i^{\theta}(t) | \mathcal{F}_{t-1}),\ p_{i,t} = P(\theta_1(t) > y_i | \mathcal{F}_{t-1}) \tag{3}$$

Доказательство: Положим, что $E_i^{\mu}(t)$ - верно (иначе левая часть равно 0 и неравенство выполнено). Тогда достаточно доказать, что

$$P(i(t) = i | E_i^{\theta}(t), \mathcal{F}_{t-1}) \leqslant \frac{1 - p_{i,t}}{p_{i,t}} P(i(t) = 1 | E_i^{\theta}(t), \mathcal{F}_{t-1}) \tag{4}$$

$E_i^{\theta}(t), i(t) = i$ только если $\theta_j(t) \leqslant y_i, \forall j$.

Значит, для любого $i \neq 1$:

$$P(i(t)=i|E_i^\theta(t),\mathcal{F}_{t-1}) \leqslant P(\theta_j(t)\leqslant y_i,\forall j|E_i^\theta(t),\mathcal{F}_{t-1}) = P(\theta_1(t)\leqslant y_i|\mathcal{F}_{t-1})P(\theta_j(t)\leqslant y_i,\forall j\neq 1|E_i^\theta(t),\mathcal{F}_{t-1}) = \\ = (1-p_{i,t})P(\theta_j(t)\leqslant y_i,\forall j\neq 1|E_i^\theta,\mathcal{F}_{t-1}) \tag{5}$$

$$P(i(t)=1|E_i^\theta(t),\mathcal{F}_{t-1}) \geqslant P(\theta_1(t)>y_i\geqslant\theta_j(t),\ \forall j\neq 1|E_i^\theta(t),\mathcal{F}_{t-1}) = \\ = P(\theta_1(t)>y_i|\mathcal{F}_{t-1})P(\theta_j(t)\leqslant y_i,\ \forall j\neq 1|E_i^\theta(t),\mathcal{F}_{t-1}) = p_{i,t}P(\theta_j(t)\leqslant y_i,\ \forall j\neq 1|E_i^\theta(t),\mathcal{F}_{t-1}) \tag{6}$$

$\Rightarrow\ P(i(t)=i|E_i^\theta(t),\mathcal{F}_{t-1}) \leqslant \dfrac{1-p_{i,t}}{p_{i,t}}P(i(t)=1|E_i^\theta(t),\mathcal{F}_{t-1}).\quad \square$

Доказательство Теоремы:

$$\mathbb{E}[k_i(T)] = \sum_{t=1}^{T}P(i(t)=i) = \sum_{t=1}^{T}P(i(t)=i,E_i^\mu(t),E_i^\theta(t)) + \sum_{t=1}^{T}P(i(t)=i,E_i^\mu(t),\overline{E_i^\theta(t)}) + \sum_{t=1}^{T}P(i(t)=i,\overline{E_i^\mu(t)}) \tag{7}$$

Пусть τ_k - это момент времени, когда $k-$ый раз была сыграна первая ручка, $\tau_0=0$. Заметим, что для любого i, для любого $k>k_i(T)$, верно $\tau_k>T$ (также $\tau_T\geqslant T$).

Используем Лемму 1:

$$\sum_{t=1}^{T}P(i(t)=i,E_i^\mu(t),E_i^\theta(t)) = \sum_{t=1}^{T}\mathbb{E}[P(i(t)=i,E_i^\mu(t),E_i^\theta(t)|\mathcal{F}_{t-1})] \leqslant \sum_{t=1}^{T}\mathbb{E}\big[\frac{1-p_{i,t}}{p_{i,t}}P(i(t)=1,E_i^\mu(t),E_i^\theta(t)|\mathcal{F}_{t-1})\big] = \\ = \sum_{t=1}^{T}\mathbb{E}\big[\mathbb{E}\big[\frac{1-p_{i,t}}{p_{i,t}}I(i(t)=1,E_i^\mu(t),E_i^\theta(t)|\mathcal{F}_{t-1})\big]\big] = \sum_{t=1}^{T}\mathbb{E}\big[\frac{1-p_{i,t}}{p_{i,t}}I(i(t)=1,E_i^\mu(t),E_i^\theta(t))\big] \leqslant \\ \leqslant \sum_{k=0}^{T-1}\mathbb{E}\big[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}}\sum_{t=\tau_k+1}^{\tau_k+1}I(i(t)=1)\big] = \sum_{k=0}^{T-1}\mathbb{E}\big[\frac{1}{p_{i,\tau_k+1}}-1\big] \tag{8}$$

Лемма 2.

$$\mathbb{E}\big[\frac{1}{p_{i,\tau_k+1}}\big] \leqslant \begin{cases} 1+\dfrac{3}{\Delta_i'}, & for\ k<\dfrac{8}{\Delta_i'} \\ 1+\Theta(e^{-\Delta_i'^2k/2}+\dfrac{1}{(k+1)\Delta_i'^2}e^{-D_ik}+\dfrac{1}{e^{\Delta_i'^2k/4}-1}), & for\ k\geqslant\dfrac{8}{\Delta_i'}, \end{cases} \tag{9}$$

где $\Delta_i'=\mu_1-y_i,\ D_i=y_iln\dfrac{y_i}{\mu_1}+(1-y_i)ln\dfrac{1-y_i}{1-\mu_1}$

Лемма 3.

$$\sum_{t=1}^{T}P(i(t)=i,\overline{E_i^\mu(t)}) \leqslant \frac{1}{d(x_i,\mu_i)}+1 \tag{10}$$

, где $d(x_i,\mu_i)=x_ilog\dfrac{x_i}{\mu_i}+(1-x_i)log\dfrac{1-x_i}{1-\mu_i}$

Лемма 4.

$$\sum_{t=1}^{T}P(i(t)=i,\overline{E_i^\theta(t)},E_i^\mu(t)) \leqslant L_i(T)+1, \tag{11}$$

где $L_i(T)=\dfrac{lnT}{d(x_i,y_i)}$

Пусть

$$x_i=\mu_i+\frac{\Delta_i}{3},\ y_i=\mu_1-\frac{\Delta_i}{3} \tag{12}$$

$\Delta_i'^2 = (\mu_1 - y_i)^2 = \dfrac{\Delta_i^2}{9}$

$d(x_i, \mu_i) \geqslant 2(x_i - \mu_i)^2 = \dfrac{2\Delta_i^2}{9},\ d(x_i, y_i) \geqslant 2(y_i - x_i)^2 \geqslant \dfrac{2\Delta_i^2}{9}$. Тогда:

$L_i(T) = \dfrac{lnT}{d(x_i, y_i)} \leqslant \dfrac{9lnT}{2\Delta_i^2}$ и $\dfrac{1}{d(x_i, \mu_i)} \leqslant \dfrac{9}{2\Delta_i^2}$

$$\begin{aligned} \mathbb{E}[k_i(T)] &\leqslant \frac{24}{\Delta_i'^2} + \sum_{j=0}^{T-1} \Theta\big(e^{-\Delta_i'^2 j/2} + \frac{1}{(j+1)\Delta_i'^2} e^{-D_i j} + \frac{1}{e^{\Delta_i'^2 j/4} - 1}\big) + L_i(T) + 1 + \frac{1}{d(x_i, \mu_i)} + 1 \leqslant \\ &\leqslant \sum_{j=0}^{T-1} \Theta\big(e^{\Delta_i'^2 j/2} + \frac{1}{(j+1)\Delta_i'^2} + \frac{4}{\Delta_i'^2 j}\big) + O\big(\frac{lnT}{\Delta_i^2}\big) = \Theta\big(\frac{1}{\Delta_i'^2} + \frac{lnT}{\Delta_i'^2}\big) + O\big(\frac{lnT}{\Delta_i^2}\big) = O\big(\frac{lnT}{\Delta_i^2}\big) \end{aligned} \tag{13}$$

Для каждой ручки i с $\Delta_i \geqslant \sqrt{\dfrac{NlnT}{T}}$ выполнено $\Delta_i \mathbb{E}[k_i(T)] = O\big(\sqrt{\dfrac{TlnT}{N}}\big)$. Для ручек с $\Delta_i \leqslant \sqrt{\dfrac{NlnT}{T}}$ выполнено $\Delta_i \mathbb{E}[k_i(T)] = O\big(\sqrt{NTlnT}\big) \Rightarrow \mathbb{E}[R(T)] = O(\sqrt{NTlnT})$

Further Optimal Regret Bounds for Thompson Sampling, Shipra Agrawal, Navin Goyal