Stochastic Finite-Time Multi-Armed Bandits

- High-level Motivation - Till now we assumed all distributions were known while optimizing for revenue.
  - How can we simultaneously learn these distributions while optimizing rewards?
  - Applications - dynamic pricing with joint market-response ~~forecasting~~ forecasting
    - A/B testing and randomized trials
    - Assortment optimization
    - etc.

Bandit Settings

- Sequential decision making with incomplete information and learning
- 'Exploration vs. Exploitation'
- Different approaches
  - 'Markovian' (discounted, Gittin's Index)
  - 'Stochastic' (finite-time, regret)
  - 'Adversarial' (finite-time, minmax)

* The finite-time stochastic MAB problem

Setup – K 'arms' (Possible set of actions), T unknown time horizon

– $X_{i,1}, X_{i,2}, \ldots X_{i,T} \equiv$ Payoff from arm i in T rounds

• $X_{i,t} \in [0,1]$, iid, $\mathbb{E}[X_{i,t}] = \mu_i$, (unknown)

– $\mu^* \triangleq \max_{i \in [K]} \mu_i$, $i^* \triangleq \arg\max_{i \in [K]} \mu_i$

– $I_t \in [K] \equiv$ Arm chosen in t^{th} round

$T_i(n) \equiv \sum_{t=1}^{n} \mathbb{1}_{\{I_t = i\}} \equiv$ # of times arm i is played in n rnds

– Regret - $R_T = \max_{i \in [K]} \left(\sum_{t=1}^{T} X_{i,t} \right) - \sum_{t=1}^{T} X_{I_t, t}$

Expected regret - $\mathbb{E}[R_T] = \mathbb{E}\left[\max_{i \in [K]} \left(\sum_{t=1}^{T} X_{i,t} \right) - \sum_{t=1}^{T} X_{I_t, t} \right]$

Pseudo regret $\boxed{\bar{R}_T = \max_{i \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} X_{i,t} - \sum_{t=1}^{T} X_{I_t, t} \right]}$

Note: $\bar{R}_T \leq \mathbb{E}[R_T]$

We focus on minimizing $\bar{R}_T$

– $\boxed{\bar{R}_T = T\mu^* - \sum_{i \in [K]} \mu_i \, \mathbb{E}_{\pi}[T_i(T)]}$ ← Policy π

- Why pseudo regret?
  - Even if we knew $\{\mu_i\}$, the expected regret is still $\Theta(\sqrt{T})$ because of randomness
  - Pseudo-regret however can be much smaller ($\Theta(\log T)$) in spite of not knowing $\{\mu_i\}$
  - More natural comparison - Given all information, we would play i^*

Key algorithmic ideas

- Optimism in the face of uncertainty
  - Given data, construct a 'prior' over possible 'states of the world'
  - Use this prior to pick actions
    - greedy over prior $\equiv$ UCB style strategies
    - sample from prior $\equiv$ Thompson Sampling
- Use knowledge of lower bounds to guide choices

Thm (Lai & Robbins '85). For any policy Π, $\bar{R}_T(\pi) = \Omega(\log T)$

- To get optimal regret, we first need some concentration results for sums of random variables

* <u>Lemma (Hoeffding)</u> - Given any rv X s.t $\mathbb{E}[X]=0$, $a \le X \le b$ a.s

then $\forall \lambda \in \mathbb{R}$, $\mathbb{E}\left[e^{\lambda X}\right] \le \exp\left(\frac{\lambda^2 (b-a)^2}{8}\right)$

Pf - $e^{\lambda x}$ is convex $\Rightarrow e^{\lambda x} \le \frac{b-x}{b-a} e^{\lambda a} + \frac{x-a}{b-a} e^{\lambda b} \quad \forall x \in [a,b]$ (Jensen's)

$$\Rightarrow \mathbb{E}\left[e^{\lambda X}\right] \le \frac{b e^{\lambda a} - a e^{\lambda b}}{b-a} = \exp\left[\lambda(b-a)\left(\frac{a}{b-a}\right) + \log\left(1 - \left(\frac{a}{b-a}\right)\left(e^{\lambda(b-a)} - 1\right)\right)\right]$$

$$= \exp\Big[\underbrace{-h\theta + \log\left(1-\theta+e^{h}\theta\right)}_{g(h)}\Big] \quad \text{where } h = \lambda(b-a),\ \theta = \frac{-a}{b-a}$$

- $g'(0) = -\theta + \frac{\theta e^{h}}{1-\theta+e^{h}\theta}\Big|_{h=0} = 0$, $g''(h) = \frac{\theta e^{h}(1-\theta)}{1-\theta+\theta e^{h}} \le \frac{1}{4}$

$$\Rightarrow [g(h)] = g(0) + h g'(0) + \frac{h^2}{2} g''(u) \quad \text{for some } u \in [0,h]$$
$$\le h^2/8$$

$$\Rightarrow \mathbb{E}\left[e^{\lambda X}\right] \le \exp\left(h^2/8\right) = \exp\left(\frac{\lambda^2}{8}(b-a)^2\right)$$

- <u>Hoeffding's Inequality</u> - Let $X_1, X_2, \ldots X_n$ be iid rv,

$X_i \in [a_i, b_i]$ a.s, $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

Then $$\mathbb{P}\left[\bar{X} - \mathbb{E}[\bar{X}] \geq t\right] \leq \exp\left(-\frac{2n^2 t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

$$\mathbb{P}\left[\bar{X} - \mathbb{E}[\bar{X}] \leq -t\right] \leq \exp\left(\frac{-2n^2 t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

<u>Pf</u> - $$\mathbb{P}\left[\bar{X} - \mathbb{E}[\bar{X}] \geq t\right] = \mathbb{P}\left[\exp\left(\lambda \sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right) \geq \exp(\lambda n t)\right]$$

$$\leq \left[\prod_{i=1}^n \mathbb{E}\left[e^{\lambda (X_i - \mathbb{E}[X_i])}\right]\right] e^{-\lambda n t}$$

$$\leq \left[\prod_{i=1}^n \exp\left(\frac{\lambda^2}{8}(b_i - a_i)^2\right)\right] e^{-\lambda n t}$$

$$\leq \min_{\lambda \geq 0} \left[\exp\left(\frac{\lambda^2}{8} \sum_{i=1}^n (b_i - a_i)^2 - \lambda n t\right)\right]$$

$$\leq \exp\left(\frac{-2n^2 t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \qquad \left(\text{Using } \lambda = \frac{4nt}{\sum (b_i - a_i)^2}\right)$$

Similarly for $\{\bar{X} - \mathbb{E}[\bar{X}] \leq -t\}$

The UCB 1 Algorithm (Auer, Cesa-Bianchi, Fischer '02)

- Algorithm
  - Play each arm once (in first k rounds)
  - At round $n+1 \in \{K+1, K+2, \ldots, T\}$
    - Define $n_i(n) = T_i(n) = \#$ of pulls of arm i

$$\bar{X}_i(n) = \frac{1}{n_i} \sum_{t=1}^{n} X_{i,t} \, \mathbb{1}\{I_t = i\} \equiv \text{Empirical mean of arm } i$$

$$\boxed{UCB_i(n) = \bar{X}_i(n) + \sqrt{\frac{2\log(n)}{n_i(n)}}}$$

    - Pull arm i with highest $UCB_i(n)$

Thm - For all $T > K$, the regret of UCB1 satisfies

$$\bar{R}_T \leq \sum_{i : \mu_i < \mu^*} \left(\frac{\log T}{(\mu^* - \mu_i)} + 2 \right)$$

Pf - We first need some definitions

$$\Delta_j = \mu^* - \mu_j \quad \forall \, j \notin \arg\max_i \{\mu_i\}$$

$$C_{t,s} = \sqrt{\frac{2\log t}{s}} \quad \left(\text{hence } UCB_i(n) = \bar{X}_i(n) + C_{n_i, n}\right)$$

- From the Hoeffding bound, we have
  - For i^*, $(\forall s \leq t)$
  $$\mathbb{P}\left[\bar{X}_{i^*}(s) \leq \mu^* - C_{t,s}\right] \leq \exp\left(\frac{-2s^2\left(\frac{2\log t}{s}\right)}{s}\right) = t^{-4}$$
  - For $i \neq i^*$, $\mathbb{P}\left[\bar{X}_i(s) \geq \mu_i + C_{t,s}\right] \leq t^{-4}$

- Via the union bound, we have $\forall t \geq 1$

1. $$\mathbb{P}\left[\exists s \in \{1,2,\dots,t\} \text{ s.t } \bar{X}_{i^*}(s) \leq \mu^* - C_{t,s}\right] \leq \sum_{s=1}^{t} \mathbb{P}\left[\bar{X}_{i^*}(s) \leq \mu^* - C_{t,s}\right] \leq \sum_{s=1}^{t} t^{-4} = t^{-3}$$

2. $$\mathbb{P}\left[\exists s \in \{1,2,\dots,t\} \text{ s.t } \bar{X}_i(s) \geq \mu_i + C_{t,s}\right] \leq t^{-3} \quad \forall i \neq i^*$$

- Consider $s_i \geq \dfrac{8\log T}{\Delta_i^2}$ $\qquad \underbrace{\mu_i \;\; C_{t,s_i} \;\; \| \;\; C_{t,s_i} \;\; \mu^*}_{\Delta_i}$

3. $\Rightarrow$ $\forall t \leq T$, $C_{t,s_i} + \mu_i \leq \mu^* - C_{t,s_i}$

- Now we will show that for arm $i \neq i^*$, after $s_i = \dfrac{8\log T}{\Delta_i^2}$ pulls, it does not get pulled again with high probability

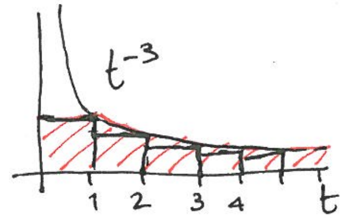
- Formally, we have - (with $s_i = 8\log T/\Delta_i^2$)

$$T_i(T) \le s_i + \sum_{t=K+1}^{T} \mathbb{1}\{I_t = i \mid n_i(t) \ge s_i\}$$

$$\le s_i + \sum_{t=K+1}^{T} \mathbb{1}\{X_i(t) > \mu_i + c_{t,s_i} , X_{i^*}(t) \le \mu^* + c_{t,s_i}\}$$

$$\Rightarrow \mathbb{E}[T_i(T)] \le s_i + \sum_{t=K+1}^{T} \mathbb{P}[X_i(t) \ge \mu_i + c_{t,s_i}] + \mathbb{P}[X_{i^*}(t) \le \mu^* + c_{t,s_i}]$$

$$\le s_i + \sum_{t=1}^{\infty} 2t^{-3}$$



$$\le \frac{8\log T}{\Delta_i^2} + \left(1 + \int_1^{\infty} 2t^{-3}dt\right) = \frac{8\log T}{\Delta_i^2} + 2$$

- Finally, for arm i, the regret incurred is $\Delta_i \mathbb{E}[T_i(T)]$

$$\Rightarrow \bar{R}_T = \sum_{i \ne i^*} \Delta_i \mathbb{E}[T_i(T)]$$

$$\le \sum_{i \ne i^*} \left(\frac{8\log T}{\Delta_i} + 2\Delta_i \right)$$