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
  - 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' (actions), T unknown horizon

- \( X_{1,1}, X_{1,2}, \ldots, X_{1,T} \) Payoff from arm i in T rounds
  - \( X_{i,t} \in [0,1] \text{ i.i.d, } E[X_{i,t}] = \mu_i \) (unknown)
  - \( \mu^* = \max_{i \in [K]} \mu_i \), i* = arg max_{i \in [K]} \mu_i

- \( I_t \in [K] \) = Arm chosen in \( t^{th} \) round

- \( T_i(n) = \sum_{t=1}^{n} \mathbb{I}\{I_t = i\} \) = # of times arm i is played in n rounds

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

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

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

Note: \( \bar{R}_T \leq E[R_T] \)

We focus on minimizing \( \bar{R}_T \)

- \( \bar{R}_T = T \mu^* - \sum_{i \in [K]} \mu_i E[T_i(1)] \)
• 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...

---

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 \( \pi_1 \), \( \overline{R}_T(\pi_1) \geq \Omega(\log T) \)
To get optimal logdet, we first need some concentration results for sums of random variables.

**Lemma (Hoeffding)** - Given any rv $X$ s.t. $E[x] = 0$, $a \leq x \leq b$ a.s.

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

**Proof** - $e^{\lambda x}$ is convex $\implies e^{\lambda x} \leq \frac{b-x}{b-a} e^{\lambda a} + \frac{x-a}{b-a} e^{\lambda b} \quad \forall x \in [a,b] \quad$ (Jensen's)

$$E[e^{\lambda x}] \leq \frac{b e^{\lambda a} - a e^{\lambda b}}{b-a} = \exp\left[\frac{\lambda}{b-a} \left(\frac{a}{b-a}\right) + \log\left(1 - \frac{a}{b-a}\right)\right]$$

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

where $h = \lambda (b-a)$, $\theta = -\frac{a}{b-a}$

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

$$\Rightarrow [g(h)]'' = \frac{1}{2} g(0) + h g'(0) + \frac{h^2}{2} g''(0) \leq \frac{h^2}{8}$$

$$\Rightarrow E[e^{\lambda x}] \leq \exp\left(\frac{h^2}{8}\right) = \exp\left(\frac{\lambda^2 (b-a)^2}{8}\right)$$
Haeflend's Inequality - Let $X_1, X_2, \ldots, X_n$ be iid so,

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

Then

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

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

**Proof**

$\Pr \left[ \overline{X} - \mathbb{E}[\overline{X}] \geq t \right] = \Pr \left[ \exp \left( \frac{\sum_{i=1}^{n} X_i - \mathbb{E}[X_i]}{\mathbb{E}[X_i]} \right) \geq \exp(2nt) \right]$

$\leq \left[ \prod_{i=1}^{n} \mathbb{E} \left[ e^{\frac{2}{\mathbb{E}[X_i]} (X_i - \mathbb{E}[X_i])} \right] \right] e^{-2nt}$

$\leq \left[ \prod_{i=1}^{n} \exp \left( \frac{\mathbb{E} \left[ (X_i - \mathbb{E}[X_i])^2 \right]}{\mathbb{E}[X_i]} \right) \right] e^{-2nt}$

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

$\leq \exp \left( -\frac{2n^2 t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right)$

(Using $\lambda = \frac{4nt}{\sum_{i=1}^{n} (b_i - a_i)^2}$)

Similarly for $\left\{ \overline{X} - \mathbb{E}[\overline{X}] \leq -t \right\}$
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 $N_i(n) = T_i(n) = \# \text{ of pulls of arm } i$
    - $\bar{X}_i(n) = \frac{1}{N_i(n)} \sum_{t=1}^{n} X_{i,t} \quad \text{if } \epsilon_t = i$ 
      \[ \text{Empirical mean of arm } i \]
    - $UCB_i(n) = \bar{X}_i(n) + \sqrt{\frac{2\log(n)}{N_i(n)}}$
  - Pull arm $i$ with highest $UCB_i(n)$

---

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

\[ R_T \leq \sum_{i : \mu_i < \mu^*} \left( \frac{\log T}{n_i} + \frac{1}{2} \right) \]

**Pf** - We first need some definitions

\[ \Delta_i = \mu^* - \mu_i + i \notin \arg\max \{ \mu_i \} \]

\[ C_{n,t,s} = \sqrt{\frac{2\log t}{s}} \quad (\text{hence } UCB_i(n) = \bar{X}_i(n) + C_{n_i,n}) \]
• From the Hoeffding bound, we have

- For $i^*$, $\Pr[\bar{X}_{i^*}(s) \leq \mu^* - C_{t,s}] \leq \exp\left(-\frac{2s^2(2\log t)}{s}\right) = t^{-4}$

- For $i \neq i^*$, $\Pr[\bar{X}_i(s) \geq \mu_i + C_{t,s}] \leq t^{-4}$

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

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

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

• Consider $S_i \geq \frac{8\log T}{\Delta_i^2}$

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

• Now we will show that for any $i \neq i^*$, after $S_i = \frac{8\log T}{\Delta_i^2}$ pulls, it does not get pulled again with high probability.
Formally, we have \( T_i(T) \leq S_i + \sum_{t = k+1}^{T} 1 \{ I_k = i, \mu_i(t) \geq S_i \} \) \( \equiv \sum_{t = k+1}^{T} 1 \{ X_i(t) \geq \mu_i + C_{t,s_i}, X_i(t) \leq \mu_i + C_{t,s_i} \} \)

\[ E[T_i(T)] \leq S_i + \sum_{t = k+1}^{T} P[X_i(t) \geq \mu_i + C_{t,s_i}] + P[X_i(t) \leq \mu_i + C_{t,s_i}] \]

\[ \leq S_i + \sum_{t = 1}^{\infty} 2t^{-3} \]

\[ \leq 8 \frac{\log T}{\Delta_i^2} + \left( 1 + \int_1^{\infty} 2t^{-3} \, dt \right) = \frac{8 \log T}{\Delta_i^2} + 2 \]

Finally, for arm \( i \), the regret incurred is \( \Delta_i E[T_i(t)] \)

\[ R_T = \sum_{i \neq i^*} \Delta_i E[T_i(T)] \leq \sum_{i \neq i^*} \left( \frac{8 \log T}{\Delta_i} + 2 \Delta_i \right) \]