Multi-Armed Bandit Problem

Key Problem

Exploration vs exploitation dilemma

- inspect new arms with possibly better rewards.
- use existing information to select best arm.

Stochastic Bandit

- K arms: for each arm $i \in \{1, ..., K\}$.
  - reward distribution P_i .
  - reward mean μ_i.
  - gap to best: $\Delta_i = \mu^* - \mu_i$, where $\mu^* = \max_{i \in [1,K]} \mu_i$.
- Bandit Setting: For $t = 1$ to T do
  - players selects action $\mathrm{I}_t \in \{1, \cdots, K\}$(*randomized*)
  - player receives reward $\mathrm{X}_t \sim \mathrm{P}_{\mathrm{I}_t}$

Objectives

1. Expected Regret

$$\mathrm{E}[\mathrm{R}_T] = \mathrm{E}\left[\max_{i\in[1,K]} \sum_{t=1}^{T} \mathrm{X}_{i,t} - \sum_{t=1}^{T} \mathrm{X}_{\mathrm{I}_t,t}\right]$$

2. Pseudo-regret

$$\begin{aligned}
\overline{\mathrm{R}_T} &= \max_{i\in[1,K]} \mathrm{E}\left[\sum_{t=1}^{T} \mathrm{X}_{i,t} - \sum_{t=1}^{T} \mathrm{X}_{\mathrm{I}_t,t}\right] \\
&= \max_{i\in[1,K]} \mathrm{E}\left[\sum_{t=1}^{T} \mathrm{X}_{i,t} - \mathrm{E}\left[\sum_{t=1}^{T} \mathrm{X}_{\mathrm{I}_t,t}\right]\right] \\
&= \max_{i\in[1,K]} \sum_{t=1}^{T} \mathrm{E}\left[\mathrm{X}_{i,t}\right] - \mathrm{E}\left[\sum_{t=1}^{T} \mathrm{X}_{\mathrm{I}_t,t}\right] \\
&= \mu^* T - \mathrm{E}\left[\sum_{t=1}^{T} \mathrm{X}_{\mathrm{I}_t,t}\right]
\end{aligned}$$

3. By Jensen's inequality, $\overline{\mathrm{R}_T} \leq \mathrm{E}[\mathrm{R}_T]$

Pseudo Regret

1. Expression in terms of Δ_is:

$$\overline{\mathrm{R}_T} = \sum_{t=1}^{K} \mathrm{E}[T_i(T)]\Delta_i$$

$T_i(T)$: number of times arm i was pulled up to time t,
$T_i(t) = \sum_{s=1}^{t} 1_{\mathrm{I}_s = i}$

Pseudo Regret

Proof.

$$\overline{\mathrm{R}_T} = \mu^* T - \mathrm{E}\left[\sum_{t=1}^{T} \mathrm{X}_{\mathrm{I}_t,t}\right] = \mathrm{E}\left[\sum_{t=1}^{T} (\mu^* - \mathrm{X}_{\mathrm{I}_t,t})\right]$$

$$= \mathrm{E}\left[\sum_{t=1}^{T} \sum_{i=1}^{K} (\mu^* - \mathrm{X}_{\mathrm{I}_t,t}) 1_{\mathrm{I}_t=i}\right] = \sum_{t=1}^{T} \sum_{i=1}^{K} \mathrm{E}[(\mu^* - \mathrm{X}_{\mathrm{I}_t,t})]\, \mathrm{E}[1_{\mathrm{I}_t=i}]$$

$$= \sum_{i=1}^{K} (\mu^* - \mu_i)\, \mathrm{E}[\sum_{t=1}^{T} 1_{\mathrm{I}_t=i}] = \sum_{i=1}^{K} \mathrm{E}[T_i(T)] \Delta_i$$

$\square$

ϵ-Greedy Strategy

Input: $\epsilon_t \in (0, 1]$

for $t \leftarrow 1$ to T do
    $\mathrm{I}_t = \mathrm{argmax}_{j=1,\cdots,K}\, \hat{\mu}_j$
    Draw u uniformly from [0,1]
    if $u > \epsilon_t$ then
        Play arm I_t
    else
        Play a random arm
    end if
end for

Thompson Sampling

For the Bernoulli bandit, X_t follows a Beta distribution, as X_t is essentially the success probability θ in Bernoulli distribution. The value of *Beta*(α, β) is within the interval [0, 1]; α and β correspond to the counts when we succeeded or failed to get a reward respectively.

$$\alpha_i = \alpha_i + \mathrm{X}_{\mathrm{I}_t,t} 1_{\mathrm{I}_t=i}$$
$$\beta_i = \beta_i + (1 - \mathrm{X}_{\mathrm{I}_t,t}) 1_{\mathrm{I}_t=i}$$

UCB Strategy

Intuition: we are drawn to the bandits that are paying out large rewards and those that we know little about. We want to upper bound the number of times we will pull arm i, so we will attempt to compute $\mathrm{E}[T_i(T)]$.

- For each $t \in [1, T]$, compute an upper confidence bound estimate on the mean of each arm at some fixed confidence level.
- select arm with largest UCB.

UCB Strategy

- With the rewards distributions in [0,1]. from Hoeffding's inequality we have:

$$\log \mathrm{E}[e^{t(X-\mathrm{E}[X])}] \leq \Psi(t)$$

- Then

$$\begin{aligned} Pr[X - \mathrm{E}[X] > \epsilon] &= Pr[e^{t(X-\mathrm{E}[X])} > e^{t\epsilon}] \\ &\leq \inf_{t>0} e^{-t\epsilon}\, \mathrm{E}[e^{t(X-\mathrm{E}[X])}] \\ &\leq \inf_{t>0} e^{-t\epsilon} e^{\Psi(t)} \\ &= e^{-\sup_{t>0}(t\epsilon - \Psi(t))} \\ &= e^{-\Psi^*(\epsilon)} \end{aligned}$$

UCB Strategy

1. Average reward estimate for arm i by time t:

$$\widehat{\mu_{i,t}} = \frac{1}{\mathrm{T}_i(t)} \sum_{s=1}^{t} X_{i,s} 1_{\mathrm{I}_s = i}$$

2. Concentration inequality:

$$Pr[\mu_i - \frac{1}{t} \sum_{s=1}^{t} X_{i,s} > \epsilon] \leq e^{-\Psi^*(\epsilon)}$$

3. Thus, for any $\delta > 0$, with probability at least $1 - \delta$:

$$\mu_i < \frac{1}{t} \sum_{s=1}^{t} X_{i,s} + {\Psi^*}^{-1}\left(\frac{1}{t} \log \frac{1}{\delta}\right)$$

(α,Ψ)-UCB Strategy

- Parameter $\alpha > 0$; (α,Ψ)-UCB Strategy: at time t, select:

$$\mathrm{I}_t \in \operatorname{argmax}\left[\hat{\mu_{i,t-1}} + {\Psi^*}^{-1}\left(\frac{\alpha \log t}{\mathrm{T}_i(t-1)}\right)\right]$$

- For $\Psi(\lambda) = \frac{\lambda^2}{8}$, then ${\Psi^*}^{-1} = 2\epsilon^2$ and substituting for $\alpha = 4$, we obtain that at time t, plays the arm (UCB1):

$$\mathrm{I}_t \in \operatorname{argmax}\left[\hat{\mu_{i,t-1}} + \sqrt{\frac{2 \log t}{\mathrm{T}_i(t-1)}}\right]$$