EE6106 Lecture 8 (Date : $13^{th}$ FEB 2024)

## STOCHASTIC MAB

Instance : $\nu = (\nu_1, \ldots, \nu_i, \ldots \nu_K)$

$\nu_i$ : 1-subGaussian (Mean $\mu_i$)

$$\Delta_i = \mu^* - \mu_i, \qquad \mu^* = \left(\max_j \mu_j\right)$$

Regret: $R_n = n\mu^* - \mathbb{E}\left[\sum^n X_t\right]$

UCB Algorithm

First, recall :

For $\varepsilon \geq 0$, $\mathbb{P}(\hat{\mu}_{i,n} \leq \mu_i - \varepsilon) \leq e^{-n\varepsilon^2/2}$ ($\sigma = 1$ by our assumptions)

$\Rightarrow$ w.p. $\geq 1-\delta$,

$$\mu_i \leq \underbrace{\hat{\mu}_{i,n} + \sqrt{\frac{2\log(1/\delta)}{n}}}_{UCB}$$

# UCB Algorithm

$$UCB_i(t-1) = \begin{cases} \infty & \text{if } T_i(t-1) = 0 \\ \hat{\mu}_i(t-1) + \sqrt{\dfrac{2\log(1/\delta)}{T_i(t-1)}} \end{cases}$$

$\hat{\mu}_i(t-1) \rightarrow$ empirical estimate

$T_i(t)$ : number of times the arm has been pulled upto time $t$

Algo :

At time $t \geq 1$,

- $A_t = \underset{i}{\operatorname{argmax}} \; UCB_i(t-1)$
- Observe rewards, update UCBs

Idea : UCB is an optimistic estimate of the mean of arm $i$.
We are pulling the arm with the highest optimistic estimate
"Optimism in the face of Uncertainty"

$T_i(t-1)$ is a random variable and not deterministic, unlike the $n$ in the earlier setting.

$\therefore$ $T_i(t-1)$ is a random quantity dependent on the past

$\therefore$ we cannot blindly say that this UCB is a perfectly bound

Intuition : In the line $\mu_i \leq \hat{\mu}_{i,n} + \sqrt{\dfrac{2\log(1/\delta)}{n}}$

if the $n$ was dependent on $\mu_i$, then this bound will not be valid.

Theorem : With $\delta = \frac{1}{n^2}$, UCB has regret

$$R_n \leq 3\sum \Delta_i + \sum_{i:\Delta_i > 0} \frac{16\log(n)}{\Delta_i}$$

Note : The algorithm is not "ANYTIME", i.e. it requires knowledge of the horizon 'n'

Note : The regret is logarithmic wrt to the horizon 'n'

Note : The bound is instance dependent bound (i.e. depends on $\Delta_i$)

Recall : $R_n = \sum \Delta_i \mathbb{E}[T_i(n)]$

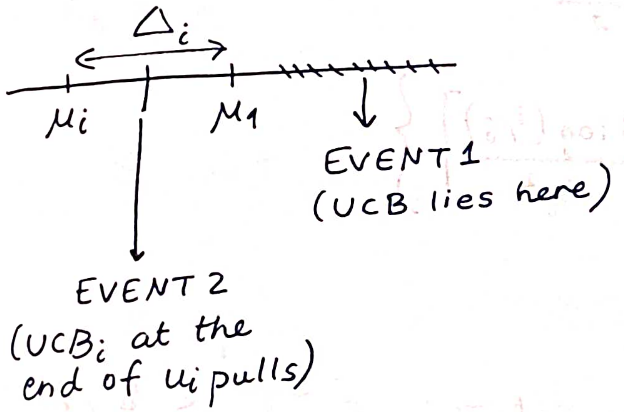
Proof : Fix suboptimal arm $i$. WLOG, arm 1 is optimal arm.

$$G_i = \Big\{ \mu_1 < \min_{t \in [n]} UCB_1(t) \Big\} \cap \Big\{ \hat{\mu}_{i,u_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} < \mu_1 \Big\}$$

$G_i$ ↓ good event associated with arm $i$

$\min_{t\in[n]} UCB_1(t)$ ↓ the UCB of arm 1 is never "violated"

$\sqrt{\frac{2\log(1/\delta)}{u_i}}$ ↓ to be specified later (think of it like a constant)




$$\mathbb{E}[T_i(n)] = \mathbb{P}(G_i)\,\mathbb{E}[T_i(n) \mid G_i] + \mathbb{P}(G_i^c)\,\mathbb{E}[T_i(n) \mid G_i^c]$$

$$\leq 1\, \mathbb{E}[T_i(n) \mid G_i] + \mathbb{P}(G_i^c)\, n \quad (\#)$$

Claim : Under $G_i$, $T_i(n) \leq u_i$

Proof : Suppose $G_i$ occurs and $T_i(n) > u_i$

$\Rightarrow \exists\, t$ s.t. $T_i(t) = u_i$, $A_i(t+1) = i$

Now, $\hat{\mu}_{i,u_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} < \mu_1 < UCB_1(t)$ ... under $G_i$

$\Longrightarrow$ Arm $i$ not pulled

(#)

$$\mathbb{E}[T_i(n)] \leq u_i + n\,\mathbb{P}(G_i^c)$$

Bounding $\mathbb{P}(G_i^c)$

$$G_i^c = \underbrace{\Big\{ \mu_1 \geq \min_{t} UCB_1(t) \Big\}}_{J} \cup \underbrace{\Big\{ \hat{\mu}_{i,u_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} > \mu_1 \Big\}}_{K}$$

... (DeMorgan Laws)

$\mathbb{P}(J)$ :

$$J \subseteq \left\{ \mu_1 \geq \min_{s \in [n]} \hat{\mu}_{1,s} + \sqrt{\frac{2\log(1/\delta)}{s}} \right\}$$

$$= \bigcup_{s \in [n]} \left\{ \mu_1 \geq \hat{\mu}_{1,s} + \sqrt{\frac{2\log(1/\delta)}{s}} \right\}$$

$\Rightarrow \quad P(J) \leq n\delta \qquad$ (union bound, further $\mathbb{P}$ (sub-event) $\leq \delta$)

$\mathbb{P}(K)$ :

$$\mathbb{P}(K) = \mathbb{P}\left( \hat{\mu}_{i,u_i} - \mu_i \geq \Delta_i - \sqrt{\frac{2\log(1/\delta)}{u_i}} \right)$$

Set $u_i$ s.t.

$$\Delta_i - \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq \frac{\Delta_i}{2}$$

$$\left\{ u_i = \left\lceil \frac{8\log(1/\delta)}{\Delta_i^2} \right\rceil \right\}$$

$$\mathbb{P}(K) \leq \mathbb{P}\left( \hat{\mu}_{i,u_i} - \mu_i \geq \frac{\Delta_i}{2} \right)$$

$$\leq e^{-\frac{u_i \Delta_i^2}{8}} \leq n^{-2}$$

(Note: $e^{-\frac{u_i \Delta_i^2}{8}} \leq \delta$, here $\delta = \frac{1}{n^2}$)

$$\therefore \ \mathbb{E}[T_i(n)] \leq \left\lceil \frac{8\log(1/\delta)}{\Delta_i^2} \right\rceil + n\left(n\delta + \frac{1}{n^2}\right)$$

$$\boxed{\mathbb{E}[T_i(n)] \leq 3 + \frac{16\log n}{\Delta_i^2}}$$

$$\Rightarrow \quad R_n \leq 3\sum \Delta_i + \sum_{i:\Delta_i > 0} \frac{16\log(n)}{\Delta_i}$$

... (Regret Decomposition Function)

Q : Can we achieve the Lai & Robbins' Bound by choosing $u_i$ even more smartly?

Q : Is $\delta = \frac{1}{n^2}$ the most optimal choice?

If $\delta$ is not chosen to be powerlaw $(n)$ and something like $e^{-n}$, we shall end up exploring too much.

## Morals

1) $\mathbb{E}[T_i(n)] = O\left(\frac{\log(n)}{\Delta_i^2}\right)$

2) Note that the way we have derived this bound tries to suggest that for a modest $(n)$, $R_n$ increases very fast when $\Delta_i$ decreases. This is not very intuitive & logical.

**Thm** For $\delta = \frac{1}{n^2}$, for UCB,

$$R_n \leq 8\sqrt{nK\log(n)} + 3\sum \Delta_i$$

(Removing inverse-dependence on the sub-optimality gap)

This bound is also sometimes called the Worst-Case bound or the instance-independent bound, though it is not completely fitting to say so

**Proof** : $R_n = \sum_{i:\Delta_i < \Delta} \Delta_i \mathbb{E}[T_i(n)] + \sum_{i:\Delta_i \geq \Delta} \Delta_i \mathbb{E}[T_i(n)]$

$$\leq n\Delta + \sum_{i:\Delta_i \geq \Delta} \left(3\Delta_i + \frac{16\log n}{\Delta_i}\right)$$

$$\leq 3\sum \Delta_i + n\Delta + \frac{16K\log(n)}{\Delta}$$

If we set, $\Delta = \sqrt{\frac{16K\log(n)}{n}}$, we get

$$R_n \leq 3\sum \Delta_i + 8\sqrt{nK\log(n)}$$ ... Chapter 7, Lattimore