Week 7 - Lecture 32

1) Greedy Algorithm

$$I_t = \arg\max_i \hat{\mu}_i(t-1)$$

Arm 1 : Ber (0.5)

Arm 2 : Ber (0.8)

$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	
$I_1=1$	$I_2=2$	$I_3=1$	$I_4=1$	$I_5=1$	(mean for Arm 1 is true no., and mean for Arm 2 is zero)
1	0	0	0		

$\therefore$ Can get stuck with a bad arm

2) ε-Greedy Algorithm

Input : ε

$$I_t = \begin{cases} \arg\max_i \hat{\mu}_i(t-1) & \text{w.p. } (1-\varepsilon_t) \leftarrow \text{exploit} \\ \text{select an arm uniformly at random} & \text{w.p. } \varepsilon_t \leftarrow \text{explore} \end{cases}$$

$\varepsilon = 0$ is equivalent to Greedy

$\varepsilon = 1$ always select arms uniformly at random.

This will also give linear regret.

Can decrease ε with increasing t $\therefore$ ε_t decays as t increases.

In initial rounds better to explore more, then when t is large and we have lots of samples of arms then decrease exploration and increase exploitation.

3)

$$P_i = \frac{e^{\eta \hat{\mu}_i(t)}}{\sum_{j=1}^{K} e^{\hat{\mu}_j(t)}}$$

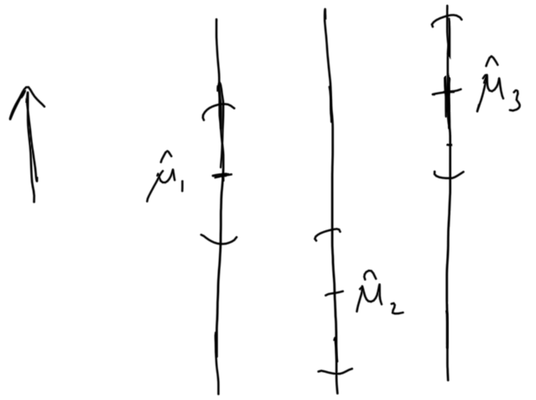
Similar to weighted majority

$P = (P_1, P_2, \ldots P_K)$

$I_t \sim P$

Upper Confidence Band (UCB) Algorithm

Introduced in 2002



The estimates are obtained after certain number of arms are played.

The larger the band the lesser number of times that arm has been played.

Three arms with 3 estimates of the mean and bands.

$\rightarrow$ Treat the upper confidence band as the actual mean, at any round, and play the arm that has highest upper confidence bound.

→ Thus estimate the confidence intervals and apply greedy algorithm

We know:

$$\hat{\mu}_i = \frac{1}{n} \sum_{t=1}^{n} x_{ti}$$

$$\Pr\left\{ \hat{\mu}_i - \mu_i \geq \varepsilon \right\} \leq \exp\left\{ \frac{-n\varepsilon^2}{2\sigma^2} \right\}$$

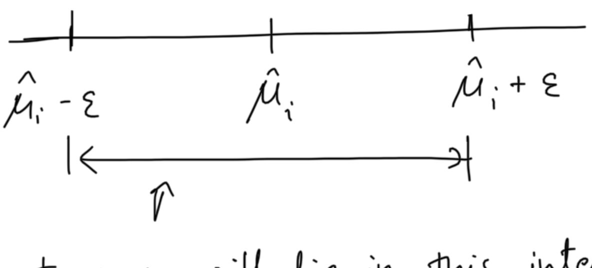
Let $\varepsilon = \sqrt{\dfrac{2\sigma^2 \log 1/\delta}{n}}$

Then,

$$\Pr\left\{ \hat{\mu}_i - \mu_i \geq \varepsilon \right\} \leq \delta$$

$$\Pr\left\{ \hat{\mu}_i - \mu_i \leq \varepsilon \right\} \leq \delta$$

Thus can estimate the confidence bound for μ_i



true μ_i will lie in this interval w.p. $(1 - 2\delta)$

Thus confidence intervals will be:

$$\hat{\mu}_i + \sqrt{\frac{2\sigma^2 \log 1/\delta}{n}} \quad , \quad \hat{\mu}_i - \sqrt{\frac{2\sigma^2 \log 1/\delta}{n}}$$

Thus after estimating the confidence intervals, & choose greedily the one with the highest upper confidence limit.

This is thus called:

<u>Optimism in the face of Uncertainity</u> (OFU)

Assuming $\sigma^2 = 1$

$$UCB_i(t-1) = \hat{\mu}_i(t-1) + \sqrt{\frac{2 \log t}{T_i(t-1)}}$$

where $T_i(t-1)$ = no. of samples of i till round $(t-1)$

$$I_t = \arg\max_i \; UCB_i(t-1)$$

<u>UCB Algo</u>

<u>Input</u>: No. of arms $= K$

Play each arm once

$t = K+1, K+2, \ldots$

$$I_t = \arg\max_i \; \mu_i(t-1) + \sqrt{\frac{2 \log t}{T_i(t-1)}}$$

How does this combine exploration & exploitation :

→ If we have played an arm less then $T_i(t-1)$ is going to be smaller ∴ UCB_i will be larger
∴ may be forced to choose that arm.
But by choosing this arm we are getting a better estimate

$$\underbrace{\hat{\mu}_i(t-1)}_{\text{exploit (term i)}} + \underbrace{\sqrt{\frac{2 \log t}{T_i(t-1)}}}_{\text{explore (term ii)}}$$

→ Due to <u>log t</u> term, the exploration never stops, as log t keeps increasing, so it makes term ii larger and thus exploration never stops.

→ But log t increases slowly, thus sub-optimal arms, even if played will be less no. of times

→ Even when sufficient rounds have been played and a good estimate of $\hat{\mu}_i$ has been obtained, the exploration never stops due to log t term in term ii.