Lec 7

Page 1

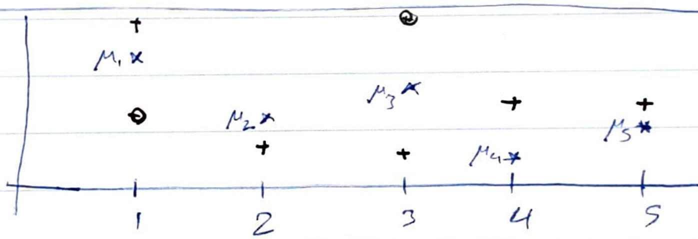
CS 59000 RL

Exploration and exploitation.

From last lecture we prove that:

$$R_n(\pi, \nu) = R_n = \sum_{i}^{K} \underbrace{\Delta_i}_{\text{supoptimality gap}} E[\underbrace{T_i(n)}_{\text{the number of time are } i \text{ is pulled.}}]$$

$$\Delta_i = \mu_1 - \mu_i$$



Explore - Then - Commit.

Choose each arm/action sequentially m times.

Then follow the best one estimated.

Two phases

- Pure exploration phase: $1 \leq t \leq mK$

$$A_t = t \mod K + 1$$

- Pure exploitation phase: for each arm estimate μ_i $_{mK<n}$: $\hat{\mu}_i = \frac{1}{m} \sum_{t=1}^{mK} X_t \, I\{A_t = i\}$

when $mK \geq n$: $\hat{\mu}_i = \frac{1}{m \wedge \lceil \frac{n}{K} \rceil} \sum_{t=1}^{n} X_t \, I\{A_t = i\}$

2)

Then find $a \in \arg\max \hat{\mu}_i$
and follow a for the remaining part
i.e. $A_t = a \quad t > mK$

Theorem: For ν, with R-sub Gaussian arms, the regret of ETC (Explore then Commit) satisfies:

$$R_n \leq m \wedge \left\lceil \frac{n}{K} \right\rceil \sum_{i=1}^{K} \Delta_i + (n-mK)^+ \sum \Delta_i \exp\left(-\frac{m\Delta_i^2}{4R}\right)$$

when $mK < n \rightarrow R_n \leq m \sum_{i=1}^{K} \Delta_i + (n-mK) \sum_{i=1}^{K} \Delta_i \exp\left(-\frac{m\Delta_i^2}{4R}\right)$

$a \wedge b$: $\min\{a, b\}$, $a \vee b$: $\max\{a, b\}$

Proof: $R_n = \sum_{i=1}^{K} \Delta_i \, E[T_i(n)]$

$$E[T_i(n)] \leq m \wedge \left\lceil \frac{n}{k} \right\rceil + (n-mk)^+ \mathbb{P}(A_{mK+1} = i)$$

$$\leq m \wedge \left\lceil \frac{n}{k} \right\rceil + (n-mk)^+ \mathbb{P}\left(\hat{\mu}_i(mK) \geq \max_{j \neq i} \hat{\mu}_j(mK)\right)$$

$$\leq m \wedge \left\lceil \frac{n}{K} \right\rceil + (n-mK)^+ \mathbb{P}\left(\hat{\mu}_i(mK) \geq \hat{\mu}_1(mK)\right)$$

$$= m \wedge \left\lceil \frac{n}{K} \right\rceil + (n-mK)^+ \mathbb{P}\left(\hat{\mu}_i - \mu_i - (\hat{\mu}_1(mK) - \mu_1) \geq \underset{\mu_1 - \mu_i}{\Delta_i}\right)$$

3)

Note the $\underbrace{\hat{\mu}_i(mK) - \mu_i}_{\frac{R}{m}\text{-sub Gausian}} - \underbrace{(\mu_1(mK) - \mu_1)}_{\frac{R}{m}\text{-sub-Gaussian}}$

$\Rightarrow$ the whole thing is $\frac{2R}{m}$ - sub-Gaussian.

↳ application of Hoeffding inequality:

$$\mathbb{P}\left(\hat{\mu}_i(mK) - \mu_i - (\hat{\mu}_1(mK) - \mu_1) \geq \Delta_i\right) \leq \exp\left(-\frac{m\Delta_i^2}{4R}\right)$$

↳ $R_n \leq m \wedge \lceil \frac{n}{K} \rceil \sum_{i=1}^{K} \Delta_i + (n - mK)^+ \sum_{i=1}^{K} \Delta_i \exp\left(-\frac{m\Delta_i^2}{4R}\right)$

If the learner explores alot $(m \gg 1)$, the second term vanishes. But, the first term blows up. We just explored too much. If m is small, we explore a little, the first term is small but we most likely will choose a wrong arm for the exploitation and the second term is over kill.

$\boxed{\text{Trade off Between Exploration and Exploitation}}$

4)

Conside a 2 armed bandit: $(mK > n)$

$$R_n \leq m\Delta_2 + n\Delta_2 \exp\left(-\frac{m\Delta_2^2}{4R}\right)$$

$\hookrightarrow$ choose $m = \max\left\{1, \frac{4R}{\Delta_2^2}\log\left(\frac{n\Delta_2^2}{4}\right)\right\}$

$\min\{n\Delta_2,$

$$\Rightarrow R_n \leq \Delta_2 + \frac{4}{\Delta_2}\left(1 + \max\left\{0, \log\frac{n\Delta_2^2}{4R}\right\}\right)$$

Important note: the bound is logarithmin in n, nice algorithm.

How come? we were able to do so, because we are assuming we know Δ_2, what a bumer.

5)

The Upper Confidence bound (UCB)

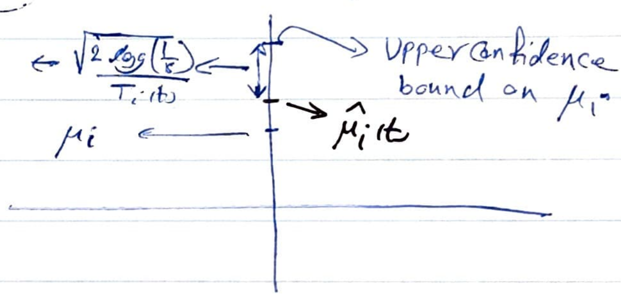
Intuition: Let pull arm i for $T_i(t)$ times.
loosely speaking (using Hoeffding's)

$$\underbrace{\mu_i - \mu_i(t)}_{\text{emperical estimate}} \leq \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{\underbrace{T_i(t)}_{\text{the number of sample.}}}} \rightarrow \text{at least } 1-\delta$$

$\hat{\mu}_i(t) = \mu_{i, \underbrace{T_i(t)}}$

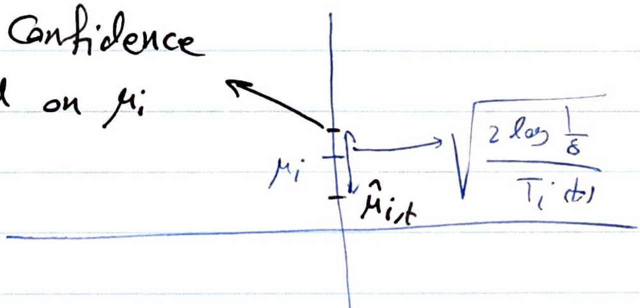
bonus

what it means:



$$\mu_i \leq \mu_{i,t} + \sqrt{\frac{2\log\frac{1}{\delta}}{T_i(t)}}$$

$\boxed{\text{assuming } R=1 \text{, arms are 1-sub-Gaussian}}$

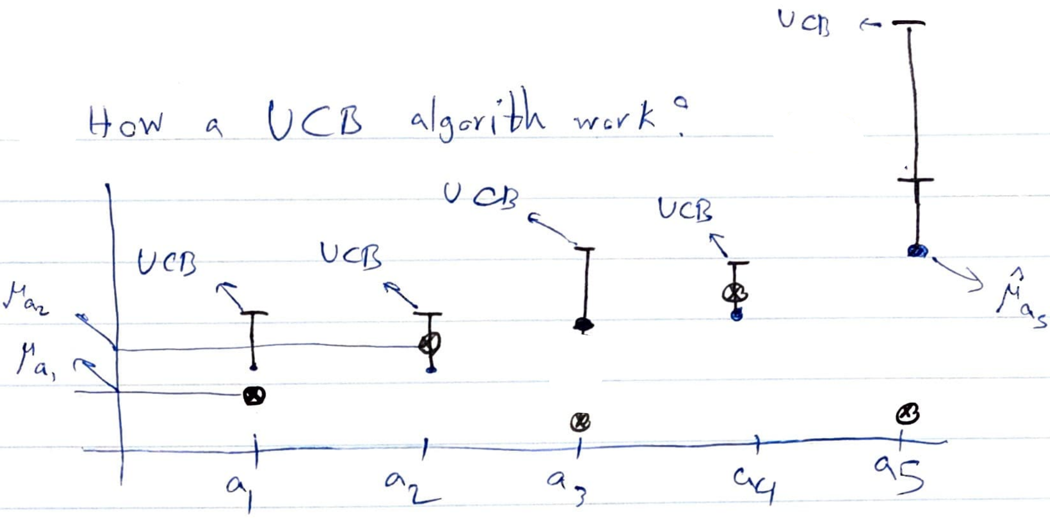


$\mu_i(t) + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{T_i(t)}}$ represent upper confidence bound!

$$UCB(t-1, \delta) = \hat{\mu}(t) + \sqrt{\frac{2\log\frac{1}{\delta}}{T_i(t)}}$$

6)

How a UCB algorith work?



At time t, compute $UCB_i(t-1, \delta)$ for all the arms. The choos the most promissing arm.

Despite the presence of uncertainty, we act according to the most optimistic senario.

It is known as *Optimism in the Face of Uncertainty (OFU)*

Theorem: UCB algorithm on a stochastic K-armed bandit with 1-sub-Gaussian reward acheives regret of

$$R_n \leq 3 \sum_{i=1}^{K} \Delta_i + \sum_{i:\Delta_i > 0} \frac{16 \log(n)}{\Delta_i}$$

when $\delta = \frac{1}{n^2}$