Lec 8

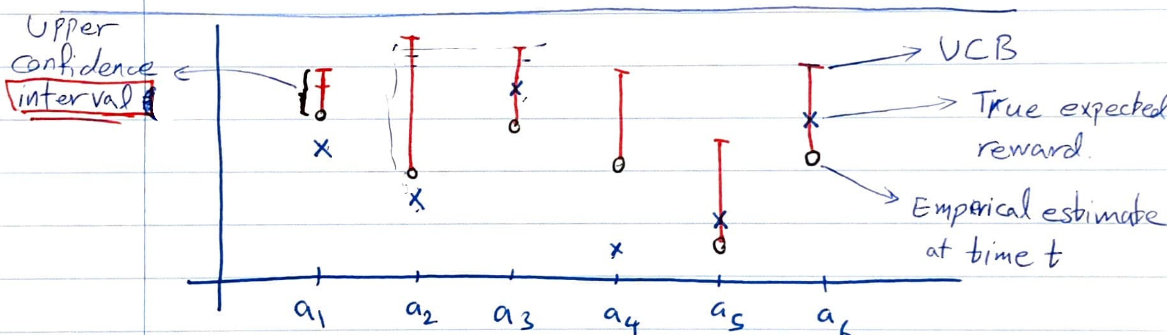
Page 1

CS 59000 RL

Finit_armed bandit

Agend:

- UCB
- Lower bound.
- Minimax optimal
- Some more side information
- Adverserial bandit.



Recall the UCB type algorithm we talk in the previouse lecture.

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

$$R_n(\underset{\substack{\downarrow \\ \text{Policy} \\ \downarrow \\ \text{UCB algorithm}}}{\pi}, \underset{\substack{\downarrow \\ \text{envionment}}}{\nu}) \leq 3\sum_{i=1}^{K} \Delta_i + \sum_{i:\Delta_i>0} \frac{16\log(n)}{\Delta_i}$$

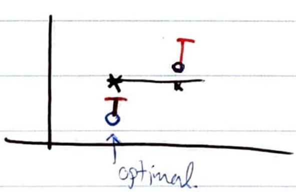
when $\delta = \frac{1}{n^2}$

Proof: Recall that $R_n = \sum_{i=1}^{K} \Delta_i \, E[T_i(n)]$
we construct proof
by bounding $E[T_i(n)]$ for each sub optimal arm

When we are sure that we are not mistakenly pulling arm $i > 1$? (For simplicity, assume arm 1 is one of the optimal arms)

Consider the following events.

$$\mu_1 < \min_{t \in [n]} UCB_1(t, \delta)$$; Upper Confidence bound on the optimal arm is a valid bound for $1 \le t \le n$



- For some number u_i

$$\underbrace{\hat{\mu}_{i,u_i}}_{\text{emperical mean after pulling i'th arm for } u_i \text{ times}} + \sqrt{\frac{2}{u_i} \log \frac{1}{\delta}} < \mu_1$$

After u_i times of pulling i'th arm the upper confidence bound is smaller μ_1

Consider the event G_i

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

First we show that in event G_i, $T_i(n) < u_i$ using contradiction.

If $T_i(n) > u_i$, then there is a $t \leq n$ where $T_i(t-1) = u_i$ and $A_t = i$.

By definition of $UCB_i(t-1, \delta)$, we have

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

$$= \hat{\mu}_{i,u_i} + \sqrt{\frac{2\log(\frac{1}{\delta})}{u_i}}$$

$$\xrightarrow{\text{under event } G_i} < \mu_1$$

$$< UCB_1(t-1,\delta)$$

therefore at time t, we would not choose ith arm $\rightarrow$ contradiction $\Rightarrow$ Therefor $T_i(n) \leq u_i$.

— Using this inequality we have:

$$E[T_i(n)] = \int_{G_i} T_i(n)\, d\mathbb{P}_{\nu,\pi} + \int_{G_i^c} T_i(n)\, d\mathbb{P}_{\nu,\pi}$$

$$= E[I_{G_i} T_i(n)] + E[I_{G_i^c} T_i(n)]$$

$$\Rightarrow E\left[T_i(n)\right] \leq E\left[I_{G_i}\, u_i\right] + n\, E\left[I_{G_i^c}\right]$$

$$= u_i \underbrace{E\left[I_{G_i}\right]}_{\mathbb{P}(G_i)} + n\, \mathbb{P}(G_i^c)$$

$$\leq \underbrace{u_i}_{\text{we will choose } u_i \text{ later.}} + n \underbrace{\mathbb{P}(G_i^c)}_{\text{Let us upper bound this.}}$$

$$G_i^c = \left\{ \left\{ \mu_1 \geq \min_{t \in [n]} UCB_1(t, \delta) \right\} \right.$$

$$\left. \cup \left\{ \mu_{i,u_i} + \sqrt{\frac{2 \log \frac{1}{\delta}}{u_i}} \geq \mu_i \right\} \right\}$$

Let's study the first event

$$\left\{ \mu_1 \geq \min_{t \in [n]} UCB_1(t, \delta) \right\}$$

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

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

$\Rightarrow$ Using union bound $\rightarrow$

$$\mathbb{P}\left(\mu_1 \geq \min_{t \in [n]} UCB_1(t)\right) \leq \sum_{s=1}^{n} \mathbb{P}\left(\mu_1 > \hat{\mu}_{1,s} + \sqrt{\frac{2 \log \frac{1}{\delta}}{s}}\right) \leq n\delta$$

Now Let's stude $\{ \hat{\mu}_{i,u_i} + \sqrt{\frac{2 \log \frac{1}{\delta}}{u_i}} \geq \mu_1 \}$

Let u_i be Large enough such that

$$\Delta_i - \sqrt{\frac{2 \log \frac{1}{\delta}}{u_i}} \geq c \Delta_i$$

$$\Rightarrow \mathbb{P}\left(\hat{\mu}_{i,u_i} + \sqrt{\frac{2 \log \frac{1}{\delta}}{u_i}} \geq \mu_1 \right) = \mathbb{P}\left(\hat{\mu}_{i,u_i} - \mu_i \geq \Delta_i - \sqrt{\frac{2 \log \frac{1}{\delta}}{u_i}} \right)$$

$$\leq \mathbb{P}\left(\mu_{i,u_i} - \mu_i \geq c\Delta_i \right) \leq \exp\left(- \frac{u_i c^2 \Delta_i^2}{2} \right)$$

using union bound

$$\Rightarrow \mathbb{P}(G_i^c) = n\delta + \exp\left(- \frac{u_i c^2 \Delta_i^2}{2} \right)$$

$\Rightarrow$ therefore

$$E[T_i(n)] \leq u_i + n\left(n\delta + \exp\left(\frac{-u_i c^2 \Delta_i^2}{2} \right) \right)$$

Let us set $u_i = \left\lceil \frac{2 \log (\frac{1}{\delta})}{(1-c)^2 \Delta_i^2} \right\rceil$ and set $\delta = \frac{1}{n^2}$

$$E[T_i(n)] \leq \frac{2 \log \frac{1}{\delta}}{(1-c)^2 \Delta_i^2} + 1 + n^{\frac{1-2c^2}{1-c}}$$

let be set $c = \frac{1}{2} \Rightarrow$ then

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

$$R_n = \sum \Delta_i \, E[T_i(n)]$$

$$\Rightarrow R_n \leqslant 3\sum_{i=1}^{K} \Delta_i + \sum_{\substack{i=1 \\ \Delta_i > 0}}^{k} \frac{16 \log(n)}{\Delta_i^2}$$

Theorem: The regret of UCB algorith on 1-sub-Gaussian stachastic K-armed bandit is upper bounded by

$$R_n \leqslant 8\sqrt{nK \log(n)} + 3\sum_{i=1}^{K} \Delta_i$$

proof: For a parameter $\Delta > 0$

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

$$= \sum_{i:\Delta_i < \Delta} \Delta_i \, E[T_i(n)] + \sum_{i:\Delta_i \geqslant \Delta} \Delta_i \, E[T_i(n)]$$

$$\leqslant n\Delta + \frac{16 K \log(n)}{\Delta} + 3\sum_i \Delta_i$$

$$\Rightarrow \text{Let's set } \Delta = \sqrt{\frac{16 K \log n}{n}}$$

$$\Rightarrow R_n \leqslant 8\sqrt{nK \log n} + 3\sum_{i=1}^{K} \Delta_i$$

$$\frac{R_n}{\log} \rightarrow$$

Lower Bound: The worst case regret of a policy π on a class of environment $\mathcal{E}$ is:

$$R_n(\pi, \mathcal{E}) = \sup_{\nu \in \mathcal{E}} R_n(\pi, \nu)$$

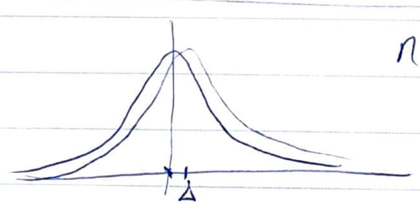
Let Π be the set of all policies, The minimax regret of Π on $\mathcal{E}$ is

$$R_n^*(\Pi, \mathcal{E}) = \inf_{\pi \in \Pi} R_n(\pi, \mathcal{E}) = \inf_{\pi \in \Pi} \sup_{\nu \in \mathcal{E}} R_n(\pi, \nu)$$

A policy in minimax optimal if it acheives minimax regret

Theorem. Let $\mathcal{E}^K$ be the set of K-armed stochastic Gaussian bandits, with unit variance and mean $\mu \in [0,1]^K$. Then there exist a constant $c > 0$ such that for $K > 1$ and $n \geq K$, it hold's

$$R_n^*(\mathcal{E}^K) \geq c\sqrt{(K-1)n}$$



n

$1 \quad \mu: (\Delta, 0 \ldots 0 \ldots)$

$\mu(\Delta, \ldots 2\Delta, \ldots)$