Lecture 2

1. sign-up sheet
2. PTE
3. Intro of me

4. MAB

(1) Recall of RL & MAB setting

RL: $\mathcal{S}$, $\mathcal{A}$, $\mathcal{P}$, r, γ

MAB: $\times$ $\mathcal{A}$ $\times$, r, n

(*) time step t: $a_k \in \mathcal{A} = \{a_1, \dots, a_K\}$

$r_a \sim P(x \mid \theta_a)$

we know the reward follows a distribution

(e.g. $r_{a_i} \sim N(\mu_{a_i}, 1)$)

$r_{a_i} = \begin{cases} 1 & \text{w.p. } p_i \\ 0 & \text{w.p. } 1-p_i \end{cases}$

The optimal arm is $i^* = \arg\max_i \mathbb{E}[r_{a_i}]$

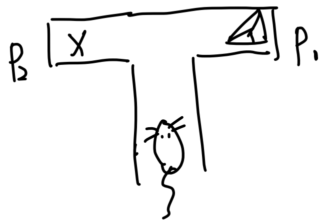
(e.g. $i^* = \arg\max_i \mu_{a_i}$, $i^* = \arg\max_i p_i$)

(*) at the beginning of step t, we need to make a decision based on the history data observed at step $1, \dots, t-1$. $\pi: \{(A_i, r_{A_i})\}_{i=1}^{t-1} \to \mathcal{A}$

(*) Goal: $\max \mathbb{E}[\sum_{i=1}^{n} r_{A_i}]$

2) History.

- Indroduced by William R. Thompson in 1933.
- Name comes from 1950s by Mosteller & Bush studying animal learning



- Become popular in different applications.
  - adaptive experimental design (recognized by US Food & Drug Administration)
  - new recommendation
  - dynamic pricing.
  - ad placement → e.g.

3). A simple example & a naive approach.

Q: I can put product 1 or 2 at an advertisement place, Not sure the customer likes which. Measure it by click rate. For the nxt 100 customers who visit the website, what is a good ~~strategy~~ strategy that I can get the most clicks?

why it can be viewed as an MAB prob

a_1: product 1: $r(a_1) \sim Ber[p_1]$ ← The prob of clicking the product is p_i

a_2: product 2: $r(a_2) \sim Ber[p_2]$

- Naive approach:

  - Try a_1 25 times
  - Try a_2 25 times.
  - Pull the one with higher emprical means for the
  - rest of 50 times.

- why it is not optimal?

  - when $\mu_1 = 10$, $\mu_2 = -10$ -
  - when $\mu_1 = 1$, $\mu_2 = -0.9$

  trade off btw exploration & exploitation.

- A better Algorithm:

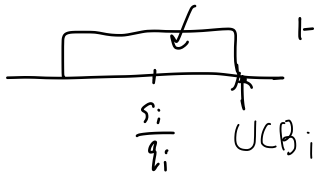
  - $\frac{S_i}{q_i} + \sqrt{\frac{c}{q_i}}$

- why it is better? < when $|\mu_1 - \mu_2|$ is large / when $|\mu_1 - \mu_2|$ is small.

4). UCB

$$\text{UCB}: \quad UCB_i = \frac{s_i}{q_i} + \sqrt{\frac{2\log(\frac{1}{\delta})}{q_i}}$$

μ_i is in this interval with $1-\delta$ prob



Lemma: For 1- subguassian r.v. $(X - \mathbb{E}X)$,

$$\mathbb{P}\left[\left|\mathbb{E}X - \frac{\sum_{t=1}^{i} X_t}{i}\right| \leq \sqrt{\frac{2\log(\frac{1}{\delta})}{i}}\right] \geq 1-\delta$$

Intro of σ- subguassion

Formal Def

$\hookrightarrow$ σ- subguassian : $\iff$ $\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$ for $\forall \lambda \in \mathbb{R}$

$\hookrightarrow$ $\mathbb{P}(|X| \geq \varepsilon) \leq e^{-\frac{\varepsilon^2}{2\sigma^2}}$ $\leftarrow$ The tail decays exponentially fast

σ is larger, decays faster.

useful ineq.

Property of σ- subguassian :

① $V[X] \leq \sigma^2$,

② cX is $|c|\sigma$ - subguassian

③ If X_1, X_2 are σ_1, σ_2 - subguassian

then $X_1 + X_2$ is $\sqrt{\sigma_1^2+\sigma_2^2}$ - subguassian pf of Lemma: $\sum_{t=1}^{i}(X_t-\bar{X})$ is $\sqrt{i}$ - sub guassian

$$\frac{1}{i}\sum_{t=1}^{i}(X_t-\bar{X}) \text{ is } \frac{1}{\sqrt{i}} \text{ - sub guassian}$$

$$\mathbb{P}\left(\left|\frac{1}{i}\sum_{t=1}^{i}(X_t-\bar{X})\right| \geq \sqrt{\frac{2\log(\frac{1}{\delta})}{i}}\right) \leq e^{-\frac{\frac{2\log(\frac{1}{\delta})}{i}}{2/i}}$$

$$= e^{-\log(\frac{1}{\delta})} = \delta$$

(HW: prove that Normal distribution, Bernoulli distribution, bounded distribution are all σ- sub guassian)

Extension: If $X-\bar{X}$ is σ- sub guassian, then $UCB_k(t)$ should be

$$\frac{S_k(t)}{q_k(t)} + \sqrt{\frac{2\sigma^2\log(\frac{1}{\delta})}{q_k(t)}}$$

- What theoretical guarantee do we have?

  - $\mathbb{A} = \{a_1, \cdots, a_k\}$.

  - reward of a_i : $X_{a_i} \sim p(x \mid \theta_i)$

  - Horizon : total # of pulls.

  - policy : a mapping from the history data → distribution is action space.

    $h_{i-1} \rightarrow A_i$

  - Goal : maximize $\mathbb{E}\left[\sum_{i=1}^{n} X_{A_i}\right]$

    e.g. UCB : history data At the beginning of each round i, is summarized by $(s_{t-1}^i, q_{t-1}^i)_{i=1}^k$

    $$A_t = \max_i \left\{ \frac{s_{t-1}^i}{q_{t-1}^i} + \sqrt{\frac{2\log(1/\delta)}{q_{t-1}^i}} \right\}.$$

  - regret : $R_n = n \max_a \mu_a - \mathbb{E}\left[\sum_{i=1}^{n} X_{A_i}\right]$

    $\Leftrightarrow$ min regret

→ Thm : k-armed 1-subgaussian bandit prob.

for $\forall$ horizon n, if $\delta = \frac{1}{n^2}$, then

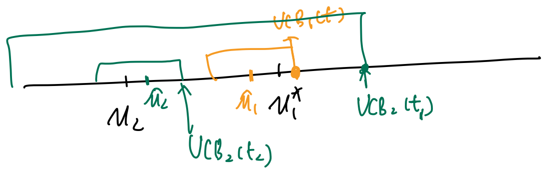
$$R_n = n \max_{a \in A} \mu_a - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right] \leq 3 \sum_{i=1}^{k} \Delta_i + \sum_{i:\Delta_i > 0} \frac{16 \log(n)}{\Delta_i}$$

(where $\Delta_i = \mu_i - \max_{a \in A} \mu_a$)

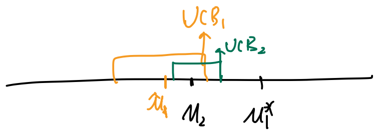
High-level idea: WLOG, μ_1 is the optimal arm

- When the regret > 0? when suboptimal arm $i > 1$ will be selected

$UCB_1 > u_1^*$



$UCB_1 < u_1^*$



selected at least one of them happens.

1) $UCB_i(t-1) \geq \mu_*$ $\rightarrow$ when this happens sufficiently large, $UCB_i \rightarrow u_i < \mu_*$ and then it won't happen

2) $UCB_{i^*}(t-1) < \mu_*$ $\rightarrow$ this will unlikely happen b/c UCB is the ucb of i^*-th arm.

It happens when $UCB_i > UCB_1$

① $UCB_1 > \mu_1$ $\leftarrow$ this happens w.p. $1-\delta$

$UCB_i > UCB_1 > \mu_1$

↳ why the # of times this will happen has an upper bd?

↑ subguassian could give it an explicit upper bd.

$UCB_i \rightarrow \mu_i < \mu_1 \rightarrow X$

② $UCB_1 < \mu_1$ $\leftarrow$ this happens w.p. $\frac{\delta}{2}$

The UCB value of the optimal arm is always > n.

"good event"

$$G_i = \{ \mu_1 < \min_{t\in[n]} UCB_1(t,\delta) \} \cap \{ \underbrace{\hat{\mu}_{i u_i}}_{\text{The empirical mean of i-th arm with } u_i \text{ pulls.}} + \sqrt{\frac{2}{u_i}\log(\tfrac{1}{\delta})} < \mu_1 \}$$

After U_i pulls of i-th arm, its UCB value $< \mu_1$

u_i to be determined later.

(*) Key pt: exploring i-th arm u_i times, its UCB is smaller than the smallest UCB for optimal arm.

We will show two things:

1). If G_i occurs, then arm i will be played at most u_i times : $T_i(n) \le u_i$

2). ~~The~~ G_i^c occurs with low prob. $\mathbb{P}(G_i^c) \le n\delta + e^{-\frac{u_i c^2 \Delta_i^2}{2}}$

holds for $\forall u_i,\ c\in(0,1)$

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

$$\le u_i + \mathbb{P}(G_i^c(n))\, n$$

$$\le u_i + \left(n\delta + e^{-\frac{u_i c^2 \Delta_i^2}{2}}\right)$$

plug in $u_i = \left\lceil \frac{2\log(1/\delta)}{(1-c)^2\Delta_i^2} \right\rceil$ (if $u_i \ge n$, (*) always holds). & $\delta = \frac{1}{n^2}$

$$\Rightarrow \mathbb{E}[T_i(n)] = \left\lceil \frac{2\log(n^2)}{(1-c)^2\Delta_i^2} \right\rceil + 1 + n^{1-2c^2/(1-c)^2}$$

plug in $c = \frac{1}{2}$

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

$$R_n = \sum_i \Delta_i \mathbb{E}[T_i(n)] = \sum_i 3\Delta_i + \frac{16\log(n)}{\Delta_i}$$

- proof of 1)

- Intro of σ-subguassian

- proof of 2).
- generalization

pf of 1): Contradiction: If i-th arm is pulled u_i times, then $UCB_i < \mu_1$, it won't be pulled any more

If $T_i(n) > u_i \Rightarrow \exists t\ T_i(t-1) = u_i,\ A_t = i$

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

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

$$< \mu_1 < UCB_1(t-1, \delta)$$

$\Rightarrow A_t = \arg\max_j UCB_j(t-1,\delta) \neq i$ ⊗

pf of 2)

$$G_i^c = \underbrace{\{\mu_1 \geq \min_{t \in [n]} UCB_1(t,\delta)\}}_{A} \cup \underbrace{\{\hat{\mu}_{1 u_i} + \sqrt{} \geq \mu_1\}}_{B}.$$

at least one of UCB_1 value is less than μ_1

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

$$\leq \sum_{s=1}^{n} \mathbb{P}\left(\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right)$$

$$\leq n\delta$$

when first arm follows 1-sub guassian,
This prob $\leq \delta$

$$B = \mathbb{P}\left(\hat{\mu}_{i u_i} + \sqrt{} \geq \mu_1 \right)$$

want the empirical mean close to true mean w.h.p

$$= \mathbb{P}\left(\underline{\hat{\mu}_{i u_i} - \mu_i} \geq \mu_1 - \mu_i - \underbrace{\sqrt{\frac{2\log(1/\delta)}{u_i}}} \right)$$

u_i large enough s.t. $< (1-c)(\mu_1 - \mu_i)$

for c determined later

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

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

when the i-th arm follows 1-subgaussian. The prob has this upper bd.

$\left(\hat{\mu}_{i u_i} - \mu_i \text{ is } \frac{1}{\sqrt{u_i}}\text{-subgaussian} \right)$

$$\mathbb{P}(G_i^c) \leq n\delta + e^{-\frac{u_i c^2 \Delta_i^2}{2}}$$

Thm 2 K-armed 1-subgaussian bandit prob.

for $\forall$ horizon n, if $\delta = \frac{1}{n^2}$, then

$$R_n = 8\sqrt{kn\log(n)} + 3\sum_i \Delta_i$$

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

$$= \sum_{i:\Delta_i<\Delta} \Delta_i \mathbb{E}[T_i(n)] + \sum_{i:\Delta_i\geq\Delta} 3\Delta_i + \frac{16\log(n)}{\Delta_i}$$

$$= n\Delta + \frac{k\,16\log(n)}{\Delta} + \sum_{\Delta_i>\Delta} 3\Delta_i$$

$$\leq 8\sqrt{kn\log(n)} + 3\sum_i \Delta_i$$

UCB Algo

δ smaller $\rightarrow$ more exploitation

δ larger $\rightarrow$ more exploration

$$\mathrm{UCB}_i(t-1,\delta) = \begin{cases} \infty \\ \underbrace{\hat{\mu}_i(t-1)}_{\text{Empirical mean}} + \underbrace{\sqrt{\frac{2\log(\frac{1}{\delta})}{T_i(t-1)}}}_{1-\delta \text{ upper bound for 1-subguassian}} \end{cases}$$

HW: Try UCB on $N(\mu_1,\sigma)$, $N(\mu_2,\sigma)$ for different $\Delta = \mu_1 - \mu_2$

$\mathrm{Ber}(p_1)$, $\mathrm{Ber}(p_2)$. for different $\Delta = p_1 - p_2$.