Multi-Armed Bandits

Given : K arms ; for each arm α, reward dist D_α w/ mean μ_α. (unknown) $\in [0,1]$

T rounds ($T \gg K$)

~~Output : Arms $\alpha_1, \alpha_2, \cdots, \alpha_T$ played.~~

Procedure In each round $t = 1, \cdots, T$,

1. ALG plays an arm α_t
2. Reward r_t is sampled indep from D_{α_t}
3. ALG learns r_t.

Defn (Regret)

$$R(T) = \mu^* \cdot T - \sum_{t=1}^{T} \mu_{\alpha_t}, \quad \text{where } \mu^* = \max_{\alpha \in K} \mu_\alpha .$$

Generally, consider $\mathbb{E}[R(T)]$ where the expectation is over the randomness of D_α's & the alg's choices.

Fact (Hoeffding's ineq)

Given mutually indep (not necessarily identically dist) $X_1, X_2, \cdots, X_n$, let

$$\overline{X}_n := \frac{X_1 + \cdots + X_n}{n} \quad \text{and} \quad \overline{\mu}_n := \mathbb{E}[\overline{X}_n] = \frac{\mu_1 + \cdots + \mu_n}{n}.$$

We then have, for any T, $\Pr\left[|\overline{X}_n - \mu_n| \le \sqrt{\frac{2 \log T}{n}} \right] \ge 1 - \frac{2}{T^4}$.

Alg 1 (Uniform exploration)

1. Try each arm N times.
2. Select one w/ highest avg reward.
3. Play the chosen one $\hat{\alpha}$ in the remaining rounds.

No.

Obsv At the end of step 4, for any arm α,

$$\Pr[|\bar{\mu}_\alpha - \mu_\alpha| \le rad] \ge 1 - \frac{2}{T^4},$$ where

$\bar{\mu}_\alpha$ denotes the observed avg reward of α, and $rad := \sqrt{\frac{2\lg T}{N}}$.

Let "clean event" be ~~the event that~~ $\bigwedge_{\alpha \in K} (|\bar{\mu}_\alpha - \mu_\alpha| \le rad)$.

Obsv

By the union bound, $\Pr[$~~clean~~ bad event$] \le \frac{2K}{T^4}$.

Lem Cond on "clean event", if ~~the chosen arm~~ $\hat{\alpha} \ne \alpha^*$, ~~we have~~

~~[illegible]~~ $\mu_{\alpha^*} - \mu_{\hat{\alpha}} \le 2\,rad$.

pf) Alg chose $\hat{\alpha}$ instead of α^* since $\bar{\mu}_{\hat{\alpha}} \ge \bar{\mu}_{\alpha^*}$. Due to the condition, $\mu_{\hat{\alpha}} + rad \ge \bar{\mu}_{\hat{\alpha}}$ & $\bar{\mu}_{\alpha^*} \ge \mu_{\alpha^*} - rad$. $\boxed{}$

Lem $\mathbb{E}[R(T)] \le NK + 2rad \cdot T + o(1)$.

pf) $$\mathbb{E}[R(T)] = \mathbb{E}[R(T) \mid clean] \Pr[clean] + \mathbb{E}[R(T) \mid bad] \Pr[bad]$$
$$\le \mathbb{E}[R(T) \mid clean] + T \cdot O\left(\frac{K}{T^4}\right).$$

Obsv $\mathbb{E}[R(T) \mid clean] \le N(K-1) + 2rad(T - NK)$. $\boxed{}$

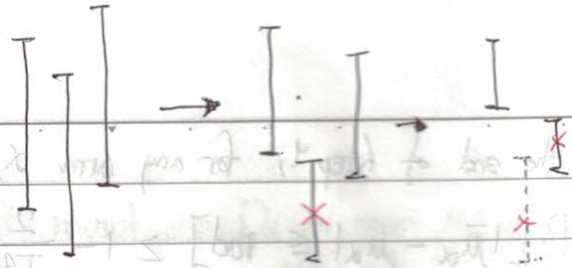
Choosing $N = O\left(\left(\frac{T}{K}\right)^{2/3} \cdot (\lg T)^{1/3}\right)$, we have $\mathbb{E}[R(T)] \le O\left(T^{\frac{2}{3}} \cdot K^{\frac{1}{3}} \cdot (\lg T)^{\frac{1}{3}}\right)$.

Obsv Fix round t and arm α. If α is played $n_t(\alpha)$ times where $\bar{\mu}_t(\alpha)$ is the observed avg reward, we then have

$$\Pr[|\bar{\mu}_t(\alpha) - \mu(\alpha)| \le rad_t(\alpha)] \ge 1 - \frac{2}{T^4},$$ where

$$rad_t(\alpha) := \sqrt{\frac{2\lg T}{n_t(\alpha)}}.$$

Set $UCB_t(\alpha) := \bar{\mu}_t(\alpha) + rad_t(\alpha)$ & $LCB_t(\alpha) := \bar{\mu}_t(\alpha) - rad_t(\alpha)$.



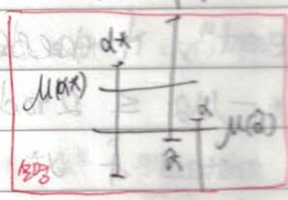
Alg 2 (Successive elimination)

1. Activate all arms.
2. For each phase :
3. play all active arms & update [LCB, UCB].
4. Deactivate all arms whose CB does not overlap the "highest" CB.

Clean event ? $\bigwedge_{\alpha \in K, t \in T} \left(|\bar{\mu}_t(\alpha) - \mu(\alpha)| \le \mathrm{rad}_t(\alpha) \right)$.

$\hookrightarrow \mu(\alpha) \in [LCB_t(\alpha), UCB_t(\alpha)]$

By the union bound, $\Pr[\text{bad event}] \le T \cdot K \cdot \frac{2}{T^4} = O\left(\frac{1}{T^2}\right)$.



Cond on clean event,

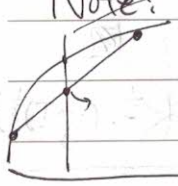
- α^* never deactivated.
- For each arm α, the CB of α overlaps CB of α^* at the $n_t(\alpha)$'th phase.

$\Rightarrow \mu(\alpha^*) - \mu(\alpha) \le 2\,\mathrm{rad}_t(\alpha) = 2\sqrt{\frac{2 \lg T}{n_t(\alpha)}}$

$\therefore$ Contribution of α to $\mathbb{E}[R(t)|\text{clean}] = n_t(\alpha) \cdot O\left(\sqrt{\frac{\lg T}{n_t(\alpha)}}\right) = O\left(\sqrt{n_t(\alpha) \lg T}\right)$

$\Rightarrow \mathbb{E}[R(t)|\text{clean}] = \sum_{\alpha \in K} O\left(\sqrt{n_t(\alpha) \lg T}\right) = O(\sqrt{\lg T}) \sum_{\alpha \in K} \sqrt{n_t(\alpha)} = O\left(\sqrt{Kt \lg T}\right)$.

Note: by Jensen's ineq.,



$$\frac{1}{K} \sum_{\alpha \in K} \sqrt{n_t(\alpha)} \le \sqrt{\frac{\sum n_t(\alpha)}{K}} = \sqrt{\frac{t}{K}} \Rightarrow \sum_{\alpha \in K} \sqrt{n_t(\alpha)} = \sqrt{Kt}$$

Lower bounds

Thm Fix T & K. For any alg, $\exists$ family of instances s.t. $\mathbb{E}[R(T)] \ge \Omega(\sqrt{KT})$.

Today, wts for $K=2$!

No.

Consider ① I_1 - $D_{\alpha_1} = Ber(\frac{1+\varepsilon}{2})$ ② I_2 - $D_{\alpha_1} = Ber(\frac{1}{2})$
- $D_{\alpha_2} = Ber(\frac{1}{2})$ - $D_{\alpha_2} = Ber(\frac{1+\varepsilon}{2})$.

Outline: think of MAB as a seq of "best-arm identification"!
in the same setting, the goal is to choose the max-reward arm! at round t

Lem Consider a "best-arm identification" w/ $t \le \frac{1}{16\varepsilon^2}$. Fix any alg.
$\exists$ arm $\alpha \in \{\alpha_1, \alpha_2\}$ s.t.

$$\Pr[\text{Alg chose } \alpha \mid I_\alpha] < 3/4.$$

~~(Using this lem, we have)~~ $T \le \frac{1}{16\varepsilon^2} \Rightarrow \boxed{\varepsilon \le \sqrt{\frac{1}{16T}}}$

Thm Fix T & any ALG. Choose an arm α uar, & run ALG on I_α w/ $\varepsilon = O(\sqrt{\frac{1}{T}})$
Then, $\mathbb{E}[R(T)] \ge \Omega(\sqrt{KT})$.

pf) By choice of ε, we can use Lem for each round $t \le T = \frac{1}{16\varepsilon^2}$.

$$\underbrace{\Pr[\alpha_t \ne \alpha]}_{\ge 1/8} = \Pr[\alpha_t \ne \alpha_1 \mid I_1] \underbrace{\Pr[I_1]}_{=\frac{1}{2}} + \Pr[\alpha_t \ne \alpha_2 \mid I_2] \underbrace{\Pr[I_2]}_{=\frac{1}{2}}.$$

one would be $\ge 1/4$

$$\therefore \mathbb{E}[R(T)] = \sum_{t=1}^{T} \Pr[\alpha_t \ne \alpha] \cdot \frac{\varepsilon}{2} \ge \frac{\varepsilon T}{16} = \Omega(\sqrt{T}).$$ ☒

KL-divergence

$$KL(p, q) = \sum_{x \in \Omega} p(x) \cdot \ln \frac{p(x)}{q(x)}.$$

Fact (Chain rule) For $p = p_1 \times p_2 \times \cdots \times p_n$ & $q = q_1 \times q_2 \times \cdots \times q_n$ where p_i, q_i on the same sample space, $KL(p, q) = \sum_{i=1}^{n} KL(p_i, q_i)$.

Fact (Pinsker's ineq) for any event $A \subset \Omega$, $2(p(A) - q(A))^2 \le KL(p, q)$!

Fact $KL(Ber(\frac{1+\varepsilon}{2}), Ber(\frac{1}{2})) \le 2\varepsilon^2$ & $KL(Ber(\frac{1}{2}), Ber(\frac{1+\varepsilon}{2})) < \varepsilon^2 \quad \forall \varepsilon < \frac{1}{2}$.

pf of Lem) Consider all possible outcomes. For example, $t=3$,

$$\begin{array}{c} \text{arm} \rightarrow \\ \text{round} \downarrow \end{array} \quad \Omega = \left\{ \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}, \cdots, \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \right\}$$

Let $p_1(\omega)$ & $p_2(\omega)$ be the prob of outcome ω in I_1 & I_2, resp.

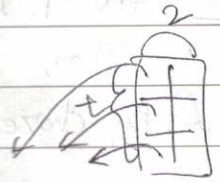
e.g., $$p_1\left(\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}\right) = \left(\frac{1-\varepsilon}{2}\right)^3 \cdot \left(\frac{1}{2}\right)^3 \qquad p_2\left(\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}\right) = \left(\frac{1}{2}\right)^3 \left(\frac{1-\varepsilon}{2}\right)^2 \left(\frac{1+\varepsilon}{2}\right)$$

As ALG is not aware of the real dist, the output of ALG depends on ~~ω~~. the outcome, i.e., $ALG : \Omega \to \{a_1, a_2\}$.

Let $\mathcal{E} := \{\omega : ALG(\omega) = a_1\}$. Then $\bar{\mathcal{E}} := \{\omega : ALG(\omega) = a_2\}$.

* Spse t.c. that $p_1(\mathcal{E}) \geq 3/4$ & $p_2(\bar{\mathcal{E}}) \geq 3/4$. We then have

$$p_1(\mathcal{E}) - p_2(\mathcal{E}) \geq 3/4 - 1/4 = 1/2. \quad \cdots (a)$$



However, observe that

$$2(p_1(\mathcal{E}) - p_2(\mathcal{E}))^2 \leq KL(p_1, p_2) = \sum_{\substack{\text{each cell} \\ \text{in } t \times 2}} KL(p_1^{\text{cell}}, p_2^{\text{cell}})$$

$$\leq 2t \cdot 2\varepsilon^2$$

$$\Rightarrow p_1(\mathcal{E}) - p_2(\mathcal{E}) \leq \varepsilon\sqrt{2t} \leq \frac{1}{2\sqrt{2}} < \frac{1}{2} \quad \text{since } t \leq \frac{1}{16\varepsilon^2}. \Rightarrow\Leftarrow \quad \square$$

~~For general K,~~

For general K, I_a : $p_a = Ber\left(\frac{1+\varepsilon}{2}\right)$, the others $Ber\left(\frac{1}{2}\right)$.

<u>Lem</u> Consider a "best-arm identification" w/ $t \leq \frac{cK}{\varepsilon^2}$ (for a small enough absolute constant c). Fix any Alg. $\exists$ at least $\lceil \frac{K}{3} \rceil$ arms a s.t.

$$Pr[\text{Alg chose } a \mid I_a] < 3/4.$$