Littemore Szepesvari

- Let $X \in [0,1]$ with mean μ;

$\forall \lambda \in \mathbb{R}$

then $$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \le e^{\lambda^2/8};$$

and $$\mathbb{E}\left[e^{\lambda(\mu-X)}\right] \le e^{\lambda^2/8}.$$

In general:

Def: A r.v is σ-subgaussian if $\forall \lambda \in \mathbb{R}$

$$\mathbb{E}\left[\exp(\lambda X)\right] \le \exp\left(\lambda^2 \sigma^2/2\right)$$

Cor: Let $X_i - \mu$ be independent σ-subgaussian r.v. Set $X = \underbrace{\frac{X_1 + \dots + X_n}{n}}$;

$\frac{\sigma}{\sqrt{n}}$

$cX \to |c|\sigma$

Then

(a) $\mathbb{E}[X] = \mu$;

(b) Set $\hat{\mu} = \frac{X_1 + \dots + X_n}{n}$;

then $$\boxed{\Pr\left[\hat{\mu} \ge \mu + \varepsilon\right] \le \exp\left(\frac{-n\varepsilon^2}{2\sigma^2}\right)}$$

and $$\Pr\left[\hat{\mu} \le \mu - \varepsilon\right] \le \underbrace{\exp\left(\frac{-n\varepsilon^2}{2\sigma^2}\right)}_{\exp(-\log(1/\delta))} \checkmark$$

<u>In particular</u>: $\forall\, \delta \in [0,1]$

$$\boxed{\Pr\left[\mu \le \hat{\mu} + \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}\right] \ge 1-\delta}$$

$$\& \ \Pr\left[\mu \ge \hat{\mu} - \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}\right] \ge 1-\delta$$

- <u>Ex of subGaussian</u> r.v.:

(a) If X is Gaussian with mean 0 & variance σ^2, X is σ-subgaussian

(b) If X has mean zero & $|X| \le B$, $B \ge 0$, X is <u>B-subgaussian</u>. ✓

(c) X has mean zero & <u>$X \in [a,b]$</u> then X is $\frac{b-a}{2}$ subgaussian

- Explore then commit bandit:

- Assume all reward distributions for all arms is 1 - Subgaussian; (ie) $\sigma = 1$ ✓ $\Leftarrow$

- Algorithms make use of this knowledge of $\boxed{\sigma}$

There are k actions; The algorithm will explore for mk rounds & choose a single action for remaining rounds.

Let $\hat{\mu}_i(t)$ = avg reward obtained from arm i, after round t;

$$(ie) \quad \hat{\mu}_i(t) = \frac{1}{T_i(t)} \sum_{s=1}^{t} \mathbb{1}\{A_s = i\} X_s .$$

Here $T_i(t) = \sum_{s=1}^{t} \mathbb{1}\{A_s = i\}$ = # action i has been played till round t;

$$\overset{\nu}{(\,,\,,\,,\,,\,,\,,\,)}$$

Explore then commit:

Input m

$$\overbrace{1\ 2\ 3 \cdots k\ \ 1\ 2 \cdots k\ \ 1\ 2 \cdots k\ \ 1\ 2 \cdots n}^{mk}$$

In round t choose

$$A_t = \begin{cases} t \bmod k & \text{if } t \le mk \\ \arg\max_i \hat{\mu}_i(mk), & t > mk \end{cases}$$

Recall: $\boxed{\Delta_i(\nu) = \mu^*(\nu) - \mu_i(\nu)}$;

Suppose $\mu_1 = \mu^*$ then $\Delta_i = \mu_1 - \mu_i$ ✓

- The above is deterministic till round mk and chooses each action m times;

Recall: $\boxed{\bar{R}_n = \sum_{i=1}^{k} \Delta_i \mathbb{E}[T_i(n)]}$;

Clearly:

$$\mathbb{E}[T_i(n)] = \underbrace{m}_{} + \underbrace{(n-mk)}_{} \Pr[\underbrace{A_{mk+1} = i}_{}]$$

$$\leq m + (n-mk)\Pr\Big[\hat{\mu}_i(mk) \geq \max_{j \neq i} \hat{\mu}_j(mk)\Big]$$

Now

$$\boxed{\Pr\Big(\hat{\mu}_i(mk) \geq \max_{j\neq i} \hat{\mu}_j(mk)\Big)} \Leftarrow$$

$$\leq \Pr\Big(\hat{\mu}_i(mk) \geq \hat{\mu}_1(mk)\Big)$$

$$= \Pr\Big[\hat{\mu}_i(mk) - \hat{\mu}_1(mk) \geq 0\Big]$$

$$= \Pr\Big[\hat{\mu}_i(mk) - \hat{\mu}_1(mk) \geq \underbrace{\Delta_i - \Delta_i}_{}\Big]$$

$$= \Pr\Big[\hat{\mu}_i(mk) - \hat{\mu}_1(mk) + \Delta_i \geq \Delta_i\Big]$$

$$= \Pr\Big[\underbrace{(\hat{\mu}_i(mk) - \mu_i)}_{\substack{\uparrow \\ \text{subgaussian} \\ \sigma = 1}} - \underbrace{(\hat{\mu}_1(mk) - \mu_1)}_{\substack{\uparrow \\ \text{subgaussian} \\ \sigma = 1}} \geq \Delta_i\Big]$$

$$\boxed{\Delta_i = \mu_1 - \mu_i}$$

H.W. $\underbrace{X_1 \text{ subgaussian } \sigma_1}_{}$ $\underbrace{X_2 \text{ subgaussian } \sigma_2}_{}$ then

X_1+X_2 is $\sqrt{\sigma_1^2+\sigma_2^2}$ subgaussian;

first arm is pulled m times till round mk;

$$\hat{\mu}_i = \frac{X_i + X_{i+k} + \cdots + X_{i+(m-1)k}}{m}$$

$$X_i + X_{i+k} + \cdots + X_{i+(m-1)k} \text{ is } \sqrt{\underbrace{1+1+\cdots+1}_{m}}$$

$$= \sqrt{m} \text{ subgaussian;}$$

$$\therefore \frac{X_i + X_{i+k} + \cdots + X_{i+(m-1)k}}{m} \text{ is } \frac{1}{\sqrt{m}} \text{ subgaussian;}$$

$$\therefore (\underbrace{\hat{\mu}_i}_{\frac{1}{\sqrt{m}}} - \underbrace{\mu_i}_{}) - (\underbrace{\hat{\mu}_1}_{\frac{1}{\sqrt{m}}} - \mu_1) .$$

H.W: If X is σ subgaussian then ∀ $c \in \mathbb{R}$ cX is $|c|\sigma$ subgaussian.

✓

$\therefore \sqrt{\frac{1}{m}+\frac{1}{m}} = \boxed{\sqrt{\frac{2}{m}}}$ subgaussian.

$$\therefore \Pr\Big[\underbrace{\hat{\mu}_i(mk)-\mu_i-\big(\hat{\mu}_1(mk)-\mu_1\big)} \ge \underline{\Delta_i}\Big]$$

$$\le \exp\left(\frac{-\Delta_i^2}{2\left(\sqrt{\frac{2}{m}}\right)^2}\right)$$

$$= \underbrace{\exp\left(\frac{-\Delta_i^2 m}{4}\right)} \checkmark$$

(m+-)

$$R_n = \sum_{a=1}^{k} \Delta_a E(T_a(n))$$

$$\therefore \boxed{R_n \le \underbrace{m\sum_{i=1}^{k}\Delta_i} + (n-mk)\sum_{i=1}^{k}\Delta_i \exp\left(\frac{-\Delta_i^2 m}{4}\right)}$$

- If m large first term dominates
- If m small, prob of commiting wrong arm ↑ & second dominates

k=2 case: say $\Delta_1 = 0$, $\Delta = \Delta_2$

$$\boxed{R_n \le m\Delta + (n-2m)\Delta\exp\left(\frac{-m\Delta^2}{4}\right)}$$

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

$$\cancel{\Delta} + n\cancel{\Delta}\, e^{\left(-m\frac{\Delta^2}{4}\right)} \cdot \left(-\frac{\Delta^2}{4}\right) = 0$$

$$1 + n e^{-m\frac{\Delta^2}{4}} \cdot \left(-\frac{\Delta^2}{4}\right) = 0$$

$$\Rightarrow \frac{1}{n} = \frac{\Delta^2}{4} \cdot \frac{1}{e^{m\Delta^2/4}}$$

$$e^{m\Delta^2/4} = \frac{\Delta^2 n}{4}$$

$$\frac{m\Delta^3}{4} = \log\left(\frac{n\Delta^2}{4}\right)$$

$$\boxed{\Rightarrow m = \frac{4}{\Delta^2}\log\left(\frac{n\Delta^2}{4}\right)} \leftarrow$$

$$\boxed{\text{Set } m = \max\left\{1, \left\lceil \frac{4}{\Delta^2}\log\left(\frac{m\Delta^2}{4}\right)\right\rceil\right\}}$$

$$\uparrow$$

- $\boxed{R_n \le \Delta + c\sqrt{n}.}$ ✓ $\bar{R}_n \le n^{2/3}$

  Sublinear in n.

- **OPTIMISM in the face of uncertainity.**

- Use the observed data to assign to each arm an "upper confidence bound", which with high probability is an overestimate of unknown mean.

- So the other arm can be played only if its UCB is greater than that of optimal arm. But that is larger than the mean of the optimal arm.

This can't happen too often – as that arm is played more its data will force us to reduce the value we have assigned to its UCB. Eventually it will fall below UCB of optimal arm.

Recall: $(X_t)_{t=1}^{n}$ a seq of 1-subgaussian r.v, mean $=\mu$ & set $\hat{\mu} = \frac{1}{n}\sum X_t$;

$$\Pr\left[\mu > \hat{\mu} + \sqrt{\frac{2\log(1/\delta)}{n}}\right] \le \delta \quad \forall \delta \in (0,1)$$

- In round t, learner has seen $T_i(t-1)$ samples of arm i; empirical mean $\hat{\mu}_i(t-1)$

Then

$$UCB_i(t-1,\delta) = \begin{cases} \infty & \text{if } T_i(t-1) = 0 \\ \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}} & \end{cases}$$

is a reasonable assignment of a UCB for the unknown mean of arm i

$\hat{\mu}_i$ → empirical mean of rewards of arm i

$\sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}$ confidence width

Algorithm:

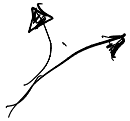
```
Input  $k, \delta$
for $t \in 1, \dots, n$ do
    Choose $A_t = \text{argmax}_i \; UCB_i(t-1, \delta)$
    observe $X_t$ and update UCB
end
```

- $(1/\delta)$ – called confidence level ✓

= δ: Can width of confidence interval for given confidence level, be significantly decreased? Turns out we cannot improve it significantly. →

- δ – should be small for ensuring optimism with high probability!

- But then the confidence interval becomes large and suboptimal arms may be played excessively.



- Idea: If the confidence interval fails & UCB_{i^*} drops below μ_{i^*} - we may not play i^* anymore & could get linear regret; select $\delta = 1/n$ in expectation the contribution to regret of this case is small.

PROBLEM: $T_i(t-1)$ is a r.v.

So δ should be smaller - $1/n^2$

THM: Stochastic k-armed 1-subgaussian bandit. For any horizon n, if $\delta = 1/n^2$

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

$R_n = O(\sqrt{n})$