Page No.: Date:

EE6106 Lecture 7 (Date: 6th Feb 2024)

STOCHASTIC BANDITS

HW1 has been posted (due in class Next Tuesday – 13th Feb 2024)

k arms (equivalently, experts)

- for each arm i, is associated a reward distribution ν_i
  mean reward of arm i $= \mu_i$, $\mu^* = \max_i \mu_i$
  $\therefore \mu_i = \int x \, d\nu_i$
- The MAB instance is $\nu = (\nu_i, 1 \le i \le k)$
- Horizon of n pulls

Anytime you pull an arm, you get an i.i.d sample from ν_i

Protocol:

At time $t \geq 1$,

- Algo chooses arm $A_t \in [k]$ to pull
- Algo gets reward $X_t \sim \nu_{A_t}$

Regret

Note: Earlier the randomness in the reward was due to the algorithm choosing randomly. Here, however, the interaction mechanism itself induces randomness. Thus, in this case, even a deterministic algorithm will incur a random regret.

Mean regret $R_m(\pi, \nu) = n\mu^* - \mathbb{E}\left[\sum_1^n X_t\right]$

π ↓ policy, ν ↓ instance

The 'learning' comes into picture because we do not know the actual reward distributions of the arms.

Define

$\Delta_i = \mu^* - \mu_i \geq 0$ known as the 'sub-optimality gap'

(Not assuming that there is a unique optimal arm)

$$T_i(t) = \sum_{1}^{t} \mathbb{1}_{\{A_t = i\}}$$ ... number of times an arm is pulled

<u>Lemma</u> :

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

<u>THE FULL-INFORMATION SETTING</u>

You "see" reward samples from ALL arms

You "get" reward X_t

$X_{1,1}$	$X_{1,2}$	
$X_{2,1}$	$X_{2,2}$	
$\vdots$	$\vdots$	
$X_{k,1}$	$X_{k,2}$	

$\longrightarrow t$

underlying assumption : the table is filled apriori, but the algorithm sees them sequentially

(this is still not the bandit setting)

logical approach: At time t, pull the arm corresponding to which the empirical mean is highest

<u>Algo</u> : Pull ~~each~~ arm ~~once~~ 1

At time $t \geq$ ~~k~~ 1,

$$A_{t+1} = \underset{i}{\operatorname{argmax}} \ \hat{\mu}_{i,t}$$

we obviously cannot use this algo' in the bandit setting.

<u>Assumption</u> : The arms are 1-subGaussian

i.e. $M_{\nu_i}(s) = \mathbb{E}[e^{sV_i}] \leq e^{s^2/2} \quad \forall s$

this is a nicely pedagogical but no so realistic assumption

<u>Lemma</u> : Under above algo,

$$R_n = O(1)$$

i.e. the regret does not grow unboundedly wrt the horizon.

<u>Proof</u> : For suboptimal i, we bound $E[T_i(n)]$

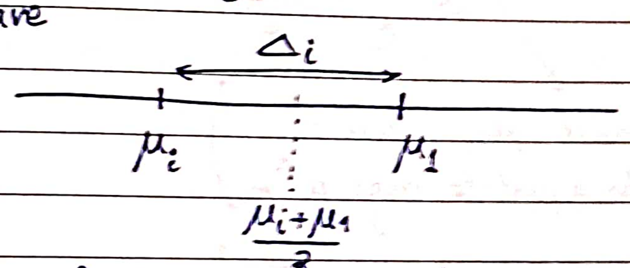
WLOG, assume that arm 1 is optimal (one of the optimal arms)

$$T_i(n) \le 1 + \sum_{2}^{n} \mathbb{1}_{\{A_t = i\}}$$

$$\therefore E[T_i(n)] \le 1 + \sum_{2}^{n} P(A_t = i)$$

$$\le 1 + \sum_{2}^{n} P(\hat{\mu}_{i,t} \ge \hat{\mu}_{1,t})$$

① If we did not have sub-Gaussianity



$$P(\hat{\mu}_{i,t} \ge \hat{\mu}_{1,t}) \le P\left(\hat{\mu}_{i,t} \ge \frac{\mu_i+\mu_1}{2}\right) + P\left(\hat{\mu}_{1,t} \le \frac{\mu_1+\mu_i}{2}\right)$$

$$= P\left(\hat{\mu}_{i,t} - \mu_i \ge \frac{\Delta_i}{2}\right) + P\left(\hat{\mu}_{1,t} - \mu_i \le \frac{\Delta_i}{2}\right)$$

② But we do have sub-Gaussianity

$$\therefore E[T_i(n)] \le 1 + \sum_{2}^{n} P(\hat{\mu}_{1,t} - \hat{\mu}_{i,t} - (\mu_1 - \mu_i) \le -\Delta_i)$$

$(\hat{\mu}_{1,t} - \hat{\mu}_{i,t}) - (\mu_1 - \mu_i)$ is $\sqrt{2/t}$ - subG

since $\frac{1}{t}\sum_{s=1}^{t} X_{1,s} = \hat{\mu}_{1,t}$ is $\sqrt{\frac{1}{t}}$ -sub G

$$\therefore E[T_i(n)] \le 1 + \sum_{2}^{n} e^{\frac{-\Delta_i^2(t-1)}{4}} < \infty$$

$$\therefore R_n = \sum_{i=1}^{k} \Delta_i E[T_i(n)] < \infty \text{ and hence } O(1)$$

□

<u>Q</u> : What about the high probability bounds on the regret? Will it be $O(1)$?

<u>Back to the MAB setting</u>

EXPLORE THEN COMMIT (ETC)

- Give m pulls to each arm
- Subsequently, pull $\arg\max_i \hat{\mu}_{i,m}$

The crux of the theorem is the regret incurred over the commit phase. What is the probability of committing to a sub-optimal arm?

Thm: Under ETC,

$$R_n \leq m \sum_{1}^{k} \Delta_i + (n-mk) \sum \Delta_i e^{-\frac{m\Delta_i^2}{4}}$$

Pick suboptimal arm i.

$$P(i \text{ is picked in commit phase}) \leq P(\hat{\mu}_{i,m} \geq \hat{\mu}_{1,m}) \leq e^{-\frac{m\Delta_i^2}{4}}$$

(because $\hat{\mu}_{i,m}$ has to be equal to $\max_j \hat{\mu}_{j,m}$)

Then, $R_n \leq m \sum_{1}^{k} \Delta_i + (n-mk) \sum \Delta_i e^{-\frac{m\Delta_i^2}{4}}$

Q: How do we determine what the optimal m is?

Case 1) $K=2$, and $\Delta_2 = \Delta > 0$

$$R_n \leq m\Delta + n\Delta e^{-\frac{m\Delta^2}{4}}$$

Differentiating wrt ~~R_n~~, m

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

$$\Rightarrow m^* \approx \frac{4}{\Delta^2} \log\left(\frac{n\Delta^2}{4}\right)$$

Then $\boxed{R_n \leq \frac{4}{\Delta}\left(1 + \log\left(\frac{n\Delta^2}{4}\right)\right)}$

Structurally, $\boxed{\text{m should be logarithmic wrt the horizon}}$

In this case, the regret is also logarithmic wrt the horizon

M T W T F S S

Page No.:

Date:

YOUVA

Note: We do not actually know Δs but we can actually replace Δ by an appropriate <u>lower bound</u>

i.e. saying that 'every suboptimal arm is at least $(\cdot)$ away from the optimal arm'

Moral: Need lower bound on smallest non-zero suboptimality gap in general

Q: Can choosing $m = c \log n$ ALWAYS lead to a sub-linear regret? What about $m = c\sqrt{n}$? (Hint: No and Yes?)

(Next: UCB-type algorithms)