EE6106 LECTURE 12 (Date : 6 March 2024)

Recall $\mu^* > \mu(\nu)$

$d_{inf}(\nu, \mu^*, \mathcal{E}) = \inf\{D(\nu_i, \nu_i') : \nu_i' \in \mathcal{E}, \mu(\nu_i') > \mu^*\}$ ①

π is consistent if $R_n(\pi, \nu) = o(n^a) \; \forall \; a > 0, \nu \in \mathcal{E}^k$

Thm : Given consistent policy π on $\mathcal{E}^k$, given ν in $\mathcal{E}^k$

$$\liminf_{n \to \infty} \frac{R_n(\pi, \nu)}{\log n} \geq \sum_{i : \Delta_i > 0} \frac{\Delta_i}{d_{inf}(\nu_i, \mu^*, \mathcal{E})} \quad ②$$

where μ^* is the mean of the optimal arm

① Here ν is a scalar (distribution of a single arm)

② Here ν is a vector (distributions of all arms, so an instance)

- Rich families have consistent policies, although they may not have logarithmic regret. Something like $(\log)^{(1+\epsilon)}$ if possible.

Proof : Suffices to show that for suboptimal i,

$$\liminf_{n \to \infty} \frac{\mathbb{E}_\nu[T_i(n)]}{\log n} \geq \frac{1}{d_i}$$

where $d_i = d_{inf}(\nu_i, \mu^*, \mathcal{E})$

$\exists \, \nu_i' \in \mathcal{E}$ s.t. $\mu(\nu_i') > \mu^*$, $D(\nu_i, \nu_i') \leq d_i + \epsilon$

The above follows from the definition of dinf

Define alternative instance $\nu' = (\nu_1, \nu_2, \dots, \nu_i', \dots, \nu_K)$ such that the i^{th} arm is changed and ν_i' is as defined above.

We now define A s.t. it is a bad event in one instance & A^c is bad in the other

$$A = \{T_i(n) > \tfrac{n}{2}\}$$

$$R_n(\pi, \nu) \geq \mathbb{P}_\nu(A) \cdot \frac{n\Delta_i}{2}$$

In the alternative instance, $R_n(\pi, \nu') \geq \mathbb{P}_{\nu'}(A^c) \frac{n}{2}(\mu_i' - \mu^*)$

$$\therefore R_n(\pi, \nu) + R_n(\pi, \nu') \geq c \cdot n [\mathbb{P}_\nu(A) + \mathbb{P}_{\nu'}(A^c)]$$

where $c = \min\left(\frac{\Delta_i}{2}, \frac{\mu_i' - \mu^*}{2}\right)$.

Using the BH inequality,

$$R_n(\pi, \nu) + R_n(\pi, \nu') \geq \frac{cn}{2} \exp(-D(\mathbb{P}_\nu, \mathbb{P}_{\nu'}))$$

$$D(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_j \mathbb{E}_\nu[T_j(n)] D(\nu_j, \nu_j') = \mathbb{E}_\nu[T_i(n)] D(\nu_i, \nu_i')$$

$$D(\nu_i, \nu_i') \leq d_i + \epsilon$$

$$\therefore \quad R_n(\pi, \nu) + R_n(\pi, \nu') \geq c' n \exp\left(-(d_i+\epsilon) E_{\nu}[T_i(n)]\right)$$

where $c' = c/2$

$$\frac{\log\left[R_n(\pi, \nu) + R_n(\pi, \nu')\right]}{\log n} \geq \frac{\log c'}{\log n} + \frac{1}{\log n} - \frac{(d_i+\epsilon) E_{\nu}[T_i(n)]}{\log n} \quad \text{--- (1)}$$

Now we know that $R_n(\pi, \nu) = o(n^a) \ \forall a > 0$
Hence $\log[R_n(\pi, \nu)] = o(\log n^{a}) \ ~~\forall a>0~~$ } EXERCISE

This is not obvious, because, for example $f = o(n) \not\Rightarrow \log f = o(\log n)$ when $f = \sqrt{n}$

Taking liminf of both sides in (1),

$$\liminf_{n \to \infty} \frac{E_{\nu}[T_i(n)]}{\log n} \geq \frac{1}{d_i + \epsilon}$$

Since this holds $\forall \epsilon > 0$, $\liminf_{n \to \infty} \frac{E_{\nu}[T_i(n)]}{\log n} \geq \frac{1}{d_i}$

<u>Note</u>: d_i tries to capture how easy it is to confuse this arm with the optimal arm. If d_i is low, the possibility of confusion is high.

<u>Note</u>

"LEARNABILITY": $d_i > 0$

We can have $d_i = 0$ if

- $\mathcal{E}$ = set of all bounded distributions (don't know $[a, b]$ apriori)
- $\mathcal{E}$ = set of all subGaussian variable (σ is not known)
  Here, we cannot estimate σ^2 along the way and still achieve logarithmic regret

This concept is known as "Statistical Robustness of the Environment"

BEST ARM IDENTIFICATION (PURE EXPLORATION)

In this paradigm, the learning is NOT minimization of regret, BUT identifying the best arm.

1) Fixed budget : The number of pulls is fixed. The algorithm outputs the arm it thinks is the best.

2) Fixed confidence : First fix the accuracy. The algorithm can make the mistake w.p. $\leq \delta$. Here we evaluate the algorithm based on how ~~it~~ long it took for identification

These two are duals but <u>not</u> interconvertible

It turns out that the theory is easier for the fixed confidence setting

FIXED BUDGET SETTING

Given a budget of T pulls, where T is the horizon.
The algorithm gives an output $\hat{i} \in \{1, 2, \dots, k\}$. Single arm $\hat{i}$, not a distribution
We want to minimize $\mathbb{P}(\hat{i} \neq i^*)$.

<u>Model</u>: ① K arms
② All 1-sub Gaussian
③ Horizon T
④ There is a unique best arm i^*

Note that assumption ④ is quite an important assumption in fixed ~~budget~~ confidence settings. It is just written for convenience here.

<u>Note</u>: We can also design a metric which penalizes instead of $\mathbb{P}(\hat{i} \neq i^*)$ by the amount of suboptimality of the selected arm.

UNIFORM EXPLORATION ALGORITHM

- Pull each arm $\lfloor \frac{T}{k} \rfloor$ times.
- Output $\underset{i}{\operatorname{argmax}} \, \hat{\mu}_i$

We shall try to bound the probability of an error of UEA.
Assume that arm 1 is the optimal arm.

$$\mathbb{P}(\hat{i} \neq 1) = \sum_{j \neq 1} \mathbb{P}(\hat{i} = j)$$

$$\therefore \mathbb{P}(\hat{i} \neq 1) \leq \sum_{j \neq 1} \mathbb{P}(\hat{\mu}_j \geq \hat{\mu}_1)$$

The inequality is because the condition is necessary but not sufficient.

$$\therefore \mathbb{P}(\hat{i} \neq 1) \leq \sum_{j \neq 1} \mathbb{P}\left(|\hat{\mu}_j - \mu_j| \geq \frac{\Delta_i}{2}\right) + \mathbb{P}\left(|\hat{\mu}_i - \mu_i| \geq \frac{\Delta_i}{2}\right)$$

$$\leq 2(k-1)\exp\left(-\frac{\lfloor T/k \rfloor \Delta_2^2}{8}\right)$$

WLOG $\Delta_2 = \min(\Delta_i : \Delta_i \neq 0)$. Δ_2 is the term with the smallest decay rate

The probability of error decays exponentially with respect to the horizon.
We hence care about the decay rate, which should be as large as possible.
Here, the decay rate is $\frac{\Delta_2^2}{8k} = \frac{(\min \Delta_i)^2}{8k}$

Let us call the decay rate as α.

$$\boxed{\therefore\ \alpha(UEA) = \frac{\Delta_2^{\,2}}{8K}}$$

The simplest algorithm can be considered as a default baseline & now we want algorithm which have higher baselines.

<u>Note</u>: It turns out that superexponential decay is not possible. For example, Something like $\exp(-\alpha T^2)$ is not possible.