Week 8 - Lecture 38

Minimax Lower bounds

Till now have looked at different Bandit settings, discussed algorithms and their bounds.

But to check whether these algorithms are optimal, need to check the bounds for the algorithms with the lower bound that can be obtained in a particular bandit setting.

$$\sup_{\nu \in \mathcal{E}} R_T(\overset{\text{learner chooses}}{\pi}, \overset{\text{environment chooses}}{\nu})$$

worst possible regret for policy π for all possible environments.

Our goal is to minimize this regret.

$$\therefore \inf_{\pi \in \Pi} \left[\sup_{\nu \in \mathcal{E}} R_T(\pi, \nu) \right] = R_T^*(\mathcal{E})$$

This is problem independent

Thus called as <u>minimax criteria</u>.

A policy π is called minimax optimal for $\mathcal{E}$ if $R^*_T(\mathcal{E}) = R_T(\pi, \mathcal{E})$

Theorem: Let $\mathcal{E}^K$ be the K-armed Stochastic Bandits with finite support (σ-subgaussian).

Then $\exists$ a universal constant $c > 0$ s.t.

$\forall\ K > 1,\ T > K$, it holds that

$$R^*_T(\mathcal{E}^K) \geq c\sqrt{KT}$$

Proof Idea :

Let $X_1, X_2 \dots X_n$ be an iid sequence of r.v.s. with mean μ & variance 1.

Assume that $\mu = 0$ or Δ

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

To find whether $\mu = 0$ or Δ, will estimate $\hat{\mu}$ and set a threshold like $\Delta/2$, if $\hat{\mu}$ is closer to 0, then mean $= 0$ otherwise Δ.

But there could be error in the estimate.

Let $\mu = 0$, (If $\mu = \Delta$, similar argument follows by symmetry)

Then $\mathbb{E}[\hat{\mu}] = 0$ $\left[\begin{array}{l}\hat{\mu} \text{ is } \therefore \text{ a gaussian r.v.} \\ \text{with mean} = 0 \text{ \& var} = 1/n\end{array}\right]$

$\mathrm{Var}[\hat{\mu}] = 1/n$

Decision criteria : whether $\hat{\mu} \geq \Delta/2$

$$\Pr(\hat{\mu} - 0 \geq \Delta/2) \leq \frac{1}{\sqrt{n\Delta^2} + \sqrt{n\Delta^2 + 8/\pi}} \sqrt{\frac{2}{\pi}} \exp(-n\Delta^2/8)$$

$$\Pr(\hat{\mu} - 0 \geq \Delta/2) \geq \frac{1}{\sqrt{n\Delta^2} + \sqrt{n\Delta^2 + 4}} \sqrt{\frac{2}{\pi}} \exp(-n\Delta^2/8)$$

(Tail probabilities, derived from the fact that $\hat{\mu}$ has $\mathbb{E}[\hat{\mu}] = 0$ & $\mathrm{Var}[\hat{\mu}] = 1/n$)

If for a fixed Δ, n is not very large then the tail probabilities can be large $\therefore$ high chance for $\hat{\mu}$ to be wrong.

↑

Even if n is large but Δ is small then similar case can occur.

Select two bandit instances such that :

a) Optimal action in the two instances are different

b) The instance are 'close enough' that the policy cannot statistically differentiate between them.

Then the policy might confuse between the two and ends up playing sub-optimal action.

Let one problem instance be:

$$\nu = (P_i)_{i=1}^{k} \quad \text{where } P_i = \mathcal{N}(\mu_i, 1)$$

and another problem instance:

$$\nu' = (P_i')_{i=1}^{k} \quad \text{where } P_i' = \mathcal{N}(\mu_i', 1)$$