Lec 6

Page 1)

CS 59000_Reinforcement Learning (RL)

- Finite_Armed Stochastic Bandit

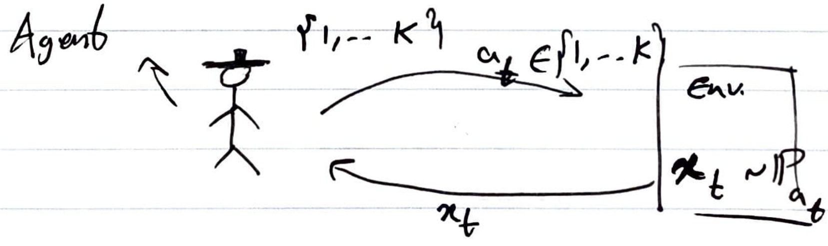
Agenda:

- Problem settup
- Regret definition
- Explore - then - Commit.

Finite_Armed Stachastic Bandit:

A K-armed stachastic bandit is a tuple of probability measures $\nu = (\mathbb{P}_1, \mathbb{P}_2, \ldots \mathbb{P}_K)$ where $\mathbb{P}_i$ is a probability measure over reals. for each $i \in \underbrace{[K]}_{\{1, \ldots K\}}$

The ~~ag~~ agent/learner and environment interact sequentially with each others.



In each round t, the agent chooses an action $A_t \in [K]$ which is fed to the environment. Then, the environment samples an reward $X_t \in \mathbb{R}$ from $\mathbb{P}_{A_t}$ and reveals it to the agent.

2)

The agent does not know ν in advance and is intrested in the best arms, the arms with highest $\mu_i(\nu) = \int_{-\infty}^{\infty} x \, dP_i(x)$ in other words, arms with highest expected rewards.

In the following, for notation simplicity, we assume one of the best arms with expected reward μ^* is the first arm.

Lets imagine our ~~a~~ agent interacts with the environment ν for n timesteps. If the agent knew ν in advane, it could go after the best arm and receive an expected return of

$n\mu^*$ ☺

But, the agent does not have this knowledge. This is the knowledge that $\boxed{\text{oracle}}$ has.
→ what is orcale's expected return?

↳ $n\mu^*$

Since the agent does not know the ν, it needs to learn its property in order to thrive. (What does it mean?)

3)

The agent makes decision A_1 at time 1, and observes X_1. Then, the agent uses (A_1, X_1) to decide on A_2. The agent receive X_2 after committing to A_2 and based on (A_1, X_1, A_2, X_2) decides on $A_3, \dots$

Let's denot the agent's strategy, Π

It means. $A_t \sim \pi(A_1, \dots X_{t-1})$

In other words, A_t is $\mathcal{F}_t$-measurable with respect to proper definition of $\mathcal{F}_t$.

$$A_t \sim \Pi(\mathcal{F}_{t-1})$$

---

what is our agent's expected return?

$$E\left[\sum_{t=1}^{n} X_t\right]$$

Def. [Regret]: How much more agent could have earned if it knew the environment in advance.

$$R_n(\Pi, \nu) = \underbrace{n\mu^*(\nu)}_{\substack{\text{How much} \\ \text{we could earn} \\ \text{(oracle)}}} - \underbrace{E\left[\sum_{t=1} X_t\right]}_{\substack{\text{How much} \\ \text{we actually} \\ \text{earned}}}$$

4) - $R_n(\pi, \nu) \geq 0$ for all π.

- $R_n(\pi, \nu) = 0 \rightarrow$ if π choose $A_t \in \operatorname{argmax}_i \mu_i$

$$A_t \in \underset{i \in [k]}{\operatorname{argmax}}\ \mu_i$$

$\rightarrow$ we are interested in strategies that have small regret.

$$- \begin{cases} \text{random regret} = n\mu^*(\nu) - \sum X_t \\ \text{Pseudo regret} = n\mu^*(\nu) - \sum \mu_{A_t} \end{cases}$$

$$E(X_t \mid \mathcal{F}_{t-1} \cup A_t)$$

Let $\Delta_i(\nu) = \mu^*(\nu) - \mu_i(\nu)$ be suboptimality gap of action i in environment ν.

Let, at time t, $T_i(t) = \sum_{s=1}^{t} \mathbb{I}\{A_s = i\}$

be a counting random variable for action/arm $i \in [K]$.

5)

Theorem: For any policy π, a k-armed stochastic bandit ν, and horizon $n \in \mathbb{N}$, the regret of π in ν is equal to

$$R_n(\pi,\nu) = \sum_{i=1}^{K} \underbrace{\Delta_i}_{\text{how bad i'th arm is}} \underbrace{E[T_i(n)]}_{\text{How many times i-th arm has pulled}}$$

proof. $S_n = \sum_{t=1}^{n} X_t$ → return random variable.

For round t, we have $\sum_{i=1}^{K} \mathbb{I}\{A_t = i\} = 1$

$$S_n = \sum_{t=1}^{n} \sum_{i=1}^{K} X_t \, \mathbb{I}\{A_t = i\}$$

$$R_n(\pi,\nu) = \underbrace{n\mu^*}_{\sum_{t=1}^{n}\sum_{i=1}^{K} \mu^* \mathbb{I}\{A_t=i\}} - E[S_n]$$

$$\Rightarrow R_n(\pi,\nu) = \sum_{t=1}^{n} \sum_{i=1}^{K} E\left[(\mu^* - X_t)\, \mathbb{I}\{A_t = i\}\right]$$

$$= \sum_{t=1}^{n} \sum_{i=1}^{K} E\left[E\left[(\mu^* - X_t)\, \mathbb{I}\{A_t = i\} \mid A_t\right]\right]$$

$$= \sum_{t=1}^{n} \sum_{i=1}^{K} E\left[\mathbb{I}\{A_t = i\}\, \underbrace{E\left[\mu^* - X_t \mid A_t\right]}_{\Delta_{A_t}}\right] \quad \sigma(A_t)$$

$$= \sum_{t=1}^{n} \sum_{i=1}^{K} E\left[\mathbb{I}\{A_t = i\}\, \Delta_{A_t}\right]$$

$$= \sum_{i=1}^{K} \sum_{t=1}^{n} E\left[\mathbb{I}\{A_t = i\}\, \Delta_{A_t}\right]$$

b)

$$\to R_n(\pi, \nu) = \sum_{i=1}^{K} \sum_{t=1}^{n} E\left[I\{A_t = i\} \underset{\Delta_i}{\underset{\downarrow}{\Delta_{A_t}}} \right]$$

$$= \sum_{i=1}^{K} \sum_{t=1}^{n} E\left[I\{A_t = i\} \Delta_i \right]$$

$$= \sum_{i=1}^{K} \Delta_i \sum E\left(I\{A_t = i\} \right)$$

$$= \sum_{i=1}^{K} \Delta_i \underbrace{E\left[\sum_{t=1}^{n} I\{A_t = i\} \right]}$$

$$= \sum_{i=1}^{K} \Delta_i E\left[T_i(n) \right]$$