CS 747, Autumn 2020: Week 2, Q&A

Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

Autumn 2020

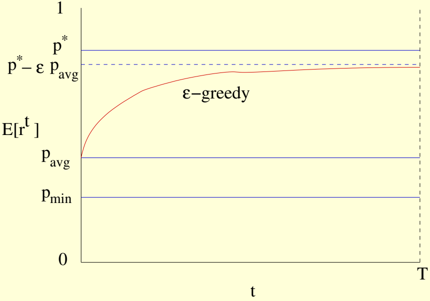
Erratum

- Typo on slide 4 in Week 2 Lecture 1.

Erratum

- Typo on slide 4 in Week 2 Lecture 1.

  ϵ-greedy: On each step explore (uniform sampling) w.p. ϵ, exploit w.p. $1 - \epsilon$.

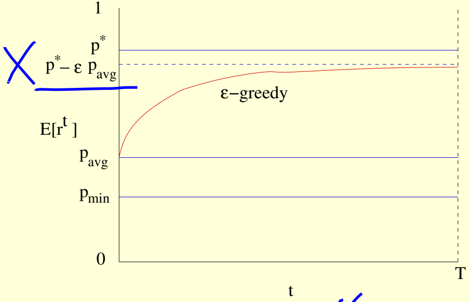


  $\mathbb{E}[r^t]$ can never exceed $p^\star - \epsilon p_{avg}$!

Erratum

- Typo on slide 4 in Week 2 Lecture 1.

ϵ-greedy: On each step explore (uniform sampling) w.p. ϵ, exploit w.p. $1 - \epsilon$.



$\mathbb{E}[r^t]$ can never exceed $p^\star - \epsilon p_{\text{avg}}$!

Must be $p^*(1-\varepsilon) + \varepsilon p_{avg} = p^* - \varepsilon(p^* - p_{avg})$.

Question 1

- What is the formal definition of “exploit” and *exploit*(T)?

Question 1

- What is the formal definition of "exploit" and $exploit(T)$?

$$exploit^t = \begin{cases} 1 & \text{if for all } a \in A : \hat{p}^t_{a^t} \geq \hat{p}^t_a, \\ 0 & \text{otherwise.} \end{cases}$$

$$exploit(T) = \sum_{t=0}^{T-1} exploit^t.$$

Question 2

- Can we **prove** that the ϵ_t-greedy strategy with $\epsilon_t = \frac{1}{(t+1)^2}$ can incur linear regret?

Question 2

- Can we **prove** that the ϵ_t-greedy strategy with $\epsilon_t = \frac{1}{(t+1)^2}$ can incur linear regret?

- Consider 2-armed bandit with arms 1 and 2.
  Means p_1, p_2, respectively, satisfying $1 > p_1 > p_2 > 0$.

Question 2

- Can we **prove** that the ϵ_t-greedy strategy with $\epsilon_t = \frac{1}{(t+1)^2}$ can incur linear regret?

- Consider 2-armed bandit with arms 1 and 2.
  Means p_1, p_2, respectively, satisfying $1 > p_1 > p_2 > 0$.
- Let E_1 be the event that $\hat{p}_1^{100} = 0$ and $\hat{p}_2^{100} > 0$.
  Let E_2 be the event that for all $t \geq 100$, *exploit*$^t = 1$.
  $E_1 \wedge E_2 \implies$ Arm 1 is pulled no more than 100 times.

Question 2

- Can we **prove** that the ϵ_t-greedy strategy with $\epsilon_t = \frac{1}{(t+1)^2}$ can incur linear regret?

- Consider 2-armed bandit with arms 1 and 2.
  Means p_1, p_2, respectively, satisfying $1 > p_1 > p_2 > 0$.
- Let E_1 be the event that $\hat{p}_1^{100} = 0$ and $\hat{p}_2^{100} > 0$.
  Let E_2 be the event that for all $t \geq 100$, $exploit^t = 1$.
  $E_1 \wedge E_2 \implies$ Arm 1 is pulled no more than 100 times.
- $\mathbb{P}\{E_1 \wedge E_2\} = \mathbb{P}\{E_1\} \cdot \mathbb{P}\{E_2|E_1\} = \mathbb{P}\{E_1\} \cdot \mathbb{P}\{E_2\}$.

Question 2

- Can we **prove** that the ϵ_t-greedy strategy with $\epsilon_t = \frac{1}{(t+1)^2}$ can incur linear regret?

- Consider 2-armed bandit with arms 1 and 2.
  Means p_1, p_2, respectively, satisfying $1 > p_1 > p_2 > 0$.
- Let E_1 be the event that $\hat{p}_1^{100} = 0$ and $\hat{p}_2^{100} > 0$.
  Let E_2 be the event that for all $t \geq 100$, $exploit^t = 1$.
  $E_1 \wedge E_2 \implies$ Arm 1 is pulled no more than 100 times.
- $\mathbb{P}\{E_1 \wedge E_2\} = \mathbb{P}\{E_1\} \cdot \mathbb{P}\{E_2 | E_1\} = \mathbb{P}\{E_1\} \cdot \mathbb{P}\{E_2\}$.
  $\mathbb{P}\{E_1\} = C_1 > 0$.

Question 2

- Can we **prove** that the ϵ_t-greedy strategy with $\epsilon_t = \frac{1}{(t+1)^2}$ can incur linear regret?

- Consider 2-armed bandit with arms 1 and 2.
  Means p_1, p_2, respectively, satisfying $1 > p_1 > p_2 > 0$.
- Let E_1 be the event that $\hat{p}_1^{100} = 0$ and $\hat{p}_2^{100} > 0$.
  Let E_2 be the event that for all $t \geq 100$, $exploit^t = 1$.
  $E_1 \wedge E_2 \implies$ Arm 1 is pulled no more than 100 times.
- $\mathbb{P}\{E_1 \wedge E_2\} = \mathbb{P}\{E_1\} \cdot \mathbb{P}\{E_2 | E_1\} = \mathbb{P}\{E_1\} \cdot \mathbb{P}\{E_2\}$.
  $\mathbb{P}\{E_1\} = C_1 > 0$.
  $\mathbb{P}\{E_2\} = 1 - \mathbb{P}\{\exists t \geq 100, \neg exploit^t\} \geq 1 - \sum_{t=100}^{\infty} \mathbb{P}\{\neg exploit^t\}$

$$= 1 - \sum_{t=100}^{\infty} \frac{1}{2(t+1)^2} = C_2 > 0.$$

Question 2

- Can we **prove** that the ϵ_t-greedy strategy with $\epsilon_t = \frac{1}{(t+1)^2}$ can incur linear regret?

- Consider 2-armed bandit with arms 1 and 2.
  Means p_1, p_2, respectively, satisfying $1 > p_1 > p_2 > 0$.
- Let E_1 be the event that $\hat{p}_1^{100} = 0$ and $\hat{p}_2^{100} > 0$.
  Let E_2 be the event that for all $t \geq 100$, $exploit^t = 1$.
  $E_1 \wedge E_2 \implies$ Arm 1 is pulled no more than 100 times.
- $\mathbb{P}\{E_1 \wedge E_2\} = \mathbb{P}\{E_1\} \cdot \mathbb{P}\{E_2 | E_1\} = \mathbb{P}\{E_1\} \cdot \mathbb{P}\{E_2\}$.
  $\mathbb{P}\{E_1\} = C_1 > 0$.
  $$\mathbb{P}\{E_2\} = 1 - \mathbb{P}\{\exists t \geq 100, \neg exploit^t\} \geq 1 - \sum_{t=100}^{\infty} \mathbb{P}\{\neg exploit^t\}$$
  $$= 1 - \sum_{t=100}^{\infty} \frac{1}{2(t+1)^2} = C_2 > 0.$$
- $\mathbb{P}\{E_1 \wedge E_2\} > C_1 C_2 = C_3 > 0 \implies$ linear regret.

Question 3

- Does Lai and Robbins' lower bound need the "if" condition?

If sub-polynomial regret on all instances,
then super-logarithmic regret on all instances.

Question 3

- Does Lai and Robbins' lower bound need the "if" condition?

If sub-polynomial regret on all instances,
then super-logarithmic regret on all instances.

Is it true that every algorithm has super-logarithmic regret on all instances?

Question 3

- Does Lai and Robbins' lower bound need the "if" condition?

If sub-polynomial regret on all instances,
then super-logarithmic regret on all instances.

Is it true that every algorithm has super-logarithmic regret on all instances?

No! What about that algorithm that always pulls arm 3?

Question 4

- Why did we exclude bandit instances with optimal mean reward of 1 in our result that GLIE $\iff$ sub-linear regret?

- Recall $\bar{\mathcal{I}} = [0, 1)^n$. We showed that GLIE($L$, I) $\forall I \in \bar{\mathcal{I}} \iff$ subLinearRegret(L, I) $\forall I \in \bar{\mathcal{I}}$.

Question 4

- Why did we exclude bandit instances with optimal mean reward of 1 in our result that GLIE $\iff$ sub-linear regret?

- Recall $\bar{\mathcal{I}} = [0, 1)^n$. We showed that GLIE($L$, I) $\forall I \in \bar{\mathcal{I}} \iff$ subLinearRegret(L, I) $\forall I \in \bar{\mathcal{I}}$.
- Let $\mathcal{I} = [0, 1]^n$.

Question 4

- Why did we exclude bandit instances with optimal mean reward of 1 in our result that GLIE $\iff$ sub-linear regret?

- Recall $\bar{\mathcal{I}} = [0, 1)^n$. We showed that GLIE($L$, I) $\forall I \in \bar{\mathcal{I}} \iff$ subLinearRegret(L, I) $\forall I \in \bar{\mathcal{I}}$.
- Let $\mathcal{I} = [0, 1]^n$.
- It is **true** that GLIE(L, I) $\forall I \in \mathcal{I} \implies$ subLinearRegret(L, I) $\forall I \in \mathcal{I}$.

Question 4

- Why did we exclude bandit instances with optimal mean reward of 1 in our result that GLIE $\iff$ sub-linear regret?

- Recall $\bar{\mathcal{I}} = [0, 1)^n$. We showed that GLIE($L$, I) $\forall I \in \bar{\mathcal{I}} \iff$ subLinearRegret(L, I) $\forall I \in \bar{\mathcal{I}}$.
- Let $\mathcal{I} = [0, 1]^n$.
- It is **true** that GLIE(L, I) $\forall I \in \mathcal{I} \implies$ subLinearRegret(L, I) $\forall I \in \mathcal{I}$.
- It is **not true** that subLinearRegret(L, I) $\forall I \in \mathcal{I} \implies$ GLIE(L, I) $\forall I \in \mathcal{I}$.

Question 4

- Why did we exclude bandit instances with optimal mean reward of 1 in our result that GLIE $\iff$ sub-linear regret?

- Recall $\bar{\mathcal{I}} = [0,1)^n$. We showed that GLIE($L$, I) $\forall I \in \bar{\mathcal{I}} \iff$ subLinearRegret(L, I) $\forall I \in \bar{\mathcal{I}}$.
- Let $\mathcal{I} = [0,1]^n$.
- It is **true** that GLIE(L, I) $\forall I \in \mathcal{I} \implies$ subLinearRegret(L, I) $\forall I \in \mathcal{I}$.
- It is **not true** that subLinearRegret(L, I) $\forall I \in \mathcal{I} \implies$ GLIE(L, I) $\forall I \in \mathcal{I}$. Why? Consider L such that

$$L(h^t) = a^t = \begin{cases} a^{t-1} & \text{if } r^{t-1} = 1, \\ UCB(h^{t-1}) & \text{otherwise.} \end{cases}$$

Question 5

- Here's the formula given for ucb-kl$_a^t$:

$$\text{ucb-kl}_a^t = \max(S) \text{ where}$$
$$S = \{q \in [\hat{p}_a^t, 1] \text{ s.t. } u_a^t KL(\hat{p}_a^t, q) \leq \ln(t) + c\ln(\ln(t))\},$$

where $c \geq 3$.

Can we instead write

$$\text{ucb-kl}_a^t = \text{ the solution } q \in [\hat{p}_a^t, 1] \text{ of}$$
$$u_a^t KL(\hat{p}_a^t, q) = \ln(t) + c\ln(\ln(t))?$$

Question 5

- Here's the formula given for ucb-kl_a^t:

$$\text{ucb-kl}_a^t = \max(S) \text{ where}$$

$$S = \{q \in [\hat{p}_a^t, 1] \text{ s.t. } u_a^t KL(\hat{p}_a^t, q) \leq \ln(t) + c\ln(\ln(t))\},$$

where $c \geq 3$.

Can we instead write

$$\text{ucb-kl}_a^t = \text{ the solution } q \in [\hat{p}_a^t, 1] \text{ of}$$

$$u_a^t KL(\hat{p}_a^t, q) = \ln(t) + c\ln(\ln(t))?$$

Yes!