CS 747, Autumn 2020: Week 1, Q&A

Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

Autumn 2020

Question 1

- Why did we define algorithms as mappings from the set of histories to the set of (probability distributions over) actions? What if decisions had to take other details (apart from history) into consideration?

Question 1

- Why did we define algorithms as mappings from the set of histories to the set of (probability distributions over) actions? What if decisions had to take other details (apart from history) into consideration?

Sufficient to:

achieve optimal regret,

describe many existing algorithms.

Question 2

- What is the number of histories of length T?

Question 2

- What is the number of histories of length T?

Say algorithm assigns "a" arms non-zero probability for each history, and rewards can take "b" possible values.

Answer: $(ab)^T$.

Question 3

- In our definitions of algorithms and regret, why does the index stop at $T-1$ (rather than T)?

  Here is what an algorithm does—

  For $t = 0, 1, 2, \ldots, T-1$:

  - Given the history $h^t = (a^0, r^0, a^1, r^1, a^2, r^2, \ldots, a^{t-1}, r^{t-1})$,
  - Pick an arm a^t to sample (or "pull"), and
  - Obtain a reward r^t drawn from the distribution corresponding to arm a^t.

  $R_T = Tp^* - \sum_{t=0}^{T-1} \mathbb{E}[r^t]$.

Question 3

- In our definitions of algorithms and regret, why does the index stop at $T-1$ (rather than T)?

  Here is what an algorithm does—

  For $t = 0, 1, 2, \ldots, T-1$:

  - Given the history $h^t = (a^0, r^0, a^1, r^1, a^2, r^2, \ldots, a^{t-1}, r^{t-1})$,
  - Pick an arm a^t to sample (or "pull"), and
  - Obtain a reward r^t drawn from the distribution corresponding to arm a^t.

  $R_T = Tp^* - \sum_{t=0}^{T-1} \mathbb{E}[r^t]$.

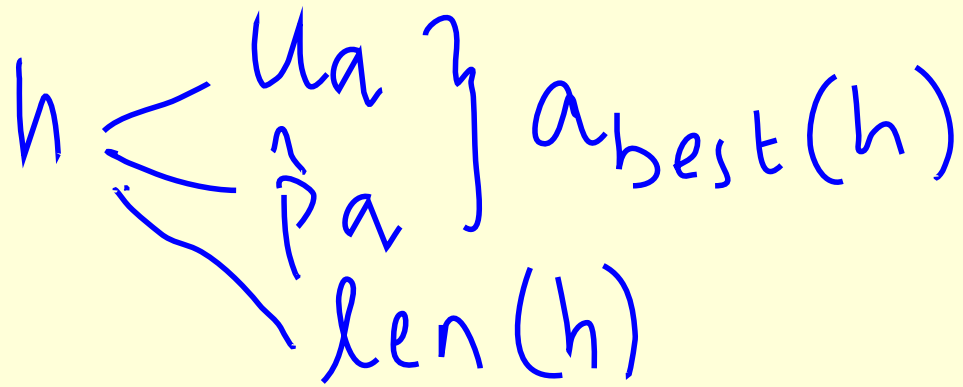
T pulls, rewards counted in horizon T.
Matter of convention.

Question 4

- How can ϵ-greedy algorithms be written in the form of $\mathbb{P}\{\text{arm}|\text{history}\}$?

Question 4

- How can ϵ-greedy algorithms be written in the form of $\mathbb{P}\{\text{arm}|\text{history}\}$?



Say 3 arms

$\epsilon/2$

if $len(h) \le \epsilon$ [1/3, 1/3, 1/3]

else 1 to $a_{best}(h)$, 0 to other arms.

$\epsilon/3$

$\left[\frac{\epsilon}{3}, 1-\epsilon+\frac{\epsilon}{3}, \frac{\epsilon}{3}\right]$

$a_{best}(h)$

Question 5

- Wouldn't a finite (but large) amount of exploration suffice?

Question 5

- Wouldn't a finite (but large) amount of exploration suffice?

In practice, maybe.

In theory, NO! Will have linear regret.

Question 6

- How is RL different from supervised learning?

Question 6

- How is RL different from supervised learning?