8/8/19

AI 5001: Intro to Modern AI

- Review & discussion ($A_{t+1} \rightarrow A_t$)
- Exploration vs exploitation: ε-greedy method
- Estimating action value: simple averaging, incremental, non stationarity
- $A_t = \underset{a}{\text{argmax}}\, Q_t(a)$: greedy
- ε-greedy method: $(1-\varepsilon)$ times for exploitation.
  ε times for exploration $(0 < \varepsilon < 1)$

→ HW problem to understand this better.

Estimating action value:

a: action

$q(a)$: true value for 'a'

$Q_t(a)$: estimated value at 'a' after 't' timesteps

Assume: stochastic nature of reward. (Stationary or rewards coming from a fixed distribution)

Recall: Our goal is to maximise expected value/reward

$$Q_t(a) = \frac{1}{N_t(a)} \cdot \left[R_1(a) + R_2(a) \cdots + R_{N_t(a)}(a) \right]$$

Note: this operation is expensive in terms of memory, so use incremental method

$$Q_t = \frac{1}{N_t} \sum_{i=1}^{N_t} R_i$$

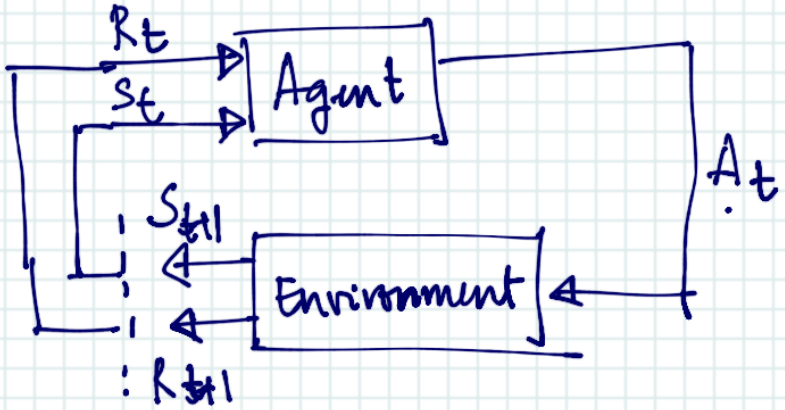
$$= \frac{1}{N_t} \left[R_{N_t} + \sum_{i=1}^{N_t - 1} R_i \right]$$

$$= \frac{R_{N_t}}{N_t} + \frac{N_t - 1}{N_t} \cdot Q$$

$$Q_t = Q_{t-1} + \frac{1}{N_t} \left[R_{N_t} - Q_{t-1} \right]$$

Nonstationary bandits: Weighted sum of rewards. We weight nearer time rewards (recent rewards) more that older rewards

Define the RL problem:



Policy: $\Pi_t(a|s)$