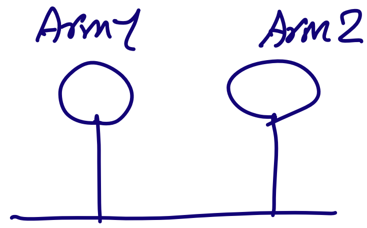
Lecture 2 - Jan 17, 2023

- Multi-Arm Bandits
  - Introduction
  - Exploration - Exploitation Delima
  - Epsilon - Greedy Policy
  - Optimistic Initial Values
  - Upper Confidence Bound Selection Policy
  - Gradient-Based Selection Policy
  - Thompson Sampling

HW1 is assigned → Due Jan 27

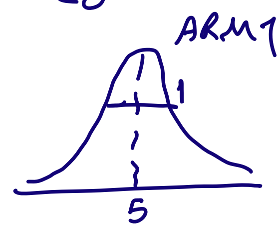
TA's first office hour: Friday, Jan 20, 12pm - 1pm

Multi-Arm Bandit

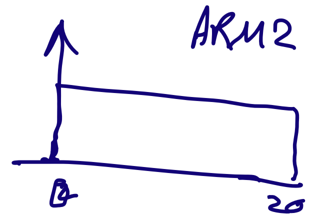


ARM1 = Reward $R^1 \,|\, a^1 \sim N(\mu = 5, \sigma = 1)$

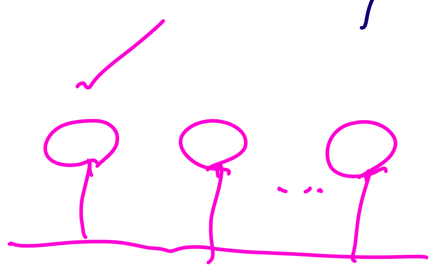
ARM2 = $R^2 \,|\, a^2 \sim \text{Uniform}[\underset{0}{a} \quad \underset{20}{b}]$



⇒ Goal: Maximizing the total reward



$Q^*(a) = E[R_t \,|\, a_t = a]$ $\quad a \in A = \{a^1, a^2, \dots, a^k\}$



Averaging Learning Rule:

$Q_2(a) = R_1$

$Q_3(a) = \frac{R_1 + R_2}{2}$

$$Q_n(a) = \frac{R_1 + R_2 + \dots + R_{n-1}}{n-1}$$

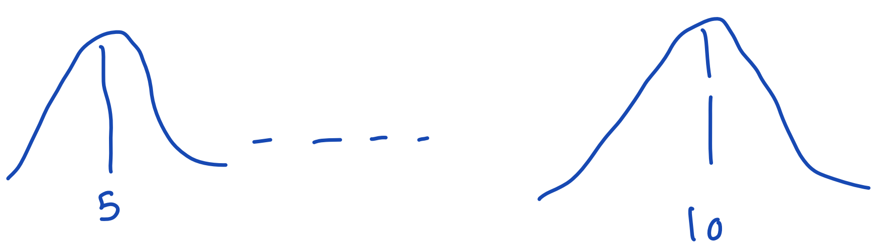
$$Q_{n+1}(a) = \frac{R_1 + R_2 + \dots + R_{n-1} + R_n}{n} = \underbrace{\frac{R_1 + R_2 + \dots + R_{n-1}}{n}}_{\times \frac{n-1}{n-1}} + \frac{R_n}{n} \times \frac{n-1}{n-1}$$

$$= Q_n(a)\frac{n-1}{n} + \frac{1}{n} R_n$$

$$Q_{n+1}(a) = \underbrace{Q_n(a)}_{\text{previous estimate}} + \frac{1}{n}\Big[\overbrace{R_n}^{\text{last reward}} - \underbrace{Q_n(a)}_{\text{Previous estimate}}\Big] \qquad (I)$$

$n=10 \longrightarrow \quad Q_n(a) + \frac{1}{10}[\bigcirc - Q_n(a)]$

$n=1000000 \longrightarrow \quad Q_n(a) + \frac{1}{1000,000}[\bigcirc - Q_n(a)]$



Nonstationarity

Learning Rate

$$Q(a) \leftarrow Q(a) + \alpha \left[R - Q(a) \right]$$

$$\sum_{t=1}^{\infty} \alpha_t = \infty \qquad \sum_{t=1}^{\infty} \alpha_t^2 < \infty \implies Q^*(a)$$

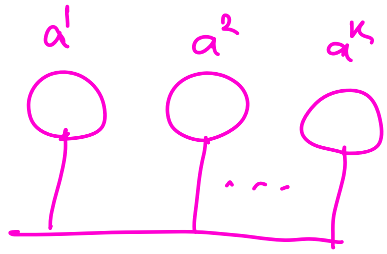
$\alpha_t = \frac{1}{t}$

$\alpha = 0.1$

$R = 100 \qquad Q(a) = 0$

Policy 1: Epsilon-Greedy Policy

ϵ-greedy

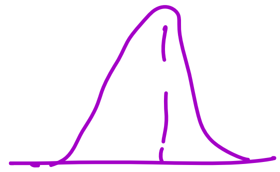


$Q(a^1) = Q(a^2) = \dots = Q(a^K) = 0$

$$a_t \sim \begin{cases} \text{Greedy. } \underset{a \in \{a^1, \dots, a^K\}}{\operatorname{argmax}} Q(a) & \text{w.p. } 1-\epsilon \Rightarrow \text{Exploitation} \\ \text{Random}\{a^1, a^2, \dots, a^K\} & \text{w.p. } \epsilon \Rightarrow \text{Exploration} \end{cases}$$

Example: $\varepsilon = 0.1$

$\alpha = 0.5$



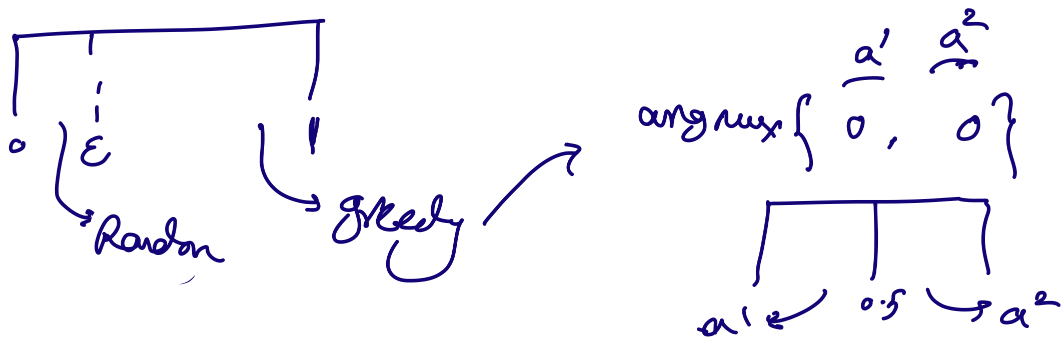
$R^1 \sim N(5, \sigma = 10)$

$R^2 \sim \text{Unif}[8, 12]$

$Q(a^1) = Q(a^2) = 0$

1

$$\pi^{\varepsilon\text{-greedy}} = \begin{cases} \underset{a \in \{a^1, a^2\}}{\text{argmax}}\, Q(a) & \text{w.p. } 0.9 \\ \text{Random}\{a^1, a^2\} & \text{w.p. } 0.1 \end{cases} \longrightarrow a^1 \rightsquigarrow$$



$R = 10 \sim N(5, 10)$

$$Q(a^1) = Q(a^1) + \alpha[R - Q(a^1)]$$

$$= 0 + 0.5[10 - 0] = 5$$

$$\boxed{Q(a^1)=5\ ,\ Q(a^2)=0}$$

(2)

$$\pi^{\varepsilon\text{-greedy}} = \begin{cases} \operatorname{argmax}\{\overset{a^1}{5},\ \overset{a^2}{0}\} = a^1 & \text{w.p. } 0.9 \\ \text{Random} & 0.1 \end{cases} \sim a^1$$

$R = 3 \leftarrow N(5, 10)$

$$Q(a^1) = \underline{Q(a^1)} + \alpha\,[R - Q(a^1)]$$
$$= 5 + 0.5\,[3 - 5] = 4$$

$$\boxed{Q(a^1)=4,\ Q(a^2)=0}$$

(3)

$$\pi^{\varepsilon\text{-greedy}} = \begin{cases} \operatorname{argmax} Q(a) = a^1 & 0.9 \\ \text{Random}\{a^1, a^2\} & 0.1 \end{cases} \sim a^2$$

Exploration

$R = 9 \sim \text{Uniform}\,[8, 12]$

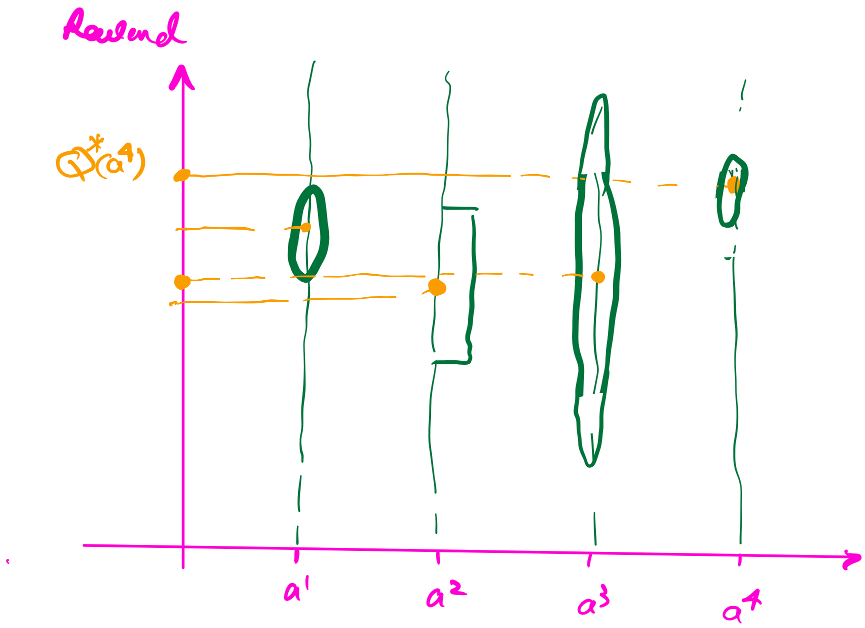
$$Q(a^2) = Q(a^2) + \alpha\left[\underline{R} - Q(a^2)\right]$$

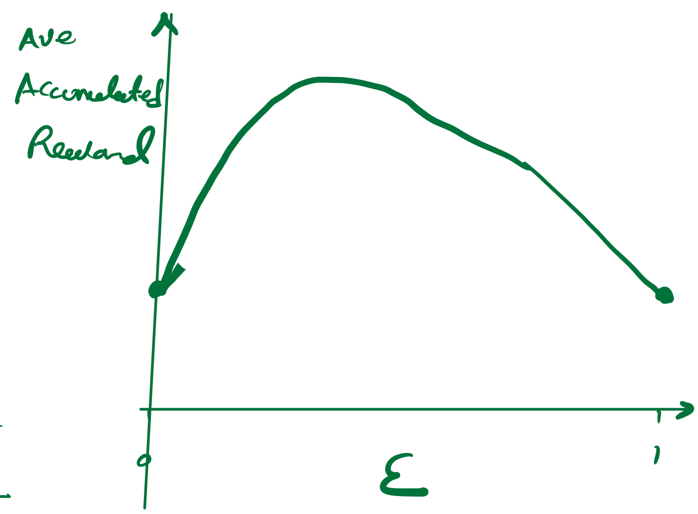
$$= 0 + 0.5\,[9 - 0] = 4.5$$

$$\boxed{Q(a^1) = 4, \quad Q(a^2) = 4.5}$$

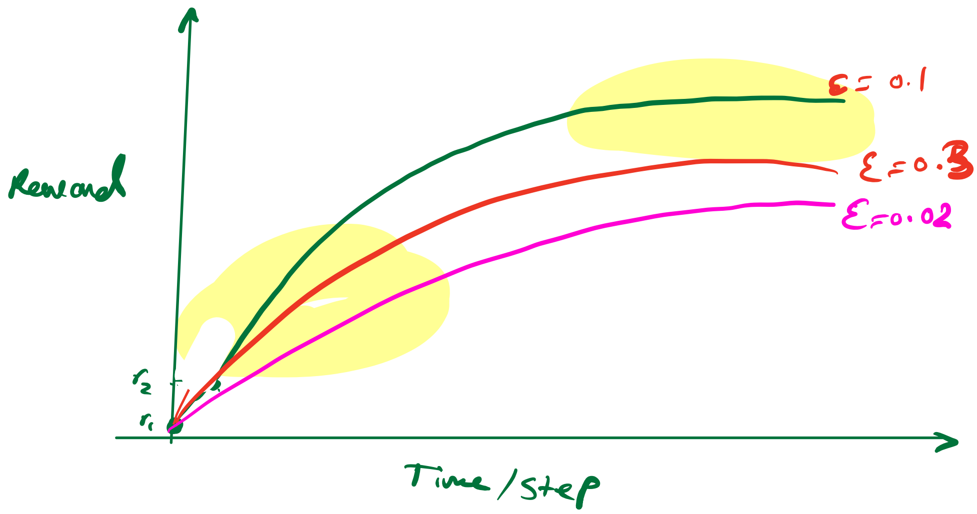
4

$$\pi^{\varepsilon\text{-}g} = \begin{cases} \text{argmax } Q(a) = a^2 & 0.9 \\ \text{Radm}\{a^1, a^2\} & 0.1 \end{cases}$$





$$a_t \sim \begin{cases} \text{Greedy: } \underset{a\in\{a^1,\dots,a^k\}}{\arg\max}\, Q(a) & \text{w.p. } 1-\varepsilon \Rightarrow \text{Exploitation} \\ \text{Random}\{a^1, a^2, \dots, a^k\} & \text{w.p. } \varepsilon \Rightarrow \text{Exploration} \end{cases}$$



episode 1

a^1 a^1

r_1 r_2