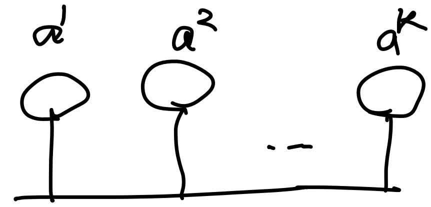
Lecture 3 - Jan 20, 2023

- 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 → Due Jan 27

Project 1 is Posted → Due Feb 7

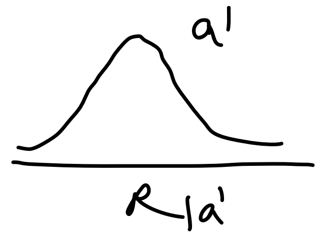
TA's first office hour: Friday, Jan 20, 12pm - 1pm overview:



$$Q^*(a) = E[R \mid a]$$

$R \mid a$ ← Random Variable

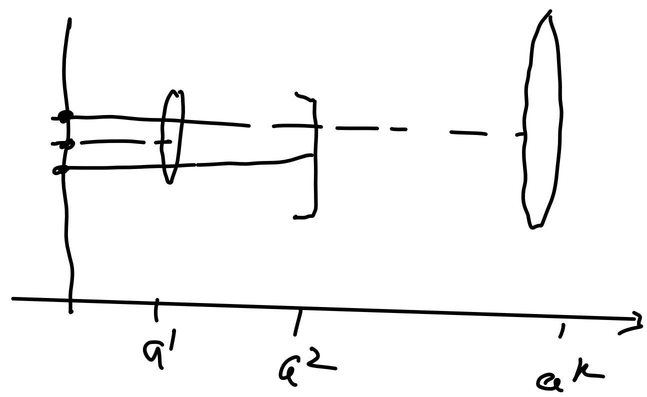
$$a^* = \underset{a \in A}{\operatorname{argmax}}\, Q^*(a)$$



Learning ← Distribution are unknown

→ Policy: $\underline{a}$

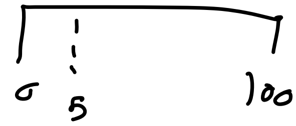
→ Estimation $Q(a)$



$Q(a) = 0$ for all $a \in A$

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

Policy 1: ε-greedy

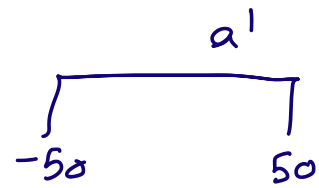


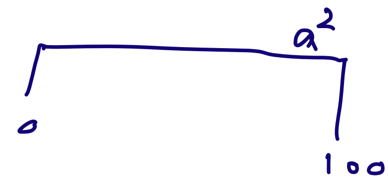
$$a \leftarrow \begin{cases} \underset{a \in A}{\operatorname{argmax}}\, Q(a) & 1-\varepsilon \quad \leftarrow \text{greedy} \\ \text{Random}\{a^1, \dots, a^k\} & \varepsilon \quad \leftarrow \text{Random} \end{cases}$$

$a^1 \longrightarrow Q(a^1) = 20, \quad Q(a^2) = 0$

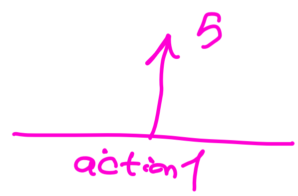
$\hookrightarrow a^2 \xrightarrow{R=90} Q(a^1) = 20, \quad Q(a^2) = 45$

40 ↩

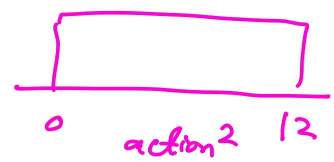




Optimistic Initial Value: $Q^*(a^1) = 5$, $Q^*(a^2) = 6$



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



Time 1

action a^1

Reward 5

$Q(a^1) = \underbrace{Q(a^1)}_{0} + \overset{0.5}{\alpha}\,[\underbrace{R}_{5} - \underbrace{Q(a^1)}_{0}] = 2.5$

$\pi^{\varepsilon\text{-}g} \rightarrow \begin{cases} \text{greedy} = a^1 & 1-\varepsilon \\ \text{Random}\{a^1, a^2\} & \varepsilon \end{cases}$

$\boxed{\begin{matrix} Q(a^1) = 2.5 \\ Q(a^2) = 0 \end{matrix}}$

$\varepsilon = 0.1 \rightarrow \begin{cases} a^1 \rightarrow 0.95 \\ a^2 \rightarrow 0.05 \end{cases}$

$Q(a^1) = Q(a^2) = 15$

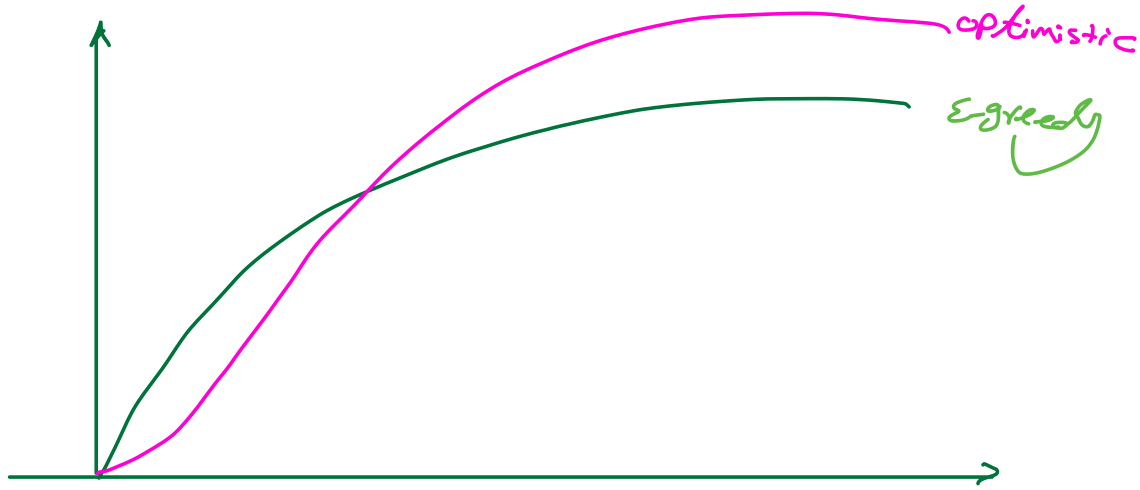
$Q(a^1) = \underbrace{Q(a^1)}_{15} + \underbrace{\alpha}_{0.5}\,[\underbrace{R}_{5} - \underbrace{Q(a^1)}_{15}] = 10$

$\pi^{\varepsilon\text{-}g} \rightarrow \begin{cases} \text{greedy} = a^2 \\ \text{Random}\{a^1, a^2\} \end{cases}$

$Q(a^1) = 10$

$Q(a^2) = 15$

$\varepsilon = 0.1 \begin{cases} a^1 \leftarrow 0.05 \\ a^2 \leftarrow 0.95 \end{cases}$



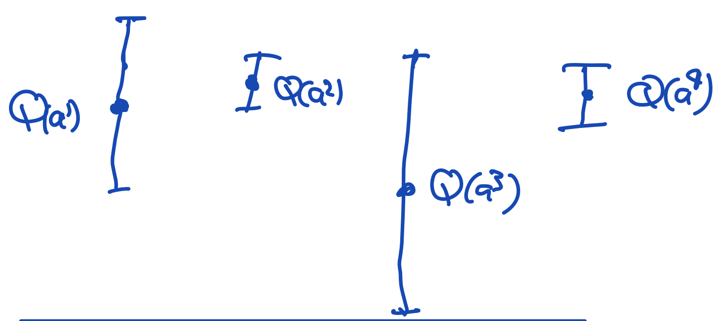
Upper Confidence Bound (UCB) ⇐ Policy

$$a_t = \underset{a\in\{a^1,a^2,\dots,a^k\}}{\arg\max} \left[\hat{Q}(a) + \underbrace{c\sqrt{\frac{\log t}{N_t(a)}}}_{\text{addition to the current estimate of } \hat{Q}(a)} \right]$$

t: time step

$N_t(a)$: # that action a is selected up to time t

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



$Q(a^1) = 10$ $\quad\quad$ $Q(a^2) = 1$

$N(a^1) = 100$ $\quad\quad$ $N(a^2) = 1$ $\rightsquigarrow$ $t = 101$

$t = 102 \rightarrow$ $a_t = \arg\max \left[Q(a) + c\sqrt{\frac{\log t}{N_t(a)}} \right]$

$$\underset{a \in \{a^1, a^2\}}{\arg\max} \left\{ \underbrace{10 + c\sqrt{\frac{\log 102}{100}}}_{a^1}, \underbrace{1 + c\sqrt{\frac{\log 102}{1}}}_{a^2} \right\}$$

$c = 1 \leftarrow$ $10 + 0.213$ $\quad\quad$ $1 + 1.52$

$c = 10 \leftarrow$ $10 + 2.13$ $\quad\quad$ $1 + 15.2$

$c = 0 \Longrightarrow$ greedy

c larger $\rightarrow$ More exploration

Online Learning Algorithms $\left\{ \begin{array}{l} \text{Exp3} \\ \text{Hedge} \\ \text{Regret Matching} \end{array} \right.$

Policy: Gradient-Bandit Policy

Directly Parametrizing the Policy:

$H_t(a)$: numeric preference for action a

$$P(a_t=a) = \frac{e^{H_t(a)}}{\sum_{i=1}^{k} e^{H_t(a^i)}} = \pi_t(a)$$

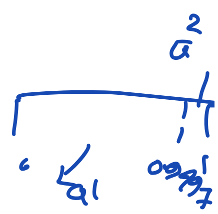
Probability of taking action a at time t

Boltzman or Softmax Policy

Gibbs distribution

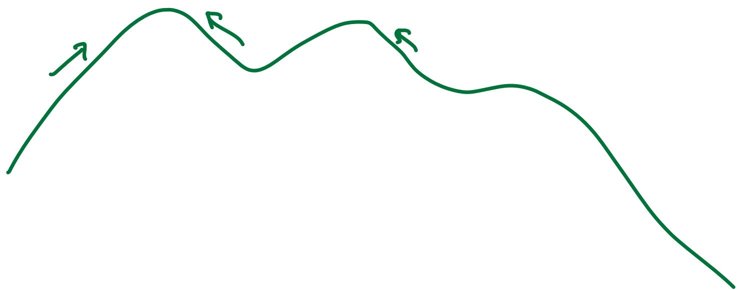
$$\left.\begin{array}{l} H_1(a^1)=0 \\ H_1(a^2)=0 \end{array}\right\} \rightarrow \left\{\begin{array}{l} \pi_1(a^1)= \dfrac{e^{\overset{=0}{H_1(a^1)}}}{e^{H_1(a^1)}+e^{H_1(a^2)}} = \dfrac{1}{1+1} = \dfrac{1}{2} \\ \pi_1(a^2)= \dfrac{e^0}{e^0+e^0} = \dfrac{1}{2} \end{array}\right.$$

$$\left.\begin{array}{l} H_5(a^1)=8 \\ H_5(a^2)=0 \end{array}\right\} \rightarrow \left\{\begin{array}{l} \pi_5(a^1)= \dfrac{e^8}{e^8+e^0} = 0.9997 \\ \pi_5(a^2)= \dfrac{e^0}{e^8+e^0} = 0.0003 \end{array}\right.$$



Average Reward

$$H_{t+1}(a) = H_t(a) + \alpha \frac{\partial E[R_t]}{\partial H_t(a)}$$



$$\begin{cases} H_{t+1}(A_t) = H_t(A_t) + \alpha (R_t - \bar{R}_t)(1 - \pi_t(A_t)) \\ H_{t+1}(a) = H_t(a) - \alpha (R_t - \bar{R}_t)\, \pi_t(a) \quad \text{for all } a \in A - A_t \end{cases}$$

A_t : action selected at time t

a : all actions except A_t

$\bar{R}_t$: Average Reward up to current time t including "t"

$\bar{R}_t = 0$, $R_t = 10$, $A_t = a^1$ $\quad A = \{a^1, a^2\}$

$\pi_t(a^1) = 0.6$

$\pi_t(a^2) = 0.4$

Action 1 $\rightarrow$ $\quad + \alpha (\underbrace{R_t}_{10} - \underbrace{\bar{R}_t}_{0}) (1 - \underbrace{\pi_t(A_t)}_{0.6})$

action 2 $\rightarrow$ $\quad - \alpha (\underbrace{R_t}_{10} - \underbrace{\bar{R}_t}_{0}) \underbrace{\pi_t(a)}_{0.4}$

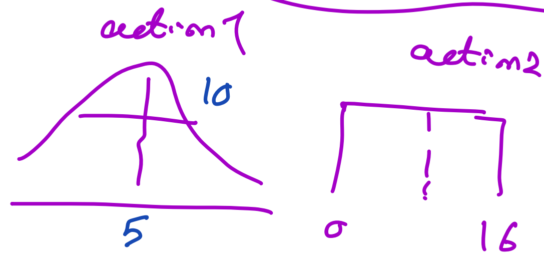
Action 1 → $+\alpha(\underbrace{R_t}_{10} - \underbrace{\bar{R}_t}_{0})\overbrace{(1-\underbrace{\pi_t(A_t)}_{0.9999})}^{0.0001}$

$$\frac{e^{H_t(a)}}{\sum_{a'} e^{H_t(a')}}$$

$\pi_t(a^1) = 0.9999$

$\pi_t(a^1) = 0.0001$

action 2 → $-\alpha(\underbrace{R_t}_{10} - \underbrace{\bar{R}_t}_{0})\underbrace{\pi_t(a)}_{0.0001}$



$R^{a^1} \sim N(5, 10)$

1

$H_1(a^1) = H_1(a^2) = 0$

policy

$$\pi_1(a^1) = \frac{e^{H_1(a^1)}}{e^{H_1(a^1)} + e^{H_1(a^2)}} = \frac{e^0}{e^0+e^0} = \frac{1}{2}$$

$$\pi_1(a^2) = \frac{e^0}{e^0+e^0} = 1/2$$

$\sim A_1 = a^1$

$A_1 = a^1,\ R_1 = 10 \Rightarrow \bar{R}_1 = \frac{10}{1} = 10$ $\qquad \frac{r_1+r_2+r_3+\dots+r_T}{T}$

preference update

$$H_2(a^1) = \underbrace{H_1(a^1)}_{0} + \underbrace{\alpha}_{0.5}[\underbrace{R_1}_{10} - \underbrace{\bar{R}_1}_{10}](1-\underbrace{\pi_1(a^1)}_{1/2}) = 0$$

$$H_2(a^2) = \underbrace{H_1(a^2)}_{0} - \underbrace{\alpha}_{0.5}[\underbrace{R_1}_{10} - \underbrace{\bar{R}_1}_{10}]\underbrace{\pi_1(a^2)}_{1/2} = 0$$

(2)

$H_2(a^1)=0 \quad , \quad H_2(a^2)=0$

$$\hookrightarrow \left.\begin{aligned} \pi_2(a^1) &= \frac{e^0}{e^0+e^0} = \frac{1}{2} \\ \pi_2(a^2) &= \frac{1}{2} \end{aligned}\right\} \longrightarrow A_t = a^2$$

$A_2 = a^2, \; R_2 = 2 \quad \rightarrow \bar{R}_2 = \frac{10+2}{2} = 6$

preference update

$$\hookrightarrow \left\{\begin{aligned} H_3(a^2) &= \underbrace{H_2(a^2)}_{0} + \underbrace{\alpha}_{0.5}\Big[\underbrace{R_2}_{2} - \underbrace{\bar{R}_2}_{6}\Big]\Big(1 - \underbrace{\pi_2(a^2)}_{1/2}\Big) = -1 \\ H_3(a^1) &= \underbrace{H_2(a^1)}_{0} - \underbrace{\alpha}_{0.5}\Big[\underbrace{R_2}_{2} - \underbrace{\bar{R}_2}_{6}\Big]\underbrace{\pi_2(a^1)}_{1/2} = 1 \end{aligned}\right.$$

$H_3(a^1)=1 \; , \quad H_3(a^2)=-1$

$$\hookrightarrow \left.\begin{aligned} \pi_3(a^1) &= \frac{e^1}{e^1+e^{-1}} = 0.88 \\ \pi_3(a^2) &= \frac{e^{-1}}{e^1+e^{-1}} = 0.12 \end{aligned}\right\}$$