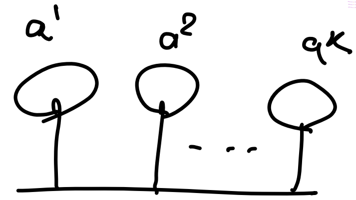
Lecture 4 - Jan 24, 2023

- Multi Arm Bandits
  - Introduction
  - Exploration - Exploitation Delima
  - Epsilon- Greedy Policy
  - Optimistic Initial Values
  - Upper Confidence Bound Selection Policy
  - Gradient-Based Selection Policy
  - Thompson Sampling
- Reinforcement Learning Preliminaries
  - State, Action, Reward, Policy

HW1 → Due Jan 28
Project 1 → Due Feb 7

TA's office hour: Wendsdays, 12pm - 1pm (in-person)
Fridays, 12pm - 1pm (virtual)

Policy Q-Estimate



1- ε-greedy

$$\pi \quad a \sim \begin{cases} \underset{a\in A}{\operatorname{argmax}}\, Q(a) & \text{w.p } 1-\varepsilon \\ \text{Random}\{a^1, \dots, a^k\} & \varepsilon \end{cases}$$

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

$\rightarrow$ Larger $\varepsilon \rightarrow$ more exploration

Optimistic initial Value $\Rightarrow$ $Q(a) = 0 \nearrow$ Large value

2- Upper Confidence Bound policy

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

$$a_{t+1} = \underset{a\in A}{\operatorname{argmax}} \left[Q(a) + c\sqrt{\frac{\log t}{N_a(t)}} \right]$$

Larger $c \longrightarrow$ More exploration

3- Gradient-Bandit policy (does not Q)

$\rightarrow$ preference

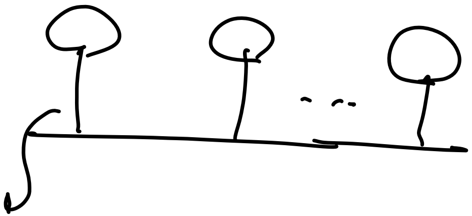
$H_0(a) = 0$ for all $a \in A$

$$\pi_t(a) = \frac{e^{H_t(a)}}{\sum_{b\in A} e^{H_t(b)}}$$

$$\begin{cases} H_{t+1}(A_t) = H_t(A_t) + \alpha\,[R_t - \bar{R}_t]\,(1-\pi_t(A_t)) \\ H_{t+1}(a) = H_t(a) - \alpha\,[R_t - \bar{R}_t]\,\pi_t(a) \end{cases}$$

4- Thompson Sampling → Bayesian Bandit Algorithms

Bernouli (P)
0.9



R^a = Distribution

R^{a^i}: Bernouli $\left(\underbrace{\dfrac{\alpha_i}{\alpha_i+\beta_i}} \right)$

Beta Distribution for likelihood of observed win/loss

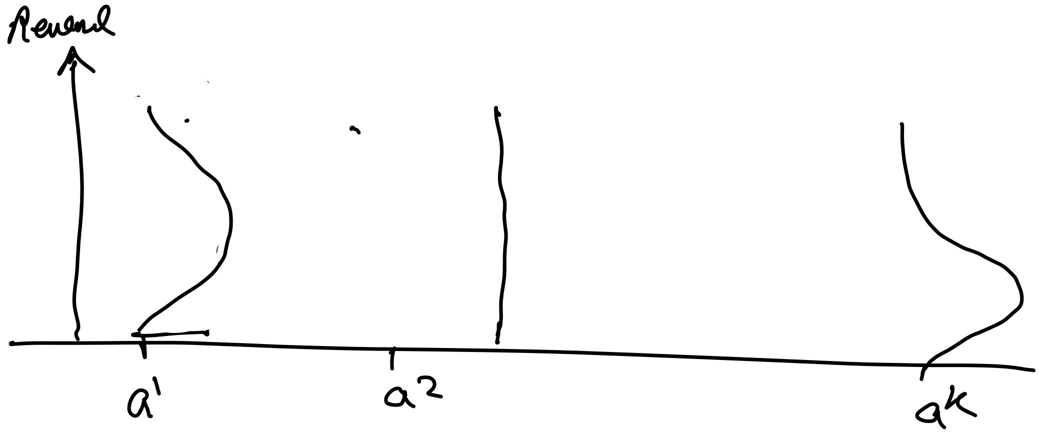
Beta(α_i, β_i)

$$P(\theta_a | D) = \frac{P(D|\theta^a)\,P(\theta^a)}{P(D)}$$

arem $\underline{2}$ seleted

Win → $\alpha_2 = \alpha_2 + 1$

Loss → $\beta_2 = \beta_2 + 1$

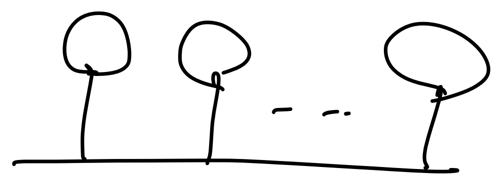


$r_i \sim R^{a^1}$ $\quad$ $r_2 \sim R^{a^2}$ $\quad$ $r_k \sim R^{a^k}$

$a_* = \underset{i}{\arg\max}\ r_i$

Proof of Gradient Bandit Policy

$$\pi_t(a) = \frac{e^{H_t(a)}}{\sum_{b \in A} e^{H_t(b)}}$$



$H_t(a^1) = 0$, $H_t(a^2) = 0$ $\rightarrow$ $\left\{ \begin{array}{l} \pi_t(a^1) = \frac{1}{2} \\ \pi_t(a^2) = \frac{1}{2} \end{array} \right.$

$H_t(a^1) = 8$, $H_t(a^2) = 0$ $\rightarrow$ $\left\{ \begin{array}{l} \pi_t(a^1) = 0.9997 \\ \pi_t(a^2) = 0.0003 \end{array} \right.$

Objective: maximizing Accumulated Reward

$E[R_t]$

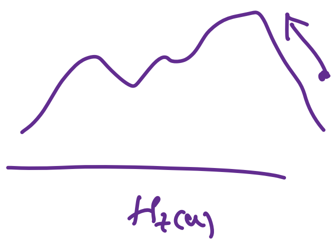
$\hookrightarrow E[R_t \mid \pi_t(a^1), \pi_t(a^2)]$

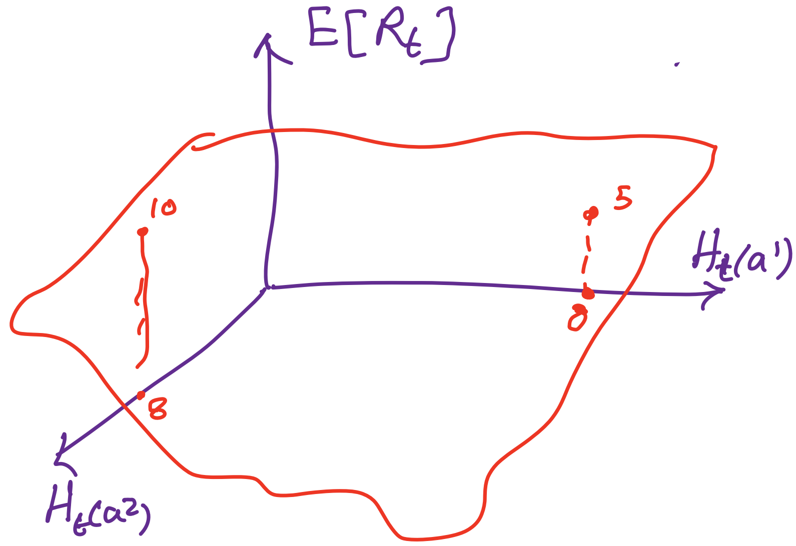
$= E[R_t \mid H_t(a^1), H_t(a^2)]$

$Q(a^1) = 5$

$Q(a^2) = 10$

$= \sum_{b \in A} Q(b)\, \pi_t(b)$





$$H_{t+1}(a) = H_t(a) + \alpha \frac{\partial E[R_t]}{\partial H_t(a)} \quad \text{for all } a \in A$$

$$\frac{\partial E[R_t]}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \left[\sum_{b \in A} Q(b)\, \pi_t(b) \right]$$

$$= \sum_{b \in A} Q(b) \frac{\partial \pi_t(b)}{\partial H_t(a)}$$

$$= \sum_{b \in A} \left(Q(b) - \overset{\bar{R}_t}{X_t} \right) \frac{\partial \pi_t(b)}{\partial H_t(a)}$$

any scalar

why?

$$\pi_t(a^1) = \frac{e^{H_t(a^1)}}{e^{H_t(a^1)} + e^{H_t(a^2)}} \qquad x = H_t(a^1),\ y = H_t(a^2)$$

$$\pi_t(a^2) = \frac{e^{H_t(a^2)}}{e^{H_t(a^1)} + e^{H_t(a^2)}}$$

$$\rightarrow \sum_{b \in A} \frac{\partial \pi_t(b)}{\partial H_t(a^1)} = \frac{\partial}{\partial x} \frac{e^x}{e^x + e^y} + \frac{\partial}{\partial x} \frac{e^y}{e^x + e^y}$$

$$\frac{\partial E[R_t]}{\partial H_t(a)} = \sum_{b \in A} \pi_t(b) \left(Q(b) - \bar{R}_t\right) \frac{\partial \pi_t(b)}{\partial H_t(a)} \Big/ \pi_t(b)$$

$$= E\left[\left(Q(b) - \bar{R}_t\right) \frac{\partial \pi_t(b)}{\partial H_t(a)} \Big/ \pi_t(b)\right]$$

$$A_t \sim \pi_t = \pi_t(a^1),\ \pi_t(a^2)$$

$$\frac{\partial E[R_t]}{\partial H_t(a)} \approx \left(\underbrace{Q(A_t)}_{R_t} - \bar{R}_t\right) \underbrace{\frac{\partial \pi_t(A_t)}{\partial H_t(a)}} \Big/ \pi_t(A_t)$$

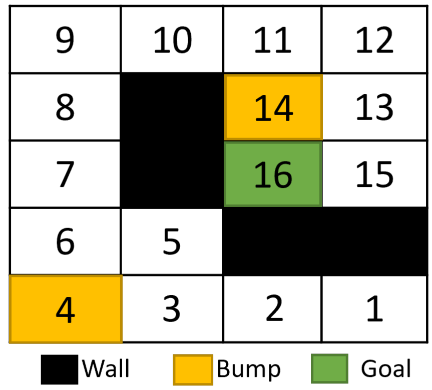
$$\frac{\partial \pi_t(A_t)}{\partial H_t(a)} \rightarrow a \neq A_t \cdot \quad \pi(A_t)\left(-\pi(a)\right)$$

$$\hookrightarrow a = A_t \rightarrow \pi(A_t)\left(1 - \pi(A_t)\right)$$

$$\left.\begin{aligned} \pi(A_t) &= \frac{e^x}{e^x + e^y} \\ \pi(a) &= \frac{e^y}{e^x + e^y} \end{aligned}\right\} \rightarrow \frac{\partial \pi_t(A_t)}{\partial H_t(A_t)} = \frac{\partial}{\partial \underbrace{H_t(A_t)}_{x}} \frac{e^x}{e^x + e^y}$$

$$\begin{cases} H_{t+1}(A_t) = H_t(A_t) + \alpha\left[R_t - \bar{R}_t\right]\left(1 - \pi_t(A_t)\right) \\ H_{t+1}(a) = H_t(a) - \alpha\left[R_t - \bar{R}_t\right]\pi_t(a) \end{cases}$$

Reinforcement Learning – Preliminaries



State Space $\underline{\mathcal{S}}$ $\underline{s} \to$ state

$s \in \mathcal{S}$

Action Space $\mathcal{A}$

$a \in \mathcal{A}$

$\mathcal{A} = \{Up, Down, Left, Right\}$

Immediate Reward $R: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \text{Real}$

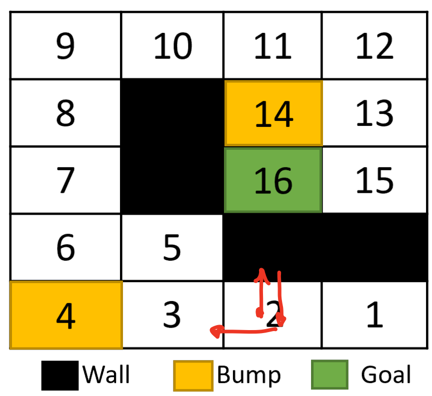
$R(s, a, s')$

$$\begin{cases} -1 & \text{any movement} \\ -10 & \text{Bump} \\ 100 & \text{Goal} \end{cases}$$

$R(3, a = L, 4) = \underbrace{-1}_{\text{movement}} \underbrace{-10}_{\text{Bump}} = -11$

$R(3, U, 1)$

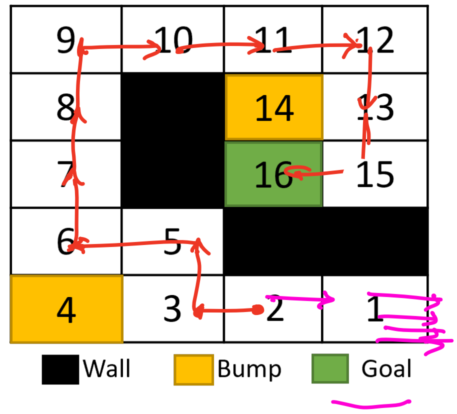
Goal: Find the best ~~path~~ sequence of actions that results in highest Accumulated Reward



$$\underset{\underline{a_{0:T}}}{\operatorname{argmax}} \ E\left[\sum_{t=0}^{T-1} R(s_t, a_t, s_{t+1}) \,\Big|\, s_0 = 2,\ a_{0:T-1},\ s_{t+1} \sim P(s' | s_t, a)\right]$$

$2 \xrightarrow{a_0} \bigcirc \xrightarrow{a_1} \bigcirc$ - - -

a_{20} s_T



-1+100

$\underbrace{-1 \; -1 \; -1 \; -1 \; -1 \; -1 \; - \; -}_{-11} + 99 + 100 + 100 \; -$

$+ 99 + \underbrace{\qquad\qquad}_{800}$

-20