Lecture 20 - March 31, 2023

I - Function Approximation in Reinforcement Learning

- Basics of Function Approximations
- Least square Policy Iteration (LSPI)
- Neural Fitted Q-Iterations (NFQI)
- Deep Q-Network (DQN) → Finite action
- Deep Policy Gradients (DPG) → large/continus Action

Exam 2 → Tues, April 4

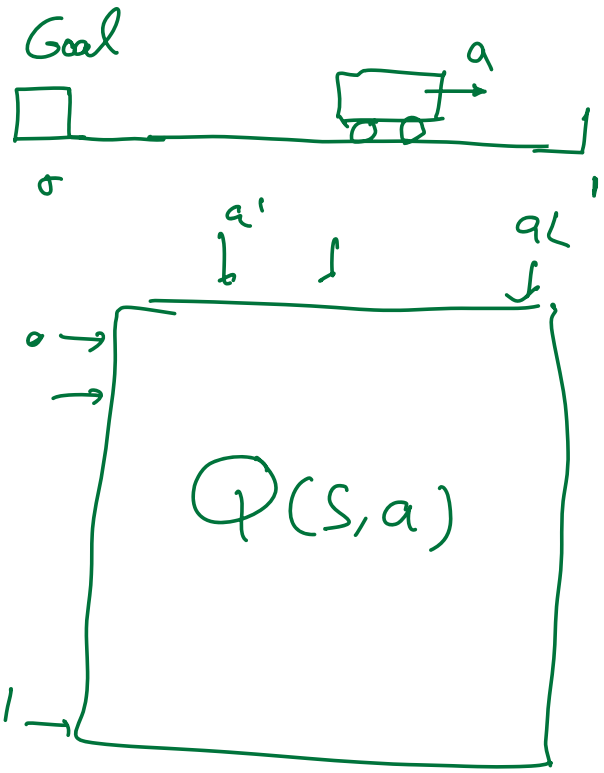
Project 3 → Due April 14

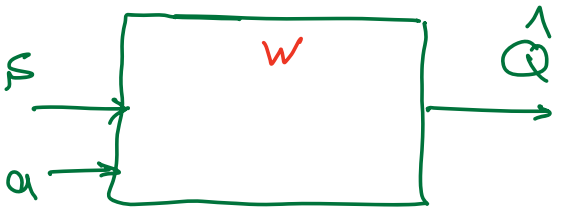
TA's office hour: Wendsdays, 2pm-3pm (in-person)
Fridays, 2pm-3pm (virtual)

$s \in \{[0 \quad 1]\}$

$v_{min} < v < v_{max}$

$a \in \{-1, 0, +1\}$





$\hat{Q}(s, a; w) = s\,a\,w$

$w = 1$

$\hat{Q}(s, a^1) = s \times (-1) \times 1 = -s$

$\hat{Q}(s, a^2) = s \times (0) \times 1 = 0$

$\hat{Q}(s, a^3) = s \times 1 \times 1 = s$

$s = 0.5 \rightarrow a^* = \underset{a \in A}{\operatorname{argmax}}\, Q(s, a)$

$= \operatorname{argmax}\{-0.5, 0, 0.5\} = a^3 = 1$

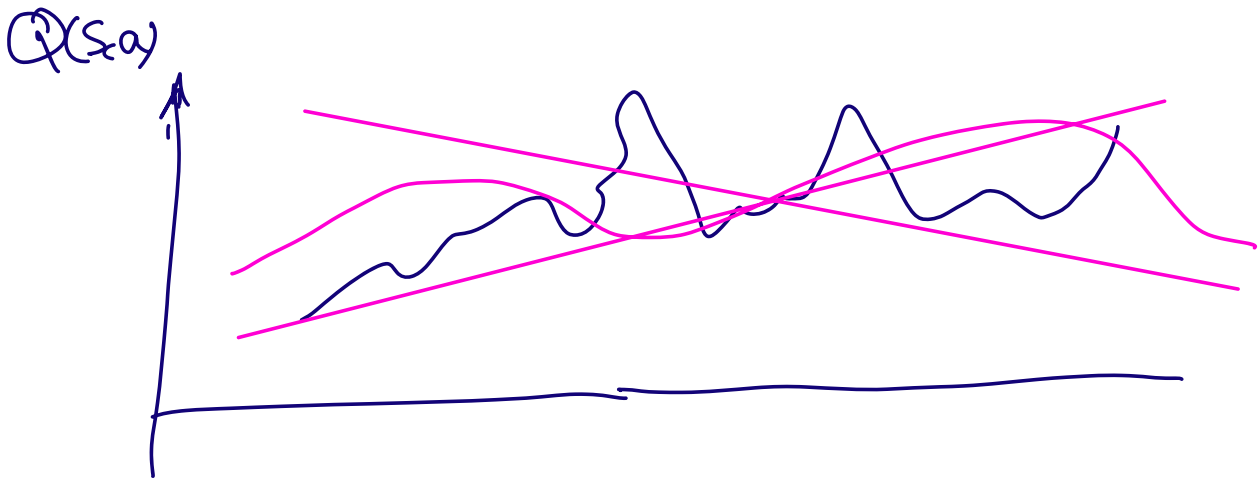
$\hat{Q}_w(s,a;w) = \phi^T(s,a)\, w$: Linear Function Approx

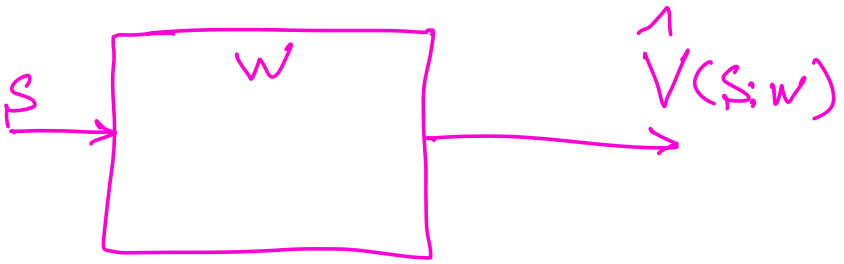
↳ Feature or Basis function

$$\phi(s,a) = \begin{bmatrix} 1 \\ s \\ sa \\ s^2a \\ sa+a^2 \end{bmatrix}$$

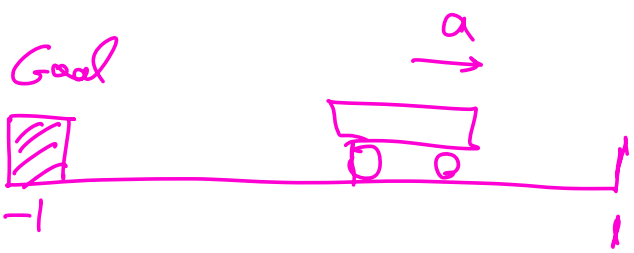
$$\hat{Q}_w(s,a) = \begin{bmatrix} 1 & s & sa & s^2a & sa+a^2 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{bmatrix}$$

$$= w_1 + s w_2 + sa\, w_3 + s^2a\, w_4 + (sa+a^2)\, w_5$$





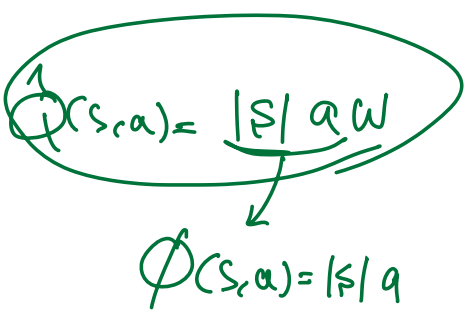
$$\hat{V}(s;w) = \phi^T(s)\, w$$



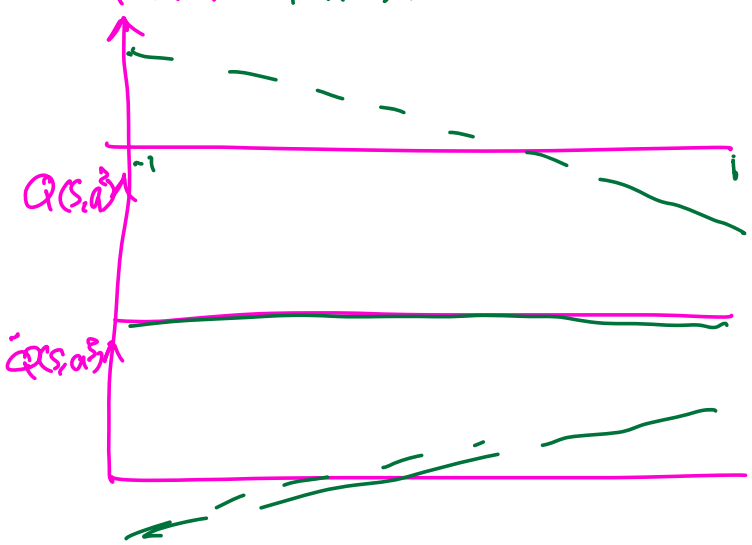
$$\pi^*(s) = -1$$

$$a = \{-1, \ldots, 1\}$$

$$\hat{Q}(s,a) = s\,a\,w$$



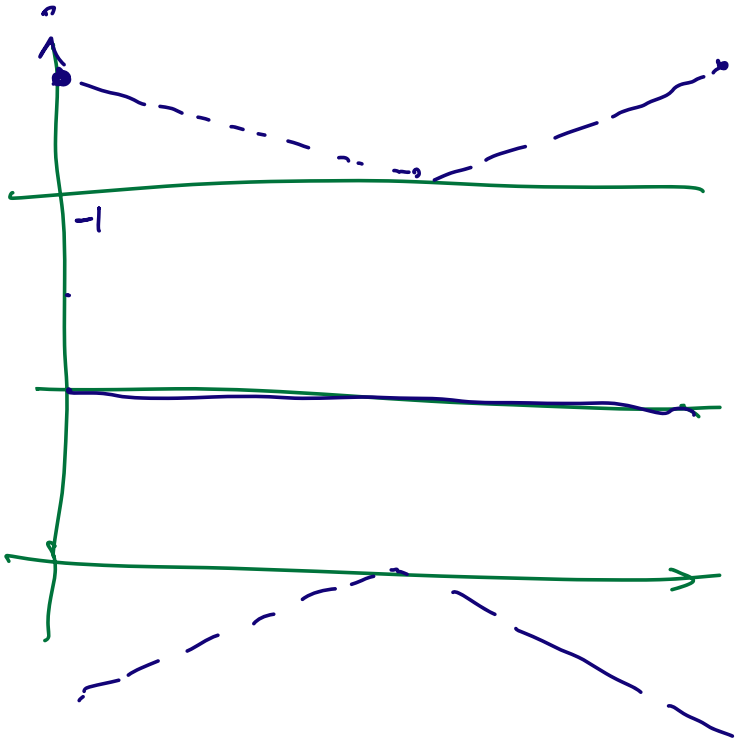
$$Q(s,a)^{w=1} = s \times (-1) \times 1 = -s$$



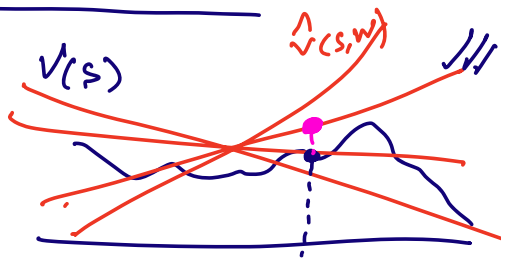
$\hat{Q}(s,a,w) = |s|\,a\,w$

$w = -1$

$\hat{Q}(s,a,w) = -|s|\,a$



$\pi \rightsquigarrow V^{\pi}$

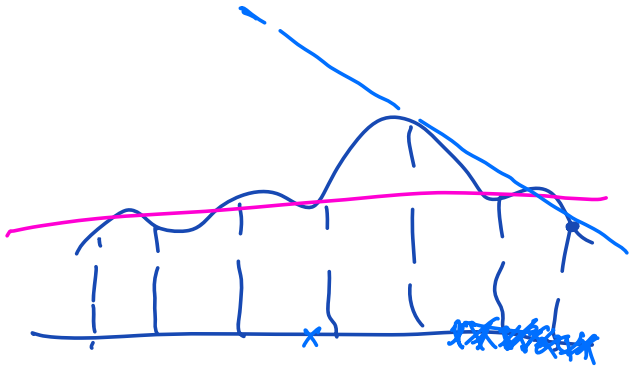


$$V_w(s) = \phi^T(s)\,w$$

$$J(w) = E_{\pi}\left[\left(V^{\pi}(s) - \hat{V}(s;w) \right)^2 \right]$$

$\Rightarrow$ Learn $w^* = \underset{w \in \mathcal{W}}{\operatorname{argmin}}\, J(w)$

① ⇐ Degree of freedom



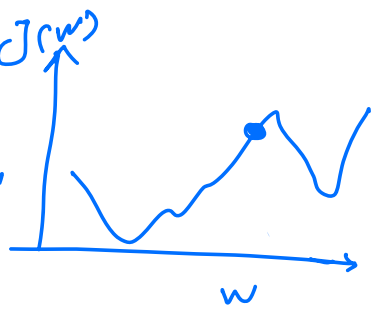
$$J(w) = E_{\pi}\left[\left(V^{\pi}(s) - \hat{V}(s;w) \right)^2 \right]$$

$$= \sum_{s \in S} \mu(s) \left(V^{\pi}(s) - \hat{V}(s;w) \right)^2$$

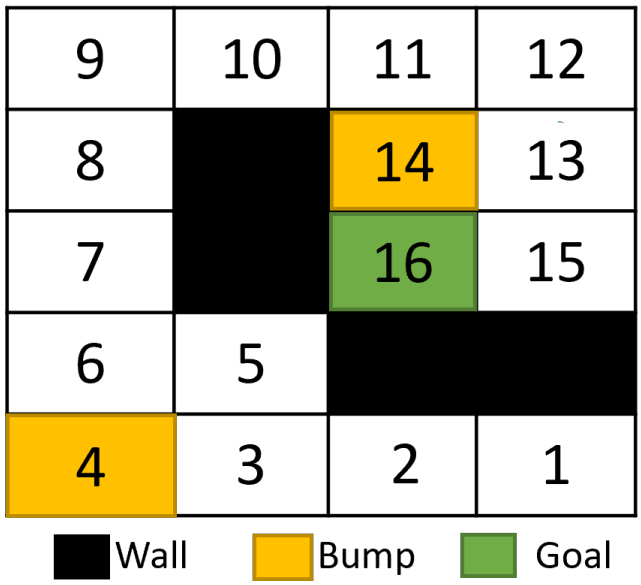
distribution of states under policy π

Gradient descend ⟹ ⇒ Learn $w^* = \underset{w \in W}{\operatorname{argmin}} J(w)$

$$w_t \leftarrow w_{t-1} - \frac{1}{2} \alpha \nabla_w J(w) \Big|_{w = w_{t-1}}$$



$$w_t \leftarrow w_{t-1} + \frac{1}{2} \alpha 2 \left(V^{\pi}(s) - \hat{V}(s;w) \right) \nabla_w \hat{V}(s;w)$$



$V^{\pi}(s) \Rightarrow \hat{V}(s;w)$

$$\hat{V}(s;w) = \underbrace{\begin{bmatrix} 1_{s=s^1} & 1_{s=s^2} & \cdots & 1_{s=s^{15}} \end{bmatrix}}_{\phi^T(s)} \begin{bmatrix} w_1 \\ \\ \\ w_{15} \end{bmatrix}$$

$s=1 \rightarrow \hat{V}(s;w) = \phi^T(s=1)\, w = w_1$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots \end{bmatrix}^T \begin{bmatrix} w_1 \\ \vdots \\ w_{15} \end{bmatrix}$$

$s=15 \rightarrow \hat{V}(s;w) = w_{15}$

Monte-Carlo $\Rightarrow$ large/continuous space

$\pi \longrightarrow V^{\pi}$ $\quad V(s;w)$

$s_0, G(s_0)$ $\quad s_0 \xrightarrow{\pi} s_1, r_1 \rightsquigarrow \hat{V}^{\pi}(s_0) = E[G_{s_0} | \cdots]$

$G(s_0) = r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots + \gamma^{T-1} R_T$

$s_1, G(s_1)$ $\quad s_1 \xrightarrow{\pi} s_2, r_2$

$G(s_1) = r_2 + \gamma r_3 + \gamma^2 r_4 + \cdots + \gamma^{T-2} R_T$

$s_2, \xrightarrow{\pi} s_3, r_3$

$\vdots$

$s_{T-1}, G(s_{T-1})$ $\quad s_{T-1} \xrightarrow{\pi} s_T, r_T$ $\quad G(s_T) = R_T$

$$J(w) = E_{\pi}\left[\left(\hat{V}^{\pi}(s) - \hat{V}(s;w)\right)^2\right]$$

$$= \frac{1}{|D|} \sum_{s \in D} \left(G_t - \hat{V}(s;w)\right)^2$$

$$w_t = w_{t-1} - \frac{1}{2} \alpha \nabla J(w)\Big|_{w_{t-1}}$$

$$w_t = w_{t-1} + \frac{1}{2} \times 2\alpha \left(G_t - \hat{V}(s;w)\right) \nabla \hat{V}(s;w)\Big|_{w_{t-1}}$$

First-visit MC prediction, for estimating $V \approx v_\pi$

Input: a policy π to be evaluated

Initialize:

$V(s) \in \mathbb{R}$, arbitrarily, for all $s \in \mathcal{S}$

$Returns(s) \leftarrow$ an empty list, for all $s \in \mathcal{S}$

Loop forever (for each episode):

Generate an episode following π: $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow \gamma G + R_{t+1}$

Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:

Append G to $Returns(S_t)$

$V(S_t) \leftarrow$ average$(Returns(S_t))$

Gradient Monte Carlo Algorithm for Estimating $\hat{v} \approx v_\pi$

Input: the policy π to be evaluated

Input: a differentiable function $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$

Algorithm parameter: step size $\alpha > 0$

Initialize value-function weights $\mathbf{w} \in \mathbb{R}^d$ arbitrarily (e.g., $\mathbf{w} = \mathbf{0}$)

Loop forever (for each episode):

Generate an episode $S_0, A_0, R_1, S_1, A_1, \ldots, R_T, S_T$ using π

Loop for each step of episode, $t = 0, 1, \ldots, T-1$:

$\mathbf{w} \leftarrow \mathbf{w} + \alpha\big[G_t - \hat{v}(S_t,\mathbf{w})\big]\nabla\hat{v}(S_t,\mathbf{w})$

$\pi \longrightarrow V^\pi \Longrightarrow \hat{V}(s;w)$ TD-version

$S_0, \pi(S_0), S_1, r_1$

$\hat{V}(s) = V(s) + \alpha [R + \gamma V(s') - V(s)]$

$$J(w) = E_\pi \left[\underbrace{\overbrace{V^\pi(s)}^{G_t} - \hat{V}(s;w)}_{R+\gamma \hat{V}(s';w)} \right]^2$$

$$w_t = w_{t-1} - \frac{1}{2}\alpha \nabla J(w)\Big|_{w_{t-1}}$$

$$= w_{t-1} + \frac{1}{2}\alpha 2 \left(R + \gamma \hat{V}(s_1;w) - \hat{V}(s_0;w) \right) \nabla \hat{V}(s_0;w)\Big|_{w_{t-1}}$$

$$\nabla \hat{V}(s;w) = \nabla (\phi^T(s) w) = \phi(s)$$

Semi-gradient TD(0) for estimating $\hat{v} \approx v_\pi$

Input: the policy π to be evaluated
Input: a differentiable function $\hat{v} : \mathcal{S}^+ \times \mathbb{R}^d \to \mathbb{R}$ such that $\hat{v}(\text{terminal},\cdot) = 0$
Algorithm parameter: step size $\alpha > 0$
Initialize value-function weights $\mathbf{w} \in \mathbb{R}^d$ arbitrarily (e.g., $\mathbf{w} = \mathbf{0}$)

Loop for each episode:
  Initialize S
  Loop for each step of episode:
    Choose $A \sim \pi(\cdot|S)$
    Take action A, observe R, S'
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha\big[R + \gamma\hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})\big]\nabla\hat{v}(S,\mathbf{w})$
    $S \leftarrow S'$
  until S is terminal

Fixed Policy $\pi \longrightarrow v^\pi := \hat{v}(s;w)$

$$Q(S,a) = Q(S,a) + \alpha\left[R + \gamma Q(S',a') - Q(S,a)\right]$$

$S_0 \xrightarrow{a \sim \pi^{\varepsilon g}} S_1, R \rightarrow a' \sim \pi^{\varepsilon g}$

$$J(w) = E_\pi\left[\left(Q(S,a) - \hat{Q}(S,a;w)\right)^2\right]$$

$R + \gamma \hat{Q}(S',a';w)$

$\phi^T(S,a)w$

$$w_t = w_{t-1} - \tfrac{1}{2}\alpha \nabla J(w)$$

$$= w_{t-1} + \tfrac{1}{2} 2\alpha \left(R + \gamma \hat{Q}(S',a';w) - \hat{Q}(S,a;w)\right) \nabla \hat{Q}(S,a;w)$$

Episodic Semi-gradient Sarsa for Estimating $\hat{q} \approx q_*$

Input: a differentiable function $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$

Initialize value-function weights $\mathbf{w} \in \mathbb{R}^d$ arbitrarily (e.g., $\mathbf{w} = \mathbf{0}$)
Repeat (for each episode):
  $S, A \leftarrow$ initial state and action of episode (e.g., ε-greedy)
  Repeat (for each step of episode):
    Take action A, observe R, S' → $\pi^{\varepsilon g}(S) = \begin{cases} \operatorname{argmax}_a Q(S,a) & 1-\varepsilon \\ \operatorname{argmax}_{a \in \mathcal{A}} \phi(S,a)^T w & \\ \text{Random} & \varepsilon \end{cases}$
    If S' is terminal:
      $\mathbf{w} \leftarrow \mathbf{w} + \alpha\left[R - \hat{q}(S, A, \mathbf{w})\right] \nabla \hat{q}(S, A, \mathbf{w})$
      Go to next episode
    Choose A' as a function of $\hat{q}(S', \cdot, \mathbf{w})$ (e.g., ε-greedy)
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha\left[R + \gamma \hat{q}(S', A', \mathbf{w}) - \hat{q}(S, A, \mathbf{w})\right] \nabla \hat{q}(S, A, \mathbf{w})$
    $S \leftarrow S'$
    $A \leftarrow A'$

Batch Learning

$D = \{ (s_0, a_0, r_1, s_1) \ldots \}$

$\Downarrow$

π^*

LSPI

$\pi^0 \xrightarrow{PE} V^{\pi^0} \xrightarrow{PI} \pi^1 \rightsquigarrow V^{\pi^1} \rightsquigarrow \pi^2 \rightsquigarrow \cdots$

$\pi^{T+1} = \pi^T$

Policy Evaluation:

$$V_{k+1}(s) = \sum_{s'} P(s'|s, \pi(s)) \left[R + \gamma V_k(s') \right]$$

Policy Improvement

$$\pi'(s) = \underset{a \in A}{\operatorname{argmax}} \sum_{s'} P(s'|s,a) \left[R + \gamma V_\pi(s') \right]$$

$$Q^\pi(s,a) = \sum_{s'} P(s'|s,a) \left[R + \gamma V_\pi(s') \right]$$

$$= \sum_{s'} P(s'|s,a) \left[R + \gamma Q_\pi(s', \pi(s')) \right]$$

$$Q^\pi = R + \gamma M Q^\pi$$

$\downarrow$

$$\begin{bmatrix} Q^{\pi}(s_1, a^1) & \cdots & Q^{\pi}(s_1, a^{|A|}) \\ Q(s^{|S|}, a^1) & - & Q^{\pi}(s^{|S|}, a^{|A|}) \end{bmatrix}$$

LSPI