Lecture 22 - April 7, 2023

- Function Approximation in Reinforcement Learning
  - Basics of Function Approximations
  - Least square Policy Iteration (LSPI)
  - Deep Q-Network (DQN) → Finite action
    - Deep Q-Network (DQN)
    - Double DQN
    - Prioritized DQN
    - Dueling DQN
  - Deep Policy Gradients (DPG) → large/continuous Action

Project 3 → Due April 14

HW5 → Due April 18

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

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:

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

Policy Improvement

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

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

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

$$\underset{|S| \times |A|}{Q^\pi} = \underset{|S| \times |A|}{R} + \gamma \, M \, Q^\pi$$

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

LSPI

$$Q(s, a; w) = \sum_{j=1}^{n} \phi_j(s,a)\, w_j = \phi^T_{(s,a)}\, w$$

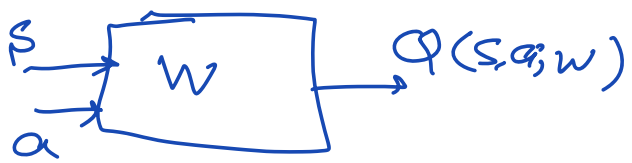
(vector; Basis function)

$$\phi_{(s,a)} = \begin{bmatrix} 1 \\ s \\ a \\ s \otimes a \\ s^2 \otimes a \\ \exp(s,a) \end{bmatrix}_{N \times 1}$$

$$Q(s, a; w) = \phi^T_{(s,a)}\, w$$

$$Q(s,a;w) = \begin{bmatrix} 1 & s & a & sa & s^2a & \exp(sa) \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_6 \end{bmatrix}$$

$$= w_1 + w_2 s + w_3 a + w_4 sa + w_5 s^2 a + w_6 \exp(sa)$$



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

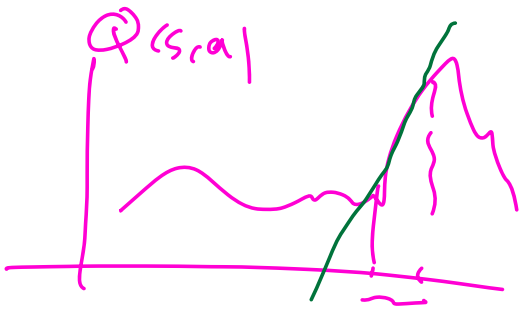
$$\Phi^T \Phi w^\pi = \Phi^T R + \gamma M \Phi w^\pi$$

$$\Phi = \begin{bmatrix} \phi^T(s^1, a^1) \\ \vdots \\ \phi^T(s^{|S|}, a^{|A|}) \end{bmatrix}$$

$$\Phi^T(\Phi - \gamma M \Phi) w^\pi = \Phi^T R$$

$$w^\pi = \underbrace{\left[\Phi^T \mu (\Phi - \gamma M \Phi)\right]^{-1}}_{A} \underbrace{\Phi^T \mu R}_{b}$$

μ: State distribution under policy π



$$w^\pi = A^{-1} b \rightarrow \begin{cases} A = \Phi^T \mu (\Phi - \gamma M \Phi) \\ b = \Phi^T \mu R \end{cases}$$

$$D = \{ (s_i, a_i, r_{i+1}, s'_i) \}_{i=1}^{L} \Leftarrow \text{Batch Data available}$$

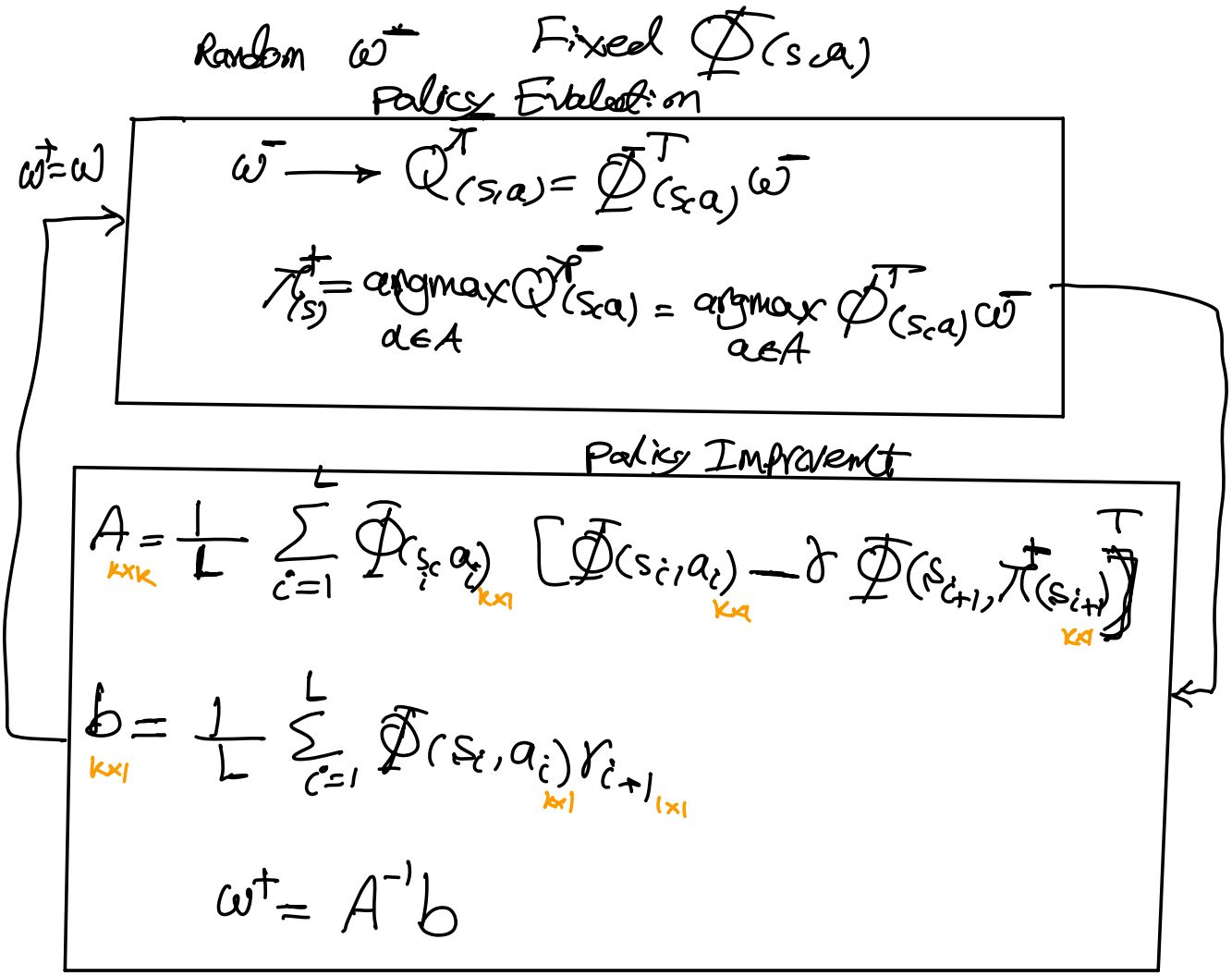
$$\hat{A}_{N\times N} = \frac{1}{L} \sum_{i=1}^{L} \phi(s_i, a_i)_{N\times 1} \left[\phi(s_i, a_i)_{N\times 1} - \gamma \phi(s'_i, \pi(s'_i))_{N\times 1} \right]^T$$

$$\hat{b}_{N\times 1} = \frac{1}{L} \sum_{i=1}^{L} \phi(s_i, a_i)_{N\times 1} r_{i+1}$$

$$w^{\pi}_{N\times 1} = \hat{A}^{-1}_{N\times N} \hat{b}_{N\times 1}$$

$$D \rightarrow \underbrace{w_0 \xrightarrow{\pi^0} A^{\pi^1} b}_{} \rightarrow w_1 \rightarrow A^{\pi^2} b \rightarrow w_2$$

$$\pi^1(s) = \text{argmax}_a Q^{\pi^0}(s, a)$$


Random ω
Fixed $\phi(s,a)$
Policy Evaluation
$\omega^- = \omega$
$\omega \rightarrow Q^{\pi}(s,a) = \phi^T(s,a)\omega$
$\pi(s) = \underset{a \in A}{\operatorname{argmax}} Q^{\pi}(s,a) = \underset{a \in A}{\operatorname{argmax}} \phi^T(s,a)\omega$
Policy Improvement
$A = \frac{1}{L}\sum_{i=1}^{L} \phi(s_i,a_i)\,[\phi(s_i,a_i) - \gamma\,\phi(s_{i+1},\pi(s_{i+1}))]^T$
$b = \frac{1}{L}\sum_{i=1}^{L} \phi(s_i,a_i)\,r_{i+1}$
$\omega = A^{-1}b$

Example of LSPI:

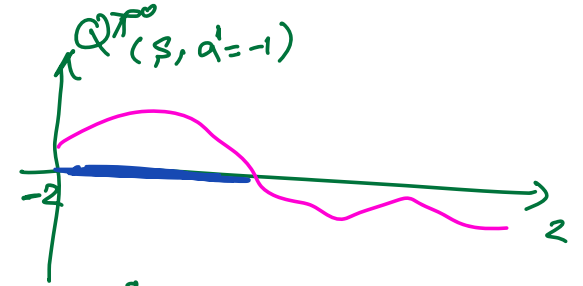
$S \in [-2 \quad 2]$ $\qquad \phi(s,a) = \begin{bmatrix} sa \\ |s+1|\, a^2 \end{bmatrix}$

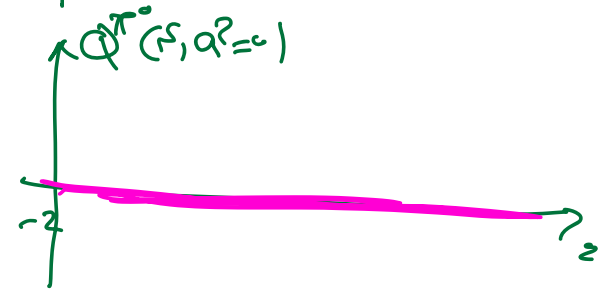
$A = \{-1, 0, 1\}$

$D = \{(s_0=0, a_0=-1, r_1=1, s_1=1), (s_1=1, a_1=1, r_1=0.8, s_2=-1)\}$

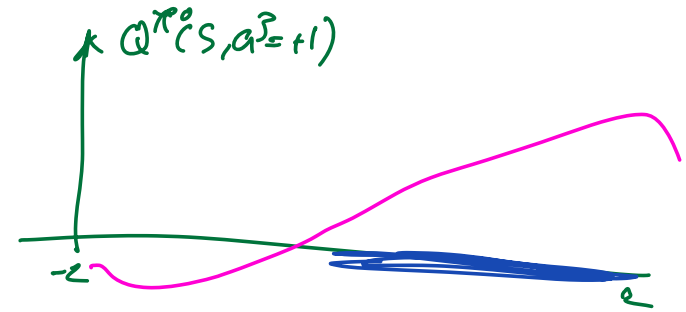
Random $w_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow Q^{\pi_0}(s,a) = \phi^T(s,a)\, w_0 = \begin{bmatrix} sa & |s+1|\, a^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$= sa + |s+1|\, a^2$





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



$$D = \{ (s_0=0, a_0=-1, r_1=1, s_1=1), (s_1=1, a_1=1, r_1=0.5, s_2=-1) \}$$

$$\hat{A} = \frac{1}{L} \sum_{i=0}^{\#} \phi^T(s_i, a_i) \left(\phi(s_i, a_i) - \gamma \phi(s_{i+1}, \pi'(s_{i+1})) \right)^T$$

$$= \frac{1}{2} \left[\phi^T(\underbrace{s_0, a_0}_{\begin{bmatrix}0\\1\end{bmatrix}}) \left(\phi(s_0, a_0) - \gamma \phi(s_1, \pi'(s_1)) \right)^T \right.$$

$\phi(s_1, a_0) = \begin{bmatrix} s_0 \\ a_0 \end{bmatrix} = \begin{bmatrix} 0 \times -1 \\ (-1)^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ $\quad \phi(s_1, \pi'(s_1)) = \begin{bmatrix} s_1 \\ \pi'(s_1) \\ s_1^2 \\ \pi'(s_1)^2 \end{bmatrix} = \begin{bmatrix} 1 \times 1 \\ 1^2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$$\left. + \phi^T(s_1, a_1) \left(\phi(s_1, a_1) - \gamma \phi(s_2, \pi'(s_2)) \right)^T \right]$$

$\phi(s_1, a_1) = \begin{bmatrix} s_1 \\ a_1 \\ s_1^2 \\ a_1^2 \end{bmatrix} = \begin{bmatrix} 1 \times 1 \\ 1^2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $\quad \phi(s_2, \pi'(s_2)) = \begin{bmatrix} s_2 \\ \pi'(s_2) \\ s_2^2 \\ \pi'(s_2)^2 \end{bmatrix} = \begin{bmatrix} -1 \times -1 \\ 0 \times (-1)^2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$$= \begin{bmatrix} -0.05 & 1 \\ -0.35 & 1.6 \end{bmatrix}$$

$$\pi'(s) = \underset{a \in A}{\operatorname{argmax}} Q^{\pi_0}(s, a) = \underset{a \in A}{\operatorname{argmax}} \phi^T(s, a) w_0$$

↳ I only need these for s_i'

$$\pi'(s_1) = \pi'(1) = \underset{a \in \{-1, 0, 1\}}{\operatorname{argmax}} \phi^T(s_1, a) \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$= \underset{a\in\{-1,0,1\}}{\operatorname{argmax}} \begin{bmatrix} s_1 a & |s_1+1|a^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$= \underset{a\in\{-1,0,1\}}{\operatorname{argmax}} \underbrace{s_1 a + |s_1+1| a^2}_{a+2a^2}$$

$$= \operatorname{argmax} \left\{ \underbrace{-1+2(-1)^2}_{1} \quad \underbrace{0+2\cdot 0^2}_{0} \quad \underbrace{1+2(1)^2}_{3} \right\} = 1$$

$$\pi(s_2) = \underset{a}{\operatorname{argmax}}\, \phi^T(s_2, a) \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$= \underset{a\in\{-1,0,1\}}{\operatorname{argmax}} s_2 a + |s_2+1| a^2 = \operatorname{argmax} -a + 0 \times a^2$$

$$= \operatorname{argmax} \left\{ \underbrace{1}_{a=-1} \quad \underbrace{0}_{a=0} \quad \underbrace{-1}_{a=1} \right\} = -1$$

$$b = \frac{1}{L} \sum_{i=0}^{1} \phi(s_i, a_i)\, r_{i+1}$$

$$= \frac{1}{2} \left[\underbrace{\phi(\overset{0}{s_0}, \overset{-1}{a_0})}_{\begin{bmatrix} 0 \\ 1 \end{bmatrix}} \underbrace{r_1}_{1} + \underbrace{\phi(\overset{1}{s_1}, \overset{0}{a_1})}_{\begin{bmatrix} 1 \\ 2 \end{bmatrix}} \underbrace{r_2}_{0.9} \right] = \begin{bmatrix} 0.45 \\ 1.4 \end{bmatrix}$$

$=$

$$w' = \hat{A}^{-1}\hat{b} = \begin{bmatrix} -0.05 & 1 \\ 0.35 & 1.6 \end{bmatrix}^{-1} \begin{bmatrix} 0.43 \\ 1.4 \end{bmatrix} = \begin{bmatrix} -1.58 \\ 0.53 \end{bmatrix}$$

$$Q^{\pi}(s,a) = \phi^T(s,a)\, w' = \begin{bmatrix} s \\ a \\ |s+1| \\ a^2 \end{bmatrix}^T \begin{bmatrix} -1.58 \\ 0.53 \end{bmatrix}$$

$$= -1.58\, s_a + 0.53\, |s+1|_{a^2}$$

$$D = \{ (s_0=0,\ a_0=-1,\ r_1=1,\ s_1=1),\ (s_1=1,\ a_1=1,\ r_1=0.5,\ s_2=-1) \}$$

$$\hat{A} = \frac{1}{L} \sum_{i=0}^{L} \phi(s_i, a_i) \left(\phi(s_i, a_i) - \gamma \phi(s_{i+1}, \pi(s_{i+1})) \right)^T$$

$$= \frac{1}{2} \left[\phi^T(s_0, a_0) \left(\phi(s_0, a_0) - \gamma \phi(s_1, \pi^2(s_1)) \right)^T \right.$$

$$\phi(s_1, a_0) = \begin{bmatrix} s_0 \\ a_0 \\ |s_0+1| \\ a_0^2 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ |0+1| \\ (-1)^2 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 1 \\ 1 \end{bmatrix} \qquad \phi(s_1, \pi^1(s_1)) = \begin{bmatrix} s_1 \\ \pi^2(s_1) \\ |s_1+1| \\ \pi^2(s_1)^2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ |1+1| \\ 1^2 \end{bmatrix}$$

$$\left. + \phi^T(s_1, a_1) \left(\phi(s_1, a_1) - \gamma \phi(s_2, \pi^2(s_2)) \right)^T \right]$$

$$\phi(s_1, a_1) = \begin{bmatrix} s_1 \\ a_1 \\ |s_1+1| \\ a_1^2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ |1+1| \\ 1^2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 1 \end{bmatrix} \qquad \phi(s_2, \pi^1(s_2)) = \begin{bmatrix} s_2 \\ \pi^2(s_2) \\ |s_2+1| \\ \pi^2(s_2)^2 \end{bmatrix}$$

$$= \begin{bmatrix} -0.05 & 1 \\ -0.35 & 1.6 \end{bmatrix}$$

$$= \begin{bmatrix} -1 \times 1 \\ 0 \times (1)^2 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

$$\pi^2(s) = \arg\max_{a \in A} Q^1(s,a) = \arg\max_{a \in A} \phi(s,a)^T w_1$$

$$= \arg\max_{a \in A} \{ -1.58\, s a + 0.53\, |s+1|\, a^2 \}$$

$$\pi^2(s_1 = 1) = \arg\max \{ -1.58a + 2 \times 0.53\, a^2 \} = -1$$

$$\pi^2(s_2 = -1) = \arg\max \{ +1.58a \} = 1$$

$$A = \begin{bmatrix} 0.95 & 1 \\ 2.35 & 1.6 \end{bmatrix}$$

b = same

$$w^2 = A^{-1} b = \begin{bmatrix} 0.82 \\ -0.33 \end{bmatrix}$$

LSPI $(D,\ k,\ \phi,\ \gamma,\ \epsilon,\ \pi_0)$ // Learns a policy from samples

// D : Source of samples (s, a, r, s')
// k : Number of basis functions
// ϕ : Basis functions
// γ : Discount factor
// ϵ : Stopping criterion
// π_0 : Initial policy, given as w_0 (default: $w_0 = 0$)

$\pi' \leftarrow \pi_0$ // $w' \leftarrow w_0$

repeat
  $\pi \leftarrow \pi'$ // $w \leftarrow w'$
  $\pi' \leftarrow$ **LSTD**Q $(D,\ k,\ \phi,\ \gamma,\ \pi)$ // $w' \leftarrow$ **LSTD**Q $(D,\ k,\ \phi,\ \gamma,\ w)$
until $(\pi \approx \pi')$ // **until** $(\|w - w'\| < \epsilon)$

return π // **return** w

LSTDQ $(D,\ k,\ \phi,\ \gamma,\ \pi)$ // Learns $\widehat{Q}^{\pi}$ from samples

// D : Source of samples (s, a, r, s')
// k : Number of basis functions
// ϕ : Basis functions
// γ : Discount factor
// π : Policy whose value function is sought

$\widetilde{\mathbf{A}} \leftarrow \mathbf{0}$ // $(k \times k)$ matrix
$\widetilde{b} \leftarrow \mathbf{0}$ // $(k \times 1)$ vector

for each $(s, a, r, s') \in D$
  $\widetilde{\mathbf{A}} \leftarrow \widetilde{\mathbf{A}} + \phi(s, a)\Big(\phi(s, a) - \gamma\phi\big(s', \pi(s')\big)\Big)^{\mathsf{T}}$
  $\widetilde{b} \leftarrow \widetilde{b} + \phi(s, a)r$

$\widetilde{w}^{\pi} \leftarrow \widetilde{\mathbf{A}}^{-1}\widetilde{b}$

return $\widetilde{w}^{\pi}$

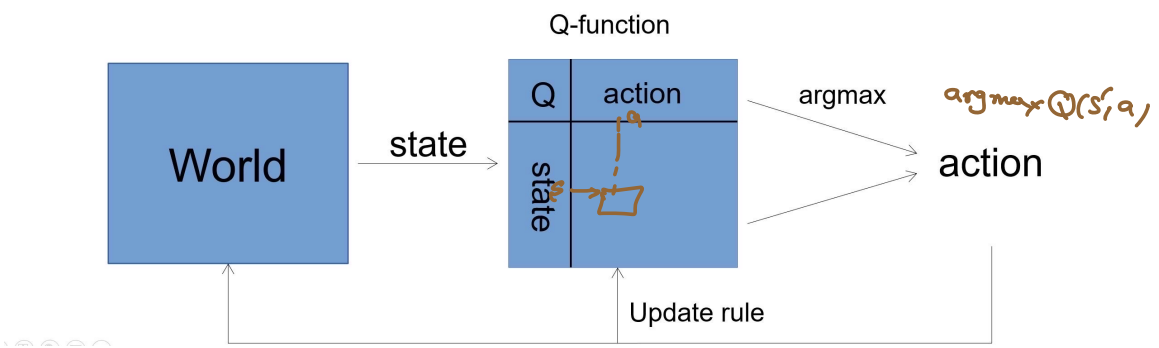
argmax $Q(s', a)$ → $\phi(s', a)w$

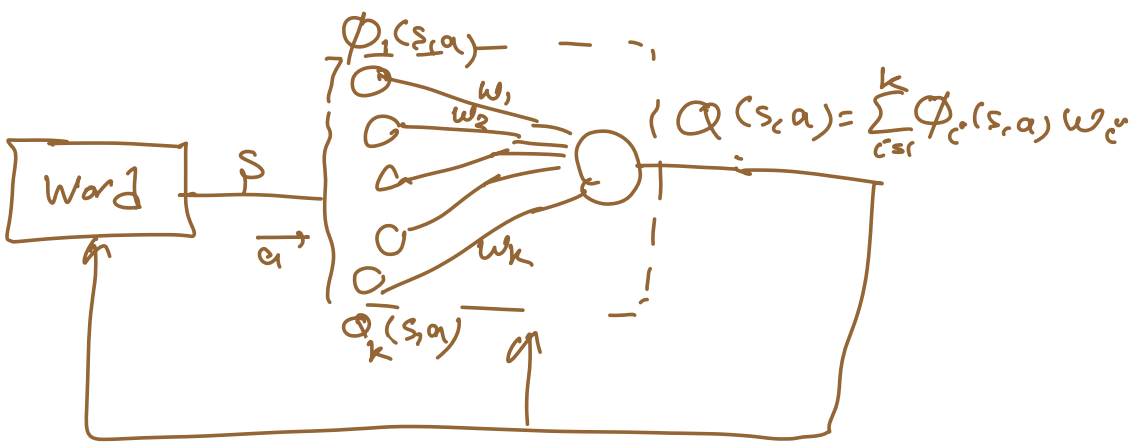
Figure 5: The LSTDQ algorithm.

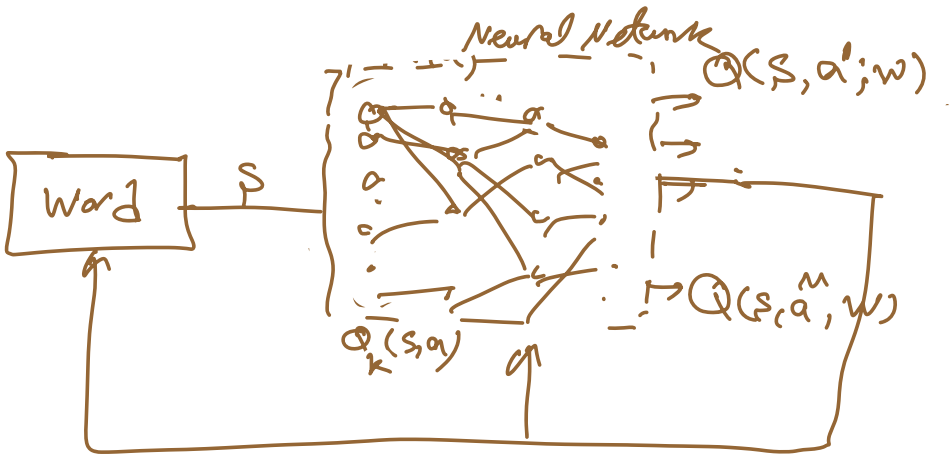
Deep Reinforcement Learning

Tabular Q-Learning

Initialize $Q(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$, arbitrarily, and $Q(\textit{terminal-state}, \cdot) = 0$
Repeat (for each episode):
  Initialize S
  Repeat (for each step of episode):
    Choose A from S using policy derived from Q (e.g., ϵ-greedy)
    Take action A, observe R, S'
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma \max_a Q(S',a) - Q(S,A)\big]$
    $S \leftarrow S'$
  until S is terminal







Q-learning

Target

$$Q(s;a) \leftarrow Q(s,a) + \alpha \left[\underbrace{r + \gamma \max_{a'} Q(s';a')}_{\text{Target}} - Q(s,a) \right]$$

Q-learning Error

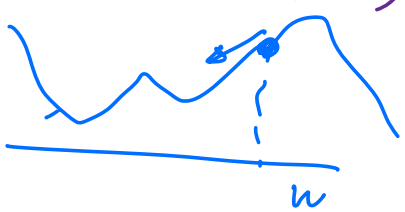
① Model $Q_w(s,a)$

② Loss function

$$L(s,a,s';w) = \frac{1}{2}\left(r + \gamma \max_{a'} Q_w(s';a') - Q_w(s,a)\right)^2$$

↳ decrease

Stochastic Gradient Descent



$$w \leftarrow w - \alpha \nabla_w L(s,a,s';w)$$

$$\nabla L(s,a,s';w) = -\left(r + \gamma \max Q_w(s',a') - Q_w(s,a)\right) \nabla_w Q_w(s,a)$$

$$w \leftarrow w + \alpha \left(r + \gamma \max Q_{\bar{w}}(s',a') - Q_w(s,a)\right) \nabla_w Q_w(s,a)\Big|_{w=\bar{w}}$$

Bare bone DQN (Does NOT Converge)

Initialize Q(*s,a;w*) with random weights
Repeat (for each episode):
  Initialize *s*
  Repeat (for each step of the episode):
    Choose *a* from *s* using policy derived from Q (e.g. e-greedy)
    Take action *a*, observe *r, s'*
    $w \leftarrow w - \alpha \nabla_w L(s, a, s'; w)$
    $s \leftarrow s'$
  Until *s* is terminal

Where:

$$\nabla_w L(s, a, s'; w) \approx -\left(r + \gamma \max_{a'} Q_w(s', a') - Q_w(s, a)\right) \nabla_w Q_w(s, a)$$

$s, a, s', r \quad y = r + \gamma \max_{a'} Q_w(s', a')$

;

}