Sharmil. Sheet1

(1) One arm

$R_1 = 1$, $R_2 = 2$, $R_3 = 3$, $R_4 = 4$, $R_5 = 5$, <Repeats>

a) $$Q_n = \frac{\sum_{i=1}^{n} R_i}{n}$$

b) $Q_1 = 1$; $Q_{(n+1)} = Q_n + \frac{1}{n}[R_n - Q_n]$

$$Q_2 = 1 + \frac{1}{2}[2-1] = 1.5$$

$$Q_3 = 1.5 + \frac{1}{3}[3-1.5] = 2$$

$$Q_4 = 2 + \frac{1}{4}[4-2] = 2.5$$

$$Q_5 = 2.5 + \frac{1}{5}[5-2.5] = 3$$

$$Q_6 = 3 + \frac{1}{6}[1-3] = 2.67$$

$$Q_7 = 2.67 + \frac{1}{7}[2-2.67] = 2.57$$

$$Q_8 = 2.57 + \frac{1}{8}[3-2.57] = 2.62$$

$$Q_9 = 2.62 + \frac{1}{9}[4-2.62] = 2.77$$

$$Q_{10} = 2.77 + \frac{1}{10}[5-2.77] = 2.993$$

Sharmili - sheet 2

$Q_{11} = 2.993 + \frac{1}{11}[1 - 2.993] = 2.81$

$Q_{12} = 2.81 + \frac{1}{12}[2 - 2.81] = 2.74$

$Q_{13} = 2.74 + \frac{1}{13}[3 - 2.74] = 2.76$

$Q_{14} = 2.76 + \frac{1}{14}[4 - 2.76] = 2.84$

$Q_{15} = 2.84 + \frac{1}{15}[5 - 2.84] = 2.984$

1c) Based on above 15 observations, we already find Q value to hover over 3 (± 0.5).

Because the sequence is stationary and (stepsize) α value is $\frac{1}{n}$, we can see the sequence converging to its expected value of 3 (± 0.5).

1d) From above observation, we see sequence converges to ~ 3.

This is because the average of (1, 2, 3, 4, 5) reward sequence is 3 and this sequence is repeated.

Shamili-S3

1e) Yes, we have used increamental Q_{n+1} in above calculations.

$$Q_{n+1} = \frac{\sum_{k=1}^{n+1} R_k}{(n+1)} = \frac{\sum_{k=1}^{n} R_k + R_{(n+1)}}{(n+1)}$$

$$Q_{n+1} = \frac{\sum_{k=1}^{n} R_k}{n} \left[\frac{n}{n+1}\right] + \frac{R_{(n+1)}}{(n+1)}$$

$$Q_{n+1} = Q_n\left[1 - \frac{1}{n+1}\right] + \frac{R_{(n+1)}}{n+1}$$

$$\boxed{Q_{n+1} = Q_n + \frac{1}{n+1}\left[R_{(n+1)} - Q_n\right]}$$

1f) $Q_1 = 1$; $Q_2 = \frac{1+2}{2} = 1.5$; $Q_3 = \frac{1+2+3}{3} = 2$;

$Q_4 = \frac{1+2+3+4}{4} = 2.5$; $Q_5 = \frac{1+2+3+4+5}{5} = 3$

ll^{ly} Using Increamental Update, refer values in (1b)

We see both values are equal.

Sharmili-4

1g) The new Update equation,

$$Q_{n+1} = Q_n + \frac{1}{(n+1)^{0.8}} \left[R_{n+1} - Q_n \right]$$

(ie) Step size $\alpha = \frac{1}{(n+1)^{0.8}}$

Trying observations,

$Q_1 = 1$; $Q_2 = 1 + \frac{1}{2^{0.8}} [2-1] = 1 + \frac{1}{1.74} = 1.57$

$Q_3 = 1.57 + \frac{1}{3^{0.8}} [3 - 1.57] = 2.16$

$Q_4 = 2.16 + \frac{1}{4^{0.8}} [4 - 2.16] = 2.77$

$Q_5 = 2.77 + \frac{1}{5^{0.8}} [5 - 2.77] = 3.39$

Thus, we observe as α (stepsize) value is increased, the convergence rate is slow. But, still we can assume the series will converge to value ~ 3.39.

Shamnil-5

2) Given $\gamma \in (0,1)$, π, $v_\pi(s)$, $q_\pi(s,a)$, $p(s',r|s,a)$

a)

$$q_\pi(s,a) = \sum_{s',r} p(s',r|s,a)\,[r + \gamma\, \underline{v}_\pi[s']]$$

b)

$$v_\pi(s) = \sum_a \pi(a/s)\, q_\pi(s,a)$$

c)

$$v_\pi(s) = \sum_a \pi(a/s) \sum_{s',r} p(s',r|s,a)[r + \gamma\, \underline{v}_\pi[s']]$$

d) Writing Bellman equation in Vector matrix Notation

$$v_\pi = r^\pi + \gamma p^\pi v_\pi$$

$$\Rightarrow (I - \gamma p^\pi) v_\pi = r^\pi$$

$$v_\pi = (I - \gamma p^\pi)^{-1} r_\pi$$

$(I - \gamma p^\pi)$ is invertible, as Eigen values are $(0,1)$

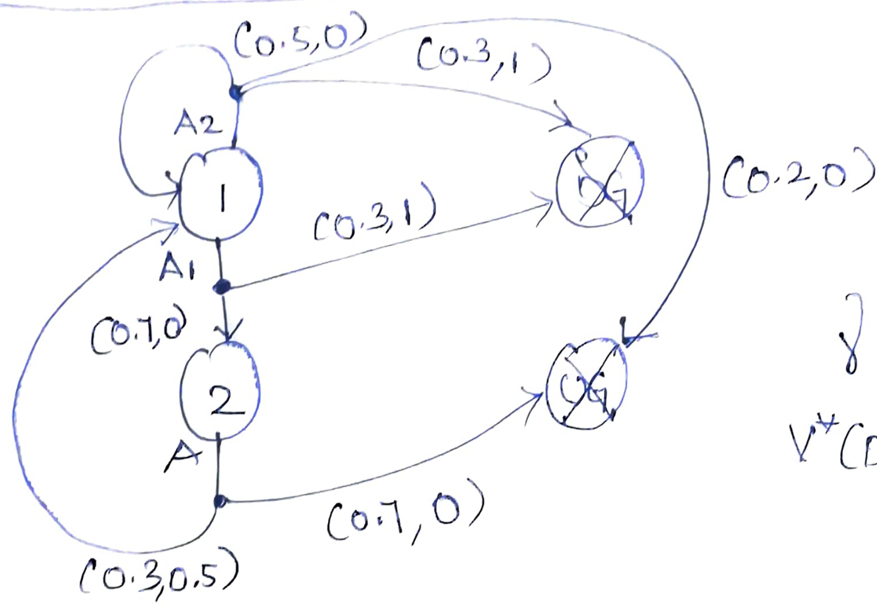
$\Rightarrow$ Solution is Unique.

Thus Bellman equation has an unique solution.

2e) Q-Bellman equation.

$$q_\pi(s,a) = \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma \sum_b \pi(b/s')\, q_\pi(s',b)\right]$$

3.



$\gamma = 1$

$V^*(DG) = V^*(UG) = 0$

a) $V(\text{Room }1) = \frac{1}{2}[\text{Value from Action } 1] +$

$\frac{1}{2}[\text{Value from Action } 2]$

[Assuming Equiprobability of actions]

$$V(\text{Room }1) = \frac{1}{2}\left[((1+\gamma 0)\,0.3) + 0.7\,(0+\gamma V(\text{Room}2))\right]$$
$$+ \frac{1}{2}\left[0.5(0+\gamma V(R1)) + 0.3(1+\gamma 0) + 0.2(0+\gamma 0)\right]$$

$$2\,V(R1) = 0.3 + 0.7\,V(R2) + 0.5\,V(R1) + 0.3.$$

Sharmila 7

$$2V(R_1) = 0.6 + 0.7V(R_2) + 0.5V(R_1)$$

$$\boxed{1.5V(R_1) - 0.7V(R_2) = 0.6} \quad \text{--- (1)}$$

$V(R_2) =$

$$V(R_2) = \left(0.0 + \gamma V(UG)\right)0.7 + \left(0.5 + \gamma V(R_1)\right)0.3$$

$$V(R_2) = 0.15 + 0.3V(R_1)$$

$$\boxed{V(R_2) - 0.3V(R_1) = 0.15} \quad \text{--- (2)}$$

Solving (1) & (2)

$$V(R_1) = 0.546$$

$$V(R_2) = 0.314$$

b) Value Iteration Scheme

$$V(R1) = \max\Big(\big(0.3(1+0) + 0.7\,(0 + V(R2))\big)/2,$$
$$\big(0.5\,(0+V(R1)) + 0.3\,(1+0) + 0.2(0+0)\big)/2\Big)$$

$$\boxed{V(R1) = \max\Big(\big(0.3 + 0.7V(R2)\big)/2,\ \big(0.5V(R1) + 0.3\big)/2\Big)}$$

$\cancel{V(R2) = \max(0.6 + 0.7 \times 8}$

$$V(R2) = \max\big(0.7\,(0+0) + 0.3\,(0.5 + V(R1))\big)$$

$$\boxed{V(R2) = 0.15 + 0.3V(R1)}$$

c) $$V(R1) = \max\big((0.15 + 0.35\,V(R2)),\ (0.25V(R1) + 0.15)\big)$$

$$V(R2) = 0.15 + 0.3V(R1)$$ # No max as only one action

Sharmili-9

Step 0 , $V_0(R1) = 0$; $V_0(R2) = 0$

Step 1

$V_1(R1) = \max((0.15 + 0), (0 + 0.15))$

$V_1(R1) = 0.15$

$V1(R2) = 0.15$

Step 2

$V_2(R1) = \max((0.15 + (0.35 \times 0.15)),$
$((0.25 \times 0.15) + 0.15))$

$= \max(0.2025, 0.1875)$

$= 0.2025$

$V_2(R2) = 0.15 + (0.3 \times 0.15)$

$= 0.195$

Step 3

$V_3(R1) = \max((0.15 + (0.35 \times 0.195)),$
$((0.25 \times 0.2025) + 0.15))$

$= \max(0.218, 0.201) = 0.218$

$V_3(R2) = 0.15 + 0.3(0.2025)$

$V_3(R_2) = 0.21075$

Step 4

$V_4(R1) = \max\left((0.15 + (0.35 \times 0.201)),\ ((0.25 * 0.218) + 0.15)\right)$

$= \max(0.224, 0.205) = 0.224$

$V_4(R2) = 0.15 + (0.3 * 0.218) = 0.2154$

$= \cancel{0.154}$

Step 5

$V_5(R1) = \max\left((0.15 + (0.35 \times 0.2154)),\ ((0.25 \times 0.224) + 0.15)\right)$

$= \max(0.2254, 0.206)$

$= 0.2254$

$V_5(R2) = 0.15 + (0.3 \times 0.224) = 0.2172$

Sharmili-11

3d) $V_{\pi}(R1) = 0.3(1+\gamma 0) + 0.7(0+\gamma V_{\pi}(R2))$

$$\boxed{V_{\pi}(R1) = 0.3 + 0.7 V_{\pi}(R2)} \rightarrow ①$$

$V_{\pi}(R2) = 0.7(0+\gamma 0) + 0.3(0.5+\gamma V_{\pi}(R1))$

$$\boxed{V_{\pi}(R2) = 0.15 + 0.3 V(R1)} - ②$$

Solving ① & ②,

$V_{\pi}(R1) = 0.513$; $V_{\pi}(R2) = 0.304$

3e) $R1 \xrightarrow{0} R2 \xrightarrow{0.5} R1 \xrightarrow{0} R2 \xrightarrow{0.5} R1 \xrightarrow{0} R2 \xrightarrow{0.5} R1 \xrightarrow{0} R2 \xrightarrow{0.5} R1 \xrightarrow{1} DG$

$\gamma = 1$

i) First visit method

$V_{\pi}(R1) = \frac{G_1}{1} = \frac{1+0.5+0.5+0.5+0.5}{1}$

$V_{\pi}(R1) = 3$

$V_{\pi}(R2) = \frac{G_1}{1} = \frac{1+0.5+0.5+0.5+0.5}{1}$

$= 3$

Sharmili: 12

ii) Every Visit method

$$V_\pi(\cancel{R}) = \frac{\sum_{i=1}^{k} G_i}{k}$$

k: Numbes of occurences

G_i : Return at i^{th} Occurence

~~$V_\pi(R1)$~~

For R1

$k = 5$

$G_1 = ((0.5)\times 4) + 1 = 3$

$G_2 = ((0.5)\times 3) + 1 = 2.5$

$G_3 = ((0.5)\times 2) + 1 = 2$

$G_4 = 1.5$

$G_5 = 1$

$$\therefore V_\pi(R1) = \frac{1+1.5+2+2.5+3}{5} = 2$$

For R2

$k = 4$

$G_1 = ((0.5)\times 4) + 1 = 3$

$G_2 = 2.5$; $G_3 = 2$; $G_4 = 1.5$

$$V_{\pi}(R2) = \frac{1.5+2+2.5+3}{4} = \frac{9}{4} = 2.25$$

---

4)

$$\boxed{E} \xleftarrow{0} A \xleftrightarrow{0} B \xleftrightarrow{0} C \xleftrightarrow{0} D \xrightarrow{1} \boxed{F}$$

Action Probability $\boxed{\gamma = 1}$

$$L \xleftarrow{\;1/2\;} \bullet \xrightarrow{\;1/2\;} R$$

Episode

$$A \xrightarrow{0} B \xrightarrow{0} C \xrightarrow{0} D \xrightarrow{1} \boxed{F}$$

$V(A) = V(B) = V(C) = V(D) = 0.6$ $\boxed{\alpha = 0.1}$

a) n-step TD with $n=1$

$$V(S_t) \leftarrow V(S_t) + \alpha\left[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right]$$

3-iterations of Regular TD

e.g. $V(S_1) \leftarrow V(S_1) + \alpha\left[R_2 + \gamma V(S_2) - V(S_1)\right]$

Sharmili-14

Iteration-1

$$V(A) = V(A) + \alpha\left[R_{A\to B} + \gamma V(B) - V(A)\right]$$

$$V(A) = 0.6 + 0.1[0 + 0.6 - 0.6]$$

$$\boxed{V(A) = 0.6}$$

$$V(B) = 0.6 + 0.1[0 + 0.6 - 0.6]$$

$$\boxed{V(B) = 0.6}$$

$$V(C) = 0.6 + 0.1[0 + 0.6 - 0.6]$$

$$\boxed{V(C) = 0.6}$$

$$V(D) = 0.6 + 0.1[1 + 0 - 0.6]$$

$$V(D) = 0.6 + 0.1[0.4]$$

$$\boxed{V(D) = 0.64}$$

Iteration-2

$$V(A) = 0.6 + 0.1[0 + 0.6 - 0.6] = 0.6$$

$$V(B) = 0.6 + 0.1[0 + 0.6 - 0.6] = 0.6$$

$$V(C) = 0.6 + 0.1[0 + 0.64 - 0.6] = 0.604$$

$$V(D) = 0.64 + 0.1[1 + 0 - 0.64] = 0.676$$

Iteration - 3

$$V(A) = 0.6 + 0.1[0 + 0.6 - 0.6] = 0.6$$

$$V(B) = 0.6 + 0.1[0 + 0.604 - 0.6] = 0.6004$$

$$V(C) = 0.604 + 0.1[0 + 0.676 - 0.604] = 0.6112$$

$$V(D) = 0.676 + 0.1[1 + 0 - 0.676] = 0.7084$$

4b) 2-step TD

$$V(S_t) \leftarrow V(S_t) + \alpha\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+3}) - V(S_t)\right]$$

Iteration-1

$$V(A) \leftarrow V(A) + \alpha\left[R_{A\to B} + \gamma R_{B\to C} + \gamma^2 V(C) - V(A)\right]$$

$$V(A) = 0.6 + 0.1[0 + 0 + 0.6 - 0.6] = 0.6$$

$$V(B) = 0.6 + 0.1[\overset{B\to C}{0} + \overset{C\to D}{0} + \overset{V'}{0.6} - 0.6] = 0.6$$

$$V(C) = 0.6 + 0.1[\overset{C\to D}{0} + \overset{D\to T}{1} + \overset{V(T)}{0} - 0.6] = 0.64$$

$$V(D) = 0.6 + 0.1[\overset{D\to T}{1} + 0 + 0 - 0.6] = 0.64$$

Iteration - 2

$$V(A) = 0.6 + 0.1[\overset{A\to B}{0} + \overset{B\to C}{0} + \overset{V(C)}{0.64} - 0.6] = 0.604$$

$$V(B) = 0.6 + 0.1[\overset{B\to C}{0} + \overset{C\to D}{0} + \overset{V(D)}{0.64} - 0.6] = 0.604$$

$$V(C) = 0.64 + 0.1[\overset{C\to D}{0} + \overset{D\to F}{1} + \overset{V(F)}{0} - 0.64] = 0.676$$

$$V(D) = 0.64 + 0.1[\overset{D\to F}{1} + \overset{T}{0} + \overset{T}{0} - 0.64] = 0.676$$

Iteration - 3

$$V(A) = 0.604 + 0.1[\overset{A\to B}{0} + \overset{B\to C}{0} + \overset{V(C)}{0.676} - 0.604] = 0.6112$$

$$V(B) = 0.604 + 0.1[\overset{B\to C}{0} + \overset{C\to D}{0} + \overset{V(D)}{0.676} - 0.604] = 0.6112$$

$$V(C) = 0.676 + 0.1[\overset{C\to D}{0} + \overset{D\to F}{1} + \overset{V(F)}{0} - 0.676] = 0.7084$$

$$V(D) = 0.676 + 0.1[\overset{D\to F}{1} + \overset{T}{0} + \overset{T}{0} - 0.676] = 0.7084$$

4c) 3-step TD

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 V(S_{t+4}) - V(S_t)]$$

Iteration - 1

$$V(A) \leftarrow V(A) + \alpha[R_{A\to B} + \gamma R_{B\to C} + \gamma^2 R_{C\to D} + \gamma^3 V(D) - V(A)]$$

$V(A) = 0.6 + 0.1\,[0+0+0+0.6-0.6] = 0.6$

$V(B) = 0.6 + 0.1\,[\overset{B\to C}{0} + \overset{C\to D}{0} + \overset{D\to A}{1} + \overset{V(A)}{0} - 0.6] = 0.64$

$V(C) = 0.6 + 0.1\,[\overset{C\to D}{0} + \overset{D\to A}{1} + 0 + \overset{V(A)}{0} - 0.6] = 0.64$

$V(D) = 0.6 + 0.1\,[1+0+0+0-0.6] = 0.64$

Iteration - 2

$V(A) = 0.6 + 0.1\,[\overset{A\to B}{0} + \overset{B\to C}{0} + \overset{C\to D}{0} + \overset{V(D)}{0.64} - 0.6] = 0.604$

$V(B) = 0.64 + 0.1\,[0+0+1+0-0.64] = 0.676$

$V(C) = 0.64 + 0.1\,[0+1+0+0-0.64] = 0.676$

$V(D) = 0.64 + 0.1\,[1+0+0+0-0.64] = 0.676$

Iteration - 3

$V(A) = 0.604 + 0.1\,[0+0+0+0.676-0.604] = 0.6112$

$V(B) = 0.676 + 0.1\,[0+0+1+0-0.676] = 0.7084$

$V(C) = 0.676 + 0.1\,[0+1+0+0-0.676] = 0.7084$

$V(D) = 0.676 + 0.1\,[1+0+0+0-0.676] = 0.7084$

Sharmili-18

4d) From above 3 steps, these are the observations

1-step TD

	A	B	C	D
Iteration-1	0.6	0.6	0.6	0.64
Iteration-2	0.6	0.6	0.604	0.676
Iteration 3	0.6	0.6004	0.6112	0.7084

2-step TD

Iteration	A	B	C	D
1	0.6	0.6	0.64	0.64
2	0.604	0.604	0.676	0.676
3	0.6112	0.6112	0.7084	0.7084

3-step TD

Iteration	A	B	C	D
1	0.6	0.64	0.64	0.64
2	0.604	0.676	0.676	0.676
3	0.6112	0.7084	0.7084	0.7084

We see 3-step TD is converging / Learning at a faster rate [As 3 steps are updated in single iteration]

SSM

Sharmili - 19

5a) Semi Gradient TD(0) Update

$$\boxed{w_{(t+1)} = w_t + \alpha \left[R_{t+1} + \gamma \hat{v}(S_{t+1}, w_t) - \hat{v}(S_t, w_t) \right] * \nabla \hat{v}(S_t, w_t)}$$

Here, instead of,

$V(S_{t+1})$, we use function approximation $\hat{v}(S_{t+1}, w_t)$. And Hence, it is called "Semi-gradient".

5b) Using Linear function approximation, we will have

$$\hat{v}(S_t, w_t) = w_t^T X(S_t)$$

$$\hat{v}(S_{t+1}, w_t) = w_t^T X(S_{t+1})$$

$$\nabla \hat{v}(S_t, w_t) = X(S_t)$$

Hence, Update rule of TD(0) under a linear approximation architecture:

$$w_{t+1} = w_t + \alpha \left[R_{t+1} + \gamma w_t^T x(S_{t+1}) - w_t^T x(S_t)\right] x(S_t)$$

$$= w_t + \alpha \left[R_{t+1} + \gamma x(S_{t+1})^T w_t - x(S_t)^T w_t\right] x(S_t)$$

$$= w_t + \alpha \left[R_{t+1} x(S_t) + \gamma x(S_{t+1})^T w_t x(S_t) - x(S_t)^T w_t x(S_t)\right]$$

$\because x(S_{t+1})^T w_t$ is a scalar,

we can write

$$x(S_{t+1})^T w_t x(S_t) = x(S_t) x(S_{t+1})^T w_t$$

$$\therefore \boxed{w_{t+1} = w_t + \alpha \left[R_{t+1} x(S_t) - x(S_t) \left[\left(x(S_t)^T - \gamma x(S_{t+1})^T\right) w_t\right]\right]}$$

Sharmili - 21

5c) For 1-step SARSA

$$\boxed{\omega_{t+1} = \omega_t + \alpha\left[R_{t+1} + \gamma\, \hat{q}(S_{t+1}, A_{t+1}, \omega_t) - \hat{q}(S_t, A_t, \omega_t\right] * \nabla \hat{q}(S_t, A_t, \omega_t)}$$

(or)

Using Linear function Approximation

$$w_{t+1} = w_t + \alpha\left[R_{t+1}\, x(S_t, A_t) - x(S_t, A_t)\left[x(S_t, A_t)^T - \gamma\, x(S_{t+1}, A_{t+1})^T\right] w_t\right]$$

5d) Q-learning Algorithm with linear function approximation

Input: a differential function

$\hat{q} : S^+, A \times \mathbb{R}^d \rightarrow \mathbb{R}$ such that

$\hat{q}(\text{terminal}, \cdot, \cdot) = 0$

Algorithm parameter: stepsize $\alpha > 0$,

small $\varepsilon > 0$

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

Loop for each episode:

Initializes S

Loop for each step of episode:

choose A from S using policy derived from $\hat{q}$ (eg: ε-greedy)

Take action A, observe R, s'

$w \leftarrow w + \alpha(R + \gamma \max_a \hat{q}(s', a, w) - \hat{q}(S, A, w)] \nabla \hat{q}(S, A, w)$

$S \leftarrow s'$

Until S is terminal

5e) Input: 2 differential functions

$\hat{q}_1 : S^+, A \times \mathbb{R}^d \rightarrow \mathbb{R}$ such that

$\hat{q}_1(\text{terminal}, \cdot, \cdot) = 0$

$\hat{q}_2 : S^+, A \times \mathbb{R}^d \rightarrow \mathbb{R}$ such that

$\hat{q}_2(\text{terminal}, \cdot, \cdot) = 0$

Algorith parameter: stepsize $\alpha \in (0,1]$,
Small $\varepsilon > 0$

Initialize: Value-function weights $\omega_1 \in \mathbb{R}^d$ arbitrarily (eg., $\omega_1 = 0$)
||ly $\omega_2 \in \mathbb{R}^d$

Loop for each episode:
Initialize S
Loop for each step of episode:
Choose A from S using policy ε-greedy in $\hat{q}_1 + \hat{q}_2$
Take action A, observe R, s'
With 0.5 probability

$$\omega_1 \leftarrow \omega_1 + \alpha \left(R + \gamma \hat{q}_2(s', \underset{a}{\text{argmax}}\, \hat{q}_1(s', a, \omega_1)) - \hat{q}_1(S, A, \omega_1)\right] \nabla \hat{q}_1(S, A, \omega_1)$$

else:

$$w_2 \leftarrow w_2 + \alpha \Big(R + \gamma \, \hat{q}_1 \big(S', \operatorname*{argmax}_a \hat{q}_2 (S', a, w_2) \big) - \hat{q}_2 (S, A, w_2) \Big) \nabla \hat{q}_2 (S, A, w_2)$$

$S \leftarrow S'$

Untill S is terminal.