INDIAN INSTITUTE OF TECHNOLOGY MADRAS

B

Roll No. M M 1 6 B 0 2 3

Name : H. Vishal

Total No. of Pages

Quiz I ☐ ~~Quiz II~~/ Mid-Sem ☑ End-Semester ☐ Make-up ☐ Date : 24th Feb. 2020

Semester & Degree : 1000 DS 8th Sem Course No. CS 6700 Part :

Question No.	1	2	3	4	5	6	7	8	9	10
Marks	0	0	6	2	2	0	0	1	1.5	0

11	12	13	14	15	16	17	18	19	20	Total
										12.5

Answer on both sides of the paper including the space below

③ (a) $v(s) = \theta^T \phi(s)$ (Linear f^n approximator)

where $\phi(s_1) = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}$, $\phi(s_2) = \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}$, $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$

$$\Rightarrow \quad V(s) = \begin{bmatrix} V(s_1) \\ V(s_2) \end{bmatrix} = \begin{bmatrix} \theta_0 - \theta_1 - \theta_2 \\ -\theta_0 - \theta_1 + \theta_2 \end{bmatrix}$$

Consider $\theta_0 - \theta_2 = a$

$$\Rightarrow \quad V(s) = \begin{bmatrix} a - \theta_1 \\ -(a + \theta_1) \end{bmatrix}$$

The terms $a - \theta_1$, $-(a+\theta_1)$ are linearly independent as $\alpha(a - \theta_1) + \beta(-a - \theta_1) = 0$

$\Rightarrow \alpha = 0, \ \beta = 0$

$\Rightarrow$ $\text{rank}(V) = 2$, $\dim(V) = 2$

$\Rightarrow$ Any point in 2D space ($\mathbb{R}^2$) can be represented and the corresponding value function can be learnt.

This is similar to a look-up table, where instead of hot-encodings, we use a different formulation.

b) TD (0) :

$$\theta_{t+1} \leftarrow \theta_t + \alpha\left(R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)\right) \nabla_{\theta_t} V_t(S_t)$$

$$\Rightarrow \theta_{t+1} \leftarrow \theta_t + \alpha\left(-5 + \gamma V_t(S_1) - V_t(S_2)\right) \nabla_{\theta_t} V_t(S_2)$$

$$V(S_2) = \theta \phi(S_2)$$

$$\nabla_{\theta_t} V_t(S_t) = \phi(S_2) = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}$$

$$\Rightarrow \theta_{t+1}^{[1]} \leftarrow \theta_t^{[1]} - \alpha\left(-5 + \gamma V_t(S_1) - V_t(S_2)\right)$$

$$\theta_{t+2}^{[2]} \leftarrow \theta_t^{[2]} - \alpha\left(-5 + \gamma V_t(S_1) - V_t(S_2)\right)$$

$$\theta_{t+3}^{[3]} \leftarrow \theta_t^{[3]} + \alpha\left(-5 + \gamma V_t(S_1) - V_t(S_2)\right)$$

where $\theta_t^{[i]}$ denotes the i^{th} element of θ_t

(2) In case of policy improvement,

$$\pi_{n+1} \in \arg\max_{\pi} \left\{ r_{\pi_n} + \gamma P_{\pi} v^{\pi_n} \right\}$$

This is equivalent to saying

$$\pi^{*}(s) = \arg\max_{a} q^{*}(s,a)$$

$$\Rightarrow \pi^{*}(s) = \max_{\pi} \sum_{s'} p(s'|s,a) \left[E[r|s,a,s'] + \gamma v^{\pi}(s') \right]$$

The max can be taken inside

Also, if there exists a state s s.t. $q^{\pi}(s,a) > v^{\pi}(s)$

$\rightarrow$ choosing that state would $\uparrow$ v^{π} contradicting the fact that π is optimal

$\Rightarrow$ ~~False~~

(5)

$$G_t = R_{t+1} + R_{t+2} + \dots R_{t+k}$$

$$v^{\pi}(s) = E_{\pi}[G_t \mid S_t = s]$$

$$= E_{\pi}\left[[R_{t+1} + R_{t+2} + \dots R_{t+k}] \mid S_t = s \right]$$

$$= E_{\pi}\left[R_{t+1} + v^{\pi}(s') - R_{t+k+1} \mid S_t = s \right]$$

$$= E_\pi\left[(R_{t+1} - R_{t+k+1}) + v^\pi(s') \mid S_t = s\right]$$

$$\Rightarrow v^*(s) = \max_a \sum_{s'} p(s'|s,a)\left[E\left[r-r' | s, a, s'\right] + v^*(s)\right]$$

$r' = R_{t+k+1}$ is the end of episode reward which we can't determine knowing s, a, s'

$\Rightarrow$ We can't write Bellman optimality eqn in this case

2

(4) SARSA:

$$q_{new}(s,a) = q_{old}(s,a) + \alpha\left[R_{t+1} + \gamma\, q_{old}(s', a') - q_{old}(s,a)\right]$$

Benefit: implicit evaluation and greedification, On-policy

Pitfall: explores too much sometimes

Expected SARSA:

$$q(S_t, a_t) = q(S_t, a_t) + \alpha\left[R_{t+1} + \gamma \sum_{a'} \pi(S_{t+1}, a')\, q(S_{t+1}, a') - q(S_t, a_t)\right]$$

Benefit: Converges to a more optimal q value

Pitfall: More sampling required ?

Q - Learning

$$q(S_t, a_t) = q(S_t, a_t) + \alpha\left[R_{t+1} + \gamma \max_{a'} q(S_{t+1}, a') - q(S_t, a_t)\right]$$

Benefit: Converges to the optimal solution more quickly

Pitfall: Sometimes, if is too greedy

(6)

$$\eta(\theta) = E[R_t]$$

$$= \sum_a q_*(a)\ \pi(a, \theta)$$

$$Pr(a_t = a) = \frac{e^{q_t^*(a)/\beta}}{\sum_{b=1}^{n} e^{q_t(b)/\beta}} = \pi(a, \beta)$$

Thus, we have a definition for the policy π parametrized by the temperature parameter β

The updates are

$$q_{t+1}(a_t) \doteq q_t(a_t) + \alpha(R_t - b_t)\left(1 - \pi_t(a_t)\right) \text{ and} \quad (1)$$

$$q_{t+1}(a) \doteq q_t(a) - \alpha(R_t - b_t)\ \pi_t(a) \quad \forall\ a \neq a_t \quad (2)$$

~~The~~ ~~above~~ And the parameters can be updated as

$$\Delta\beta_t = \alpha_n(R_t - b_t)\ \frac{\partial \ln \pi(a_t; \beta_t)}{\partial\theta}$$

Thus, we have casted this as a policy gradient problem.

6

1

3

(7) Yes, in some sense γ clarifies the eligibility of a particular reward term to the total return G_t

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$$

γ gives indication of how much of R_{t+3} is valued (taken into account) into G_t. This (γ) gives some kind of a spacial eligibility definition, whereas λ gives temporal eligibility

(8) There are total 16 policies possible, out of which following are Blackwell optimal

(i)

T_2	S_1	S_2	S_3	S_4	T_1
T_2	→	←	←	←	T_1

1

$V(s_1) = \frac{a\gamma}{1-\gamma^2}$, $V(s_2) = \frac{a}{1-\gamma^2}$, $V(s_3) = \frac{a}{1-\gamma^2}$

→ optimal for $a > 0$, $\gamma > \frac{-a+\sqrt{a^2+400}}{20}$

(2)

		↓			
T_1	→	→	→	→	T_2

$V(s_1) = 10\gamma^2$

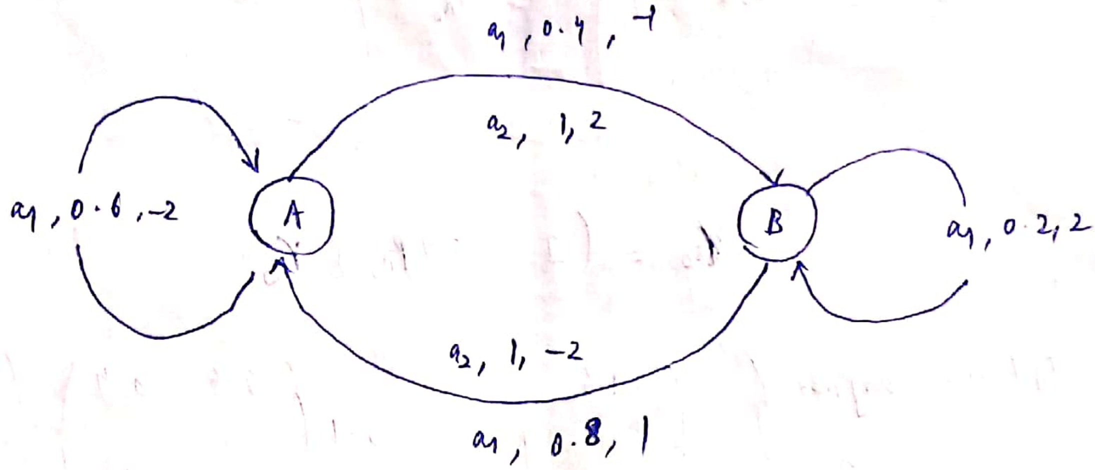
$V(s_2) = 10\gamma^2$

$V(s_3) = 10$

$V(s_4) = 10$

op timal for $a < \frac{5}{\sqrt{2}}$, $\gamma > \frac{1}{\sqrt{2}} = k$ $\boxed{k = 0.5}$

(10)



(a) Initial value f : { A : 0, B : 0 } $\gamma = 1$

$$\cancel{v_1(A) = \max_a \{ a_1 : -1.6 + (1)(0.4)(0) + (1)(0.6)(0) \}}$$

$$v_1(A) = \max \begin{cases} a_1 : -1.6 + 1(0.4)(0) + (1)(0.6)(0) \\ a_2 : 2 + (1)(1)(0) \end{cases} = 2$$

$$\cancel{v_1(B) = \max \{ a_1 : +1.2 + (1)(0.2)(2) + (1)(0.8)(2) \}}$$

$$v_1(B) = \max \begin{cases} a_1 : 1.2 + 0 + 0 \\ a_2 : -2 + 0 + 0 \end{cases} = 1.2$$

$$v_2(A) = \max \begin{cases} a_1 : -1.6 + 0.4(2) + 0.6(2) \\ a_2 : 2 + 2 \end{cases} = 4$$

$$v_2(B) = \max \begin{cases} a_1 : -1.6 + 0.2(1.2) + 0.8(1.2) \\ a_2 : 2 + (1)(1.2) \end{cases} = 3.2$$

(b)

$$P_{\pi_0} = \begin{pmatrix} 0.6 & 0.4 \\ 0 & 1 \end{pmatrix} \qquad r_{\pi_0} = \begin{pmatrix} -1.6 \\ 2 \end{pmatrix}$$

$$V_{\pi_0} = (I - 0.9 P_{\pi_0})\, r_{\pi_0}$$

$$\pi_1(A) = \arg\max_a \left\{ \begin{pmatrix} -1.6 \\ 2 \end{pmatrix} + 0.9 \begin{pmatrix} 0.6 & 0.4 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 1.2 \end{pmatrix} \right\}$$

$$= \arg\max_a \left\{ \begin{pmatrix} -1.6 \\ 2 \end{pmatrix} + \begin{pmatrix} 1.51 \\ 1.33 \end{pmatrix} \right\} = \arg\max_a \left\{ \begin{matrix} -0.09 \\ 3.53 \end{matrix} \right\} = a_2$$

$$\cancel{\pi_1(B) = \arg\max_a \left\{ \begin{pmatrix} 1.2 \\ 2 \end{pmatrix} + 0.9 \begin{pmatrix} 0.2 & 0.8 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \\ \end{pmatrix} \right.}$$

Sly, other iteration can be done

(1)

<~~...~~, $\langle x, f(x) \rangle$

use errors as rewards