Birla Institute of Technology & Science - Pilani, Hyderabad Campus
First Semester 2020-2021
CS F441: Selected Topics (Reinformcent Learning)
Comprehensive Examination

Type: Open Book **Time: 2 hours** **Max Marks: 70** **Date: 02/03/2021**

Answer *all* questions. All parts of the same question should be answered together.

Q. (**Total Marks: 15)** Consider a case of space garbage collector, a mobile unmanned spaceship that collects space garbage. It has sensors to detect space garbage and equipment to pick them. However, this spaceship works on battery where there are three charge levels: {high, medium, low}. Based on the battery level, the spaceship can search for the garbage or it can come back to the space station to recharge.

Initially, the battery level is high and the spaceship is searching. While searching, spaceship checks battery level at some intervals. A period of searching that begins with a high energy level, leaves the energy level high with probability α, reduces it to medium with probability β and reduces it to low with probability $1-\alpha-\beta$. A period of searching that begins with a medium energy level, leaves the energy level medium with probability δ and leaves the energy level low with probability $1-\delta$. While the battery level is high and medium, spaceship never goes to space station to recharge. However, when the battery level is low, and spaceship decides to search, it leaves battery level to low with probability μ and depletes the battery with probability $1-\mu$. On the other hand, when the battery level is low, it can also come back to space station to recharge. In this case, after recharge the battery level is high with probability 1.

Note that when a battery level is higher, the space ship can collect more garbage. Let us refer reward r_s, r_m, r_l as the expected number of garbage spaceship can collect when the battery level is high, medium and low, respectively. In case, the battery depletes then the spaceship needs to be rescued and hence the reward is -5.

Using Bellman Optimality equation, answer the following questions with proper derivation by assuming ($\alpha = \beta = \frac{1}{3}, \delta = \mu = \frac{1}{2}$, $r_s = 3$, $r_m = 2$, $r_l = 1$):

(a) (**Marks: 4)** Find the optimal value function $v_*(h)$.
(b) (**Marks: 4)** Find the optimal value function $v_*(m)$.
(c) (**Marks: 4)** Find the optimal value function $v_*(l)$.
(d) (**Marks: 3)** What is the optimal action selected by the agent when the battery level is low ? Justify with the above optimal value functions.

Solution:

Answers with $\boldsymbol{\gamma = 0.5}$ **or** $\boldsymbol{\gamma = 1}$ will be considered.

Considering $\boldsymbol{\gamma = 0.5}$:

For State *h*:

$$v_*(h) = \max\left[p(h\,|h,s)[r_s + \gamma v_*(h)] + p(m\,|h,s)[r_s + \gamma v_*(m)] + p(l\,|h,s)[r_s + \gamma v_*(l)]\right]$$

$$v_*(h) = \max\left[\alpha\,[r_s + \gamma v_*(h)] + \beta[r_s + \gamma v_*(m)] + (1-\alpha-\beta)[r_s + \gamma v_*(l)]\right]$$

$v_*(h) = \alpha\,[r_s + \gamma v_*(h)] + \beta[r_s + \gamma v_*(m)] + (1-\alpha-\beta)[r_s + \gamma v_*(l)]$

= 1/3 [3 r_s+1/2($v_*(h)$+ $v_*(m)$+ $v_*(l)$)]

5/6 $v_*(h)$ = r_s+ 1/6 ($v_*(m)$+ $v_*(l)$), keep r_s=3,

5 $v_*(h)$ = 18+ $v_*(m)$+ $v_*(l)$ - Eq 1

For State *m*:

$$v_*(m) = \max\left[p(m\,|m,s)[r_m + \gamma v_*(m)] + p(l\,|m,s)[r_m + \gamma v_*(l)]\right]$$

$v_*(m) = \delta[r_m + \gamma v_*(m)] + (1-\delta)[r_m + \gamma v_*(l)]$

= r_m + ¼ ($v_*(m)$+ $v_*(l)$), keep r_m=2,

¾ $v_*(m)$ = 2 + $v_*(l)$/4

$v_*(m) = \frac{v_*(l)+8}{3}$ - Eq 2

Using Eq. 1 and Eq. 2:

5 $v_*(h)$ = 18+$\frac{v_*(l)+8}{3}$+ $v_*(l)$

$v_*(h)$ = (62+4 $v_*(l)$)/15 - Eq 3

For State *l*:

$$v_*(l) = \max\begin{bmatrix} p(l\,|l,s)[r_l + \gamma v_*(l)] + p(h\,|l,s)[-5 + \gamma v_*(h)], \\ p(h\,|l,r)[0 + \gamma v_*(h)] \end{bmatrix}$$

$$= \max\begin{bmatrix} \mu[r_l + \gamma v_*(l)] + (1-\mu)[-5 + \gamma v_*(h)], \\ v_*(h)/2 \end{bmatrix}, \text{ keep } r_l\text{=1}$$

$$= \max\begin{bmatrix} 1/2\left[1 + \left(\frac{1}{2}\right)v_*(l)\right] + (1/2)[-5 + (1/2)v_*(h)], \\ v_*(h)/2 \end{bmatrix}$$

$$v_*(l) = \max\begin{bmatrix} (-8 + v_*(l) + v_*(h))/4 \\ v_*(h)/2 \end{bmatrix} \begin{matrix} \text{- Eq 4} \\ \text{- Eq 5} \end{matrix}$$

Rewriting Eq. 4, $v_*(l) = \frac{-8+v_*(h)}{3}$

With Eq. 5, $v_*(l) = v_*(h)/2$ ⟶ max

Solving Eq. 3 and Eq. 5, we get:

$v_*(h)$ = 62/13, $v_*(l)$= 31/13 and $v_*(m) = \frac{v_*(l)+8}{3} = 45/13$

Considering $\gamma = 1$:

For State *h*:

$$v_*(h) = \max\left[p(h\,|h,s)[r_s + \gamma v_*(h)] + p(m\,|h,s)[r_s + \gamma v_*(m)] + p(l\,|h,s)[r_s + \gamma v_*(l)]\right]$$

$$v_*(h) = \max\left[\alpha\,[r_s + \gamma v_*(h)] + \beta[r_s + \gamma v_*(m)] + (1-\alpha-\beta)[r_s + \gamma v_*(l)]\right]$$

$v_*(h) = \alpha\,[r_s + \gamma v_*(h)] + \beta[r_s + \gamma v_*(m)] + (1-\alpha-\beta)[r_s + \gamma v_*(l)]$

= 1/3 [3 r_s+($v_*(h)$+ $v_*(m)$+ $v_*(l)$)], keep r_s=3,

2 $v_*(h)$ = 9+ $v_*(m)$+ $v_*(l)$ - Eq 1

For State *m*:

$$v_*(m) = \max\left[p(m\,|m,s)[r_m + \gamma v_*(m)] + p(l\,|m,s)[r_m + \gamma v_*(l)]\right]$$

$v_*(m) = \delta[r_m + \gamma v_*(m)] + (1-\delta)[r_m + \gamma v_*(l)]$

= r_m + 1/2 ($v_*(m)$+ $v_*(l)$), keep r_m=2,

$v_*(m) = 4 + v_*(l)$- Eq 2

Using Eq. 1 and Eq. 2:

$2v_*(h)$ = 13+2 $v_*(l)$ - Eq 3

For State *l*:

$$v_*(l) = \max\begin{bmatrix} p(l\,|l,s)[r_l + \gamma v_*(l)] + p(h\,|l,s)[-5 + \gamma v_*(h)], \\ p(h\,|l,r)[0 + \gamma v_*(h)] \end{bmatrix}$$

$$= \max\begin{bmatrix} \mu[r_l + \gamma v_*(l)] + (1-\mu)[-5 + \gamma v_*(h)], \\ v_*(h)/2 \end{bmatrix}, \text{keep } r_l=1$$

$$= \max\begin{bmatrix} 1/2[1 + v_*(l)] + (1/2)[-5 + v_*(h)], \\ v_*(h) \end{bmatrix}$$

$$v_*(l) = \max\begin{bmatrix} (-4 + v_*(l) + v_*(h))/2 \\ v_*(h) \end{bmatrix}$$ - Eq 4

- Eq 5

Using (3), (4) or (3), (5), the problem is not solvable. You will get marks if your derivations till Eq. 3, 4, 5 are correct.

Q. (**Total Marks: 15)** Let us assume the MDP with three states: S1, S2 and S3 and three actions: A1, A2 and A3. Perform Q-learning with $\alpha = 0.5, \gamma = 0.5$ for the following six steps:

- S1, A1, S1, 2
- S1, A1, S1, -5
- S1, A1, S2, +10
- S2, A3, S3, -10
- S3, A2, S1, +20
- S1, A3, S1, **X**

Answer the following questions (provide the proper derivation for all the three questions).

- (**Marks: 5)** What is the value of Q(S1,A1) after six steps ?
- (**Marks: 5)** What is the value of Q(S3,A2) after six steps ?
- (**Marks: 5)** Find the minimum value of reward **X**, such that after six steps action **A3** is the optimal action to choose when you are in state **S1**.

Solution:

- Initial entries

Q	S1	S2	S3
A1	0	0	0
A2	0	0	0
A3	0	0	0

- First Step: S1, A1, S1, 2

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right].$$

$$Q(S1, A1) = 0 + 0.5[2 + 0.5 \max[0,0,0] - 0]$$

$$= 1$$

Q	S1	S2	S3
A1	1	0	0
A2	0	0	0
A3	0	0	0

- Initial entries

Q	S1	S2	S3
A1	1	0	0
A2	0	0	0
A3	0	0	0

- Second Step: S1, A1, S1, -5

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right].$$

$$Q(S1, A1) = 1 + 0.5[-5 + 0.5 \max[1,0,0] - 1]$$

$$= -1.75$$

Q	S1	S2	S3
A1	-1.75	0	0
A2	0	0	0
A3	0	0	0

- Initial entries

Q	S1	S2	S3
A1	-1.75	0	0
A2	0	0	0
A3	0	0	0

- Third Step: S1, A1, S2, 10

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right].$$

$$Q(S1, A1) = -1.75 + 0.5[10 + 0.5 \max[0,0,0] + 1.75]$$

= 4.125

Q	S1	S2	S3
A1	4.125	0	0
A2	0	0	0
A3	0	0	0

- Initial entries

Q	S1	S2	S3
A1	4.125	0	0
A2	0	0	0
A3	0	0	0

- Fourth Step: S2, A3, S3, -10

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right].$$

$$Q(S2, A3) = 0 + 0.5[-10 + 0.5 \max[0,0,0] + 0]$$

= -5

Q	S1	S2	S3
A1	4.125	0	0
A2	0	0	0
A3	0	-5	0

- Initial entries

Q	S1	S2	S3
A1	4.125	0	0
A2	0	0	0
A3	0	-5	0

- Fifth Step: S3, A2, S1, 20

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right].$$

$$Q(S3, A2) = 0 + 0.5[20 + 0.5 \max[4.125,0,0] + 0]$$

= 11.03

Q	S1	S2	S3
A1	4.125	0	0
A2	0	0	11.03
A3	0	-5	0

- Initial entries

Q	S1	S2	S3
A1	4.125	0	0
A2	0	0	11.03
A3	0	-5	0

- Sixth Step: S1, A3, S1, 5

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right].$$

$$Q(S1, A3) = 0 + 0.5[X + 0.5 \max[4.125,0,0] + 0]$$

Solving, Q(S1,A3) > Q(S1,A1)

Q(S1,A3) > 4.125

Gives X > 6.1875

Q. (**Total Marks: 18**) Let us consider the gird world MDP as shown in the figure below.

1,1	1,2	1,3	+5 1,4	1,5
2,1	2,2	2,3	2,4	+10 2,5
S 3,1	3,2	3,3	3,4	-5 3,5

The states are grid squares and they are represented by (i,j) where (i,j) is specified in the grid. The agent always start in state S represented by (3,1). There are three terminal states, (1,4) with reward +5, (2,5) with reward +10 and (3,5) with reward -5. Rewards are 0 in non-terminal states. The agent can move up, down, right and left. If the collision with wall happens, the agent stays in the same state. Let us assume that agent chooses the following episodes:

- First Episode: (3,1) – (3,2) – (3,3) – (3,4) – (3,5)
- Second Episode: (3,1) – (2,1) – (2,2) – (1,2) – (1,3) – (1,4)
- Third Episode: (3,1) – (2,1) – (2,2) – (2,3) – (2,4) – (2,5)
- Fourth Episode: (3,1) – (3,2) – (3,3) – (3,4) – (2,4) – (2,5)
- Fifth Episode: (3,1) – (2,1) – (2,2) – (2,3) – (2,4) – (3,4) – (3,5)

Given this MDP and these five episodes, answer the following questions:

a) **(Marks: 6)** What are the monte-Carlo estimates of state values (V(I,j)) for states (2,1), (2,4) and (1,3) ?

b) **(Marks: 2+10)** Assuming initial values as zero and learning rate of 0.2 and discount factor of 0.9, which states will have non-zero state values (V(I,j)) after five episodes using Temporal-Difference Learning TD(0) ? Find these values.

1,1	1,2	1,3	+5 1,4	1,5
2,1	2,2	2,3	2,4	+10 2,5
S 3,1	3,2	3,3	3,4	-5 3,5

First Episode: (3,1) – (3,2) – (3,3) – (3,4) – (3,5)
Second Episode: (3,1) – (2,1) – (2,2) – (1,2) – (1,3) – (1,4)
Third Episode: (3,1) – (2,1) – (2,2) – (2,3) – (2,4) – (2,5)
Fourth Episode: (3,1) – (3,2) – (3,3) – (3,4) – (2,4) – (2,5)
Fifth Episode: (3,1) – (2,1) – (2,2) – (2,3) – (2,4) – (3,4) – (3,5)

Monte Carlo Estimates:

- V(2,1) = (5+10-5)/3 = 10/3
- V(2,4) = (10+10-5)/3 = 15/3 = 5
- V(1,3) = 5

1,1	1,2	1,3	+5 1,4	1,5
2,1	2,2	2,3	2,4	+10 2,5
S 3,1	3,2	3,3	3,4	-5 3,5

First Episode: (3,1) – (3,2) – (3,3) – (3,4) – (3,5)
Second Episode: (3,1) – (2,1) – (2,2) – (1,2) – (1,3) – (1,4)
Third Episode: (3,1) – (2,1) – (2,2) – (2,3) – (2,4) – (2,5)
Fourth Episode: (3,1) – (3,2) – (3,3) – (3,4) – (2,4) – (2,5)
Fifth Episode: (3,1) – (2,1) – (2,2) – (2,3) – (2,4) – (3,4) – (3,5)

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

After first episode, all updates are zero, except:

- V(3,4) = 0 + 0.2(-5+0.9x0-0). = -1

After second episode, all updates are zero, except:

- V(1,3) = 0+0.2(5+0.9x0-0) = 1

After third episode, all updates are zero, except:

- V(2,4) = 0+0.2(10+0.9x0-0)= 2

After fourth episode, all updates are zero, except:

- V(3,3) = 0+0.2(0+0.9x(-1)-0) = -.18
- V(3,4) = -1+0.2(0+0.9x(2)-(-1)) = -0.44
- V(2,4) = 2+0.2(10+0.9x0-2) = 3.6

After fifth episode, all updates are zero, except:

- V(2,3) = 0+0.2(0+0.9x3.6-0) = 0.648
- V(2,4) = 3.6+0.2(0+0.9x(-0.44)-3.6) = 2.8008
- V(3,4) = -.44+0.2(-5+0.9x(0)-(-.44)) = -1.352

Q. (**Total Marks: 12)** Consider a multi-arm bandit with k=6 actions, denoted by 1, 2, 3, 4, 5 and 6. Let us assume that you are using $\epsilon - greedy$ action selection and the initial sequence of actions and rewards are:

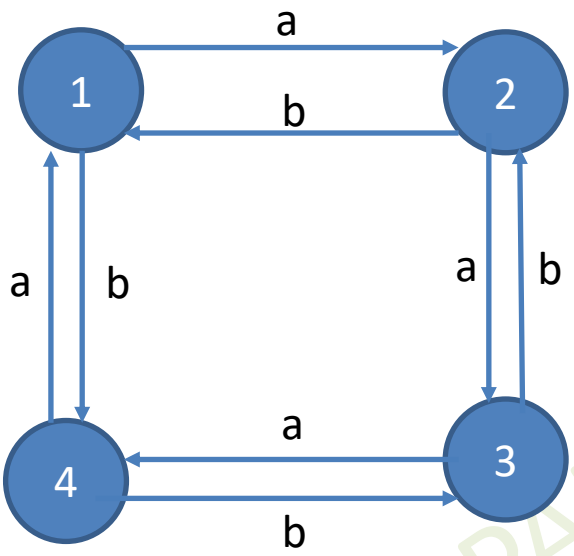
- Step 1: Action 1 with Reward -2
- Step 2: Action 1 with Reward 4
- Step 3: Action 1 with Reward 1
- Step 4: Action 3 with Reward 2
- Step 5: Action 3 with Reward 2.5
- Step 6: Action 4 with Reward -0.4
- Step 7: Action 3 with Reward -5

(a) (**Marks: 9)** Assuming initial estimates as 0, find the time steps where the action has been selected at random. Justify with the reason.
(b) (**Marks: 3)** Which action will be taken after step 7 and why ?

Solution:

- At Step 1, the action has been taken as random since all the initial estimates are zero.
- At Step 2, the action has been taken as random since the reward -2 is less than 0. So Action 1 should not be selected.
- At Step 4, the action has been taken as random since the average estimate for action 1 is (1+4-2)/3 = 1 which is more than the initial values of all the other states.
- At Step 6, the action has been taken as random since the average estimate for action 2 is (2+2.5)/2.= 2.25 which is more than the initial estimate for action 4 (which is zero).
- There are two possibilities. It may be a random action depending on ϵ or it may be action 1 since it is the only state with positive estimate.

Q. (**Total Marks: 10)** Let us consider the MDP shown in the figure below. There are 4 states and 2 actions. When the agent takes action "a" and moves from state 1 to state 2, it receives reward of +2. Similarly, if the agent takes action "a" at state 2 and moves from state 2 to state 3, it receives reward of -1. For any other transitions, the reward is 0. The learning rate is 0.5 and discount factor is 1. Assuming that agent starts at state 2 and initial Q-values are zero, find all the Q-values after actions (a,a,a,a,a,a) are applied with Sarsa.



$Q(2,a) = 0 + 0.5[-1 + 0 - 0] = -0.5$

$Q(3,a) = 0 + 0.5[0 + 0 - 0] = 0$

$Q(4,a) = 0 + 0.5[0 + 0 - 0] = 0$

$Q(1,a) = 0 + 0.5[2 + (-0.5) - 0] = 0.75$

$Q(2,a) = -0.5 + 0.5[-1 + (0) - (-0.5)] = -3/4$

$Q(3,a) = 0 + 0.5[0 + (0) - (0)] = 0$