CMPUT 366 A2

Yiyang Wang

TOTAL POINTS

93.5 / 115

QUESTION 1

Question 1 15 pts

1.1 Part a **2 / 2**

✓ **- 0 pts** Correct

1.2 Part b **2 / 2**

✓ **- 0 pts** Correct

1.3 Part c **2 / 2**

✓ **- 0 pts** Correct

1.4 Pard d **2 / 2**

✓ **- 0 pts** Correct

1.5 Part e **0.5 / 2**

✓ **- 1.5 pts Only has a final answer.**

1.6 Part f **5 / 5**

✓ **- 0 pts** Correct

QUESTION 2

Question 2 85 pts

2.1 Part a **6 / 6**

✓ **- 0 pts** Correct

2.2 Part b **6 / 6**

✓ **- 0 pts** Correct

2.3 Part c **6 / 6**

✓ **- 0 pts** Correct

2.4 Part d **9 / 9**

✓ **- 0 pts** Correct

2.5 Part e **12 / 12**

✓ **- 0 pts** Correct

2.6 Part f **9 / 9**

✓ **- 0 pts** Correct

2.7 Part g **0 / 9**

✓ **- 9 pts The provided answer is incorrect**

2.8 Part h **12 / 12**

✓ **- 0 pts** Correct

2.9 Part i **2 / 8**

✓ **- 2 pts Summation form written in terms of successor q-values**

✓ **- 4 pts Expected value form incorrect**

The back-up diagram stops at the action, so there should be no rewards or explicit return in the expectation. Also based on the back-up diagram, it should be written in terms of q_\pi of the current state, not the next state.

2.10 Part j **8 / 8**

✓ **- 0 pts** Correct

QUESTION 3

3 Question 3 (Bonus) 10 / 10

✓ **- 0 pts** Correct

QUESTION 4

4 Question 4 (Bonus) 0 / 5

✓ **- 2.5 pts episodic return expression incorrect**

✓ **- 2.5 pts continuing return expression incorrect**

QUESTION 5

5 Late Penalty 0 / 0

✓ **- 0 pts** correct

CMPUT 366 Assignment 2

Question 1

(a) $\overset{S_0}{X}$, $\overset{a_0}{\text{left}}$, $\overset{r_1}{0}$, $\overset{S_1}{X}$, $\overset{a_1}{\text{left}}$, $\overset{r_2}{0}$, X ...... (continues)

(b) $\overset{S_0}{X}$, $\overset{a_0}{\text{right}}$, $\overset{r_1}{-1}$, $\overset{S_1}{Y}$ $\overset{a_1}{\text{right}}$, $\overset{r_2}{+3}$, terminal state.

(c) $\gamma = 0.5$

$G_0 = R_1 + \gamma G_1 = R_1 + \gamma(R_2 + \gamma G_2) = -1 + (0.5)(3 + (0.5)(0)) = 0.5$

(d) $V_{\pi_1}(Y) = \mathbb{E}_\pi[G_0 | S = Y] = 3 + 0.5(0) = 3$

(e) $q_{\pi_1}(X, \text{left}) = 0$

(f) $V_{\pi_2}(X) = \sum_a \pi(a|s) \sum_{s',r} p(s', r | s, a)[r + \gamma v_\pi(s')]$

$= \frac{2}{3}[1 + 0.5(V_{\pi_2}(X)] + \frac{1}{3}[-1 + 0.5(V_{\pi_2}(Y)]$

$= \frac{2}{3} + \frac{1}{3} V_{\pi_2}(X) - \frac{1}{3} + \frac{1}{6} V_{\pi_2}(Y)$

$= \frac{2}{3} + \frac{1}{3} V_{\pi_2}(X) - \frac{1}{3} + \frac{1}{2}$

$\frac{2}{3} V_{\pi_2}(X) = \cancel{-\frac{1}{6}}\ \frac{5}{6}$

$V_{\pi_2}(X) = \frac{5}{4}$

Question 2

(a) 3.1

Example 1: Brightness controller (For electrical devices)

states: the ~~dif~~ brightness difference between device and outside environmen

actions: lighter or darker the device

rewards: agent sensor the brightness of outside, set a best difference value. Then compare the value with ~~outeer~~ current difference, positive reward if the difference between [current difference] and value is smaller.

limitation: The agent doesn't consider power saving.

Example 2: Automatic package sorting maching

states: remainder items in storage

action: sort an item to its elements

rewards: (+1) if correctly sort an item, (-1) if it put the item in wrong elements.

Example 3: Self-dring car

states: the distance from your destination

action: choose an direction at crossroad

rewards: The shorter path, the higher reward

limitations: The car may stuck in traffic jam, waste time

Scanned with CamScanner

(b) 3.7 maze running

We only give +1 reward for a successful escapation and 0 reward at all other time. As a result, the agent can keep exploring the route and find a way out without reward loss. And the agent didn't learn and shows no improvement. We have not communicated to the agent about shortest path ~~and~~ or fastest time. We want it to find shortest path and use less time to escape. From this idea, we can set a negative reward (eg. -1, -2) to the agent, for every timestep it explores in maze, then we can achieve our goal.

c) 3.8

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \dots$$
$$= R_{t+1} + \gamma G_{t+1} \qquad (3.9 \text{ SB})$$

$G_5 = 0$

$G_4 = R_5 + 0.5 G_5 = 2 + 0 = 2$

$G_3 = R_4 + 0.5 G_4 = 3 + 1 = 4$

$G_2 = R_3 + 0.5 G_3 = 6 + 2 = 8$

$G_1 = R_2 + 0.5 G_2 = 2 + 4 = 6$

$G_0 = R_1 + 0.5 G_1 = -1 + 3 = 2$

d) 3.9

from (SB.3.8) $\quad \gamma = 0.9 \quad R_1 = 2 \quad 7s \quad G_1, G_0 ?$

$G_1 = (0.9)(7) + (0.9^2)(7) + (0.9^3)(7) + \cdots$

Geometric $\Rightarrow 7 \dfrac{1}{1-0.9} = 70$

$G_0 = R_1 + \gamma G_1 = 2 + 0.9(70) = 65$

(e) 3.14

$V_\pi(s) \doteq \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')]$ $\quad$ north, south, east, west

For all 4 directions, we have $p = \frac{1}{4}$, $r = 0$

Then we input value, get:

	2.3	
0.7	0.7	0.4
	-0.4	

north $\frac{1}{4}(0 + 0.9(2.3))$

east $\frac{1}{4}(0 + 0.9(0.4))$

south $\frac{1}{4}(0 + 0.9(-0.4))$

west $\frac{1}{4}(0 + 0.9(0.7))$

$+0.675 \approx +0.7$

Sum up, we get +0.7, holds for center.

[

(f) 3.15

The signs of rewards are not important, only the intervals between them matters. We prove in following that adding a constant does not have effect. (Where adding a constant to rewards may change signs).

using SB 3.8 $\Rightarrow$ $G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

adding c

$$\Rightarrow \quad G_t = R_{t+1} + c + \gamma(R_{t+2} + c) + \dots$$

$$= \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c)$$

seperate

$$\Rightarrow \sum_{k=0}^{\infty} (\gamma^k R_{t+k+1} + \gamma^k c)$$

$$= \sum_{k=0}^{\infty} (\gamma^k R_{t+k+1}) + \sum_{k=0}^{\infty} \gamma^k \cdot c$$

$$\overset{\text{Geometric}}{=} \sum_{k=0}^{\infty} (\gamma^k R_{t+k+1}) + \frac{c}{1-\gamma}$$

So V_c is $\frac{c}{1-\gamma}$

(g) 3.16

This would have no effect, it leaves the task unchanged as above.
We can just regard the continuing task as several episodic tasks.
We just add these tasks together.
For example, add a constant c to maze running rewards.
$\frac{c}{1-\gamma}$ is added for each episodic task, and it won't change the variance of ~~our~~ the performance. Add them together, V_c sum up, and it extend learning time for ~~two~~ larger process.
And we still ~~achive~~ achieve goals.

(h) 3.17 $\quad q_\pi(s', a')$

$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t | S_t = s, A_t = a]$

$= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a]$ by 3.9

$= \sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')]$ by 3.14

$= \sum_{s',r} p(s',r|s,a)[r + \gamma \sum_{a'} \pi(a'|s') q_\pi(s',a')]$ 3.14 with action values

(i) 3.18

prob. $\pi(a|s)$ $\quad v_\pi(s)$ $\quad q_\pi(s,a)$ $\quad S_t = s.$ $\quad a_t = a$.

$v_\pi(s) = \mathbb{E}_\pi[R_t | S_t = s]$

$= \sum_a \mathbb{E}_\pi[R_t | S_t = s, a_t = a] \pi(s,a)$

∴ no expected value : ↓ $q_\pi(s,a)$

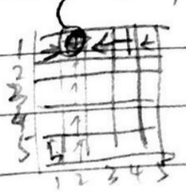
$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a)$

(j) 3.24

$v_*(s) = \max_{a \in A(s)} q_{\pi_*}(s,a)$ (SB 3.9)

$= \max_a \sum_{s',r} p(s',r|s,a)[r + \gamma v_*(s')]$

(1,2)



The max $q_{\pi_*}(s,a)$. From (5,2) to (1,2)

in position (1,2) always optimal

$v_*(S=(1,2)) = \cancel{+}10 + \underbrace{0+0+0+0+0}_{5\ 0s'} + \gamma^5 10 + \cdots$

every 5 0s we have a value.

So

$V^*(s) = 10 + \gamma^5 10 + \gamma^{10} 10 + \gamma^{15} 10 + \cdots$

$= \sum_{k=0}^{\infty} \gamma^{5k} 10$

Geometric $\longrightarrow = 10 \frac{1}{1-\gamma^5}$

input $\gamma = 0.9$

$\frac{10}{1-\gamma^5} = \frac{10}{1-0.9^5} = 24.419428$

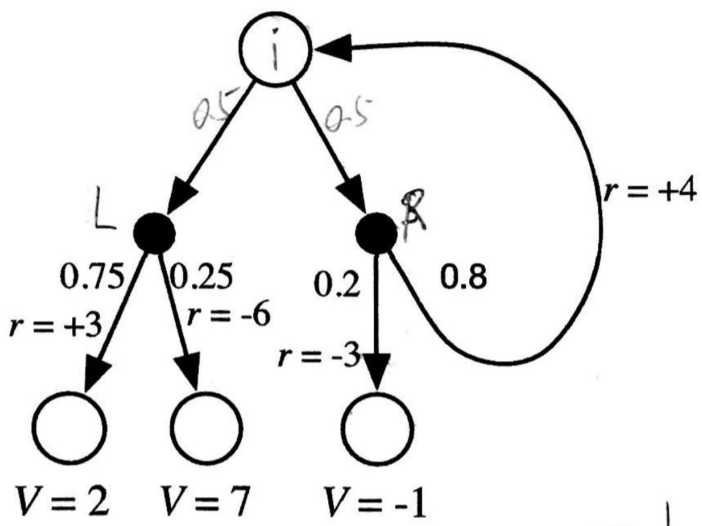
≈ 24.419

Bonus Questions [total 15 points available]. There are **two** bonus questions.

Question 3: Trajectories, returns, and values (**10 Bonus points**)

Consider the following fragment of an MDP graph. The fractional numbers indicate the world's transition probabilities and the whole numbers indicate the expected rewards. The three numbers at the bottom indicate what you can take to be the value of the corresponding states. The discount is 0.8. What is the value of the top node for the equiprobable random policy (all actions equally likely) and for the optimal policy? Show your work.

$\gamma = 0.8$



$v_\pi = 4.25735$

≈ 4.26

$v_* = 6.78$

$V_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p\,[r + \gamma V_\pi(s')]$

$V_\pi(L) = (0.5)(0.75)(+3 + 0.8(2))$

$+ (0.5)(0.25)(-6 + 0.8(7))$

$= 1.675$

$V_\pi(R) = (0.5)(0.2)(-3 + 0.8(-1))$

$+ (0.5)(0.8)(+4 + 0.8\,V(\pi))$

$= -0.38 + 1.6 + 0.32\,V(\pi)$

$= 1.22 + 0.32\,V(\pi)$

$V(\pi) = V_\pi(L) + V_\pi(R)$

$= 1.675 + 1.22 + 0.32\,V(\pi)$

$\cancel{V(\pi)}\ V_\pi = 4.25735$

$V_*^L = (0.75)(3 + 0.8(2))$

$+ (0.25)(-6 + 0.8(7))$

$= 3.35$

$V_*^R = (0.2)(-3 + 0.8(-1))$

$+ (0.8)(4 + 0.8\,V_*^R)$

$V_*^R \approx 6.78$

$V_* = \max_{a \in A(s)} q_{\pi_*}(s, a)$

$= V_*^R$

$= 6.78$

Scanned with CamScanner

Question 4 [5 bonus points]. Complete Exercise 3.6 (episodic pole balancing). See SB textbook, second ed.

0. -1

failure

not fail return 0

fail return -1

for continuing: return $-\gamma^k$, for k steps until a failure occurs at each time

for an episodic task with γ:

as most conditions are same,

it still return $-\gamma^k$.

Difference: for the episodic task, there are finite time steps and terms.

So the value of k, should be [# of time steps before failure -1]

Thus, $\underset{\text{episodic}}{-\gamma^{k'}} \le \underset{\text{continuing}}{-\gamma^{k}}$

Scanned with CamScanner