ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

SEMESTER FINAL EXAMINATION **WINTER SEMESTER, 2020-2021**

DURATION: 1 Hour 30 Minutes **FULL MARKS: 75**

CSE 4711: Artificial Intelligence

Programmable calculators are not allowed.

There are **3 (three)** questions. Answer all **3 (three)** of them.

Figures in the right margin indicate marks of each question.

The square brackets on the start of each question denotes the corresponding CO and PO.

---

1. Consider that an agent is moving in the Gridworld shown in Figure **??**. The only action available in cells A, F, G, and H is `Exit` with rewards $2X$, X, 0, and 0, respectively. Here, X = The last digit of your student ID + 1.

   When the agent exits from a cell, it gets the reward for that cell as specified. From the other cells, the agent can only take the action `Left` or `Right`, which results in the agent moving to the immediate left or right cell, respectively.

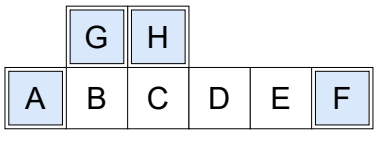

Figure 1: Gridworld for Question 1

If the agent is in cell B or C, and it takes a `Left` or `Right` action, it might fail sometimes. The agent will move in the desired direction with probability p, and it will fail and move up with probability $1 - p$. If the agent is in any other cell ($A, D, E, F, G,$ or H), the action will always be successful.

Assume that the agent plays optimally, there is no living reward/penalty, and the discount factor is $\gamma \in [0, 1]$.

a) [CO1, PO1] For each of the following policies, determine the value of each non-terminal state. 6×2

i. $\pi_{\text{Right}}(S) = \begin{cases} \texttt{Exit}, & \text{if } S \text{ is a Terminal State} \\ \texttt{Right}, & \text{otherwise} \end{cases}$

Solution:

- $B \rightarrow X\gamma^4 p^2$
- $C \rightarrow X\gamma^3 p$
- $D \rightarrow X\gamma^2$
- $E \rightarrow X\gamma$

Rubric:

- 0.5 points for each correct integer
- 0.5 points for each correct discount
- 0.5 points for each correct probability

ii. $\pi_{\text{Left}}(S) = \begin{cases} \texttt{Exit}, & \text{if } S \text{ is a Terminal State} \\ \texttt{Left}, & \text{otherwise} \end{cases}$

Solution:

- $B \to 2X\gamma p$
- $C \to 2X\gamma^2 p^2$
- $D \to 2X\gamma^3 p^2$
- $E \to 2X\gamma^4 p^2$

Rubric:

- 0.5 points for each correct integer
- 0.5 points for each correct discount
- 0.5 points for each correct probability

b) [CO3, PO2, PO3] For what range of value for p will the agent choose π_{Left} over π_{Right}? 7

Solution: For π_{Left} to be the optimal policy, for each of the four cells, π_{Left} needs to be greater than π_{Right}. Intuitively, the farther Right we are, the higher the value of moving Right, and the lower the value of moving Left (since the discount factor penalizes far-away rewards).
Thus if moving Left is the optimal policy, we want to focus our attention on the Rightmost cell, E. We need:

$$\pi_{\text{Left}}(E) > \pi_{\text{Right}}(E)$$
$$2X\gamma^4 p^2 > X\gamma$$
$$p > \frac{1}{\sqrt{2\gamma^3}}$$

Considering the probability range $[0, 1]$, we get: $p \in \left(\frac{1}{\sqrt{2\gamma^3}}, 1\right]$.
However, $\lim_{\gamma \to 0} \frac{1}{\sqrt{2\gamma^3}} = \infty$. So the policy will not be optimal unless $\gamma > 0$.

Rubric:

- 2 points for describing the correct condition for π_{Left} being optimal.
- 2 points for the correct lower limit.
- 2 points for the correct upper limit.
- 1 point for the impossible condition.

c) [CO3, PO2, PO3] For what range of value for p will the agent choose π_{Right} over π_{Left}? 6

Solution: Following the similar logic from the previous question, moving Right to be the

optimal policy, we want to focus our attention on the Leftmost cell, B. We get:

$$\pi_{\text{Right}}(B) > \pi_{\text{Left}}(B)$$
$$X\gamma^4 p^2 > 10\gamma p$$
$$p > \frac{2}{\gamma^3}$$

However, since γ ranges from 0 to 1, the right side of the expression ranges from 2 to ∞, which means p (a probability, and thus bounded by 1) has no valid value. So the range is $\emptyset$.

Rubric:

- 2 points for describing the correct condition for π_{Right} being optimal.
- 2 points for the correct lower limit.
- 2 points for showing that it's not possible.

2. a) [CO1, PO1] Two students, Rabbi and Sajal, are on their way to their class. The class is starting right now, so both of them are running late. It takes 40 minutes to get to the classroom by walking, and only 10 minutes to get there by bus. Sajal prefers to go on a bus given the choice, so the bus incurs an additional utility bonus of 5. However, to get on the bus, they need to wait at the bus stand. Due to traffic congestion, the bus might come 20, 40, or 60 minutes, each with equal probability. For each of the following utility functions, determine whether that person should walk or take the bus: 8×2

i. Rabbi does not like being late. For him, the utility function for being late by time t is:

$$U_R(t) = \begin{cases} 0 & t \leq 0 \\ -2^{t/5} & t > 0 \end{cases}$$

Solution:

$$\begin{aligned} EU(walk) &= -2^{40/5} \\ &= -256 \\ EU(bus) &= \frac{1}{3}\left(-2^{(20+10)/5} - 2^{(40+10)/5} - 2^{(60+10)/5}\right) \\ &= -5824 \end{aligned}$$

Rabbi should walk.

Rubric:

- 3 points for expected utility of walking
- 3 points for expected utility of bus
- 2 point for the correct choice

ii. Once Sajal is late, it does not matter to him how late he is. For him, the utility function for

being late by time t is:

$$U_S(t) = \begin{cases} 0 & t \leq 0 \\ -50 & t > 0 \end{cases}$$

Solution:

$$\begin{aligned} EU(walk) &= -50 \\ EU(bus) &= -50 + 5 \\ &= -45 \end{aligned}$$

Sajal should take the bus.

Rubric:

- 3 points for expected utility of walking
- 3 points for expected utility of bus
- 2 point for the correct choice

b) [CO3, PO2, PO3] Assume that you have designed a deep learning model for determining whether a tomato leaf is infected by disease or not. Your system takes tomato leaf images as input and outputs "Positive" denoting that the leaf is infected, and "Negative" otherwise. After testing your model with a dataset, you have generated the confusion matrix to see how your model performs. You see that in 97% of the cases your model identifies a leaf as positive when the leaf is actually infected. And in 95% of the cases, your model identifies a leaf as negative, when the leaf is not infected. Suppose that, we know that 0.5% of the tomato leaf images in your dataset were infected by diseases. Explain why you should or should not trust the prediction of your system. 9

Solution: Let, I be the random variable denoting whether the leaf is infected or not. Let, P be the random variable denoting the output of the model.

Using Bayes' rule, we can write:

$$\begin{aligned} P(+i|+p) &= \frac{P(+p|+i)P(+i)}{P(+p)} \\ &= \frac{P(+p|+i)P(+i)}{P(+p|+i)P(+i) + P(+p|-i)P(-i)} \\ &= \frac{0.97 \times 0.005}{0.97 \times 0.005 + 0.05 \times 0.995} \\ &= 0.089 \end{aligned}$$

Given the model predicts a leaf to be infected, only 8.9% of the time the leaf will be actually infected! For example, if 1000 leaves are tested, it is expected that 995 of them are healthy leaves and 5 of them are infected by disease. From the 995 healthy leaves your model will predict $0.05 \times 995 \approx 50$ will be classified as "Positive". From the 5 infected leaves, $0.95 \times 5 \approx 5$ will be classified as "Positive". That means, out of the 55 positive results, only 5 are genuine.

Rubric:

- 2 points for the correct equation

- 4 points for the analysis
- 3 points for the conclusion
- Any correct argument (positive or negative) as conclusion will be accepted

3. [CO2, PO4] Consider that Pacman is moving in an unknown grid with only two cells X and Y. From each state, the Pacman can either move Left or Right. According to some policy π, Pacman got the following sequence of actions and rewards: 25

t	s_t	a_0	s_{t+1}	r_t
0	X	$Right$	Y	2
1	Y	$Right$	Y	-4
2	Y	$Left$	Y	0
3	Y	$Left$	X	3
4	X	$Left$	X	-1

Table 1: Samples table for Question 3

Model the scenario as Reinforcement Learning problem. Mention the states, actions, and Q-values for each state-action pair after performing sample-based Q-Learning. Recall the update function for sample-based Q-Learning is:

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha(r_t + \gamma \max_{a'} Q(s_{t+1}, a'))$$

Initially all Q-values are set to 0. The discount factor $\gamma = 0.5$ and the learning rate $\alpha = 0.5$.

Solution:

- States: $\{X, Y\}$
- Actions: $\{Left, Right\}$

At $t = 0$, we get:

$$\begin{aligned} Q(X, Right) &\leftarrow (1-\alpha)Q(X, Right) + \alpha(r + \gamma \max_{a'} Q(Y, a') \\ &\leftarrow (1-0.5) \times 0 + 0.5 \times (2 + 0.5 \times 0) \\ &\leftarrow 1 \end{aligned}$$

At $t = 1$, we get:

$$\begin{aligned} Q(Y, Right) &\leftarrow (1-\alpha)Q(Y, Right) + \alpha(r + \gamma \max_{a'} Q(Y, a')) \\ &\leftarrow (1-0.5) \times 0 + 0.5 \times (-4 + 0.5 \times 0) \\ &\leftarrow -2 \end{aligned}$$

At $t = 2$, we get:

$$\begin{aligned} Q(Y, Left) &\leftarrow (1-\alpha)Q(Y, Left) + \alpha(r + \gamma \max_{a'} Q(Y, a')) \\ &\leftarrow (1 - 0.5) \times 0 + 0.5 \times (0 + 0.5 \times 0) \\ &\leftarrow 0 \end{aligned}$$

At $t = 3$, we get:

$$\begin{aligned} Q(Y, Left) &\leftarrow (1-\alpha)Q(Y, Left) + \alpha(r + \gamma \max_{a'} Q(X, a')) \\ &\leftarrow (1 - 0.5) \times 0 + 0.5 \times (3 + 0.5 \times 1) \\ &\leftarrow \frac{7}{4} \end{aligned}$$

At $t = 4$, we get:

$$\begin{aligned} Q(X, Left) &\leftarrow (1-\alpha)Q(X, Left) + \alpha(r + \gamma \max_{a'} Q(X, a')) \\ &\leftarrow (1 - 0.5) \times 0 + 0.5 \times (-1 + 0.5 \times 1) \\ &\leftarrow -\frac{1}{4} \end{aligned}$$

So the final values are:

- $Q(X, Left) = -\frac{1}{4}$
- $Q(X, Right) = 1$
- $Q(Y, Left) = \frac{7}{4}$
- $Q(Y, Right) = -2$

Rubric:

- 1 point for each item from states and actions
- 4 points for each update
- 0.25 points for each correct value of state-action pair