Indian Institute of Technology Kharagpur
Department of Computer Science and Engineering

Reinforcement Learning (CS60077) **Autumn Semester, 2022-2023**

Mid-Semester Examination **29-Sep-2022 (Thursday), 09:00–11:00** **Maximum Marks: 60**

Instructions:

- Write your answers in the answer booklet provided to you in the examination hall.
- There are a total of FOUR questions. Marks are indicated in parentheses. All questions are compulsory. Write the answers for all parts of a question together.
- Be brief and precise. Organize your work, in a reasonably neat and coherent way. Work scattered all across the answer script without a clear ordering will receive very little marks.
- If you use any theorem / result / formula covered in the class, just mention it and do not elaborate and/or prove (if not explicitly asked to do so).
- Write all the derivations / proofs / deductions in mathematically and/or logically precise language. Unclear and/or dubious statements would be severely penalized.
- Mysterious or unsupported answers will not receive full marks. A correct answer, unsupported by calculations, explanation, will receive no marks; an incorrect answer supported by substantially correct calculations and explanations may receive partial marks.

– The question paper starts from the next page –

Q1. [16 Marks] Consider a 4×4 Grid-world problem where the goal is to reach either the *top-left corner* or the *bottom-right corner* (refer to Table 1). The agent can choose from four actions $\langle$`up`, `down`, `left`, `right`$\rangle$ which deterministically cause the corresponding state transitions, except that actions that would take the agent off the grid leave the state unchanged. We model this as an undiscounted, episodic task, where the reward is -1 for all transitions. Suppose that the agent follows the equiprobable random policy, π.

(Hint: The Bellman equation must hold for every state.)

Table 1: 4×4 Grid-world

0		-20	-22
-14	-18	-20	
	-20	-18	-14
-22	-20		0

(a) Given above is the partial value function for this problem. Calculate respectively, the missing values in the first and second row? **(3 + 3)**

(b) What are the respective values of $q_\pi(s_1, \texttt{down})$ and $q_\pi(s_2, \texttt{down})$ given that s_1 is the last cell in the third row (value is -14) and s_2 is the last cell in the second row? **(2 + 2)**

(c) We defined the operator $L_\pi : \mathcal{V} \to \mathcal{V}$ as $L_\pi v = r_\pi + \gamma P_\pi v$, for all $v \in \mathcal{V}$. Given a value, $v \in \mathcal{V}$, let $L_\pi v = v'$. Then can we conclude the following? Give justifications to your answer. **(1.5 + 1.5)**

(i) Is $v = v'$? **(ii)** Is $||L_\pi v - L_\pi v'|| \leq \lambda ||v - v'||,\ 0 \leq \lambda < 1$?

(d) In a particular grid-world example, rewards are positive for goals, negative for running into the edge of the world, and zero the rest of the time. Prove, using the discounted return equation,

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

that adding a constant C to all the rewards adds a constant, K, to the values of all states, and thus does not affect the relative values of any states under any policies. Derive the value of K in terms of C and γ. **(3)**

Solution:

(a) Recall Bellman Expectation Equation,

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s'} \mathbb{P}[s'|s, a]\big[r + v_\pi(s')\big]$$

Let the state value in the grid at first row second column be v_1. Then, applying Bellman Expectation Equation, we have:

$$\begin{aligned} v_1 &= 0.25 \times \big[(-1+0) + (-1-20) + (-1-18) + (-1+v_1)\big] \\ &= 0.25 \times \big[-42 + v_1\big] \\ \implies v_1 &= -14 \end{aligned}$$

Let the state value in the grid at second row fourth column be v_2. Then, applying Bellman Expectation Equation, we have:

$$\begin{aligned} v_2 &= 0.25 \times \big[(-1+20) + (-1-22) + (-1-14) + (-1+v_2)\big] \\ &= 0.25 \times \big[-60 + v_2\big] \\ \implies v_2 &= -20 \end{aligned}$$

(b) For s_1, we have: $q_\pi(s_1, \texttt{down}) = \sum_{s'} \mathbb{P}[s'|s_1, \texttt{down}]\big[r + v_\pi(s')\big]$

Therefore, $q_\pi(s_1, \texttt{down}) = -1 + 0 = -1.$

Similarly, for s_1, we have: $q_\pi(s_2, \texttt{down}) = \sum_{s'} \mathbb{P}[s'|s_2, \texttt{down}]\big[r + v_\pi(s')\big]$

Therefore, $q_\pi(s_2, \texttt{down}) = -1 - 14 = -15.$

(c) **(i)** *No.* For v to be equal to v', it has to be a fixed point.

(ii) *Yes.* Because, L_π is a contraction.

(d) Assume that the grid-world problem is a continuing task. For some policy π and state s, the value function can be give as, $v_\pi(s) = \mathbb{E}_\pi\big[G_t \mid s_t = s\big].$

Using the discounted reward equation, we have, $v_\pi(s) = \mathbb{E}\big[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid s_t = s\big].$

Adding a constant C to all rewards, we have,

$$\begin{aligned} v_\pi(s) &= \mathbb{E}\big[\sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + C) \mid s_t = s\big] \\ &= \mathbb{E}\big[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + C\sum_{k=0}^{\infty} \gamma^k \mid s_t = s\big] \\ &= v_\pi(s) + \frac{C}{1-\gamma} \end{aligned}$$

We see that adding a constant C to all rewards does not affect the relative values of any states under any policies. Here, $K = \frac{C}{1-\gamma}$

Q2. [14 Marks] Consider a finite, episodic and undiscounted Markov Decision Process (MDP) with states P and Q apart from the terminal state. Let the following two samples are observed when a Monte-Carlo (MC) evaluation is being carried out.

- $(P, +3) \rightarrow (P, +2) \rightarrow (Q, -4) \rightarrow (P, +4) \rightarrow (Q, -3)$
- $(Q, -2) \rightarrow (P, +3) \rightarrow (Q, -3)$

For example, a sample such as, $(Q, -2) \rightarrow (P, +3) \rightarrow (Q, -3)$, means that the episode starts at Q then goes to P again, then goes to Q and then terminates. On the way, the agent gets rewards of -2, $+3$ and -3, respectively.

(a) Estimate the state value of both P and Q using *first-visit* Monte-Carlo evaluation. **(2)**

(b) Estimate the state value of both P and Q using *every-visit* Monte-Carlo evaluation. **(3)**

(c) Construct a Markov model that best explains the observations given in the question. **(4)**

(d) What would be the value function estimate of P and Q (call it as $v(P)$ and $v(Q)$, respectively) if batch TD(0) (Temporal Difference learning) were applied to the above transaction data? **(3)**

(e) In solving an episodic problem we observe that all trajectories from the start state to the goal state pass through a particular state exactly twice. In such a scenario, is it preferable to use first-visit or every-visit Monte-Carlo for evaluating the policy? Choose the appropriate answer(s) with a brief justification. **(2)**

(i) first-visit Monte-Carlo
(ii) every-visit Monte-Carlo
(iii) every-visit Monte-Carlo with exploring starts
(iv) neither, as there are issues with the problem itself

Solution:

(a) The first visit return of state P in the first trajectory is $3 + 2 - 4 + 4 - 3 = 2$. The same for the second trajectory is $3 - 3 = 0$.

So, the first-visit MC estimate of the state-value of P is $\frac{2+0}{2} = 1$.

The first visit return of state Q in the first trajectory is $-4 + 4 - 3 = -3$. The same for the second trajectory is $-2 + 3 - 3 = -2$.

So, the first-visit MC estimate of the state-value of Q is $\frac{-3-2}{2} = -\frac{5}{2}$.

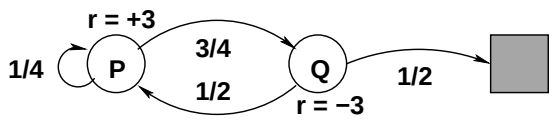
(b) There are 3 visits of the state P in the first trajectory with the corresponding returns of $3+2-4+4-3 = 2$, $2 - 4 + 4 - 3 = -1$ and $4 - 3 = 1$. There is only one visit of the state P in the second trajectory. The return for the second trajectory thus, is $3 - 3 = 0$.

So, the every-visit MC estimate of the state-value of P is $\frac{2-1+1+0}{4} = \frac{1}{2}$.

There are 2 visits of the state Q in the first trajectory with the corresponding returns of $-4+4-3 = -3$ and -3. There are 2 visits of the state Q in the second trajectory. The returns for the second trajectory thus, are $-2 + 3 - 3 = -2$ and -3.

So, the every-visit MC estimate of the state-value of Q is $\frac{-3-3-2-3}{4} = -\frac{11}{4}$.

(c) The Markov model is given as follows.



(d) We can solve Bellman equations directly from the above Markov model to get,

$$\begin{aligned} v(P) &= 3 + \frac{1}{4} \times v(P) + \frac{3}{4} \times v(Q) \\ v(Q) &= -3 + \frac{1}{2} \times v(P) \end{aligned}$$

$$\begin{aligned} \implies v(P) &= 2 \\ v(Q) &= -2 \end{aligned}$$

(e) Correct Choice: **(iv)**

Justification: A state having to be visited exactly twice in any trajectory from the start state to the goal state indicates that the problem environment does not follow the Markov property.

Q3. [16 Marks] A rat is involved in an experiment. It experiences one episode. At the first step it hears a bell. At the second step it sees a light. At the third step it both hears a bell and sees a light. It then receives some food, worth $+1$ reward, and the episode terminates on the fourth step. All other rewards were zero. The experiment is undiscounted (i.e., $\gamma = 1$).

We may represent the rat's state s by a vector of two binary features, $bell(s) \in \{0, 1\}$ and $light(s) \in \{0, 1\}$. So, the sequence of feature vectors corresponding to the four steps of this episode can be expressed as, $[1, 0]^T$, $[0, 1]^T$, $[1, 1]^T$ and $[0, 0]^T$.

(a) Approximate the state-value function by a linear combination of these features with two parameters: $\big[b \times bell(s) + l \times light(s)\big]$. If $b = 2$ and $l = -2$, then write down the sequence of approximate values corresponding to this episode. **(2)**

(b) Write down the sequence of λ-returns G_t^λ $(1 \leq t \leq 3)$ corresponding to this episode, for $\lambda = \frac{1}{2}$ and $b = 2,\ l = -2$. Clearly show the detailed evaluations. **(6)**

(c) Using the forward-view TD(λ) algorithm and your linear function approximator, what are the sequence of updates to weight b? What is the total update to weight b? Use $\lambda = \frac{1}{2},\ \gamma = 1,\ \alpha = \frac{1}{2}$ and start with $b = 2,\ l = -2$. **(3)**

(d) Using linear value function approximation, write down the sequence of TD(λ) accumulating eligibility trace e_t corresponding to the bell, using $\lambda = \frac{1}{2},\ \gamma = 1$. **(2)**

(e) Using the backward-view TD(λ) algorithm and your linear function approximator, what are the sequence of updates to weight b? (Use offline updates, i.e., do not actually change your weights, just accumulate your updates). What is the total update to weight b? Use $\lambda = \frac{1}{2},\ \gamma = 1,\ \alpha = \frac{1}{2}$ and start with $b = 2,\ l = -2$. **(3)**

Solution:

(a) Figure 1 expresses the activity of the presented episode.

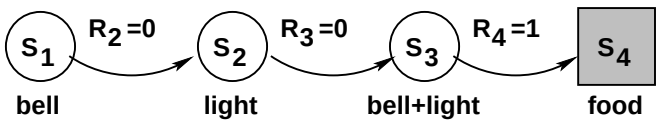


Figure 1: Episode Activity

Given, $b = 2$ and $l = -2$. Let the feature vectors corresponding to each episode are given as,

$$\phi(s_1) = [1, 0]^T, \quad \phi(s_2) = [0, 1]^T, \quad \phi(s_3) = [1, 1]^T, \quad \phi(s_4) = [0, 0]^T.$$

Therefore, the sequence of approximate values corresponding to each step in this episode is denoted as,

$$\begin{aligned} V(s_1) &= [b, l].\phi(s_1) = [2, -2].[1, 0]^T = 2 \\ V(s_2) &= [b, l].\phi(s_2) = [2, -2].[0, 1]^T = -2 \\ V(s_3) &= [b, l].\phi(s_3) = [2, -2].[1, 1]^T = 0 \\ V(s_4) &= [b, l].\phi(s_4) = [2, -2].[0, 0]^T = 0 \ \text{(for terminal state)} \end{aligned}$$

It may also be noted that, $R_2 = 0$, $R_3 = 0$, but $R_4 = +1$.

(b) The n-step return and λ-return G_t^λ are given as,

$$\begin{aligned} G_t^{(n)} &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n V(s_{t+n}) \\ G_t^\lambda &= (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} . G_t^{(n)} \end{aligned}$$

where, $G_t^{(\infty)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-1} R_T$, with multiplication factor being λ^{T-t-1}

Here, the calculation for $G_1^{(\lambda)}$ goes as follows. In our case, $T = 4$ and hence $R_T = R_4 = +1$.

$$\begin{aligned} G_1^{(1)} &= R_2 + \gamma V(s_2) = 0 + 1 \times (-2) = -2 \\ G_1^{(2)} &= R_2 + \gamma R_3 + \gamma^2 V(s_3) = 0 + 1 \times 0 + 1 \times 0 = 0 \\ G_1^{(\infty)} &= R_2 + \gamma R_3 + \gamma^2 R_4 = 0 + 1 \times 0 + 1 \times 1 = 1 \\ \therefore\ G_1^\lambda &= (1 - \frac{1}{2}) \times \Big[(\frac{1}{2})^0 G_1^{(1)} + (\frac{1}{2})^1 G_1^{(2)} \Big] + (\frac{1}{2})^{4-1-1} G_1^{(\infty)} = \frac{1}{2} \Big[1 \times (-2) + \frac{1}{2} \times 0 + \frac{1}{2} \times 1 \Big] = -\frac{3}{4} \end{aligned}$$

Similarly, the calculation for $G_2^{(\lambda)}$ goes as follows.

$$\begin{aligned} G_2^{(1)} &= R_3 + \gamma V(s_3) = 0 + 1 \times 0 = 0 \\ G_2^{(\infty)} &= R_3 + \gamma R_4 = 0 + 1 \times 1 = 1 \\ \therefore\ G_2^\lambda &= (1 - \frac{1}{2}) \times \Big[(\frac{1}{2})^0 G_1^{(1)} \Big] + (\frac{1}{2})^{4-2-1} G_2^{(\infty)} = \frac{1}{2} \Big[1 \times 0 + 1 \times 1 \Big] = \frac{1}{2} \end{aligned}$$

And, the calculation for $G_3^{(\lambda)}$ goes as follows.

$$\begin{aligned} G_3^{(\infty)} &= R_4 = 1 \\ \therefore\ G_3^\lambda &= (\frac{1}{2})^{4-3-1} G_3^{(\infty)} = (\frac{1}{2})^0 \times 1 = 1 \end{aligned}$$

(c) The sequence of updates to weight b is given as,

$$\begin{aligned} \Delta b_1 &= \alpha \Big(G_1^\lambda - V(s_1) \Big) bell(s_1) = \frac{1}{2} \times \Big(-\frac{3}{4} - 2 \Big) \times 1 = -\frac{11}{8} \\ \Delta b_2 &= \alpha \Big(G_2^\lambda - V(s_2) \Big) bell(s_2) = \frac{1}{2} \times \Big(\frac{1}{2} - (-2) \Big) \times 0 = 0 \\ \Delta b_3 &= \alpha \Big(G_3^\lambda - V(s_3) \Big) bell(s_3) = \frac{1}{2} \times \Big(1 - 0 \Big) \times 1 = -\frac{1}{2} \end{aligned}$$

The total update to weight b is, $\sum \Delta b = (\Delta b_1 + \Delta b_2 + \Delta b_3) = -\frac{11}{8} + 0 + \frac{1}{2} = -\frac{7}{8}$.

(d) The equation for eligibility trace is given as, $e_t = \gamma \alpha e_{t-1} + bell(s_t)$.

So, the sequence of eligibility traces e_t corresponding to $bell(s_t)$ are,

$$\begin{aligned} e_1 &= 1 \times \frac{1}{2} \times 0 + bell(s_1) = 1 \\ e_2 &= 1 \times \frac{1}{2} \times 1 + bell(s_2) = \frac{1}{2} \\ e_2 &= 1 \times \frac{1}{2} \times \frac{1}{2} + bell(s_3) = \frac{5}{4} \end{aligned}$$

(e) The sequence of updates to weight b is given as,

$$\begin{aligned} \Delta b_1 &= \alpha \delta_1 e_1 = \alpha \Big[R_2 + \gamma V(s_2) - V(s_1) \Big] e_1 = \frac{1}{2} \times \Big[0 + 1 \times (-2) - 2 \Big] \times 1 = -2 \\ \Delta b_2 &= \alpha \delta_2 e_2 = \alpha \Big[R_3 + \gamma V(s_3) - V(s_2) \Big] e_2 = \frac{1}{2} \times \Big[0 + 1 \times 0 - (-2) \Big] \times \frac{1}{2} = \frac{1}{2} \\ \Delta b_3 &= \alpha \delta_3 e_3 = \alpha \Big[R_4 + \gamma V(s_4) - V(s_3) \Big] e_3 = \frac{1}{2} \times \Big[1 + 1 \times 0 - 0 \Big] \times \frac{5}{4} = \frac{5}{8} \end{aligned}$$

The total update to weight b is, $\sum \Delta b = (\Delta b_1 + \Delta b_2 + \Delta b_3) = -2 + \frac{1}{2} + \frac{5}{8} = -\frac{7}{8}$.

Q4. [14 Marks] You are given an environment with one state, X, and two actions, b and c. T is the terminal state. Your Temporal Difference (TD) algorithm generates the following episode using the policy π when interacting with its environment:

Timestep	Reward	State	Action
0		X	b
1	16	X	c
2	12	X	b
3	16	T	

- The policy π is given by: $\pi(b|X) = 0.9,\ \pi(c|X) = 0.1$.
- The current values of q are: $q(X, b) = 1$ and $q(X, c) = 2$.
- The discount factor, $\gamma = \frac{1}{2}$.
- The step size, $\alpha = 0.1$.

Show the values of $q(X, b)$ and $q(X, c)$ after their first update using the following approaches:

(a) 1-step SARSA **(2 + 2)**

(b) 2-step SARSA **(2 + 2)**

(c) 2-step Full Tree Backup **(3 + 3)**

Note: You should update $q(X, b)$ and $q(X, c)$ only once per learning algorithm. Show your work and carry out your calculations to two decimal places.

Solution:

(a) *1-step SARSA*:

After the first timestep,

$$\begin{aligned} q(S_0, A_0) &= q(S_0, A_0) + \alpha\big[R_1 + \gamma q(S_1, A_1) - q(S_0, A_0)\big] \\ \therefore\ q(X, b) &= q(X, b) + \alpha\big[R_1 + \gamma q(X, c) - q(X, b)\big] \\ &= 1 + 0.1 \times \big[16 + 0.5 \times 2 - 1\big] \quad = 2.60 \end{aligned}$$

After the second timestep,

$$\begin{aligned} q(S_1, A_1) &= q(S_1, A_1) + \alpha\big[R_2 + \gamma q(S_2, A_2) - q(S_1, A_1)\big] \\ \therefore\ q(X, c) &= q(X, c) + \alpha\big[R_2 + \gamma q(X, b) - q(X, c)\big] \\ &= 2 + 0.1 \times \big[12 + 0.5 \times 2.6 - 2\big] \quad = 3.13 \end{aligned}$$

(b) *2-step SARSA*:

After the second timestep,

$$\begin{aligned} q(S_0, A_0) &= q(S_0, A_0) + \alpha\big[R_1 + \gamma R_2 + \gamma^2 q(S_2, A_2) - q(S_0, A_0)\big] \\ \therefore\ q(X, b) &= q(X, b) + \alpha\big[R_1 + \gamma R_2 + \gamma^2 q(X, b) - q(X, b)\big] \\ &= 1 + 0.1 \times \big[16 + 0.5 \times 12 + 0.5^2 \times 1 - 1\big] \quad = 3.13 \end{aligned}$$

After the third timestep,

$$\begin{aligned} q(S_1, A_1) &= q(S_1, A_1) + \alpha\big[R_2 + \gamma R_3 + \gamma^2 q(S_3, A_3) - q(S_1, A_1)\big] \\ \therefore\ q(X, c) &= q(X, c) + \alpha\big[R_2 + \gamma R_3 + \gamma^2 q(T, \cdot) - q(X, c)\big] \\ &= 2 + 0.1 \times \big[12 + 0.5 \times 16 + 0.5^2 \times 0 - 2\big] \quad = 3.80 \end{aligned}$$

(c) *2-step Full Tree Backup*:

After the second timestep,

$$\begin{aligned}
q(S_0,A_0) &= q(S_0,A_0)+\alpha\Big[R_1+\gamma\sum_{a\neq A_1}\pi(a|S_1)q(S_1,a) \\
&\qquad +\gamma\pi(A_1|S_1)\big[R_2+\gamma\sum_a\pi(a|S_2)q(S_2,a)\big]-q(S_0,A_0)\Big] \\
\therefore\ q(X,b) &= q(X,b)+\alpha\Big[R_1+\gamma\sum_{a\neq c}\pi(a|X)q(X,a) \\
&\qquad +\gamma\pi(c|X)\big[R_2+\gamma\sum_a\pi(a|X)q(X,a)\big]-q(X,b)\Big] \\
&= q(X,b)+\alpha\Big[R_1+\gamma\pi(b|X)q(X,b) \\
&\qquad +\gamma\pi(c|X)\big[R_2+\gamma\big(\pi(b|X)q(X,b)+\pi(c|X)q(X,c)\big)\big]-q(X,b)\Big] \\
&= 1+0.1\times\big[16+0.5\times0.9\times1+0.5\times0.1\times(12+0.9\times1+0.1\times2)-1\big] \quad = 2.61
\end{aligned}$$

After the third timestep,

$$\begin{aligned}
q(S_1,A_1) &= q(S_1,A_1)+\alpha\Big[R_2+\gamma\sum_{a\neq A_2}\pi(a|S_2)q(S_2,a) \\
&\qquad +\gamma\pi(A_2|S_2)\big[R_3+\gamma\sum_a\pi(a|S_3)q(S_3,a)\big]-q(S_1,A_1)\Big] \\
\therefore\ q(X,c) &= q(X,c)+\alpha\Big[R_2+\gamma\sum_{a\neq b}\pi(a|X)q(X,a) \\
&\qquad +\gamma\pi(c|X)\big[R_3+\gamma\sum_a\pi(a|T)q(T,a)\big]-q(X,c)\Big] \\
&= q(X,c)+\alpha\Big[R_2+\gamma\pi(c|X)q(X,c)+\gamma\pi(c|X)\big[R_3+0\big]-q(X,b)\Big] \\
&= 2+0.1\times\big[12+0.5\times0.1\times2+0.5\times.9\times(16+0)-2\big] \quad = 3.73
\end{aligned}$$