
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 (up, down, left, right) 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 $q_\pi(s_1, \text{down})$ and $q_\pi(s_2, \text{down})$ 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 $L_\pi : \mathcal{V} \rightarrow \mathcal{V}$ 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,

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

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)**

- 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

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: $[b \times bell(s) + l \times light(s)]$. 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)

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.