Indian Institute of Technology, Kanpur
Department of Electrical Engineering
Introduction to Reinforcement Learning (EE932)
Theory Assignment 3

Deadline: 25th May 2024 | 2023-24 Quarter 4 | Max points: 12

All the Questions are Objective type

1. Policies found by value iteration are superior to policies found by policy iteration. **(1 pt)**

   (a) True

   (b) False

2. What is the update equation for the value function in n-step TD for $n = 2$? **(1 pt)**

   (a) $V(s_t) = V(s_t) + \alpha[R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2}) - V(s_t)]$

   (b) $V(s_t) = V(s_t) + \alpha[R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+1}) - V(s_t)]$

   (c) $V(s_t) = V(s_t) + \alpha[R_{t+1} + \gamma R_{t+2} + \gamma^2 G_{t+2} - V(s_t)]$

   (d) $V(s_t) = V(s_t) + \alpha[R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2}) - V(s_{t+1})]$

3. Given a reinforcement learning problem, algorithm A will return the optimal state-value function $v^*(s)$ for that problem and algorithm B will return the optimal action-value function $q^*(s, a)$. Your aim is to find an optimal policy. Assuming that you know the expected rewards but not the transition probabilities corresponding to the problem in question, which algorithm would you prefer to use for finding π^*? **(1 pt)**

   (a) Algorithm A

   (b) Algorithm B

4. Consider the MDP in Figure 1 with discount factor $\gamma = 0.5$. Uppercase letters A, B, and C represent states; lowercase letters ab, ba, bc, ca, and cb represent actions; signed integers represent rewards, and fractions represent transition probabilities. **(1+1+1)**

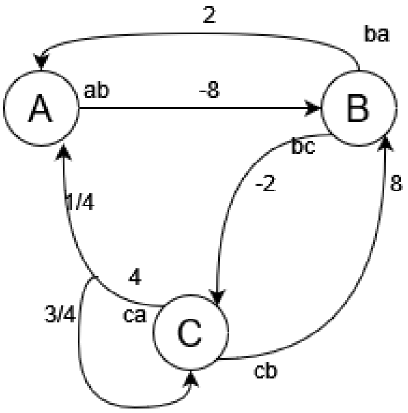


Figure 1: MDP

(i) Consider uniform random policy $\pi_1(a|s)$ that takes all actions from any state s with equal probability. If we start with an initial value of $V_1(s) = 2, \forall s \in \{A, B, C\}$, and apply one step of iterative policy evaluation, what will be the new value function $V_2(s)$?**(1 pt)**

(a) $V_2(A) = -7, V_2(B) = -2, V_2(C) = 7$

(b) $V_2(A) = -8, V_2(B) = 1, V_2(C) = 8$

(c) $V_2(A) = -7, V_2(B) = 1, V_2(C) = 7$

(d) $V_2(A) = 2, V_2(B) = 1, V_2(C) = 2$

(ii) If you apply policy improvement on V_2, what will be the resulting policy $\pi_2(s)$? **(1 pt)**

(a) $\pi_2(A) = ab, \pi_2(B) = ba, \pi_2(C) = ca.$

(b) $\pi_2(A) = ab, \pi_2(B) = bc, \pi_2(C) = cb.$

(c) $\pi_2(A) = ab, \pi_2(B) = bc, \pi_2(C) = ca.$

(d) $\pi_2(A) = ab, \pi_2(B) = ba, \pi_2(C) = cb.$

(iii) Is your new value function V_2 optimal? **(1 pt)**

(a) Yes

(b) No

(c) Given information is insufficient to answer

5. Let us revisit Assignment 2's grid problem in Figure 2. The states are grid squares, identified by their row and column numbers. The agent always starts in the bottom left state (1,1), marked with the letter S. (**Note that the bottom row is denoted by number 1, the top row by number 2**). There are two terminal goal states, (2,3) with reward +5 and (1,3) with reward -5. Rewards are 0 in non-terminal states. (The reward for a state is received as the agent moves into the state.) The transition function is such that the intended agent movement (UP, Down, Left, or Right) happens with probability 0.8. With probability 0.1 each, the agent ends up in one of the states perpendicular to the intended direction. Please refer Figure 3.

(1+1+1+1 pts)

(i) The agent starts with the policy that always chooses to go right, and executes the following three trials:
Trial 1) (1,1)–(1,2)–(1,3),
Trial 2) (1,1)–(1,2)–(2,2)–(2,3), and
Trail 3) (1,1)– (2,1)–(2,2)–(2,3).
What are the *first-visit* Monte Carlo value estimates for states (1,1) and (2,2), given these traces? Assume a discount factor of $\gamma = 1$. Choose the correct answer. **(1 pt)**

(a) $V(1, 1) = 5/3, V(2, 2) = 5/2$

(b) $V(1, 1) = 5, V(2, 2) = 5/2$

(c) $V(1, 1) = 5/3, V(2, 2) = 5$

(b) $V(1, 1) = 5, V(2, 2) = 10$

(ii) If we had used *every-visit* Monte Carlo method, what would be the answer to the previous question? **(1 pt)**

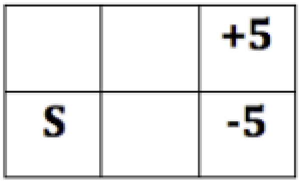
(a) $V(1, 1) = 5/3, V(2, 2) = 5/2$

(b) $V(1, 1) = 5, V(2, 2) = 5/2$

(c) $V(1, 1) = 5/3, V(2, 2) = 5$

(b) $V(1, 1) = 5, V(2, 2) = 10$

(iii) Consider the TD method with a learning rate of $\alpha = 0.1$, a discount factor of $\gamma = 0.9$, and initial values for all states equal 0. The agent starts with the policy that always chooses to go right and executes the following trials:



		+5
S		-5

Figure 2: Grid-World

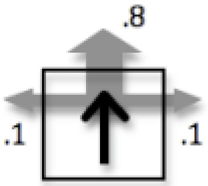


Figure 3: Tranistion probabilities for an ‘UP’ action

Trial 1) (1,1)–(1,2)–(1,3),
Trial 2) (1,1)–(1,2)–(2,2)–(2,3).
After **Trial 1**, what are the updated values of these states? **(1 pt)**

(a) $V(1,1) = 0, V(1,2) = 0, V(1,3) = 0.$
(b) $V(1,1) = -5, V(1,2) = -5, V(1,3) = 0.$
(b) $V(1,1) = 0, V(1,2) = -4.5, V(1,3) = 0.$
(d) $V(1,1) = 0, V(1,2) = -0.5, V(1,3) = 0.$

(iv) After completing both **Trail 1** and **Trail 2**, what are the updated values of these states given by the TD method described in the previous question? **(1 pt)**

(a) $V(1,1) = 0, V(1,2), = -0.5, V(2,2) = 0, V(2,3) = 0$
(b) $V(1,1) = -0.045, V(1,2), = -0.045, V(2,2) = 0.5, V(2,3) = 0$
(c) $V(1,1) = 0, V(1,2), = 0, V(2,2) = 5, V(2,3) = 0$
(c) $V(1,1) = -0.045, V(1,2), = -0.5, V(2,2) = 0.5, V(2,3) = 0$

6. Consider an MDP with two states A and B and two actions a and b in each state. We are following SARSA algorithm to find the optimal policy using GPI. Assume $\gamma = 0.8$ and $\alpha = 0.2$, $\epsilon = 0.1$. Suppose the initial Q-values are shown in the table below. **(1+1 pt)**

$Q(A, a)$	2.0
$Q(A, b)$	2.0
$Q(B, a)$	4.0
$Q(B, b)$	2.0

Suppose that we were intially in state A, we took action b, received reward 1, and moved to state B, and then took action b again to get a reward of -1 and landed up in state A. In other words, the trajectory observed so far $S_0, A_0, R_1, S_1, A_1, R_2, S_2$ is

$$A, b, 1, B, b, -1, A.$$

(i) After the first SARSA update, what is the value of $Q(A, b)$? **(1 pt)**

(a) 2.12
(b) 2.6

(c) 2

(d) 0

(ii) As observed from the trajectory, at time $t = 2$, we are at state A, i.e., $S_2 = A$. What is the probability of choosing $A_2 = b$? **(1 pt)**

(a) 0.1

(b) 0.9

(c) 0.95

(d) 0.15