INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR	
	Stamp / Signature of the Invigilator

EXAMINATION (End Semester)	SEMESTER (Autumn 2023-2024)

Roll Number									**Section**		**Name**	
Subject Number	C	S	6	0	0	7	7	**Subject Name**	*REINFORCEMENT LEARNING*			
Department / Center of the Student								**Additional sheets**				

Important Instructions and Guidelines for Students

1. You must occupy your seat as per the Examination Schedule/Sitting Plan.
2. Do not keep mobile phones or any similar electronic gadgets with you even in the switched off mode.
3. Loose papers, class notes, books or any such materials must not be in your possession, even if they are irrelevant to the subject you are taking examination.
4. Data book, codes, graph papers, relevant standard tables/charts or any other materials are allowed only when instructed by the paper-setter.
5. Use of instrument box, pencil box and non-programmable calculator is allowed during the examination. However, exchange of these items or any other papers (including question papers) is not permitted.
6. Write on both sides of the answer script and do not tear off any page. **Use last page(s) of the answer script for rough work.** Report to the invigilator if the answer script has torn or distorted page(s).
7. It is your responsibility to ensure that you have signed the Attendance Sheet. Keep your Admit Card/Identity Card on the desk for checking by the invigilator.
8. You may leave the examination hall for wash room or for drinking water for a very short period. Record your absence from the Examination Hall in the register provided. Smoking and the consumption of any kind of beverages are strictly prohibited inside the Examination Hall.
9. Do not leave the Examination Hall without submitting your answer script to the invigilator. **In any case, you are not allowed to take away the answer script with you.** After the completion of the examination, do not leave the seat until the invigilators collect all the answer scripts.
10. During the examination, either inside or outside the Examination Hall, gathering information from any kind of sources or exchanging information with others or any such attempt will be treated as **'unfair means'**. Do not adopt unfair means and do not indulge in unseemly behavior.

Violation of any of the above instructions may lead to severe punishment.

Signature of the Student

To be filled in by the examiner											
Question Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Marks Obtained											

Marks obtained (in words)	Signature of the Examiner	Signature of the Scrutineer

Indian Institute of Technology Kharagpur
Department of Computer Science and Engineering

End-Semester Exam	**Reinforcement Learning (CS60077)**	**Autumn 2023-2024**
Date: 24-Nov-2023 (FN)	**Answer *all* questions.**	**Maximum Marks:** 80

— Write your answers at indicated places inside the question paper. —

— This page is kept blank intentionally. —

— The question paper starts from the next page. —

Q1. **[Markov Decision Process and Bellman Equation]** **14 marks**

Assume an underlying Markov Decision Process (MDP), $M = (S, A, P, R, \gamma)$, where $s, s' \in S$ are the states of an MDP, $a \in A$ denotes an action in MDP, $r(s,a) = R_s^a$ denotes the reward accumulated when applying action a to state s, $\mathbb{P}\big[s' \mid s, a\big] = P_{ss'}^a$ denotes the transition probability to s' from s upon executing action a, and $\gamma \in (0, 1]$ is the discount factor.

(a) The Bellman equations give us the iterative relations that relate the state value functions and the action value functions. The Bellman expectation equations provide the relation between state value function v_π and action value function q_π for any policy π, while the Bellman optimality equations provide the relation between the optimal state value function v_* and the optimal action value function q_*. The two tables below require you to fill out the relations between the quantities marked in the row and the column headings.

(i) Fill up the missing entries from the table for *Bellman Expectation Equations* as given below.

	v_π	q_π
v_π	① $v_\pi(s) = ?$	$v_\pi(s) = \sum_{a \in A} \pi(a\|s) q_\pi(s,a)$
q_π	② $q_\pi(s,a) = ?$	③ $q_\pi(s,a) = ?$

(3)

(ii) Fill up the missing entries from the table for *Bellman Optimality Equations* as given below.

	v_*	q_*
v_*	④ $v_*(s) = ?$	$v_*(s) = \max_{a \in A} q_*(s,a)$
q_*	⑤ $q_*(s,a) = ?$	⑥ $q_*(s,a) = ?$

(3)

Answer:

① $v_\pi(s) =$

② $q_\pi(s,a) =$

③ $q_\pi(s,a) =$

---

④ $v_*(s) =$

⑤ $q_*(s,a) =$

⑥ $q_*(s,a) =$

(b) Let π be an ε-greedy policy. Let π' be the ε-greedy policy inferred from the action value function q_π (ε-greedy Policy Improvement from π to π'), i.e.,

$$\pi'(a|s) = \begin{cases} 1-\varepsilon+\frac{\varepsilon}{|A|}, & \text{if } a = arg\max_{b\in A} q_\pi(s,b) \\ \frac{\varepsilon}{|A|}, & \text{otherwise} \end{cases}$$

Prove that, $\sum_{a\in A} \pi'(a|s).q_\pi(s,a) \geq v_\pi(s)$, for all $s \in S$,

where $v_\pi(s)$ is state value function for policy π. **(8)**

Answer:

Q2. **[Model-free Control: Monte-Carlo vs. SARSA]** **10 marks**

In this problem, we deal with the batch Monte-Carlo algorithm and the online Expected-SARSA algorithm. We set the discount factor to $\gamma = 1$ and run 5 episodes (in all episodes first action is always a_1) in a given environment as follows:

- *Episode 1:* $s_1\ a_1\ 0\ s_4\ a_2\ 0\ s_7$
- *Episode 2:* $s_1\ a_1\ 0\ s_4\ a_3\ 0.4\ s_8$
- *Episode 3:* $s_2\ a_1\ 0\ s_4\ a_2\ 2\ s_6$
- *Episode 4:* $s_2\ a_1\ 0\ s_4\ a_3\ 0.4\ s_8$
- *Episode 5:* $s_3\ a_1\ 0.5\ s_4\ a_2\ 0\ s_7$

Here, each episode starts in one of the states $s_1,\ s_2,\ s_3$ and with action a_1. Also, $s_6,\ s_7,\ s_8$ are terminal states. In s_4 there is a choice between actions a_2 and a_3 which are taken with equal probability $\pi(a_2|s_4) = \pi(a_3|s_4) = 0.5$. The rewards are deterministic as mentioned within the episodes with numeric values and only depend on the transition (s, a, s').

(a) Calculate the Q-values in states $s_1,\ s_2,\ s_3,\ s_4$ using online Expected-SARSA. For a given Q-value $Q(s,a)$, use $\eta = 1$ the FIRST TIME you update this value and $\eta \in [0, 0.5]$ for all LATER (subsequent) update steps. Assume that $Q(s,a)$ values in all states are zero intitially. (**Hint:** You can neglect terms of order η^2.) **(5)**

Answer:

(b) Calculate the Q-values in states s_1, s_2, s_3, s_4 using batch Monte-Carlo control (i.e., average total returns from each starting state). **(3)**

Answer:

(c) You can choose the initial state for episode 6. Which initial state looks best in Part (a), i.e. for online Expected SARSA? Which initial state looks best in Part (b), i.e. for batch Monte-Carlo? **(2)**

Answer:

Q3. **[TD(λ) and Eligibility Traces]** **16 marks**

Sitting at hall on a rainy evening, you hear an episode of experience as follows: At the first step you saw a lightning. At the second step you hear a thunder with a drizzle of rain. At the third step you saw only a drizzle of rain. Then you had a powercut, 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 state s that you witnessed by a vector of three binary features, $light(s) \in \{0,1\}$, $thunder(s) \in \{0,1\}$ and $drizzle(s) \in \{0,1\}$. So, the sequence of feature vectors corresponding to the four steps of this episode can be expressed as, $[1,0,0]^\intercal$, $[0,1,1]^\intercal$, $[0,0,1]^\intercal$ and $[0,0,0]^\intercal$.

(a) Approximate the state-value function by a linear combination of these features with two parameters: $\phi(s) = \left[l \times light(s) + t \times thunder(s) + d \times drizzle(s)\right]$. If $l = -3$, $t = -2$ and $d = 1$, then write down the sequence of approximate values corresponding to this episode. **(2)**

Answer:

(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 $l = -3,\ t = -2,\ d = 1$. Clearly show the detailed evaluations. **(6)**

Answer:

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

Answer:

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

Answer:

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

Answer:

Q4. **[Policy Gradient Methods]** **12 marks**

(a) Let us consider the linear MDP as shown below. Here the (shaded) states, s_1 and s_7, are

s1	**s2**	**s3**	**s4**	**s5**	**s6**	**s7**
+1	0	−1	0	−1	0	+10

terminal states. The rewards presented below are received when you enter a particular state. There are two actions, `Left` and `Right`. The action, `Left`, transitions from state s_i to s_{i-1} with 0.5 probability and stays in state s_i with 0.5 probability. Similarly, the action, `Right`, transitions from state s_i to s_{i+1} with 0.5 probability and stays in state s_i with 0.5 probability. Let $\gamma = 1$. We want to apply Monte-Carlo policy gradient algorithm, REINFORCE, to learn a policy in this linear MDP setting. Let our feature representation be a one-hot encoding using the state, action pair. More concretely, let us denote $a_1 =$ `Left` and $a_2 =$ `Right`. Then, assuming the vector is 0-indexed, our feature representation is, $\phi(s_i, a_j)_k = \begin{cases} 1, & \text{if } 7(j-1)+(i-1)=k \\ 0, & \text{otherwise} \end{cases}$. Let us use a softmax policy parameterized by θ: $\pi_\theta(s,a) = \frac{\exp\left(\phi(s,a)^\top\theta\right)}{\sum_a \exp\left(\phi(s,a)^\top\theta\right)}$ and run REINFORCE algorithm. Assume θ is initialized to be all zeros. We execute one rollout of the policy π_θ to obtain the following episode: $(s_4, a_1, -1, s_3, a_2, 0, s_4, a_2, -1, s_5, a_2, 0, s_6, a_1, 0, s_6, a_2, +10)$. Run REINFORCE to update θ *three times* using the provided episode. For simplicity, let $\alpha = 1$. **(6)**

Answer:

(b) Suppose you have a Gaussian policy, π_θ, that samples actions a from a normal distribution with mean $\mu = \phi(s)^\intercal\theta$ and variance σ^2.

(**Hint:** Recall that, the Gaussian PDF is as follows: $f(x;\ \mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.)

(i) Prove that, $\nabla_\theta \log \pi_\theta(s,a) = \frac{\left(a-\phi(s)^\intercal\theta\right)\phi(s)}{\sigma^2}$. (show your derivation in details) **(3)**

Answer:

(ii) Prove that, $\nabla_\sigma \log \pi_\theta(s,a) = \frac{\left(a-\phi(s)^\intercal\right)^2}{\sigma^3} - \frac{1}{\sigma}$. (show your derivation in details) **(3)**

Answer:

Q5. **[Multi-arm Bandits]** **12 marks**

(a) In the *2-armed bandit* problem, one has to choose one of 2 actions. Assume action a_1 yields a reward of $r = 1$ with probability $p = 0.25$ and 0 otherwise. If you take action a_2, you will receive a reward of $r = 0.4$ with probability $p = 0.75$ and 0 otherwise. The *2-armed bandit* game is played several times and Q-values are updated using the update rule, $\Delta Q(s,a) = \eta \left[r_t - Q(s,a)\right]$.

(i) Assume that you initialize all Q-values at zero. You first try both actions: in trial 1, you choose a_1 and get $r = 1$; in trial 2, you choose a_2 and get $r = 0.4$. Update your Q-values ($\eta = 0.2$). **(3)**

Answer:

(ii) In trials 3 to 5, you play greedy and always choose the action which looks best (i.e., has the highest Q-value). Which action has the higher Q-value after trial 5? (Assume that the actual reward is $r = 0$ in trials 3-5.). **(3)**

Answer:

(iii) Calculate the expected reward for both actions. Which one is the best? **(2)**

Answer:

(b) After 12 iterations of the UCB 1 algorithm applied on a 4-arm bandit problem, we have $n_1 = 3,\ n_2 = 4,\ n_3 = 3,\ n_4 = 2$ and $Q_{12}(1) = 0.55,\ Q_{12}(2) = 0.63,\ Q_{12}(3) = 0.61,\ Q_{12}(4) = 0.40$. Which arm should be played next? Show the calculations to justify your answer. **(4)**

Answer:

Q6. **[Approximation Guarantees over MDPs]** **16 marks**

(a) For an MDP $\langle S, A, P, R, \gamma \rangle$, let $V_0 : S \to \mathbb{R}$ be an initial guess of the optimal value function V^*. Let this guess be progressively updated using Value Iteration: i.e., by setting $V_{t+1} \leftarrow T^*(V_t)$ for $t = 0, 1, 2, \ldots$. Recall that T^* is the Bellman optimality operator.

In this question, we examine the design of a stopping condition for Value Iteration. As usual, let $||\cdot||_\infty$ denote the max norm. We would like that our computed solution, V_u for some $u \in \{1, 2, \ldots\}$, be within ε of V^* for some given tolerance $\varepsilon > 0$. In other words, we would like to stop after u applications of T^*, so long as we can guarantee $||V_u - V^*||_\infty \leq \varepsilon$. Naturally, we cannot use V^* itself in our stopping rule, since it is not known! *Prove that* it suffices to stop when $||V_u - V_{u-1}||_\infty \leq \frac{\varepsilon(1-\gamma)}{\gamma}$ and thereafter return V_u as the answer. **(5)**

You are likely to find two results handy: (1) that T^* is a contraction mapping with contraction factor γ, and (2) the triangle inequality: for $X : S \to \mathbb{R}$, $Y : S \to \mathbb{R}$, $||X + Y||_\infty \leq ||X||_\infty + ||Y||_\infty$.

Answer:

(b) The *future state distribution* gives the probability of a state s appearing anywhere in a trajectory τ when a policy π is followed. It is denoted by $T^{\pi}(s) = \sum_{t=0}^{\infty} \mathbb{P}\big[S_t = s \mid \pi\big]$. Similarly,the *discounted future state distribution* provides the probability of a state s appearing anywhere in a trajectory but discounted by when is the state visited. It is denoted by, $d^{\pi}(s)$ and is defined as, $d^{\pi}(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}\big[S_t = s \mid \pi\big]$ where γ is the discount factor of an infinite horizon MDP. $\mathbb{P}\big[S_t = s \mid \pi\big]$ denotes the probability of the state s to appear at timestep t.

We can use the discounted future state distribution to rewrite the objective function of a RL problem in the infinite horizon discounted reward setting. The objective i.e., the expectation of the discounted sums over trajectories can be rewritten in terms of expectations over states and actions. For any policy π and any reward function $r : S \times A \to \mathbb{R}$, the relation can be written as,

$$\mathop{\mathbb{E}}_{\tau\sim p(\tau)}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\Big] = \frac{1}{1-\gamma}\sum_{s\in S} d^{\pi}(s) \sum_{a\in A} \pi(a|s) r(s,a) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\,\sim\, d^{\pi}(\cdot) \\ a\,\sim\,\pi(\cdot|s)}} \big[r(s,a)\big]$$

where, $s \sim d^{\pi}(\cdot)$ is a shorthand to denote the fact that states are drawn according to the discounted future state distribution and similarly $a \sim \pi(\cdot|s)$ is a shorthand to denote the fact that actions are distributed according to π. Your task is to *prove this above relation*. **(7)**

(**Hint:** Start proving from the opposite direction!)

Answer:

(c) Suppose an MDP $\langle S, A, R, P, \gamma \rangle$ is defined such that $R(s,a) \geq 0$ for all state action pairs $(s,a) \in S \times A$. Furthermore, suppose that for every state $s \in S$, there is some action $a_s \in A$ such that $\mathbb{P}(s' = s \mid s, a_s) \geq p$, where $0 \leq p \leq 1$ is some constant probability. Consider performing value iteration on this MDP. Let $V_t(s)$ be the value of state s after t iterations. We initialize to $V_0(s) = 0$ for all states $s \in S$. *Prove that* for all states $s \in S$ and $t \geq 0$, $\quad V_{t+1}(s) \geq p\gamma V_t(s)$. **(4)**

Answer:

— The question paper ends here. —