[This question paper contains 8 printed pages.]

Your Roll No.....

L

Sr. No. of Question Paper : 7968

Unique Paper Code : 2343010023

Name of the Paper : Reinforcement Learning

Name of the Course : **Bsc. Computer Science**

Semester : VIII

Duration : 3 Hours

Maximum Marks : 90

Instructions for Candidates

1. Write your Roll No. on the top immediately on receipt of this question paper.
2. **Section A** is compulsory.
3. Attempt **any 4** questions from **Section B.**
4. Parts of a question must be answered together.
5. Use of Scientific Calculator is allowed.

Section A

1. (a) What is Reinforcement Learning? How is it different from Supervised learning? (3)

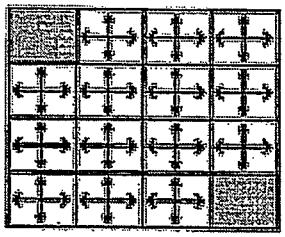
   (b) Differentiate between a Reward Signal and Value Function. Which one of them represents the long-term potential of a state? (3)

   (c) A robot's next position depends only on its current coordinates and its current velocity. Does this system satisfy the Markov Property? Justify your answer. (3)

   (d) An agent receives a sequence of rewards $R_1 = 2$, $R_2 = 4$, $R_3 = 10$. If the discount factor $\gamma = 0.5$, calculate the total discounted return G_0. (3)

(e) Consider a 4x4 grid world. The top-left and bottom-right are the terminal states, while the rest are non-terminal states. Initial state-values v are all 0s and corresponding greedy policy is shown using arrows for all directions achieving the maximum. (3)

0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0



After a series of policy iteration steps, the new state-values are shown as below. Show the greedy policy (using arrows) with respect to these given state values:

0.0	-1.7	-2.0	-2.0
-1.7	-2.0	-2.0	-2.0
-2.0	-2.0	-2.0	-1.7
-2.0	-2.0	-1.7	0.0

(f) An agent is using ϵ-greedy policy with $\epsilon = 0.1$. If there are 5 possible actions, compute the probabilities of choosing the greedy (best-known) action and the non-greedy (random) actions. (3)

(g) In the Actor-Critic algorithm, which component is responsible for the policy, and which is responsible for value function computation? (3)

(h) Using a suitable example, describe the trade-off between exploration and exploitation. What are the consequences of favoring one over the other? (3)

(i) In a TD(λ) algorithm, the eligibility trace is updated as: (3)

$$z_t = \gamma\lambda z_{t-1} + 1$$

Given $\gamma = 0.9$, initial trace $z_0 = 0$, compute z_1 and z_2 when $\lambda = 1$.

(j) What is the difference between n-step TD and TD(λ) in terms of how they use future rewards? (3)

Section B

2 (a) Describe the Agent-Environment interaction in a Reinforcement Learning Task. Draw a diagram showing the interaction loop of S_t, A_t, and R_t and explain how the time step progresses from one interaction to the next. (5)

(b) Consider a state s_0 with two possible actions a_1 and a_2: (5)

- Action a_1 leads to state s_1 with reward = 4
- Action a_2 leads to state s_2 with reward = 2

The agent follows a policy:

$$\pi(a_1|s_0) = 0.6, \pi(a_2|s_0) = 0.4$$

Given $\gamma = 0.9$, $V(s_1) = 5$, $V(s_2) = 3$, compute $V(s)$ using the Bellman Expectation Equation.

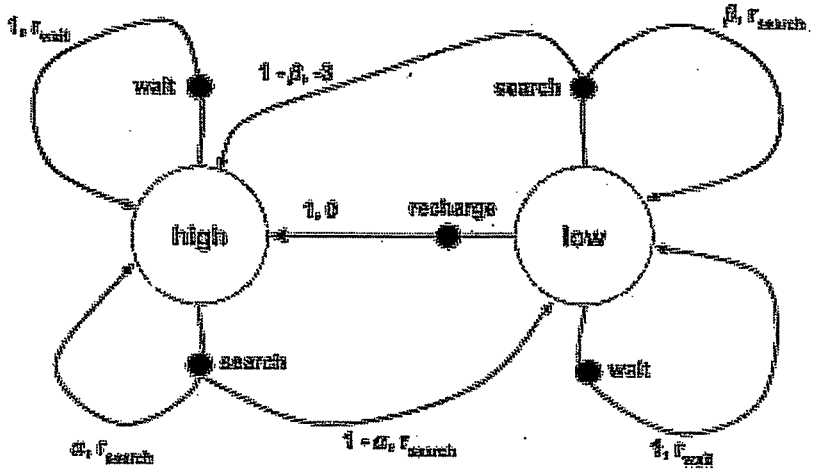
(c) Consider a Reinforcement learning agent being trained to play the game of Chess. Identify and briefly describe the following components of this RL system: Agent, Environment, State, Action, Reward. (5)

3 (a) Define Value Functions in Reinforcement Learning. Write the Bellman Equation for the state-value function $v_\pi(s)$. Identify and explain the meaning of each mathematical term used in the equation such as π, s, a, r, and γ. (5)

(b) Explain the difference between Policy Iteration and Value Iteration algorithms. Describe the primary steps involved in each. Which one of the two is generally faster in terms of the number of iterations to converge? (5)

(c) Consider a state s where taking action 'a' leads to state s_1 with probability 0.7 (reward +5) and state s_2 with probability 0.3 (reward -1). If discount factor $\gamma = 0.9$, $v(s_1) = 10$, and $v(s_2) = 2$, calculate $q(s, a)$. (5)

4 (a) (5)



Above diagram shows the transition graph for a finite MDP with following details:

States, $S = \{high, low\}$

Actions, $A = \{search, wait, recharge\}$

Starting in a state s and taking action a moves you along the line from state node s to action node (s, a). Each arrow corresponds to a triple (s, s', a), where s' is the next state and the arrow is labelled with transition probability, $p(s'|s, a)$, where s' is the next state and the expected reward for that transition $r(s, a, s')$.

For this finite MDP system, fill in the transition probabilities and the expected rewards in the table below:

s	a	s'	p(s'|s, a)	r(s, a, s')
high	search	high		
high	search	low		
low	search	high		
low	search	low		
high	wait	high		
high	wait	low		
low	wait	high		
low	wait	low		
low	recharge	high		
low	recharge	low		

(b) Explain the assumption of *exploring starts* in Monte Carlo estimation of Action values? Why is it difficult to satisfy in practice, and how does an ϵ-greedy policy address this issue? (5)

(c) An agent generates two episodes. Compute the value estimate $V(S_1)$ using First-visit Monte Carlo (MC) assuming a discount factor $\gamma = 1$. (5)

Episode 1: $(S_1, A_1, 2) \rightarrow (S_2, A_2, 3) \rightarrow (S_1, A_2, 5) \rightarrow$ Terminated.

Episode 2: $(S_1, A_1, 1) \rightarrow (S_3, A_2, 9) \rightarrow$ Terminated.

5 (a) Compare SARSA and Q-learning in the context of temporal-difference learning. Write their update rules and justify why SARSA is classified as an on-policy method while Q-learning is considered off-policy. (5)

(b) Discuss how n-step TD methods are a generalization of Reinforcement Learning algorithms. Show how by taking specific values of parameter 'n' allows these methods converge to either TD(0) or Monte-Carlo. (5)

(c) An agent is in state S_t with $V(S_t) = 10$. It takes an action, receives a reward of 2, and transitions to S_{t+1} where $V(S_{t+1}) = 12$. If the learning rate $\alpha = 0.5$ and discount factor $\gamma = 0.9$, calculate the updated value $V(S_t)$. (5)

6 (a) Why is it impractical to use a simple lookup table (tabular method) for complex games like Chess? Compare Function Approximation with Tabular methods in terms of memory constraints, and the ability to generalize to new situations. (5)

(b) State the Policy Gradient Theorem. Why do policy gradient methods work well when actions are continuous or when policy needs to be stochastic? (5)

(c) Consider a short corridor problem with switched actions consisting of 3 non-terminal states S_1, S_2, S_3 and a terminal goal state G. The agent always starts from S_1. The reward is −1 per step, and reaching G ends the episode. (5)

Given Action dynamics:

In S_1 and S_3:

Action Right moves right

Action Left moves left (no movement if already at boundary)

In S_2: actions are switched

Right moves left

Left moves right

A policy selects Right with probability p and Left with probability 1−p at every state.

Assume:

Discount factor gamma(γ)=1

From experiments, the expected number of steps to reach terminal Goal state, i.e. T(p) is:

T(p = 0.4) =25

T(p = 0.6) = 18

T(p = 0.8) =30

Perform the following tasks: -

i. Create a suitable diagram for the above problem.

ii. Compute the expected return for each value of p.

iii. Identify which policy (value of p) is best.

Comment briefly on why performance decreases when p is too small or too large.

7 (a) Given: $Q(S_t, A_t) = 2$, $R_{t+1} = 3$, $Q(S_{t+1}, a_1) = 7$, $Q(S_{t+1}, a_2) = 8$, $\gamma = 0.9$, $\alpha = 0.1$ (5)
Assume that under the current behavior policy, the next action selected at state S_{t+1} is a_1.

- Compute the updated value of $Q(S_t, A_t)$ using the SARSA update rule.
- Compute the updated value of $Q(S_t, A_t)$ using the Q-learning update rule (assume greedy action selection).

(b) Compare Dynamic Programming (DP), Monte Carlo (MC) , and Temporal Difference (TD) methods. Evaluate the methods based on following basis: (5)

- Model Requirement
- Bootstrapping
- Suitability for Episodic/Continuous tasks.

(c) In the Monte Carlo (MC) Prediction task, an agent has visited state S five times, resulting in an average return of $V(S) = 2.0$. On the sixth visit to state S, the agent observes a new return (G_t) of 4.0. (5)

Calculate the new value estimate $V(S)$ using the incremental update rule.