BIRLA INSTITUTE OF TECHNOLOGY, MESRA, RANCHI
(END SEMESTER EXAMINATION)

CLASS: BTECH **SEMESTER : VII**
BRANCH: AIML **SESSION : MO/2025**

SUBJECT: AI401 REINFORCEMENT LEARNING

TIME: 3 Hours **FULL MARKS: 50**

INSTRUCTIONS:
1. The question paper contains 5 questions each of 10 marks and total 50 marks.
2. Attempt all questions.
3. The missing data, if any, may be assumed suitably.
***4*. Before attempting the question paper, be sure that you have got the correct question paper.**
5. Tables/Data hand book/Graph paper etc. to be supplied to the candidates in the examination hall.

			CO	BL
Q.1(a)	Derive the Bellman equation, which characterises the relationships of state values. Explain briefly the terms of the equation	[5]	1	3
Q.1(b)		[5]	1	4

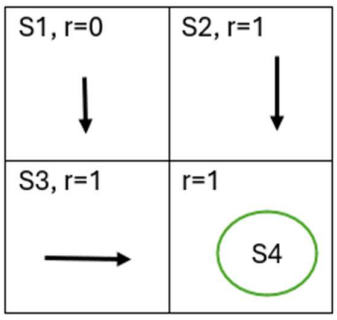

Fig1

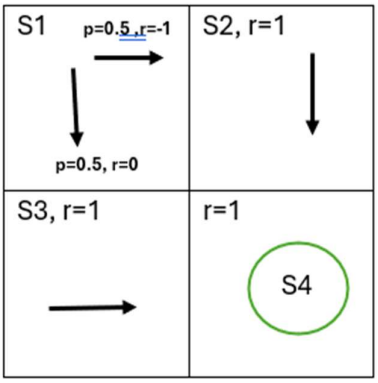

Fig2

Why do we care about the values of the actions that a given policy cannot select? Determine which policy(Fig1/Fig2) is better by calculating state values, assuming ¥ = 0.89.

Q.2(a)	If we increase all the rewards by the same amount, will the optimal state value change? Will the optimal policy change? Prove mathematically	[5]	2	3
Q.2(b)		[5]	2	4

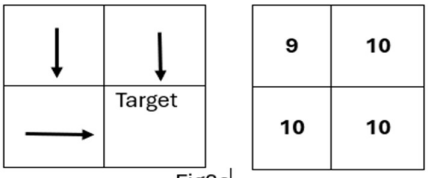

Fig2a

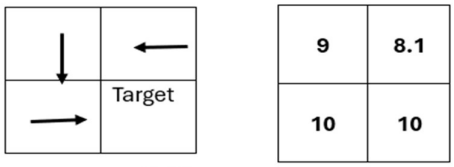

Fig2b

Determine which policy is optimal? Explain the significance in terms of immediate and discounted rewards,¥=0.9

PTO

Q.3(a) [5] 3 5

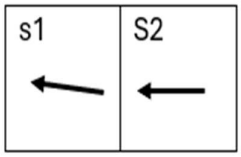

Fig3a

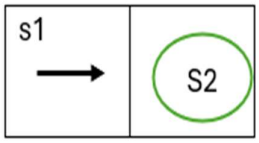

Fig3b

Determine the state values for j=1 to 3 using the policy iteration algorithm , assuming more iterations $v_{\pi 0}^{j}(s1) \rightarrow, v_{\pi 0}(s1) \ldots \ldots$and $v_{\pi 0}^{j}(s2) \rightarrow, v_{\pi 0}(s2) \ldots \ldots$as j increases. Obtain the q-value tables to determine the optimal policy. There are two states with three possible actions: A = al, a0 ar . The three actions represent moving leftward, staying unchanged, and moving rightward. The reward settings are $r_{boundary}$ = 1 and r_{target} = 1. The discount rate is = 0.89

Q.3(b) What is the core idea of model-free MC-based reinforcement learning? What are initial-visit, first-visit, and every-visit strategies? Can €-greedy policy be optimal? Give reasons [5] 3 3

Q.4(a) Explain Robbins-Monro's Algorithm. State its convergence steps. What is the advantage of the RM algorithm over other root-finding algorithms? [5] 4 3

Q.4(b) Explain the differences between SGD,BGD and MBGD algorithms [5] 4 3

Q.5(a) Explain the differences between TD learning and MC learning [5] 5 3

Q.5(b) Write the steps of Q-learning.Obtain the optimal policy and update the Q(s,a) values for the given grid,assume ¥=0.9,Goal as absorbing state. [5] 5 5

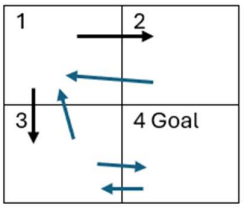

Fig1 Non-Optimal Policy

	1	2	3	4
1	-1	0	0	-1
2	0	-1	-1	10
3	0	-1	-1	10
4	-1	-1	-1	0

Fig2 Reward matrix

	1	2	3	4
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0

Fig3 Initial Q (s ,a) values

::::::19/11/2025::::::M