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

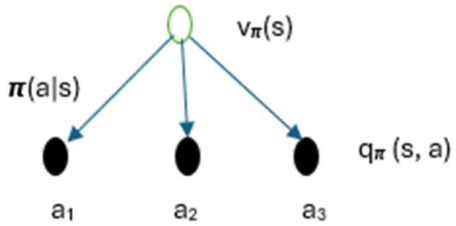
CLASS: MTECH **SEMESTER : II**
BRANCH: AIML **SESSION : SP/2025**

SUBJECT: AI626 CONCEPTS OF 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)	Explain the RL framework with a block diagram showing states, actions, rewards using mathematical notations.	[5]	1	2
Q.1(b)	For a multi arm bandit derive an expression for estimated Action value in stationary and non-stationary environment. For k=5 arm bandit the actions A1=a1,A2=a,A3=a2 and A4=a1, Rewards R1=1, R2=2, R3=5, R4=4 respectively. Implies R1,1=1,R1,2=2,R1,3=0, R1, 4=4.Determine Q1,5. Ri,j (ith action and jth step)	[5]	1	3
Q.2(a)	Write the Bellman's equation for action values($q_\pi(s,a)$).Also using the given backup diagram write $v_\pi(s)$ in terms of action values. Derive Bellman's Optimality equations for state values and actional values	[5]	2	3

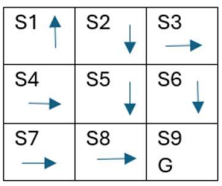


Q.2(b)	(i) Determine V(S) for the given grid world using Bellman's equation assuming four action only(L,R,U,D)	[5]	2	5

3.0	2.3	0.5
0.6	S	0.4
-0.4	-0.4	-0.6

(ii)Suppose Y=0.9 and the reward sequence is R_1=3 followed by infinite sequence of 7s.Determine G_1 and G_2.for given Y=0.9 If the reward is 1 the Find G_t

Q.3(a)	Explain the differences (i)Deterministic Policy vs Stochastic Policy,(ii)Model based policy vs Model free policy (ii)Off policy vs On Policy	[5]	3	2
Q.3(b)	Explain mathematically the policy iteration and value iteration in context of dynamic programming.Why the given policy(see Fig) is not optimal ? The reward settings are $r_{boundary}$ = $r_{forbidden}$ = 1 and r_{target} = 1. The discount rate is =0.9. Determine the optimal policy and the corresponding action values by assuming 4 actions in each state. Use MC Basic algorithm.	[5]	3	5



PTO

Q.4(a)	Explain the MC epsilon-greedy algorithm in context of exploration and exploitation.	[5]	4	3
Q.4(b)		[5]	4	5

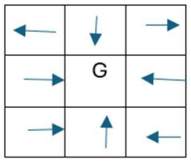


Fig 1 Policy

2.5	10	5.9
10	10	10
9	10	9

Fig2 State values

Fig1 shows ε-greedy policy and Fig 2 shows the state values for ε=0. Obtain the optimal ε-greedy policy and sate values for ε=0.1,0.2 and 0.5 and comment on the results. Draw a grid diagram to show the difference between target and behavior policy.

Q.5(a)	Explain the differences between Tabular TD algorithm and Linear TD algorithm in terms of true state values and 3D visualization	[5]	5	4
Q.5(b)	Draw the architecture of Dyna.Explain the REINFORCE algorithm	[5]	4,5	4

::::::01/05/2025::::::E