Reg. No.															

B.Tech. DEGREE EXAMINATION, DECEMBER 2023

Sixth Semester

18AIC304J – REINFORCEMENT LEARNING TECHNIQUES

(For the candidates admitted from the academic year 2020-2021 & 2021-2022)

Note:

(i) **Part - A** should be answered in OMR sheet within first 40 minutes and OMR sheet should be handed over to hall invigilator at the end of 40th minute.

(ii) **Part - B & Part - C** should be answered in answer booklet.

Time: 3 hours | Max. Marks: 100

PART – A (20 × 1 = 20 Marks)

Answer **ALL** Questions

	Marks	BL	CO	PO
1. Consider Monte-Carlo approach for policy evaluation suppose the states are S_1, S_2, S_3, S_4, S_5, S_6 and terminal state. You sample one trajectory as follows $S_1 \rightarrow S_3 \rightarrow S_5 \rightarrow S_2 \rightarrow$ terminal state. Which among the following states can be updated from this sample? (A) S_1 (B) S_6 (C) S_4 (D) S_7	1	2	1	1
2. If we follow Boltzmann exploration strategy then how should the temperature (T) be varied? (A) Start off with low value and increase gradually (B) Start off with high value and decay it gradually (C) Keep it fixed at a small value (D) Keep it fixed at a higher value	1	2	1	1
3. The matrix created during the Q-learning algorithm is commonly known as _____. (A) Query-table (B) Q-table (C) Quick-matrix (D) Table	1	1	1	1
4. How many tuples does MDP have? (A) 2 (B) 3 (C) 4 (D) 5	1	1	1	1
5. How many parameters are needed to be estimated during Q-learning in a world with S states each having A different action? (A) O (S.A) (B) O(S) (C) O(A) (D) $O(A^S)$	1	2	2	2
6. Gamma (γ) in the Bellman equation is known as? (A) Value factor (B) Discount factor (C) Environment factor (D) State factor	1	1	2	2
7. Which of the following could be an application of reinforcement learning? (A) Image classification (B) Self driving cars (C) Pattern recognition (D) Market based analysis	1	2	2	1

8. Which of the following statement is true for model-based and model-free reinforcement learning method? 1 2 2 3

(A) Model-based learning requires more parameters and data to learn
(B) Model-free learning can exploit the underlying MDP structure
(C) Model-free learning can simulate new episode from past experience
(D) Model-based learning needs minimum data to learn

9. A learning algorithm that evaluates and improves a policy which is dissimilar from the policy used for action selection is called the _____. 1 1 3 2

(A) Target policy
(B) Behaviour policy
(C) Off-policy
(D) On-policy

10. RL is learning what to do and how to map situations to action therefore to maximize _____. 1 2 3 1

(A) Actions
(B) Decisions
(C) Rewards
(D) Learn from prior experience

11. Through _____ element will the agent take action based on analyzing the current state within the environment? 1 1 3 1

(A) Policy
(B) Action
(C) State
(D) Environment

12. In a Bandit problem, let's consider a single optimum arm a^* , will it be possible to eliminate a^* using median elimination algorithm? In this case, the algorithm can output an arm e-close to a^*, is that true? 1 2 3 2

(A) No, yes
(B) Yes, no
(C) Yes, yes
(D) No, no

13. The optimal value of the discount factor lies within the range of _____. 1 2 4 1

(A) 1.2 to 1.8
(B) 0.2 to 0.8
(C) 1.0 to 1.9
(D) 0.1 to 0.9

14. Which of the following is not true about upper confidence bound? 1 1 4 1

(A) It's a deterministic algorithm
(B) It follows principle of optimism in the face of uncertainty
(C) It does not allow delayed feedback
(D) It is based on Bayes inference

15. Reinforce algorithm belongs to the special class of reinforcement learning algorithms known as _____. 1 1 4 1

(A) Q-learning
(B) Value-based learning
(C) Policy gradient algorithm
(D) Temporal difference algorithm

16. Which of the following is the basic form of reinforcement learning? 1 2 4 1

(A) Reward values
(B) Quality values
(C) Q-values
(D) State values

17.	The _____ algorithm is used to train a Markov decision process on a new policy. (A) Q-learning (B) Temporal difference algorithm (C) State-action-reward-state-action (D) Software agents	1	1	5	1
18.	Among the following definitions, which option exactly represents Q-learning? (A) It's a type of reinforcement learning that forces on rewards (B) It's a supervised machine learning with rewards (C) A type of unsupervised learning that relies heavily on a well established model (D) A type of reinforcement learning where accuracy degrades over time	1	2	5	2
19.	Which of the following is true about Q-table? (A) Guide us to the best action at each state (B) Guide us to the best rewards at each state (C) Guide us to the penalty at each iteration (D) Guide us to the maximum future record	1	2	5	2
20.	Which of the following is true about temporal difference learning? (A) Learn Q-function (B) Employs deep neural network to an appropriate value (C) Calculates all possible action based on Q-value (D) Learn how to predict a quantity that depends on future values	1	2	5	1

PART – B (5 × 4 = 20 Marks)
Answer **ANY FIVE** Questions

		Marks	BL	CO	PO
21.	What are the elements of reinforcement learning and provide an intuitive explanation for the policy in reinforcement learning.	4	3	1	1
22.	Write a reinforcement learning algorithm to balance a pole when the cart moves left and right within a certain angle.	4	1	2	1
23.	Distinguish PMF and PDF.	4	2	3	1
24.	Devise reinforcement learning based strategy for the working of pick and place operations in Robotics interpret the action, state and reward with the dynamically changing constraints.	4	3	4	2
25.	Demonstrate the importance of on-policy and off-policy learning with an example.	4	2	5	1
26.	Elaborate the steps in Monte Carlo learning through sampling rewards from the environment averaging over obtained rewards.	4	2	3	2
27.	Construct a model-free reinforcement learning algorithm.	4	3	3	1

PART – C (5 × 12 = 60 Marks)

Answer **ALL** Questions

		Marks	BL	CO	PO
28. a.	Demonstrate self-learning Tic-Tac-Toe game using reinforcement learning techniques.	12	2	1	1
	(OR)				
b.	Devise an agent for solving armed bandits big slot machine with k arms end each arm you pull has a different reward associated with it. You're given 100 quarters, so you need to develop some kind of strategy to get most reward.	12	4	1	1
29. a.	Illustrate the steps involved in policy evaluation and compute the state-value function using Markov Decision process.	12	3	2	1
	(OR)				
b.	Explain the working of Bellman equation for decomposing the value function into immediate reward and future values.	12	2	2	1
30. a.	Examine and explain the steps in policy improvement by computing value function to improve an original policy.	12	3	3	1
	(OR)				
b.	Explain the steps to train Markov Decision process on a new policy using SARSA on-policy algorithm.	12	2	3	1
31. a.	Explain the process of Q-learning to learn the value of an action in a particular state.	12	3	4	1
	(OR)				
b.	Identify the conditions for convergence of temporal difference learning to predict the total rewards expected over the future and explain it.	12	3	4	2
32. a.	Explain the memory structure of actor critic method to represent the policy independent of the value function.	12	4	5	2
	(OR)				
b.	Determine the effect of Naïve reinforcement learning to formalize sequential decision making in stock market analysis.	12	3	5	1

* * * * *