25.	Illustrate the weighing on the sequence of N-step returns in the λ-Return.	4	2	5	1
26.	Consider a hole of golf as a reinforcement learning task. Discuss about the state-value function and optimal action-value function with neat diagram.	4	3	2	2
27.	Write a pseudo code for least-squares temporal-difference algorithm for estimating value function.	4	2	5	2

PART – C (5 × 12 = 60 Marks)

Answer **ALL** Questions

		Marks	BL	CO	PO
28. a.	Consider the following learning problem. You are faced repeatedly with a choice among K different options, or actions. After each choice, you receive a numerical reward chosen form a stationary. Develop an algorithm for the above learning scenario.	12	4	1	4
	(OR)				
b.	Imagine an online advertising trial where an advertiser wants to measure the click-through rate of four different ads for the same product. Explain how the advertiser finds the best performing advertisement using upper confidence bound algorithm.	12	4	1	4
29. a.	A mobile robot has the job of collecting empty soda cans in an office environment. High_level decisions about how to search for cans are made by a reinforcement learning agent based on the current charge level of the battery. Formulate the above problem using Markov decision process.	12	4	2	4
	(OR)				
b.	Explain how Bellman optimality equations are used to find the optimal policy and optimal value function.	12	2	2	3
30. a.	Explain the convergence of iterative policy evaluation with the 4 × 4 grid world problem.	12	3	3	4
	(OR)				
b.	State and derive Banach's fixed point theorem.	12	3	3	3
31. a.	Explain the detail about the Mote Carlo prediction and discuss how Monte Carlo estimation can be used in control that is to approximate optimal policies.	12	3	4	3
	(OR)				
b.	Explain in detail about state-Action-Reward-State-Action (SARSA) and Q-Learning temporal difference control techniques with an example.	12	3	4	3
32. a.	Discuss in detail about the actor critic methods for episodic and continuing problems.	12	3	5	3
	(OR)				
b.	Consider a 1000-state version of the random walk task. The states are numbered from 1 to 1000, left to right and all episodes being near the center, in state 500. State transitions are from the 100 neighboring states to its left or to one of the 100 neighboring states to its right, all with equal probability. Perform state aggregation using semi-gradient method and explain it.	12	4	5	4

* * * * *

Reg. No.														

B.Tech. DEGREE EXAMINATION, JUNE 2023

Sixth Semester

18AIC304J – REINFORCEMENT LEARNING TECHNIQUES

(For the candidates admitted during the academic year 2018-2019 to 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.	Which of the following is true of the upper confidence bound algorithm? (A) The action with the highest Q value is chosen at every iteration (B) After a very large number of iterations, the confidence intervals of unselected actions will not change much (C) The true expected value of an action always lies within its estimated confidence interval (D) With a small probability ε, random action is selected to ensure adequate exploration of the action space	1	2	1	2
2.	Which of the following is true for the ε-greedy approach for a stationary environment? (A) Always keeping ε as constant is a good approach (B) Small values of ε will lead to unnecessary exploration in the long run (C) Always keeping ε as large is a good approach (D) Cooling down ε too fast is problematic as it cannot guarantee correctness in value estimates	1	2	1	2
3.	Which of the following is the practical example of reinforcement learning? (A) House pricing prediction (B) Market basket analysis (C) Text classification (D) Driverless cars	1	1	1	1
4.	What is the environment in reinforcement learning? (A) Environment is a situation that is based on the current state (B) Environment is a situation in which an agent is present (C) Environment is similar to feedback (D) Environment is a situation that the agent returns as a result	1	1	1	2
5.	Which of the following is true for an MDP? (A) $P_r(s_{t+1}, r_{t+1} \mid s_t, a_t) = P_r(s_{t+1}, r_{t+1})$ (B) $P_r(s_{t+1}, r_{t+1} \mid s_t, a_t, s_{t-1}, a_{t-1}, s_{t-2}, a_{t-2} \ldots\ldots, s_0, a_0) = P_r(s_{t+1}, r_{t+1} \mid s_t, a_t)$ (C) $P_r(s_{t+1}, r_{t+1} \mid s_t, a_t) = P_r(s_{t+1}, r_{t+1} \mid s_0, a_0)$ (D) $P_r(s_{t+1}, r_{t+1} \mid s_t, a_t) = P_r(s_t, r_t \mid s_{t-1}, a_{t-1})$	1	2	2	2

6. Let's assume for full RL problem with policy 'π'. At some time 't', the state is 's' with action 'A_1'. After few time steps 't'. the same state 's' was reached where we performed an action $a_2(\neq a_1)$. Which of the following is true. 1 2 2 2

(A) π is definitely a stationary policy (B) π is definitely a non-stationary policy
(C) π can be stationary or non-stationary (D) π cannot be defined

7. Gamma (γ) in the Bellman equation is known as 1 1 2 1

(A) Discount factor (B) Value factor
(C) Environment factor (D) Policy

8. Which of the following is a benefit of using reinforcement learning (RL) algorithms for solving Markov decision process (MDP)? 1 2 2 2

(A) They donot require the state of the agent for solving a MDP (B) They donot require the action taken by the agent for solving a MDP
(C) They donot require the state transition probability matrix for solving a MDP (D) They donot require the reward signal for solving a MDP

9. Assertion: Monte Carlo evolution must use exploring starts in the case of non-deterministic policies 1 2 3 2
Reason: They have to rely upon exploring starts in case of non-deterministic policies to ensure adequate sampling

(A) Assertion and reason are both true (B) Assertion is false and reason are is true
(C) Assertion is true and reason is false (D) Both assertion and reason are false

10. If V satisfies $LV = L_\pi V = V$ (where L is the Bellman optimality operator L_π is the Bellman operator for value function of π), then 1 2 3 2

(A) π is ε optimal and V is the fixed point of π (B) π is not optimal and V is not the point of π
(C) π is an optimal policy and V is the fixed point of π (D) π is not optimal and V is the fixed point of the optimal policy

11. In policy iteration, which of the following is true of the policy evaluation (PE) and policy improvement (PI) steps? 1 2 3 2

(A) The values of states that are returned by PE may fluctuate between high and low values (B) PE does not returns the fixed point of L_{x_R}
(C) PI can randomly select any greedy policy for a given value function V^n (D) Policy iteration always converges for a finite Markov decision process (MDP)

12. The value of any state under an optimal policy is _____ the value of the state under a non-optimal policy 1 2 3 1

(A) Strictly greater than (B) Greater than or equal to
(C) Strictly less than (D) Less than or equal to

13. Which of the following is true? 1 2 4 1

(A) Dynamic programming methods use full backups and no bootstrapping
(B) Temporal difference methods use sample backups and bootstrapping
(C) Monte Carlo methods use sample backups and bootstrapping
(D) Monte Carlo methods use backups and no bootstrapping

14. Assertion: SARSA is an on-policy method 1 2 4 2
Reason: In SARSA, we do not update the action that was actually used, so it is on-policy

(A) Both assertion and reason are true (B) Assertion is false, reason is true
(C) Assertion is true, reason is false (D) Both assertion and reason are false

15. Which of the following statement is true? 1 2 4 2

(A) TD(0) methods uses unbiased sample of the return (B) TD(0) method does not uses a sample of the reward from the distribution of rewards
(C) TD(0) methods uses the current estimate of value function (D) TD(0) methods uses unbiased sample from the distribution of rewards

16. Actor-critic algorithm can be viewed as a method for training a _____ deep policy network by using a group of related source tasks. 1 1 4 1

(A) Single (B) Double
(C) Triple (D) Multiple

17. SARSA follows an _____ learning algorithm. 1 1 5 1

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

18. Actor-critic algorithm learns policies in _____. 1 2 5 2

(A) High dimensional, continuous action spaces (B) High dimensional, discrete action spaces
(C) Low dimensional, discrete action spaces (D) Low dimensional, continuous action spaces

19. Which of the following type of policy is a learning algorithm in which the same policy is improved and evaluated? 1 1 5 1

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

20. Which of the following algorithms will find the best course of action, based on the agents current state without using a model and off-policy reinforcement learning? 1 1 5 1

(A) Q-learning (B) Markov property
(C) State action reward state action (D) Deep Q-neural network

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

	Marks	BL	CO	PO

21. Illustrate the general idea of reinforcement learning with Tic-Tac-Toe and contrast it with other approaches. 4 2 1 1

22. Illustrate the episodic task using pole-balancing example. 4 3 2 2

23. Explain about generalized policy iteration. 4 2 3 2

24. Compare the prediction abilities of TD (0) and constant-∝ Monte Carlo method of the following Markov reward process. 4 3 4 3