National University of Computer and Emerging Sciences
Peshawar Campus

Reinforcement Learning (CS3018)

Final Exam

Total Time (Hrs): 3
Total Marks: 100
Total Questions: 9

Date: 19th May, 2025

Course Instructor(s)
Ms. Sara Rehmat

228- — BAI-6A
Roll No Section

Student Signature

Do not write below this line

Attempt all the questions.

CLO 1: Explain the foundational concepts and mathematical tools of reinforcement learning.

Question 1	Marks: 2*3=6	Estimated Time: 15 minutes

a. How is Policy Gradient algorithm different from Deep Q-learning?
b. Is REINFORCE a policy-based method or value-based method? Justify your answer.
c. Why is the approximation method using Temporal Difference called Semi-Gradient method?

Question 2	Marks: 8	Estimated Time: 15 minutes

Consider the following description of Tic-tac-toe game:

Tic-tac-toe is a simple two-player turn-based strategy game played on a 3×3 grid. Players take turns marking empty squares with their symbol: Player 1: "X", Player 2: "O". The goal is to be the first to form a straight line of three of your own symbols—either horizontally, vertically, or diagonally. If all 9 squares are filled and no player has formed such a line, the game ends in a draw.

a. Explain the markov property in the Markov Decision Process. **[2]**
b. Describe how a state can be represented in the game of Tic-tac-toe. Give the size of state space. **[4]**
c. Considering your state description, justify if the game of Tic-tac-toe can be classified as an MDP. **[2]**

Question 3	Marks: 10	Estimated Time: 20 minutes

a. Differentiate between on policy and off policy algorithms. Give an example of on-policy control and off-policy control algorithms. **[2]**

Consider the following algorithm:

```
Initialize:
  For all states s:
    V(s) ← 0
    C(s) ← 0     # Cumulative weight for state s
Loop forever (or for each episode):
  Generate an episode following behavior policy b:
  Episode = [(S0, A0, R1), (S1, A1, R2), ..., (ST-1, AT-1, RT)]

  G ← 0
  W ← 1
  For t = T-1 down to 0:
    G ← γ * G + R_{t+1}

    C(St) ← C(St) + W
    V(St) ← V(St) + [W / C(St)] * (G − V(St))

    If π(St) ≠ At:
      break

    W ← W * [π(At | St) / b(At | St)]
```

b. What does the term W represent and what role it plays in given algorithm? [4]

c. Classify the given as an on policy or off policy algorithm. Justify your answer. [2]

d. Classify the given as a control or prediction algorithm. Justify your answer. [2]

Question 4	Marks: 10	Estimated Time: 15 minutes

a. What are the limitations of using tabular Q-learning in large or continuous state spaces? [3]

b. How does Deep Q-Network (DQN) address this issue? [2]

c. Explain the role of experience replay (memory) in DQN. [2]

d. How is deep Q-learning different from deep supervised learning? [3]

CLO 2: Implement basic RL algorithms to solve standard benchmark problems..

GridWalker Scenario:

A mobile robot named GridWalker operates in a 3×3 indoor grid environment as shown below. Each cell represents a distinct location:

S_0	S_1	S_2
S_3	S_4	S_5
S_6	S_7	S_8 (terminal)

GridWalker receives a reward of −1 per step and 0 upon reaching the terminal state S_8. Transitions are deterministic. Invalid moves taking the agent off the grid leave the robot in the same state. GridWalker follows a stochastic policy where all actions (Up, Down, Left, Right) are chosen with equal probability. Consider learning rate to be 0.5 and discount factor to be 1.0. All state values V(s) are initialized to 0.

Question 5	Marks: 10	Estimated Time: 20 minutes

Considering the GridWalker scenario, use the Iterative Policy Evaluation to answer the following questions:

a. For two iterations, find the value of each state. **[5]**
b. Extract the policy assuming the values of states you determined in part a. **[3]**
c. Now consider that you assume the policy, determined in part b, to estimate the values of states until they converge using Iterative policy evaluation. You then again extract the policy based on the updated values and continue till there is no change in policy, what algorithm you'll be following. Moreover, will you classify the algorithm as prediction or control? Justify your answer. **[2]**

CLO 3: Evaluate and compare the performance of RL algorithms.

Question 6	Marks: 12	Estimated Time: 15 minutes

Considering the above scenario, GridWalker runs two episodes of experience:

Episode number	Trajectory	Rewards
1	S0 → S1 → S2 → S5 → S8	−1, −1, −1, 0
2	S0 → S3 → S6 → S7 → S8	−1, −1, −1, 0

Using Temporal Difference TD(0):

a. Apply updates for each transition in Episode 1, showing updated values of states after each step. **[4]**
b. Using the updated values from (a), apply updates for all states for Episode 2. **[4]**
c. Compare TD(0) with the Monte Carlo method for value estimation. **[2]**
d. Compare TD(0) with Iterative Policy Evaluation. **[2]**

Question 7	Marks: 4*3=12	Estimated Time: 20 minutes

a. Differentiate between planning and learning algorithms in the context of Reinforcement Learning.
b. Explain Dyna architecture and elaborate how it helps the agent.
c. Which problem of Dyna Q does Dyna Q+ solve and how?

CLO 4: Apply RL methods to real-world scenarios

Question 8	Marks: 12	Estimated Time: 20 minutes

A small e-commerce startup wants to determine the most effective marketing strategy to maximize customer engagement. The marketing team can choose from three different online ad strategies, each with an unknown but fixed probability of getting a customer to click: Ad A (Arm 1), Ad B (Arm 2) and Ad C (Arm 3).

To optimize performance, the team decides to model this as a multi-armed bandit problem, using the ε-greedy algorithm with epsilon = 0.2. After running for 10 rounds, they record the following rewards (clicks):

Round	Ad Shown	Click (Reward)
1	A	1
2	C	0
3	B	1
4	A	0
5	B	1
6	A	1
7	B	0
8	A	1
9	C	0
10	B	1

a. Compute the action values for choosing each ad (A, B, C). [6]
b. If the team continues to use ε-greedy for round 11, which ad is most likely to be selected (under exploitation)? What is the probability that a random ad is explored instead? [3]
c. How can exploration be further encouraged in the given scenario? [3]

Question 9	Marks: 20	Estimated Time: 30 minutes

You are given the following three episodes collected by a robot navigating a warehouse. Each step is represented as a (state, action, reward) tuple, and the arrow (→) shows the transition to the next state. Assume the discount factor γ=1.

Episode 1: (S1, a1, -1) → (S2, a1, -1) → (S3, a1, 0)

Episode 2: (S1, a1, -1) → (S4, a2, -1) → (S1, a1, -1) → (S3, a1, 0)

Episode 3: (S1, a1, -1) → (S2, a1, -1) → (S5, a1, 1)

a. Estimate the Q-value for the state-action pair (S1, a1) using both:

- First-Visit Monte Carlo Method
- Every-Visit Monte Carlo Method

b. Estimate the state-value function V(s) for each visited state using the First-Visit Monte Carlo.
c. Can we extract the optimal policy from the values of states using the Monte Carlo method? Justify your answer.