Birla Institute of Technology & Science, Pilani

Work Integrated Learning Programmes Division
First Semester 2023-2024
Comprehensive Test (EC-3 Regular)

Course No. : AIMLCZG512 **Course Title** :Deep Reinforcement Learning
Nature of Exam : Open Book **Weightage** : 40% **No. of Pages** = 2 ; **No. Of Questions =** 4;
Duration : 2:30 Hours / 150 Mins **Date of Exam** : 06-06-2024 (AN)

Note to Students:

1. Answer all the questions. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
2. **Write all the answers neatly in A4 papers, scan and upload them.**
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q1) [Answer parts and their subparts in the same sequence.]
Imagine you are an investor trying to optimize your trading strategy for four different stocks, labeled A, B, C, and D. Each stock has its own unique potential for profit, which is unknown to you. To maximize your returns over a series of 100 trades, you decide to implement an ε-greedy strategy with ε being 0.1. The actual returns from each stock follow these distributions:

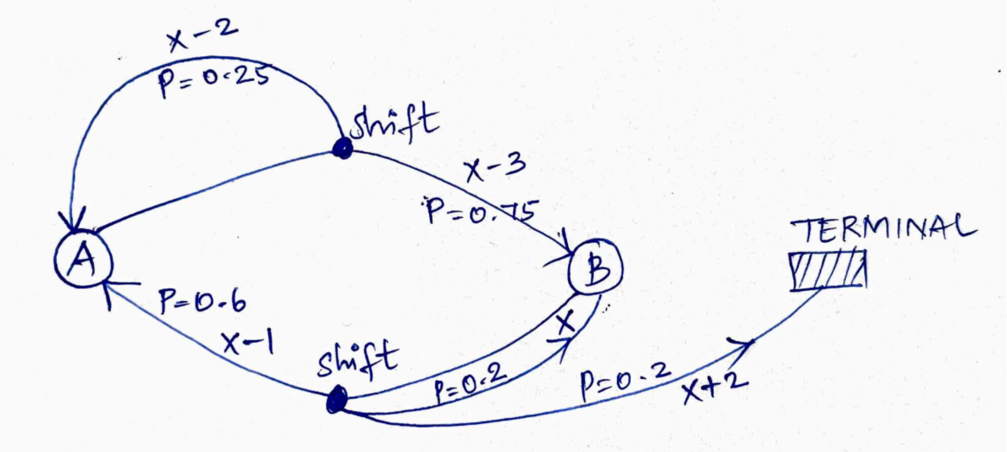
Stock A: 70% chance of +1 return, 30% chance of 0.
Stock B: 50% chance of +2 return, 50% chance of 0.
Stock C: 10% chance of +5 return, 90% chance of 0.
Stock D: Guaranteed return of +0.5.

Given this, answer the following questions

(a) Show how do you model this as a Reinforcement Learning Problem. [1 M]
(b) An investor intends to buy 100 times (each time buying one share of one stock). The strategies the investor chooses are (i) follow ε-greedy for the initial 25 tradings and only exploit the information for the next 75 purchases [1.5M] (ii) follow ε-greedy for the initial 75 tradings and only exploit the information for the next 25 purchases [1.5M] (iii) Follow ε-greedy for all the 100 purchases.[1.5M] Support the investor with your analysis. Show all the steps, tabulate answers to all the options and write your conclusion. [1M]
(c) What are MDP, POMDP and CMDP? What are they? Suggest one RL technique that is used to solve problems stated using them. It is adequate if you write just one/two lines for each. [1.5M]

[1+1.5+1.5+1+1.5 = 8 M]

Q2) [Answer parts and their subparts in the same sequence.]
Consider the MDP given below containing 2 states A and B with action *Shift* that may result in A,B or terminal state.. The rewards obtained are as indicated along the edges in the figure (X-2,X-3,X-1,X,X+2). *Treat the value of X to be 6.* The transition probabilities are as given along the edges. Let the discount factor be 0.4.



(a) Evaluate the given deterministic policy where the shift always executes the higher probability action. Improve it up to 1 iterations. Use Dynamic Programming solution to MDP.[4.0 M]

(b) Using value iteration of dynamic programming, determine the values of states A and B. Let the values of A and B be initialized to 1. Show 1 Iteration. [4.0 M]

[4+4= 8 Marks]

Q3) [Answer parts and their subparts in the same sequence.]

(a) What are the two most important issues when you have to learn the value function using a first-visit Monte Carlo using for a deterministic policy.[2.0 M] Explain. Also, provide possible solutions. [1.5 M].

(b) Explain any 3 most significant action selection strategies used in RL and mention how each selection method balances exploration and exploitation. Provide your answer as a table. [3 M]

(c) If we utilize a policy gradient method to address a reinforcement learning problem and find that the policy it provides is not optimal, what could be the possible explanations for this? State the most relevant 3 reasons. [1.5 M]

[2 + 1.5+ 3+ 1.5= 8 Marks]

Q4) [Answer parts and their subparts in the same sequence.] For each of the questions answer in not more than 4 precise statements. Vague Answers will not be accepted.

(a) Why AlphaGo use a separate policy network and a separate value network? [1.0 M]

(b) How does the MCTS ensure an action with the highest value is found in real-time? If the best action can be selected only by MCTS, why is any prior learning of Q(s,a) required? [2.0 M]

(c) We have learned that Supervised Learning that learns with samples from a given distribution does not capture the online nature of interactions as required for reinforcement learning quite well.

(i) Why does AlphaGo use supervised learning to learn the initial policy (and even further)? [1.5 M]

(ii) In what ways the shortcomings of supervised learning are mitigated in AlphaGo? [2.0 M]

(d) How does DQN handle the challenges referred to in the c part of this question? [1.5 M]

[1+2+1.5+2+1.5 = 8 Marks]

Q5) [Answer parts and their subparts in the same sequence.]

(a) Consider the following ways of organizing reinforcement learning techniques. (i) Model-Based vs. Model Free;(ii) Value-based vs. Policy-Based; (iii) On-Policy vs. Off-Policy. Write a statement or two on each of the points(for both categories) explaining the kind of problems those RL techniques are suited to. Provide your response in a neatly organized table. [3 M]

(b) Consider the learning scenario. A human expert is presented with two trajectories taken by two drivers in a highway stretch. The human expert marks which of the trajectories is better. The agent learns this expertise (to decide a better trajectory by giving two unseen trajectories) observing the expert's decision from many such examples. Explain how you precisely model this as an appropriate RL problem [3 M]. Show all the elements of your modeling making necessary assumptions [2 M]. [Note: Only the most appropriate modeling gets the credit.]

[3+3+2 = 8 Marks]