Time: 03 Hours **Marks: 80**

Note: 1. Question 1 is compulsory

2. Answer any three out of the remaining five questions.

3. Assume any suitable data wherever required and justify the same.

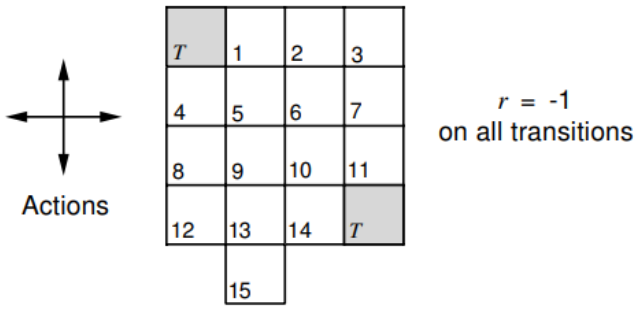
Q1 a) What are model-based and model-free reinforcement methods? [5]

b) Suppose $\gamma = 0.5$ and the following sequence of rewards is received R1 = 1, R2 = 2, R3 = 6, R4 = 3, and R5 = 2, with T = 5. What are G0, G1, ..., G5? [5]

c) What is policy iteration? Explain policy iteration algorithm. [5]

d) Explain first-visit Monte Carlo and every-visit Monte Carlo methods. [5]

Q2 a) Suppose a new state 15 is added to the gridworld just below state 13, and its actions, left, up, right, and down, take the agent to states 12, 13, 14, and 15, respectively. Assume that the transitions from the original states are unchanged. What, then, is V^{π} (15) for the equiprobable random policy? Now suppose the dynamics of state 13 are also changed, such that action down from state 13 takes the agent to the new state 15. What is V^{π} (15) for the equiprobable random policy in this case? [10]



b) What is agent and environment? Explain agent–environment interaction in a Markov decision process. [10]

Q3 a) Explain how upper confidence bound (UCB) action selection generally performs better than ε-greedy action selection with a suitable example. [10]

b) Explain Q-learning algorithm to learn optimal action value function. [10]

Q4 a) Consider a k-armed bandit problem with k = 4 actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ε -greedy action selection, sample-average action-value estimates, and initial estimates of Q1(a) = 0, for all a. Suppose the initial sequence of actions and rewards is A1 = 1, R1 =1, A2 = 2, R2 = 1, A3 = 2, R3 =2, A4 = 2, R4 = 2, A5 = 3, R5 = 0. On some of these time steps the ε case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred? [10]

b) What are Goals and Rewards? Explain with a suitable example. [10]

Q5 a) Differentiate between Monte-Carlo prediction and TD prediction. [10]

b) Describe asynchronous dynamic programming with an example. [10]

Q6 a) Give similarities and differences between Q-learning and SARSA algorithms. [10]

b) Describe the application of reinforcement learning to the real world problem of elevator dispatching. [10]

****