Indian Institute of Technology, Kanpur
Department of Electrical Engineering
Introduction to Reinforcement Learning (EE932)
Theory Assignment 2

Deadline: 4th May 2024 | 2023-24 Quarter 4 | Max points: 25

1 Objective Questions [8 points]

1. Suppose we want an RL agent to learn to play the game of golf. For training purposes, we make use of a golf simulator program. Assume that the original reward distribution gives a reward of +10 when the golf ball is hit into the hole and -1 for all other transitions. To aide the agents learning process, we propose to give an additional reward of +3 whenever the ball is within a 1 meter radius of the hole. Is this additional reward a good idea or not? Why?
   (a) Yes, the additional reward will help speed-up learning.
   (b) Yes, getting the ball to within a metre of the hole is like a sub-goal and hence, should be rewarded.
   (c) No, the additional reward may actually hinder learning.
   (d) No, it violates the idea that a goal must be outside the agent's direct control. **(1 pt)**

2. Which among the following is/are the differences between contextual bandits and full RL problems?
   (a) The states in contextual bandits share features, but not in full RL problems.
   (b) The actions in contextual bandits do not determine the next state, but typically do in full RL problems.
   (c) Full RL problems can be modeled as MDPs whereas contextual bandit problems cannot.
   (d) No difference **(1 pt)**

3. Recall the definition of the return $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$. Suppose $\gamma = 0.5$, and the following sequence of rewards is received in an episode: $R_1 = 1, R_2 = 2, R_3 = 6, R_4 = 3, R_5 = 2$. What are the returns $G_0, G_1, G_2, G_3, G_4, G_5$? Hint: Work backwards. **(2 pt)**

4. Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2$ followed by an infinite sequence of $7, 7, \ldots$. What are G_0 and G_1? **(2 pt)**

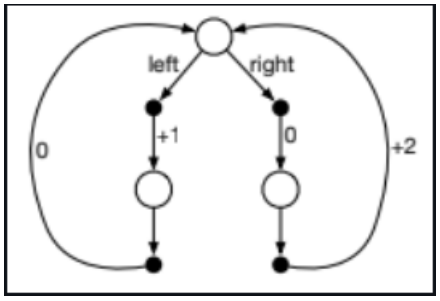
5. Consider a 100x100 grid world domain where the agent starts each episode in the bottom-left corner, and the goal is to reach the top-right corner in the least number of steps. To learn an optimal policy to solve this problem you decide on a reward formulation in which the agent receives a reward of +1 on reaching the goal state and 0 for all other transitions. Suppose you try two variants of this reward formulation, (P_1), where you use discounted returns with $\gamma < 1$, and (P_2), where no discounting is used, i.e., $\gamma = 1$. Which among the following would you expect to observe?
   (a) the same policy is learned in (P_1) and (P_2)
   (b) no learning in (P_1)
   (c) no learning in (P_2)
   (d) policy learned in (P_2) is better than the policy learned in (P_1) **(2 pt)**

2 Subjective Questions [17 points]

6. The weather of a day could be either cloudy or sunny. Assume that a day's weather can be modelled using a Markov Chain, i.e., Given today's weather, tomorrow's weather does not depend on yesterday's (or any other previous day's) weather. Let the probability that tomorrow's weather is the same as today be 0.7. Taking 'Sunny' and 'Cloudy' as the states, draw the transition diagram of the Markov Chain. Upload the answer as an image. **(2 pt)**

7. Devise two example tasks of your own that fit into the MDP framework, identifying its states, actions, and rewards for each. Make the two examples as different from each other as possible. The framework is abstract and flexible and can be applied in many ways. Stretch its limits in some way in at least one of your examples. **(2 pt)**

8. Consider the continuing MDP shown. The only decision to be made is in the top state, where two actions are available, left and right. In the other two states, only one action is available, and hence there is nothing to decide. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, π_{left} and π_{right}. What policies are optimal for the three cases given below? Show your calculations and upload an image.
   Case 1: $\gamma = 0$, Case 2: $\gamma = 0.9$, Case 3: $\gamma = 0.5$. **(3 pt)**



9. In the Bellman expectation equation, we related V^π in terms of V^π. Write down an equation that expresses Q^π in terms of Q^π. Hint: Refer to the Week 3 (Part 2) slide with the title "Relating Q^π and V^π. **(2 pt)**

10. Consider the grid shown in Figure 1. The states are grid squares, identified by their row and column numbers. The agent always starts in the bottom left state (1,1), marked with the letter S. (Note that the bottom row is denoted by number 1, the top row by number 2). There are two terminal goal states, (2,3) with reward +5 and (1,3) with reward -5. Rewards are 0 in non-terminal states. (The reward for a state is received as the agent moves into the state.) The transition function is such that the intended agent movement (UP, Down, Left, or Right) happens with probability 0.8. With probability 0.1 each, the agent ends up in one of the states perpendicular to the intended direction. Please refer Figure 2.

    1. Draw the usual MDP state diagram like this figure containing all the states with clearly showing the actions and transition probabilities for one of its state (1,2), i.e., the state in between source cell and -5 cell. (2 pt)
    2. Assume $\gamma = 1$, and make an intelligent guess for the optimal policy for this grid. (2 pt)
    3. What is the optimal value function for each state? You can answer this by intuition. (2 pt)
    4. Verify that the optimal value function that you intuitively guessed is indeed correct by using the Bellman optimality equations. (2 pt)

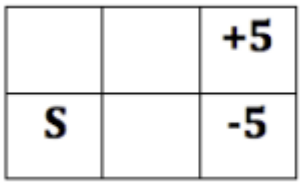


Figure 1: Grid-World

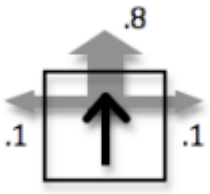


Figure 2: Tranistion probabilities for an 'UP' action