Mid-term Exam I: Reinforcement Learning

Answer all the Questions; Total Marks: 30

1. Suppose two identical and perfectly balanced coins are tossed once.

   (a) Find the conditional probability that both coins show a head given that the first shows a head. (2.5) ✓

   (b) Find the conditional probability that both are heads given that at least one of them is a head. (2.5) ✓

2. Consider a multi-armed bandit with three arms. We denote by r_1 the reward obtained by pulling arm 1, r_2 the reward obtained by pulling arm 2 and r_3 the reward obtained by pulling arm 3. The distribution of reward r_1 of the first arm is given by

$$r_1 = \begin{cases} 4 & \text{w.p. } 3/4 \\ 100 & \text{w.p. } 1/4 \end{cases}$$

   Similarly, the distribution of reward r_2 of the second arm is given by

$$r_2 = \begin{cases} 4 & \text{w.p. } 1/2 \\ 100 & \text{w.p. } 1/2 \end{cases}$$

   Finally, the distribution of reward r_3 of the third arm is given by

$$r_3 = \begin{cases} 4 & \text{w.p. } 1/4 \\ 100 & \text{w.p. } 3/4 \end{cases}$$

   (a) Find the expected rewards $E[r_1]$, $E[r_2]$ and $E[r_3]$ of the three arms 1, 2 and 3? (1.5)

   (b) Which arm would you like to pull more often and why? (1)

   (c) Find the reward-variance for each of the three arms? (1.5)

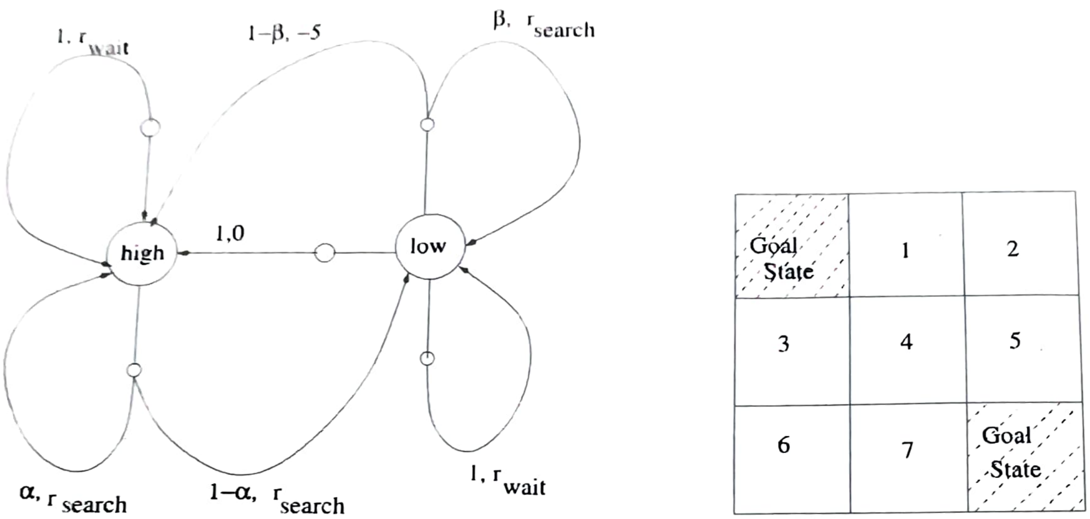
   (d) Which arm would you pull more often so as to minimize the reward-variance? (1)

   (e) Consider the situation where you don't know the reward distributions but can only observe the rewards obtained by pulling any of the three arms. Describe an incremental update procedure that estimates the variance of the arms alongwith an ϵ-greedy action selection to choose the arm with least variance? (3)

3. Recall the example of recycling robot from Chapter 3. The state transition diagram along with the possible actions is given in the Figure 1(a) below.

   (a) Write down the Bellman equation for this problem (for both states high and low) for the following parameter settings: $r_{\text{search}} = 5$, $r_{\text{wait}} = 1$, $\gamma = 0.9$, $\alpha = 0.3$ and $\beta = 0.5$, respectively. (4)

   (b) Suppose it is optimal for the robot to actively search for cans when the state is high and recharge when the state is low. Find in this case, the optimal values $v^*(\text{high})$ and $v^*(\text{low})$ for the two states. (2)



(a) The Recycling Robot

(b) The 3 × 3 Grid World Example

Figure 1: Changing operating conditions for autonomous agents.

4. Consider the 3×3 grid world example shown in the Figure 1(b) above. Here the two shaded states are the goal states. For the edge states (i.e., all states except state 4), an action that can take the agent out of the grid is penalized with a reward of -2 but with no change in state. Thus, if the state is 2, then either the right or the top action will result in a reward of -2 with the next state remaining as 2 itself. Consider the equiprobable policy for deciding on the actions to be chosen. For all other actions in any state (including state 4), the reward is -1.

   (a) Using relative policy evaluation starting with an initial value estimate of 0 for all states, for the equiprobable policy, find the value function estimates for three successive steps, i.e., find $v_1(s)$, $v_2(s)$ and $v_3(s)$, respectively, for each of the states $s = 1, \ldots, 7$. (3)

   (b) At each of the three stages of update of the value function above, i.e., $k = 1, 2, 3$, identify the greedy action as suggested by the correspomnding value function updates. (3)

5. Suppose we have a machine that is either running or is broken down. If it runs throughout one week, it makes a gross profit of \$100. If it fails during the week, gross profit is zero. If it is running at the start of the week and we perform preventive maintenance, the probability that it will fail during the week is 0.4. If we do not perform such maintenance, the probability of failure is 0.7. However, maintenance will cost \$20. When the machine is broken down at the start of the week, it may either be repaired at cost of \$40, in which case it will fail during the week with a probability of 0.4, or it may be replaced at a cost of \$150 by a new machine that is guaranteed to run through its first week of operation. Let the discount factor be $\gamma < 1$.

   (a) Formulate the problem in the framework of infinite-horizon MDP. Clearly specify State space, Action Space, transition probabilities and single stage rewards. (3)

   (b) Write down the Bellman equation for the MDP formulated above. (2)