Practice Questions: Reinforcement Learning

1. Consider a multi-armed bandit with five arms with the ith arm associated with a parameter p_i that decides the reward which one gets by pulling arm i. Specifically assume that the reward obtained from pulling arm i is obtained as follows: When arm i is pulled at time t, a coin for which heads occurs with probability p_i is tossed once. The reward R^i_{t+1} obtained by pulling arm i with $i = 1, \ldots, 5$, is obtained according to

$$r^i_{t+1} = \begin{cases} 1 - p_i & \text{if coin toss results in heads} \\ 0 & \text{otherwise} \end{cases}$$

Let $p_1 = 1/4$, $p_2 = 1/3$, $p_3 = 1/2$, $p_4 = 2/3$ and $p_5 = 3/4$ denote the values of p_i for the five arms respectively.

   (a) Suppose the goal of the decision maker is to pull the arm that gives the maximum expected reward. Assume the decision maker knows the reward distribution as described above. Then find the arm that the decision maker should pull? In other words, find the arm that gives the maximum expected reward.

   (b) For a given arm i, suppose we can decide the parameter to be any value $p_i \in [0, 1]$, not necessarily restricted to one of the above mentioned values. Using basic calculus, find the parameter p_i that maximizes the expected reward $E[R^i_{t+1}]$ for this arm?

2. Consider the 3×3 grid world example shown in the figure below. 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.

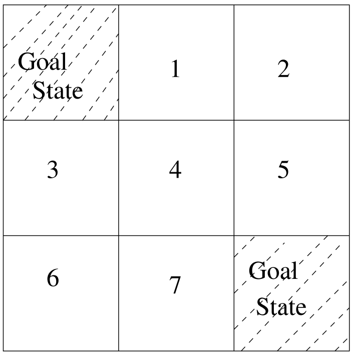

Figure 1: The 3×3 Grid World Example

   (a) Using 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 $s = 1, \ldots, 7$.

   (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 corresponding value function updates.

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

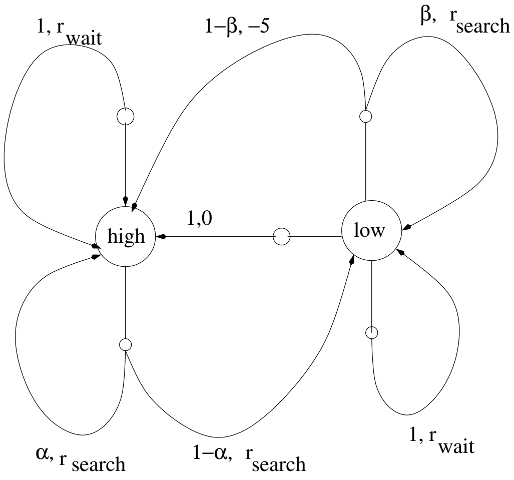

Figure 2: The Recycling Robot

(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.

(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 values $v^*(\text{high})$ and $v^*(\text{low})$ for the two states.