Quiz 5

Please write all answers on these pages.

1. Describe the four components of a Markov Decision Process. For components that are functions, describe also what their arguments are.

Solution:
State space $S$
Action space $A$
Transition function $\delta(s, a) \rightarrow s'$
Reward function $r(s, a) \rightarrow \text{reward}$
2. Consider the following $Q$ matrix for Robby the Robot:

<table>
<thead>
<tr>
<th>Actions</th>
<th>MoveN</th>
<th>MoveS</th>
<th>MoveE</th>
<th>MoveW</th>
<th>Pick Up Can</th>
</tr>
</thead>
<tbody>
<tr>
<td>State: North, South, East, West, Here</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>W, C, E, E, E</td>
<td>-5</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>W, E, C, W, E</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>W, C, E, E, C</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The notation for state “W, C, E, E, C” represents the following: **North:** Wall; **South:** Can, **East:** Empty; **West:** Empty; **Here:** Can.

Suppose Robby is in state “W, C, E, E, C” and he picks up a can for a reward of 10 points. Given the update formula $Q(s, a) = Q(s, a) + \eta (r + [\gamma \max_{a'} Q(s', a')] - Q(s, a))$, what is the updated entry for this state, action pair?

Assume $\gamma = 0.9$ and $\eta = 0.2$.

**Solution:**

$s' = W, C, E, E, E$

$Q(s, a) = 0 + 0.2(10 + (0.9)(10) - 0) = 3.8$
3. Referring to the Q-matrix of Question 2, suppose Robby is in state W, C, E, E, E, and is following an epsilon-greedy action-selection strategy, with \( \epsilon = 0.4 \).

(a) Explain how Robby will choose an action.

**Solution:**

With probability \( \epsilon = 0.4 \), Robby will choose a random action from his list of possible actions. With probability \( (1 - \epsilon) = 0.6 \), Robby will choose the action with the highest \( Q \) value, namely MoveS.

(b) What is the probability Robby will choose action MoveNorth?

**Solution:**

\[ (0.4) (0.2) = 0.08. \]
4. Agent A lives in a $3 \times 2$ grid (pictured below), and is using Q-learning, with the usual update formula:

$$Q(s,a) \leftarrow Q(s,a) + \eta (r + \gamma \max_{a'} Q(s',a') - Q(s,a)),$$

The states correspond to the numbered squares. The agent receives $5$ every time she lands in state 2, and $10$ every time she lands in state 6. There are no other rewards or penalties.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>$5$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

Her possible actions are MoveNorth, MoveSouth, MoveEast, MoveWest.

The Q matrix is initialized to all zeros.

(a) During episode 1, Agent A starts in square 1 and performs the following actions: MoveSouth, MoveEast, MoveEast, MoveNorth, MoveWest, Move West.

Assuming a learning rate $\eta = 1$ and discount rate $\gamma = 0.9$, give the nonzero entries of the Q matrix at the end of this episode.

**Solution:**

$$Q(1, \text{South}) = 0 + (10 + (0.9)(0) - 0) = 10$$
$$Q(3, \text{West}) = 0 + (5 + (0.9)(0) - 0) = 5$$
$$Q(2, \text{West}) = 0 + (0 + (0.9)(10) - 0) = 9$$

(b) During episode 2, Agent A performs the same sequence of actions as in episode 1. Give the nonzero entries of the Q matrix at the end of this episode.

**Solution:**

$$Q(1, \text{South}) = 10 + (10 + (0.9)(0) - 10) = 10$$
$$Q(4, \text{North}) = 0 + (0 + (0.9)(5) - 0) = 4.5$$
$$Q(3, \text{West}) = 5 + (5 + (0.9)(9) - 5) = 13.1$$
$$Q(2, \text{West}) = 9 + (0 + (0.9)(10) - 9) = 9$$