Assignment 4
Reinforcement Learning
Prof. B. Ravindran

1. State True/False

   The state transition graph for any MDP is a directed acyclic graph.

   (a) True

   (b) False

   Sol. (b)
   The statement is false. There is a possibility of transitioning to the same state, as well as having other cycles.

2. Consider the following statements:

   (i) The optimal policy of an MDP is unique.

   (ii) We can determine an optimal policy for a MDP using only the optimal value function(v^*), without accessing the MDP parameters.

   (iii) We can determine an optimal policy for a given MDP using only the optimal q-value function(q^*), without accessing the MDP parameters.

   Which of these statements are true?

   (a) Only (ii)

   (b) Only (iii)

   (c) Only (i), (ii)

   (d) Only (i), (iii)

   (e) Only (ii), (iii)

   Sol. (b)
   Optimal policy can be recovered from an optimal q-value function. Also, a given MDP can have multiple optimal policies.

3. Which of the following is a benefit of using RL algorithms for solving MDPs?

   (a) They do not require the state of the agent for solving a MDP.

   (b) They do not require the action taken by the agent for solving a MDP.

   (c) They do not require the state transition probability matrix for solving a MDP.

   (d) They do not require the reward signal for solving a MDP.

   Sol. (c)
   RL algorithms require to know the state the agent is in, the action it takes and a reward signal from the environment to solve the MDP. However, they do not need to know the state transition probability matrix.

4. Consider the following equations:

(i) $v^{\pi}(s) = \mathbb{E}_{\pi}[\sum_{i=t}^{\infty} \gamma^{i-t} R_{i+1} | S_t = s]$

(ii) $q^{\pi}(s,a) = \sum_{s'} p(s'|s,a) v^{\pi}(s')$

(iii) $v^{\pi}(s) = \sum_{a} \pi(a|s) q^{\pi}(s,a)$

Which of the above are correct?

(a) Only (i)

(b) Only (i), (ii)

(c) Only (ii), (iii)

(d) Only (i), (iii)

(e) (i), (ii), (iii)

Sol. (d)
(i) is the definition of $v^{\pi}(s)$ and (iii) follows from definition of $v^{\pi}(s)$ and $q^{\pi}(s,a)$. (ii) doesn't contain the immediate reward term and hence is wrong.

5. State True/False

While solving MDPs, in case of discounted rewards, the value of γ(discount factor) cannot affect the optimal policy.

(a) True

(b) False

Sol. (b)
With the change in γ value, the expected return of any state could change and thus, the optimal policy could change.

6. Consider the following statements for a finite MDP (I is an identity matrix with dimensions $|S| \times |S|$(S is the set of all states) and P_{π} is a stochastic matrix):

(i) MDP with stochastic rewards may not have a deterministic optimal policy.

(ii) There can be multiple optimal stochastic policies.

(iii) If $0 \leq \gamma < 1$, then rank of the matrix $I - \gamma P_{\pi}$ is equal to $|S|$.

(iv) If $0 \leq \gamma < 1$, then rank of the matrix $I - \gamma P_{\pi}$ is less than $|S|$.

Which of the above statements are true?

(a) Only (ii), (iii)

(b) Only (ii), (iv)

(c) Only (i), (iii)

(d) Only (i), (ii), (iii)

Sol. (a)
Check the lectures, it states that there always exists a deterministic optimal policy.
Lectures provide an example of multiple stochastic optimal policies.
$I - \gamma P_{\pi}$ will have non zero eigenvalues, so the rank will be equal to the number of rows which is $|S|$

7. Consider an MDP with 3 states A, B, C. From each state, we can go to either of the two states, i.e, from state A, we can perform 2 actions, that lead to state B and C respectively. The rewards for all the transitions are: $r(A, B) = 2$ (reward if we go from A to B), $r(B, A) = 5$, $r(B, C) = 7$, $r(C, B) = 10$, $r(A, C) = 1$, $r(C, A) = 12$. The discount factor is 0.7. Find the value function for the policy given by: $\pi(A) = C$ (if we are in state A, we choose the action to go to C), $\pi(B) = A$ and $\pi(C) = B$ ($[v^\pi(A), v^\pi(B), v^\pi(C)]$).

   (a) [10.2, 16.7, 20.2]

   (b) [14.2, 16.5, 15.1]

   (c) [15.9, 16.1, 21.3]

   (d) [12.2, 6.2, 14.5]

   Sol. (c)
   We can just substitute the options to find out which one is a fixed point of the Bellman equation, alternatively, compute $(I - \gamma P_\pi)^{-1} r_\pi$

   Note: $P_\pi = [[0\ , 0\ , 1], [1\ , 0\ , 0], [0\ , 1\ , 0]]$; $r_\pi = [1\ , 5\ , 10]^T$; $\gamma = 0.7$

8. Suppose x is a fixed point for the function A, y is a fixed point for the function B, and $x = BA(x)$, where BA is the composition of B and A. Consider the following statements:

   (i) x is a fixed point for B

   (ii) $x = y$

   (iii) $BA(y) = y$

   Which of the above must be true?

   (a) Only (i)

   (b) Only (ii)

   (c) Only (i), (ii)

   (d) (i), (ii), (iii)

   Sol. (a)
   $x = B(A(x)) \implies x = B(x)$. Therefore, x is a fixed point of B.
   However, that does not mean $x = y$. The function B could have multiple fixed points (consider the identity function), so (ii) is False.
   There is no guarantee that y is a fixed point for A. So, we cannot say (iii) is True.

9. Which of the following is not a valid norm function? (x is a D dimensional vector)

   (a) $max_{d \in \{1,\ldots,D\}} |x_d|$

   (b) $\sqrt{\Sigma_{d=1}^{D} x_d^2}$

   (c) $min_{d \in \{1,\ldots,D\}} |x_d|$

   (d) $\Sigma_{d=1}^{D} |x_d|$

   Sol. (c)
   (c) can be zero when x is not the zero vector.

10. Which of the following is a contraction mapping in any norm?

    (a) $T([x_1, x_2]) = [0.5x_1, 0.5x_2]$

    (b) $T([x_1, x_2]) = [2x_1, 2x_2]$

    (c) $T([x_1, x_2]) = [2x_1, 3x_2]$

    (d) $T([x_1, x_2]) = [x_1 + x_2, x_1 - x_2]$

    Sol. (a)
    (a) is a contraction mapping in any norm as $\|Tu - Tv\| = 0.5\|u - v\|$.