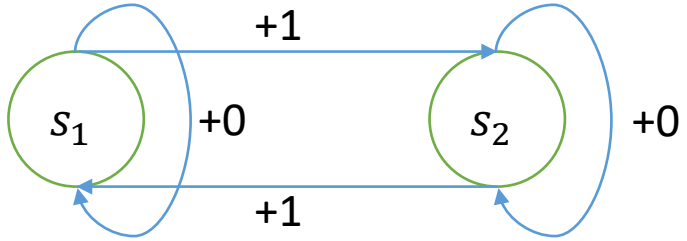
Q1-1: Assume that we have the current $\hat{Q}(s,a)$ as follows, and we are using a greedy update, i.e. $\hat{Q}(s,a) = r + \gamma \max_{a'} \hat{Q}(s',a')$ in the Q learning process, for the following MDP. Here we choose $\gamma = 0.9$, and the MDP has two actions: a_1 (move) and a_2 (stay), with rewards $r_1 = 1$ and $r_2 = 0$ respectively.
Suppose we are currently at the state s_1, and selecting the action a_1, please calculate the new $\hat{Q}(s_1, a_1)$.

1. 9.1
2. 8.1
3. 10
4. 9

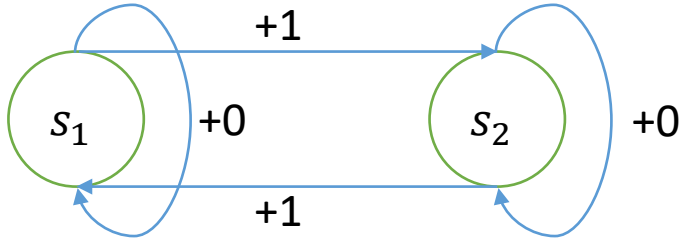


$\hat{Q}(s,a)$	a_1	a_2
s_1	10	9
s_2	9	10

Q1-1: Assume that we have the current $\hat{Q}(s,a)$ as follows, and we are using a greedy update, i.e. $\hat{Q}(s,a) = r + \gamma \max_{a'} \hat{Q}(s',a')$ in the Q learning process, for the following MDP. Here we choose $\gamma = 0.9$, and the MDP has two actions: a_1 (move) and a_2 (stay), with rewards $r_1 = 1$ and $r_2 = 0$ respectively.
Suppose we are currently at the state s_1, and selecting the action a_1, please calculate the new $\hat{Q}(s_1,a_1)$.

1. 9.1
2. 8.1
3. 10 ⬅
4. 9

$$\hat{Q}(s_1,a_1) = r_1 + \gamma \max_{a'} \hat{Q}(s_2,a')$$
$$= 1 + 0.9 * 10 = 10$$



$\hat{Q}(s,a)$	a_1	a_2
s_1	10	9
s_2	9	10

Q1-2: Are these statements true or false?
(A) When we select a random action (do exploration), the outcome would be better than the outcome when following the current policy.
(B) Suppose that we have an optimal policy π^* over the history of the environment response, then it's optimal for us to keep doing exploitation with π^*.

1. True, True
2. True, False
3. False, True
4. False, False

Q1-2: Are these statements true or false?
(A) When we select a random action (do exploration), the outcome would be better than the outcome when following the current policy.
(B) Suppose that we have an optimal policy π^* over the history of the environment response, then it's optimal for us to keep doing exploitation with π^*.

1. True, True
2. True, False
3. False, True
4. False, False ⟵

(A) Since it is a random action, we don't really know about its outcome, it could be better or it could be worse than following the current policy.
(B) Since π^* is only optimal for some history of the environment response, it is usually NOT optimal for true environment response. Moreover, the environment can also be changing (not fixed) over time.

Q2-1: Are these statements true or false?
(A) Compared to Q learning with a table, using a compact Q function can save us memory usage.
(B) We need to set up different compact Q function representations for different states.

1. True, True
2. True, False
3. False, True
4. False, False

Q2-1: Are these statements true or false?
(A) Compared to Q learning with a table, using a compact Q function can save us memory usage.
(B) We need to set up different compact Q function representations for different states.

1. True, True
2. True, False ⟵
3. False, True
4. False, False

(A) A compact Q function usually provides a compact representation with fewer parameters compared to Q tables, so it saves us memory usage.
(B) Since compact Q function can generalize across states, it is possible for different states to use the same compact representation.

Q2-2: Are these statements true or false?
(A) We can use linear regression to represent Q functions.
(B) When doing Q learning with function approximation, we should select the action which maximizes the Q function prediction given the current state.

1. True, True
2. True, False
3. False, True
4. False, False

Q2-2: Are these statements true or false?
(A) We can use linear regression to represent Q functions.
(B) When doing Q learning with function approximation, we should select the action which maximizes the Q function prediction given the current state.

1. True, True
2. True, False ⟵
3. False, True
4. False, False

(A) Linear approximation can be simple and efficient for Q learning for many problem. But sometimes it may not be that effective because many problems are non-linear.
(B) We need to ensure some randomization for exploration.

Q3-1: Are these statements true or false for the autonomous helicopter example?
(A) The reward function is the similarity to the trajectory demonstrated by expert pilots.
(B) We learn a dynamic model to directly predict the next state by linear regression.

1. True, True
2. True, False
3. False, True
4. False, False

Q3-1: Are these statements true or false for the autonomous helicopter example?
(A) The reward function is the similarity to the trajectory demonstrated by expert pilots.
(B) We learn a dynamic model to directly predict the next state by linear regression.

1. True, True
2. True, False
3. False, True
4. False, False ⟵

(A) The trajectory demonstrated by expert pilots are usually sub-optimal, so we need to infer the implicit desired trajectory and learn a reward function on the desired trajectory.
(B) We learn a dynamic model to predict the accelerations, and then we integrate the accelerations to get the next state.

Q3-2: Are these statements true or false for the autonomous helicopter example?
(A) We can use EM to infer the desired trajectory based on the trajectory demonstrated by expert pilots.
(B) We can directly use the RL methods we learnt so far to find the optimal control policy for autonomous helicopter.

1. True, True
2. True, False
3. False, True
4. False, False

Q3-2: Are these statements true or false for the autonomous helicopter example?
(A) We can use EM to infer the desired trajectory based on the trajectory demonstrated by expert pilots.
(B) We can directly use the RL methods we learnt so far to find the optimal control policy for autonomous helicopter.

1. True, True
2. True, False ⟵
3. False, True
4. False, False

(A) As is shown in the lecture.
(B) The RL methods we learnt so far are for problems where the state and action space are discrete, but here the helicopter problem has continuous states and actions.