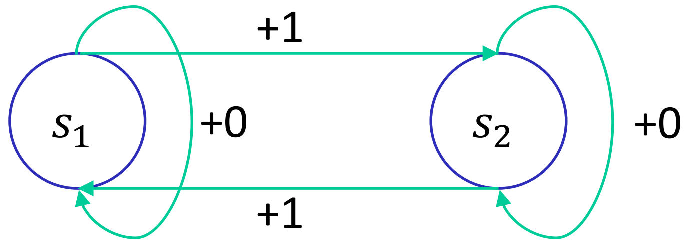
Q1-1: Assume that we have the current $\hat{Q}(s,a)$ as follows, and we are using a greedy update, i.e. $\hat{Q}(s,a) = r + \gamma \max_{a'} \hat{Q}(s',a')$ in the Q learning process, for the following MDP. Here we choose $\gamma = 0.9$, and the MDP has two actions: a_1 (move) and a_2 (stay), with rewards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



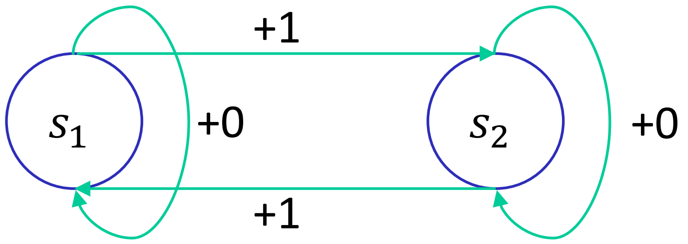
1. 9.1
2. 8.1
3. 10
4. 9

$\hat{Q}(s,a)$	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

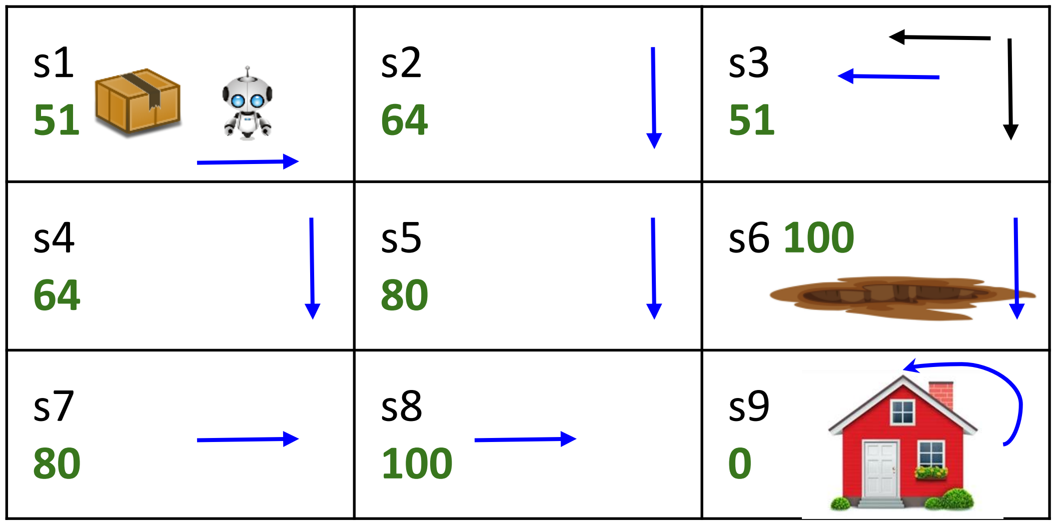
$$\hat{Q}(s_1,a_1) = r_1 + \gamma \max_{a'} \hat{Q}(s_2,a')$$
$$= 1 + 0.9 * 10 = 10$$



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

Q2-1: A robot wants to deliver a package from warehouse at s1 to a home at s9. However, it wants to avoid trench (present at s6). In the figure, the green numbers are the optimal V*(s), the blue arrows are the optimal policy, and the black arrows are the possible actions from s3. How can you get V*(s3) using Q(s, a)? Assume discount factor γ = 0.8 and rewards as follows:

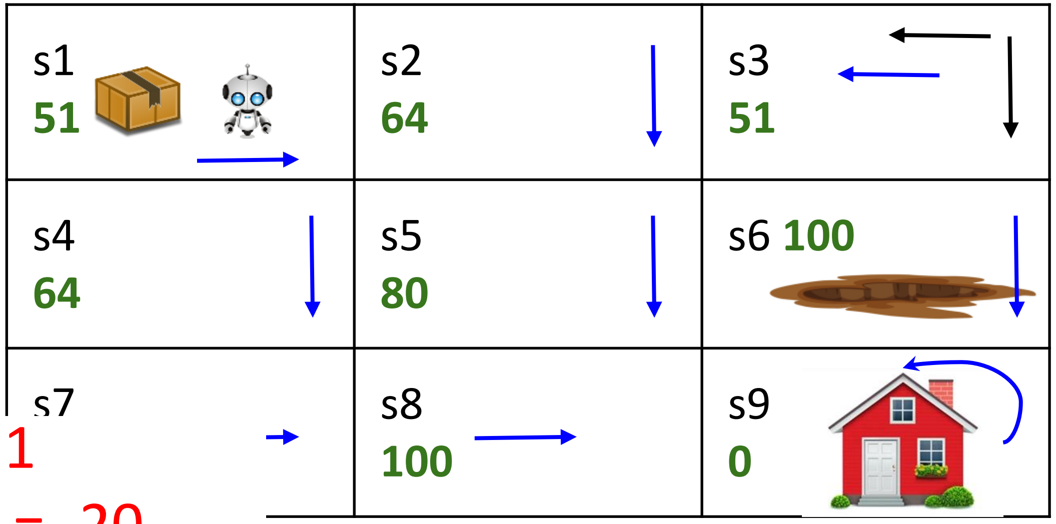
- r(s, a) = -100 if entering the trench 
- r(s, a) = +100 if entering home 
- r(s, a) = 0 otherwise



1. max {51, 0}
2. max {51, -20}
3. max {51, -80}
4. max {51, -100}

Q2-1: A robot wants to deliver a package from warehouse at s1 to a home at s9. However, it wants to avoid trench (present at s6). In the figure, the green numbers are the optimal V*(s), the blue arrows are the optimal policy, and the black arrows are the possible actions from s3. How can you get V*(s3) using Q(s, a)? Assume discount factor γ = 0.8 and rewards as follows:

- r(s, a) = -100 if entering the trench 
- r(s, a) = +100 if entering home 
- r(s, a) = 0 otherwise



1. max {51, 0}
2. max {51, -20}
3. max {51, -80}
4. max {51, -100}

$Q(s3, \leftarrow) = 0 + 0.8 * 64 = 51$

$Q(s3, \downarrow) = -100 + 0.8*100 = -20$

$V^*(s3) = \max \{Q(s3, \leftarrow), Q(s3, \downarrow)\}$

$= \max \{51, -20\}$