SUTD 2021 50.007 Homework 6

James Raphael Tiovalen 1004555

Markov Decision Process & Reinforcement Learning

Question 1

Assuming that we only consider synchronous updates (i.e., we update the current iteration's Q-values using the previous iteration's Q-values), and using the Q-learning algorithm's update equation as such for each (s, a):

$$Q_1^*(s,a) \leftarrow \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma \max_{a'} Q_0^*(s',a')],$$

we would obtain:

$$
\begin{aligned}
Q_1^*(0, J) &= T(0, J, 0) \times [R(0, J, 0) + 0.5 \times \max_{a'} Q_0^*(0, a')] \\
&= 1 \times (0 + 0) = 0 \\
Q_1^*(0, W) &= T(0, W, 0) \times [R(0, W, 0) + 0.5 \times \max_{a'} Q_0^*(0, a')] \\
&= 1 \times (0 + 0) = 0 \\
Q_1^*(1, J) &= T(1, J, 0) \times [R(1, J, 0) + 0.5 \times \max_{a'} Q_0^*(0, a')] \\
&+ T(1, J, 1) \times [R(1, J, 1) + 0.5 \times \max_{a'} Q_0^*(1, a')] \\
&= 0.5 \times (1 + 0) + 0.5 \times (0 + 0) = 0.5 \\
Q_1^*(1, W) &= T(1, W, 0) \times [R(1, W, 0) + 0.5 \times \max_{a'} Q_0^*(0, a')] \\
&= 1 \times (1 + 0) = 1 \\
Q_1^*(2, J) &= T(2, J, 0) \times [R(2, J, 0) + 0.5 \times \max_{a'} Q_0^*(0, a')] \\
&+ T(2, J, 2) \times [R(2, J, 2) + 0.5 \times \max_{a'} Q_0^*(2, a')] \\
&= 0.5 \times (4 + 0) + 0.5 \times (0 + 0) = 2 \\
Q_1^*(2, W) &= T(2, W, 1) \times [R(2, W, 1) + 0.5 \times \max_{a'} Q_0^*(1, a')] \\
&= 1 \times (1 + 0) = 1 \\
Q_1^*(3, J) &= T(3, J, 1) \times [R(3, J, 1) + 0.5 \times \max_{a'} Q_0^*(1, a')] \\
&+ T(3, J, 3) \times [R(3, J, 3) + 0.5 \times \max_{a'} Q_0^*(3, a')] \\
&= 0.5 \times (4 + 0) + 0.5 \times (0 + 0) = 2 \\
Q_1^*(3, W) &= T(3, W, 2) \times [R(3, W, 2) + 0.5 \times \max_{a'} Q_0^*(2, a')] \\
&= 1 \times (1 + 0) = 1 \\
Q_1^*(4, J) &= T(4, J, 2) \times [R(4, J, 2) + 0.5 \times \max_{a'} Q_0^*(2, a')] \\
&+ T(4, J, 4) \times [R(4, J, 4) + 0.5 \times \max_{a'} Q_0^*(4, a')] \\
&= 0.5 \times (4 + 0) + 0.5 \times (0 + 0) = 2 \\
Q_1^*(4, W) &= T(4, W, 3) \times [R(4, W, 3) + 0.5 \times \max_{a'} Q_0^*(3, a')] \\
&= 1 \times (1 + 0) = 1
\end{aligned}
$$

Hence, our Q-values would be:

	$s = 0$	$s = 1$	$s = 2$	$s = 3$	$s = 4$
J	0	0.5	2	2	2
W	0	1	1	1	1

Question 2

Since the action should be chosen based on $\arg\max_a Q_1^*(s, a)$ for each state s, we would get:

$s=1$	$s=2$	$s=3$	$s=4$
W	J	J	J

Question 3

Since the value for state s should be $\max_a Q_1^*(s,a)$, we would get:

$s=0$	$s=1$	$s=2$	$s=3$	$s=4$
0	1	2	2	2

Question 4

No. Conducting a second iteration of the Q-Value Iteration Algorithm, we would obtain these corresponding values of $Q_2^*(s,a)$ for each (s,a) tuple:

$$
\begin{aligned}
Q_2^*(0,J) &= T(0,J,0) \times [R(0,J,0) + 0.5 \times \max_{a'} Q_1^*(0,a')] \\
&= 1 \times (0+0) = 0 \\
Q_2^*(0,W) &= T(0,W,0) \times [R(0,W,0) + 0.5 \times \max_{a'} Q_1^*(0,a')] \\
&= 1 \times (0+0) = 0 \\
Q_2^*(1,J) &= T(1,J,0) \times [R(1,J,0) + 0.5 \times \max_{a'} Q_1^*(0,a')] \\
&+ T(1,J,1) \times [R(1,J,1) + 0.5 \times \max_{a'} Q_1^*(1,a')] \\
&= 0.5 \times (1+0) + 0.5 \times (0 + 0.5 \times 1) = 0.75 \\
Q_2^*(1,W) &= T(1,W,0) \times [R(1,W,0) + 0.5 \times \max_{a'} Q_1^*(0,a')] \\
&= 1 \times (1+0) = 1 \\
Q_2^*(2,J) &= T(2,J,0) \times [R(2,J,0) + 0.5 \times \max_{a'} Q_1^*(0,a')] \\
&+ T(2,J,2) \times [R(2,J,2) + 0.5 \times \max_{a'} Q_1^*(2,a')] \\
&= 0.5 \times (4+0) + 0.5 \times (0 + 0.5 \times 2) = 2.5 \\
Q_2^*(2,W) &= T(2,W,1) \times [R(2,W,1) + 0.5 \times \max_{a'} Q_1^*(1,a')] \\
&= 1 \times (1 + 0.5 \times 1) = 1.5 \\
Q_2^*(3,J) &= T(3,J,1) \times [R(3,J,1) + 0.5 \times \max_{a'} Q_1^*(1,a')] \\
&+ T(3,J,3) \times [R(3,J,3) + 0.5 \times \max_{a'} Q_1^*(3,a')] \\
&= 0.5 \times (4 + 0.5 \times 1) + 0.5 \times (0 + 0.5 \times 2) = 2.75 \\
Q_2^*(3,W) &= T(3,W,2) \times [R(3,W,2) + 0.5 \times \max_{a'} Q_1^*(2,a')] \\
&= 1 \times (1 + 0.5 \times 2) = 2 \\
Q_2^*(4,J) &= T(4,J,2) \times [R(4,J,2) + 0.5 \times \max_{a'} Q_1^*(2,a')] \\
&+ T(4,J,4) \times [R(4,J,4) + 0.5 \times \max_{a'} Q_1^*(4,a')] \\
&= 0.5 \times (4 + 0.5 \times 2) + 0.5 \times (0 + 0.5 \times 2) = 3 \\
Q_2^*(4,W) &= T(4,W,3) \times [R(4,W,3) + 0.5 \times \max_{a'} Q_1^*(3,a')] \\
&= 1 \times (1 + 0.5 \times 2) = 2
\end{aligned}
$$

As demonstrated, the derived optimal policy is as follows and does not change:

$s=1$	$s=2$	$s=3$	$s=4$
W	J	J	J