SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

50.007 Machine Learning, Fall 2020
Homework 5
Sample Solutions

Consider the following Markov decision process (MDP). It has states $\{0, 1, 2, 3, 4\}$ with 4 as the starting state. In every state, you can take one of two possible actions: walk (W) or jump (J). The Walk action decreases the state by one. The Jump action has probability 0.5 of decreasing the state by two, and probability 0.5 of leaving the state unchanged. Actions will not decrease the state below zero: you will remain in state 0 no matter which action you take (i.e., state 0 is a terminal state). Jumping in state 1 leads to state 0 with probability 0.5 and state 1 with probability 0.5. This definition leads to the following transition functions:

- For states $k \geq 1$, $T(k, W, k-1) = 1$
- For states $k \geq 2$, $T(k, J, k-2) = T(k, J, k) = 0.5$
- For state $k = 1$, $T(k, J, k-1) = T(k, J, k) = 0.5$
- For state $k = 0$, $T(k, J, k) = T(k, W, k) = 1$

The reward gained when taking an action is the distance travelled squared, i.e., $R(s, a, s') = (s - s')^2$. The discount factor is $\gamma = 0.5$.

1. Suppose we initialize $Q_0^*(s, a) = 0$ for all $s \in \{0, 1, 2, 3, 4\}$ and $a \in \{J, W\}$. Evaluate the Q-values $Q_1^*(s, a)$ after exactly one iteration of the Q-Value Iteration Algorithm. Write your answers in the table below.

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

Explanations: As mentioned in class, we consider synchronous updates in this course. This is also reflected by the algorithm described in the notes. Specifically, when updating the current iteration's Q-values (or values, when the value iteration algorithm is considered), we refer to the previous iteration's Q-values (or values).

The update equations are as follows, for each (s, a):

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

This leads to:

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

2. What is the policy that we would derive from $Q_1^*(s,a)$? Answer by filling in the action that should be taken at each state in the table below.

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

Explanations: The action should be chosen based on $\arg\max_a Q_1^*(s,a)$ for each state s.

3. What are the values $V_1^*(s)$ corresponding to $Q_1^*(s,a)$?

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

Explanations: The value for state s should be $\max_a Q_1^*(s,a)$.

4. Will the policy change after the second iteration? If your answer is"yes", briefly describe how.

Ans: No.

Explanations: we can follow the above procedure to calculate the $Q_2^*(s,a)$ for each (s,a) tuple. Next we can derive the policy after the second iteration.

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

The policy is as follows. As we can see, there is no change.

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