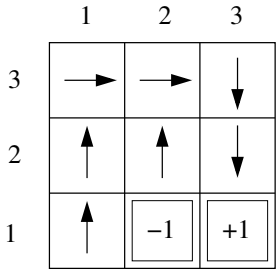
CS 6300 **TD Learning Practice** **March 2, 2021**

In the mini grid world shown below, there are two terminal states: state (2,1) with a negative reward of -1, and (3,1) with a positive reward of +1. The transition model is the same as the grid world in the course slides: an action succeeds with probability 0.8, and goes to the left or right of the intended direction with probability 0.1, respectively. However, moves into the wall are not allowed. The optimal policy π is in the left figure, and the correct utility function the optimal policy is in the right figure.



0.554	0.639	0.735
0.497	0.581	0.849
0.294	−1.0	1.0

Below is a series of three trials (1, 2, and 3 left to right) in this environment. Starting in state (1,3), actions were taken according to the fixed policy π above, and ended once a terminal state is reached. The trials are as follows:

S	A	R	S	A	R	S	A	R
(1,3)	E	0	(1,3)	E	0	(1,3)	E	0
(2,3)	E	0	(2,3)	E	0	(1,2)	N	0
(3,3)	S	0	(2,2)	N	0	(1,3)	E	0
(3,2)	S	1	(2,3)	E	0	(2,3)	E	0
(3,1)			(3,3)	S	0	(3,3)	S	0
			(3,2)	S	1	(3,2)	S	1
			(3,1)			(3,1)		

1. Estimate the transition function $T(s, a, s')$ as much as possible given these limited trials. Zeros have been put in for non-neighboring states.

	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
(1,2),N	0	0	1	0	0	0	0	0	0
(1,3),E	0	0.25	0	0	0	0.75	0	0	0
(2,2),N	0	0	0	0	0	1	0	0	0
(2,3),E	0	0	0	0	0.25	0	0	0	0.75
(3,2),S	0	0	0	0	0	0	1	0	0
(3,3),S	0	0	0	0	0	0	0	1	0

2. Perform policy evaluation with your estimated MDP to find the state values. Assume $\gamma = 0.9$. Set up the linear equations to solve exactly without iteration.

   The equation to use is:

$$V^\pi(s) = \sum_{s'} T(s,a,s')(R(s,a,s') + \gamma V^\pi(s'))$$

$$\begin{aligned}
V^\pi(3,1) &= 0 \\
V^\pi(3,2) &= 1 * (1 + 0.9 * V^\pi(3,1)) \\
&= 1 \\
V^\pi(3,3) &= 1 * (0 + 0.9 * V^\pi(3,2)) \\
&= 0.9 \\
V^\pi(2,2) &= 1 * (0 + 0.9 * V^\pi(2,3)) \\
V^\pi(2,3) &= 0.75 * (0 + 0.9 * V^\pi(3,3)) + 0.25 * (0 + 0.9 * V^\pi(2,2)) \\
&= 0.75 * (0.9 * 0.9) + 0.25 * (0 + 0.9 * 0.9 * V^\pi(2,3)) \\
&= 0.6075/(1 - 0.25 * 0.81) \\
&= 0.762 \\
V^\pi(2,2) &= 1 * (0 + 0.9 * 0.762) \\
&= 0.6858 \\
V^\pi(1,2) &= 1 * (0 + 0.9 * V^\pi(1,3)) \\
V^\pi(1,3) &= 0.75 * (0 + 0.9 * V^\pi(2,3)) + 0.25 * (0 + 0.9 * V^\pi(1,2)) \\
&= 0.75 * (0.9 * 0.762) + 0.25 * (0 + 0.9 * 0.9 * V(1,1)) \\
&= 0.5143/(1 - 0.25 * 0.81) \\
&= 0.645 \\
V^\pi(1,2) &= 1 * (0 + 0.9 * 0.645) \\
&= 0.581
\end{aligned}$$

3. Perform TD learning for the three trials left to right. Use $\alpha(n) = 1/n$, where n is the trial number. Only write the non-trivial updates.

   All values are initialized to 0. The update equation to use is:

$$V^\pi(s) \leftarrow (1 - \alpha)V^\pi(s) + \alpha(R(s,a,s') + \gamma V^\pi(s'))$$

   Trial 1:

$$\begin{aligned}
V^\pi(3,2) &\leftarrow (1-1) * V^\pi(3,2) + 1 * (1 + 0.9 * V^\pi(3,1)) \\
&\leftarrow 0 + (1 + 0.9 * 0) = 1
\end{aligned}$$

Trial 2:

$$\begin{aligned} V^{\pi}(3,3) &\leftarrow \frac{1}{2} * V^{\pi}(3,3) + \frac{1}{2} * (0 + 0.9 * V^{\pi}(3,2) \\ &\leftarrow 0 + \frac{1}{2}(0 + 0.9 * 1) = 0.45 \end{aligned}$$

Trial 3:

$$\begin{aligned} V^{\pi}(2,3) &\leftarrow \frac{2}{3} * V^{\pi}(2,3) + \frac{1}{3} * (0 + 0.9 * V^{\pi}(3,3)) \\ &\leftarrow 0 + \frac{1}{3} * (0 + 0.9 * 0.45) = 0.135 \\ V^{\pi}(3,3) &\leftarrow \frac{2}{3} * V^{\pi}(3,3) + \frac{1}{3} * (0 + 0.9 * V^{\pi}(3,2)) \\ &\leftarrow \frac{2}{3} * 0.45 + \frac{1}{3} * (0 + 0.9 * 1) = 0.6 \end{aligned}$$