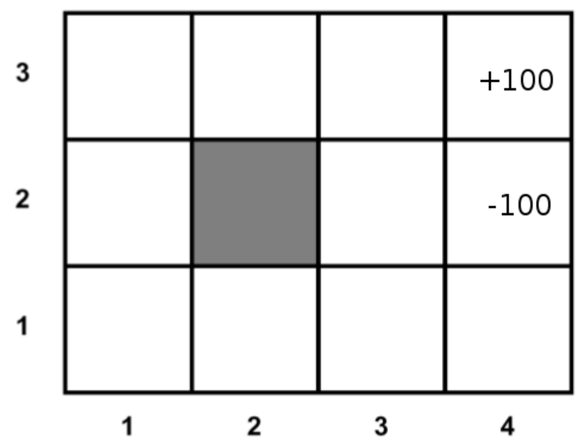
HW6: Reinforcement Learning (Solution)

CS 4300: Artificial Intelligence **Tucker Hermans**
University of Utah

1 TD and Q in Blockworld

Consider the following gridworld:



Suppose that we run two episodes that yield the following sequences of (state, action, reward) tuples:

S	A	R	S	A	R
(1,1)	up	-1	(1,1)	up	-1
(2,1)	left	-1	(1,2)	up	-1
(1,1)	up	-1	(1,3)	right	-1
(1,2)	up	-1	(2,3)	right	-1
(1,3)	up	-1	(2,3)	right	-1
(2,3)	right	-1	(3,3)	right	-1
(3,3)	right	-1	(4,3)	exit	+100
(4,3)	exit	+100	(done)		
(done)					

1. According to model-based learning, what are the transition probabilities for every (state, action, state) triple. Don't bother listing all the ones that we have no information about.

   - T((1,1),up ,(2,1)) =1/3
   - T((1,1),up ,(1,2)) =2/3
   - T((2,1),left ,(1,1)) =1
   - T((1,2),up ,(1,3)) =1
   - T((1,3),up ,(2,3)) =1
   - T((1,3),right,(2,3)) =1
   - T((2,3),right,(2,3)) =1/3

- T((2,3),right,(3,3)) =2/3
- T((3,3),right,(4,3)) =1

2. What would the Q-value estimate be if SARSA were run to generate these same trajectories? Assume all Q-value estimates start at 0, a discount factor of 0.9 and a learning rate of 0.5. Again, don't bother listing all of the cases where we don't have data.

Remember: $Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r(s,a,s') + \lambda Q(s',a'))$; $\alpha \leftarrow 0.5$; $\lambda \leftarrow 0.9$

Sequence of updates:

(a) Trial 1

i. $Q((1,1),up) = (1.0 - 0.5) \cdot 0.0 + 0.5 \cdot (-1 + 0.9 \cdot 0.0) = -0.5$

ii. $Q((2,1),left) = 0.5 \cdot 0.0 + 0.5 \cdot (-1 + 0.9 \cdot (-0.5)) = -0.725$

iii. $Q((1,1),up) = 0.5 \cdot (-0.5) + 0.5 \cdot (-1 + 0.9 \cdot 0.0) = -0.75$

iv. $Q((1,2),up) = 0.5 \cdot 0.0 + 0.5 \cdot (-1 + 0.9 \cdot 0.0) = -0.5$

v. $Q((1,3),up) = 0.5 \cdot 0.0 + 0.5 \cdot (-1 + 0.9 \cdot 0.0) = -0.5$

vi. $Q((2,3),right) = 0.5 \cdot 0.0 + 0.5 \cdot (-1 + 0.9 \cdot 0.0) = -0.5$

vii. $Q((3,3),right) = 0.5 \cdot 0.0 + 0.5 \cdot (-1 + 0.9 \cdot 0.0) = -0.5$

viii. $Q((4,3),exit) = 0.5 \cdot 0.0 + 0.5 \cdot (100 + 0.9 \cdot 0.0) = 50$

(b) Trial 2

i. $Q((1,1),up) = 0.5 \cdot (-0.75) + 0.5 \cdot (-1 + 0.9 \cdot -0.5) = -1.1$

ii. $Q((1,2),up) = 0.5 \cdot (-0.5) + 0.5 \cdot (-1 + 0.9 \cdot 0.0) = -0.75$

iii. $Q((1,3),right) = 0.5 \cdot 0.0 + 0.5 \cdot (-1 + 0.9 \cdot -0.5) = -0.725$

iv. $Q((2,3),right) = 0.5 \cdot -0.275 + 0.5 \cdot (-1 + 0.9 \cdot -0.275) = -1.2125$

v. $Q((2,3),right) = 0.5 \cdot -0.76125 + 0.5 \cdot (-1 + 0.9 \cdot -0.5) = -1.105625$

vi. $Q((3,3),right) = 0.5 \cdot -0.5 + 0.5 \cdot (-1 + 0.9 \cdot 50) = 21.75$

vii. $Q((4,3),exit) = 0.5 \cdot 50 + 0.5 \cdot (100 + 0.9 \cdot 0.0) = 75$

Final values:

- $Q((1,1),up) = 0.5 \cdot (-0.75) + 0.5 \cdot (-1 + 0.9 \cdot -0.5) = -1.1$
- $Q((1,2),up) = 0.5 \cdot (-0.5) + 0.5 \cdot (-1 + 0.9 \cdot 0.0) = -0.75$
- $Q((1,3),right) = 0.5 \cdot 0.0 + 0.5 \cdot (-1 + 0.9 \cdot -0.5) = -0.725$
- $Q((2,1),left) = 0.5 \cdot 0.0 + 0.5 \cdot (-1 + 0.9 \cdot (-0.5)) = -0.725$
- $Q((2,3),right) = 0.5 \cdot -0.76125 + 0.5 \cdot (-1 + 0.9 \cdot -0.5) = -1.2125$
- $Q((3,3),right) = 0.5 \cdot -0.5 + 0.5 \cdot (-1 + 0.9 \cdot 50) = 21.75$
- $Q((4,3),exit) = 0.5 \cdot 50 + 0.5 \cdot (100 + 0.9 \cdot 0.0) = 75$

3. Suppose that we run Q-learning. However, instead of initializing all our Q values to zero, we initialize them to some large positive number ("large" with respect to the maximum reward possible in the world: say, 10 times the max reward). I claim that this will cause a Q-learning agent to initially explore a lot and then eventually start exploiting. Why should this be true? Justify your answer in a short paragraph.

If we start all the Q values out higher than the max reward, then for most of them, as we learn and experience the world, the values will decrease. So if theres some state-action pair (s, a) that weve already explored, our Q value will have likely dropped from its initial value. This means that for some other, unexplored action a 0 , the Q value for (s, a 0) will remain large and therefore well choose to take a 0 instead of a. This leads to a large amount of exploration.