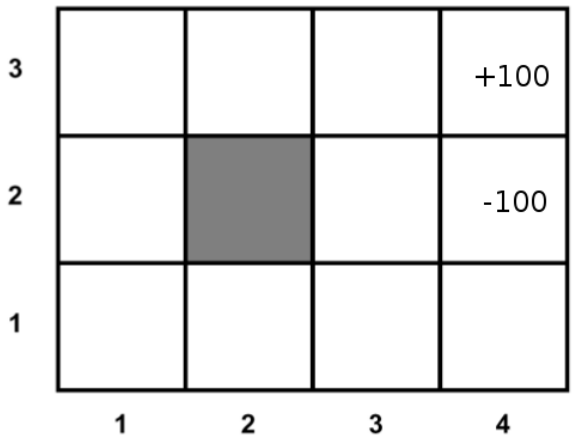
HW6: Reinforcement Learning

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.
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.
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.