William Frank
3/6/19
HW6: Reinforcement Learning

1. What are the transition probabilities for every (state, action, state) triple?

   T((1, 1), up, (2, 1)) = 0.33
   T((1, 1), up, (1, 2)) = 0.67
   T((2, 1), left, (1, 1)) = 1.0
   T((1, 2), up, (1, 3)) = 1.0
   T((1, 3), up, (2, 3)) = 1.0
   T((1, 3), right, (2, 3)) = 1.0
   T((2, 3), right, (2, 3)) = 0.33
   T((2, 3), right, (3, 3)) = 0.67
   T((3, 3), right, (4, 3)) = 1.0
   T((4, 3), exit, (done)) = 1.0

2. What would the Q-value estimate be if SARSA were run to generate these same trajectories?

   Q((1, 1), up) = 0 → -0.5
   Q((2, 1), left) = 0 → -0.5
   Q((1, 1), up) = -0.5 → -0.75
   Q((1, 2), up) = 0 → -0.5
   Q((1, 3), up) = 0 → -0.5
   Q((2, 3), right) = 0 → -0.5
   Q((3, 3), right) = 0 → -0.5
   Q((4, 3), exit) = 0 → 50

   Q((1, 1), up) = -0.75 → -0.875
   Q((1, 2), up) = -0.5 → -0.75
   Q((1, 3), right) = 0 → -0.5
   Q((2, 3), right) = -0.5 → -0.75
   Q((2, 3), right) = -0.75 → -0.875
   Q((3, 3), right) = -0.5 → 21.75
   Q((4, 3), exit) = 50 → 75

3. Why would initializing Q-values to something relatively large cause the agent to initially explore a lot then eventually start exploiting?

   By initializing Q-values to something large the agent will be incentivized to visit all the tiles. As it starts exploring the negative rewards will decrease these Q-values, and the agent will explore elsewhere. In this way the agent explore these disproportionately-high-valued tiles a lot, until their Q-value more accurately reflects their long term reward. At this point the agent can start exploiting the more accurate model it has learned.