Reinforcement Learning—Quiz 7 Solution

March 30, 2020

Name: _____ Score:_____

1. N-step TD methods span a spectrum from Monte Carlo method at one end to one-step TD at the other.

   A. True

   B. False

2. Assume you are using an n-step TD method with $n = 4$. If you've just taken action A_t from state S_t at timestep t, receiving reward R_{t+1}, what is the most recent Q value that can now be updated?

   A. $Q(S_{t-5}, A_{t-5})$

   B. $Q(S_{t-4}, A_{t-4})$

   C. $Q(S_{t-3}, A_{t-3})$

   D. $Q(S_{t-2}, A_{t-2})$

   E. $Q(S_{t-1}, A_{t-1})$

   F. $Q(S_t, A_t)$

   G. $Q(S_{t+1}, A_{t+1})$

   Solution: Assume we're using Q-learning. Then, in a 1-step case, if we've just received reward R_{t+1}, then we can update $Q(S_t, A_t)$. The 4-step case, then, should allow us to update three-steps back from S_t, or S_{t-3}. Thus, we can update $Q(S_{t-3}, A_{t-3})$.

3. (a) Assume the maze MDP has a discount factor $\gamma = 0.9$ and gives 0 reward for everything but leaving the maze, and gives +1 reward for leaving the maze.

   If you have an episode of length 100 that starts at the maze entrance and leaves at the exit, what is the maximum number of non-trivial (not updated to have the same value) state value updates?

   A. 1

   B. 10

   C. 100

   Solution: We update each state based on the next 10 transitions. For the first 90 states in the episode, we'll be doing trivial updates. The last 10 states in our episode will include the reward for leaving the maze, so will be non-trivial updates.

   (b) Assume the maze MDP has a discount factor $\gamma = 1.0$ and gives -1 reward for everything but leaving the maze, and gives 0 reward for leaving the maze.

   If you have an episode of length 100 that starts at the maze entrance and leaves at the exit, what is the maximum number of non-trivial (not updated to have the same value) state value updates?

A. 1

B. 10

C. 100

Solution: Each state in the episode will update based on the next 10 rewards which will be non-zero. Thus, each state value will get an updated value.