Assignment 6

Reinforcement Learning

Prof. B. Ravindran

1. In the search procedure listed in the lecture for Monte-Carlo tree search what is/are the uses of the depth parameter?

   (a) allows us to identify leaf states

   (b) allows us to identify terminal states

   (c) can be used to impact the choice of action selection

   (d) allows us to specialise value functions based on the number of steps that have been taken

2. Suppose you are given a finite set of transition data. Assuming that the Markov model that can be formed with the given data is the actual MDP from which the data is generated, will the value functions calculated by the MC and TD methods (in a manner similar to what we saw in the lectures) necessarily agree?

   (a) no

   (b) yes

3. In the iterative policy evaluation process, we have seen the use of different update equations in DP, MC, and TD methods. With regard to these update equations

   (a) DP and TD make use of estimates but not MC

   (b) TD makes use of estimates but not DP and MC

   (c) MC and TD make use of estimates but not DP

   (d) all three methods make use of estimates

4. Is it necessary for the behaviour policy of an off-policy learning method to have non-zero probability of selecting all actions?

   (a) no

   (b) yes

5. With respect to the Expected SARSA algorithm, is exploration (using for example ϵ-greedy action selection) required as it is in the normal SARSA and Q-learning algorithms?

   (a) no

   (b) yes

6. Assume that we have available a simulation model for a particular problem. To learn an optimal policy, instead of following trajectories end-to-end, in each iteration we randomly supply a state and an action to the model and receive the corresponding reward. This information is used for updating the value function. Which method among the following would you expect to work in this scenario?

   (a) SARSA
   (b) Expected SARSA
   (c) Q-learning
   (d) none of the above

7. Consider the following transitions observed for an undiscounted MDP with two states P and Q.

   P, +3, P, +2, Q, -4, P, +4, Q, -3

   Q, -2, P, +3, Q, -3

   Estimates the state value function using first-visit Monte-Carlo evaluation.

   (a) v(P) = 2, v(Q) = -5/2
   (b) v(P) = 2, v(Q) = 0
   (c) v(P) = 1, v(Q) = -5/2
   (d) v(P) = 1, v(Q) = 0

8. Considering the same transition data as above, estimate the state value function using the every-visit Monte-Carlo evaluation.

   (a) v(P) = 2, v(Q) = -5/2
   (b) v(P) = 2, v(Q) = -11/4
   (c) v(P) = 1/2, v(Q) = -11/4
   (d) v(P) = 1/4, v(Q) = -5/2

9. Construct a Markov model that best explains the observations given in question 7. In this model, what is the probability of transitioning from state P to itself? What is the expected reward received on transitioning from state Q to state P?

   (a) 1/4, -4
   (b) 1/4, -3
   (c) 1/2, -4
   (d) 1/2, -3

10. What would be the value function estimate if batch TD(0) were applied to the above transaction data?

    (a) 1, -1/2
    (b) 1, -2
    (c) 2, -1/2
    (d) 2, -2