Assignment 6 (Sol.)

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

   Sol. (a), (c), (d)

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

   Sol. (a)
   In our lecture example, we saw that the value functions calculated by the MC and TD methods did not agree. The same would still hold if the MDP that generated the data was the MDP that we formed from the data in applying the TD method.

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

   Sol. (d)
   DP methods update estimates based on other learned estimates, i.e., they bootstrap. MC methods, while they do not bootstrap, make use of estimates because the target in MC methods, i.e., the sample return, is an estimate of the actual expected return; the expected value of course, is not known. TD methods use both the above, i.e., they make use of samples of the expected values as well as bootstrap.

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

   Sol. (a)
   If the probability of selecting certain actions in the estimation policy, i.e., the policy which is being evaluated and/or improved, is zero, then the corresponding probability of selecting those same actions in the behaviour policy can also be zero without causing any problem to the off-policy learning procedure.

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

   Sol. (b)
   The difference in the update rules that differentiate Expected SARSA from the SARSA algorithm do not obviate the need for exploration in the former, since without exploration the algorithm would, in general, miss out on large parts of the state space, preventing it from correctly converging (in the limit) to an optimal policy.

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

   Sol. (d)
   Note that each of the three algorithms listed above require, in addition to the reward, the next state information, which is not provided by the described simulation model.

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

Sol. (c)
For first-visit MC, we consider only the first occurrence of each state in each transition. Thus, we have

v(P) = (2 + 0)/2 = 1

v(Q) = (-3 - 2)/2 = -5/2

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

Sol. (c)
In the every-visit case we consider each occurrence of each state in the transitions. Thus, we have

v(P) = (2 + -1 + 1 + 0)/4 = 1/2

v(Q) = (-3 - 3 - 2 - 3)/4 = -11/4

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?

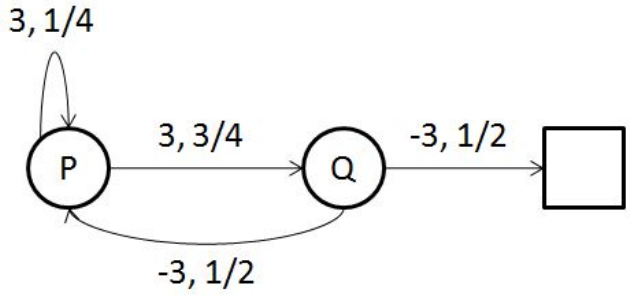
(a) 1/4, -4

(b) 1/4, -3

(c) 1/2, -4

(d) 1/2, -3

Sol. (b)
Following is the model that can be constructed from the available data.



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

    Sol. (d)
    We can solve the Bellman equations directly based on the above model to get

    v(A) = 3 + 1/4 * v(A) + 3/4 * v(B)

    v(B) = -3 + 1/2 * v(A)

    v(A) = 2

    v(B) = -2