EE675A: Solutions for Quiz 2

03/04/2023

Lecturer: Dr. S.S. Peruru *Scribe: Chethan R (22104131)*

Question 1

Let $v_1, v_2, v_3, \ldots$ be the value iteration algorithm for an MDP. Let $\pi_1, \pi_2, \pi_3, \ldots$ be the policies obtained by behaving greedily w.r.t those value iterates. Then is it true that $v_{\pi_1} \leq v_{\pi_2} \leq v_{\pi_3}, \ldots$?. Explain.

Solution:

No, it is not true. Since in the value iteration, we do not evaluate the policy fully before improving it, the policy improvement theorem does not hold here.

Question 2

Consider the following transitions observed from two episodes 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

1. Estimate the state value function using first-visit Monte-Carlo evaluation.
2. Considering the same transition data as above, estimate the state value function using the every-visit Monte-Carlo evaluation.
3. Construct a Markov model that best explains the given data and draw the transition diagram. In this model, what is the probability of transitioning from P to itself? What is the expected reward for transitioning from state Q to state P?
4. Considering the same transition data, what would be the value function estimate if the batch TD algorithm was applied?

Solution:

1. For the first-visit MC method, in the first transition from P and in the second transition from 2 the value estimate will be:

$$V(P) = ((3 + 2 - 4 + 4 - 3) + (+3 - 3))/2$$

$$V(P) = (2/2) = 1$$

For state Q,

$$V(Q) = ((-4 + 4 - 3) + (-2 + 3 - 3))/2$$

$$V(Q) = (-3 - 2)/2 = -2.5$$

2. For every visit MC, for state P, there are there sums from first transition, and one sum from the second transition.

$$V(P) = ((3 + 2 - 4 + 4 - 3) + (2 - 4 + 4 - 3) + (4 - 3) + (+3 - 3))/4$$

$$V(P) = (2 - 1 + 1 + 0)/4 = 0.5$$

For state Q, two sums in first transition and two from the second,

$$V(Q) = ((-4 + 4 - 3) + (-3) + (-2 + 3 - 3) + (-3))/4$$

$$V(Q) = (-3 - 3 - 2 - 3)/4 = -11/4$$

3. See Figure 1.

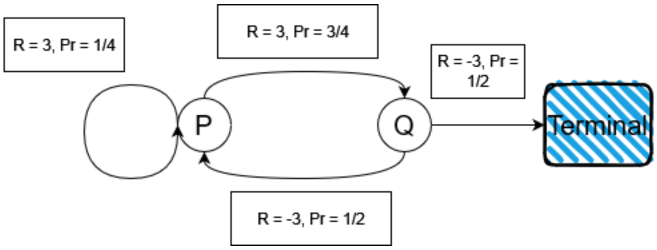

Figure 1: Best MDP estimate

4. Value function using TD(0)

$$V(P) = 3 + (1/4) * V(P) + (3/4) * V(Q)$$

$$V(Q) = -3 + (1/2) * V(P)$$

Solving the linear equations,

$$V(P) = 2$$

$$V(Q) = -2$$

Question 3

Which among DP, MC, and TD methods use bootstrapping?

Solution:

DP and TD methods use bootstrapping.

Question 4

Write the update equation of n-step TD for $n = 2$. Consider TD(λ) for $\lambda \in (0, 1)$, is it true that a one-step return always gets assigned the maximum weight? Explain.

Solution:

The update equation for $n = 2$ is,

$$V(s_t) = V(s_t) + \alpha[R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2}) - V(s_t)]$$

No, a one-step return doesn't necessarily get assigned the maximum weight because it will depend on the λ value and the length of the episode.

Question 5

Consider the Markov Decision Process (MDP) with discount factor $\gamma = 0.5$ as shown. Uppercase letters A, B and C represents states, arcs represent state actions, lowercase letters ab, ba, bc, ca, cb represents actions; signed integers represent rewards, and fractions represent transition probabilities.

1. Consider uniform random policy $\pi_1(a|s)$ that takes all actions from any state s with equal probability. Starting with initial value of $V_1(s) = 2, \forall s \in \{A, B, C\}$, apply one synchronous iteration of iterative policy evaluation (i.e, one backup for each state) to compute a new value function $V_2(s)$.

2. Apply one iteration of policy improvement to compute a new, deterministic policy $\pi_2(s)$.

3. Is your new value function $V_2(s)$ optimal? Justify your answer.

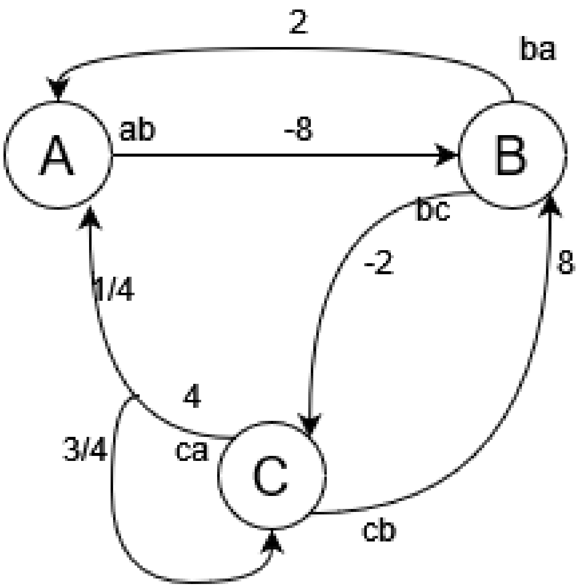

Figure 2: MDP

Solution:

1. The update rule for the policy evaluation step is:

$$V_{k+1}(s) = \mathbf{E}_{\pi}[R_{t+1} + \gamma V_k(s_{t+1}|s_t = s)]$$

$$V_{k+1}(s) = \sum_{a} \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma V_k(s')]$$

Using this equation: state A has only one action, hence,

$$V_2(A) = 1(1[-8 + 0.5 * 2])$$

$$V_2(A) = -7$$

For state B,

$$V_2(B) = 0.5 * (1[2 + 0.5 * 2] + 0.5 * (1[-2 + 0.5 * 2])$$

$$V_2(B) = 1.5 - 0.5 = 1$$

For State C,

$$V_2(C) = 0.5[1(8 + 0.5 * 2)] + 0.5(\frac{1}{4}[4 + 0.5 * 2] + \frac{3}{4}[4 + 0.5 * 2])$$

$$V_2(C) = 4.5 + 2.5 = 7$$

2. Policy improvement step using the updated values functions, the new greedy policy is given by,

$$\pi'(s) = \underset{a}{\operatorname{argmax}}\, \mathbf{E}[R_{t+1} + \gamma V_k(s_{t+1}|s_t = s, A_t = a)]$$

$$\pi'(s) = \underset{a}{\operatorname{argmax}} \sum_{s',r} p(s', r|s, a)[r + \gamma V_\pi(s')]$$

So, for state A, only one action is available, so the action will be ab.

$$\pi'(A) = ab$$

For state B,

$$\pi'(B) = \underset{a}{\operatorname{argmax}}\{(1[2 + 0.5 * (-7)]), (1[-2 + 0.5 * 7])\}$$

$$\pi'(B) = \underset{a}{\operatorname{argmax}}\{(-1.5), (1.5)\}$$

$$\pi'(B) = bc$$

For state C,

$$\pi'(C) = \underset{a}{\operatorname{argmax}}\{(\frac{1}{4}[4 + 0.5(-7)] + \frac{3}{4}[4 + 0.5 * 7]), (1[8 + 0.5 * 1])\}$$

$$\pi'(C) = \underset{a}{\operatorname{argmax}}\{(\frac{1.5}{4} + \frac{22.5}{4}), (8.5)\}$$

$$\pi'(C) = \underset{a}{\operatorname{argmax}}\{(5.75), (8.5)\}$$

$$\pi'(C) = cb$$

This is the updated policy.

If V_2 satisfies the bellman optimality equation, then its an optimal value.

$$V_{\pi^*}(s) = max_a \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')]$$

$$V_2^*(B) = max_a\{((1)[2 + 0.5 * (-7)), ((1)[-2 + 0.5 * (7))]\} = 1.5 \neq 1$$

Hence, $V_2(s)$ is not optimal.

Question 6

Consider the Markov Decision Process (MDP) with two states S_1 and S_2 and two actions a and b. Assume that $\gamma = 0.8$ and $\alpha = 0.2$. Suppose the current values of Q are shown in the table below. Suppose that we were in state S_1, we took action b, received reward 1.0, and moved to state S_2, and then took action b again.

$Q(S_1, a)$	2.0
$Q(S_1, b)$	2.0
$Q(S_2, a)$	4.0
$Q(S_2, b)$	2.0

1. We are using Q-learning to learn a policy. Which of the item of the Q table will change after one update of Q-learning and what is the new value ?
2. Repeat your answer if we had used expected SARSA to learn the policy with $\epsilon = 0.1$.
3. Repeat your answer if we had used SARSA to learn the policy with $\epsilon = 0.1$.

Solution:

1. Trajectory: $S_1, b, 1, S_2, b$.

$$Q_{new} = Q_{old} + \alpha[R_{t+1} + \gamma max_{a'} Q_{old}(S_{t+1}, a') - Q_{old}(S_t, A_t)]$$

   $Q(S_1, b)$ will get updated.

$$Q_{new} = 2 + 0.2[1 + 0.8 * max_a'\{Q(S_2, a), Q(S_2, b)\} - Q(S_1, b))]$$

$$Q_{new} = 2 + 0.2[1 + 0.8 * max\{4, 2\} - 2 = 2.44$$

2. Expected SARSA:
   Greedy policy w.r.t current Q-values is

$$\pi(s) = \underset{a}{\operatorname{argmax}}' Q(s, a')$$

$$\pi(s_1) = a$$

   or

$$\pi(s_1) = b$$

$$\pi(s_2) = a$$

   ϵ greedy with ϵ = 0.1 At S_1: $\pi(a|s_1) = 0.05$ and $\pi(b|s_1) = 0.95$ or $\pi(b|s_1) = 0.05$ and $\pi(a|s_1) = 0.95$.
   At S_2: $\pi(a|s_2) = 0.95$ and $\pi(b|s_2) = 0.05$.

$$Q_{new} = Q_{old} + \alpha[R_{t+1}\gamma \mathbf{E}_{a'\ \epsilon greedy}[Q_{old}(S_{t+1}, a') - Q_{old}(S_t, A_t)]$$

$$Q_{new}(S_1, b) = Q_{old}(S_1, b) + 0.2\{1 + 0.8 * [\pi(a|S_2) * Q_{old}(S_2, a) + \pi(b|S_2) * Q_{old}(S_2, b)] - Q_{old}(S_1, b)\}$$

$$Q_{new}(S_1, b) = 2 + 0.2\{1 + 0.8 * [(0.95 * 4) + (0.05 * 2)] - 2)\} = 2.424$$

3. SARSA

$$Q_{new} = Q_{old} + \alpha[R_{t+1} + \gamma Q_{old}(S_{t+1}, A_{t+1}) - Q_{old}(S_t, A_t)]$$

$$Q_{new} = 2 + 0.2[1 + 0.8 * 2 - 2] = 2.12$$