Assignment 5 (Sol.)

Reinforcement Learning

Prof. B. Ravindran

1. For a particular finite MDP with bounded rewards, let V be the space of bounded functions on S, the state space of the MDP. Let Π be the set of all policies, and let v_π be the value function corresponding to policy π where $\pi \in \Pi$. Is it true that $v_\pi \in V$, $\forall \pi \in \Pi$?

   (a) no

   (b) yes

   Sol. (b)
   The question essentially asks whether V contains all value functions, which we have seen in the lectures to be true.

2. In the proof of the value iteration theorem, we saw that $Lv^{n+1} = L_\pi v^{n+1}$. Is it true, in general, that for an arbitrary bounded function v, $Lv = L_\pi v$ (disregarding any special conditions that may be existing in the aforementioned proof)?

   (a) no

   (b) yes

   Sol. (a)

3. Continuing with the previous question, why is it the case that $Lv^{n+1} = L_\pi v^{n+1}$ in the proof of the value iteration theorem?

   (a) because the equality holds in general

   (b) because v^{n+1} is the optimal value function

   (c) because we are considering only deterministic policies choosing a max valued action in each state

   (d) because v^{n+1} is not a value function

   Sol. (c)

4. Given that $q_\pi(s, a) > v_\pi(s)$, we can conclude

   (a) action a is the best action that can be taken in state s

   (b) π may be an optimal policy

   (c) π is not an optimal policy

(d) none of the above

Sol. (c)
The inequality indicates that there exists an action that if taken in state s, the expected return would be higher than the expected return of taking actions in state s as per policy π. While this indicates that π is not an optimal policy, it does not indicate that a is the best action that can be taken in state s, since there may exist another action a' such that $q_\pi(s, a') > q_\pi(s, a)$.

5. Recall the problem described in the first question of the previous assignment. Use the MDP formulation arrived at in that question and starting with policy $\pi(\text{laughing}) = \pi(\text{silent}) =$ (incense, no organ), perform a couple of policy iterations or value iterations (by hand!) until you find an optimal policy (if you are taking a lot of iterations, stop and reconsider your formulation!). What are the resulting optimal state-action values for all state-action pairs?

   (a) $q_*(s, a) = 8$, $\forall a$
   (b) $q_*(s, a) = 10$, $\forall a$
   (c) $q_*(s, a^*) = 10$, $q_*(s, a) = -10$, $\forall a \neq a^*$
   (d) $q_*(s, a^*) = 10$, $q_*(s, a) = 8$, $\forall a \neq a^*$

Sol. (d)
First consider policy iteration.

Initialisation: $V(L) = V(Q) = 0$; $\pi(L) = \pi(Q) = \neg O \wedge I$

Assuming $\theta = 0.7$ and given, $\gamma = 0.9$

Evaluation:
$\Delta = 0$
$V(L) = P_{LL}^{\pi(L)} * [R_{LL}^{\pi(L)} + \gamma V(L)] + P_{LQ}^{\pi(L)} * [R_{LQ}^{\pi(L)} + \gamma V(Q)] = 1 * (-1 + 0.9 * 0) + 0 = -1$
$V(Q) = P_{QL}^{\pi(Q)} * [R_{QL}^{\pi(Q)} + \gamma V(L)] + P_{QQ}^{\pi(Q)} * [R_{QQ}^{\pi(Q)} + \gamma V(Q)] = 0 + 1 * (1 + 0.9 * 0) = +1$
$\Delta = 1$
$V(L) = P_{LL}^{\pi(L)} * [R_{LL}^{\pi(L)} + \gamma V(L)] + P_{LQ}^{\pi(L)} * [R_{LQ}^{\pi(L)} + \gamma V(Q)] = -1.9$
$V(Q) = P_{QL}^{\pi(Q)} * [R_{QL}^{\pi(Q)} + \gamma V(L)] + P_{QQ}^{\pi(Q)} * [R_{QQ}^{\pi(Q)} + \gamma V(Q)] = +1.9$
$\Delta = 0.9$
$V(L) = P_{LL}^{\pi(L)} * [R_{LL}^{\pi(L)} + \gamma V(L)] + P_{LQ}^{\pi(L)} * [R_{LQ}^{\pi(L)} + \gamma V(Q)] = -2.71$
$V(Q) = P_{QL}^{\pi(Q)} * [R_{QL}^{\pi(Q)} + \gamma V(L)] + P_{QQ}^{\pi(Q)} * [R_{QQ}^{\pi(Q)} + \gamma V(Q)] = +2.71$
$\Delta = 0.81$
$V(L) = P_{LL}^{\pi(L)} * [R_{LL}^{\pi(L)} + \gamma V(L)] + P_{LQ}^{\pi(L)} * [R_{LQ}^{\pi(L)} + \gamma V(Q)] = -3.439$
$V(Q) = P_{QL}^{\pi(Q)} * [R_{QL}^{\pi(Q)} + \gamma V(L)] + P_{QQ}^{\pi(Q)} * [R_{QQ}^{\pi(Q)} + \gamma V(Q)] = +3.439$
$\Delta = 0.729$
$V(L) = P_{LL}^{\pi(L)} * [R_{LL}^{\pi(L)} + \gamma V(L)] + P_{LQ}^{\pi(L)} * [R_{LQ}^{\pi(L)} + \gamma V(Q)] = -4.0951$
$V(Q) = P_{QL}^{\pi(Q)} * [R_{QL}^{\pi(Q)} + \gamma V(L)] + P_{QQ}^{\pi(Q)} * [R_{QQ}^{\pi(Q)} + \gamma V(Q)] = +4.0951$
$\Delta = 0.6561$
Improvement:
$\pi(L) = O \wedge I$ (or $O \wedge \neg I$)
$\pi(Q) = \neg O \wedge I$
Evaluation:
$\Delta = 0$
$V(L) = P_{LL}^{\pi(L)} * [R_{LL}^{\pi(L)} + \gamma V(L)] + P_{LQ}^{\pi(L)} * [R_{LQ}^{\pi(L)} + \gamma V(Q)] = +4.6856$

$V(Q) = P_{QL}^{\pi(Q)} * [R_{QL}^{\pi(Q)} + \gamma V(L)] + P_{QQ}^{\pi(Q)} * [R_{QQ}^{\pi(Q)} + \gamma V(Q)] = +4.6856$
$\Delta = 0.5905$
Improvement:
$\pi(L) = O \wedge I$ (or $O \wedge \neg I$)
$\pi(Q) = \neg O \wedge I$
No change in policy. This is an optimal policy.

Now consider value iteration.
Initialisation: $V(L) = V(Q) = 0$; Assuming $\theta = 0.9$
$\Delta = 0$
$V(L) = max_a\{P_{LL}^a * [R_{LL}^a + \gamma V(L)] + P_{LQ}^a * [R_{LQ}^a + \gamma V(Q)]\}$
For $a = O \wedge I$ or $O \wedge \neg I$:
$V(L) = +1$
$V(Q) = max_a\{P_{QL}^a * [R_{QL}^a + \gamma V(L)] + P_{QQ}^a * [R_{QQ}^a + \gamma V(Q)]\}$
For $a = \neg O \wedge I$: $V(Q) = +1$
$\Delta = 1$
$V(L) = +1.9$ for $a = O \wedge I$ or $O \wedge \neg I$
$V(Q) = +1.9$ for $a = \neg O \wedge I$
$\Delta = 0.9$
$V(L) = +2.71$ for $a = O \wedge I$ or $O \wedge \neg I$
$V(Q) = +2.71$ for $a = \neg O \wedge I$
$\Delta = 0.81$

Deterministic policy:
$\pi(L) = O \wedge I$ (or $O \wedge \neg I$)
$\pi(Q) = \neg O \wedge I$

Continuing evaluation will lead to convergence values of ± 10 ($\pm 1(1+\gamma+\gamma^2+...) = \frac{\pm 1}{1-\gamma} = \frac{\pm 1}{.1}$).
Optimal state-action values:

Current state	Action	Next state	$q_\star(s,a)$
L	$O \wedge I$	Q	$+10$
L	$O \wedge \neg I$	Q	$+10$
L	$\neg O \wedge I$	L	$+8$
L	$\neg O \wedge \neg I$	L	$+8$
Q	$O \wedge I$	L	$+8$
Q	$O \wedge \neg I$	L	$+8$
Q	$\neg O \wedge I$	Q	$+10$
Q	$\neg O \wedge \neg I$	L	$+8$

Note: $q_\star(s,a) = +10$ when an optimal action is taken in any state, whereas $+8 (= -1 + 0.9 * 10)$ results when an initial sub-optimal action is followed by actions taken according to an optimal policy.

6. In the previous question, what does the state value function converge to for the policy we started off with?

   (a) $v_\pi(laughing) = v_\pi(silent) = 10$
   (b) $v_\pi(laughing) = 8$, $v_\pi(silent) = 10$
   (c) $v_\pi(laughing) = -10$, $v_\pi(silent) = 10$
   (d) $v_\pi(laughing) = -8$, $v_\pi(silent) = 10$

Sol. (c)
Refer to the solution of the previous question.

7. In solving an episodic problem we observe that all trajectories from the start state to the goal state pass through a particular state exactly twice. In such a scenario, is it preferable to use first-visit or every-visit MC for evaluating the policy?

   (a) first-visit MC
   (b) every-visit MC
   (c) every-visit MC with exploring starts
   (d) neither, as there are issues with the problem itself

   Sol. (d)
   A state having to be visited exactly twice in any trajectory from the start state to the goal state indicates that the problem environment does not follow the Markov property.

8. Which of the following are advantages of Monte Carlo methods over dynamic programming techniques?

   (a) the ability to learn from actual experience
   (b) the ability to learn from simulated experience
   (c) the ability to estimate the value of a single state independent of the number of states
   (d) the ability to show guaranteed convergence to an optimal policy

   Sol. (a), (b), (c)
   Option (c) is an advantage, since it allows us to estimate values of only specific states of an MDP if those are the only states we care to know about.

9. For a specific MDP, suppose we have a policy that we want to evaluate through the use of actual experience in the environment alone and using Monte Carlo methods. We decide to use the first-visit approach along with the technique of always picking the start state at random from the available set of states. Will this approach ensure complete evaluation of the action value function corresponding to the policy?

   (a) no
   (b) yes

   Sol. (a)
   Depending upon the policy, starting from random states alone will not ensure observing returns for each state-action pairs, which is required to fully evaluate q_π.

10. Assuming an MDP where there are n actions $a \in A$ each of which is applicable in each state $s \in S$, if π is an ϵ-soft policy for some $\epsilon > 0$, then

    (a) $\pi(a|s) = \epsilon$, $\forall a, s$
    (b) $\pi(a|s) = \frac{\epsilon}{n}$, $\forall a, s$
    (c) $\pi(a|s) >= \frac{\epsilon}{n}$, $\forall a, s$
    (d) $\pi(a'|s) = 1 - \epsilon + \frac{\epsilon}{n}$, $\pi(a|s) = \frac{\epsilon}{n}$, $\forall a \neq a'$, $\forall s$

    Sol. (c)
    An ϵ-soft policy is one where the probability of selecting any action is at least ϵ/n.