Assignment 4 (Sol.)

Reinforcement Learning

Prof. B. Ravindran

1. You receive the following letter:

   Dear Friend,

   Some time ago, I bought this old house, but found it to be haunted by ghostly sardonic laughter. As a result it is hardly habitable. There is hope, however, for by actual testing I have found that this haunting is subject to certain laws, obscure but infallible, and that the laughter can be affected by my playing the organ or burning incense. In each minute, the laughter occurs or not, it shows no degree. What it will do during the ensuing minute depends, in the following exact way, on what has been happening during the preceding minute:

   Whenever there is laughter, it will continue in the succeeding minute unless I play the organ, in which case it will stop. But continuing to play the organ does not keep the house quiet. I notice, however, that whenever I burn incense when the house is quiet and do not play the organ it remains quiet for the next minute.

   At this minute of writing, the laughter is going on. Please tell me what manipulations of incense and organ I should make to get that house quiet, and to keep it so.

   Sincerely, At Wits End

   Assume that we make the following decisions in formulating this problem as an MDP:

   State set: $\{L, Q\}$, where L indicates that there is laughter in the room, and Q indicates that the room is quiet.

   Action set: $\{O \wedge I, O \wedge \neg I, \neg O \wedge I, \neg O \wedge \neg I\}$, where O corresponds to playing the organ, and I corresponds to burning incense.

   We consider this as a continuing discounted problem with $\gamma = 0.9$ and we let the reward be $+1$ on any transition into the silent state, and -1 on any transition into the laughing state.

   Assuming deterministic state transitions and rewards based upon current state and action, which among the following 4-tuples (current state, action, next state, reward) represent correct state transitions and rewards?

   (a) $(L, O \wedge I, Q, +1)$
   (b) $(L, O \wedge \neg I, L, -1)$
   (c) $(L, \neg O \wedge I, Q, +1)$
   (d) $(L, \neg O \wedge \neg I, L, -1)$
   (e) $(Q, O \wedge I, Q, +1)$

(f) $(Q, O \wedge \neg I, L, -1)$

(g) $(Q, \neg O \wedge I, Q, +1)$

(h) $(Q, \neg O \wedge \neg I, L, -1)$

Sol. (a), (d), (f), (g), (h)
We know that if there is laughter and the organ is played, then in the next step laughter will stop. This contradicts option (b). Similarly, option (c) indicates that by burning incense alone laughter can be made to stop, which is incorrect. Option (e) is also not correct because we know that playing the organ when the house is quiet does not result in the house staying quiet.

2. Based on the above problem description, what advice will you give to At Wit's End?

   (a) if there is laughter, play the organ and do not burn incense; if room is quite, play the organ and burn incense

   (b) never play the organ, always burn incense

   (c) always play the organ, never burn incense

   (d) if there is laughter, play the organ; if room is quite, do not play the organ and burn incense

   Sol. (d)

3. If a policy is greedy with respect to its own value function, then it is an optimal policy.

   (a) false

   (b) true

   Sol. (b)
   Consider the value function corresponding to an arbitrary policy π. If we derive a policy that is greedy with respect to this value function, by the policy improvement theorem, we are guaranteed to get a policy which is at least as good as the policy π. This derived policy will be equivalent to the policy π if and only if π is optimal. Hence, if a policy is greedy with respect to its own value function, then it is optimal.

4. Consider a 4 X 4 grid world problem where the goal is to reach either the top left corner or the bottom right corner. The agent can choose from four actions {up, down, left, right} which deterministically cause the corresponding state transitions, except that actions that would take the agent off the grid leave the state unchanged. We model this as an undiscounted, episodic task, where the reward is -1 for all transitions. Suppose that the agent follows the equiprobable random policy. Given below is the partial value function for this problem. Calculate respectively, the missing values in the first and second row? (Hint: the Bellman equation must hold for every state.)

0.0		-20.	-22.
-14.	-18.	-20.	
	-20.	-18.	-14.
-22.	-20.		0.0

(a) -20, -14

(b) -14, -20

(c) -14, -18

(d) -20, -18

Sol. (b)
For the value in the first row, we have

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r + v_\pi(s')]$$

$$v_\pi(s) = 0.25 * (-1 + v_\pi(s) - 1 - 21 - 19)$$

$$v_\pi(s) = 0.25 v_\pi(s) - 10.5$$

$$0.75 v_\pi(s) = -10.5$$

$$v_\pi(s) = -14$$

Similarly, for the value in the second row, we have

$$v_\pi(s) = 0.25 * (-23 + v_\pi(s) - 1 - 21 - 15)$$

$$v_\pi(s) = 0.25 v_\pi(s) - 15$$

$$0.75 v_\pi(s) = -15$$

$$v_\pi(s) = -20$$

5. If π is the equiprobable random policy, what are the respective values of $q_\pi(s_1, down)$ and $q_\pi(s_2, down)$ given that s_1 is the last cell in the third row (value -14) and s_2 is the last cell in the second row?

(a) -1, -15

(b) -15, -21

(c) 0, -14

(d) -13, -19

Sol. (a)
For s_1, we have

$$q_\pi(s_1, down) = \sum_{s'} p(s'|s_1, down)[r + v_\pi(s')]$$

$$q_\pi(s_1, down) = -1 + 0 = -1$$

Similarly, for s_2, we have

$$q_\pi(s_2, down) = -1 - 14 = -15$$

6. In a particular grid-world example, rewards are positive for goals, negative for running into the edge of the world, and zero the rest of the time. Are the signs of these rewards important, or only the intervals between them? Prove, using the discounted return equation

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

that adding a constant C to all the rewards adds a constant, K, to the values of all states, and thus does not affect the relative values of any states under any policies. What is K in terms of C and γ?

(a) $K = \frac{1}{C(1-\gamma)}$
(b) $K = C(\frac{1}{1-\gamma} - 1)$
(c) $K = C(\frac{1}{1-\gamma} + 1)$
(d) $K = \frac{C}{1-\gamma}$

Sol. (d)
Assume that the grid-world problem is a continuing task. For some policy π and state s, the value function can be give as

$$v_\pi(s) = E_\pi\{G_t|s_t = s\}.$$

Using the discounted reward equation, we have

$$v_\pi(s) = E_\pi\{\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|s_t = s\}.$$

Adding a constant C to all rewards, we have

$$\begin{aligned} v'_\pi(s) &= E_\pi\{\sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + C)|s_t = s\} \\ &= E_\pi\{\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + C\sum_{k=0}^{\infty} \gamma^k|s_t = s\} \\ &= v_\pi(s) + \frac{C}{1-\gamma}. \end{aligned}$$

We see that adding a constant C to all rewards does not affect the relative values of any states under any policies. Here $K = \frac{C}{1-\gamma}$.

7. Given a reinforcement learning problem, algorithm A will return the optimal state value function for that problem and algorithm B will return the optimal action value function. Your aim is to use the value function so obtained to behave optimally in the environment. Assuming that you know the expected rewards but not the transition probabilities corresponding to the problem in question, which algorithm would you prefer to use for your control task?

   (a) algorithm A

   (b) algorithm B

   Sol. (b)
   Since algorithm B returns the optimal action value function, we can use the information provided by the optimal action value function to control the behaviour of the agent without knowledge of the transition probabilities of the underlying MDP.

8. In proving that L_π is a contraction, we had the expression

$$\gamma \sum_j p(j|s)[v(j) - u(j)] \leq \gamma ||v - u|| \sum_j p(j|s)$$

   This inequality holds because

   (a) $v(j) - u(j)$ is a component of $||v - u||$

   (b) the max norm of the difference on the LHS is less than the max norm of the difference on the RHS

   (c) the difference in the LHS can be negative but the norm in the RHS is non-negative

   (d) the max norm on the RHS can at worst be equal to the difference in the LHS

   Sol. (d)

9. We defined the operator $L_\pi : V \to V$ as $L_\pi v = r_\pi + \gamma P_\pi v$. Having seen the proof of the Banach fixed point theorem and assuming that v^π and v^* have their usual meanings, which among the following are implications of showing that L_π is a contraction?

   (a) v^π is a fixed point of L_π

   (b) v^π is a unique fixed point of L_π

   (c) repeatedly applying L_π starting with an arbitrary $v \in V$ results in convergence to v^π

   (d) repeatedly applying L_π starting with an arbitrary $v \in V$ results in convergence to v^*

   Sol. (b), (c)
   Note that while the statement of option (a) is true, it is a result of the Bellman equation and the definition of the L_π operator. Option (d) is not true since repeated application of the operator guarantees convergence only to v^π and not the optimal v^*.

10. Given a value $v \in V$, suppose $L_\pi v = v'$. Then we can conclude that

    (a) $v = v'$

    (b) $v \neq v'$

    (c) $||L_\pi v - L_\pi v'|| \leq \lambda ||v - v'||$, $0 \leq \lambda < 1$

(d) none of the above

Sol. (c)
The first option may not hold if $v \neq v_\pi$. Similarly, the second option may not hold if $v = v_\pi$. The third option is true because L_π is a contraction and in all three possible scenarios ($v \neq v' \neq v_\pi$, $v \neq v' = v_\pi$, and $v = v' = v_\pi$), the statement holds.