Multi-Agent Systems Assignment 5

Konstantin Mihhailov, Haohui Zhang, Roel van de Laar, Henrik Kathmann

December 10, 2021

1 Bellman equations

Rewrite the Bellman equations for v_π and q_π for the following special cases:

1.1 Deterministic policy π: each state is mapped to a single action (say as);

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

$$q_\pi(s, a) = \sum_{s',r} p(s', r|s, a_s)[r + \gamma q_\pi(s', a_s)]$$

1.2 Combination of deterministic policy and deterministic transition

$$V_\pi(s) = r + \gamma V_\pi(s_a)$$

$$q_\pi(s, a) = r + \gamma q_\pi(s_a, a_s)]$$

2 MDP 1

Consider an MDP with a circular state space with an odd number of nodes (i.e. the nodes are positioned along a circle and labeled 0 through n, with n even). Assume that the 0-node is an absorbing terminal state and arriving at this state yields a one-time reward of 10. In the other nodes, one can go in either one of the two circle directions, resulting in reward of 0 (unless you transition to the terminal state). Assume an equiprobable policy π (i.e. going in either direction with prob 1/2) and no discounting (i.e. $\gamma = 1$).

2.1 What would be the corresponding values functions v_π and q_π?

$$\upsilon_\pi(s) = E_\pi(\sum_{n=1}^{n} 10 \mid s_0 = s)$$

Since the probability of going in each direction is 0.5 and each non-terminal state has a value of 0, the R will always be 10 for each starting state S because the agent will always end up in the absorbing state with the same reward.

2.2 What would be an optimal policy? Is this unique? What are the corresponding value functions $v*$ and $q*$?

$$\upsilon_*(s) = \max \upsilon_\pi(s)$$

Assuming that the agent does not know in which state it is, one optimal policy would be to stick to going in one direction with probability 1, until you reach the absorbing state.

2.3 How would your answer for (2) change if each non-terminal step accrued a reward of $r_{NT} = -1$?

By changing the reward of each non-terminal step to -1 it becomes possible to calculate the shortest route to the absorbing state by maximizing the reward, using v_π, from the current state to the absorbing state.

2.4 How would your answer for (2) change if $\gamma < 1$? (Assume $r_{NT} = 0$).

Depending on the starting point, the $\gamma < 1$ of for example, $\gamma = 0.9$ would be able to identify the direction toward the ultimate reward of 10. By applying the below formula to get $v(s)$ for the states:

$$v(s) = \max(R(s, a) + \gamma v(s')$$

2.5 How would your answer for (2) change if the number of non-terminal states was odd? (Assume $r_{NT} = 1$ and $\gamma = 1$)

If the number of non-terminal states is odd, that means that there is one non-terminal state that is exactly in the middle. Therefore, because the reward of all non-terminal states is equal to -1, when the agent starts in the state that is exactly in the middle it should go in either direction with probability 0.5 again.

3 MDP 2

3.1 Values of non-terminal states

In calculating the reward, we can easily see that the value function 1 is the same as the value function 6, 2 is the same as 5 and 3 is the same as 4.
Now, if we substitute v(1) and v(3) into v(2), we obtain:

$$v(2) = v(5) = 0.5(10 + 0.5v(2)) + 0.5(0.5v(2)) = 10$$

as well as:

$$v(3) = v(4) = 0.5 * 10 = 5$$

3.2 Optimal policy

The optimal strategy would be to end up in the terminal node A regardless of the starting node. A possible optimal strategy is:

$$\pi(a|s) = \begin{cases} 1, & \text{if s' is closer to A} \\ 0, & \text{otherwise} \end{cases}$$

This $\pi*$ is not unique. If, for example, the agent starts in node 6, it will be going to absorbing terminal node A and get the optimal reward. However, if it would go to node 5 first (and even 4) and subsequently go back to A, the agent would receive an equal optimal reward.

3.3

Since, $\gamma = 1$ en $R_{NT} = -1$ the optimal strategy will still be that beginning from each node will end up in terminal absorbing node A. Therefore, the optimal reward is:

$$v_\pi(1) = v_\pi(6) = 20$$
$$v_\pi(2) = v_\pi(5) = 19$$
$$v_\pi(3) = v_\pi(4) = 18$$

A possible optimal strategy is:

$$\pi(a|s) = \begin{cases} 1, & \text{if s' is closer to A} \\ 0, & \text{otherwise} \end{cases}$$

This $\pi*$ is unique.

3.4

Since, $\gamma = 1$ en $R_{NT} = -10$ the optimal strategy will be that beginning from each other node than 3 and 4, the agent will end up in terminal absorbing node A. Therefore, the optimal reward is:

$$v_\pi(1) = v_\pi(6) = 20$$
$$v_\pi(2) = v_\pi(5) = 10$$
$$v_\pi(3) = v_\pi(4) = 0$$

A possible optimal strategy is:

$$\pi(a|s) = \begin{cases} 1, & \text{if s' is closer to A} \\ \frac{1}{2}, & \text{if the current s is node 3 or 4} \\ 0, & \text{otherwise} \end{cases}$$

This $\pi*$ is not unique. If, for example, the agent starts in node 4 it receives the same (optimal) reward if it moves to absorbing terminal node B directly or goes via node 5, and 6 and ends up in A.

4 4 Q-learning and SARSA

4.1 4.1

The Q-learning function is:

$$Q(s_t, a_t) \to Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a(Q(s_{t+1}, a)) - Q(s_t, a_t)]$$

The next state for $q_\pi(2, R)$ is state 3, because $q(3, R) = 6 > q(3, L) = 3$, for Q-learning, we will select the action with the highest q. Now the equation becomes:

$$Q(2, R) \to Q(2, R) + 0.9 * [-1 + \frac{2}{3}Q(3, R) - Q(2, R)] = 3.2$$

4.2 4.2

The expected SARAS formula is:

$$q_\pi(s_t, a_t) \to q_\pi(s_t, a_t) + \alpha[r_{t+1} + \gamma \sum_a \pi(a|s_{t+1})q_\pi(s_{t+1}, a) - q_\pi(s_t, a_t)]$$

Thus:

$$q_\pi(2, R) \to q_\pi(2, R) + 0.9 * [-1 + \frac{2}{3}(\frac{1}{2}q_\pi(3, R) + \frac{1}{2}q_\pi(3, L)] - q_\pi(2, R)] = 2.3$$