Multi-Agent Systems

Homework Assignment 5

MSc AI, VU

E.J. Pauwels

Version: December 3, 2021— Deadline: Friday, December 10, 2021 (23h59)

NB: Unless otherwise indicated, the problems below can be solved using pen and paper.

1 Bellman equations

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

1. Deterministic policy π: each state is mapped to a single action (say a_s);

$$\pi(a \mid s) = \begin{cases} 1 & \text{if } a = a_s \\ 0 & \text{otherwise} \end{cases}$$

2. Combination of deterministic policy and deterministic transition $p(s' \mid s, a)$. The latter is characterized by the fact that applying an action a to a state s results each time in the same successor state s_a;

$$p(s' \mid s, a) = \begin{cases} 1 & \text{if } s' = s_a \\ 0 & \text{otherwise} \end{cases}$$

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. $\gamma = 1$).

1. What would be the corresponding values functions v_π and q_π?

2. What would be an optimal policy? Is this unique? What are the corresponding value functions v^* and q^*?

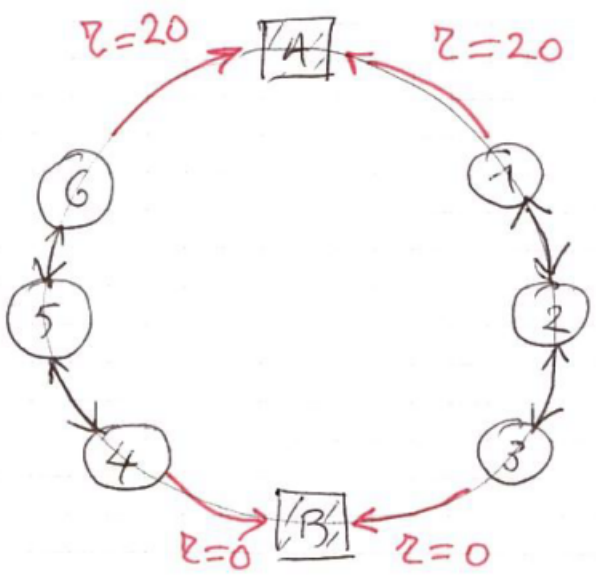
3. How would your answer for (2) change if each non-terminal step accrued a reward of $r_{NT} = -1$?

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

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

3 MDP 2

Consider the following MDP with circular state space (see figure below). The top and bottom states (A and B) are absorbing, terminal states. The immediate reward for moving to the terminal state A at the top is $+20$. The immediate reward for moving to the bottom terminal state B is 0. Transitions between two non-terminal states yield an immediate reward r_{NT}. From non-terminal states one can move in both directions (but staying in place is not allowed). We assume for all the subquestions below that there is **no discounting**, i.e. $\gamma = 1$.



1. Assume $r_{NT} = 0$. Consider a policy π which assigns equal probabilities to the two possible "directions" in each of the nodes 1 through 6. What would be the values $v_\pi(1), \ldots, v_\pi(6)$, i.e. the long-term return for each of the six non-terminal states? Explain.

2. For the setup defined above, what would be an optimal policy π^* and the corresponding optimal values $v^*(1), \ldots, v^*(6)$. Is π^* unique?

3. Now assume that $r_{NT} = -1$. Again, what would be an optimal policy π^* and the corresponding optimal values $v^*(1), \ldots, v^*(6)$. Is π^* unique?

4. Finally, assume that $r_{NT} = -10$. Again, what would be an optimal policy π^* and the corresponding optimal values $v^*(1), \ldots, v^*(6)$. Is π^* unique?

4 Q-learning and SARSA

Consider the MDP with a linear state space, i.e. states are located on a line and agents can only move to the immediate neighbours. In each state there are two possible actions: move left (a = L) or right (a = R) and transitions are deterministic. After a number of iteration steps, some of the action values, immediate rewards and current q-values are given by the table below. Consider an equiprobable policy π that picks actions L and R with equal probability $1/2$. Furthermore, assume throughout a learning rate $\alpha = 0.9$ and discount factor $\gamma = 2/3$.

$state(s)$	$action(a)$	$reward(r)$	$q(s, a)$
2	R	-1	5
2	L	0	4
3	R	1	6
3	L	-2	3

1. Compute the next value for $q_\pi(2, R)$ under one **Q-learning** iteration (i.e. only update this state-action pair).

2. **Expected SARSA** is a variation on SARSA which computes the update using the following formula:

$$q_\pi(S_t, A_t) \leftarrow q_\pi(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \sum_a \pi(a \mid S_{t+1}) q_\pi(S_{t+1}, a) - q_\pi(S_t, A_t) \right]$$

   Compute the next value for $q_\pi(2, R)$ under one iteration step of expected SARSA (using the equiprobable policy π).