# Multi-Agent Systems

## Homework Assignment 5

## MSc AI, VU

E.J. Pauwels

Version: December 2, 2024— Wednesday, December 11, 2024 (23h59)

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

## 1 Bellman equations

Rewrite the Bellman equations for $v_\pi$ and $q_\pi$ for the following special cases:

1. Deterministic policy $\pi$: 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 $\pi$ (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 $v_\pi$ and $q_\pi$?

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

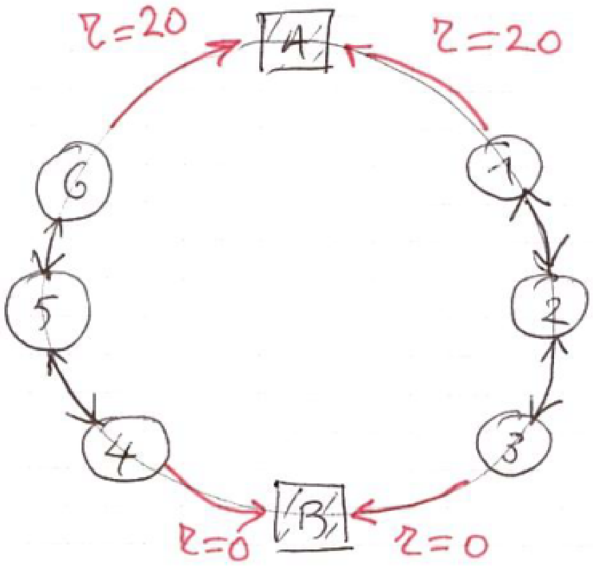
3. How would your answer for (2) change if each non-terminal step accrued a reward of $r_{NT} = -1$?

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 $\pi$ 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.

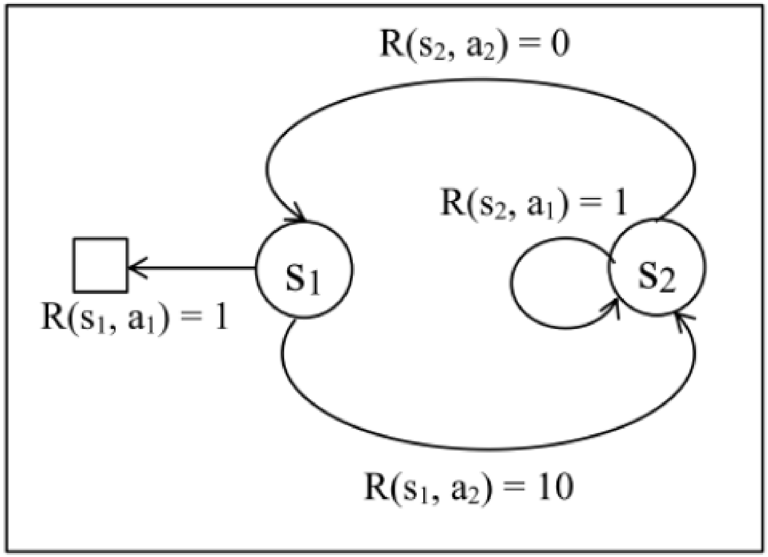
2. For the setup defined above, what would be an optimal policy $\pi^*$ and the corresponding optimal values $v^*(1), \ldots, v^*(6)$. Is $\pi^*$ unique?

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

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

# 4 MDP 3

Consider the following 2-state MDP:



In both states ($s_1$ and $s_2$), there are two possible actions ($a_1$ and $a_2$). The actions result in deterministic transitions. Taking action $a_1$ in state $s_1$ results in a reward of 1, and ends the episode. Taking action $a_2$ in state $s_1$ results in a reward of 10, and brings the agent to state $s_2$. In state $s_2$ action $a_1$ results in a reward of 1 and the agent stays in state $s_2$. Action $a_2$ results in a reward of 0 and brings the agent to state $s_1$. The agent will act (and continues to receive rewards) until the episode ends.

### Questions

- For a discount factor $\gamma = 0.9$, what is the optimal policy $\pi^*$?
- Provide the corresponding optimal value $v^*(s_2) = v_{\pi^*}(s_2)$ in state $s_2$ . Please explain your reasoning and provide your derivation.
- Is it possible to adjust the discount factor $\gamma$ in such a way that the optimal policy changes? Explain how you would decide whether this is possible or not. If the answer is affirmative, provide an example $\gamma$, the corresponding optimal policy, and its corresponding optimal value-function. If you think the answer is negative, provide argument(s),