Multi-Agent Systems
VU AI MSc
Final Exam RESIT

E.J. Pauwels

16 February 2022, 18h45 – 21h00

General Remarks

BEFORE YOU START

- Check if your version of the exam is complete. Your copy should have **4 questions**.
- Write down your **name and student ID number** on each (or at least the first) sheet.
- The use of a calculator is allowed (but isn't really necessary).

PRACTICAL MATTERS

- You are obliged to identify yourself at the request of the examiner (or his representative) with a proof of your enrollment or a valid ID.
- During the examination it is not permitted to visit the toilet, unless the invigilator gives permission to do so.
- **DO NO REMOVE THE STAPLE!** Write your answers and solutions on the blank pages in this exam bundle.

GOOD LUCK!

1 Game Theory (25%)

Consider the following ranked coordination game in which players can choose between options A and B. They get positive utility when they coordinate, and negative utility when they anti-coordinate. However, coordinating at option A is preferable (whence: ranked). The pay-off matrix is given below:

		Player_2	
		A	B
Player_1	A	$2, 2$	$-1, -1$
	B	$-1, -1$	$1, 1$

Questions

1. What are the Nash equilibria (pure and/or mixed) for this game? Explain.

2. Compute the (expected) utilities for all of the equilibria. How do the equilibria compare in terms of Pareto-dominance?

3. What if coordination at A (i.e. both players select A) yields less utility for player 2, say: $u_2(A, A) = 2 - b$ where $0 < b < 1$. How does that affect the Nash equilibria (indicating trends suffices, no need for detailed computations).

2 Sequential games (25%)

Consider the following sequential game (see Fig. 1 below). Player 1 takes the first move (actions A or B), whereupon player 2 takes the second move. In the case player 1 started with action B, the subsequent action taken by player 2 is not revealed to player 1 when it is his turn again to take the next action. Put differently, at that point, the information set of player 1 is non-trivial (as indicated by the dotted line). Further actions by both players and the ensuing rewards are indicated in the figure.

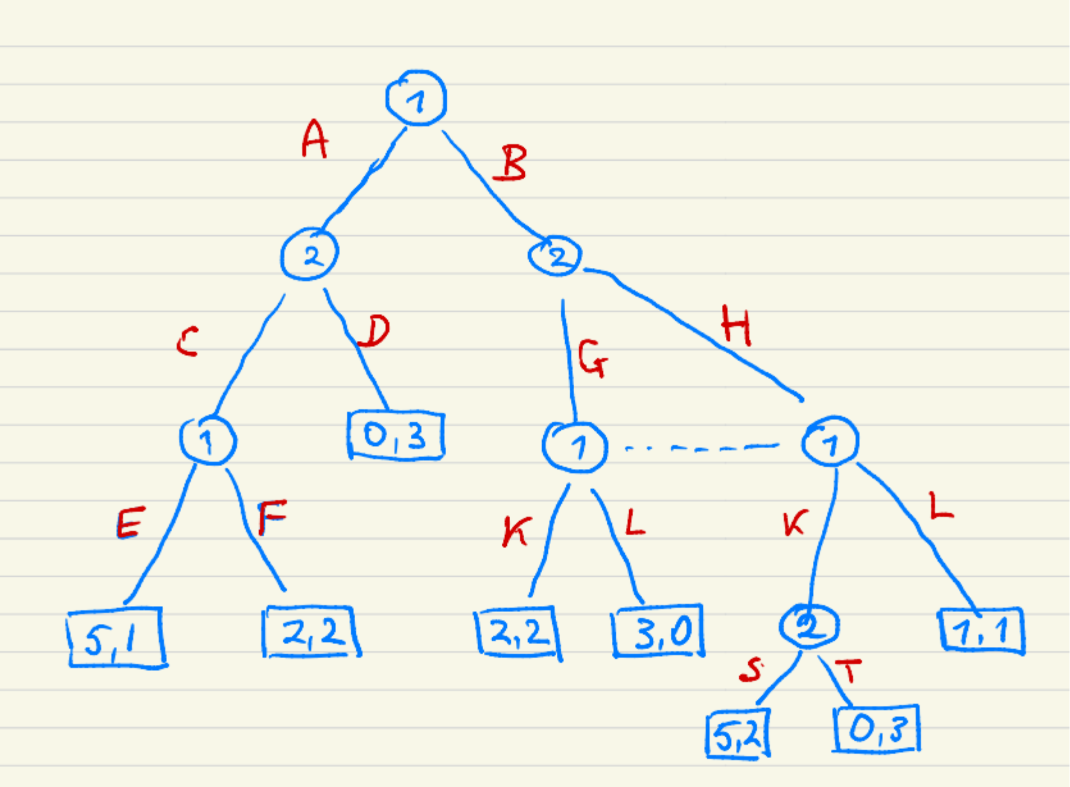

Figure 1: Sequential game

Questions

1. Use backward induction to solve this game (i.e. what actions will players take, and what is the corresponding payoff).
2. Indicate (e.g. in the above figure) all the valid subgames.
3. When you want to write down this game in normal form, you need to list all the pure strategies for each of the players. Write down the list of pure strategies for each player (**no need** to write down the normal form matrix).

3 Markov Decision Process (25%)

Consider the following 2-state MDP (i.e. non-absorbing states) depicted in Fig. 2. 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. Similarly, taking action a_1 in s_2 yields an immediate reward of 3 and ends the episode. Action a_2 in s_1 sends the agent to s_2 and yields an immediate reward of 10. Similarly, action a_2 in s_2 sends the agent to s_1 and yields an immediate reward of -2. The agent will act (and continue to receive rewards) until the episode ends.

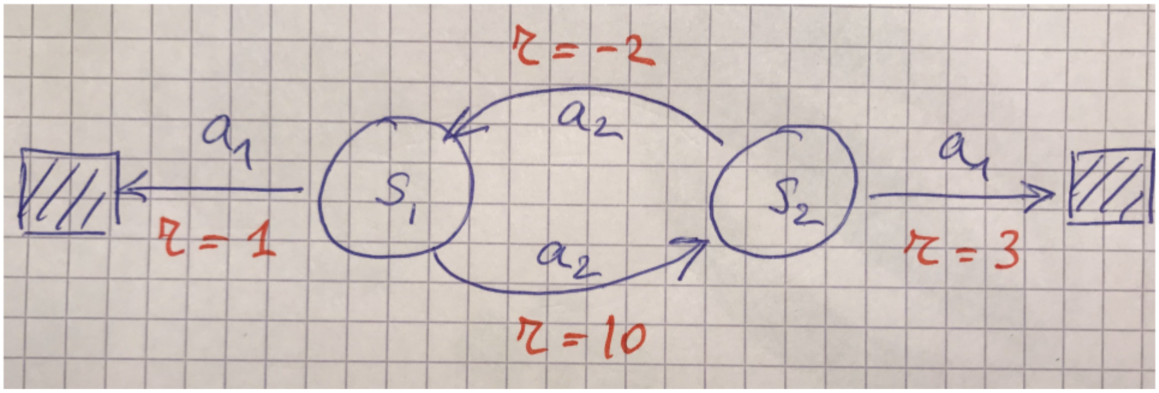

Figure 2: MDP

Questions

1. Is this an infinite horizon, or a finite horizon problem?

2. For a discount factor $\gamma = 0.9$, what is the optimal policy π^*. Provide the corresponding optimal value $v^*(s_2) = v_{\pi^*}(s_2)$ in state s_2. Please explain your reasoning and provide your derivation.

3. Same question as above, but now for discount factor $\gamma = 0.1$.

4. Based on the conclusions above, can you determine (or at least estimate) the threshold value for γ at which the optimal policy switches?

PS: Provide the correct formulae for your answers, even if you can't compute the corresponding numerical results.

4 Reinforcement Learning (25%)

Consider the MDP with a linear state space, i.e. all the states are positioned along a horizontal line. In each state there are two possible actions: move left ($a = L$) or right ($a = R$). After a number of iteration steps, some of the action values, immediate rewards and current v and q-values are given by the table below.

Consider a policy π that picks actions L and R according to the policy probabilities $\pi(a \mid s)$ listed in the table below. Furthermore, assume throughout a learning rate $\alpha = 0.9$ and discount factor $\gamma = 2/3$. Notice that some values in the table are actually missing (as indicated by double question marks "??").

$state(s)$	$action(a)$	$next\,state(s')$	$reward(r)$	$q(s,a)$	$v(s)$	$\pi(a \mid s)$
2	R	3	-1	??	5	1/4
2	L	1	0	4	5	3/4
3	R	4	1	6	7	2/3
3	L	2	-2	??	7	1/3

Questions

1. What is your best estimate for the two missing values in the table. Explain your reasoning.
2. Compute the next value for $q_\pi(2, R)$ under one **Q-learning** iteration (i.e. only update this state-acion pair). Specify both the update formula you're using, and the numerical value that you obtain.
3. Why is Q-learning called **off-policy**? Explain.