MAS exam 2019 Solutions

Multiagent systems (Vrije Universiteit Amsterdam)

Multi-Agent Systems
VU AI MSc
Final Exam

17 December 2019, 8:45 – 11:30

General Remarks

BEFORE YOU START

- Check if your version of the exam is complete. Your copy should have 13 printed pages, this one included. The last page is blank and can be used as scrap paper.
- Write down your **name and student ID number** on each sheet.
- **Do NOT remove the staple!**
- The blank space provided for each question should be (more than) sufficient for your answer. You can also use the blank last page, if necessary.
- Your mobile phone has to be switched off and in your coat or bag. Your coat and bag must be under your table.
- 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.
- 15 minutes before the end, you will be warned that the time to hand in is approaching.

GOOD LUCK!

1 Copying from Wikipedia for homework

Student life is hectic, and there are many essential life skills to be acquired in limited time: throwing and enjoying great parties, conducting (more or less) profound philosophical discussions into the morning hours, exploring the teeming metropolitan bio-sphere, ...to name just a few. It is therefore completely understandable that homework assignments are seen as an unwelcome distraction and need to be dealt with as efficiently as possible. Fortunately, quite often you can simply copy the relevant answers from Wikipedia, saving a lot of valuable time. Unfortunately, the TAs, who in a recent past used to be students themselves, are aware of these time-saving practices and are prone to check the homeworks for plagiarism. It requires more effort on their part, but they get a lot of satisfaction from catching cheating students. In fact, this situation can be interpreted as a simultaneous game with the following pay-off matrix:

		Student	
		Honest	*Wikipedia*
TA	*Check*	$5, \underline{0}$	$\underline{7}, -20$
	No_check	$\underline{10}, 0$	$2, \underline{10}$

Questions

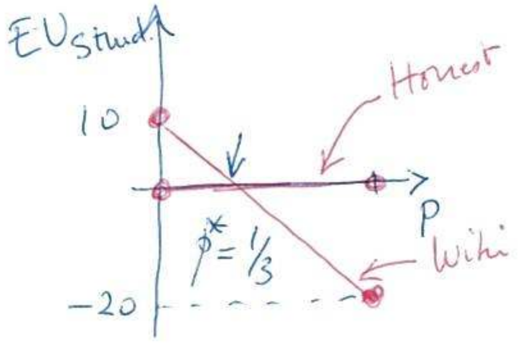
1 - (4pts) Determine all the Nash equilibria (NE) for this game.

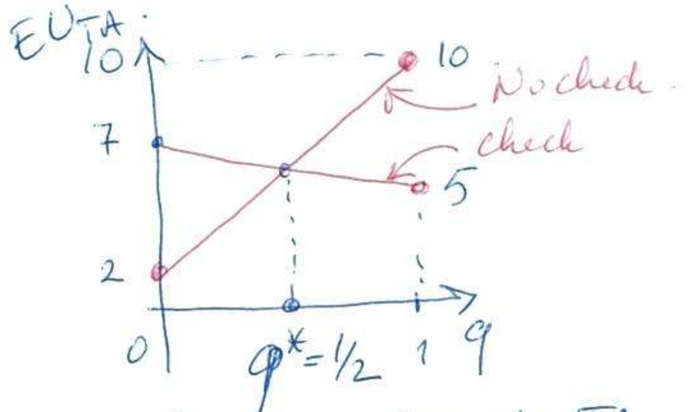
2 - (2pts) For each of the NE, compute the expected utility for both student and TA.

3 - (4pts) What can be done (e.g. in terms of pay-offs) to reduce the probability that a student will cheat?

① No pure NE.

Mixed NE : TA mixes with prob p (Check), 1-p (No-check)
Student - - - - - - q (Honest), 1-q (Wiki).





Expected utility for student

$EU_{st}(\overset{Honest}{\cancel{check}}) = EU_{st}(\overset{Wiki}{\cancel{no\ check}})$

$\overset{||}{0} = -20p + 10(1-p)$

$\Rightarrow \boxed{p^* = 1/3}$

Expected utility for TA

$EU_{TA}(Check) = EU_{TA}(nocheck)$

$5q + 7(1-q) = 10q + 2(1-q)$

$\Rightarrow \boxed{q^* = 1/2}$

Solution page (continued)

2) Expected utility in Mixed NE $(p^* = 1/3,\ q^* = 1/2)$

Probabilities for each ~~state~~ action profile:

	$1/2$ H	$1/2$ W
$1/3$ C	$1/6$	$1/6$
$2/3$ N	$2/6$	$2/6$

utilities

5, 0	7, -20
10, 0	2, 10

$$EU_{TA} = \frac{1}{6}\cdot 5 + \frac{2}{6}\cdot 10 + \frac{1}{6}\cdot 7 + \frac{2}{6}\cdot 2 = \frac{36}{6} = 6.$$

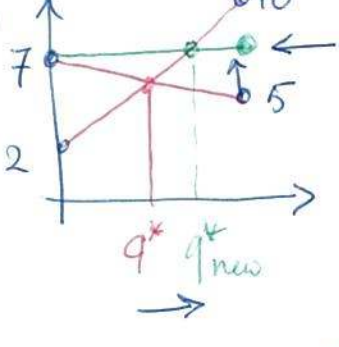
$$\parallel$$

$$EU_{TA}(p^*, q^*)$$

$$EU_{st.}(p^*, q^*) = \frac{1}{6}\cdot 0 + \frac{2}{6}\cdot 0 + \frac{1}{6}\cdot(-20) + \frac{2}{6}(10) = 0.$$

3) To improve fraction of non-cheating students (q) we can change the utilities of the TA in such a way that the intersection point moves to the right.

Eg



← new payoff for TA for (C, H)

+ other possibilities
10↓, 2↓ of 7↑.

Similar for other possible changes.

2 Markov Decision Processes (MDP)

Consider an MDP with a finite number of states $s_1, s_2, \ldots, s_n$ and actions $a_1, a_2, \ldots, a_k$. For this MDP we define a policy π that specifies the conditional probabilities $\pi(a \mid s)$. The state value function $\mathbf{v}_\pi$ satisfies the matrix form of the Bellman equation:

$$\mathbf{v}_\pi = \gamma P \mathbf{v}_\pi + \mathbf{r}$$

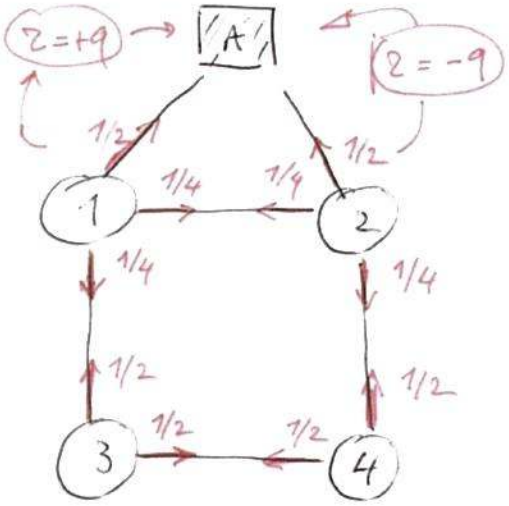
where

- $P(s, s') = \sum_a \pi(a \mid s) p(s' \mid s, a)$
- $\mathbf{r}(s) = \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a) r(s, a, s')$,

Questions:

1. (2pts) Explain in words the meaning of $P(s, s')$ and $\mathbf{r}(s)$;
2. (2pts) Explain in words the meaning of the *product* $P(s, s')P(s', s'')$. How is this different from $P^2(s, s'')$, i.e. the (s, s'') entry of the matrix $P^2 = P \cdot P$?
3. (2pts) Consider the MDP depicted in the figure below. State A is absorbing. Transition to A from state 1 yields an immediate reward of 9. Transition to A from state 2 yields an immediate reward of -9. All other transitions incur a reward of -1. On this MDP we consider a policy π that assigns transition probabilities as indicated in the figure below. E.g.: $\pi(\text{move to A} \mid \text{currently in state 1}) = 1/2$ and $\pi(\text{move to 4} \mid \text{currently in state 2}) = 1/4$, etc. Transitions are deterministic (i.e. each action maps a state s to a unique successor state s').

   What are P and $\mathbf{r}$ in this concrete case? Make sure to include the absorbing state A in both P and $\mathbf{r}$.
4. (2pts) Assuming γ is sufficiently small (e.g. $\gamma = 0.1$). How would you calculate an *approximate* solution for $\mathbf{v}_\pi$?
5. (2pts) Determine the optimal state value function $\mathbf{v}^*$ assuming $\gamma = 1/3$.



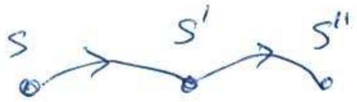
Solution page (continued)

① $\mathbb{P}(s,s')$ = prob of transition $s \to s'$ (under policy π)

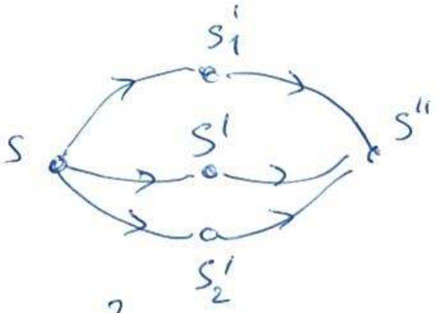
$r(s)$ = expected immediate reward in state s (under π)

② $P(s,s')P(s',s'')$ = prob of transition $s \to s' \to s''$

$\mathbb{P}^2(s,s'')$ = prob of transition $s \to s''$ in two steps (over any intermediate state)



$P(s,s')P(s',s'')$



$\mathbb{P}^2(s,s'')$

③ Transition matrix: $P(s,s') = \sum_a \pi(a,s) \underbrace{p(s'|s,a)}_{\downarrow \text{ degenerate } (1/0)}$

$$\mathbb{P} = \begin{array}{c|ccccc|} & 1 & 2 & 3 & 4 & A \\ \hline 1 & 0 & 1/4 & 1/4 & 0 & 1/2 \\ 2 & 1/4 & 0 & 0 & 1/4 & 1/2 \\ 3 & 1/2 & 0 & 0 & 1/2 & 0 \\ 4 & 0 & 1/2 & 1/2 & 0 & 0 \\ A & 0 & 0 & 0 & 0 & 1 \\ \hline \end{array}$$

Solution page (continued)

$$r(1) = \underbrace{\tfrac{1}{2}\cdot 9}_{\uparrow} + \underbrace{\tfrac{1}{4}(-1)}_{\uparrow} + \underbrace{\tfrac{1}{4}(-1)}_{\llcorner} = 4.$$

$\llcorner$ immediate reward $\rightarrow 3$.

$\uparrow$ immediate reward for transition to 2.

$\uparrow$ immediate reward for transition to A

$$r(2) = \tfrac{1}{2}(-9) + \tfrac{1}{4}(-1) + \tfrac{1}{4}(-1) = -5$$

$$r(3) = r(4) = -1 \qquad r(A) = 0.$$

④ Bellman in matrix form:

$$v = \gamma P v + r \Rightarrow (I - \gamma P)\, v = r$$

$$\Rightarrow v = (I - \gamma P)^{-1} r = (I + \gamma P + \gamma^2 P^2 + \ldots)\, r.$$

for $\gamma = 0.1 \rightarrow \gamma^2 = 0.01 \rightarrow$ can be neglected.

$$\Rightarrow v \cong (I + \gamma P)\, r \simeq r + \gamma P r.$$

⑤ The optimal policy is to go to A through 1 (not 2) as fast as possible.

Hence the corresponding value function v^* is given by:

$v^*(1) = 9$ (move to A)

$v^*(2) = \underset{\uparrow}{-1} + \gamma \cdot 9 \longrightarrow$ ~~transition~~ discounted $v^*(1)$

$\uparrow$ transition $2 \rightarrow 1$

$= -1 + 3 = 2.$

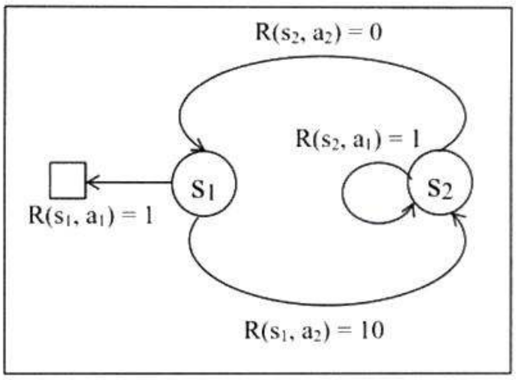
$$\boxed{v^*(A) = 0}$$

$v^*(3) = v^*(2)$

$v^*(4) = -1 + \gamma v^*(3) = -1 + \frac{1}{3}\cdot 2 = -1/3 = -1 + \gamma v^*(2)$

3 MDP 2

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

- Is this an infinite horizon, or a finite horizon problem?
- 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.
- Is it possible to adjust the discount factor γ 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 γ, the corresponding optimal policy, and its corresponding optimal value-function. If you think the answer is negative, provide argument(s),

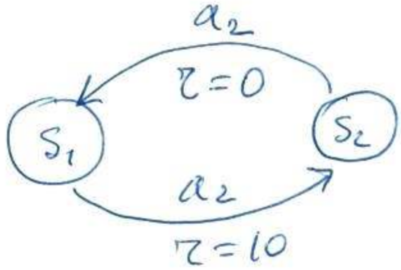
PS: provide the correct formulae for your answers, even if you can't compute the corresponding numerical result.

① Infinite horizon; process might loop for ever.

② Consider the following two policies (see overleaf).

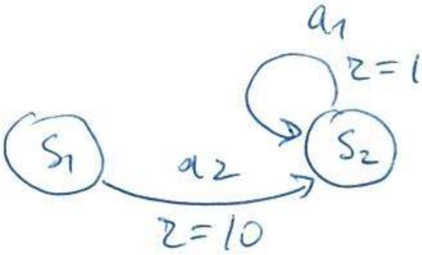
Solution page (continued)

policy 1 : π_1



$$v(S_2) = 0 + 10\gamma + 0 + 10\gamma^3 + 0 + 10\gamma^5 + \dots$$
$$= 10\gamma(1+\gamma^2 + \gamma^4 + \dots) = \frac{10\gamma}{1-\gamma^2} = \frac{9}{1-0.81} \simeq 47.4$$

policy 2 π_2



$$v(S_2) = 1 + \gamma + \gamma^2 + \dots = \frac{1}{1-\gamma} = \frac{1}{0.1} = 10.$$

Conclusion:

- $*$ Policy 1 is better for $\gamma = 0.9$.
- $*$ No need to consider S_1 since we are taking the same action (a_2) under both policies and the utility is at least 10 (better than 1 for transition to absorbing state).

③ We will switch from policy 1 to policy 2 when: $\frac{10\gamma}{1-\gamma^2} = \frac{1}{1-\gamma} \Rightarrow \frac{10\gamma}{(1-\gamma)(1+\gamma)} = \frac{1}{1-\gamma}$

$$\Rightarrow \frac{10\gamma}{1+\gamma} = 1 \Rightarrow 9\gamma = 1 \Rightarrow \boxed{\gamma = 1/9}$$

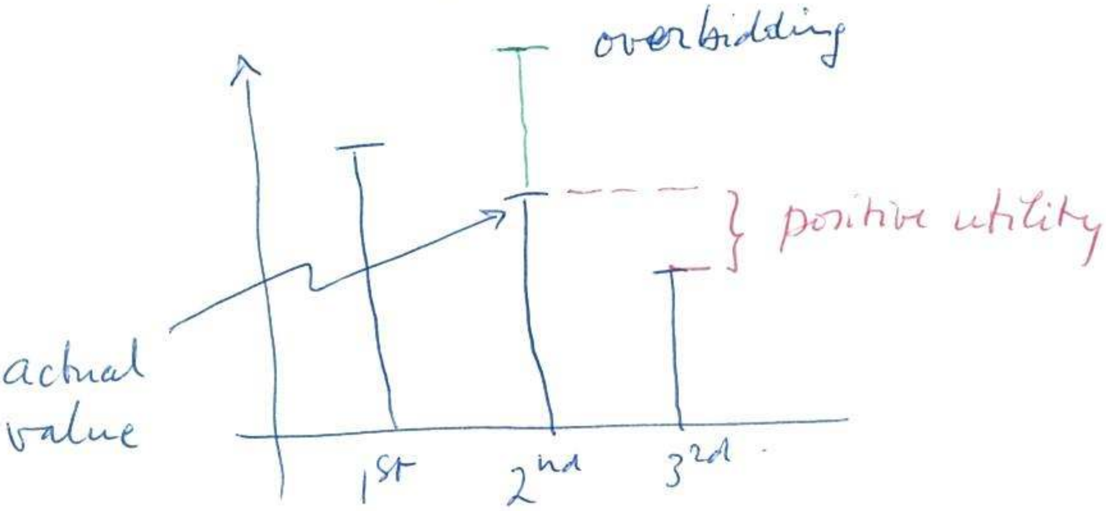
4 Vickrey auction

1. (6pts) A Vickrey auction is a sealed bid, second price auction. Explain why truth-telling is a dominant strategy for this auction.
2. (4pts) Would these properties also hold for a sealed bid, third price auction? If affirmative, why would an auctioneer prefer a Vickrey auction to a third price auction? If your answer is negative, explain the difference?

① See theory (slides)

② No: properties do not hold.

Reason: The 2nd highest bidder will be tempted to overbid. (ie, put in a bid that is higher than the winner).
This way he will win, but still have a positive utility



Solution page (continued)

5 Q-learning and SARSA

(10pts) 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 q-values are given by the tabel below. Consider a policy π that picks actions L and R according to the probabilities $\pi(a \mid s)$ listed in the table below. Furthermore, assume throughout a learning rate $\alpha = 0.9$ and discount factor $\gamma = 2/3$.

$state(s)$	$action(a)$	$next\,state(s')$	$reward(r)$	$q(s,a)$	$\pi(a \mid s)$
2	R	3	-1	5	$1/4$
2	L	1	0	4	$3/4$
3	R	4	1	6	$2/3$
3	L	~~1~~2	-2	8	$1/3$

1. (2pts) Using this table, what is the current estimate for $v_\pi(2)$ and $v_\pi(3)$ of the state value function v_π.

2. (4pts) 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. (4pts) **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 policy π specified above).

① $$v_\pi(s) = \sum_a \pi(a|s)\, q(s,a)$$

Hence: $$v_\pi(2) = \frac{1}{4} \cdot 5 + \frac{3}{4} \cdot 4 = \frac{17}{4} = 4.25$$

$$v_\pi(3) = \frac{2}{3} \cdot 6 + \frac{1}{3} \cdot 8 = \frac{20}{3} = 6.67$$

② Q-learning update

$$q(s,a) \longleftarrow q(s,a) + \alpha \left[r + \gamma \max_{a'} q(s',a') - q(s,a) \right]$$

$$\underbrace{q(2,R)}_{\text{new update value}} \longleftarrow 5 + 0.9 \left[-1 + \frac{2}{3} \underbrace{\max(6,8)}_{8} - 5 \right] = \underline{\underline{4.4}}$$

Solution question 5 (continued)

③ Expected SARSA:

$$q_\pi(2,R) \longleftarrow 5 + 0.9\left[-1 + \frac{2}{3}\Big(\underbrace{\frac{2}{3}6 + \frac{1}{3}\cdot 8}_{v(3)=\frac{20}{3}}\Big) - 5\right]$$

$$= 5 + 0.9\left[-6 + \frac{2}{3}\cdot\frac{20}{3}\right] = \underline{\underline{3.6}}\,.$$

***** MY BEST WISHES FOR A JOYFUL AND EXCITING 2020 ! *****