Eindexamen voor de cursus Multi Agent Systems Dec 2020

Multiagent systems (Vrije Universiteit Amsterdam)

Multi-Agent Systems
VU AI MSc
Final Exam

E.J. Pauwels

15 December 2020, 12h15 – 14h30

General Remarks

BEFORE YOU START

- 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.
- You can upload your solution paper (as pdf) between 14h15 and 14h45. After that you can still upload, but your paper will be marked as *too late*, and this might have an impact on your final grade.

GOOD LUCK!

1 Game Theory

In the following normal-form game, player 1 has a choice of actions U, M or D, while player 2 can choose between actions L, C and R. The corresponding pay-off matrix is given below.

	L	C	R
U	2, 0	1, 1	4, 2
M	3, 4	1, 2	2, 3
D	1, 3	0, 2	3, 0

Questions

1. What strategies survive *iterated elimination of strictly dominated strategies*?
2. What are the pure-strategy Nash equilibria?
3. Are there any mixed Nash equilibria? If affirmative, provide details.
4. What are the (expected) utilities for each of the players in each of the Nash equilibria.
5. Is it possible to confidently predict the outcome of this game?

2 Game Theory: Investment game

Consider an investment game in which there are an odd number (n) of agents (e.g. $n = 7$), Each agent has only two strategies: he can either invest 10 Euro (I) or not invest (N). The pay-offs are equal to:

$$\textit{pay-off} = \textit{(return on investment)} - \textit{(actual investment)},$$

and are computed as follows:

- An agent that did not invest gets zero return, resulting in zero pay-off;
- For the agents that did actually invest: If there is a **majority** of agents that did invest (i.e. (number of investing agents) $> n/2$) then each investing agent gets a return of 30 Euros, resulting in a net pay-off of $30 - 10 = 20$ Euros. If, on the other hand, the investing agents are in the **minority**, then they get zero return, resulting in a net pay-off of $0 - 10 = -10$ Euro.

Questions

1. What are the **pure** Nash equilibria for this game? Notice that since the number of agents is odd, the majority and minority are well defined.
2. Do the Nash equilibria change when the role of majority and minority are interchanged, i.e. there is positive return (of 30 Euro) for the investing agents when they constitute a minority, and zero return when they are in the majority?

3 Markov Decision Processes (MDP)

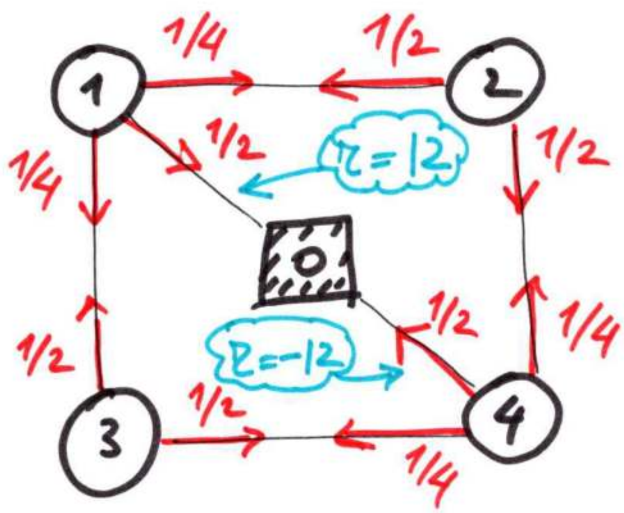
Recall that for a general MDP with a finite number of states $s_1, s_2, \ldots, s_n$ and actions $a_1, a_2, \ldots, a_k$, a policy π specifies the conditional probabilities $\pi(a \mid s)$. The state value function $\mathbf{v}_\pi$ satisfies the matrix form of the Bellman equation:

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

where

- $P_\pi(s, s') = \sum_a \pi(a \mid s) p(s' \mid s, a)$
- $\mathbf{r}_\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a) r(s, a, s')$,
- $0 \leq \gamma \leq 1$ is a discount factor.

Now, consider the specific MDP depicted in the figure below. State 0 is absorbing. Transition to state 0 from state 1 yields an immediate reward of 12. Transition from state 4 to state 0 yields an immediate reward of -12. All other transitions incur a reward of -1. Transitions are **deterministic** (i.e. each action maps a state s to a unique successor state s').
On this MDP, consider a policy π that assigns transition probabilities as indicated in the figure below. E.g.: π(move to state 0 | currently in state 1) $= 1/2$ and π(move to 1 | currently in state 2) $= 1/2$, etc.



Questions

1. For this specific MPD and policy π, write down P_π and $\mathbf{r}_\pi$ explicitly. Make sure to include absorbing state 0.

2. Determine the optimal state value function $\mathbf{v}^*$ assuming $\gamma = 2/3$. Is the corresponding optimal policy unique?

3. Let a be the action that maps state 1 into state 2. What is the optimal state-action value $q^*(1, a)$ (assuming $\gamma = 2/3$)?

4. Suppose now that we use the policy π as specified above (see figure) but that the **transitions are no longer deterministic**: More precisely, assume that with probability $3/4$ an action will induce the expected transition (with reward as above), but with probability $1/4$ will result in *"staying in place"* while picking up a reward ("cost") of -2. As an example, in state 3, the action *"go east"* would induce a transition to state 4 with probability 3/4, while the agent would stay in state 3 with probability 1/4. How would that change the row $P_\pi(1, 0:4)$, i.e. the row that corresponds to starting state $s = 1$. Under these assumptions, what are $\mathbf{r}_\pi(1)$ and $\mathbf{r}_\pi(2)$?

4 Reinforcement Learning and Exploration vs. Exploitation

Bellman equations for the value functions:

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} p(s' \mid s, a)[r(s, a, s') + \gamma v_\pi(s')]$$

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

4.1 Q-learning computation (5pts)

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 ($a = R$). The transitions are deterministic. Consider a policy π that picks actions L and R according to the probabilities $\pi(a \mid s)$ listed in the table below.

After a number of iteration steps, some of the action values, immediate rewards and current v_π and q_π-values are given by 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 "??"), if you need them, you have to compute them yourself.

$state(s)$	$action(a)$	$next\,state(s')$	$reward(r)$	$q(s,a)$	$v(s)$	$\pi(a \mid s)$
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. Compute the next value for $q_\pi(2, R)$ under one **Q-learning** iteration (i.e. only update this state-action pair). Recall that Q-learning uses the update rule:

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

2. Do you have enough information to update $q_\pi(2, R)$ using SARSA?

3. SARSA is called *on-policy* while Q-learning is called *off-policy*. Explain why.

4.2 Monte Carlo estimation of Kullback-Leibler divergence

The Kullback-Leibler divergence for two (continuous) probability distributions f and g is defined by:

$$KL(f||g) := \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx.$$

We have seen that this quantity can be estimated using a Monte Carlo sample:

$$KL(f||g) \approx \sum_{i=1}^{n} \log\left(\frac{f(X_i)}{g(X_i)}\right) \qquad \text{where } X_i \sim f \quad (i = 1\ldots, n)$$

i.e. each X_i is independently sampled from f. Use this Monte Carlo representation to make it *plausible* that the KL-divergence is always positive, i.e. $KL(f||g) \geq 0$. **NO need for a proof, just a (short!) intuitive argument.**

***** BEST WISHES FOR A JOYFUL AND EXCITING 2021 ! *****