# Multi-agent learning 2017-18, mid term exam

This exam consists of five items. You may use 2.5 hours to complete the exam. No exit in the first half hour. No internet, no notes. Calculators are allowed. Answers must be justified. In particular, numeric answers must be justified by a full computation. Less is more: incorrect answer fragments and/or unnecessary long answers may lead to subtraction. Points are evenly divided over items. Within items points are evenly divided over sub-items (if any).

Clearly circle problem numbers on answer sheets. (Facilitates finding answers. Thank you.)

Good luck!

1. Give four performance standards for MAL algorithms followed by descriptions. (Six were discussed in the lectures.)

2. In reinforcement learning, actions can be selected according to cumulative payoff matching (CPM) or according to average payoff matching (APM). Show with the help of an example that there is a difference.

3. The following game is played infinitely many times with $\epsilon$-greedy regret matching against a column player who plays $A$ in the first round and tit-for-tat thereafter.

| | A | B |
|---|---|---|
| A | $0,3$ | $-5,1$ |
| B | $2,-1$ | $4,0$ |

   .

   Play eight rounds under the assumption that $E_A = \{1,5\}$ and $E_B = \{2,7\}$. Recall that $E_x = \{\, t \mid$ player $A$ experimented in round $t$ and played $x$ $\}$.

4. Give a formula for smoothed fictitious play. Describe its relation with fictitious play, describe its relation with no-regret, and describe its convergence properties. (You may miss out on one.)

5. Determine the evolutionarily stable strategies of

| | L | R |
|---|---|---|
| L | $-2,-2$ | $5,1$ |
| R | $1,5$ | $4,4$ |

   .

End of document.

## Answers

1, p. 1: See the introductory slides, and/or the corresponding chapter of "Multi-agent systems" (Y. Shoham and K. Leyton-Brown).
Here are six:

1. **Auto-compatible**. Approximate Pareto-optimality in self-play.
2. **Safety**. At least earn the maxmin (security value).
3. **Targeted optimality**. Best response against a limited class of opponents.
4. **Efficient targeted learning**. For every $\epsilon > 0$ and $0 < \delta < 1$, there exists an $M$ polynomial in $1/\epsilon$ and $1/\delta$, such that after $M$ steps, with probability $\geq 1-\delta$, (a), (b) and (c) are achieved within $\epsilon$.
5. **Rational**. Approximate a best response if the opponents settle on stationary strategies.
6. **No regret**. At any point, earn no less than any pure strategy would have.

2, p. 1: If an action is chosen often with cumulative payoff matching (CPM), then the choice for that action is reinforced more strongly than with average payoff matching (APM). This is because the accumulated payoff is not divided by the number of times that action was chosen. Example: two actions $a_1$ and $a_2$ with constant payoffs 1 and 2, respectively. Let CPM start with initial propensities $\theta_1 = \theta_2 = 1$ and let APM start with initial averages $\mu_1 = \mu_2 = 1$. With CPM a typical run would be

| $t$: | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | … |
|---|---|---|---|---|---|---|---|---|---|---|
| $e_1^t$: | | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | … |
| $e_2^t$: | | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | … |
| $\theta_1$: | 1 | 2 | 3 | | 4 | | | | | … |
| $\theta_2$: | 1 | | | 3 | | 5 | 7 | 9 | 11 | … |

With APM a typical run would be

| $t$: | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | … |
|---|---|---|---|---|---|---|---|---|---|---|
| $e_1^t$: | | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | … |
| $e_2^t$: | | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | … |
| $\mu_1$: | 1 | 1 | 1 | | | 1 | | | | … |
| $\mu_2$: | 1 | | | 1.5 | 1.67 | | 1.75 | 1.8 | 1.83 | … |

At the outset, the averages are set to 1. Values like 0 or "undefined" are forbidden because APM must work right away. The average of $a_1$ remains of course 1 and the average of $a_2$ grows towards 2, every time $a_2$ is chosen, according to the sequence 1, $(1+2)/2$, $(1+2+2)/3$, $(1+2+2+2)/4$, etc.
The examples demonstrate (but of course do not prove) that the dynamics in CPM differs from the dynamics in APM. More specifically, CPM tends to stick to sub-optimal actions too long at the outset, because it accumulates its payoffs greedily without averaging them. For the same reason, however, CPM tends to avoid sub-optimal actions in the long run. Look at the development of the relative propensities with CPM: these tend to press out $a_1$ in the limit. On the other hand, APM leaves more room for $a_1$ in the long run, viz. one out of three in the limit ($\rightsquigarrow 1$ for $a_1$; $\rightsquigarrow 2$ for $a_2$).

3, p. 1: See the slides on no-regret and/or Ch. 2 of "Strategic Learning an its Limits (H. Peyton Young, 2004).

| Round nr.: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| experiment: | $A$ | $B$ | – | – | $A$ | – | $B$ | – |
| row: | $A$ | $B$ | $B$ | $B$ | $A$ | $B$ | $B$ | $B$ |
| column: | $A$ | $A$ | $B$ | $B$ | $B$ | $A$ | $B$ | $B$ |
| $u_t$: | 0 | 2 | 4 | 4 | −5 | 2 | 4 | 4 |
| $\bar{u}^t$: | 0 | 1 | 2 | 2.5 | 1 | 1.17 | 1.57 | 1.88 |
| $\bar{u}_A^t(E)$: | 0 | 0 | 0 | 0 | −2.5 | −2.5 | −2.5 | −2.5 |
| $\bar{u}_B^t(E)$: | – | 2 | 2 | 2 | 2 | 2 | 3 | 3 |
| $\bar{r}_A^t$: | 0 | −1 | −2 | −2.5 | −3.5 | −3.67 | −4.07 | −4.38 |
| $\bar{r}_B^t$: | – | 1 | 0 | −0.5 | 1 | 0.83 | 1.43 | 1.12 |

4, p. 1: See the slides on fictitious play and/or the corresponding chapter of "Strategic Learning an its Limits (H. Peyton Young, 2004).

- *Formula*. Let $x_i^1, \ldots, x_i^n$ the actions that are at the disposal of player $i$. Let $y_{-i}$ player $i$'s counterprofile of empirical frequencies of play. Let $u_k = u_i(x_i^k, y_{-i})$ player $i$'s utility for playing action $x_i^k$ against counterprofile $y_{-i}$. Let $\gamma > 0$ be the smoothing parameter. Let $p_k$ the probability of playing action $x_i^k$. Then
$$p_k = \frac{e^{u_k/\gamma}}{\sum_{j=1}^n e^{u_j/\gamma}}.$$
- *Relation with fictitious play*. $\gamma \downarrow 0$ approaches fictitious play.
- *Relation with no-regret*. For every $\epsilon > 0$ and sufficiently small $\gamma$, regrets are bounded above by $\epsilon$ a.s.
- *Convergence properties*. For every $\epsilon > 0$ and sufficiently small $\gamma$, the empirical frequencies of play converge to the set of coarse correlated $\epsilon$-equilibria a.s.

5, p. 1: This game has two pure equilibria $(1,0)$, $(0,1)$ and one mixed equilibrium $(1/4, 1/4)$. The corresponding mixed strategies are $p = (1/4, 3/4)$ and $q = (1/4, 3/4)$. (See the game theory slides on how to determine mixed equilibria.) The mixed equilibrium is also symmetric. Only symmetric equilibria are candidates for ESSs, so $(p, q)$ is the only equilibrium to consider.
Since $p = (1/4, 3/4)$ is fully mixed, every response $q$ is a best response to $p$ (again, see the game theory slides to see why the latter is true):

$$\text{for all } q : q^T A p \geq p^T A p. \text{ In particular for all } q : q^T A p = p^T A p.$$

So the first condition of an ESS is violated. We'll have to verify the second condition of an ESS:

$$\text{for all } q \neq p : q^T A q < p^T A q.$$

Let $q = (y, 1-y)$, $y \neq 1/4$, be arbitrary. Then

$$p^T A q = \begin{pmatrix} 1/4 & 3/4 \end{pmatrix} \begin{pmatrix} -2 & 5 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} y \\ 1-y \end{pmatrix} = 4\frac{1}{4} - 4y,$$
$$q^T A q = \begin{pmatrix} y & 1-y \end{pmatrix} \begin{pmatrix} -2 & 5 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} y \\ 1-y \end{pmatrix} = 4 - 2y - 4y^2.$$

It is easy to verify that $q \neq p \;\Rightarrow\; q^T A q < p^T q$, because

$$\begin{aligned} p^T A q - q^T A q &= \left(4\frac{1}{4} - 4y\right) - \left(4 - 2y - 4y^2\right) \\ &= 4y^2 - 2y + \frac{1}{4} \\ &= \frac{1}{4}(4y-1)^2, \end{aligned}$$

which is positive on $[0,1]\backslash\{\frac{1}{4}\}$. So the second condition of an ESS is met. It follows that $p$ is an equilibrium that corresponds to an ESS.

Last modified on Wednesday 20th May, 2020, 16:50 uur.