Multi-agent learning 2017-18, end 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. Given

	L	R
T	$2,-2$	$1,-1$
B	$-1,2$	$2,-3$

.

Suppose IGA-WoLF is played on the strategy profile $(\alpha,\beta) = (1/2, 1/2)$ with learning rate $\eta = 0.1$ and $(l_{\min}, l_{\max}) = (1,2)$. Determine the strategy profile after one iteration.

2. Suppose two agents use Bayesian learning to guide their play in the repeated coordination game with two actions C and D. Agent 1 gives equal priors to the following response rules:

> all-C, all-D, TFT, grim trigger (cooperate, but defect forever when opponent defects), mix $C = 1/2$, mix $C = 2/3$, fictitious play.

Compute Agent 1's probability distribution over Agent 2's response rules after history CC, DC, CD, CD, DC. (I.e, in action profile DC, Agent 1 plays D and Agent 2 plays C.) Which action should Agent 1 select in Round 6 and why?

3. Describe P^0 and P^ϵ of H. Peyton-Young's take on Th.C. Schelling's model of segregation. Explain that, from any state to any state which is completely segregated, there exists a sequence of transitions, each with zero or low resistance.

4. Describe hypothesis testing, regret testing, their commonalities, and their differences.

5. Show that there exist uncoupled 2-recall response rules that guarantee almost sure convergence of play to pure Nash equilibria of the stage game in every game where such equilibria exist.

End of document.

Answers

1, p. 1: If

$$M = \begin{array}{c} \\ \mathrm{T} \\ \mathrm{B} \end{array} \begin{array}{c} \begin{array}{cc} \mathrm{L} & \mathrm{R} \end{array} \\ \left(\begin{array}{cc} r_{11}, c_{11} & r_{12}, c_{12} \\ r_{21}, c_{21} & r_{22}, c_{22} \end{array} \right) \end{array}$$

then

$$\begin{cases} u \ = (r_{11} - r_{12}) - (r_{21} - r_{22}) = \ \ 4 \\ u' = (c_{11} - c_{21}) - (c_{12} - c_{22}) = -6 \end{cases}$$

so

$$\begin{aligned} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_{t+1} &= \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_t + \eta \begin{pmatrix} l_l \\ l_2 \end{pmatrix} \left(\begin{bmatrix} 0 & u \\ u' & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_t + \begin{bmatrix} r_{12} - r_{22} \\ c_{21} - c_{22} \end{bmatrix} \right) \\ &= \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_t + 0.1 \begin{pmatrix} l_l \\ l_2 \end{pmatrix} \left(\begin{bmatrix} 0 & 4 \\ -6 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_t + \begin{bmatrix} -1 \\ 5 \end{bmatrix} \right) \\ &= \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_t + 0.1 \begin{bmatrix} l_l(\ \ 4\,\beta - 1) \\ l_2(-6\,\alpha + 5) \end{bmatrix}. \end{aligned}$$

The dynamics is stationary when the gradient is zero, so when

$$0.1 \begin{bmatrix} l_l(\ \ 4\,\beta - 1) \\ l_2(-6\,\alpha + 5) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

that is when $(\alpha, \beta) = (5/6, 1/4)$. This point is in the unit square, so $(\alpha^*, \beta^*) = (5/6, 1/4)$ is a mixed Nash equilibrium.

The eigenvalues of

$$U = \begin{bmatrix} 0 & 4 \\ -6 & 0 \end{bmatrix}$$

are imaginary (solve $Ux = \lambda x$ for lambda, you will end up with $\lambda^2 + 24 = 0$) so the dynamic is centric in such a way that the strategy profiles circle around (α^*, β^*), which is the only Nash equilibrium. To determine whether the dynamics is clockwise or anti-clockwise, pick an arbitrary strategy profile $(\alpha, \beta) \neq (5/6, 1/4)$, let us say $(\alpha, \beta) = (0, 0)$. This profile happens to be on the lower left of the center (α^*, β^*), and the gradient at $(\alpha, \beta) = (0, 0)$ is

$$0.1 \begin{bmatrix} l_l \cdot -1 \\ l_2 \cdot \ \ 5 \end{bmatrix}$$

which regardless of l_1 and l_2 is up left (both l_1 and l_2 are positive, remember). So the dynamics is clockwise.

The question was about the strategy profile $(\alpha, \beta) = (1/2, 1/2)$. This point is to the upper left of the center (α^*, β^*). With clockwise movement this means that the first player (the owner of α) moves towards the center and therefore is losing, while the second player moves away from the center and therefore is winning. Following the WoLF principle (win or learn fast), $l_1 = l_{\max} = 2$, $l_2 = l_{\min} = 1$, and the exact dynamics at $(\alpha, \beta) = (1/2, 1/2)$ is

$$0.1 \begin{bmatrix} l_l(\ \ 4\,\beta - 1) \\ l_2(-6\,\alpha + 5) \end{bmatrix} = 0.1 \begin{bmatrix} 2(\ \ 4\,\beta - 1) \\ 1(-6\,\alpha + 5) \end{bmatrix}.$$

It follows that the strategy profile after one iteration is

$$\begin{aligned} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_{t+1} &= \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_t + 0.1 \begin{bmatrix} l_l(\ \ 4\,\beta - 1) \\ l_2(-6\,\alpha + 5) \end{bmatrix} \\ &= \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_t + 0.1 \begin{bmatrix} 2(\ \ 4\,\beta - 1) \\ 1(-6\,\alpha + 5) \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} + 0.1 \begin{bmatrix} 2(\ \ 2 - 1) \\ 1(-3 + 5) \end{bmatrix} \\ &= \begin{bmatrix} 7/10 \\ 7/10 \end{bmatrix} = \begin{bmatrix} 0.70 \\ 0.70 \end{bmatrix}. \end{aligned}$$

2, p. 1: This question can actually be answered quickly by ascertaining that all strategies but mix C = 1/2 and mix C = 2/3 fall off eventually. We can then compute the product probabilities for both remaining strategies, and normalise (without dividing by 7, since the prior is uniformly distributed).

Elaborated answer:

$P(h|s)$:

	all-C	all-D	TFT	GT	C=1/2	C=2/3	FP
CC	1	0	1	1	1/2	2/3	1
DC	1	0	1	1	1/4	4/9	1
CD	0	0	1	1	1/8	4/27	1/2
CD	0	0	0	1	1/16	4/81	0
DC	0	0	0	0	1/32	8/243	0

$P(h|s)P(s)$ (divide everything by 7):

	all-C	all-D	TFT	GT	C=1/2	C=2/3	FP
CC	1/7	0	1/7	1/7	1/14	2/21	1/7
DC	1/7	0	1/7	1/7	1/28	4/63	1/7
CD	0	0	1/7	1/7	1/56	4/189	1/14
CD	0	0	0	1/7	1/112	4/567	0
DC	0	0	0	0	1/224	8/1701	0

In decimals:

	all-C	all-D	TFT	GT	C=1/2	C=2/3	FP
CC	0.14	0	0.14	0.14	0.0714	0.0952	0.14
DC	0.14	0	0.14	0.14	0.0357	0.0634	0.14
CD	0	0	0.14	0.14	0.1786	0.0211	0.0714
CD	0	0	0	0.14	0.0089	0.0071	0
DC	0	0	0	0	0.0046	0.0047	0

Normalise last row (since that was the question):

	all-C	all-D	TFT	GT	C=1/2	C=2/3	FP
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
DC	0	0	0	0	0.4946	0.5054	0

In Round 6, Player 1 should play C because response rule C = 2/3 is more likely than response rule C = 1/2.

3, p. 1: P^0 is a Markov process with states all configurations of Schelling's one-dimensional model of segregation, and with transitions all beneficial trades of location. P^ϵ is like P^0 but allows all possible trades of action, including neutral and disadvantageous ones, where the likelihood of such trades is exponentially small and ordered by their negative profit.
A completely segregated state is an absorbing state in P^0 in which all elements of the first type are lined up on one side of the circle and all the elements of the other type on the other. A dispersed segregated state is an absorbing state that is not completely segregated. (Was not asked, but these definitions are repeated here for the sake of convenience.)
Suppose x is an arbitrary state. If x is completely segregated, then s can be reached through a sequence of transitions of low resistance by repeatedly putting the element at the head of the group of the first kind to the tail of that group. If x is dispersed, clusters of the first kind can be merged by a sequence of transitions of low resistance by repeatedly putting the element at the head of any group of the first kind to the tail of that same group. If a single player of the first kind remains between players of the other kind, this player can trade with the first player in his group, and this trade has zero resistance. By repeating this step, the number of groups of the first kind reduces until we arrive at a completely segregated state, from which we can reach s. Finally, if x is not absorbing, then a mutually advantageous swap with another state is possible such that the total discontent decreases. By repeating this step it is possible to travel with zero resistance from every state to an absorbing state.

4, p. 1: **Hypothesis testing**. Four parameters: smoothing γ, sample size s, tolerance level τ, and randomness, ρ, of new hypothesis. These parameters depend on the radius of the approximated equilibrium, ϵ.
Initialisation: select a hypothesis $p^{-i} \in \Delta_{-i}$ of opponent's play at random, and set mode to "not sampling".
Repeat:

1. Play a γ-smoothed reply to p^{-i}.
2. When not sampling, a new sample period of length s is started with probability $1/s$.
3. When sampling is over compare the empirical frequencies of play of the opponents with the hypothesis. If the difference exceeds the tolerance level, select a new hypothesis with probability ρ completely random and with probability $1-\rho$ equal to the opponent's empirical frequency of play.

Regret testing. Parameters: s and τ. These depend on the radius of the approximated equilibrium, ϵ.
Initialisation: select a (possibly mixed) strategy at random, $x^i \in \Delta_i$.
Repeat:

1. Play x for s rounds, experimenting ϵ of the time.
2. If s's revenue during the past s rounds plus τ does not supersede the yield of every individual action during the past s rounds, then adopt a new random strategy x.

Commonalities. In self-play, both algorithms learn to play ϵ-equilibria $1-\epsilon$ of the time, provided the parameters are sufficiently tightened.

Differences. Hypothesis testing maintain a hypothesis of opponent's play, regret testing maintains a hypothesis of best response. Hypothesis testing plays a smoothed reply, regret testing plays a fixed mixed strategy. Hypothesis testing starts sampling at random times to avoid coordination, regret testing samples continuously. Etc.

5, p. 1: We must give a 2-recall algorithm that converges to pure Nash equilibria of the stage game in every game where such equilibria exist.
Let players maintain the last two action profiles: (a_{-2}, a_{-1}). Take as learning algorithm: persist at best replies when $a_{-2} = a_{-1}$, randomise else. Now there are four regions (not classes) in the Markov chain:

1. $S_1 =_{Def} \{(a\ , a) \in A^2 \mid a \text{ is Nash}\}$.
2. $S_2 =_{Def} \{(a', a) \in A^2 \mid a \text{ is Nash}\}$, $a' \neq a$.
3. $S_3 =_{Def} \{(a', a) \in A^2 \mid a \text{ is not Nash}\}$, $a' \neq a$.
4. $S_4 =_{Def} \{(a\ , a) \in A^2 \mid a \text{ is not Nash}\}$.

Clearly, all states in S_1 are absorbing. (Check!) Further, all other states are transient: there is a positive probability of reaching a state in S_1 in finitely many periods. Indeed:

- At each state (a', a) in S_2 all players randomize; hence there is a positive probability that next period they will play a—and so the next state will be (a, a), which belongs to S_1.
- At each state (a', a) in S_3 all players randomize; hence there is a positive probability that next period they will play a pure Nash equilibrium $\bar{a}$ (which exists by assumption)—and so the next state will be $(a, \bar{a})$, which belongs to S_2.
- At each state (a, a) in S_4 at least one player is not best-replying and thus is randomizing; hence there is a positive probability that the next period play will be some $a' \neq a$—and so the next state will be (a, a'), which belongs to $S_2 \cup S_3$.

In all cases there is thus a positive probability of reaching an absorbing state in S_1 in at most three steps. Once such a state (a, a), where a is a pure Nash equilibrium, is reached (this happens eventually with probability one), the players will continue to play a every period.

Last modified on 29-6-2018.