Multi-agent learning 2016-17, 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	$1,0$	$-1,2$
B	$0,1$	$1,0$

   .

   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. Each considers the set of possible response rules to be the set of all 1-recall response rules. (So this gives 32 mappings from outcomes in the previous turn to actions plus an initial move. For instance, $CC \to D$, $CD \to D$, $DC \to D$, $DD \to D$, initial move D would be the "always defect" strategy.)

   Assume every player starts with a uniform prior over all 32 strategies. (E.g., Pr("always defect") = 1/32.) Use Bayesian updating to calculate the resulting beliefs of a player after observing the following play: CC, CD, DD, DC. Calculate the best response rule for this player to adopt given these new beliefs.

3. Consider a technology choice game in which there are three choices A, B and C and the payoffs are

$$\begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 10 & 1 & 2 \\ 1 & 6 & 3 \\ 2 & 3 & 8 \end{pmatrix} \end{array}.$$

   Each year, players choose a technology by sampling and playing a best reply almost always. Compute the stochastically stable states of this process for large enough samples.

4. We consider learning by naive hypothesis testing (Foster and Young).

   (a) Which four parameters are involved? Explain their use. Explain (in qualitative terms (i.e., without the use of numbers) what conditions these parameters must fulfill, and in which order should they must be chosen on order to ensure converge to ϵ-Nash equilibria $1-\epsilon$ of the time a.s.

   (b) – For fixed $\epsilon > 0$ and appropriate parameters, the responses constitute an ϵ-Nash $1-\epsilon$ of the time a.s.
   – If ϵ is tightened sufficiently slowly, the responses converge in probability to the Nash equilibria of the repeated game.

   Describe the convergence behaviour for both cases, i.e., for fixed ϵ and for $\epsilon \downarrow 0$. Indicate what happens if the stage game possesses multiple equilibria. (Hint: the difference in behaviour amongst both cases is not that large.)

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.

Answers

1, p. 1: If

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

then

$$\begin{cases} u &= (r_{11} - r_{12}) - (r_{21} - r_{22}) = \ \ 3 \\ u' &= (c_{11} - c_{21}) - (c_{12} - c_{22}) = -3 \end{cases}$$

so

$$\begin{aligned} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_{t+1} &= \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_t + \eta \begin{pmatrix} l_l \\ l_2 \end{pmatrix} \left(\begin{bmatrix} 0 & u \\ u' & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_t + \begin{bmatrix} r_{12} - r_{22} \\ c_{21} - c_{22} \end{bmatrix} \right) \\ &= \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_t + 0.1 \begin{pmatrix} l_l \\ l_2 \end{pmatrix} \left(\begin{bmatrix} 0 & 3 \\ -3 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_t + \begin{bmatrix} -2 \\ 1 \end{bmatrix} \right) \\ &= \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_t + 0.1 \begin{bmatrix} l_l(\ \ 3\beta - 2) \\ l_2(-3\alpha + 1) \end{bmatrix}. \end{aligned}$$

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

$$0.1 \begin{bmatrix} l_l(\ \ 3\beta - 2) \\ l_2(-3\alpha + 1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

that is when $(\alpha, \beta) = (1/3, 2/3)$. This point is in the unit square, so $(\alpha^*, \beta^*) = (1/3, 2/3)$ is a mixed Nash equilibrium.
The eigenvalues of

$$U = \begin{bmatrix} 0 & 3 \\ -3 & 0 \end{bmatrix}$$

are imaginary (solve $Ux = \lambda x$ for lambda, you will end up with $\lambda^2 + 9 = 0$) 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 $(\alpha, \beta) \neq (1/3, 2/3)$, 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

$$0.1 \begin{bmatrix} l_l \cdot -2 \\ l_2 \cdot \ \ 1 \end{bmatrix}$$

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 lower right of the center (α^*, β^*). (By now, you will have discovered that a sketch will help.) 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

$$0.1 \begin{bmatrix} l_l(\ \ 3\beta - 2) \\ l_2(-3\alpha + 1) \end{bmatrix} = 0.1 \begin{bmatrix} 2(\ \ 3\beta - 2) \\ 1(-3\alpha + 1) \end{bmatrix}.$$

It follows that the strategy profile after one iteration is

$$\begin{aligned} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_{t+1} &= \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_t + 0.1 \begin{bmatrix} l_l(\ \ 3\beta - 2) \\ l_2(-3\alpha + 1) \end{bmatrix} \\ &= \begin{bmatrix} \alpha \\ \beta \end{bmatrix}_t + 0.1 \begin{bmatrix} 2(\ \ 3\beta - 2) \\ 1(-3\alpha + 1) \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} + 0.1 \begin{bmatrix} 2(\ \ 3\frac{1}{2} - 2) \\ 1(-3\frac{1}{2} + 1) \end{bmatrix} \\ &= \begin{bmatrix} 2/5 \\ 9/20 \end{bmatrix} = \begin{bmatrix} 0.40 \\ 0.45 \end{bmatrix}. \end{aligned}$$

2, p. 1: After observing this history, our player will only consider two opponent strategies possible: opponent: CC → D; CD → D; DC → C; DD → C, initial move C or CC → D; CD → D; DC → D; DD → C, initial move C, each with probability 1/2. All other possible strategies are inconsistent with this play and hence will get zero weight.
A best response rule for our player at this point would be to echo his opponents moves insofar they are known. Further, DC → C and DC → D are equally likely, given the history of play. We end up with the set

$$\{\ \text{CC} \to \text{D};\ \text{CD} \to \text{D};\ \text{DC} \to \text{p};\ \text{DD} \to \text{C},\ \text{initial move C} \mid 0 \le p \le 1\ \}.$$

Any element of this set would yield the maximum possible payoffs for either of the two possible opponent strategies.

3, p. 1: First we identify the recurrence classes of P^0, where P^0 is the process in which players always play best responses.
One such class is the absorbing state in which almost everyone[1] plays A. Call this class z^A. Similarly, we have z^B and z^C. It can be checked that these are the only recurrence classes: from any state the probability is one of eventually landing in one of these three classes.
Now compute the "path of least resistance" amongst all recurrence classes. There are $3(3-1)=6$ such paths to be computed. Let us start with the transition from z^A to z^B. Starting from z^A we need to know how many adoptions (due to perturbations) of B are needed that lead to a critical or "tipping" state z^*, from which the process can transit to z^B with no further perturbations. If $a=|A|$, $b=|B|$ and $c=|C|$, this tipping point occurs when

$$\begin{cases} a+6b+3c \ge 10a+b+2c, \\ a+b+c=n. \end{cases} \tag{1}$$

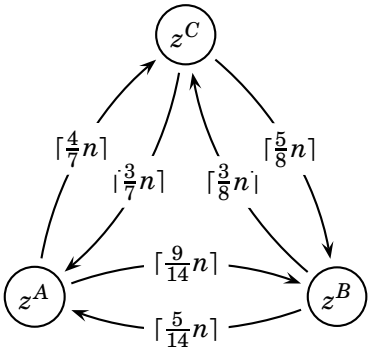
where n is the sample size. An agent who draws such a combination of individuals in his sample will choose B instead of A. Since the shortest path from z^A to z^B does not involve individuals playing C, Eq. (1) reduces to

$$\begin{cases} a+6b \ge 10a+b, \\ a+b=n. \end{cases}$$

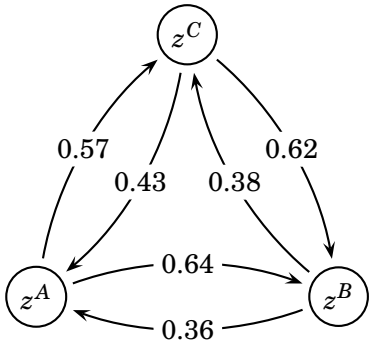
Solving for a and b yields $b \ge 9/14n$. So at least $b \ge \lceil 9/14n \rceil$ "erratic" transitions are needed to bring the process into state z^*. The probability of such an event is $(\epsilon/2)^{\lceil 9/14n \rceil}$, so the resistance from z^A to z^B, written $r(z^A \to z^B)$, is $\lceil 9/14n \rceil$. Similarly:

$$r(z^B \to z^C)=\lceil \tfrac{3}{8}n \rceil, \quad r(z^C \to z^A)=\lceil \tfrac{3}{7}n \rceil,$$
$$r(z^B \to z^A)=\lceil \tfrac{5}{14}n \rceil, \quad r(z^C \to z^B)=\lceil \tfrac{5}{8}n \rceil, \quad r(z^A \to z^C)=\lceil \tfrac{4}{7}n \rceil.$$

(The resistances back and forth add up to one. This is specific for the scenario and is not always the case.)
It is convenient to draw a complete di-graph with links back and forth between all recurrence classes.[2]

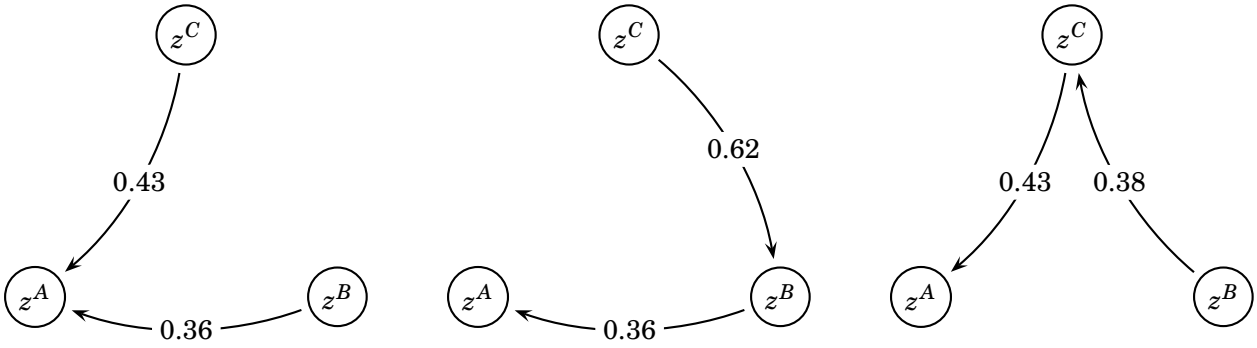


For large n, the effect of rounding disappears, so that we get (in decimals, and proportionally):



Now examine all the nine state trees to determine the recurrence classes with minimum stochastic potential. Each class had three state trees. The state trees for z^A, for example, are

[1] $N - \lfloor n/2 \rfloor$ to be precise, where N is the population size, and n is the sample size.
[2] Diagrams made with PSTricks.



These trees have resistance $0.36+0.43=0.79$, $0.36+0.62=0.98$, and $0.38+0.43=0.81$, respectively. The stochastic potential of a recurrence class is defined to be the minimum resistance over all trees rooted at that recurrence class, so the stochastic potential of z^A equals 0.79. With z^B the resistances are 1.07, 1.26 and 1.19, which gives a stochastic potential of 1.07. With z^C the resistances are 0.95, 0.93 and 1.02, which gives a stochastic potential of 0.93.
It follows that z^A is the class with minimal stochastic potential. Hence the set of stochastically stable states equals $\mathscr{S}=\{z^A\}$.

Remarks. (These are to further interpret the answer, and do not belong to the answer proper.)

1. z^A is also Pareto-optimal, with a common payoff of 10. It is not always that case that stochastically stable states are Pareto-optimal. On the contrary, this state of affairs is quite exceptional, and hinges on the form of the payoff matrix (cf. key publication, beginning of Sec. 5).

2. This game has three pure equilibria, viz. $(1,0,0)\to 10$, $(0,1,0)\to 6$ and $(0,0,1)\to 8$, and four mixed equilibria, viz. $(5/14,9/14,0)\to 43/14$, $(7/24,23/48,11/48)\to 3+41/48$, $(3/7,0,4/7)\to 53/7$ and $(0,5/8,3/8)\to 47/8$.[3] One of these equilibria is z^A. So z^A is Nash and Pareto-optimal.

3. The sample size, n, must be large enough. If not, the ceiling function may top of resistances such that some of them become equal. This already happens for $n=48$ ($\frac{5}{14}\approx\frac{3}{8}$):

$$(\lceil 48\frac{9}{14}\rceil,\lceil 48\frac{5}{14}\rceil,\lceil 48\frac{3}{8}\rceil,\lceil 48\frac{5}{8}\rceil,\lceil 48\frac{3}{7}\rceil,\lceil 48\frac{4}{7}\rceil)=(31,18,18,30,21,28).$$

   For small n a consequence is that more than one class has minimum stochastic potential.

4a, p. 1: – Smoothing γ. Responses are smoothed, so every action is played with positive probability. Must be close to zero.

– Tolerance level τ, depends on γ, hence ϵ. If the difference between the hypothesis and what is observed is larger than τ, then reject the hypothesis and adopt a new one. The tolerance must be small, in fact exponentially small in comparison to ϵ.

– Sample size s, depends on γ and τ, hence ϵ. This is the duration of sampling *and* the reciprocal of the probability to start a new sampling period (i.e., $1/s$). The sample size must be large, in fact exponentially large in comparison to ϵ.

– The probability to select a new hypothesis at random ρ. This parameter does not depend on the other three parameters. The parameter ρ must be non-zero.

4b, p. 1: – If $\epsilon>0$ is fixed, there is a small but positive probability that players make Type I errors or do not choose best responses given their hypotheses with the result that they start exploring again. If there are multiple equilibria, they may at some time converge at another equilibrium. It follows that every equilibrium is visited infinitely often. (But not necessarily with equal frequency. So it is not possible to enforce ϵ-convergence to one particular equilibrium if there are multiple equilibria.)

[3] Found with Netlogo doing a system call to Gambit's `gambit-enummixed`. Cf. `www.gambit-project.org`. Gambit returns raw (nearly unreadable) results; Netlogo is used, then, to interpret and present these results.

– If $\epsilon \downarrow 0$, same as previous answer, so convergence to the *set* of Nash equilibria (rather than pointwise to one of the equilibria), except that players players do not make Type I errors, for such eventualities are pressed out by $\epsilon \downarrow 0$. Still, it remains possible that players select sub-optimal responses. It is true that $\gamma \downarrow 0$ but the probabilities of the responses still must "add up" (in fact, integrate) to 1.

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 27-6-2017.