Name: .... Student card number: ....

Hand in this sheet only.

Rules

- ID required.
- You are not allowed to leave the exam room during the first 30 minutes.
- Scratch paper is handed out. You cannot use your own. It is possible to request additional scratch paper from the invigilator.

  The use of markers is not permitted.
- If you want to go to the toilet, raise your finger to warn a security guard. He or she will give you permission to go and walk with you to the toilet. Toilet visits are not permitted during the first and last half hour of the exam. You may only visit the toilet once.

  It is forbidden to take a telephone or similar electronic devices to the toilet.
- After you have left the examination room, you are not allowed to stay in the corridors / hall immediately outside due to noise. You follow the instructions of the invigilator.

Instructions

- There are open questions and multiple-choice questions.
- Every multiple-choice question has exactly one correct answer. In some cases, other answers may be "almost correct " or "partly correct". In such cases the best answer applies.

  Answer in the appropriate boxes by placing a cross. If you make a mistake, scratch the cross and put a cross in another box.

  Each correctly answered multiple-choice item yields one point.
- Answers to open questions are entered in the boxes (open rectangles)

  First draft your answer. Then fill the box.

  Each correctly answered open item yields two points, unless indicated otherwise.
- Because there are different versions of the exam, the order of the multiple-choice questions does not always correspond with the order of the material as discussed in the lectures.
- It is possible to request a new answer sheet as well as additional scratch paper from the invigilator. Our stock of answer sheets is finite, first come first serve.

  Good luck!

Multiple-choice answers

	A	*B*	*C*	*D*
1.				
2.				
3.				
4.				

	A	*B*	*C*	*D*
5.				
6.				
7.				
8.				

	A	*B*	*C*	*D*
9.				
10.				
11.				
12.				

Open questions—first write your answer in draft elsewhere, then fill the boxes here

1. (Regret matching.) Row is a 10% noise no-regret player. So row plays a random action 10% of the time at random moments. A normal form game is played with payoffs

	L	C	R
T	$3,2$	$1,1$	$2,0$
M	$4,1$	$0,3$	$6,2$
B	$0,1$	$4,3$	$9,1$

. Compute the strategy of Row at round eight if $h =$ TC, BC, TL, BL, TR, TR, TL.

Explanation. We will have to compute actual and hypothetical payoffs.

— Row's actual accumulated payoff: $1 + 4 + 3 + 0 + 2 + 2 + 3 = 15$.

Col's history: CCLLRRL, so 3xL, 2xC and 2xR.

— Row's hypothetical accumulated payoff for T: $3 \cdot 3 + 2 \cdot 1 + 2 \cdot 2 = 15$, regret $15 - 15 = 0$,
— Row's hypothetical accumulated payoff for M:$3 \cdot 4 + 2 \cdot 0 + 2 \cdot 6 = 24$, regret $24 - 15 = 9$,
— Row's hypothetical accumulated payoff for B: $3 \cdot 0 + 2 \cdot 4 + 2 \cdot 9 = 26$, regret $26 - 15 = 11$.

Total regret: $0 + 9 + 11 = 20$.

If Row would be a pure no-regret player, it would play T with probability $0/20$, M with probability $9/20$ and B with probability $11/20$. However, Row isn't a pure no-regret player, instead it randomizes 10% at the time. If R represents the event that Row does not randomize and plays according to regret matching, then

$$\begin{aligned}
\Pr\{T|h\} &= \Pr\{T|R,h\}\Pr\{R|h\} + \Pr\{T|\overline{R},h\}\Pr\{\overline{R}|h\} = \frac{0}{20}\frac{9}{10} + \frac{1}{3}\frac{1}{10} = \frac{1}{30} \approx 0.033,\\
\Pr\{M|h\} &= \Pr\{M|R,h\}\Pr\{R|h\} + \Pr\{M|\overline{R},h\}\Pr\{\overline{R}|h\} = \frac{9}{20}\frac{9}{10} + \frac{1}{3}\frac{1}{10} = \frac{263}{600} \approx 0.438,\\
\Pr\{B|h\} &= \Pr\{B|R,h\}\Pr\{R|h\} + \Pr\{B|\overline{R},h\}\Pr\{\overline{R}|h\} = \frac{11}{20}\frac{9}{10} + \frac{1}{3}\frac{1}{10} = \frac{317}{600} \approx 0.528.
\end{aligned}$$

2. (Reinforcement learning.) Suppose player P pulls levers A, B and C according to the Erev-Roth scheme. Compute P's strategy in round five if $\theta_0 = (8, 8, 8)$, $\lambda = 0.5$, and P's realisation of play is

round	1	2	3	4
A	12			
B		22		
C			41	1

. Sanity check: all propensities remain integers.

(With this question and the next, your answer must include a computation that leads to the outcome.)

Explanation. Development of propensities of play w.r.t. A, B and C:

round	0	1	2	3	4
θ_A	8	$8/2 + 12 = 16$	$16/2 = 8$	$8/2 = 4$	$4/2 = 2$
θ_B	8	$8/2 = 4$	$4/2 + 22 = 24$	$24/2 = 12$	$12/2 = 6$
θ_C	8	$8/2 = 4$	$4/2 = 2$	$2/2 + 41 = 42$	$42/2 + 1 = 22$

Total propensity: $2 + 6 + 22 = 30$. So

$$\Pr\{A \mid h\} = \frac{2}{30} = \frac{1}{15}, \quad \Pr\{B \mid h\} = \frac{6}{30} = \frac{3}{15}, \quad \Pr\{C \mid h\} = \frac{22}{30} = \frac{11}{15}.$$

N.B. Item 5 on page 4 suggests there is no decay for actions that are not selected. For Arthur's algorithm this is not true: also for actions not selected the propensities decay. If you used the suggestion of Item 5 in the computation of the answer here, and your computation is consistent, this is discounted positively in the grading.

Multiple-choice questions

1. (Introduction.) Targeted optimality against fictitious play would amount to playing

   $\surd$ Bully.
   (b) Fictitious play.
   (c) No-regret.
   (d) Q-learning.

> **Explanation.** Fictitious play is a pure follower algorithm. The best response against a pure follower is a pure teacher, which is the pure teacher algorithm Bully. The other algorithms listed do not exploit fictitious play as optimal as Bully does a.s

2. (The Borel-Cantelli lemma's.) Before the Primark opens at 10.00, a group of customers gather in front because they want to shop there. For one reason or another, customers cannot or do not want to line up properly.

   At 10.00 the doors open and will, for the sake of the greater good, remain open from then on.

   From 10.00 on, customers arrive one per minute. Every five minutes, a random customer is picked from the group to be given entrance. (Since they were unable to form a queue.) Selection does not depend on previous picks.

   At some time after 10.00 you join the group. Will you ever be allowed entrance?

   (a) Almost surely: apply BC1.
   $\surd$ Almost surely: apply BC2.
   (c) With probability < 1: apply BC1.
   (d) With probability < 1: apply BC2.

> **Explanation.** Inexact argument: the group grows linearly in time (effectively by 4 every five minutes), and so the sum of the probabilities that you will be picked grows at a harmonic rate, i.e. along the lines of
>
> $$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \dots \tag{1}$$
>
> Since all variants of the harmonic series diverge (i.e., time intervals and group sizes do not really matter), all tails of such series must diverge as well. Moreover, since selection does not depend on previous picks, BC2 applies, which means that, at every point in time, no matter how late you arrive, there is a later point in time at which you will be allowed entrance a.s.

3. (Multi-armed bandits.) After a next payoff is given, it is determined (i.e, there is no choice) which arm to pull next.

   (a) UCB and Q-learning.
   $\surd$ Only UCB.
   (c) Only Q-learning.
   $\surd$ None of the two.

> **Explanation.** With UCB, the arm to pull next is determined by the arm with the highest upper confidence bound, and after incorporating the last payoff, highest confidence bounds are fixed. With Q-learning, the arm to pull next is determined a probabilistic choice among alternatives weighed by estimations of arm revenues.
>
> Answer (d) is also correct because it may of course happen that multiple arms have largest UCB values. This is especially likely if payoffs are integer-valued.

4. (Multi-armed bandits.) With Exp3,

(a) probabilities are computed, then estimated rewards, then weights.

(b) probabilities are computed, then weights. then estimated rewards.

√ weights are computed, then probabilities, then estimated rewards.

(d) Another answer.

Explanation. Given r_i^t,

$$\begin{aligned} w_i^t &= w_i^{t-1} \exp\left(\gamma \frac{1}{K} \hat{r_i}^{t-1}\right), \\ p_i^t &= (1-\gamma)\frac{w_i^t}{\sum_{j=1}^k w_j^t} + \gamma \frac{1}{K}, \\ \hat{r}_i^t &= \begin{cases} \frac{r_i^t}{p_i^t} & \text{if } i \text{ is chosen at } t, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The computation of the estimated rewards must be at the end of the loop because these are used to decide which arm to pull next.

5. (Reinforcement learning.) Let θ_j^t be the value of arm j after round t, $1 \leq j \leq n$, and let arm i be pulled in round t with payoff p^t. Generally,

$$\theta_j^{t+1} = \begin{cases} \lambda\theta_j^t + \eta p^t & j = i, \\ \theta_j^t & \text{otherwise.} \end{cases}$$

Different values of the decay factor λ and the learning rate η yield different learning schemes. These include CPM: cumulative payoff matching; APM: average payoff matching; AR: Arthur; ER: Erev-Roth.

	λ	η	$\lambda+\eta$	
1.		$1/t$	1	
2.		$1/Ct^q$	1	$C, q > 0$
3.	1	1		
4.		1		$0 \leq \lambda \leq 1$

(a) 1 → APM; 2 → CPM; 3 → AR; 4 → ER.

√ 1 → APM; 2 → AR; 3 → CPM; 4 → ER.

(c) 1 → ER; 2 → APM; 3 → CPM; 4 → AR.

(d) 1 → ER; 2 → AR; 3 → APM; 4 → CPM.

Explanation. With APM, $\lambda = (t-1)/t$, so that θ_j^t is the average of the payoffs given by arm j. With CPM, no discounts are used. With Arthur, the learning rate is $1/Ct^q$. (We use q rather than p here to avoid confision with p^t.)

6. (Reinforcement learning.) Consider the following two games in which only payoffs for Col are displayed. Does L dominate R?

	L	R
T	5	3
B	3	4

	L	R
T	5	3
B	4	4

(a) Yes in 1st game, yes in 2nd game.

(b) Yes in 1st game, no in 2nd game.

(c) No in 1st game, yes in 2nd game.

√ No in 1st game, no in 2nd game.

Explanation. Action L dominates action R if there is a $\gamma > 1$ such that, for all histories of play h,

$$E[\,\pi(L) \mid h\,] \;>\; \gamma E[\,\pi(R) \mid h\,].$$

Since there are only finitely many actions in a normal form game, this amounts to

$$E[\,\pi(L) \mid h\,] \;>\; E[\,\pi(R) \mid h\,], \text{for all histories of play } h.$$

First matrix: if Row plays B throughout, Col receives $3 < 4 = \pi$(R). So L does not dominate R. Second matrix: if Row plays B throughout, Col receives $4 = \pi$(R). Domination must be strict, so L does not dominate R.

7. (Equilibria.) For which of the following games it holds that CE = CCE?

i) For the chicken game.

ii) For the fashion game (a.k.a. leader follower game).

	Red	Yellow	Blue
Red	$(1,0)$	$(0,0)$	$(0,1)$
Yellow	$(0,1)$	$(1,0)$	$(0,0)$
Blue	$(0,0)$	$(0,1)$	$(1,0)$

(a) *i)* and *ii)*.

√ Only *i)*.

(c) Only *ii)*.

(d) None.

Explanation. The chicken game is a 2x2 game. For all 2x2 games, CE = CCE. The fashion game is an example of a game where CE $\neq$ CCE. See slides “Equilibria” and PY, p. 34 ff. Of course, it is also possible to just write out the four inequalities that correspond te CE and CCE thus answering the question.

8. (Equilibria.) Regret matching leads players into a

(a) NE.

(b) CE.

√ CCE.

(d) Another answer.

Explanation. This is the result of Hart and Mas-Colell.

For people who study this exam as preparation for an upcoming exam, here is some extra explanation.

Until now, there are no natural MAL algorithms known that lead players into a NE. (Cf., e.g., “Introduction”). There are MAL algorithms that lead players into NE, however these are considered fabricated and violate reasonable constraints of autonomy and independence. The absence of natural MAL algorithms that lead players into a NE is probably due to the fact that players must be able to coordinate independent of their history of play which is considered an unreasonable assumption, at least within the framework of MAL.

There are algorithms that lead players into CE (such as internal regret matching a.k.a. conditional regret matching), however, these algorithms are more sophisticated than regret matching. Conditional regret matching was referred to in the lectures, but knowledge of its existence is not necessary for answering the question.

9. (Regret matching.) Negative regret

√ Exists.

(b) Exists, and converges to zero in the long run.

(c) Exists, and disappears in the long run.

(d) Does not exist.

Explanation. Negative regret exists: it indicates that an action played was played not too often in the past for good reason. See also the slide "Regret matching in Shapley's game" with the Netlogo plots of regret. On that slide you see that most plots are below the x-axis most of the time.

10. (Regret matching.)

i) Average RM amounts to the same thing as cumulative RM.

ii) Average RM halts when all regrets are non-positive.

(a) *i)* and *ii)*.

✓ Only *i)*.

(c) Only *ii)*.

(d) None.

Explanation. *i)* For proportional regret matching

$$q_x^{t+1} =_{Def} \frac{[\bar{r}_x^t]_+}{\sum_{x' \in X} [\bar{r}_{x'}^t]_+} = \frac{[r_x^t]_+}{\sum_{x' \in X} [r_{x'}^t]_+}.$$

(Just multiply numerator and denominator by t.)

ii) When all regrets are non-positive, an action is selected randomly and RM proceeds.

11. (Fictitious play.) Let $\epsilon > 0$ be small and fixed. For Smoothed FP to converge to the set of ϵ-CCE, it must

(a) play a best response as often as possible.

(b) play a best reply as often as possible.

(c) play a better reply as often as possible.

✓ randomize to a sufficient extent.

Explanation. For smoothed fictitious play to converge to the set of ϵ-CCE, the smoothing parameters γ_i must be sufficiently small. With soft max [a.k.a. mixed logit a.k.a. quantal response] this means that sub-optimal actions are played relatively often. Playing a best reply (which is the same as playing a best response) is not in order here.

The notion of "better reply" does not apply to SFP. It applies, among others, to exponentiated regret matching.

12. Which of the following statements are true?

i) Smoothed fictitious play depends on own past payoffs.

ii) Exponentiated regret matching depends on past play of opponents.

(a) *i)* and *ii)*.

(b) Only *i)*.

(c) Only *ii)*.

✓ None.

Explanation. It is the other way around: smoothed fictitious play depends on past play of opponents (it is a so-called best reply strategy) and exponentiated regret matching depends on own past payoffs (it is a so-called better reply strategy). In the limit, when $\gamma \downarrow 0$, smoothed fictitious play limits regret but never to the full extent. And when $a \to \infty$, exponentiated regret matching approaches fictitious play but never exactly plays like it.

Added June 1, 2021, 14.15 on occasion of a remark in the Q&A session: *ii)* claims that the regret matching *algorithm* depends on opponent's behaviour. This is indeed incorrect, because all a regret matcher does is to compare its own realized payoffs with its own hypothetical payoffs. In doing so, a regret matcher is not at all interested in opponent's behaviour. On the other hand, it is of course (and fortunately!) true that the *outcome* of regret matching *does* depend on opponent's behaviour. Luckily so, otherwise regret matching would be a lousy algorithm. Nonetheless, when Peyton Young discussed regret matching in the best reply vs. better reply framework, he obviously referred to the *process* of regret matching (the algorithm) rather than the *outcome* of regret matching (the result).

End of multiple choice questions.