UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

# REINFORCEMENT LEARNING

**Saturday 1 st April 2017**

**00:00 to 00:00**

## INSTRUCTIONS TO CANDIDATES

**Answer QUESTION 1 and ONE other question.**

**Question 1 is COMPULSORY.**

**All questions carry equal weight.**

**CALCULATORS MAY BE USED IN THIS EXAMINATION**

**This is an OPEN BOOK examination.**
**You may consult notes, books or other documents during this examination.**

***THIS IS A MOCKUP EXAM! and will not be marked***

.

***THIS IS A MOCKUP EXAM! and will not be marked***

.

***THIS IS A MOCKUP EXAM! and will not be marked***

MSc Courses

Convener: ITO-Will-Determine
External Examiners: ITO-Will-Determine

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. **THIS QUESTION IS COMPULSORY.**

Bunny wakes up in a strange room with 2 doors; one on the left, and one on the right. In front of him is a map of the building, laying out clearly what lies behind each door. Behind one of the doors is a room with the way outside, and behind the other one a room with a hungry tiger. Bunny does not like tigers. Especially hungry ones.

(a) Consider the control problem where the current state is specified by the current room Bunny is in, and the actions Bunny can take are to move through a door to another room. Assume that there are 2 doors in the starting room, each leading to its own room, and that one of these other rooms (specifically the one on the left) has a single door that leads to the exit, while the other one (specifically the one on the right) has a tiger in it. Moreover, assume that Bunny has full knowledge of these details but that there is a 10% chance that when trying to move through one of the doors, he gets confused and goes through the other door instead. Lastly, assume that the episode does *not* end when Bunny enters the room with the tiger.

i. Formulate a Markov Decision Process (MDP) for the problem of controlling Bunny's actions in order to avoid the tiger and exit the building. (Give the transition and reward functions in tabular format, or give the transition graph with rewards). [*8 marks*]

ii. If instead of the way outside, the 'exits' just transported Bunny back to the starting room (or if you prefer, "moved him to another room with two doors; a room with an 'exit' on the left, and a room with a tiger on the right"), how would you modify the above MDP? (Similarly, "How would an MDP for this modified problem differ from the MDP for the above question?"). Would your answer change if we assumed that the episode ended when Bunny entered a room with a tiger? [*7 marks*]

(b) In the example at the beginning of this question, Bunny has access to a Transition and Reward function.

i. Considering a Reinforcement Learning algorithm in general, what is the property of *needing* these two functions as input called? [*2 marks*]

ii. What is the main idea behind *Reward Shaping*? What is the main problem with using it when formulating an MDP control problem? [*5 marks*]

iii. Explain the main differences between *Inverse Reinforcement Learning* and *Behavioural Cloning*. [*3 marks*]

2. **ANSWER EITHER THIS QUESTION OR QUESTION 3.**

Assume you are playing the following game for 2 players: There are $n$ coins on the table, and starting from Player A, the players take turns picking up 1 to $k$ coins from the table. Whoever picks up the *last coin loses* the game.

(a) i. Consider a version of the game where $n = 3$ and $k = 2$. It is Player A's turn. If you were to model this problem as an MDP for Player A, what would be the possible next states from Player A taking the action "pick up 1 coin"? [*3 marks*]

ii. Describe a way to learn the transition probability to either of these states, or, if you prefer, the whole transition function. [*3 marks*]

iii. Consider the multi-agent version of the problem. Compute the Minimax policy for Player A, when starting from the state as defined in the first part of this question. [*4 marks*]

(b) i. Give and explain the equation for the linear function approximation of a state-value function. [*2 marks*]

ii. (*Maybe some exercise about picking the features for a linear function approximator*). [*5 marks*]

(c) Given a fixed policy $\pi$ over some unkown MDP, consider the following observed trajectories (state, reward, next state etc.):

$$A, 1, A, 1, C$$
$$B, 1, A, -1, C$$

i. Use first-time visit Monte Carlo to evaluate $\pi$ at state $A$. [*3 marks*]

ii. Use every-time visit Monte Carlo to evaluate $\pi$ at state $A$. [*3 marks*]

iii. Compute the 2-step return for all states appearing in the second trajectory. [*2 marks*]

3. **ANSWER EITHER THIS QUESTION OR QUESTION 2.**

(a) i. What do we mean when we say that Temporal Difference Learning algorithms *bootstrap*? [*2 marks*]

ii. Which other category of Reinforcement Learning algorithms bootstraps? [*1 mark*]

(b) Consider the original problem with Bunny in Question 1, that is, the one with one room on the left with a real exit, and a tiger in the right room. However, assume that we do not know in which of the two rooms there is a tiger, and in which one there is an exit. Furthermore, assume that Bunny can try to listen in order to get a sense of which room contains the tiger. This technique is not perfect, and Bunny will sometimes mistakenly think the sound is coming from the wrong room.

Assuming there is a 20% chance of hearing the tiger in the wrong room (and an 80% chance of hearing it in the correct room):

i. Define the new set of states and actions for this new problem. Give the new transition function. (Assume that there is still a 10% chance that when trying to move through one of the doors, he gets confused and goes through the other door instead). [*7 marks*]

ii. Give the observation function for the POMDP as defined for this new problem. [*6 marks*]

iii. If at the start of the game Bunny assigns a 50/50 chance of the tiger being in either door, takes the action to listen and receives an observation that the tiger is in the left room, compute the posterior belief (similarly, "update our belief state according to the observation"). [*4 marks*]

(c) i. Explain the difference between continuing (non-episodic) and episodic Reinforcement Learning tasks. [*3 marks*]

ii. What is the role of the discount factor in either case? [*2 marks*]