FINAL

IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2017-2018

MEng Honours Degree in Mathematics and Computer Science Part IV
MEng Honours Degrees in Computing Part IV
MSc in Advanced Computing
MSc in Computing Science (Specialist)
for Internal Students of the Imperial College of Science, Technology and Medicine

This paper is also taken for the relevant examinations for the Associateship of the City and Guilds of London Institute

PAPER C424H

REINFORCEMENT LEARNING

Tuesday 12 December 2017, 10:00
Duration: 70 minutes

Answer TWO questions

Paper contains 3 questions
Calculators not required

1 Reinforcement gambling

a The card game "Quoker" is played as follows. 4 players have each a pot of credit chips $C \in \mathbb{N}_0^+$ and start with 100 credits each. A Quoker deck of playing cards consists of 52 Cards in each of the 4 suits of Spades, Hearts, Diamonds, and Clubs. Each suit contains 13 cards: Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King. At the start each player is dealt two cards. Thereafter, a round of betting ensues: Starting to the left of the card dealer and going round, each player has four options

- Raise – A player who thinks he has a good hand (or who wants the other players to think he has a good hand) may increase the wager required to continue playing.
- Fold – A player who thinks his hand is not good enough to win and who does not want to wager the increased amount may lay down his cards. He cannot win the hand, but he also will not lose any more chips.
- Call – Once a player has raised the stakes, each player must decide whether to raise the stakes again, to give in and fold his hand, or to call, which means to equal the amount wagered by the player who raised.
- Pass – If no one has increased the wager required to continue, a player may pass on his option to bet.

These rounds continue till all players have dropped out by folding, calling or passing. The winning player receives all the wagered credits. The winning player is determined by ranking the hand of each player against all other players using a comparison function $comp(\text{player } i \text{ hand, player } j \text{ hand}) = 1$, if player i hand wins (or the value is 0 if player j hand wins. You do not know how this comparison function operates (e.g. what patterns and value of cards are good or bad), but you can trust the function's outcome to be accurate when used to compare hands. The game is over for the agent when his credit is $C = 0$ or when all other agents have no more funds to play.

Draft a model-free Reinforcement Learning agent that can participate in the game. Explain concisely your design choices and in particular,

i) What are the state space dimensions and how many dimensions are these?

ii) What are action space dimensions and how many dimensions are these?

Note: If your dimensions reach $> 100,000$ states feel free to give approximate numbers (e.g. 12.1 Mio instead of $12,345,678$) and use fractions instead of decimals.

2 Reinforcement Learning

a Given an infinite sequence drawn from a Markov Reward Process with the immediate reward on any state being $r_t = 1$ for all t, and a geometric discount $\gamma = \frac{z-1}{z}$ for integer $z \in \mathbb{N}^+$ (e.g. $z = 4 \implies \gamma = \frac{3}{4}$). What is the total return for the trace $R(T)$ from any time T? Write out your derivation.

b Suppose a Q-learning agent, with fixed α, and discount γ, was in state S_{42} did action A_8, received reward 5 and ended up in state S_{31}. What value(s) get updated? Give explicit expression(s) for the new value(s).

c Explain concisely what the SARSA and Q-Learning algorithms are and discuss their differences.

The three parts carry, respectively, 30%, 35%, and 35% of the marks.

3 **More Reinforcement Learning**

a Consider the following reinforcement learning algorithms (A-F)

- A Q-learning with $\alpha_k = \frac{1}{k}$ and 100% exploitation.
- B Q-learning with fixed α and 75% exploitation.
- C Q-learning with fixed $\alpha_k = \frac{1}{k}$ and 75% exploitation.
- D SARSA learning with soft-max action selection.
- E SARSA-learning fixed $\alpha_k = \frac{1}{k}$ and 75% exploitation.
- F SARSA-learning fixed $\alpha_k = \frac{1}{k}$ and 100% exploitation.

i) Which of the above reinforcement algorithms will find the optimal policy, given enough time. Explain briefly why.

ii) Which ones will actually follow the optimal policy. Explain briefly why.

b What are the key differences between Dynamic Programming and Temporal Difference methods for Reinforcement learning. Explain briefly.

c What is the relationship between asynchronous value iteration and Q-learning? Explain briefly.

The three parts carry, respectively, 30%, 20%, and 50% of the marks.