IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2017

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

LEARNING IN AUTONOMOUS SYSTEMS

Tuesday 13 December 2016, 11:40
Duration: 70 minutes

Answer TWO questions

Paper contains 3 questions
Calculators not required

1 Consider a self-driving car moving on the motor way. The goal of the car is to move from start to destination as fast as possible. However, the car battery has a temperature issue and can become too hot. The car can drive slowly allowing the battery to cool, and the car can drive quickly, but that puts heat stress on the battery. We want to control the car speed in an optimal manner:

Let s be the state (3 states: s_1="cool", s_2="hot" and terminal state s_3="burnout") of a Markov Decision Process (MDP) and $V^*(s)$ the optimal value function.
Let a be an action (2 actions: a_1="drive fast", a_2 ="drive slow"), γ a discount factor and $r(a)$ be the immediate reward of choosing action a (2 rewarded actions $r(a_1) = 2$ and $r(a_2) = 1$).
Finally, $P(s'|s,a)$ denotes the transition probability of going to state s', given that we come from state s and perform action a. The world has the following dynamics:

s	a	s'	$P(s'\|a,s)$
cool	slow	cool	1
hot	slow	cool	0.5
hot	slow	hot	0.5
cool	fast	cool	0.25
cool	fast	hot	0.75
hot	fast	hot	0.875
hot	fast	burnout	0.125

a Give the optimal policy for each state (any solution approach is valid). Show your derivation. Hint: start with writing out Bellmann's Optimality Equation for this MDP.

b Is it possible to change the immediate rewards without affecting the optimal policy? If so, give an example. If it is not possible, give a proof.

c The manufacturer wants that the car appears to come across as "powerful but majestic", therefore the action sequence "fast,slow,slow" is preferred to "slow, fast, fast" when starting in any of the two states "cold" or "hot". To avoid costly engineering R&D the manufacturer wants to solve this problem by tweaking the discount factor γ. What parameter range of γ would produce the desired result?

The three parts carry, respectively, 40%, 20%, and 40% of the marks.

2 **Machine Learning** Consider a Markov Reward Process with $\gamma = 1$ with two states $Jupiter$ and $Saturn$. The transition matrix and reward function are unknown, but you have observed the following sample episodes:

$$(Jupiter, 3) \Rightarrow (Jupiter, 2) \Rightarrow (Saturn, -4) \Rightarrow (Jupiter, 4) \Rightarrow (Saturn, -3) \Rightarrow Terminal$$

$$(Saturn, -2) \Rightarrow (Jupiter, 3) \Rightarrow (Saturn, -3) \Rightarrow Terminal$$

In these two traces the sample state transitions and rewards are shown at each step, i.e. we transit in the top sequence from state $Jupiter$ to state $Jupiter$ incurring a reward of 3.

a Using either first-visit Monte-Carlo or every visit Monte-Carlo evaluation, estimate the state-value function $V(\cdot)$. Show your derivation and explain your findings.

b Using batch TD learning, what value function $V(\cdot)$ would we find if TD learning was applied repeatedly for these 2 episodes. Show your derivation and explain your findings.

The two parts carry equal marks.

3 **Reinforcement learning**

a Discuss what the key similarity or differences between Temporal Difference and Monte Carlo methods are.

b Discuss the general role of the parameter factor ϵ for the behaviour of a reinforcement learning agent that uses ϵ-greedy policies.

c Discuss the difference between On-Policy and Off-Policy Methods for reinforcement learning.

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