IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2015-2016

MEng Honours Degree in Mathematics and Computer Science Part IV
MEng Honours Degrees in Computing Part IV
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 15 December 2015, 11:30
Duration: 70 minutes

Answer TWO questions

Paper contains 3 questions
Calculators not required

1 **Reward and Value**

Consider a deterministic three state Markov Decision Process (MDP) in Figure 1. It has three states "**D**ante" , "**B**eatrice" and "**P**urgatory", the latter being a terminal, absorbing state. In each non-terminal state two actions are possible ("**K**iss", "**R**un") and as the MDP is deterministic all transition probabilities are either 0 or 1. Each action is associated with a reward as shown in the figure, so e.g. "Kiss" in state "Dante" results in a reward of 2, while "Run" in state "Beatrice" results in a reward of 0.

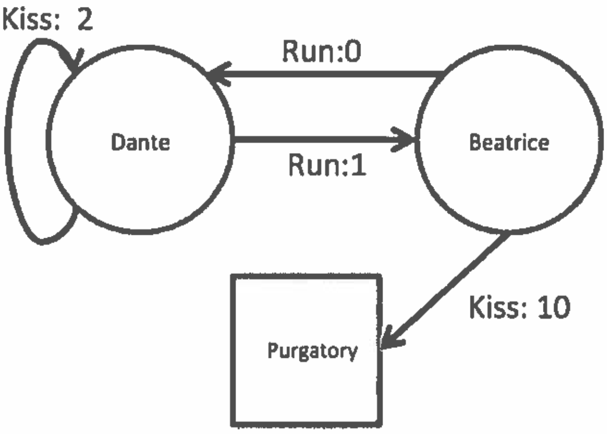

Fig. 1: The Dante & Beatrice MDP. Links denote valid transitions with "Action:Reward" values.

a Suppose an agent starts in state "Beatrice", $s_0 = B$. The agent then performs a sequence of actions: 'run','kiss','kiss','run','kiss'.

i Write out the state,action,reward trace for this episode.

ii Draw out the back-up diagram for this MDP and highlight the above sample trace. Hint: Root this tree diagram with node "Beatrice".

iii Given just this single trace, would you expect Temporal Difference or Monte Carlo policy evaluation to perform better, or equally well, in estimating the value function. Explain your answer in words (mathematics is not required).

b i Write out the Bellman equation for a fixed policy π with arbitrary γ for the above three state MDP. To do this briefly explain your notation, e.g. "$V^{\pi}(s)$ is the value of state s given policy π, which in turn is a function of state s and action a". Also, remember to expand the definitions where they are determined by this specific MDP.

ii Write out the dynamic programming update rules for the value function $V(D) \leftarrow \ldots$ and $V(B) \leftarrow \ldots$ in terms of $V(D), V(B), V(P), \gamma$ and an unbiased policy $\pi(s,a) = \frac{1}{2}$ for all s, a.

iii What are the critical values of γ which determines whether our agent stays with 'Dante' or if he goes to 'Purgatory'. You may assume that you are starting in state "Dante" and you should consider the range of $\gamma \in [0, 1[$. Show your derivation explaining in words and mathematics your reasoning. Please express the simplified answer in roots (i.e. a calculator is not necessary).

Hint: $ax^2 + bx + c = 0$ is solved by $x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

The two parts carry, respectively, 40%, and 60% of the marks.

2 **Markov Decision Processes**

Consider a three state "Life" Markov Decision Process (MDP) in Figure 2. It has three states Home,Work and Club, the latter being a terminal, absorbing state. In each non-terminal state two actions are possible ("wait", "go") and as the MDP is deterministic all transition probabilities are either 0 or 1. Each action is associated with a reward as shown in the figure, so e.g. "wait" in state "Home" results in a reward of 1, while "wait" in state "Work" results in a reward of U (a variable).

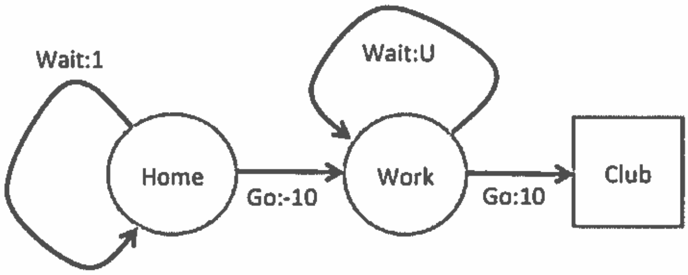

Fig. 2: Figure for question 3 (three state deterministic MDP)

a Suppose an agent starts in state "Home", $s_0 = Home$. The agent then performs a sequence of actions: 'wait','go','wait','wait','go'. In this part we assume $U = 0$.

i) Given discount factor $\gamma = \frac{1}{2}$ what is the return for this trace starting at "Home"?

ii) Estimate the value function from the above single trace of states, rewards and actions using first-visit Monte Carlo policy evaluation.

iii) Does the return you obtain in i) match your estimated $V(\text{Home})$ in ii) ? Please discuss your finding.

b **Policy evaluation:** Here we assume that U is variable and $\gamma = \frac{1}{2}$.

i) Write down the **Bellman equation** suiting our three state MDP using the following notation. Let s be the state of the MDP, a be an action, γ a discount factor and $r(s,a,s')$ be the immediate reward of moving from state s to state s' with action a under a policy $\pi(s,a)$. 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.

ii) Depending on the value of U the optimal action in state "Home" and/or "Work" may change. There are critical values of U where the optimal policy switches. Write out the optimal policies for ranges of $U \in [0, \infty]$. Hint: Calculate the optimal value function analytically using the **Bellman equation**

The two parts carry, respectively, 40%, and 60% of the marks.

3 Reinforcement learning

a i) Explain the similarities and differences between synchronous and asynchronous backup updates in Dynamic Programming.

ii) What are the key similarity or differences between Temporal Difference and Dynamic Programming methods?

iii) Explain the difference between Policy Iteration and Value Iteration.

b Consider an unknown two state MDP, which has states $\{1,2\}$ and two actions $\{eat, run\}$. We do not know the reward function or the transition function. We observe the following 4 traces, given as sequences of state, action, reward, state, action, reward, state:

$$1, eat, \frac{1}{2}, 1, eat, \frac{1}{2}, 1$$
$$2, eat, 2, 2, eat, 2, 2$$
$$1, run, 0, 2, run, 1, 1$$
$$2, run, 1, 1, run, 0, 2$$

i) What can we infer about the unknown transition and reward function given the data that we observe? To answer this question draw the minimal MDP graph consistent with the data. Please make sure to draw any self-connections. Explain your reasoning.

ii) Find the optimal value function of the MDP with your inferred transition and reward functions. Assume that the traces indeed capture all states and possible transitions and that transitions are deterministic. Derive by using **Bellman's optimality equation** or otherwise. Assume that the discount factor is $\gamma = \frac{1}{2}$. Write out the optimal greedy policy.

The two parts carry equal marks.