Imperial College London

October 17, 2019

Department of Computing

CO424/BBB – Reinforcement Learning/Bits, Brain & Behaviour – Dr Aldo Faisal

Assessed coursework 1 (Version 1.1)

To be returned as online submission.

The final document should be submitted in **PDF** format, preferably in Latex. If you need to program your answers, please paste the completed annotated source code in the appendix of your submission.

Your answers should be yours, i.e., written by you, in your own words, showing your own understanding. You must produce your own code and submit it when appropriate. Figures should be clearly readable, labelled and visible. Your coursework should not be longer than 4 single sided pages with at least 2 centimetre margins all around and 12pt font (again, 4 pages is maximum length, shorter courseworks are desirable, code appendix does not count towards page limit). You are encouraged to discuss with other students, but your answers should be yours, i.e.,written by you, in your own words, showing your own understanding. You should produce your own code and you can reuse code you developed in the lab assignments. If you have questions about the coursework please make use of the labs or Piazza, but note that GTAs cannot provide you with answers that directly solve the coursework

Marks are shown next to each question. Note that the marks are only indicative.

If you are a Dept. of Computing student, please submit this coursework on CATE. If you are a Dept. of Bioengineering student, please submit this coursework on Blackboard. All coursework will be marked together by all GTAs to ensure consistency of marks agnostic of department/course/etc.

Question 1: Understanding of MDPs (20 points)

This question is personalised by your College ID (CID) number.

1. Consider the following observed traces for an Markov Decision Process with three states $\mathbb{S} = \{s_0, s_1, s_2\}$, $\gamma = 1$, and immediate rewards specified after every state. There is only one action possible ("no choice"), which we therefore omit from the trace. The traces are composed of state $s(t)$ at time step t and immediate rewards $r(t+1)$ that are collected here upon departing from state $s(t)$. We are going to "observe" a state-reward trace computed from your CID as follows:

   - Take your CID number and omit any leading 0s. Then we call the left-most digit $CID(1)$, the subsequent digit $CID(2)$ etc.
   - Going from the leftmost to the right most digit of your CID compute a sequence of states by moving from $t = 1, 2, \ldots$

     $$s(t) = (\text{CID}(t) + 2) \mod 3 \tag{1}$$

     and the sequence of corresponding rewards is

     $$r(t+1) = \text{CID}(t) \mod 2 \tag{2}$$

For example if your CID is 012345678, then the trace of states and rewards is

$$\tau = s_2\ 0\ s_0\ 1\ s_1\ 0\ s_2\ 1\ s_0\ 0\ s_1\ 1\ s_2\ 0\ s_0\ 1\ s_0\ 0$$

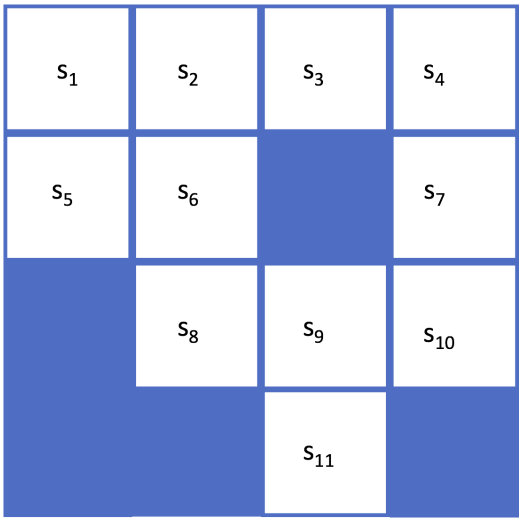
Write out your personalised trace (please use exactly the same format as in the example above).

2. Let us assume we do not know the transition matrix nor the reward matrix of this MDP, just as is the case of a naive reinforcement learning agent. What can we infer about this unknown MDP given the data that we just observe?

   (a) We want to reason about two points: 1. What can you infer about the structure of the transition matrix? 2. What can you infer about the structure of the reward function? To answer these two points draw the minimal MDP graph consistent with the data (do not add anything that is not in the data). Please make sure to draw any self-connections between states (these are typically omitted, but it will be beneficial for your learning experience to draw these in). Briefly explain your rationale for the graph you drew.

   (b) What can you say or perhaps even compute about the value of state s_0? Briefly give an explanation.

Question 2: Understanding of Grid Worlds (20 points)



This question is personalised by your CollegeID (CID) number, specifically the last 3 digits (which we call x, y, z).

Consider the following grid world. There are 11 states, corresponding to locations on a grid. This Grid World has two terminal states. The first terminal state is the reward state as reaching it yields +10 reward and ends the episode. The reward state is the state $s_j, j = ((z+1) \mod 3) + 1$, where z is the last digit of your CID. The second terminal state is the penalty state s_{11} which yields -100 reward.

Possible actions in this grid world are N, E, S and W (North, East, South, West), which correspond to moving in the four cardinal directions of the compass. The effects of actions are not deterministic, and only succeed in moving in the desired direction with probability p, in which case the action leads to the remaining 3 cardinal directions with equal probability. After the movement direction is determined, and if a wall blocks the agent's path, then the agent will stay where it is, otherwise it will move. The agent receives a reward of -1 for every transition (i.e. a movement cost), except those movements ending in the terminal state (where you collect the reward for arriving at them).

Throughout this question we set $p = 0.25 + 0.5 \times \frac{x}{10}$ and $\gamma = 0.2 + 0.5 \times \frac{y}{10}$, where x is the antepenultimate digit of your CID and y is the penultimate digit of your CID.

1. State your personalised reward state, p, and γ.

2. Compute the optimal value function and the optimal policy by any method. Briefly state how you solved the problem, including any parameters that you set or assumptions you made. Report your optimal value function by "writing in" the values into the grid world (handwritten scan is ok). Report your optimal policy function by "drawing in" arrows reflecting your optimal action for each grid world state (again, handdrawn scan is ok). For those that want to do it by hand and scan: two grid worlds are reproduced on the last page (you can use the printers in the computer lab to scan, just select option "email" after tapping in with your College card). Please write and draw clearly.

3. What action does your optimal policy execute in state s_9? Write out $p(a, s_9)$ and briefly discuss in words why that is a sensible action (consider how your personalised γ and p may have affected these).

4. Briefly discuss any general and particular features you noticed for your personal Grid World's optimal value function and optimal policy (consider how your personalised γ and p may have affected these).

Name:
CID:
reward state:
$p =$
$\gamma =$

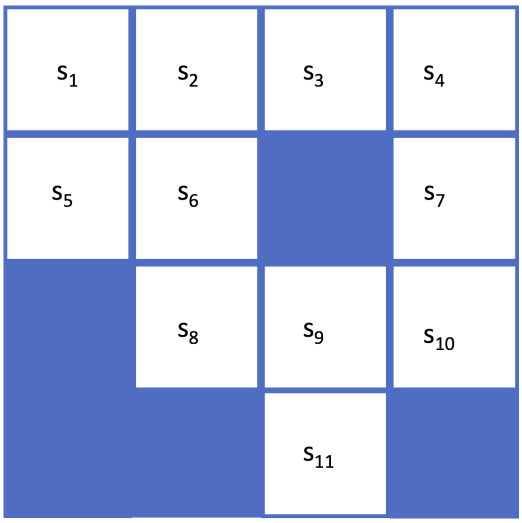


Figure 1: Optimal value function. Values for each state rounded to 2 decimal places.

.

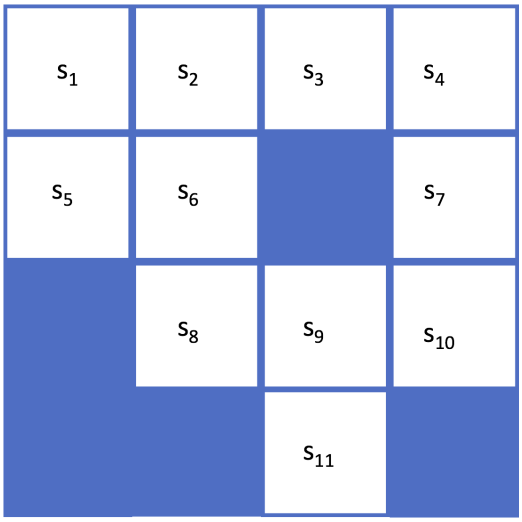


Figure 2: Optimal policy. Arrows indicate optimal action direction for each state (deterministic policy), multiple arrows from one state indicate equiprobable choice between indicated directions (stochastic policy).

.