Imperial College London

October 25, 2018

MLNC – Machine Learning & Neural Computation – Dr Aldo Faisal

Coursework 1 - Grid World

To be returned via Blackboard as indicated online.

Your coursework should contain: your name, your CID and your degree course at the top of the first page. You text should provide *brief* analytical derivations and calculations as necessary in-line, so that the markers can understand what you did. Please use *succinct* answers to the questions. Your final document should be submitted as **a single .zip file**, containing **one single PDF file**, in the format of *CID_FirstnameLastname.pdf* (example: 012345678_JaneXu.pdf), and **one single .m file**, also in the format *CID_FirstnameLastname.m*. Note, that therefore all code that you have written or modified must be within that one Matlab file. Do not submit multiple Matlab files, do not modify other Matlab files. Your Matlab script should contain a function that takes no arguments and is called `RunCoursework()`, that should produce all the Matlab-based results of your coursework (in clearly labelled text and/or figure output). This function should be able to run on its own, in a clean Matlab installation and directory with only the code we provided for the coursework present.
Please additionally paste the same **fully-commented** Matlab source code in the appendix of your PDF submission. You are allowed to use all built-in Matlab functions and any Matlab functions supplied by the course or written by you.

The markers may subtract points for badly commented code, coding that does not run and coding that does not follow the specifications. Figures should be clearly readable, labelled and visible – poor quality or difficult to understand figures may result in a loss of points.
Your coursework should not be longer than 4 single sided pages with 2 centimetre margins all around and 12pt font. 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 have to produce your own code. If you have questions about the coursework please make use of 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.

Specification

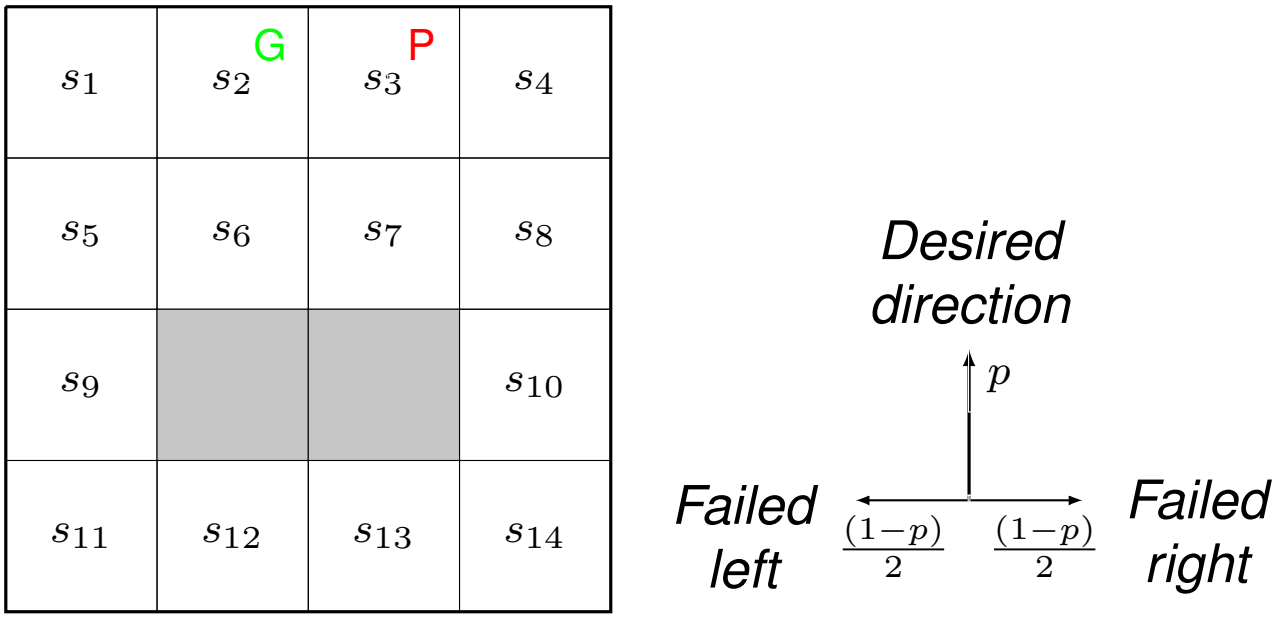

Figure 1: Grid World

This coursework uses the simple Grid World shown in Figure 1. There are 14 states, corresponding to locations on a grid – two cells (marked in grey) are walls and therefore cannot be occupied. This Grid World has two terminal states, s_2 (the Goal state, in green) and s_3 (the Penalty state, in red).

- The starting state can vary. In each simulation there is an equal probability of starting from one of the states $s_{11}, s_{12}, s_{13}, s_{14}$ (i.e. there is $\frac{1}{4}$ probability of starting from any of these states).
- Possible actions in this 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. Alternatively, the agent will move perpendicular to its desired direction in either adjacent direction with probability $\frac{(1-p)}{2}$.
- 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 to the corresponding adjacent. So for example, in the grid world where $p = 0.8$, an agent at state s_5 which chooses to move north will move north to state s_1 with probability 0.8; will move east to state s_6 with probability 0.1; or will move west staying in state s_5 with probability 0.1 (in which case it will bang into the wall and come to rest in state s_5).
- The agent receives a reward of -1 for every transition (i.e. a movement cost), except those movements ending in state s_3 (marked P for penalty) or state s_2 (marked with G for goal). For transitioning to s_3 there is a penalty of -10. For transitioning to s_2 there is a reward of 0.
- We provide the code `PersonalisedGridWorld.p`. It contains the function that sets up the Grid World with p probability of successful transition and returns the full MDP information. Note that `.p` files are similar to normal Matlab functions/scripts, but are not human-readable (i.e. you do not/should not edit it).

```
>> [NumStates, NumActions, TransitionMatrix, ...
    RewardMatrix, StateNames, ActionNames, AbsorbingStates] ...
    = PersonalisedGridWorld(p);
```

With `NumStates` being the number of states in the Grid World, and `NumActions` the number of actions the agent can take. The `TransitionMatrix` is a `NumStates` × `NumStates` × `NumActions` array of specified transition probabilities between (first dimension) successor state, (second dimension) prior state, and (third dimension) action. `RewardMatrix` is the `NumStates` × `NumStates` × `NumActions` array of reward values between (first dimension) successor state, (second dimension) prior state, and (third dimension) action. `StateNames` is a `NumStates` × `1` matrix containing the name of each state. `ActionNames` is a `NumActions` × `1` matrix containing the name of each action. Finally, `AbsorbingStates` is a `NumStates` × `1` matrix specifying which states are terminal.

The coursework is personalised by your CID number. Throughout the exercise we set $p = 0.5 + 0.5 \times \frac{x}{10}$ and $\gamma = 0.2 + 0.5 \times \frac{y}{10}$, where x is the penultimate digit of your College ID (CID), and y is the last digit of your CID. If your CID is 876543210 we have $X = 1$ and $y = 0$ resulting in $p = 0.55$ and $\gamma = 0.2$.

Questions

Points per questions are indicative only. Questions become progressively more challenging.

1. (1 point) State your CID and personalised p and γ (no need to show derivation).
2. (15 points) Assume the MDP is operating under an unbiased policy π^u, compute the value function $V^{\pi^u}(s)$ for every non-terminal state $(s_1, s_4 \ldots, s_{14})$ by any dynamic programming method of your choice. Report your result in the following format:

State	s_1	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	s_{11}	s_{12}	s_{13}	s_{14}
Value	2	2.54	$\frac{1}{3}$	...								

3. (25 points) Assume you are observing the following state transitions from the above MDP: $\{s_{14}, s_{10}, s_8, s_4, s_3\}$, $\{s_{11}, s_9, s_5, s_6, s_6, s_2\}$, $\{s_{12}, s_{11}, s_{11}, s_9, s_5, s_9, s_5, s_1, s_2\}$.
   (a) What is the likelihood that the above observed 3 sequences were generated by an unbiased policy π^u? Report the value of the likelihood.
   (b) Find a policy π^M for the observed 3 sequences that has higher likelihood than the likelihood of π^u to have generated these sequences. Report it in the following table format. Note, that as not all states are visited by these 3 sequences you only have to report the policy for visited, non-transient states. Report your result using the following format:

State	s_1	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	s_{11}	s_{12}	s_{13}	s_{14}
Action	N	S	W	...								

4. (39 points)
   (a) Assume an unbiased policy π^u in this MDP. Generate 10 traces from this MDP and write them out. When writing them out use one line for each trace, use symbols S1, S4, ..., S14, actions N, E, S, W, and the rewards in the following format (please make sure we can easily copy and paste these values from the PDF in one go), e.g. the output ***must*** be in the following format (so that we can copy and paste the text from your PDF into our automatic testing software).

```
S12,W,-1,S11,N,-1,S9,N,-1,S5,N,-1,S1,N,-1,S1,E,0
S14,E,-1,S10,E,-1,S8,W,-1,S7,S,-1,S6,N,0
```

(b) Apply First-Visit Batch Monte-Carlo Policy Evaluation to estimate the value function $\hat{V}^{\pi^u}$ from these 10 traces alone. Report the value function for every non-terminal state $(s_1, s_4 \ldots, s_{14})$ using the format specified in Question 2.

(c) Quantify the difference between $\hat{V}^{\pi^u}$ obtained from Q4.b and V^{π^u} obtained from Q2 by defining a measure that reports in a single number how similar these two value functions are. Justify your choice of measure. Then, plot the value of the proposed similarity measure against the number of traces used. Start plotting the measure using the first trace, then the first and second trace, and so forth. Comment on how increasing the number of traces affects the similarity measure.

5. (20 points)

(a) Implement ϵ-greedy first-visit Monte Carlo control. Evaluate the learning and control for two settings of ϵ, namely 0.1 and 0.75.

For each setting of ϵ, plot two types of learning curves:

i. Plot reward against episodes.

ii. Plot trace length per episode against episodes.

Note: An episode is one complete trace. A trial is many episodes starting from an initialisation of the agent. The learning curves are stochastic quantities, you may need to run a good number of repeated learning experiments to average out the variability across trials. Specify the number of trials and plot mean $\pm$ standard deviation of your learning curves.