Imperial College London

November 2, 2016

Department of Bioengineering

MLNC – Machine Learning & Neural Computation– Dr Aldo Faisal

Assessed coursework

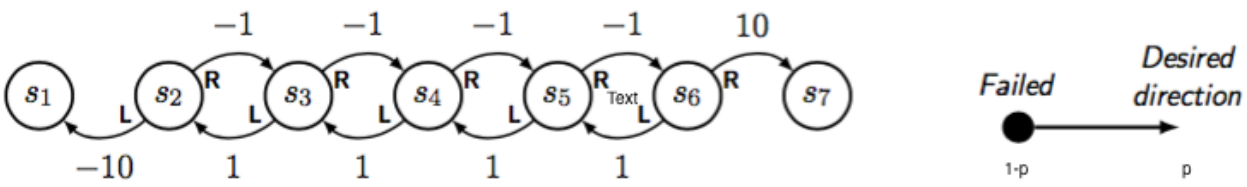
To be returned via CATE as indicated online.

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 or short pieces of code as necessary in-line, so that the markers can understand what you did. Please use succinct answers to the questions. The final document should be submitted in **PDF** format, preferably with typed text, and if necessary with scanned equations, pictures should be from Matlab or a drawing program.
If you need to program your answers, please paste the full annotated source code that generated your results in the appendix of your submission. You are allowed to use all built-in Matlab functions and any Matlab functions supplied by the course or written by you. You can solve the problems in general using computer assistance in Matlab or by any other approach (even by hand and calculator, although we do not recommend it).
Your coursework should not be longer than 4 single sided pages (excluding code appendices if really necessary). 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 and submit it where appropriate. 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.

Learning in simple systems

Consider a novel noisy stair climbing MDP (see Figure). States s_1 and s_7 are absorbing. The actions here are noisy, i.e. not deterministic. Taking a step succeeds with a probability p. So the action, L, will take the system left one state with p chance, and will remain in the same state with a $1-p$ chance. Likewise the action, R, will move right or fail with probabilities p and $1-p$ respectively. In this case here, we assume that if you attempt an action but fail you will not receive a reward, e.g. $r(s_5, R, s_5) = 0$. Note, also that the diagram below, as often done, does not show the self-connections.



Throughout the exercise we set $p = 0.6 + 0.4 \times X \times \frac{1}{10}$ and $\gamma = 0.1 + 0.9 \times Y \times \frac{1}{10}$, where X is the penultimate digit of your College Id (CID), and Y is the last digit of your CID. For example if your CID is 00123456 we have $X = 5$ and $Y = 6$ resulting in $p = 0.8$ and $\gamma = 0.64$. If your CID is 77777701 we have $X = 0$ and $Y = 1$ resulting in $p = 0.6$ and $\gamma = 0.19$.

1. (1 point) State your personalised p and γ.
2. (15 points) Assume the MDP is operating under an unbiased policy π^u, compute the value function $V^{\pi^u}(s)$ for the states $s_2, \ldots, s_6$ by any dynamic programming method of your choice.
3. (25 points) Assume you are observing the following state transitions from the above MDP: $\{s_4, s_5, s_6, s_7\}$, $\{s_4, s_5, s_6, s_7\}$, $\{s_4, s_3, s_4, s_5, s_6, s_7\}$.
   (a) What is the likelihood that the above observed 3 sequences were generated by an unbiased policy π^u?
   (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. Note, that as not all states are visited by these 3 sequences you only have to report the policy for visited, non-transient states.
4. (39 points)
   (a) Assume an unbiased policy π^u in this MDP and assume that you are starting in the middle state s_4. Generate 10 traces[1] from this MDP and write them out. When writing them out use one line for each trace, use symbols S1,S2,...,S7, actions L,R 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 should have this format:

   ```
   S4,R,-1,S5,R,-1,S6,R,10
   S4,R,-1,S5,R,0,S5,R,-1,S6,R,10
   ```

   (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 states $s_2, \ldots, s_6$.
5. (20 points)
   (a) Quantify the difference between $\hat{V}^{\pi^u}$ obtained from Q 4.b and V^{π^u} obtained from see Q 2.a by defining a measure that reports in a single number how similar these two value functions are. Justify your choice of measure and comment on the result you find.
   (b) Discuss whether setting a reasonable ϵ for ϵ-greedy Monte Carlo control is necessary (ie. $\epsilon > 0$) in worlds where the occurrence of unintended actions (e.g. in our MDP the value $1 - p$) is relatively high (e.g. $1 - p \approx \epsilon$). What are potential benefits and issues?

[1] Please make sure that you have a reasonable coverage of states S_2 to S_6, i.e. they are visited at least once across the 10 traces.