Reinforcement Learning: Coursework Assignment 2 (Semester 2, 2014)

Instructions

- This homework assignment is to be done *individually*, without help from your classmates or others. Plagiarism will be dealt with strictly as per University policy.
- Solve all problems and provide your complete solutions (with adequate reasoning behind each step) in a computer-printed or *legibly* handwritten form.
- For computational questions, include a pseudo-code in the report. Also, include the values of of all major numerical parameters involved using a reasonable accuracy and format.
- This assignment will count for 10% of your final course mark.
- Please submit your assignment by 4 pm on 27th March 2014 as a paper copy to ITO as well as an electronic version via the submit system (including code):
  submit rl 2 s1234567.pdf code.zip

Questions

Your task is to develop a program to solve a simple navigation problem using the Q_{MDP} model and algorithm. The environment has 13 discrete states. Initially, the robot is placed at a random location, chosen uniformly among the possible 13 states. Its goal is to advance to state 9, as shown in Fig. 1.

At any point in time, the robot may go along the canonical directions - north, east, south or west. Its only sensor is a bumper: when it hits a wall, the bumper triggers and the robot does not change state. Other than this piece of weak information, the robot lacks any ability to sense what state it is in. Also, it cannot sense the direction of its bumper. There is no further noise or uncertainty in this problem - just the initial location uncertainty and weak sensing.

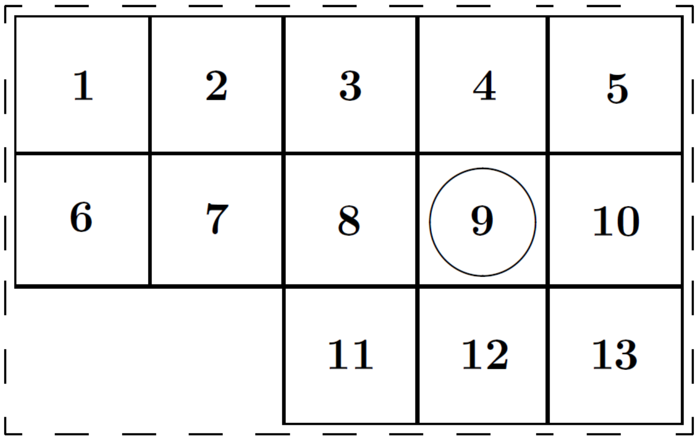

Figure 1: The environment with 13 discrete states. The target is state 9.

1. To start with, assume that the robot can observe its location. Set up a suitable MDP and find an optimal policy for reaching the target state, using value iteration. Show your value function and the resultant policy, in addition to salient design decisions related to the MDP.
2. Next, with the robot unable to observe its location, implement a Bayes filter that tracks the robot's belief over its location while it moves using a random navigation policy. The filter should use the observation of bumps, the actions taken, and the knowledge of the world map to update the belief over the state.
3. Implement the Q_{MDP} algorithm to solve the navigation POMDP. The robot starts in an unknown location, and should advance toward state 9. An episode only completes if the

robot reaches the target and its belief has converged. How many steps does it take the robot (on average) to reach the target?

4. Comment on the performance of your algorithm; does it always achieve the specified task? What improvements can you propose to make it do so?

[100 points (20 + 30 + 40 + 10 i.e. 2% + 3% +4% + 1% of the course mark)]

Hints:

- Q_{MDP} model was proposed in Ref. [1], see also the slides of Lecture 12. For the more general background you may also like read sections 4.7 and 16.2 of the Sutton-Barto book.
- See p. 502 of Thrun at al.'s Probabilistic Robotics (http://www.probabilistic-robotics.org/)
- More hints will be available in the tutorials.
- If you have any further questions related to this assignment, please contact your tutor or michael.herrmann@ed.ac.uk.

[1] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling, Learning policies for partially observable environments: Scaling up. *Proc. 12th Int. Conf. Machine Learning*. Morgan Kaufmann Publ., 1995.