INTELLIGENT SYSTEMS & ROBOTICS
ASSIGNMENT 3

shreenidhi.bharadwaj@northwestern.edu | ChristopherFiore2015@u.northwestern.edu

Objective

- Implement Q-learning and understand how to extend Q-learning to continuous actions
- Improve on how the processes interact with each other and describe the Limitations of Dynamic Programming Approaches

Submissions (1 Submission/Team)

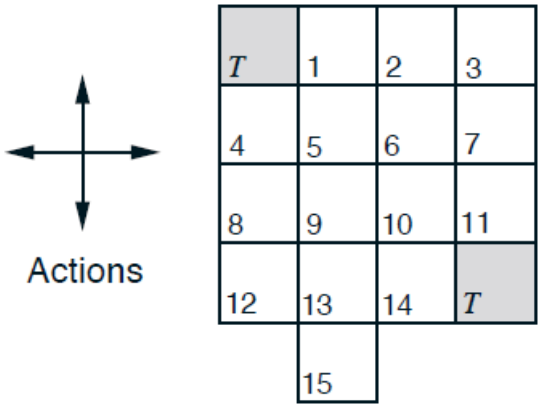
- Presentation Deck
- Python Notebook or Collab Notebook with answers to the questions below including the programming assignment.

Question 1: (Chapter 4: Exercise 4.1) {10 Points}

In Example 4.1, if Π is the equiprobable random policy, what is Q_{Π} (11; down)? What is Q_{Π} (7; down)?

Question 2: (Chapter 4: Exercise 4.2) {10 Points}

In Example 4.1, suppose a new state 15 is added to the grid world just below state 13, and its actions, left, up, right, and down, take the agent to states 12, 13, 14, and 15, respectively. Assume that the transitions from the original states are unchanged. What, then, is V_{Π} (15) for the equiprobable random policy? Now suppose the dynamics of state 13 are also changed, such that action down from state 13 takes the agent to the new state 15. What is V_{Π} (15) for the equiprobable random policy in this case?



Question 3: (Chapter 5: Exercise 5.1) {10 Points}
Consider the diagrams on the right in Figure 5.1. Why does the estimated value function jump up for the last two rows in the rear? Why does it drop off for the whole last row on the left? Why are the frontmost values higher in the upper diagrams than in the lower?

Question 4: (Chapter 6: Exercise 6.2) {10 Points}
This is an exercise to help develop your intuition about why TD methods are often more efficient than Monte Carlo methods. Consider the driving home example and how it is addressed by TD and Monte Carlo methods. Can you imagine a scenario in which a TD update would be better on average than a Monte Carlo update? Give an example scenario - a description of past experience and a current state in which you would expect the TD update to be better.
Here's a hint: Suppose you have lots of experience driving home from work. Then you move to a new building and a new parking lot (but you still enter the highway at the same place). Now you are starting to learn predictions for the new building. Can you see why TD updates are likely to be much better, at least initially, in this case? Might the same sort of thing happen in the original task?

Question 5: (Chapter 6: Exercise 6.6) {10 Points}
In Example 6.2 we stated that the true values for the random walk example are 1/6, 2/6, 3/6, 4/6 and 5/6, for states A through E. Describe at least two different ways that these could have been computed. Which would you guess we actually used? Why?

Question 6: (Chapter 6: Exercise 6.11) {10 Points}
Why is Q-learning considered an off-policy control method?

Question 7: (Chapter 6: Exercise 6.14) {10 Points}
Describe how the task of Jack's Car Rental (Example 4.2) could be reformulated in terms

of after states. Why, in terms of this specific task, would such a reformulation be likely to speed convergence?

Question 8 {30 Points}

Train a goal oriented chatbot with deep reinforcement learning. Programmatically, implement the solution and provide recommendations for management to move the solution to production.

References:

- https://towardsdatascience.com/training-a-goal-oriented-chatbot-with-deep-reinforcement-learning-part-i-introduction-and-dce3af21d383
- https://github.com/pochih/RL-Chatbot