CS 7642: Reinforcement Learning and Decision Making **Homework #4**

Q-Learning

1 Problem

1.1 Description

In this homework, you will have a complete RL experience. You will work towards implementing and evaluating the Q-learning algorithm on a simple domain. Q-learning is a fundamental RL algorithm and has been successfully used to solve a variety of decision-making problems. For this homework, you will have to think carefully about algorithm implementation, especially exploration parameters.

The domain you will be tackling is called Taxi (Taxi-v2). It is a discrete MDP which has been used for RL research in the past. This is also a chance to become more familiar with the OpenAI Gym environment https://github.com/openai/gym. The OpenAI Gym is a platform where users can test their RL algorithms over a selection of domains.

The Taxi problem was introduced by Dietterich 1998. It is a grid-based domain where the goal of the agent is to pick up a passenger at one location and drop them off at another. There are 4 fixed locations, each assigned a different letter. The agent has 6 actions; 4 for movement, 1 for pickup, and 1 for dropoff. The domain has a discrete state space and deterministic transitions.

1.2 Procedure

Implement a basic version of the Q-learning algorithm and use it to solve the taxi domain. The agent should explore the MDP, collect data to learn the optimal policy and the optimal Q-value function. (Be mindful of how you handle terminal states, typically if S_t is a terminal state, $V(S_{t+1}) = 0$). Use $\gamma = 0.90$. Also, you will see how an Epsilon-Greedy strategy can find the optimal policy despite finding sub-optimal Q-values. As we are looking for optimal Q-values you will have to carefully consider your exploration strategy.

Evaluate your agent using the OpenAI gym 0.14.0 Taxi-v2 environment. Install OpenAI Gym 0.14.0 with `pip install gym==0.14.0`

1.3 Examples

Below are the optimal Q values for 5 (state, action) pairs of the Taxi domain.

- $Q(462, 4) = -11.374402515$
- $Q(398, 3) = 4.348907$
- $Q(253, 0) = -0.5856821173$
- $Q(377, 1) = 9.683$
- $Q(83, 5) = -12.8232660372$

1.4 Resources

1.4.1 Lectures

- Lesson 4: Convergence
- Lesson 7: Exploring Exploration

1.4.2 Readings

- Asmuth-Littman-Zinkov-2008.pdf Asmuth, Michael L Littman, and Zinkov 2008
- littman-1996.pdf (Chapters 1-2) Michael Lederman Littman 1996

1.4.3 Documentation

- http://gym.openai.com/docs/

1.5 Submission Details

The due date is indicated on the Canvas page for this assignment.

Make sure you have set your timezone in Canvas to ensure the deadline is accurate. To complete the assignment calculate answers to the specific problems given and submit results to Canvas.

You will be evaluated based on optimality of results. This will be assessed by your algorithm's optimal Q-values for 10 specific state-action pairs (remember to use $\gamma = 0.90$) You will submit your results to 10 problems selected for you on Canvas. The values will be graded on a 0.01 precision threshold.

Optionally, you might want to, *with the same implementation*, solve the environment under OpenAI's criteria. If you accomplish that you will have definitely learned something about exploration vs exploitation, and the difference between an optimal policy and optimal Q-values.

References

[ALZ08] John Asmuth, Michael L Littman, and Robert Zinkov. "Potential-based Shaping in Model-based Reinforcement Learning." In: *AAAI*. 2008, pp. 604–609.

[Die98] Thomas G Dietterich. "The MAXQ Method for Hierarchical Reinforcement Learning." In: *ICML*. Vol. 98. Citeseer. 1998, pp. 118–126.

[Lit96] Michael Lederman Littman. *Algorithms for sequential decision making*. Brown University Providence, RI, 1996.