CS6700 : Reinforcement Learning
Written Assignment #4

Deadline: May 5th 2019, 11:55 pm

- This is an individual assignment. Collaborations and discussions are strictly prohibited.
- Be precise with your explanations. Unnecessary verbosity will be penalized.
- Check the Moodle discussion forums regularly for updates regarding the assignment.
- **Please start early.**

1. (4 marks) Consider the following problem design. You have a grid world with several rooms, as discussed in class. You set up an agent with options for exiting each of the rooms into the other. You also allow the agent to pick from the four primitive actions. There is a step reward of -1. The learning algorithm used is SMDP Q-learning, with normal Q learning updates for the primitive actions. You expect the agent to learn faster due to the presence of the options, but discover that it is not the case. Can you explain what might have caused this?

2. (6 marks) This question requires you to do some additional reading. Dietterich specifies certain conditions for safe-state abstraction for the MaxQ framework. Even if we do not use the MaxQ value function decomposition, the hierarchy provided is still useful. Which of the safe-state abstraction conditions are still necessary when we do not use value function decomposition.

3. (4 marks) What are some advantages and disadvantages of A3C over DQN? What are some potential issues that can be caused by asynchronous updates in A3C?

4. (3 marks) Option discovery has entailed using heuristics, the most popular of which is to identify bottlenecks. Justify why bottlenecks are useful sub-goals. Describe scenarios in which a such a heuristic would fail.

5. (4 marks) List a few ways in which a model of the world ($\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow P(\mathcal{S})$) could be used to augment the learning process.

6. (3 marks) Q-MDPs are a technique for solving the problem of behaving in POMDPs. The behavior produced by this approximation would not be optimal. In what sense is it not optimal? Are there circumstances under which it can be optimal?

7. (5 marks) Dyna-Q and Dyna-Q+ are presented in Chapter 8 of Sutton and Barto. At time-step 3000 (see Figure 1), the dynamics of the environment changes as described in Example 8.3. Why does Dyna-Q+ perform better than Dyna-Q, before and after the 3000 time-step mark? Why is the difference larger, after the 3000 time-step mark? Design an algorithm that performs better than Dyna-Q+ in the first 3000 steps (i.e dynamics of the environment is stationary).

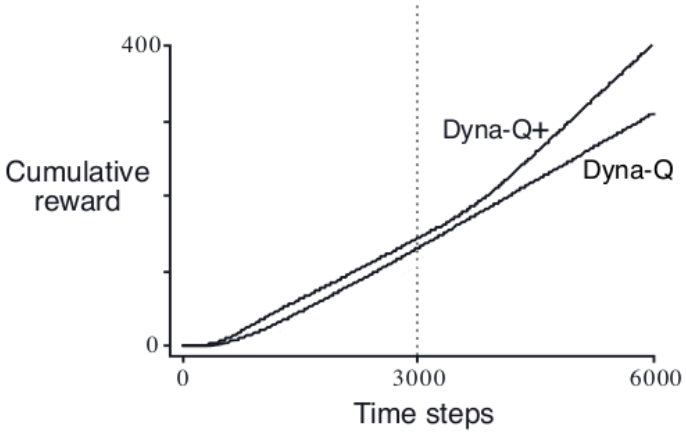

Figure 1: Dyna-Q+ vs Dyna-Q