CS6700 : Reinforcement Learning
Written Assignment #3

Deadline: ??

- 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.**

AUTHOR : Name.
ROLL NUMBER :

1. (3 marks) Consider the problem of solving POMDPs using Deep Reinforcement Learning. Can you think of ways to modify the standard DQN architecture to ensure it can remember histories of states. Does the experience replay also need to be modified? Explain.

Solution: Ref: http://cs229.stanford.edu/proj2015/363_report.pdf
In deep Q learning, a neural network is used to approximate the POMDP Q-values. For POMDP, the Q values are parameterized by either the belief and the action Q(b,a) or an action-observation history h and Q(h,a). The modified Q-values can be learned by a neural network that is characterized by weights and biases combined denoted as θ. Q values will be $Q(b, a|\theta)$. The standard update of Q values may lead to divergence. Such issues is resolved by experience replay tuples (b,a,r,b) which are recorded in a reply memory. The aim of replay is to stabilize the learning by drawing samples uniformly.

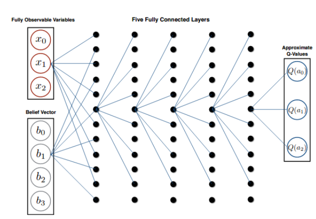

Figure 2: Five layer fully connected network that maps the concatenated fully observable variable and belief vectors to Q-values

Apart from this, a separate target network is used update state targets to the main network. An adaptive learning method can be used to regulate and adjustment the parameter rate of the network.

Changes in architecture:
The changed architecture of DQN is given in figure-1. In architecture, simulator is used to populate and experience reply dataset. The input layer consists the belief of the agent and fully observable state variables which generalize the representation of problems. The agent may have some knowledge of the system state known as mixed observability MDPs(MDOMPs). **A similar architecture was used for training the DQN on action observation histories which requires the manipulation of experience reply for efficient use of reply.** The fully observable state variables are used as inputs to the network. The current formulation uses a fully connected network that either takes the fully observable state variables and the belief or just the belief and outputs a value approximation.

2. (4 marks) Exploration is often ameliorated by the use of counts over the various states. For example, one could maintain a visitation count $N(s)$, for every state and use the same to generate an intrinsic reward $(r_i(s))$ for visiting that state.

$$r_i(s) = \tau \times \frac{1}{N(s)}$$

However, it is intractable to maintain these counts in high-dimensional spaces, since the count values will be zero for a large fraction of the states. Can you suggest a solution(s) to handle this scenario? How can you maintain an approximation of the counts for a large number of states and possibly generalize to unseen states?

Solution:

3. (5 marks) Suppose that the MDP defined on the observation space is k-th order Markov, i.e. remembering the last k observations is enough to predict the future. Consider using a belief state based approach for solving this problem. For any starting state and initial belief, the belief distribution will localize to the right state after k updates, i.e., the true state the agent is in will have a probability of 1 and the other states will have a probability of 0. Is this statement true or false? Explain your answer.

Solution:

4. (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?

Solution:

5. (3 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?

Solution:

6. (6 marks) There are a variety of very efficient heuristics available for solving deterministic travelling salesman problems. We would like to take advantage of such heuristics in solving certain classes of large scale navigational problems in stochastic domains. These problems involve navigating from one well demarcated region to another. For e.g., consider the problem of delivering mail to the office rooms in a multi storey building.

   (a) (4 marks) Outline a method to achieve this, using concepts from hierarchical RL.

   Solution:

   (b) (2 marks) What problems would such an approach encounter?

   Solution:

7. (6 marks) This question may require you to refer to https://link.springer.com/content/pdf/10.1007/BF paper on average reward RL. Consider the 3 state MDP shown in Figure 1. Mention the recurrent class for each such policies. In the average reward setting, what are the corresponding ρ^π for each such policy ? Furthermore, which of these policies are gain optimal ?

   (a) (3 marks) What are the different deterministic uni-chain policies present ?

   (b) (3 marks) In the average reward setting, what are the corresponding ρ^π for each such policy ? Furthermore, which of these policies are gain optimal ?

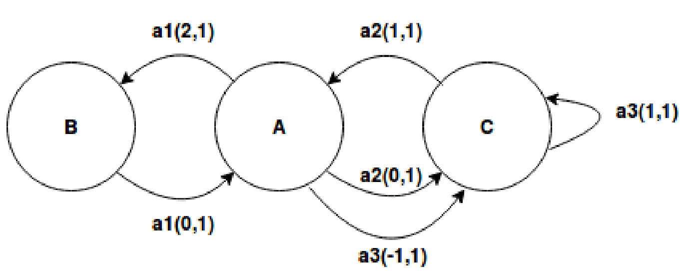

Figure 1: Notation : action(reward, transition probability). Example : a1(3, 1) refers to action a1 which results in a transition with reward +3 and probability 1