Assignment #7 Reinforcement Learning

(note: all the answers should be typed in MS-word or Latex and the pdf file is submitted. No handwritten answers are accepted.)

1- What are the two sources of error in Deep RL with function approximation?

2- In TD learning with a neural network what are we trying to minimize? What are we trying to maximize?

3- What is the main benefit of deep neural networks in function approximation compared to linear method? What is the linear function approximation useful for?

4- In DQN, what is the purpose of the target network and value network?

5- In the Deep Q-learning method, which are true:

   a. epsilon-greedy is required to ensure exploration.
   b. exploration this taken care of by the randomization provided by experience replay?

6- Value function-based RL methods are oriented towards finding deterministic policies whereas policy search methods are geared towards finding stochastic policies
   a. True
   b. False

7- What makes the optimization space smooth in policy gradient methods?

8- How does actor-critic architecture improve the “vanila” policy gradient?

9- What is the Advantage function, $A_\pi(s_t, a_t)$, in actor-critic architecture? Write down the equation for monte-carlo-based Advantage function and TD-based advantage function and comment on the pro and cons of each on.

10- Can you find a resemblance between actor-critic architecture and generalized policy iteration in chapter 4?

11- In the actor-critic architecture described in the lecture, assuming the use of differentiable function approximators, we can conclude that the use of such a scheme will result in:

a) convergence to a globally optimal policy
b) convergence to a locally optimal policy
c) cannot comment on the convergence of such an algorithm.

12- Assume that we are using the linear function approximation for the critic network and SoftMax policy with linear function approximation for the actor network. Write down the parameter update equations for the actor and the critic network (use on-line update equations similar to the algorithm in page 332)

13- Consider a trajectory rollout of (s, a_1, s, a_2, s, a_3). The initial policy depicted in the figure below for state s.The horizontal line also shows the value of the Advantage function for each (state, action) pair. After applying the policy gradient update, what do you expect the probability of each action to change to?