Assignment 1
Reinforcement Learning
Prof. B. Ravindran

1. Which among the following are features of a reinforcement learning solution to a learning problem?

   (a) trial and error approach to learning
   (b) exploration versus exploitation dilemma
   (c) learning based on rewards
   (d) absence of any feedback or supervision

2. Modelling a living organism as an RL agent, the environment encompasses

   (a) everything external to the organism
   (b) some portions internal to the organism as well

3. It was mentioned in the lecture that reinforcement learning solutions have to be able to handle delayed rewards. What exactly is/are the problems that arises due to delayed rewards?

   (a) problem in storing rewards
   (b) problem in assigning rewards to the actions that caused them
   (c) problem in mapping states to the rewards that can be achieved from those states

4. In the tic-tac-toe problem, it can be observed that many board positions (states) are symmetric. Assuming an imperfect opponent with a stationary policy/strategy who is also taking advantage of symmetries in the board positions, which of the following statements are true, if we take advantage of symmetries by maintaining a single (probability) value for all symmetric states?

   (a) time taken to learn the values reduces
   (b) space required to store the values reduces
   (c) the final policy learned in this approach is better than the policy learned when not taking advantage of symmetries

5. Considering again the question of taking advantage of symmetric board positions, is it true, that symmetrically equivalent positions should necessarily have the same value?

   (a) no

(b) yes

6. This question is related to the one discussed in class. Recall the temporal difference learning approach to the tic-tac-toe problem. Suppose that the probability of winning at a particular state is 0.6, the max probability value in the next set of states is 0.8, and based on our exploration policy, we choose a next state which has probability value 0.4. Should you backup the current state's probability value based on this choice of next state (i.e., move probability value 0.6 closer to 0.4) or not, given that the agent never stops exploring (i.e., the agent always makes an exploratory move some fraction of the time)?

   (a) backup the value

   (b) do not backup the value

7. Suppose we want an RL agent to learn to play the game of golf. For training purposes, we make use of a golf simulator program. Assume that the original reward distribution gives a reward of +10 when the golf ball is hit into the hole and -1 for all other transitions. To aide the agents learning process, we propose to give an additional reward of +3 whenever the ball is within a 1 metre radius of the hole. Is this additional reward a good idea or not? Why?

   (a) yes, the additional reward will help speed-up learning

   (b) yes, getting the ball to within a metre of the hole is like a sub-goal and hence, should be rewarded

   (c) no, the additional reward may actually hinder learning

   (d) no, it violates the idea that a goal must be outside the agents direct control.

8. Suppose that we are given a 10 armed bandit problem and are also told that the rewards for each arm are deterministic. How many times should each arm be tried before we can identify the best arm with 95% probability?

   (a) 1

   (b) 2

   (c) 95

   (d) none of the above

9. Suppose that you have been given a number of different drug formulations to treat a particular disease and your job is to identify one among them that best meets certain criteria with regards to its efficacy in treating the disease. Before you run the experiments, you need to provision for the samples that would be required. Treating this as a multi-armed bandit problem, which kind of solution method would you prefer for identifying the best option?

   (a) asymptotic correctness

   (b) regret optimality

   (c) PAC optimality

10. Suppose we have a 10-armed bandit problem where the rewards for each of the 10 arms is deterministic and in the range (0, 10). Which among the following methods will allow us to accumulate maximum reward in the long term?

(a) ϵ-greedy with $\epsilon = 0.1$

(b) ϵ-greedy with $\epsilon = 0.01$

(c) greedy with initial reward estimates set to 0

(d) greedy with initial reward estimates set to 10