Assignment 9

Reinforcement Learning

Prof. B. Ravindran

1. (1) Which among the following is/are the advantages of using the Deep Q-learning method over other learning methods that we have seen?

   (a) a faster implementation of the Q-learning algorithm

   (b) guarantees convergence to the optimal policy

   (c) obviates the need to hand-craft features used in function approximation

   (d) allows the use of off-policy algorithms rather than on-policy learning schemes

2. (1) In the Deep Q-learning method, is ϵ-greedy (or other equivalent techniques) required to ensure exploration, or is this taken care of by the randomisation provided by experience replay?

   (a) no

   (b) yes

3. (1) Value function based methods are oriented towards finding deterministic policies whereas policy search methods are geared towards finding stochastic policies. True or false?

   (a) false

   (b) true

4. (1) Suppose we are using a policy gradient method to solve a reinforcement learning problem. Assuming that the policy returned by the method is not optimal, which among the following are plausible reasons for such an outcome?

   (a) the search procedure converged to a locally optimal policy

   (b) the search procedure was terminated before it could reach the optimal policy

   (c) the sample trajectories arising in the problem were very long

   (d) the optimal policy could not be represented by the parametrisation used to represent the policy

5. (1) In using policy gradient methods, if we make use of the average reward formulation rather than the discounted reward formulation, then is it necessary to consider, for problems that do not have a unique start state, a designated start state, s_0?

   (a) no

(b) yes

6. (1) Using similar parametrisations to represent policies, would you expect, in general, MC policy gradient methods to converge faster or slower than actor-critic methods assuming that the approximation to Q^π used in the actor-critic method satisfies the compatibility criteria?

   (a) slower

   (b) faster

7. (1) If f_w approximates Q^π and is compatible with the parameterisation used for the policy, then this indicates that we can use f_w in place of Q^π in the expression for calculating the gradient of the policy performance metric with respect to the policy parameter because

   (a) $Q^\pi(s,a) - f_w(s,a) = 0$ in the direction of the gradient of $f_w(s,a)$

   (b) $Q^\pi(s,a) - f_w(s,a) = 0$ in the direction of the gradient of $\pi(s,a)$

   (c) the error between Q^π and f_w is orthogonal to the gradient of the policy parameterisation

8. (1) Suppose we use the actor-critic algorithm described in the lectures where Q^π is approximated and the approximation used is compatible with the parametrisation used for the actor. 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