Assignment 9 (Sol.)

Reinforcement Learning

Prof. B. Ravindran

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

   Sol. (c)
   As we have seen with the Atari games example, the input to the network is raw pixel data from which the network manages to learn appropriate features to be used for representing the action value function.

2. 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

   Sol. (b)
   Some technique to ensure exploration is still required. As with the original Q-learning algorithm, if we only consider transitions generated by the action value function (which is essentially what we'll get with experience replay without any exploration), a large part of the state space will likely remain unexplored.

3. 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

   Sol. (b)
   With value function based methods, policies are derived from the value function by considering, for a state, that action which gives maximum value. This leads to a deterministic policy. On the other hand, no such maximisation is at work in policy search methods, where the parameters learned, using the gradient descent method, for example, determine the agent's policy. This is likely to be stochastic if the optimal policy (global or local) is stochastic.

4. 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

   Sol. (a), (b), (d)
   Option (c) may result in an increase in the time it takes to converge to a policy, but does not necessarily affect the optimality of the policy obtained.

5. 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

   Sol. (a)
   We used the concept of a designated start state to allow a single value that can be assigned to a policy for the purpose of evaluation. The same result is obtained when using the average reward formulation, i.e, by using the average reward formulation we can compare policies according to their long term expected reward per step, $\rho(\pi)$, where

$$\rho(\pi) = lim_{n\to\infty}\frac{1}{N}E\{r_1 + r_2 + r_3 + ... + r_N|\pi\}$$

6. 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

   Sol. (a)
   As we have seen, MC policy gradient algorithms may suffer from large variance due to long episode lengths which can slow down convergence. Actor-critic methods, by relying on value function estimates can lead to reduced variance, and hence, faster convergence.

7. 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

Sol. (b), (c)
As indicated by options (b) & (c), we can use f_w in place of Q^π if the difference between the two is zero in the direction of the gradient of $\pi(s, a)$.

8. 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

   Sol. (b)
   The idea behind the two theorems (policy gradient and policy gradient with function approximation) that we saw in the lectures is to show that using such an approach will result in the policy converging. However, we can only prove convergence to a locally optimal policy.