Assignment 8 (Sol.)

Reinforcement Learning

Prof. B. Ravindran

1. Is the problem of non-stationary targets an issue when using Monte Carlo returns as targets?

   (a) no

   (b) yes

   Sol. (a)
   When Monte Carlo returns are used as targets, then the targets are stationary as the target values do not change as the parameters evolve.

2. In a parameterised representation of the value function, we use a feature which acts as a counter of some concept in the environment (number of cans the robot has collected, for example). Does such a feature used for representing the state space lead to a violation of the Markov property?

   (a) no

   (b) yes

   Sol. (a)
   As long as states contain adequate information so that the conditional probability distribution of the future states depends only upon the current state of the environment, any kind of feature can be used to describe the state space without violating the Markov property.

3. Which of the following will effect generalisation when using the tile coding method?

   (a) modify the number of tiles in each tiling (assuming the range covered along each dimension by the tilings remains unchanged)

   (b) modify the number of tilings

   (c) modify the size of tiles

   (d) modify the shape of tiles

   Sol. (a), (b), (c), (d)
   Option (a) and (c) are equivalent and effect the range over which values of one state generalise to values of other states. The same effect is also achieved by the number of tilings, since, for example, more tilings (with different overlapping regions) will result in more states being affected by updates to the value of a single state. Finally, the shape of tiles has an effect on generalisation since it allows us to vary between uniform generalisation (for example using square tiles) and non-uniform generalisation (for example using rectangular tiles) where generalisation is more pronounced along some dimensions and not so much along others.

4. For a particular MDP, suppose we use function approximation and using the gradient descent approach, converge to the value function that is the global optimum. Is this value function the same, in general, as the true value function of the MDP?

   (a) no

   (b) yes

   Sol. (a)
   The value function corresponding to the global optimum and the true value function need not be the same considering that, given the representation used, the function approximation may not even be able to express the true value function.

5. Which of the following methods would benefit from normalising the magnitudes of the basis functions?

   (a) on-line gradient descent TD(λ)

   (b) linear gradient descent Sarsa(λ)

   (c) LSPI

   (d) none of the above

   Sol. (a), (b)
   Order of magnitude differences in the values of the features can impact the performance of gradient descent procedures and hence such methods can benefit from normalisation of the inputs. On the other hand LSPI involves solving a system of linear equations, for which procedures exist which are not impacted by scaling issues.

6. Suppose that individual features, $\phi_i(s, a)$, used in the representation of the action value function are non-linear functions of s and a. Is it possible to use the LSTDQ method in such scenarios?

   (a) no

   (b) yes

   Sol. (b)
   When discussing linear methods such as LSTDQ, we are talking about methods which can approximate functions which are linear in the parameter vector, not the feature vector.

7. Which among the following statements about the LSTD and LSTDQ methods is/are correct?

   (a) LSTD learns the state value function

   (b) LSTDQ learns the action value function

   (c) both LSTD and LSTDQ can reuse samples

   (d) both LSTD and LSTDQ can be used along with tabular representations of value functions

   Sol. (a), (b), (d)
   Option (c) is not correct. Given a set of samples collected from the actual process (not from a generative model in which case reusing samples is perhaps not that important) it is useful for these samples to be reused in evaluating different policies. Recall that LSPI is a policy

iteration algorithm where the policy is constantly being improved upon (and hence changing). Such sample reuse is possible in LSTDQ if for any policy π, $\pi(s')$ is available for each s' in the set of samples. This is because for different policies, the same samples can be made use of, as individual policies only determine the $\phi(s', \pi(s'))$ component of $\tilde{A}$. This is not the case with LSTD where freedom in action choices is not available and must be determined from the policy that is being evaluated.

8. Consider the five state random walk task described in the book. There are five states, $\{s_1, s_2, ..., s_5\}$, in a row with two actions each, left and right. There are two terminal states at each end, with a reward of +1 for terminating on the right, after s_5 and a reward of 0 for all other transitions, including the one terminating on the left after s_1. In designing a linear function approximator, what is the least number of state features required to represent the value of the equi-probable random policy?

   (a) 1

   (b) 2

   (c) 3

   (d) 5

   Sol. (a)
   The value of the states from s_1 to s_5 are 1/6, 2/6, ..., 5/6. Hence a single feature, $\phi(s_i) = i$ is adequate to represent the values.