Assignment 7 (Sol.)

Reinforcement Learning

Prof. B. Ravindran

1. Consider example 7.1 and figure 7.2 in the text book. Which among the following steps (keeping all other factors unchanged) will result in a decrease in the RMS errors shown in the graphs?

   (a) increasing the number of states of the MDP

   (b) increasing the number of episodes over which error is calculated

   (c) increasing the number of repetitions over which the error is calculated

   (d) none of the above

   Sol. (b)
   Note that the graphs are generated by averaging over the first 10 episodes. If we increase the number of episodes considered, the error shown in the graphs would reduce as evaluation of the policy improves.

2. Considering episodic tasks and for $\lambda \in (0, 1)$, is it true that the one step return always gets assigned the maximum weight in the λ-return?

   (a) no

   (b) yes

   Sol. (a)
   This is not necessarily true and depends on the length of an episode (as well as the value of λ). For example, consider an episode of length 3 and a value of $\lambda = 0.7$.

3. In the TD(λ) algorithm, if $\lambda = 1$ and $\gamma = 1$, then which among the following are true?

   (a) the method behaves like a Monte Carlo method for an undiscounted task

   (b) the eligibility traces do not decay

   (c) the value of all states are updated by the TD error in each episode

   (d) this method is not suitable for continuing tasks

   Sol. (a), (b), (d)
   Note that even if $\lambda = 1$ and the eligibility traces do not decay, states must first be visited before their values can be updated.

4. Assume you have a MDP with $|S|$ states. You decide to use an n-step truncated corrected return for the evaluation problem on this MDP. Do you think that there is any utility in considering values of n which exceed $|S|$ for this problem?

   (a) no

   (b) yes

   Sol. (b)
   Note that the number of steps in an n-step truncated corrected return is related to the length of the trajectories which can exceed the number of states in the state space of a problem.

5. Which among the following are reasons to support your answer in the previous question?

   (a) only values of $n \leq |S|$ should be considered as the number of states is only $|S|$

   (b) all implementations with $n > |S|$ will result in the same evaluation at each stage of the iterative process

   (c) the length of each episode may exceed $|S|$, and hence values of $n > |S|$ should be considered

   (d) regardless of the number of states, different values of n will always lead to different evaluations (at each step of the iterative process) and hence cannot be disregarded

   Sol. (c)

6. Consider the text book figure 5.1 describing the first-visit MC method prediction algorithm and figure 7.7 describing the TD(λ) algorithm. Will these two algorithms behave identically for $\lambda = 1$? If so, what kind of eligibility trace will result in equivalence?

   (a) no

   (b) yes, accumulating traces

   (c) yes, replacing traces

   (d) yes, dutch traces, with $\alpha = 0.5$

   Sol. (a)
   The two algorithms are not identical since figure 7.7 describes the online version of the TD(λ) algorithm, whereas in the MC algorithm described in figure 5.1, updates are obviously not made after each individual reward is observed.

7. Given the following sequence of states observed from the beginning of an episode,

$$s_2, s_1, s_3, s_2, s_1, s_2, s_1, s_6$$

   what is the eligibility value, $e_7(s_1)$, of state s_1 at time step 7 given trace decay parameter λ, discount rate γ, and initial value, $e_0(s_1) = 0$, when accumulating traces are used?

   (a) $\gamma^7\lambda^7$

   (b) $(\gamma\lambda)^7 + (\gamma\lambda)^6 + (\gamma\lambda)^3 + \gamma\lambda$

   (c) $\gamma\lambda(1 + \gamma^2\lambda^2 + \gamma^5\lambda^5)$

   (d) $\gamma^7\lambda^7 + \gamma^3\lambda^3 + \gamma\lambda$

Sol. (c)
According to the non-recursive expression for accumulating eligibility trace, we have

$$e_t(s) = \sum_{k=0}^{t} (\gamma\lambda)^{t-k} I_{ss_k}$$

where I_{ss_k} is an indicator function.

Using the above expression along with the given state sequence, we have

$$e_7(s_1) = (\gamma\lambda)^{7-1} + (\gamma\lambda)^{7-4} + (\gamma\lambda)^{7-6} = \gamma\lambda + \gamma^3\lambda^3 + \gamma^6\lambda^6$$

8. For the above question, what is the eligibility value if replacing traces are used?

   (a) $\gamma^7\lambda^7$

   (b) $\gamma\lambda$

   (c) $\gamma\lambda + 1$

   (d) $3\gamma\lambda$

   Sol. (b)
   We know that when using replacing traces, the eligibility trace of a state is set to 1 if that state is visited and decayed by a factor of $\gamma\lambda$ otherwise. Thus, the latest occurrence of state s_1 just before state s_6 would cause $e_6(s_1)$ to be set to 1 and after the occurrence of state s_6, this would decay to $e_7(s_1) = \gamma\lambda$.

9. In solving the control problem, suppose that at the start of an episode the first action that is taken is not an optimal action according to the current policy. Would an update be made corresponding to this action and the subsequent reward received in Watkin's Q(λ) algorithm?

   (a) no

   (b) yes

   Sol. (b)
   This is immediately clear from the Watkin's Q(λ) algorithm described in the text.

10. Suppose that in a particular problem, the agent keeps going back to the same state in a loop. What is the maximum value that can be taken by the eligibility trace of such a state if we consider accumulating traces with $\lambda = 0.25$ and $\gamma = 0.8$?

    (a) 1.25

    (b) 5.0

    (c) ∞

    (d) insufficient data

    Sol. (a)
    For accumulating traces maximum increase in eligibility occurs if the state is selected: $e_t(s) = \gamma\lambda e_{t-1}(s) + 1$. At maximum, $e_t(s) = e_{t-1}(s)$, giving, $e_t(s) = e_{t-1}(s) = \frac{1}{1-\gamma\lambda}$.