Assignment 7

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

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

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

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

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

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 $\alpha = 0.5$

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

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) $3\gamma\lambda$

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

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