CS6700 : Reinforcement Learning

Written Assignment #2

MDPs and TD — Deadline: 17 Mar 2019, 11:55 pm

- This is an individual assignment. Collaborations and discussions are strictly prohibited.
- Be precise with your explanations. Unnecessary verbosity will be penalized.
- Check the Moodle discussion forums regularly for updates regarding the assignment.
- **Please start early.**
- Turn in **only the answers** on Turnitin.

1. Consider the following 10 episodes from a Markov reward process (A Markov reward process is specified by the states and the rewards obtained in inter state transitions):
A, 0, B, 1
A, 0, C, 1
A, 1, C, 1
C, 0
C, 1
C, 1
B, 0
B, 0
B, 1
B, 1

   (a) Estimate the values of the states A, B and C using batch TD(0) and batch MC with parameters : $\gamma = 1$ and $\alpha = 1$.

   (b) Draw the MDP for the process.

   (c) Calculate the MSE (Mean Squared Error) of the estimates from TD(0) and MC on the train data.

$$MSE = \frac{1}{\text{Number of times } S_i \text{ appears in the batch}} \sum_{i=1}^{\#S} \sum_{j=1}^{\#\text{S in batch}} (V(S_i) - R(S_{ij}))^2$$

   Which method gives lower MSE ? Based on this, which of the two methods is truer to the training data ?

   (d) Which of the two methods is truer to the model (and hence the Markov assumption) drawn in part (b) ?

   (e) Which of the two methods would produce lower error on the future data and why?

2. Consider the task of controlling a system when the control actions are delayed. The control agent takes an action on observing the state at time t. The action is applied to the system at time $t + \tau$. The agent receives a reward at each time step.

   (a) What is an appropriate notion of return for this task?

(b) Give the TD(0) backup equation for estimating the value function of a given policy.

3. Consider the case where you truncate the eligibility trace after 3 steps as shown below:

$$e_t(s) = \begin{cases} \gamma\lambda e_{t-1}(s) & \text{if } s \in [s_{t-3}, s_{t-2}, s_{t-1}], \\ 1 & \text{if } s = s_t, \\ 0 & \text{otherwise} \end{cases}$$

   Assume that a state is never revisited in a trajectory, for convenience. What is the equivalent "lambda return" that you are optimizing? Extra credit: Do this for any n.

4. Suppose that the system that you are trying to learn about (estimation or control) is not perfectly Markov. Comment on the suitability of using Temporal Difference learning. Explicitly state any assumptions that you are making.

5. For some policy π in an MDP, if there exists a constant k, such that for all $\gamma > k$, π is optimal for the discounted reward formulation, then π is said to be Blackwell optimal. Consider the gridworld problem shown below, where S denotes a starting state and T1 and T2 are terminal states. The reward for terminating in T1 is +10 and for terminating in T2 is +5. Any transition into the state marked $*$ has a reward of $a \in \mathbb{R}$. All other transitions have a reward of 0.

   For this problem, give a characterization of Blackwell optimal policies, in particular the value k, parameterized by a. In other words, for different ranges of a, give the Blackwell optimal policy, along with the value of k.

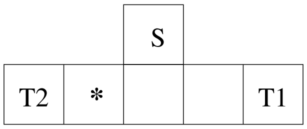

Figure 1: Gridworld for Question 5

6. Assume that a problem's dynamics is periodic, i.e., the problems dynamics changes every K steps and the cycle is repeated every $M * K$ steps. How will you design your problem representation so that the resulting dynamics are Markov? Ensure that any value function based solution method will work properly with this representation.

7. Q-learning is known as an off-policy method since it learns about the optimal policy while following an exploratory policy for generating trajectories. Can we make Q-Learning on-policy? If so, what is convergence time of on-policy variant and why is it so? Also, as described in Chapter 5 of the textbook, you can learn about any policy while following another policy. Can you learn the value function of an arbitrary policy while following an optimal policy? Explain your answer.

8. You receive the following letter:
   Dear Friend, Some time ago, I bought this old house, but found it to be haunted by

ghostly sardonic laughter. As a result it is hardly habitable. There is hope, however, for by actual testing I have found that this haunting is subject to certain laws, obscure but infallible, and that the laughter can be affected by my playing the organ or burning incense. In each minute, the laughter occurs or not, it shows no degree. What it will do during the ensuing minute depends, in the following exact way, on what has been happening during the preceding minute: Whenever there is laughter, it will continue in the succeeding minute unless I play the organ, in which case it will stop. But continuing to play the organ does not keep the house quiet. I notice, however, that whenever I burn incense when the house is quiet and do not play the organ it remains quiet for the next minute. At this minute of writing, the laughter is going on. Please tell me what manipulations of incense and organ I should make to get that house quiet, and to keep it so. Sincerely, At Wits End

(a) Formulate this problem as an MDP (for the sake of uniformity, formulate it as a continuing discounted problem, with = 0.9. Let the reward be +1 on any transition into the silent state, and -1 on any transition into the laughing state.) Explicitly give the state set, action sets, state transition, and reward function.

(b) Starting with policy π(laughing) = π(silent) = (incense, no organ). Perform a couple of policy iterations (by hand!) until you find an optimal policy. (Clearly show and label each step. If you are taking a lot of iterations, stop and reconsider your formulation!) Do a couple of value iterations as well.

(c) What are the resulting optimal state-action values for all state-action pairs?

(d) What is your advice to At Wits End ?