

---

## CS6700 : Reinforcement Learning

### Written Assignment #1

**Topics:** Intro, Bandits, MDP, Q-learning, SARSA, PG    **Deadline:** 20 March 2023, 23:55

**Name:** –Your name here–

**Roll number:** –Your roll no. here–

---

- 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.
  - Type your solutions in the provided L<sup>A</sup>T<sub>E</sub>X template file.
  - **Please start early.**
- 

1. (2 marks) [Bandit Question] Consider a N-armed slot machine task, where the rewards for each arm  $a_i$  are usually generated by a stationary distribution with mean  $Q^*(a_i)$ . The machine is under repair when a arm is pulled, a small fraction,  $\epsilon$ , of the times a random arm is activated. What is the expected payoff for pulling arm  $a_i$  in this faulty machine?

**Solution:**

2. (4 marks) [Delayed reward] 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) (2 marks) What is an appropriate notion of return for this task?

**Solution:**

- (b) (2 marks) Give the TD(0) backup equation for estimating the value function of a given policy.

**Solution:**

3. (5 marks) [Reward Shaping] Consider two finite MDPs  $M_1$ ,  $M_2$  having the same state set,  $S$ , the same action set,  $A$ , and respective optimal action-value functions  $Q_1^*$ ,  $Q_2^*$ . (For simplicity, assume all actions are possible in all states.) Suppose that the following is true for an arbitrary function  $f : S \rightarrow R$  :

$$Q_2^*(s, a) = Q_1^*(s, a) - f(s)$$

for all  $s \in S$  and  $a \in A$ .

- (a) (2 marks) Show mathematically that  $M_1$  and  $M_2$  has same optimal policies.

**Solution:**

- (b) (3 marks) Now assume that  $M_1$  and  $M_2$  has the same state transition probabilities but different reward functions. Let  $R_1(s, a, s')$  and  $R_2(s, a, s')$  give the expected immediate reward for the transition from  $s$  to  $s'$  under action  $a$  in  $M_1$  and  $M_2$ , respectively. Given the optimal state-action value functions are related as given above, what is the relationship between the functions  $R_1$  and  $R_2$  ? That is, what is  $R_1$  in terms of  $R_2$  and  $f$ ; OR  $R_2$  in terms of  $R_1$  and  $f$ .

**Solution:**

4. (10 marks) [Jack's Car Rental] Jack manages two locations for a nationwide car rental company. Each day, some number of customers arrive at each location to rent cars. If Jack has a car available, he rents it out and is credited \$ 10 by the national company. If he is out of cars at that location, then the business is lost. Cars become available for renting the day after they are returned. To help ensure that cars are available where they are needed, Jack can move them between the two locations overnight, at a cost of \$ 2 per car moved. We assume that the number of cars requested and returned at each location are Poisson random variables, meaning that the probability that the number  $n$  is  $\frac{\lambda^n}{n!}e^{-\lambda}$ , where  $\lambda$  is the expected number. Suppose  $\lambda$  is 3 and 4 for rental requests at the first and second locations and 3 and 2 for returns. To simplify the problem slightly, we assume that there can be no more than 20 cars at each location (any additional cars are returned to the nationwide company, and thus disappear from the problem) and a maximum of five cars can be moved from one location to the other in one night.

- (a) (4 marks) Formulate this as an MDP. What are the state and action sets? What is the reward function? Describe the transition probabilities (you can use a formula rather than a tabulation of them, but be as explicit as you can about the probabilities.) Give a definition of return and describe why it makes sense.

**Solution:**

- (b) (3 marks) One of Jack's employees at the first location rides a bus home each night and lives near the second location. She is happy to shuttle one car to the second location for free. Each additional car still costs \$ 2, as do all cars moved in the other direction. In addition, Jack has limited parking space at each location. If

more than 10 cars are kept overnight at a location (after any moving of cars), then an additional cost of \$ 4 must be incurred to use a second parking lot (independent of how many cars are kept there). These sorts of nonlinearities and arbitrary dynamics often occur in real problems and cannot easily be handled by optimization methods other than dynamic programming. Can you think of a way to incrementally change your MDP formulation above to account for these changes?

**Solution:**

- (c) (3 marks) Describe how the task of Jack's Car Rental could be reformulated in terms of *afterstates*. Why, in terms of this specific task, would such a reformulation be likely to speed convergence? (*Hint:- Refer page 136-137 in RL book 2nd edition. You can also refer to the video at <https://www.youtube.com/watch?v=w3wGvwi336I>*)

**Solution:**

5. (8 marks) [Organ Playing] 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) (4 marks) Formulate this problem as an MDP (for the sake of uniformity, formulate it as a continuing discounted problem, with  $\gamma = 0.9$ . 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.

**Solution:**

- (b) (2 marks) Starting with simple policy of **always** burning incense, and not playing organ, perform a couple of policy iterations.

**Solution:**

- (c) (2 marks) Finally, what is your advice to “At Wits End”?

**Solution:**

6. (4 marks) [Stochastic Gridworld] An  $\epsilon$ -greedy version of a policy means that with probability  $1-\epsilon$  we follow the policy action and for the rest we uniformly pick an action. Design a stochastic gridworld where a deterministic policy will produce the same trajectories as a  $\epsilon$ -greedy policy in a deterministic gridworld. In other words, for every trajectory under the same policy, the probability of seeing it in each of the worlds is the same. By the same policy I mean that in the stochastic gridworld, you have a deterministic policy and in the deterministic gridworld, you use the same policy, except for  $\epsilon$  fraction of the actions, which you choose uniformly randomly.

- (a) (2 marks) Give the complete specification of the world.

**Solution:**

- (b) (2 marks) Will SARSA on the two worlds converge to the same policy? Justify.

**Solution:**

7. (5 marks) [Contextual Bandits] Consider the standard multi class classification task (Here, the goal is to construct a function which, given a new data point, will correctly predict the class to which the new point belongs). Can we formulate this as contextual bandit problem (Multi armed Bandits with side information) instead of standard supervised learning setting? What are the pros/cons over the supervised learning method. Justify your answer. Also describe the complete Contextual Bandit formulation.

**Solution:**

8. (5 marks) [TD, MC, PG] Suppose that the system that you are trying to learn about (estimation or control) is not perfectly Markov. Comment on the suitability of using different solution approaches for such a task, namely, Temporal Difference learning, Monte Carlo methods, and Policy Gradient algorithms. Explicitly state any assumptions that you are making.

**Solution:**

9. (5 marks) [PG] Recent advances in computational learning theory, have led to the development of very powerful classification engines. One way to take advantage of these classifiers is to turn the reinforcement learning problem into a classification problem. Here the policy is treated as a labeling on the states and a suitable classifier is trained to learn the labels from a few samples. Once the policy is adequately represented, it can be then used in a policy evaluation stage. Can this method be considered a policy gradient method? Justify your answer. Describe a complete method that generates appropriate targets for the classifier.

**Solution:**