Assignment #1

Course: *Reinforcement Learning (CS6700)*

Instructor: *Prashanth L.A.*
Due date: *March 15th, 2021*

Instructions

1. Work on your own. You can discuss with your classmates on the problems, use books or web. However, the solutions that are submitted must be your own and you must acknowledge all the sources (names of people you worked with, books, webpages etc., including class notes.) Failure to do so will be considered cheating. Identical or similar write-ups will be considered cheating as well.

2. In your submission, add the following declaration at the outset:
   "I pledge that I have not copied or given any unauthorized assistance on this assignment."

3. The assignment has two parts. The first part involves theoretical exercises, while the second part requires programming. For the first part, write/typeset the solutions, and upload it on moodle. For the second part, you are required to submit your work in a separate interface (check the details in Section II below).

4. The submission deadline is final, and late submissions would not be considered.

I. Theory exercises

Problem 1.

Consider the finite horizon MDP setting, as formulated in Section 1.2 of the course notes. In place of the expected cost objective defined there, consider the following alternative cost objective for any policy π and initial state x_0:

$$J_\pi(x_0) = \mathbb{E}\left[\exp\left(g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, \mu_k(x_k), x_{k+1})\right)\right],$$

Answer the following: (2+1+2 marks)

(a) Show that an optimal cost and an optimal policy can be obtained by the following DP-algorithm variant:

$$\begin{aligned} J_N(x_N) &= \exp\left(g_N(x_N)\right), \\ J_k(x_k) &= \min_{a_k \in A(x_k)} \mathbb{E}_{x_{k+1}}\left(\exp\left(g_k(x_k, a_k, x_{k+1})\right) J_{k+1}(x_{k+1})\right). \end{aligned}$$

(b) Let $V_k(x_k) = \log J_k(x_k)$. Assume that the single stage cost g_k is a function of x_k and a_k only (and does not depend on x_{k+1}). Then, show that the DP algorithm, which is specified above, can be re-written as

$$\begin{aligned} V_N(x_N) &= g_N(x_N), \\ V_k(x_k) &= \min_{a_k \in A(x_k)} \left(g_k(x_k, a_k) + \log \mathbb{E}_{x_{k+1}}\left(\exp\left(V_{k+1}(x_{k+1})\right)\right)\right). \end{aligned}$$

(c) Recall the "oven problem" which was a linear system with quadratic cost problem over two stages. Using the notation from this problem, consider the following ‘exponentiated cost’ objective: For a given scalar θ, define

$$J_{a_0,a_1}(x_0) = \mathbb{E}\left[\exp\left(\theta\left(a_0^2 + a_1^2 + (x_2 - T)^2\right)\right)\right], \tag{1}$$

$$\text{where } x_{k+1} = (1-\alpha)x_k + \alpha a_k + w_k, \text{ for } k = 0, 1.$$

In the above, w_k is a Gaussian random variable with mean zero and variance σ^2.

Solve problem (1) using the DP algorithm from the part above. Identify the optimal policy, and the optimal expected ‘exponentiated’ cost.

Problem 2.

You are walking along a line of N stores in a shopping complex, looking to buy food before entering a movie hall at the end of the store line. Each store along the line has a probability p of providing the food you like. You cannot see what the next store (say $k+1$) offers, while you are at the kth store and once you pass store k, you cannot return to it. You can choose to buy at store k, if it has the food item you like and pay an amount $N-k$ (since you have to carry this item for a distance proportional to $N-k$). If you pass through all the stores without buying, then you have to pay $\frac{1}{1-p}$ at the entrance to the movie hall to get some food.

Answer the following: (2+2+3 marks)

(a) Formulate this problem as a finite horizon MDP.

(b) Write a DP algorithm for solving the problem.

(c) Characterize the optimal policy as best as you can. This may be done with or without the DP algorithm.

Problem 3.

A machine is either running or broken. If it runs for a week, it makes a profit of 1000 INR, while the profit is zero if it fails during the week. For a machine that is running at the start of a week, we could perform preventive maintenance at the cost of 200 INR, and the probability of machine failing during the week is 0.4. In case the preventive maintenance is not performed, the failure probability is 0.7. A machine that is broken at the start of a week can be repaired at a cost of 400 INR, in which case the machine would fail with probability 0.4, or replaced at a cost of 1500 INR. A replaced machine is guaranteed to run through its first week.

Answer the following: (2+3 marks)

(a) Formulate this problem as an MDP with the goal of maximizing profit for a given number of weeks.

(b) Find the optimal repair/replace/maintenance policy under a horizon of four weeks, assuming a new machine at the start of the first week.

Problem 4.

Suppose you want to travel from a start point S to a destination point D in minimum average time. You have the following route options:

1. A direct route that requires α time units.

2. A potential shortcut that requires β time units to get to a intermediate point I. From I, you can do one of the following: (i) go to D in c time units; or (ii) head back to S in β time units. The random variable c takes one of the values $c_1, \ldots, c_m$, with respective probabilities $p_1, \ldots, p_m$. The value of c changes each time you return to S, independent of the value in the previous time period.

Answer the following: (3+1+2+2 marks)

(a) Formulate the problem as an SSP problem. Write the Bellman's equation ($J = TJ$) and characterize the optimal policy as best as you can.

(b) Are all policies proper? If not, why does the Bellman equation hold?

(c) Solve the problem for $\alpha = 2, \beta = 1$, $c_1 = 0, c_2 = 5$, and $p_1 = p_2 = \frac{1}{2}$. Specify the optimal policy.

(d) Consider the following problem variant, where at I, you have the additional option of waiting for d time units. Each d time units, the value of c changes to one of $c_1, \ldots, c_m$ with probabilities as before, independently of the value at the previous time period. Each time c changes, you have the option of waiting extra d units, return to the start or go to the destination. Write down the Bellman equation and characterize the optimal policy.

II. Simulation exercises

The programming component of this assignment is available at `https://www.aicrowd.com/challenges/rliitm-1`. The total marks for this component is 25, and the grading will be done through the AIcrowd interface.

The students are required to do the following:

1. Click on participate in the link above, and enter you Roll no.

2. You only need to open the starter notebook in colab, and make a personal copy. You can work on everything in colab itself.

For submission status, check on the submissions tab.