Assignment #1

Course: *Reinforcement Learning (CS6700)*

Instructor: *Prashanth L.A.*

TAs: Nithia V, Mizhaan Maniyar, Akash Reddy, and Rebin Silva

Due date: *September 16th, 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. For the assignment problems, you could either write or typeset the solutions, and upload it on moodle.

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

Problem 1.

At each time period, an animal chooses in which of 3 patches of land to forage. In patch 1, the risk of predation is 0, the probability of finding food is 0.0, and its energy value is 0. In patch 2, the risk of predation is 0.004 in each period, the probability of finding food is 0.3, and the energy gain is 2 if food is found. In patch 3, the predation risk is 0.02, the probability of finding food is 0.5 and its energy gain is 3. Foraging in any patch uses 1 unit of energy reserve. Energy reserves below 2 indicate death and the animal's maximum energy capacity is 4 units.

Answer the following: (1+2 marks)

(a) Formulate this patch selection as a finite horizon MDP with the goal of maximizing probability of survival of the animal over 3 time periods.

(b) Solve this problem using the DP algorithm and write down the optimal policy for foraging.

Problem 2.

A farmer produces x_k units of crops. He/she stores $(1-u_k)x_k$ units of his crop production, where $0 \leq u_k \leq 1$, and invests the remaining $u_k x_k$ units to get more crops for the next year. The next year's production x_{k+1} is given by

$$x_{k+1} = x_k + w_k u_k x_k,\ k = 0, 1, ..., N-1$$

where scalars w_k are independent identical random variables with probability distribution that do not depend either on x_k or u_k. Furthermore, $\mathbb{E}\{w_k\} = \bar{w} > 0$. The problem is to find the optimal investment policy that maximize the total expected crops stored over N years

$$\mathbb{E}_{w_0, w_1, ..., w_{N-1}}\{x_N + \Sigma_{k=0}^{N-1}(1-u_k)x_k\}$$

Answer the following: (1.5+2.5 marks)

(a) Formulate this problem as an MDP.

(b) Characterize the optimal policy as best as you can.

Problem 3.

Consider an SSP with optimal expected cost $J^*(s)$, and an optimal policy π^*.

Answer the following: (2+2 marks)

(a) A new action a' becomes available in state s'. How can you determine whether π^* is still optimal in the modified SSP without re-solving the problem? If it is not, how can you find a new optimal policy?

(b) Suppose action a^* is optimal in state s^*, that is $\pi^*(s^*) = a^*$, and you find that the return in state s^* under action a^* decreases by Δ. Provide an efficient way for determining whether π^* is still optimal and, if not, for finding a new optimal policy and its corresponding expected cost.

Problem 4.

After a long disagreement, Kevin, Kanan, and Kannan agree to enter into a three-way duel. The three shooters, Kevin, Kanan, and Kannan have shooting accuracy probabilities $0 < \alpha < \beta < \gamma \leq 1$ respectively. Because of this disparity the shooters agree that Kevin shall shoot first, followed by Kanan, followed by Kannan, this sequential rotation continuing until only one man (the winner) remains standing. When all three men are standing, the active shooter must decide whom to shoot at, and he can only shoot at one opponent at a time. Kevin is allowed to miss intentionally (i.e., by shooting into the air). Kanan and Kannan are not allowed to miss intentionally. Every shooter wants to maximize his probability of winning the game. The winner get a reward of 1, and in all other cases the reward is 0.

Now, consider you are Kevin, and you want to maximize your probability of winning the game.

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

(a) Model your problem as a Markov decision process. Write down the state space, the action space, the transition probabilities, and the single stage rewards.

(b) Write down the Bellman optimality equations.

(c) Let $\alpha = 0.3, \beta = 0.5$, and $\gamma = 0.6$. Give the optimal values and the optimal policies for each state.

Problem 5.

Consider an SSP problem, where all policies to be proper, i.e., the Bellman optimality operator satisfies

$$||TJ - TJ'||_\xi \leq \rho||J - J'||_\xi, \forall J, J' \in \mathbb{R}^n,$$

for some $0 < \rho < 1$.

Consider the following variant of the value iteration algorithm, which is given an input parameter $\epsilon > 0$. This algorithm, say ϵ-VI, terminates after m-iterations, where m satisfies

$$||J_{m+1} - J_m||_\xi < \epsilon \left(\frac{1-\rho}{2\rho}\right).$$

In the above, $J_{k+1} = TJ_k$, $\forall k \geq 1$, and J_0 is initialized arbitrarily. Note that ϵ-VI algorithm assumes knowledge of ρ and ξ.

Answer the following: (3+2 marks)

(a) Recall that J_{m+1} is what we obtain after ϵ-VI terminates. Show that

$$||J_{m+1} - J^*||_\xi < \frac{\epsilon}{2}.$$

(b) Let π^ϵ be the policy that we obtain from J_{m+1} as follows:

$$T_{\pi^\epsilon} J_{m+1} = TJ_{m+1}.$$

Let J_{π^ϵ} denote the expected cost of the policy π^ϵ. Show that

$$||J_{\pi^\epsilon} - J^*||_\xi < \epsilon.$$