Assignment #2

Course: *Reinforcement Learning (CS6700)*

Instructor: *Prashanth L.A.*

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

Due date: *October 25th, 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 an MDP with a finite-horizon. For this problem, derive a policy iteration algorithm. In particular, provide the policy evaluation and policy improvement steps. Compare the computational requirements of policy iteration to that of DP algorithm, assuming horizon N, n states per stage, and m actions in each state. (5 marks)

Problem 2.

Consider an episodic MDP with a finite state space $\mathcal{X} \in \{1, \cdots, n\}$, and a special state 0 as the terminal state. Let $T \in \mathbb{N}$ be the random length of an episode. Consider a deterministic and bounded reward function $r : \mathcal{X} \to \mathbb{R}$, and assume zero reward at the terminal state. For each $x \in \mathcal{X}$, a fixed policy π determines a stochastic transition to a subsequent state $x' \in \{\mathcal{X} \cup 0\}$ with probability $\mathbb{P}(x' \mid x)$. Assume that the policy π is proper. We denote by x_t the state at time t, where $t = 0, 1, 2, \cdots$.
The cumulative discounted reward over an episode be $R \in \mathbb{R}$ is defined by

$$R = \sum_{t=0}^{T-1} \gamma^t r(x_t),\ \gamma \in (0,1).$$

Define the value function $J(\cdot)$, and the variance $V(\cdot)$ of the cumulative discounted reward as

$$J(x) = \mathbb{E}[R \mid x_0 = x], V(x) = \text{Var}[R \mid x_0 = x], \forall x \in \mathcal{X}.$$

Also define the second moment $M(\cdot)$ of the cumulative discounted reward as

$$M(x) = \mathbb{E}[R^2 \mid x_0 = x], \forall x \in \mathcal{X}.$$

Answer the following: (3+2 marks)

(a) $\forall x \in X$, express $J(x)$ and $M(x)$ in terms of Bellman equation (exclude the terminal state 0).

(b) Show that

$$\forall x \in X, V(x) = \psi(x) + \gamma^2 \sum_{x' \in \mathcal{X}} \mathbb{P}(x' \mid x) V(x'), \text{ where}$$

$$\psi(x) = \gamma^2 \left(\sum_{x' \in \mathcal{X}} \mathbb{P}(x' \mid x) J(x')^2 - \left(\sum_{x' \in \mathcal{X}} \mathbb{P}(x' \mid x) J(x') \right)^2 \right).$$

Problem 3.

Let $\mathbb{R}^d$ be a d-dimensional Euclidean space with supremum norm $\|x\|_\infty = \max_{1\le i\le d} |x_i|$, and with ordering $x \preceq y$ if $x_i \le y_i$, for all $1 \le i \le d$. Let I_d be a d-dimensional vector with all elements equal to one.

Prove the following: (1+1+3 marks)

(a) Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a monotone contraction mapping with respect to the supremum norm with contraction factor β, and let a be a scalar. Then, $x \preceq y + aI_d$ implies $f(x) \preceq f(y) + \beta|a|I_d$.

(b) Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a mapping with property $x \preceq y + aI_d$ implies $f(x) \preceq f(y) + \beta|a|I_d$, for all scalar a, and for some $0 \le \beta < 1$. Then with respect to the supremum norm, f is a monotone contraction with contraction factor β.

(c) Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a monotone contraction mapping with respect to the supremum norm with contraction factor β, and fixed point x^*. Then

$$x - \frac{1}{1-\beta}\|f(x) - x\|_\infty I_d \preceq f(x) - \frac{\beta}{1-\beta}\|f(x) - x\|_\infty I_d \preceq x^* \preceq f(x) + \frac{\beta}{1-\beta}\|f(x) - x\|_\infty I_d \preceq x + \frac{1}{1-\beta}\|f(x) - x\|_\infty I_d.$$

II. Simulation exercises

The programming component of this assignment is available at AIcrowd platform:
`https://www.aicrowd.com/challenges/iit-m-2021-assignment-2`.
The total marks for this component is 20, and the grading will be done through the AIcrowd interface.

The instructions on how to use the AIcrowd interface are available at
`https://wiki.aicrowd.com/share/2e92c0cb-870a-4d56-b8dc-959a9723da54`.