The University of Texas at Austin
Department of Electrical and Computer Engineering

EE381V: Learning-Based Optimal Control, Fall 2022

HOMEWORK 3

In this problem set we wish to move an agent in a gridworld to its goal location. The gridworld is represented as an $n \times n$ grid, i.e., the state space is

$$\mathcal{S} = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1, x_2 \in \{0, 1, ..., n-1\}\}$$

In these coordinates, $(0, 0)$ represents the bottom left corner of the map and $(n-1, n-1)$ represents the top right corner of the map. From any location $x = (x_1, x_2) \in \mathcal{S}$, the agent has four possible directions it can move in, i.e.,

$$\mathcal{A} = \{\texttt{up}, \texttt{down}, \texttt{left}, \texttt{right}\}.$$

The corresponding state changes for each action are:

- `up`: $(x_1, x_2) \mapsto (x_1, x_2 + 1)$
- `down`: $(x_1, x_2) \mapsto (x_1, x_2 - 1)$
- `left`: $(x_1, x_2) \mapsto (x_1 - 1, x_2)$
- `right`: $(x_1, x_2) \mapsto (x_1 + 1, x_2)$

Additionally there is stochasticity in the agent's dynamics. Given its current state x and action a, the agent's next state is determined as follows:

- With probability 0.4, the agent moves in a uniformly random direction.
- With probability 0.6, the agent will move in the direction specified by the action.
- If the resulting movement would cause the agent to leave $\mathcal{S}$, then it will not move at all. For example, if the agent is on the right boundary of the map, then moving right will do nothing.

The agent's objective is to reach $x_{\text{goal}} \in \mathcal{S}$, so the reward function is the indicator function $R(x) = I_{x_{\text{goal}}}(x)$. In other words, the agent will receive a reward of 1 if it reaches the $x_{\text{goal}} \in \mathcal{S}$, and a reward of 0 otherwise. The reward of a trajectory in this infinite horizon problem is a discounted sum of the rewards earned in each timestep, with discount factor $\gamma \in (0, 1)$.

Let $n = 20$, $\gamma = 0.95$ and $x_{\text{goal}} = (19, 9)$. Find the optimal value function for this problem and plot a heatmap of the optimal value function over the grid $\mathcal{S}$, with $x = (0, 0)$ in the bottom left corner, $x = (n-1, n-1)$ in the top right corner, the x_1-axis along the bottom edge, and the x_2-axis along the left edge.

Use the optimal value function to compute an optimal policy. Then, use this policy to simulate the MDP starting at $x = (0, 0)$ and proceeding till goal state is reached. Plot the policy as a heatmap where the actions $\{\texttt{up}, \texttt{down}, \texttt{left}, \texttt{right}\}$ correspond to the values $\{0, 1, 2, 3\}$, respectively. Plot the simulated agent trajectory overlaid on the policy heatmap, and briefly describe in words what the policy is doing.

Code the following two algorithms to compute the optimal value function and report the number of sweeps of the state space required for convergence in both cases.

1. Policy iteration.
2. Value iteration.