(CSC 581D/ 485D: Topics in Artificial Intelligence: Reinforcement Learning (Spring 2022, Assignment 2)

Instructor: Kui Wu

Due: 23:55 pm, Feb. 18, 2022

Each question has two marks: the first one for CSC 581D students, and the second one for CSC 485D students. A zero mark means optional.

1. (30, 30%) Consider an undiscounted Markov Reward Process with two states A and B. The transition matrix and reward function are unknown, but you have observed two sample episodes:

   $A + 3 \rightarrow A + 2 \rightarrow B - 3 \rightarrow A + 4 \rightarrow B - 4 \rightarrow$ terminate

   $B - 2 \rightarrow A + 3 \rightarrow B - 4 \rightarrow$ terminate

   In the above episodes, sample state transitions and sample rewards are shown at each step, e.g., $A + 3 \rightarrow A$ indicates a transition from state A to state A, with a reward of $+3$.

   (a) (5, 10%) Using first-visit Monte-Carlo evaluation, estimate the state-value function $V(A), V(B)$.

   (b) (5, 0%) Using every-visit Monte-Carlo evaluation, estimate the state-value function $V(A), V(B)$.

   (c) (5, 10%) Draw a diagram of the Markov Reward Process that best explains these two episodes (i.e. the model that maximises the likelihood of the data - although it is not necessary to prove this fact). Show rewards and transition probabilities on your diagram.

   (d) (5, 10%) Define the Bellman equation for your above Markov reward process. Solve the Bellman equation **directly**, rather than iteratively, to find the true state-value function $V(A), V(B)$

   (e) (10, 0%) What value function would batch TD(0) find, i.e., if TD(0) was applied repeatedly to these two episodes?

2. (70, 70%) Assume that you are required to develop an RL-based solution to driving a car over a slippery grid. Due to the slippery surface, the vehicle exhibits some stochastic behaviour as depicted in Fig. 1.

   Your goal is to drive a loop through the grid shown in Fig. 2. The car starts in the top left corner ($x = 0, y = 0$) facing South and receives a reward of 10 and terminates a training episode if it loops counterclockwise around the grip map and transitions from cell position

$(x = 1, y = 0)$ to the starting cell. The agent's absolute orientations and action set are encoded in `CarMDP` as:

```
orientations = 0: 'North', 1: 'East', 2: 'South', 3: 'West'
A = 0: 'Forward', 1: 'Left', 2: 'Right', 3: 'Brake'
```

The agent's complete state at each time step is encoded as the 3-element tuple `(x, y, Orientation)`. For example, the starting state S_0 is encoded `(0, 0, 2)`. Cells with obstacles are drawn in black in Fig. 2, and, along with cells outside of the fixed grid map, represent terminal "crash" states that reward a penalty of -5. Any other non-terminating action receives a penalty of -0.01. You must develop a **model-based** reinforcement learning algorithm to find the optimal policy. Specifically, we assume that the agent is aware of the model (e.g., the transition dynamics, the size of the state space, the rewards for actions, the obstacles). You need to (1) implement Iterative Policy Evaluation to find the state values for the Random policy (i.e., the policy that takes each action uniformly random among the four possible actions). Then, starting with the state values of the Random policy, you (2) implement the Policy Iteration method to find the optimal policy.

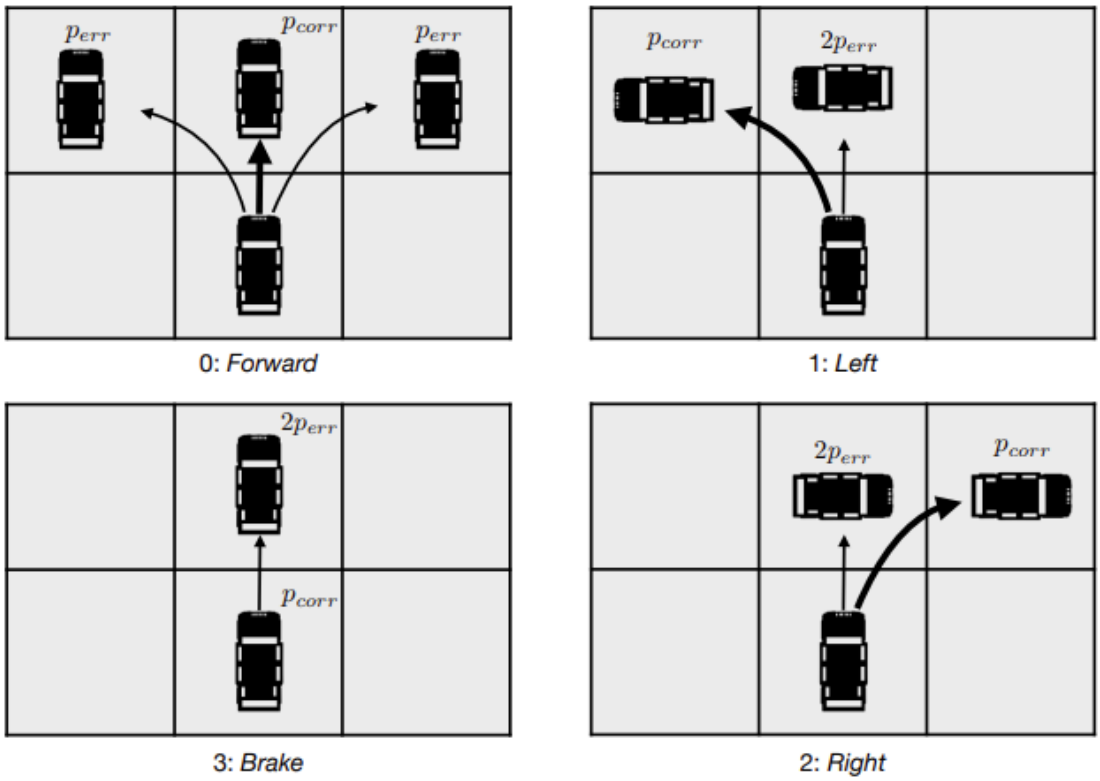

Figure 1: Probabilistic transition dynamics of the car MDP for each of its four actions (Forward, Left, Right, and Brake). The diagrams show all possible outcomes (i.e., $P_{corr} + 2P_{err} = 1$). Each action's effect is relative to the car's orientation: in the four figures shown here, the car is in the North orientation. For example, the figures would simply be rotated 90 degrees clockwise if the car were in the East orientation.

Note: `CarPilotEnv-A2.ipynb` is provided to help you finish the assignment. You need to implement the learning algorithm in `ModelBasedRLAgent`.

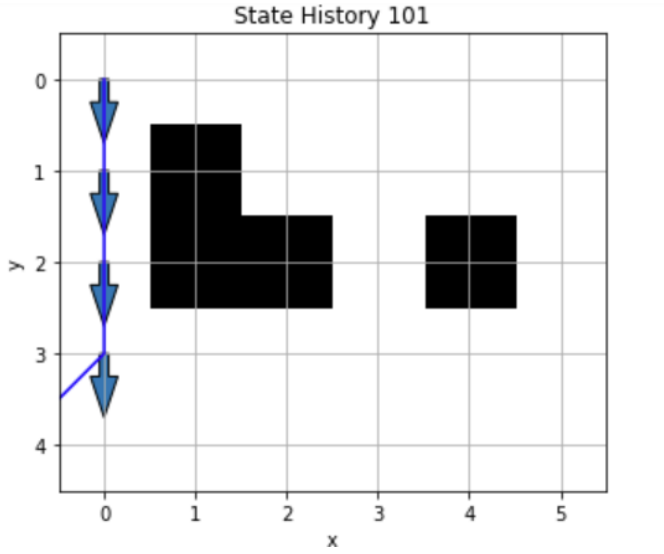

Figure 2: A sample path of the agent using the random policy.

Warning: The assignment is non-trivial. Start your assignment as early as possible!

Deliverables: Zip your answers, including your answer of Q1, a Readme file, and your jupyter notebook (following the Python code template provided with this Assignment), in one file and submit the zipped file to Brightspace. The Readme file should tell TA (1) your session (CSC 581D or 485D) and (2) how to run your Python code on the jupyter notebook.

The End
