Robot Learning	**Assignment 1**
CS 4756 Spring 2024	**Due 11:59pm Sunday, February 18**

(1) Exploring Markov Decision Processes (5 points)

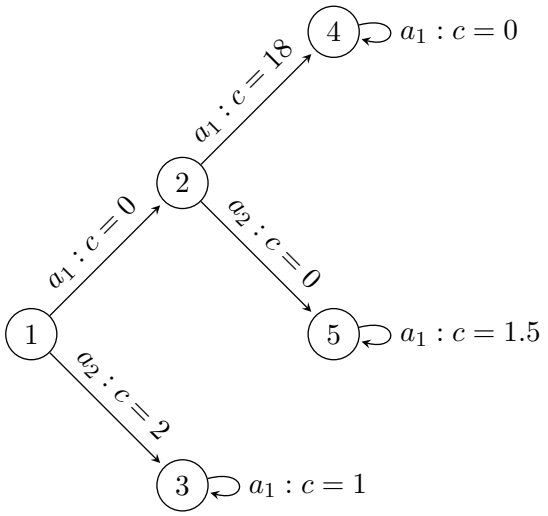


Figure 1: MDP for Problem 1

Compute the optimal value function V^* and the corresponding optimal policy π^* for each state in Fig. 1 for a discount factor of $\gamma = 0.9$ in the infinite horizon setting.

Notes:

- Initial State is always State 1.
- Each edge of the MDP is labeled in the following format: "{action} : {cost of action to complete transition}". Thus, the problem formulation involves a minimization of cost, rather than a maximization of a reward as may be seen elsewhere.
- Action a_1 at states 3, 4, and 5 must be taken infinitely if those states are ever reached.

(2) A Story of Three Cliffs: Behavior Cloning and DAgger (10 points)

In this question, we are going to be thinking about robots falling off of cliffs and trying to get back on. We will look at three types of cliffs with varying levels of difficulty, and compare how well behavior cloning (BC) performs with respect to DAgger.

In all of the following parts, we will consider an infinite horizon setting with a discount factor of γ. Each part considers a Cliff MDP, where there exists a path atop the cliff consisting of "safe" states as well as a path to fall off the cliff from any point and land at the bottom, which we denote at s_x.

For each variant of the MDP, compute the tightest possible upper bounds (in terms of big-O) for the following quantities:

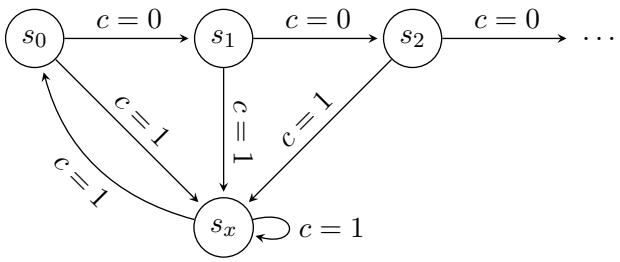
- $J(\pi_{BC})$: Expected total discounted sum of costs for Behavior Cloning
- $J(\pi_{DAgger})$: Expected total discounted sum of costs for DAgger

Important notes and assumptions:

- An agent begins at state s_0
- The expert will always follow an optimal trajectory, thus the expert policy will incur 0 cost
- At each state the learner (both BC and DAgger) visits that the expert has also visited (i.e. all the safe states), it will make a mistake with probability ϵ and fall off the cliff.
- Once the the BC learner has reached an unknown state, in the worst case with probability 1 it will continue to make mistakes and stay at the bottom of the cliff
- In contrast, the DAgger learner will query the expert to determine its next action, being able to complete a recovery action with probability $1 - \epsilon$ (if it exists)
- Assume that $0 < \epsilon \ll 1$, and $0 < 1 - \gamma \ll 1$ to simplify your calculations.

Hint: Try formulating two mutually recursive equations for $J_{\text{cliff}}(\pi)$ and $J_{\text{ditch}}(\pi)$, the expected total discounted sum of costs when the learner either starts on the cliff or starts in the ditch, respectively. This can help avoid any infinite sums.

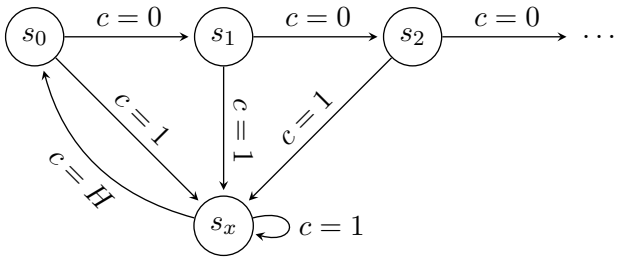
(a) Cliff-Easy



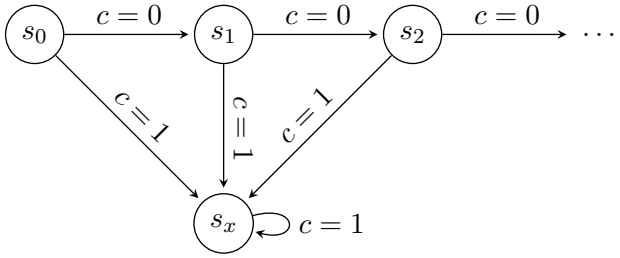
Cliff-Easy: The safe states s_i can either transition to s_{i+1} with $c = 0$ or s_x with $c = 1$, but there exists a recovery action from s_x to return to s_0 with $c = 1$. Bounds should be in terms of ϵ, γ.

(b) Cliff-Medium

Cliff-Medium: The safe states s_i can either transition to s_{i+1} with $c = 0$ or s_x with $c = 1$, but there exists a recovery action from s_x to return to s_0 with $c = H$. Bounds should be in terms of ϵ, γ, H.



(c) Cliff-Hard

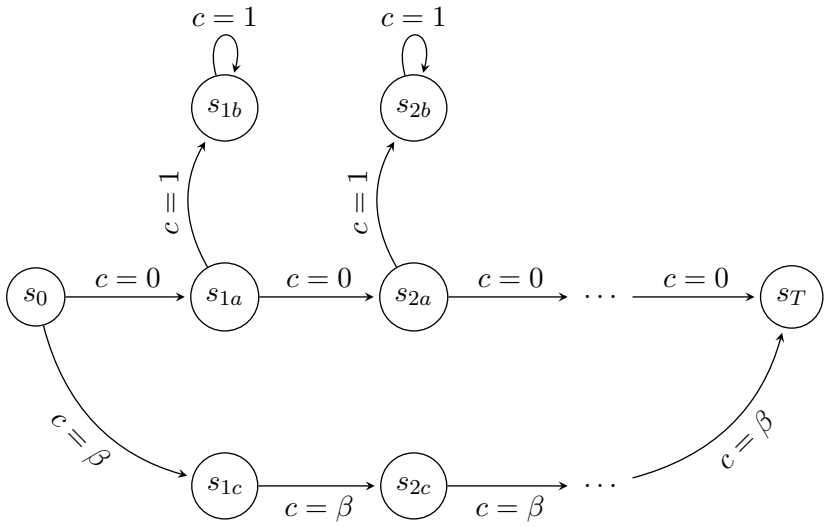


Cliff-Hard: The safe states s_i can either transition to s_{i+1} with $c = 0$ or s_x with $c = 1$, but there is NO recovery action at s_x. Bounds should be in terms of ϵ, γ.

Finally, comment on the following: How does the gap between DAgger and BC change as we vary the H in cliff-medium from 1 to $\frac{1}{1-\gamma}$? What is the explanation for this trend?

(3) Pitfalls of DAgger (5 points)

In this problem, we explore the performance bound of a policy learned by DAgger in an interesting MDP with multiple routes to a final goal state.



Two ways to the goal

The MDP has two routes from the start s_0 to the goal s_T:

1. One route is the optimal route with action cost 0 that the expert takes. There is, however, a caveat. The route passes over a bridge $(s_{1a}, s_{2a}, \dots,)$ where a mistake can result in falling off the path into a ditch $(s_{1b}, s_{2b}, \dots)$. Once in the ditch, there is no recovery.
2. One route is a long detour with action cost β.

Assume the following about our learner:

- On states on the bridge, $(s_{1a}, s_{2a}, \dots,)$, the learner makes a mistake with probability ϵ.
- On all other states, it can perfectly imitate the expert.

Compute the upper bound on $J(\pi_{DAgger})$, the expected total cost of trajectories for a DAgger policy over T timesteps. Now, compute the cost of the best policy in the learner's class, $J(\pi_L^\star)$.

- What is surprising about DAgger's performance specifically from this MDP based on this bound? Does DAgger find the best policy in the learner's policy class?
- In a few sentences, can you describe a real-world example that has a similar characteristic as this MDP?