Robot Learning	**Assignment 1**
CS 4756 Spring 2025	**Due 11:59pm Friday, February 14**

(1) Cliff Recovery: 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 compare how well behavior cloning (BC) performs with respect to DAgger.

We will consider an infinite horizon setting with a discount factor of γ. We consider 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.

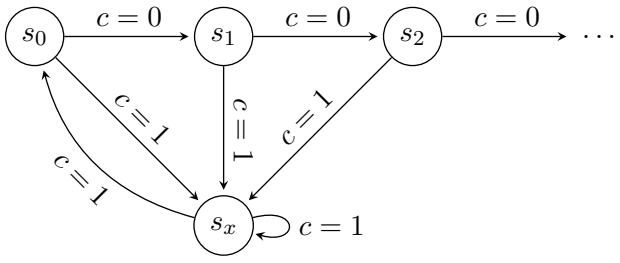
Express the following quantities in terms of ϵ and γ.

- $J(\pi_{BC})$: 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).
- To simplify your calculations, look for terms you can neglect. For example, we assume that $0 < \epsilon \ll 1$, and $0 < 1 - \gamma \ll 1$, so we can neglect any products of these two very small terms.
- This problem is in the infinite horizon setting.

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.



Cliff Recovery: 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 $c = 1$. Answers should be in terms of ϵ, γ.

(2) Exploring Markov Decision Processes (5 points)

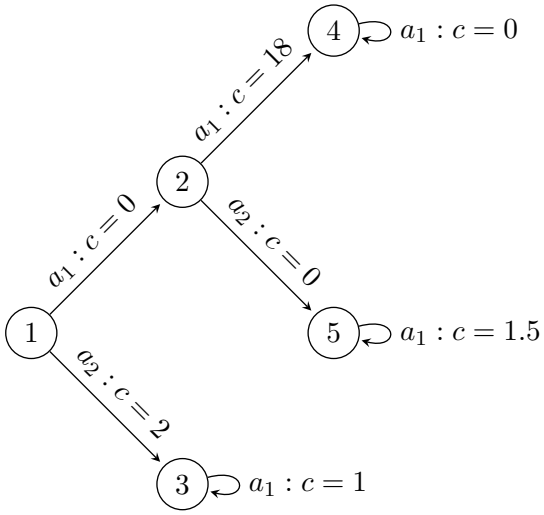


Figure 1: MDP for Problem 2

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.