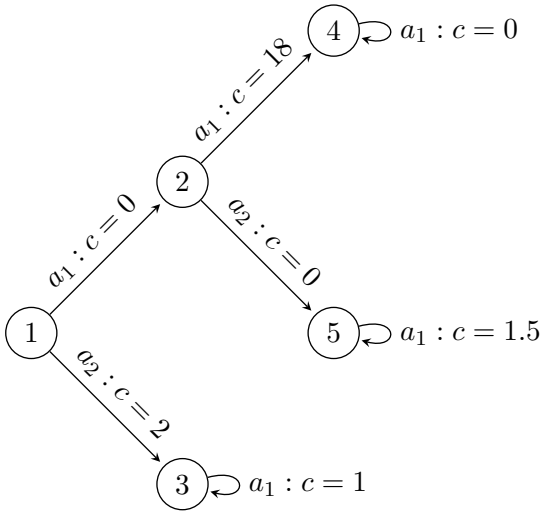
| **Robot Learning** | **Assignment 1** |
|---|---|
| **CS 4756 Spring 2024** | **Due 11:59pm Thursday, February 15** |

**(1) Exploring Markov Decision Processes (?? points)**




Compute the optimal Value Function $V^*$ and the corresponding Optimal Policy $\pi^*$ for each state in Figure 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) Behavior Cloning and DAgger on Cliff MDP (?? points)**

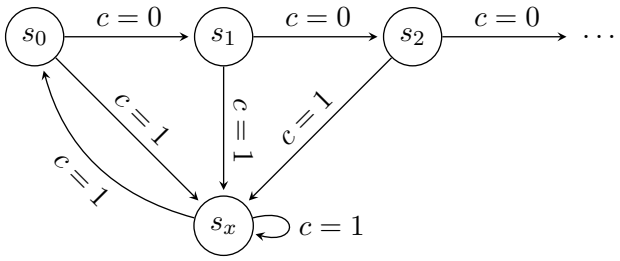
In all of the following parts, we will consider a finite horizon setting with $T$ timesteps. Each part considers a different degree of Cliff MDP, where there exists a path atop the cliff consisting of "safe" states as well as a path to fall off the cliff from any point and land at the bottom, which we 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 cost of trajectories for a Behavior Cloning policy over $T$ timesteps
- $J(\pi_{DAgger})$: Expected total cost of trajectories for a DAgger policy over $T$ timesteps

Important notes and assumptions:

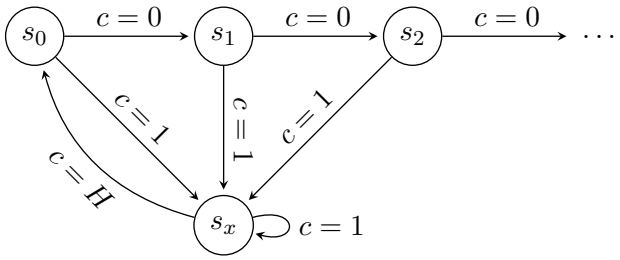
- The expert will always follow an optimal trajectory, thus the expert policy will incur 0 cost
- At each state the learner (both BC and DAgger) visits that the expert has also visited (i.e. all the safe states), it will make a mistake with probability $\epsilon$ and fall off the cliff
- Once the the 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, potentially being able to complete a recovery action

**(a) Cliff-Easy**



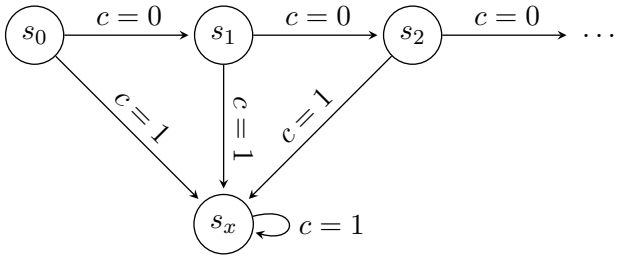

Cliff-Easy: The safe states $s_i$ can either transition to $s_{i+1}$ with $c = 0$ or $s_x$ with $c = 1$, but there exists a recovery action from $s_x$ to return to $s_0$ with $c = 1$. Bounds should be in terms of $\epsilon, T$.

**(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 $\epsilon, T, H$.

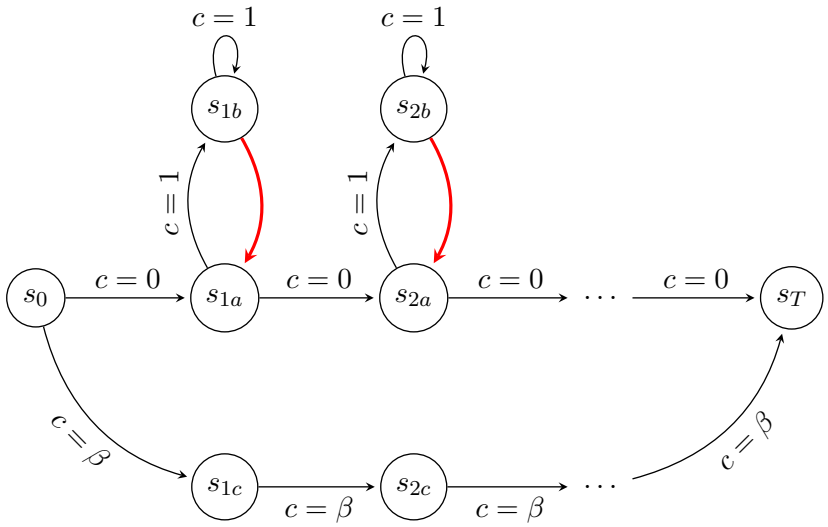
**(c) Cliff-Hard**




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 $\epsilon, T$.

**(3) [Extra Credit] Pitfalls of DAgger (?? 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.




This MDP has one optimal route (which the expert always takes) and one suboptimal route to the final state $s_T$, differing by potential transitions to costly states off of the optimal route. The main caveat in this problem is that the recovery action (highlighted by the thick red edges returning from all the intermediate $b$ states to the $a$ route are NOT REALIZABLE by the learner, meaning the learner's policy class is unable to take that action. Therefore, if the learner ever ends up falling off of the $a$ route, it can never recover, while the expert is capable of recovering for $c = 0$ (omitted from the figure as we only care about the performance of the learner).

Compute the upper bound on $J(\pi_{DAgger})$, the expected total cost of trajectories for a DAgger policy over $T$ timesteps, again assuming that the learner has an $\epsilon$ probability of deviating from the expert's optimal policy at any state the expert has visited, but you may assume that the learner follows the expert with probability 1 on the first step. What is surprising about DAgger's performance specifically from this MDP based on this bound?