1 Value Iteration

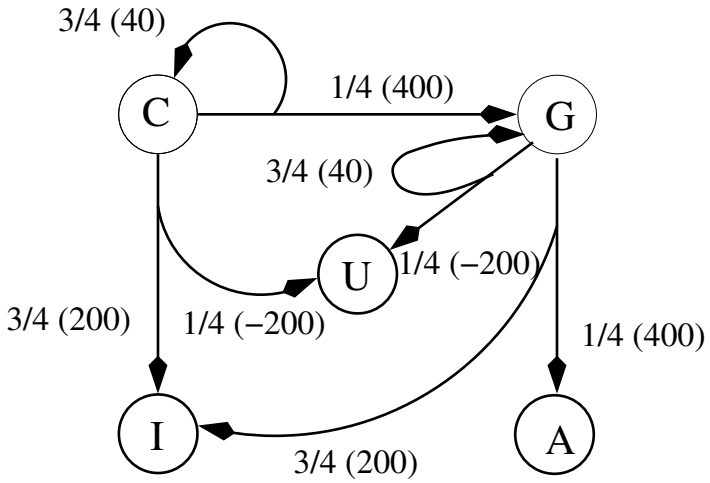
In the MDP below, there are 5 states: C(ollege), G(rad school), I(ndustry), A(cademia), and U(nemployed). States I, A and U are terminal states. Probabilities of transitions are either 1/4 or 3/4, and the values in parentheses are the rewards for that transition. The possible actions from states C and G are:

- State C: You may choose to stay in C, but with a probability of 1/4 you may end up going to state G.

  You may also choose to go to state I, but with probability 1/4 you end up in state U.

- State G: You may choose to stay in state G, but with probability 1/4 you end up in state U.

  You may also choose to go to state A, but with probability 3/4 you end up in state I.



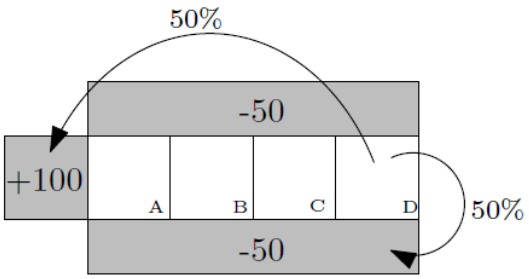
1. (6 pts) You start in state C. Perform two iterations of value iteration, where you first compute the Q values and then take the maximum of the Q values. The discount is $\gamma = 1$.

2. (4 pts) Perform policy extraction after these two iterations to find $\pi^*(s)$. Please show all work.

---

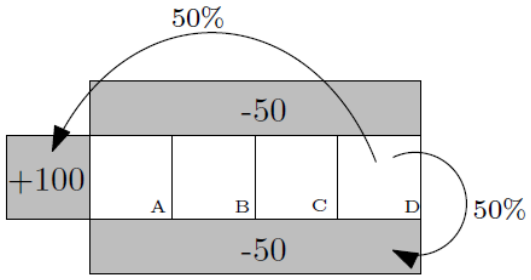
2 Policy Iteration

Consider the MDP shown below. There are three terminal states (the shaded ones) at which the agent immediately receives a reward of either 100 or -50 upon transitioning into this state. Since these states are terminal, the game "ends" once the agent reaches one of those states. There are four non-terminal states, marked A-D. At any state the agent has the option of going left or right. 80% of the time the agent moves correctly, 20% of the time, the agents slips off the bridge into the "lava" (the -50 parts). The only exception is "magic state D." In magic state D, no matter what action the agent choses, he is transitioned with probability 50% to the "good state" and with probability 50% to the lava. The rewards going between states A-D is zero, and the discount is 0.5.



1. Suppose the agent lives according to the policy $\pi = L$ (always go left). Perform two iterations of policy evaluation for each state by filling in the following table. Besides the table, show your work. Remember that the value of a terminal state is always zero.

i	$V_i^\pi(A)$	$V_i^\pi(B)$	$V_i^\pi(C)$	$V_i^\pi(D)$
0	0	0	0	0
1				
2				



2. Suppose that this agent were to execute a single policy update based on the above values. What would be the new policy (fill it in below)?

$\pi(A) =$
$\pi(B) =$
$\pi(C) =$
$\pi(D) =$

3 (CS 6300 only) Policy Evaluation as a Linear System

Consider again the following MDP. There are 5 states: C(ollege), G(rad school), I(ndustry), A(cademia), and U(nemployed). States I, A and U are terminal states. Probabilities of transitions are either 1/4 or 3/4, and the values in parentheses are the rewards for that transition. The possible actions from states C and G are:

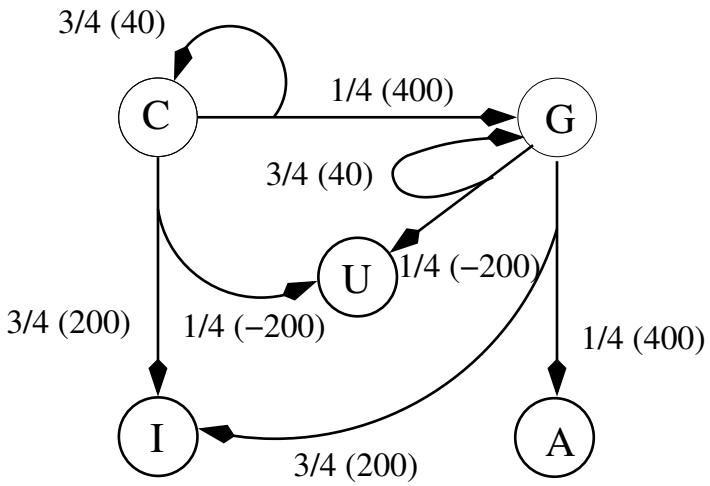
- State C: You may choose to stay s in C, but with a probability of 1/4 you may end up going g to state G.

  You may also choose to go g to state I, but with probability 1/4 you end up in state U.

- State G: You may choose to stay s in state G, but with probability 1/4 you end up in state U.

  You may also choose to go g to state A, but with probability 3/4 you end up in state I.

You start in state C. Assume your initial policy is $\pi_0(s) = s$, i.e., you wish to stay in the current state you're in. Also the discount is $\gamma = 1$.



1. Perform policy evaluation exactly by solve for the utility values $V^{\pi_0}(C)$ and $V^{\pi_0}(G)$. [Hint: remember that the utility values can be solved for analytically by writing out a system of equations with n equations and n unknowns. For this problem you just need to solve for 2 unknowns.]

   Solution:

2. You will now derive a general closed form for solving policy evaluation. Consider the value iteration equation given π:

$$V^{\pi}(s) = \sum_{s'} T(s, \pi(s), s') \left[R(s, \pi(s), s') + \gamma V^{\pi}(s') \right]$$

Write out a closed form solution for V^{π}, where V^{π} is a vector where the ith component is equal to $V^{\pi}(i)$ for state i. [Hint: You will want to first write out the linear system in matrix form. You should use the transition matrix T^{π}, where the element in the ith row and jth column, $T^{\pi}[i, j]$, denotes the probability of transitioning from state i to state j when following policy π, e.g., $T^{\pi}[i, j] = T(i, \pi(i), j) = P(s' = j | s = i, a = \pi(s))$. It will be easier if you also defining a vector of reward values, call them $\bar{R}$, where $\bar{R}[s] = \sum_{s'} T(s, \pi(s), s') R(s, \pi(s), s')$.]

Solution: