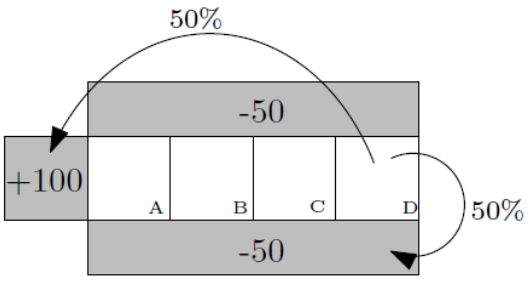
1 Monte Carlo Policy Evaluation

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 $\pi = R$ (always go right). Rather than using dynamic programming, you decide to just run the policy multiple times to get a Monte Carlo estimate of the value of this policy. You run two rollouts (episodes) shown below. Each episode is a sequence of (state, action, reward, next-state, next-action, next-reward, etc.).

   Episode 1: *(A, right, 0, B, right, 0, C, right, 0, D, right, -50, done)*

   Episode 2: *(A, right, 0, B, right, 0, C, right, 0, D, right, +100, done)*

   Using $\gamma = 0.5$, compute the Monte Carlo estimates (average over rollouts) of the values of each state below:

$V^\pi(A)$	$V^\pi(B)$	$V^\pi(C)$	$V^\pi(D)$

Solution

$V^\pi(A)$	$V^\pi(B)$	$V^\pi(C)$	$V^\pi(D)$
3.125	6.25	12.5	25

See slides for algorithm. Sutton and Barto Chapter 5 also has lots of details if you want to dig deeper.

We accumulate discounted sums of rewards for the first occurrence of each state in each trajectory. So for state A, we have a value estimate of $0 + \gamma \cdot 0 + \gamma^2 \cdot 0 - \gamma^3 \cdot 50$ from episode 1 and a value estimate of $0 + \gamma \cdot 0 + \gamma^2 \cdot 0 + \gamma^3 \cdot 100$ from episode 2. Similarly, we can do the same thing for state B and compute a value estimate of $0 + \gamma \cdot 0 - \gamma^2 \cdot 50$ from episode 1, etc. We do this for each state and compute the Monte Carlo estimate of the values by averaging the sampled values across episodes. Below we show some book keeping to make the steps clear of how we got the above solution:

	$V^\pi(A)$	$V^\pi(B)$	$V^\pi(C)$	$V^\pi(D)$
return estimate from episode 1	-50/8	-50/4	-50/2	-50
return estimate from episode 2	100/8	100/4	100/2	100
average over episodes (MC estimate)	(-50/8+100/8)/2	(-50/4+100/4)/2	(-50/2+100/2)/2	(-50+100)/2