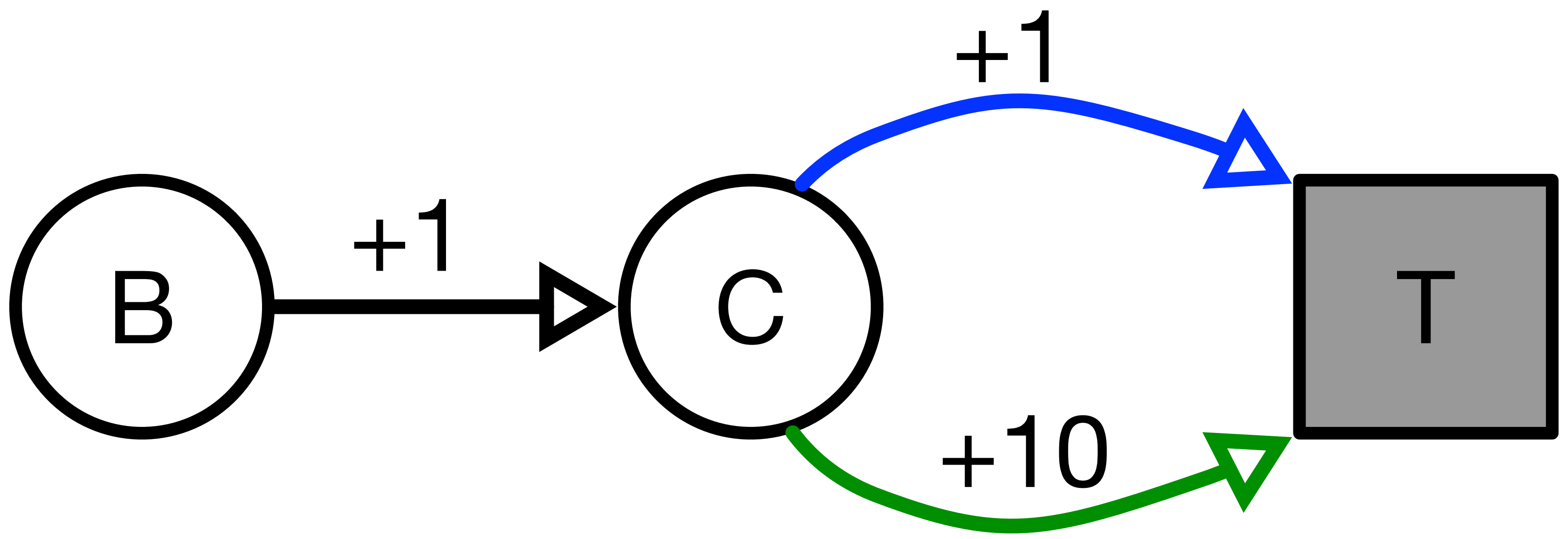
Course 2, Module 1
Monte Carlo

CMPUT 397
Fall 2019

Challenge Questions

- **Challenge Question 1**: How do we handle **continuing tasks** with MC?
- **Challenge Question 2:**
  - When we had the model, we used DP to find the value function vpi.
  - Without the model, we use sampled experience from the environment, and do Monte Carlo updates
  - **Can you use Monte Carlo updating, if you have a model?** If so, how? Is there more than one way?

Worksheet Q6

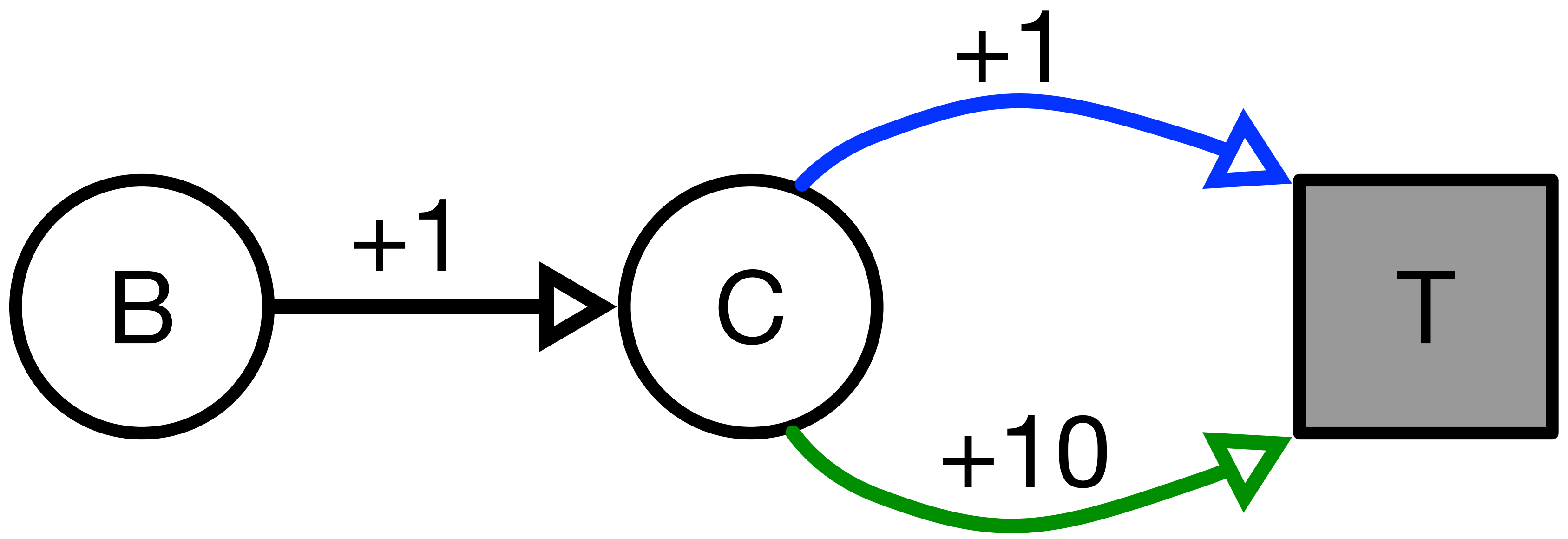


5. Off-policy Monte Carlo prediction allows us to use sample trajectories to estimate the value function for a policy that may be different than the one used to generate the data. Consider the following MDP, with two states B and C, with 1 action in state B and two actions in state C, with $\gamma = 1.0$. Assume the target policy π has $\pi(A = 1|B) = 0.9$ and $\pi(A = 2|B) = 0.1$, and that the behaviour policy b has $b(A = 1|B) = 0.25$ and $b(A = 2|B) = 0.75$.

(a) What are the true values v_π?

(b) Imagine you got to execute π in the environment for one episode, and observed the episode trajectory $S_0 = B, A_0 = 1, R_1 = 1, S_1 = C, A_1 = 1, R_2 = 1$. What is the return for B for this episode? Additionally, what are the value estimates V_π, using this one episode with Monte Carlo updates?

Worksheet Q6



(c) But, you do not actually get to execute π; the agent follows the behaviour policy b. Instead, you get one episode when following b, and observed the episode trajectory $S_0 = B, A_0 = 1, R_1 = 1, S_1 = C, A_1 = 2, R_2 = 10$. What is the return for B for this episode? Notice that this is a return for the behaviour policy, and using it with Monte Carlo updates (without importance sampling ratios) would give you value estimates for b.

(d) But, we do not actually want to estimate the values for behaviour b, we want to estimates the values for π. So, we need to use importance sampling ratios for this return. What is the return for B using this episode, but now with importance sampling ratios? Additionally, what is the resulting value estimate for V_π using this return?

Q1: The pseudocode for Monte Carlo is inefficient because, for each state, it maintains a list of all returns and repeatedly calculates their mean. How can we modify the algorithm to have incremental updates for each state?

Input: a policy π to be evaluated

Initialize:

$V(s) \in \mathbb{R}$**, arbitrarily, for all** $s \in S$

$Returns(s) \leftarrow$ **an empty list, for all** $s \in S$

Loop forever (for each episode):

Generate an episode following $\pi : S_0, A_0, R_1, S_1 \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode, $t = T-1, T-2, \ldots, 0$

$G \leftarrow \gamma G + R_{t+1}$

Append G **to** $Returns(S_t)$

$V(S_t) \leftarrow$ **average**$(Returns(S_t))$