Mini-Course 2, Module 1
Monte Carlo Methods for Prediction & Control

CMPUT 365
Fall 2021

October 8, 2021

- NO CLASS MONDAY
- Let’s chat about practice quizzes!
  - **I am increasing the number of attempts for practice quizzes to 10 attempts**
- **FIRE Drill**
- Any questions about course admin?

Terminology Review

- In Monte Carlo there are **no models, and no bootstrapping**
- **Experience**: data generate by the agent taking actions and getting reward feedback for the action it selected.
  - different from what Dynamic Programming does. DP updates the value of states using p(s',r|s,a). DP knows all the rewards in each state via p
- **Sample episodes**: starting in the start state, run policy pi (select actions according to pi) until termination, recording the states, actions, and rewards observed
- MC methods update the value estimates on an **episode-by-episode** basis. Must wait until the end of an episode to update the values of each state the agent observed

Terminology Review (2)

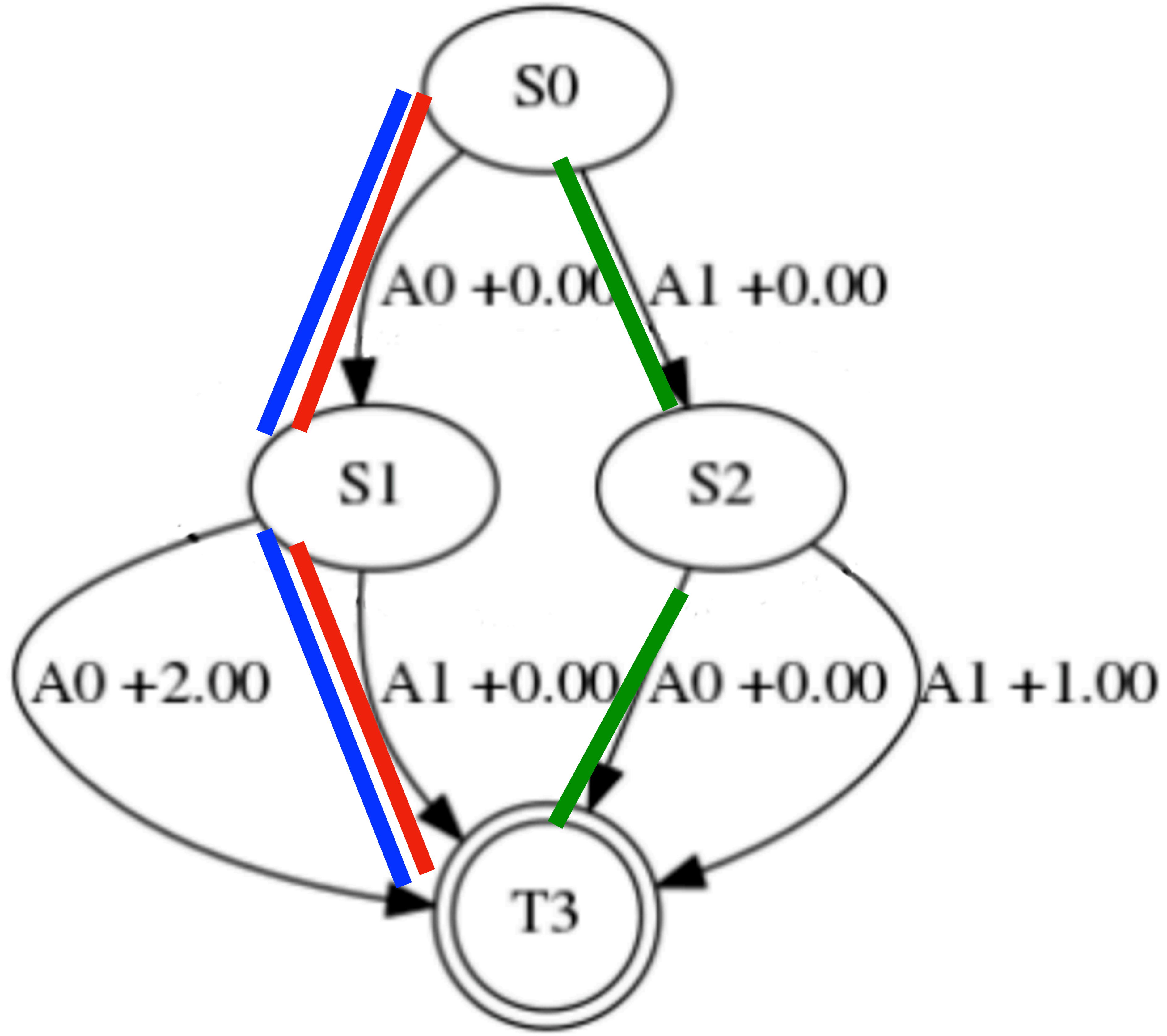
- **Maintaining exploration:** Why we need exploration in MC. Assume pi never takes action b in state S. If we want to estimate q(S,b) we will have no data about the reward you get from state S when pi chooses action b
- **Exploring starts:** every episode must begin in a random state, and the first action must be randomly selected, even if that action is not what pi would do
  - guarantees we visit every state-action pair
- **Epsilon-soft policies:** a stochastic policy. A policy where each action is selected with at least epsilon probability. (e.g., epsilon-greedy)

Terminology Review (3)

- **Off-policy:** learning about one policy, while following another
  - e.g., learning the value function for the optimal policy (q*) while following some exploration policy b (i.e. b=random_policy)
- **Target policy:** the policy you want to learn *about. W*e always call it pi. We either want to learn v_π or (q* and pi*)
- **Behavior policy:** the policy used to select actions, to generate the data. We always call it *b.* It is usually an exploratory policy (e.g., epsilon-greedy with respect to Q)
- **Importance sampling:** a statistical technique for estimating the expected value when the samples used to compute the average don't match the distribution you want.

Computing MC estimates by hand

Consider the three state MDP below with terminal state T_3 and $\gamma = 1$. Suppose you observe three episodes: $\{S_0, S_1, T_3\}$ with a return of 2, $\{S_0, S_1, T_3\}$ with a return of 2, $\{S_0, S_2, T_3\}$ with a return of 1. What is the (every-visit) Monte-carlo estimator of the value for each of state S_0, S_1, S_2? How would the Monte-Carlo estimates change if $r(S_0, A_1, S_1) = +1.00$?



Computing MC estimates by hand

Consider the three state MDP below with terminal state T_3 and $\gamma = 1$. Suppose you observe three episodes: $\{S_0, S_1, T_3\}$ with a return of 2, $\{S_0, S_1, T_3\}$ with a return of 2, $\{S_0, S_2, T_3\}$ with a return of 1. What is the (every-visit) Monte-carlo estimator of the value for each of state S_0, S_1, S_2? How would the Monte-Carlo estimates change if $r(S_0, A_1, S_1) = +1.00$?

Step 1: write down the states visited and the returns

Computing MC estimates by hand

Consider the three state MDP below with terminal state T_3 and $\gamma = 1$. Suppose you observe three episodes: $\{S_0, S_1, T_3\}$ with a return of 2, $\{S_0, S_1, T_3\}$ with a return of 2, $\{S_0, S_2, T_3\}$ with a return of 1. What is the (every-visit) Monte-carlo estimator of the value for each of state S_0, S_1, S_2? How would the Monte-Carlo estimates change if $r(S_0, A_1, S_1) = +1.00$?

Step 1: write down the states visited and the returns

S_0, S_1, T , return=2	S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, S_1, T , return=2	S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, S_2, T , return=1	S_0, A_1, **0**, S_2, A_1, **1**, T

Computing MC estimates by hand

Step 1: write down the states visited and the returns

S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, A_1, **0**, S_2, A_1, **1**, T

Step 2: write down the returns from each state in a list

Computing MC estimates by hand

Step 1: write down the states visited and the returns

S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, A_1, **0**, S_2, A_1, **1**, T

Step 2: write down the returns from each state in a list

- Start with S_2:
  - Returns(S_2) = [1]
  - V(S_2) = average (Returns(S_2)) = average([1]) = 1.0

Computing MC estimates by hand

Step 1: write down the states visited and the returns

S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, A_1, **0**, S_2, A_1, **1**, T

Do the same for V(S_1) and V(S_0)

Computing MC estimates by hand

Step 1: write down the states visited and the returns

S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, A_1, **0**, S_2, A_1, **1**, T

Answer hidden below, only visible by moonlight:

>>

Computing MC estimates by hand

S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, A_1, **0**, S_2, A_1, **1**, T

How would the Monte-Carlo estimates change if $r(S_0, A_1, S_1) = +1.00$?

>>

Computing MC estimates by hand

S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, A_1, **0**, S_2, A_1, **1**, T

How would the Monte-Carlo estimates change if $r(S_0, A_1, S_1) = +1.00$?

>>

S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, A_1, **1**, S_2, A_1, **1**, T —> G=2

Computing MC estimates by hand

How would the Monte-Carlo estimates change if $r(S_0, A_1, S_1) = +1.00$?

S_0, A_0, **0**, S_1, A_0, **2**, T	>>	S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, A_0, **0**, S_1, A_0, **2**, T		S_0, A_0, **0**, S_1, A_0, **2**, T
S_0, A_1, **0**, S_2, A_1, **1**, T		S_0, A_1, **1**, S_2, A_1, **1**, T

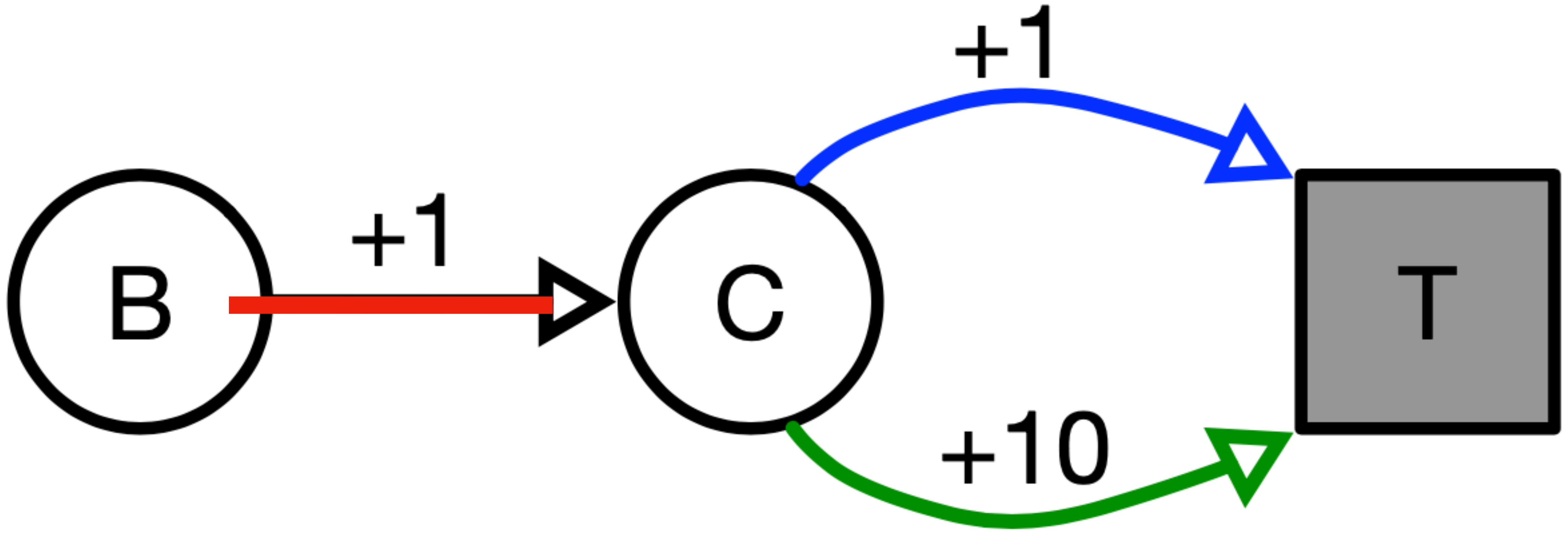
- Would V(S_2) change?
- Would V(S_1) change?
- How would V(S_0) change? What is returns(S_0) now?

Computing off-policy MC estimates

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 $\gamma = 1.0$. In state C both actions transition to the terminating state with $A = 1$ following the blue path to receive a reward $R = 1$ and $A = 2$ following the green path to receive a reward $R = 10$. Assume the target policy π has $\pi(A = 1|C) = 0.9$ and $\pi(A = 2|C) = 0.1$, and that the behaviour policy b has $b(A = 1|C) = 0.25$ and $b(A = 2|C) = 0.75$.

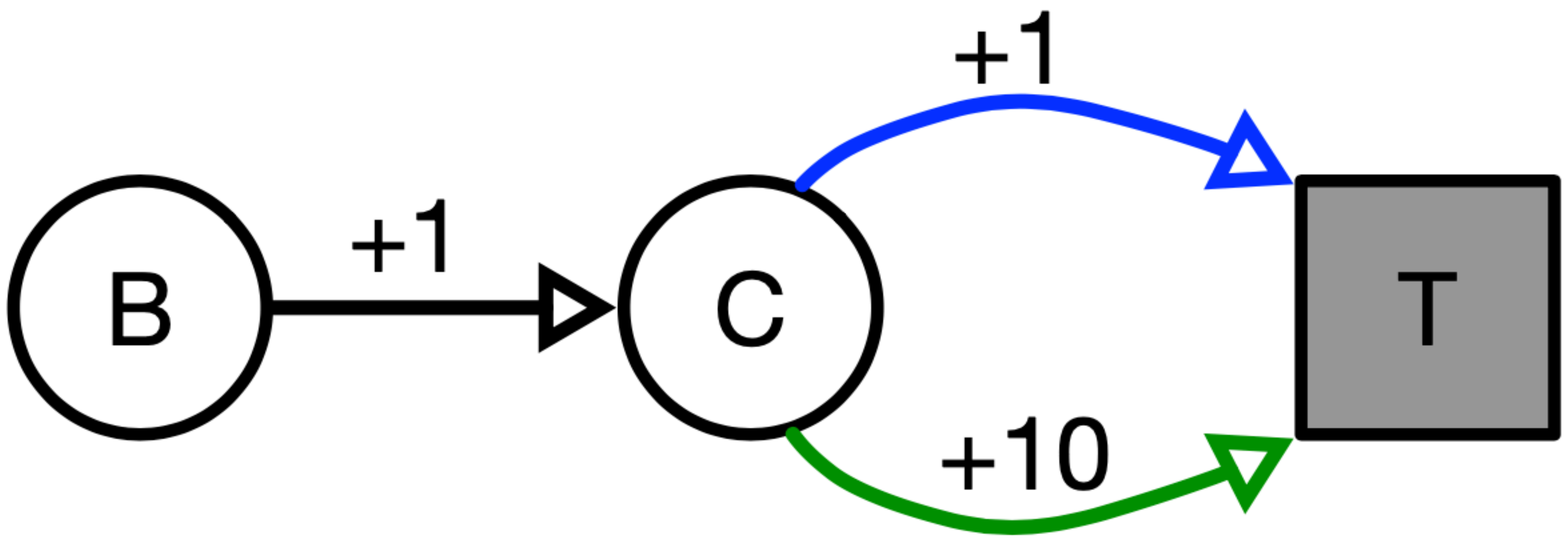
Computing off-policy MC estimates

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 $\gamma = 1.0$. In state C both actions transition to the terminating state with $A = 1$ following the blue path to receive a reward $R = 1$ and $A = 2$ following the green path to receive a reward $R = 10$. Assume the target policy π has $\pi(A = 1|C) = 0.9$ and $\pi(A = 2|C) = 0.1$, and that the behaviour policy b has $b(A = 1|C) = 0.25$ and $b(A = 2|C) = 0.75$.



Computing off-policy MC estimates

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 $\gamma = 1.0$. In state C both actions transition to the terminating state with $A = 1$ following the blue path to receive a reward $R = 1$ and $A = 2$ following the green path to receive a reward $R = 10$. Assume the target policy π has $\pi(A = 1|C) = 0.9$ and $\pi(A = 2|C) = 0.1$, and that the behaviour policy b has $b(A = 1|C) = 0.25$ and $b(A = 2|C) = 0.75$.

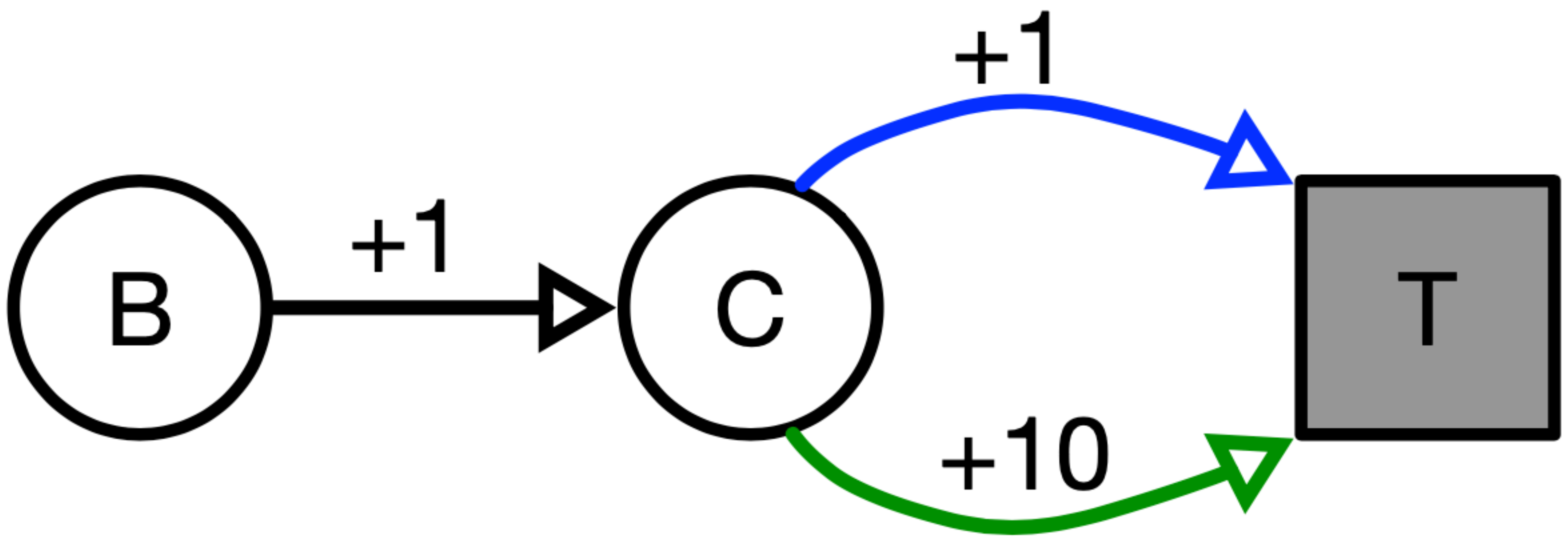


What v_\pi(C)?

hint:
v_\pi(s) = E[R] + \gamma v_\pi(s')

Computing off-policy MC estimates

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 $\gamma = 1.0$. In state C both actions transition to the terminating state with $A = 1$ following the blue path to receive a reward $R = 1$ and $A = 2$ following the green path to receive a reward $R = 10$. Assume the target policy π has $\pi(A = 1|C) = 0.9$ and $\pi(A = 2|C) = 0.1$, and that the behaviour policy b has $b(A = 1|C) = 0.25$ and $b(A = 2|C) = 0.75$.

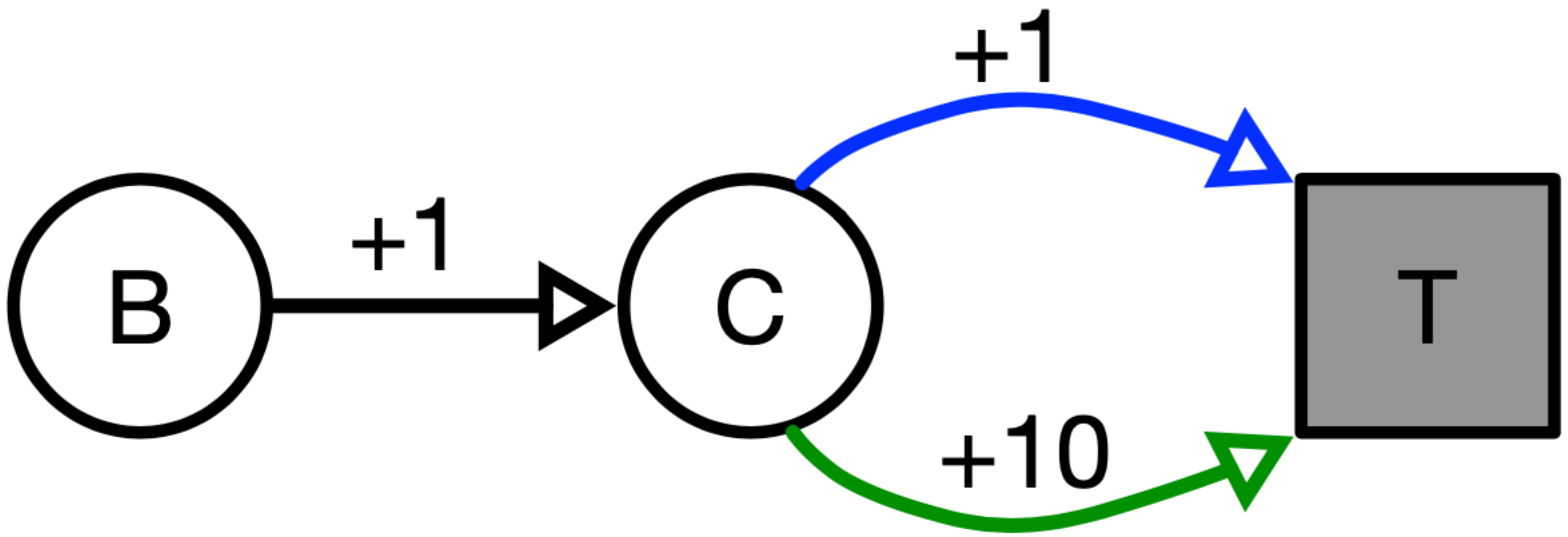


What v_\pi(C)
= 1*.9 + 10*1

=1.9

V_\pi(B) = ?

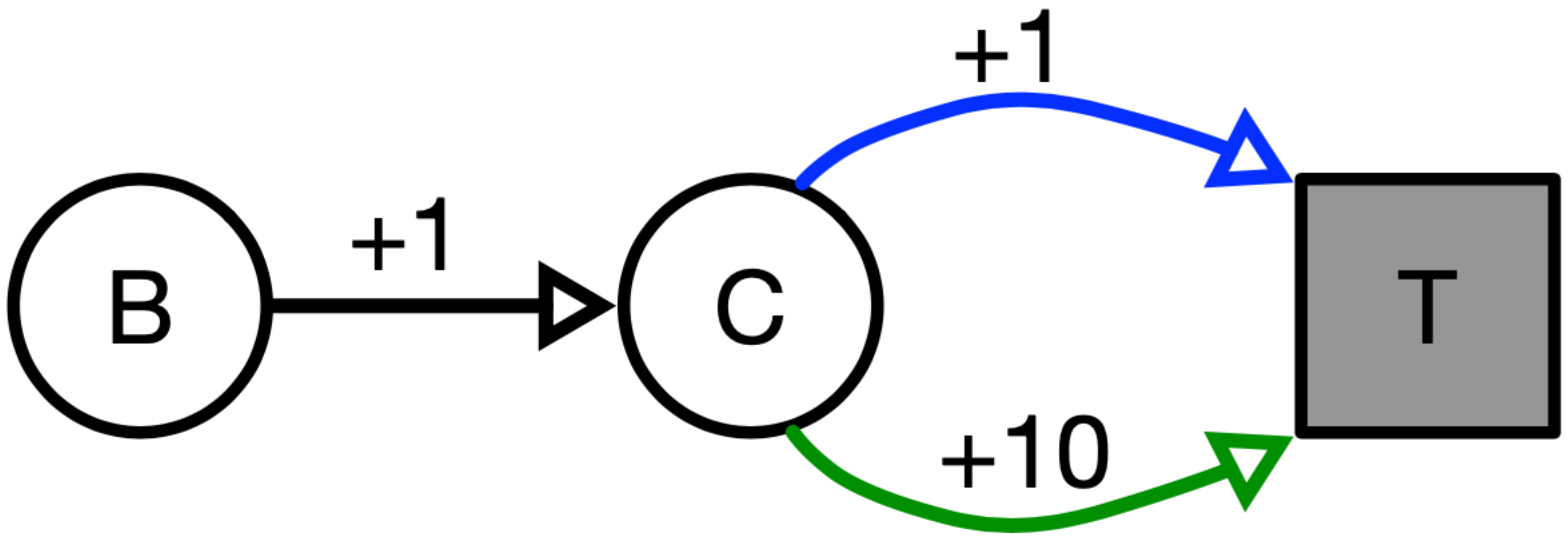
path to receive a reward $R = 10$. Assume the target policy π has $\pi(A = 1|C) = 0.9$ and $\pi(A = 2|C) = 0.1$, and that the behaviour policy b has $b(A = 1|C) = 0.25$ and $b(A = 2|C) = 0.75$



What is the return from this episode??

<S0= B, A0= 1, R1= 1, S1= C , A1= 1, R2= 1, S_2 = T>

path to receive a reward $R = 10$. Assume the target policy π has $\pi(A = 1|C) = 0.9$ and $\pi(A = 2|C) = 0.1$, and that the behaviour policy b has $b(A = 1|C) = 0.25$ and $b(A = 2|C) = 0.75$



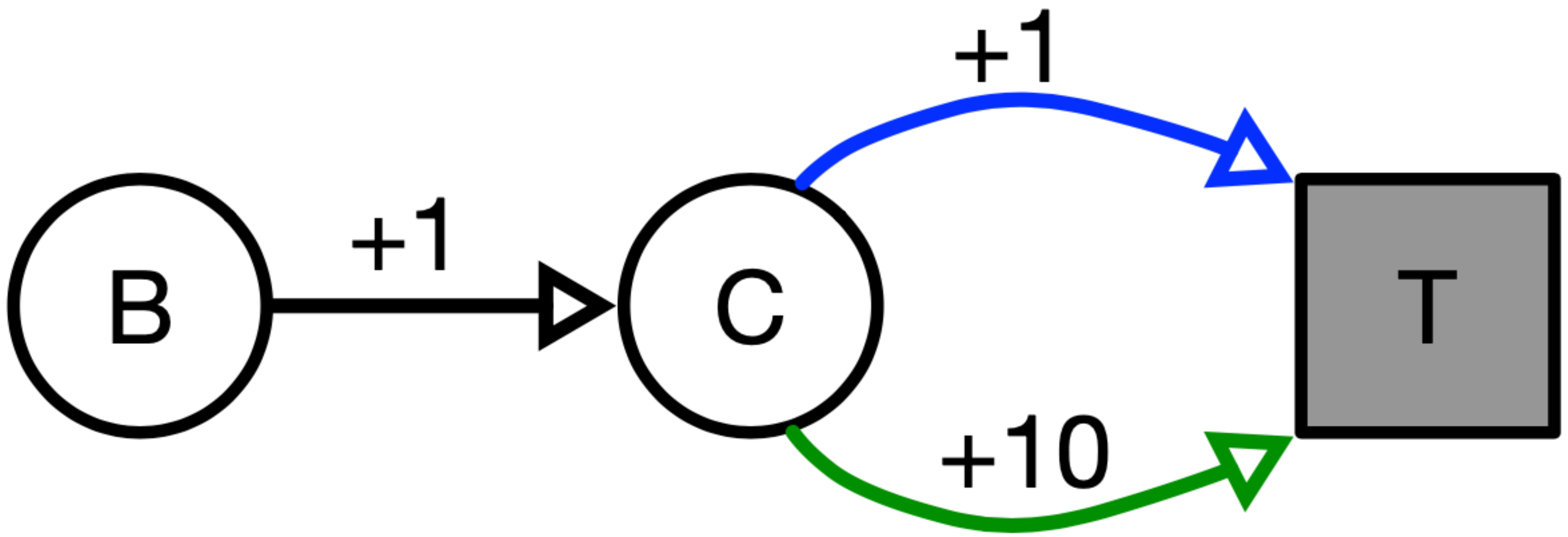
What is the return from this episode??

<S0= B, A0= 1, R1= 1, S1= C , A1= 1, R2= 1, S_2 = T>

G = 2

What is v_\pi(B) based on this one episode?

path to receive a reward $R = 10$. Assume the target policy π has $\pi(A = 1|C) = 0.9$ and $\pi(A = 2|C) = 0.1$, and that the behaviour policy b has $b(A = 1|C) = 0.25$ and $b(A = 2|C) = 0.75$



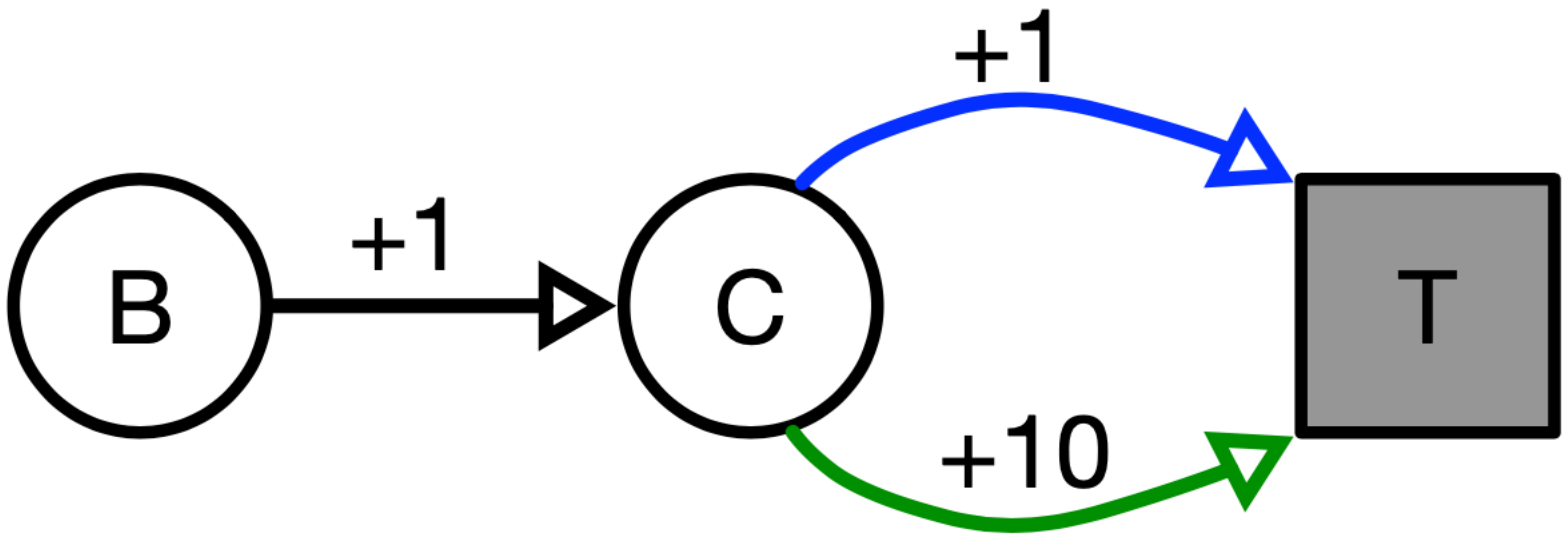
What is the return from this episode??

<S0= B, A0= 1, R1= 1, S1= C , A1= 1, R2= 1, S_2 = T>

G = 2

What is v_\pi(B) based on this one episode? 2

path to receive a reward $R = 10$. Assume the target policy π has $\pi(A = 1|C) = 0.9$ and $\pi(A = 2|C) = 0.1$, and that the behaviour policy b has $b(A = 1|C) = 0.25$ and $b(A = 2|C) = 0.75$.



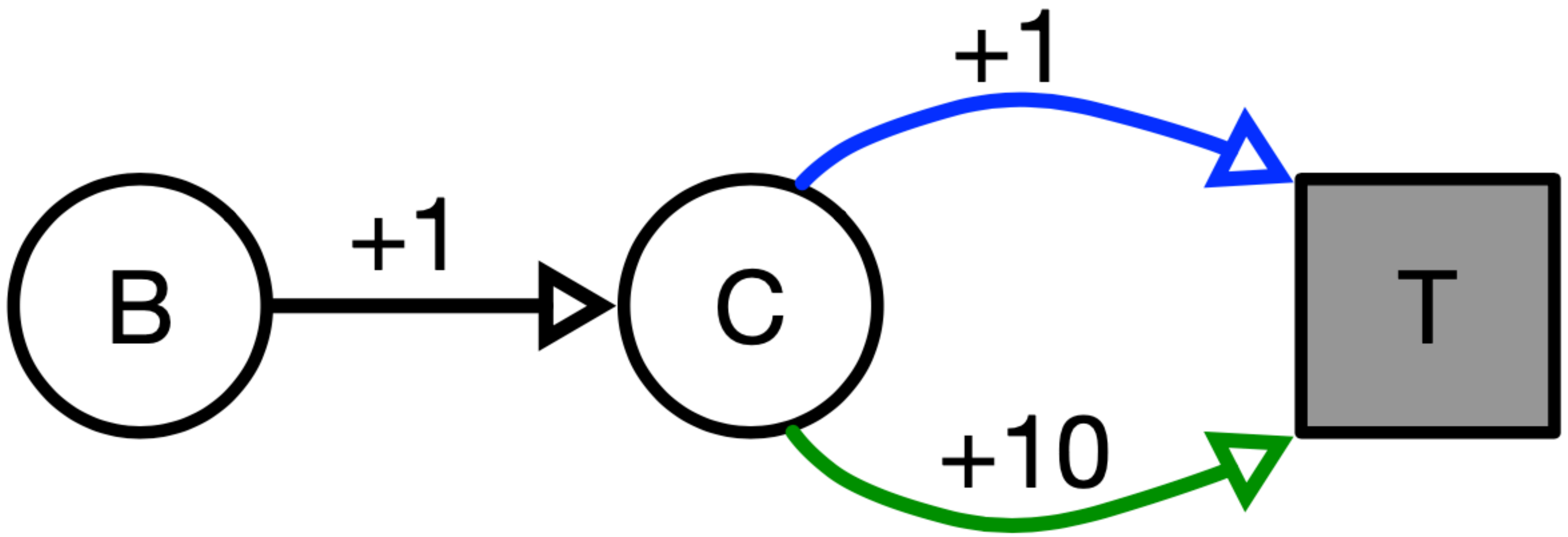
Now let's do **off-policy!!!**

Lets assume we get the following episode from policy b:

<S0= B, A0= 1, R1= 1, S1= C , **A1= 2, R2= 10**, S_2 = T>

G = 11

path to receive a reward $R = 10$. Assume the target policy π has $\pi(A = 1|C) = 0.9$ and $\pi(A = 2|C) = 0.1$, and that the behaviour policy b has $b(A = 1|C) = 0.25$ and $b(A = 2|C) = 0.75$



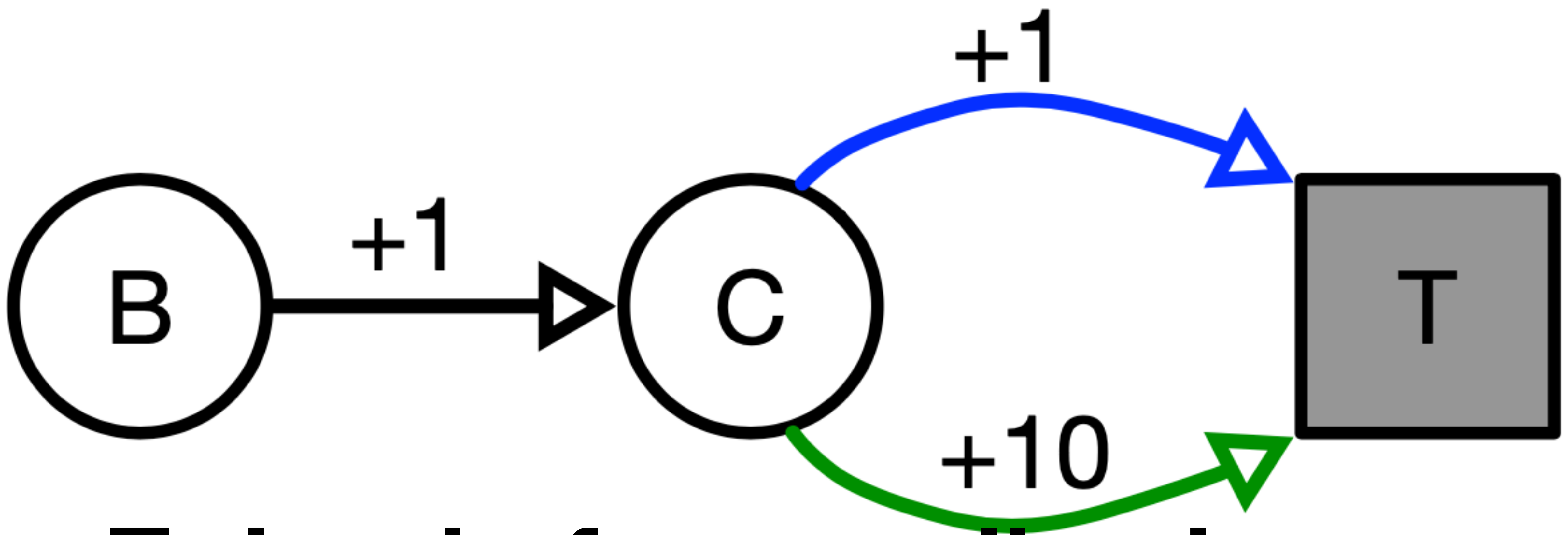
Episode from policy b:

<S0= B, A0= 1, R1= 1, S1= C , **A1= 2, R2= 10**, S_2 = T>

G = 11

What is prob of this episode under \pi?

path to receive a reward $R = 10$. Assume the target policy π has $\pi(A = 1|C) = 0.9$ and $\pi(A = 2|C) = 0.1$, and that the behaviour policy b has $b(A = 1|C) = 0.25$ and $b(A = 2|C) = 0.75$



Episode from policy b:

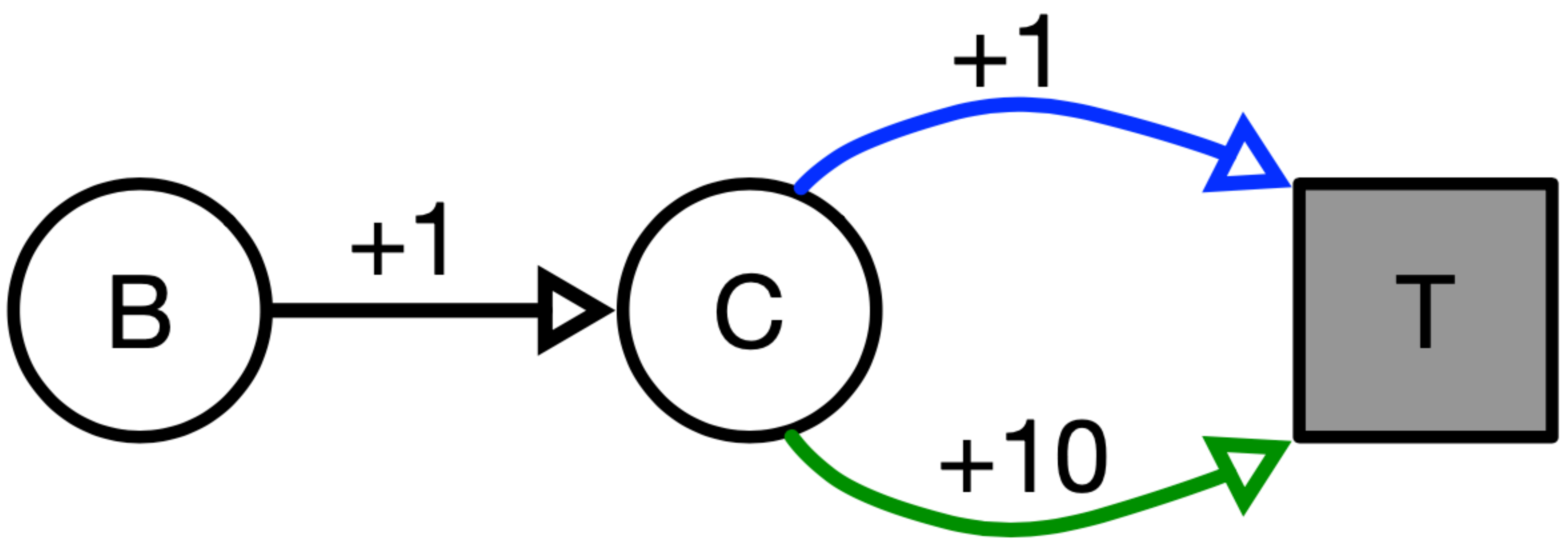
<S0= B, A0= 1, R1= 1, S1= C , **A1= 2, R2= 10**, S_2 = T>

G = 11

What is prob of this episode under \pi?

\pi(A=1| B)*\pi(A=2|C) = 1 * 0.1

path to receive a reward $R = 10$. Assume the target policy π has $\pi(A = 1|C) = 0.9$ and $\pi(A = 2|C) = 0.1$, and that the behaviour policy b has $b(A = 1|C) = 0.25$ and $b(A = 2|C) = 0.75$.



Episode from policy b:

<S0= B, A0= 1, R1= 1, S1= C , **A1= 2,** **R2= 10**, S_2 = T>

G = 11

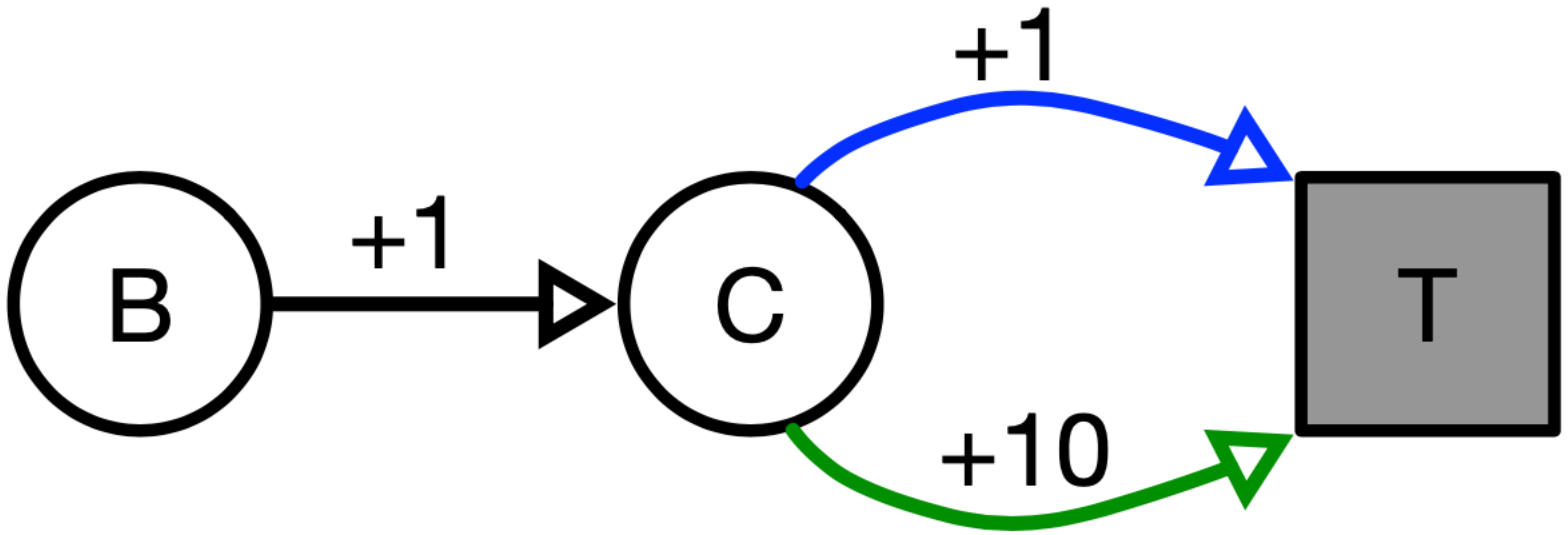
What is prob of this episode under \pi?

\pi(A=1| B)*\pi(A=2|C) = 1 * 0.1 = 0.1

What is prob of this episode under b?

b(A=1| B)*b(A=2|C) = 1 * 0.75 = 0.75

path to receive a reward $R = 10$. Assume the target policy π has $\pi(A = 1|C) = 0.9$ and $\pi(A = 2|C) = 0.1$, and that the behaviour policy b has $b(A = 1|C) = 0.25$ and $b(A = 2|C) = 0.75$.



Episode from policy b:

<S0 = B, A00 = 1, R1 = 1, S1 = C , **A1 = 2, R2 = 10**, S_2 = T>; G = 11

Prob of this episode under \pi? \pi(A=1| B)*\pi(A=2|C) = 1 * 0.1 = 0.1
Prob of this episode under b? b(A=1| B)*b(A=2|C) = 1 * 0.75 = 0.75

What would be the off-policy MC estimate of V_\pi(B) using this episode from b (using the importance sampling ratio)? **1.47**

A simple proof exercise

- Let $\rho_t = \dfrac{\pi(A_t \mid S_t)}{b(A_t \mid S_t)}$ and $\boxed{r(s,a) \doteq \mathbb{E}[R_t \mid S_{t-1}=s, A_{t-1}=a] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r \mid s, a)}$

- Show that $\mathbb{E}_b[\rho_t R_{t+1} \mid S_t = s] = \mathbb{E}_\pi[R_{t+1} \mid S_t = s]$

A simple proof exercise

- Let $\rho_t = \dfrac{\pi(A_t \mid S_t)}{b(A_t \mid S_t)}$ and $\boxed{r(s,a) \doteq \mathbb{E}[R_t \mid S_{t-1}=s, A_{t-1}=a] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r \mid s, a)}$

- Show that $\mathbb{E}_b[\rho_t R_{t+1} \mid S_t = s] = \mathbb{E}_\pi[R_{t+1} \mid S_t = s]$

- Lets write out the expectation as a sum:

    - $$\mathbb{E}_b[\rho_t R_{t+1} \mid S_t = s] = \sum_a b(a \mid s) \mathbb{E}[\rho_t R_{t+1} \mid S_t = s, A_t = a]$$

A simple proof exercise

- Let $\rho_t = \dfrac{\pi(A_t \mid S_t)}{b(A_t \mid S_t)}$ and $\boxed{r(s,a) \doteq \mathbb{E}[R_t \mid S_{t-1}=s, A_{t-1}=a] = \sum_{r\in\mathcal{R}} r \sum_{s'\in\mathcal{S}} p(s', r \mid s, a)}$
- Show that $\mathbb{E}_b[\rho_t R_{t+1} \mid S_t = s] = \mathbb{E}_\pi[R_{t+1} \mid S_t = s]$
- Lets write out the expectation as a sum:

$$
\begin{aligned}
\mathbb{E}_b[\rho_t R_{t+1} \mid S_t = s] &= \sum_a b(a \mid s)\mathbb{E}[\rho_t R_{t+1} \mid S_t = s, A_t = a] \\
&= \sum_a b(a \mid s)\frac{\pi(a \mid s)}{b(a \mid s)}\mathbb{E}[R_{t+1} \mid S_t = s, A_t = a] \\
&= \sum_a b(a \mid s)\frac{\pi(a \mid s)}{b(a \mid s)}\sum_{s',r} p(s', r \mid s, a) r \\
&= \sum_a \cancel{b(a \mid s)}\frac{\pi(a \mid s)}{\cancel{b(a \mid s)}} r(s,a) \\
&= \sum_a \pi(a \mid s) r(s,a) = \mathbb{E}_\pi[R_{t+1} \mid S_t = s]
\end{aligned}
$$