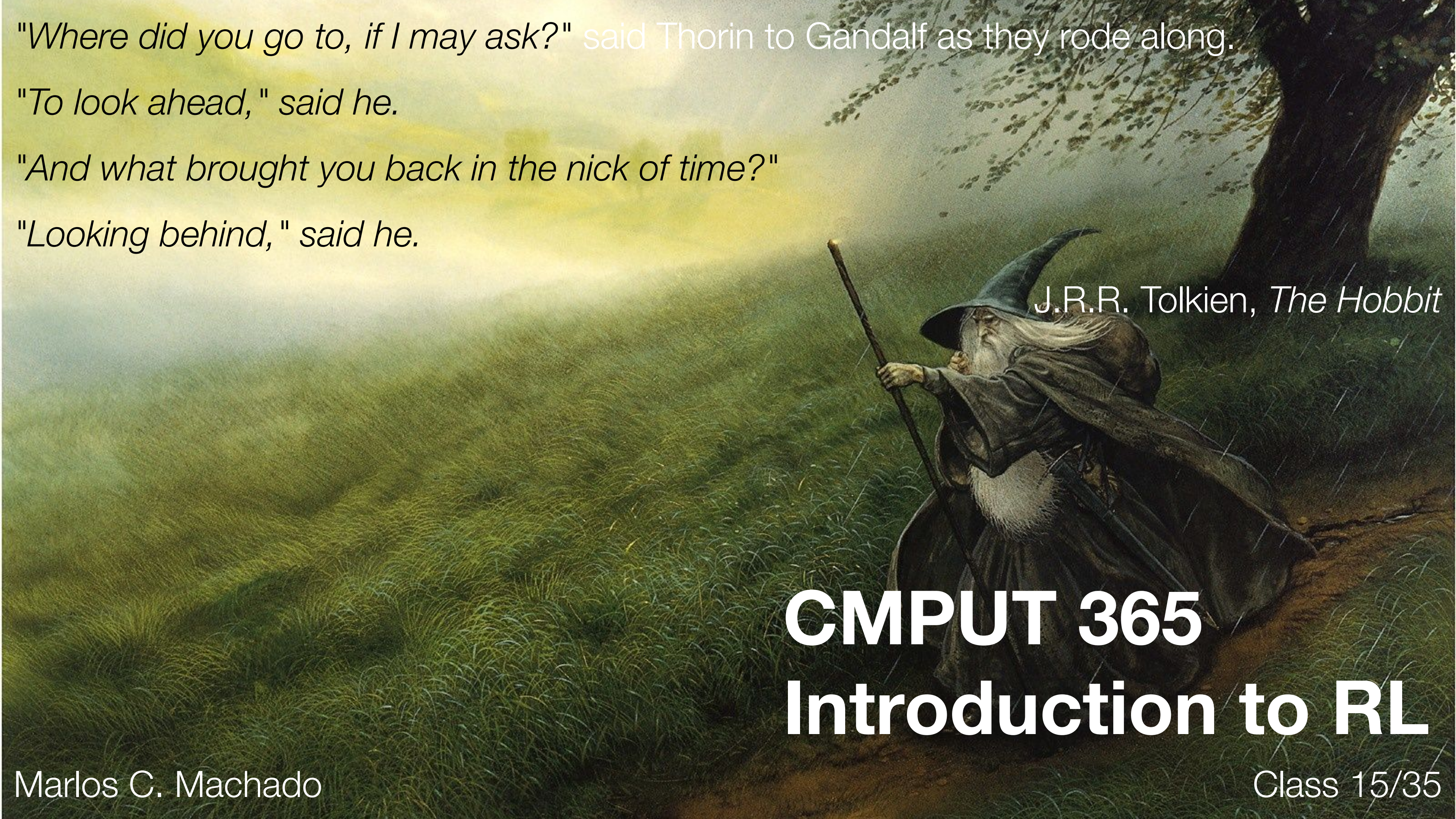


"Where did you go to, if I may ask?" said Thorin to Gandalf as they rode along.

"To look ahead," said he.

"And what brought you back in the nick of time?"

"Looking behind," said he.

J.R.R. Tolkien, *The Hobbit*

CMPUT 365 Introduction to RL

Marlos C. Machado

Class 15/35

Reminder

You **should be enrolled in the private session** we created in Coursera for CMPUT 365.

I **cannot** use marks from the public repository for your course marks.

You **need** to **check**, **every time**, if you are in the private session and if you are submitting quizzes and assignments to the private section.

There were **13 pending invitations** last time I checked ¯_(ツ)_/¯

If you have any questions or concerns, **talk with the TAs** or email us `cmput365@ualberta.ca`.

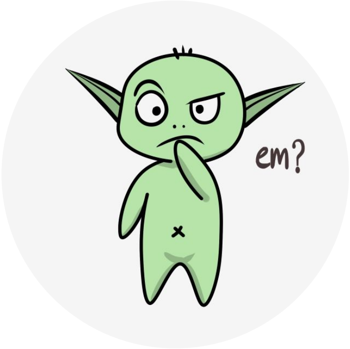
Reminders and Notes

- Exam viewing:
  - It will happen next Thursday (1 pm – 4pm) and Friday (2 pm – 5pm) at CSC 3-50.
- What I plan to do today:
  - Finish overview of Monte Carlo Methods for Prediction & Control (Chapter 5 of the textbook).
- Useful information for you:
  - Monday is a holiday – Thanksgiving.
  - The Quiz for Temporal Difference Learning is due on Wednesday.
  - Rich Sutton's guest lecture is confirmed for December 9th.

SPOT: Mid-term Course Evaluation

https://go.blueja.io/MlqAHuUezE-my_PTHx9IEg

Please, interrupt me at any time!



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Last Class: MC Control without Exploring Starts

On-policy: You learn about the policy you used to make decisions.

Off-policy: You learn about a policy that is different from the one you used to make decisions.

$\boxed{\text{On-policy}}$ first-visit MC control (for ε-soft policies), estimates $\pi \approx \pi_*$

Algorithm parameter: small $\varepsilon > 0$

Initialize:

$\pi \leftarrow$ an arbitrary ε-soft policy

$Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

$Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Repeat forever (for each episode):

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

$G \leftarrow 0$

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

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

Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:

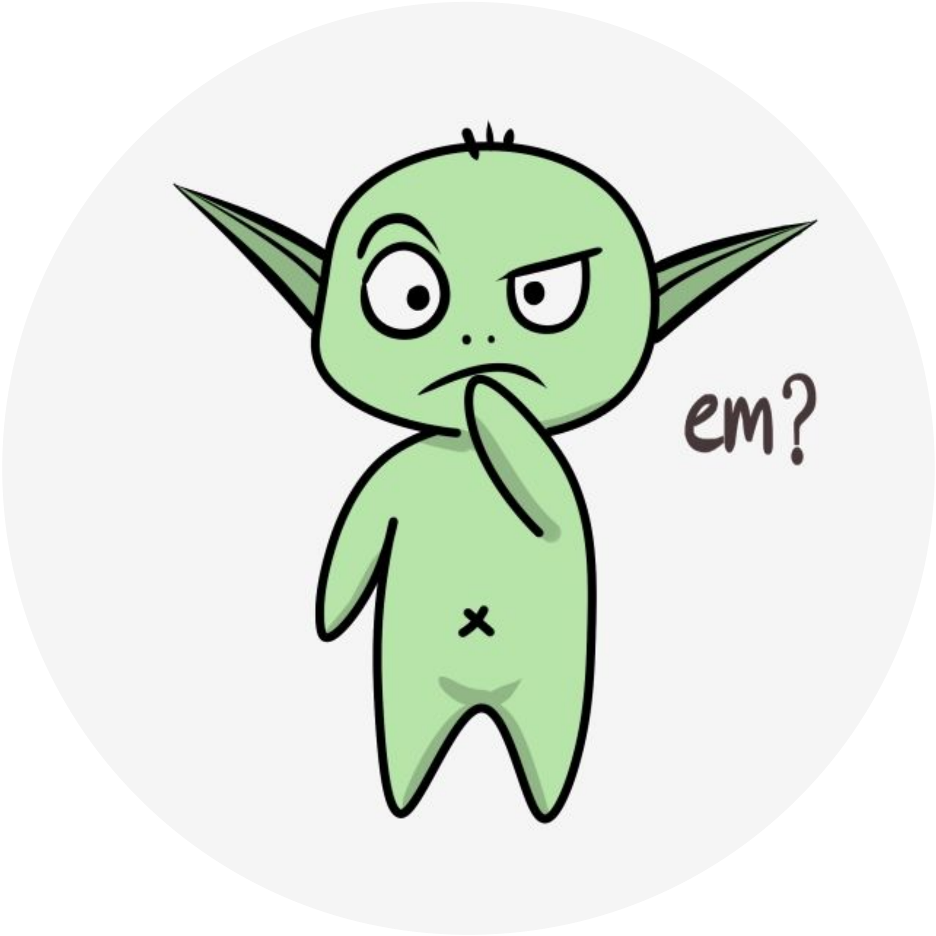
Append G to $Returns(S_t, A_t)$

$Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$

$A^* \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken arbitrarily)

For all $a \in \mathcal{A}(S_t)$:

$$\pi(a|S_t) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(S_t)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(S_t)| & \text{if } a \neq A^* \end{cases}$$



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Learning with exploration

- *On-policy first-visit MC control (for ε-soft policies)* seems great!
- … but how can we learn about the optimal policy while behaving according to an exploratory policy? We need to behave non-optimally in order to explore 🤔.
- So far we have been *on-policy*, which is a compromise: we learn about a near-optimal policy, not the optimal one.
- But what if we had two policies? We use one for exploration but we learn about another one, which would be the optimal policy?

Behaviour policy

Target policy

That's off-policy learning!

Pros and cons of off-policy learning

Pros

- It is more general.
- It is more powerful.
- It can benefit from external data
  - and other additional use cases.

Cons

- It is more complicated.
- It has much more variance.
  - Thus it can be much slower to learn.
- It can be unstable.

Check Example 5.5 in the textbook about Infinite Variance

What's the actual issue?

Let π denote the target policy, and let b denote the behaviour policy.

We want to estimate $\mathbb{E}_{\pi}[G_t]$, but what we can actually directly estimate is $\mathbb{E}_{b}[G_t]$.

In other words, $\mathbb{E}[G_t \mid S_t = s] = v_b(s)$.

Importance Sampling

A general technique for estimating expected values under one distribution given samples from another. It is based on re-weighting the probabilities of an event.

Importance Sampling

$$\mathbb{E}_\pi[X] \doteq \sum_{x \in X} x\pi(x)$$

Importance Sampling

In RL, the probability of a trajectory is:

$$
\begin{aligned}
&\Pr\{A_t, S_{t+1}, A_{t+1}, \ldots, S_T \mid S_t, A_{t:T-1} \sim \pi\} \\
&\quad = \pi(A_t|S_t)p(S_{t+1}|S_t, A_t)\pi(A_{t+1}|S_{t+1}) \cdots p(S_T|S_{T-1}, A_{T-1}) \\
&\quad = \prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k),
\end{aligned}
$$

Importance Sampling

In RL, the probability of a trajectory is:

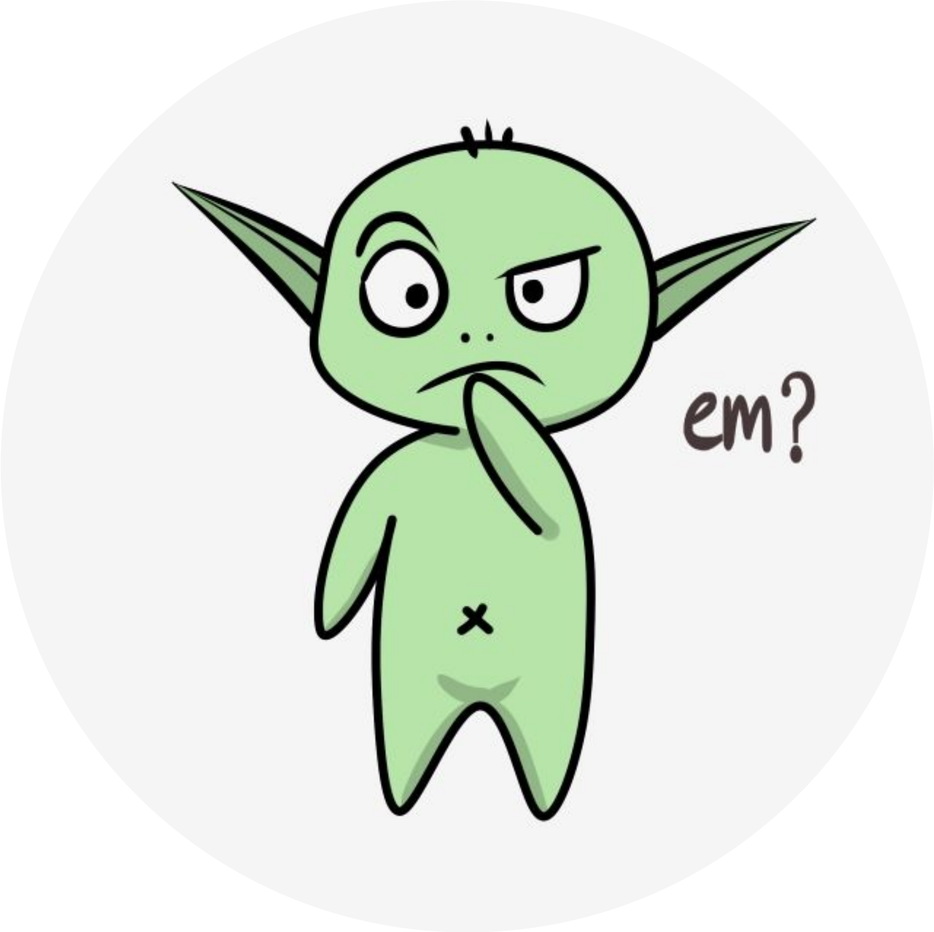
$$
\begin{aligned}
&\Pr\{A_t, S_{t+1}, A_{t+1}, \ldots, S_T \mid S_t, A_{t:T-1} \sim \pi\} \\
&\quad = \pi(A_t|S_t)p(S_{t+1}|S_t, A_t)\pi(A_{t+1}|S_{t+1}) \cdots p(S_T|S_{T-1}, A_{T-1}) \\
&\quad = \prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k),
\end{aligned}
$$

the relative prob. of the traj. under the target and behavior policies (the IS ratio) is:

We require *coverage*:
b(a|s) > 0 when π(a|s) > 0

$$
\rho_{t:T-1} \doteq \frac{\prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k)}{\prod_{k=t}^{T-1} b(A_k|S_k)p(S_{k+1}|S_k, A_k)} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}.
$$

The IS ratio does not depend on the MDP, that is, on p(s’, r | s, a)!



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

The solution

The ratio $\rho_{t:T-1}$ transforms the returns to have the right expected value:

$$\mathbb{E}[\rho_{t:T-1}G_t \mid S_t = s] = v_\pi(s).$$

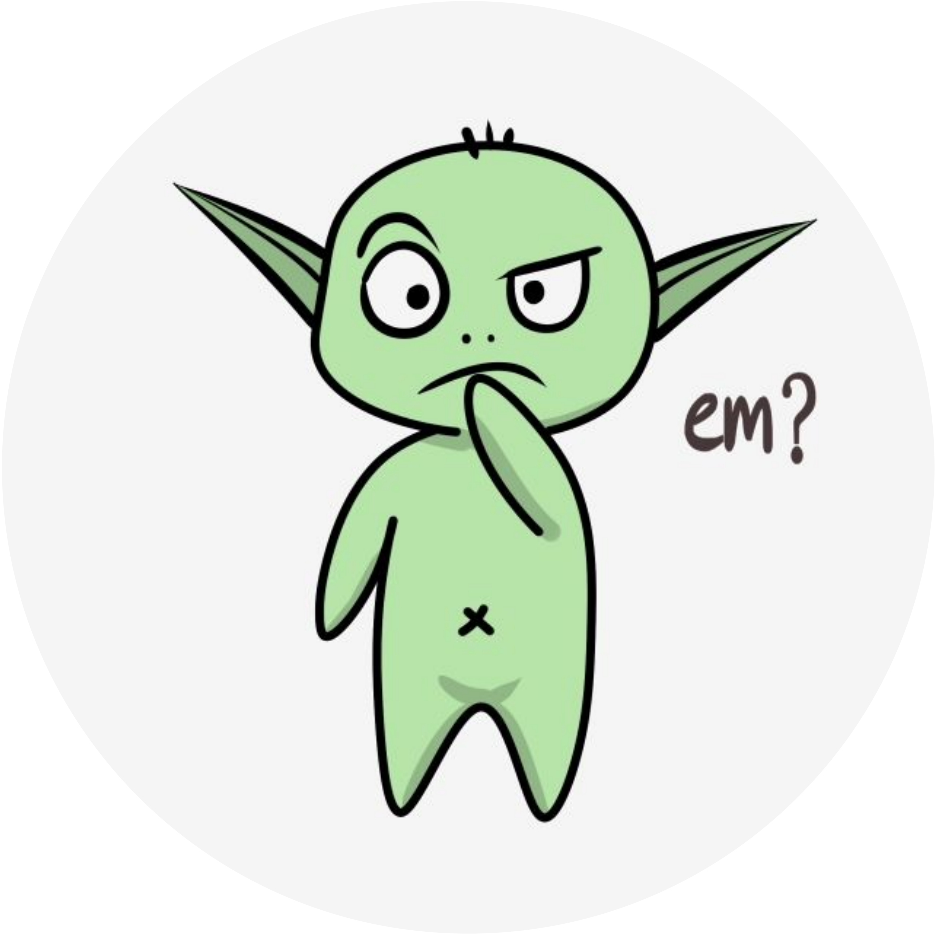
Ordinary importance sampling:

$$V(s) \doteq \frac{\sum_{t\in\mathcal{T}(s)} \rho_{t:T(t)-1}G_t}{|\mathcal{T}(s)|}.$$

Set of all time steps in which state s is visited.

Weighted importance sampling:

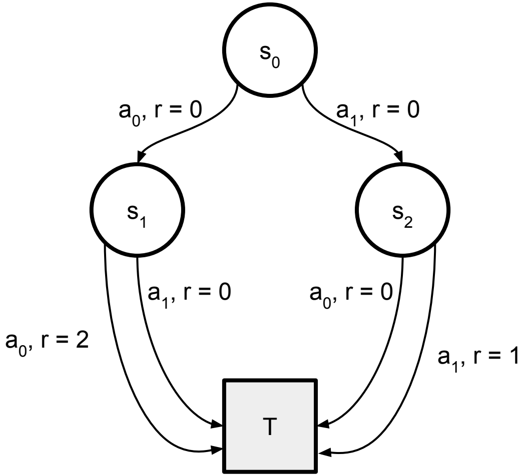
$$V(s) \doteq \frac{\sum_{t\in\mathcal{T}(s)} \rho_{t:T(t)-1}G_t}{\sum_{t\in\mathcal{T}(s)} \rho_{t:T(t)-1}}$$



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Practice Exercise 1

Consider the three-state MDP below with terminal state T and $\gamma = 1$. Suppose you observe three episodes: **$\{s_0, s_1, T\}$** with a return of 2, **$\{s_0, s_1, T\}$** with a return of 2, **$\{s_0, s_2, T\}$** with a return of 1. **What is the (every-visit) Monte-Carlo estimator of the value for each of the states, s_0, s_1, s_2?** How would the Monte-Carlo estimates change if $r(s_0, a_1, s_2) = 1$?



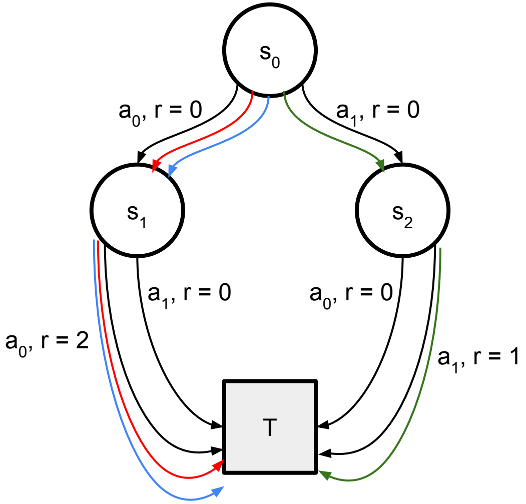
Practice Exercise 1

Consider the three-state MDP below with terminal state T and $\gamma = 1$. Suppose you observe three episodes: **$\{s_0, s_1, T\}$** with a return of 2, **$\{s_0, s_1, T\}$** with a return of 2, **$\{s_0, s_2, T\}$** with a return of 1. **What is the (every-visit) Monte-Carlo estimator of the value for each of the states, s_0, s_1, s_2?** How would the Monte-Carlo estimates change if $r(s_0, a_1, s_2) = 1$?

Trajectories:	**Returns:**
$s_0, a_0, 0, s_1, a_0, 2, T$	$0 + 2 = 2$
$s_0, a_0, 0, s_1, a_0, 2, T$	$0 + 2 = 2$
$s_0, a_1, 0, s_2, a_1, 1, T$	$0 + 1 = 1$

States Visited / Return:

s_0, s_1, T / 2
s_0, s_1, T / 2
s_0, s_2, T / 1



Monte-Carlo Estimate

Returns from s_2: [1] → $V(s_2)$ = avg([1]) = 1
Returns from s_1: [2, 2] → $V(s_1)$ = avg([2, 2]) = 2
Returns from s_0: [1, 2, 2] → $V(s_0)$ = avg([1, 2, 2]) = 5/3

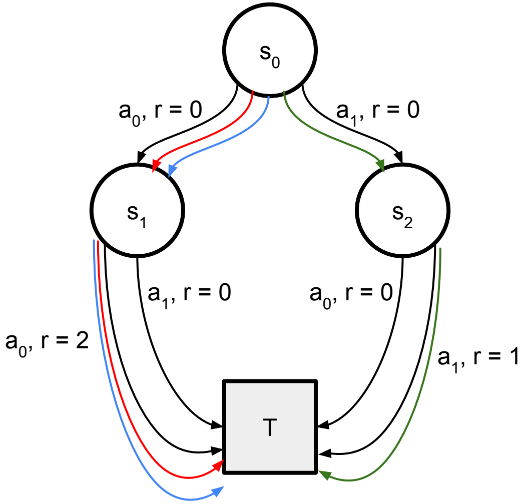
Practice Exercise 1

Consider the three-state MDP below with terminal state T and γ = 1. Suppose you observe three episodes: **{s_0, s_1, T}** with a return of 2, **{s_0, s_1, T}** with a return of 2, **{s_0, s_2, T}** with a return of 1. What is the (every-visit) Monte-Carlo estimator of the value for each of the states, s_0, s_1, s_2? **How would the Monte-Carlo estimates change if $r(s_0, a_1, s_2) = 1$?**

Trajectories:	**Returns:**
s_0, a_0, 0, s_1, a_0, 2, T	0 + 2 = 2
s_0, a_0, 0, s_1, a_0, 2, T	0 + 2 = 2
s_0, a_1, 0, s_2, a_1, 1, T	0 + 1 = 1

States Visited / Return:

s_0, s_1, T / 2
s_0, s_1, T / 2
s_0, s_2, T / 1



Monte-Carlo Estimate

Returns from s_2: [1] → $V(s_2)$ = avg([1]) = 1
Returns from s_1: [2, 2] → $V(s_1)$ = avg([2, 2]) = 2
Returns from s_0: [1, 2, 2] → $V(s_0)$ = avg([1, 2, 2]) = 5/3

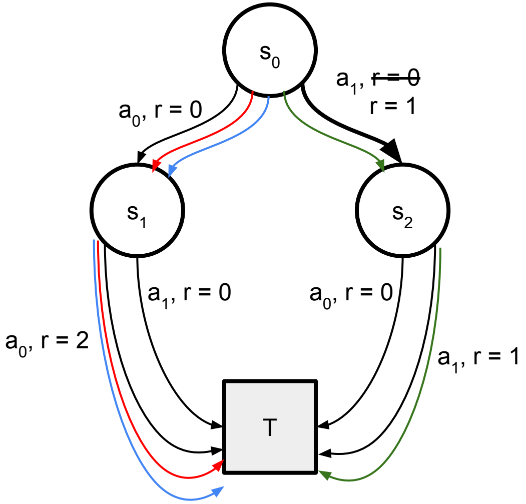
Practice Exercise 1

Consider the three-state MDP below with terminal state T and γ = 1. Suppose you observe three episodes: **{s_0, s_1, T}** with a return of 2, **{s_0, s_1, T}** with a return of 2, **{s_0, s_2, T}** with a return of 1. What is the (every-visit) Monte-Carlo estimator of the value for each of the states, s_0, s_1, s_2? **How would the Monte-Carlo estimates change if $r(s_0, a_1, s_2) = 1$?**

Trajectories:	**Returns:**
s_0, a_0, 0, s_1, a_0, 2, T	0 + 2 = 2
s_0, a_0, 0, s_1, a_0, 2, T	0 + 2 = 2
s_0, a_1, 0, s_2, a_1, 1, T	0 + 1 = 1

States Visited / Return:

s_0, s_1, T / 2
s_0, s_1, T / 2
s_0, s_2, T / 1



Monte-Carlo Estimate

Returns from s_2: [1] → $V(s_2)$ = avg([1]) = 1
Returns from s_1: [2, 2] → $V(s_1)$ = avg([2, 2]) = 2
Returns from s_0: [1, 2, 2] → $V(s_0)$ = avg([1, 2, 2]) = 5/3

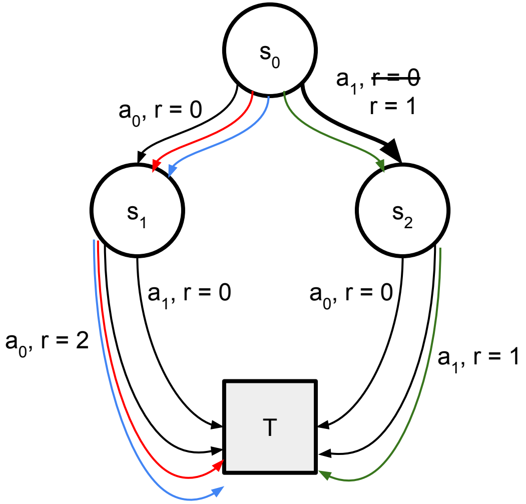
Practice Exercise 1

Consider the three-state MDP below with terminal state T and γ = 1. Suppose you observe three episodes: **{s_0, s_1, T}** with a return of 2, **{s_0, s_1, T}** with a return of 2, **{s_0, s_2, T}** with a return of 1. What is the (every-visit) Monte-Carlo estimator of the value for each of the states, s_0, s_1, s_2? **How would the Monte-Carlo estimates change if $r(s_0, a_1, s_2) = 1$?**

Trajectories:	**Returns:**
s_0, a_0, 0, s_1, a_0, 2, T	0 + 2 = 2
s_0, a_0, 0, s_1, a_0, 2, T	0 + 2 = 2
~~s_0, a_1, 0, s_2, a_1, 1, T~~	~~0 + 1 = 1~~
s_0, a_1, **1**, s_2, a_1, 1, T	1 + 1 = 2

States Visited / Return:

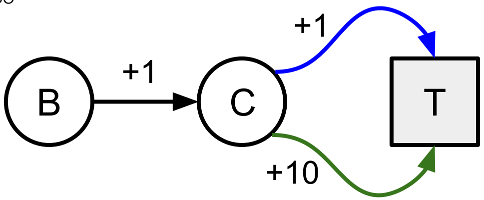
s_0, s_1, T / 2
s_0, s_1, T / 2
~~s_0, s_2, T / 1~~
s_0, s_2, T / 2



Monte-Carlo Estimate

Returns from s_2: [1] → $V(s_2)$ = avg([1]) = 1
Returns from s_1: [2, 2] → $V(s_1)$ = avg([2, 2]) = 2
~~Returns from s_0: [1, 2, 2] → $V(s_0)$ = avg([1, 2, 2]) = 5/3~~
Returns from s_0: [2, 2, 2] → $V(s_0)$ = avg([2, 2, 2]) = 2

Practice Exercise 2



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 $\pi(A = 1 \mid C) = 0.9$ and $\pi(A = 2 \mid C) = 0.1$, and that the behaviour policy *b* has $b(A = 1 \mid C) = 0.25$ and $b(A = 2 \mid C) = 0.75$.

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?

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 rations) would give you value estimates for *b*.

d) But we do not actually want to estimate the values for behaviour b, we want to estimate the values for π. So we need to use importance sampling rations 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?

Practice Exercise 2