"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 17/35

Reminder I

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 is still **1 pending invitation** for a student who is still enrolled in the course.

At the end of the term, I **will not port grades** from the public session in Coursera.

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

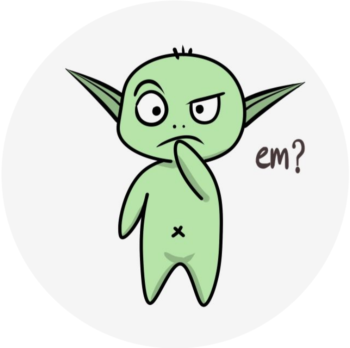
Plan / Reminder II

- The time of my office hours has changed.
  - Thursday 10:00am - 12:00pm in ATH 3-08.
- On the midterm:
  - It is marked, I need to put the grades in eClass.
- **I want your feedback!**
  - Mid-term Course and Instruction Feedback online evaluation opened today.
  - It will be open for this week only: Oct 16 – Oct 20.
  - 4 of 90 students responded so far 😭

Plan / Reminder III

- What **I** plan to do today:
  - Finish talking about Monte Carlo Methods for Prediction & Control (Chapter 5 of the textbook).
  - Some exercises.
- What I recommend **YOU** to do for Friday:
  - Read Chapter 6 up to Section 6.3.
  - Practice Quiz (Advantages of TD).

Please, interrupt me at any time!



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

Last class: 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, $\mathbb{E}[G_t \mid S_t = s] = v_b(s)$.

Last class: 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.

Last class: Importance Sampling

In RL, the probability of a trajectory is:

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

Last class: Importance Sampling

In RL, the probability of a trajectory is:

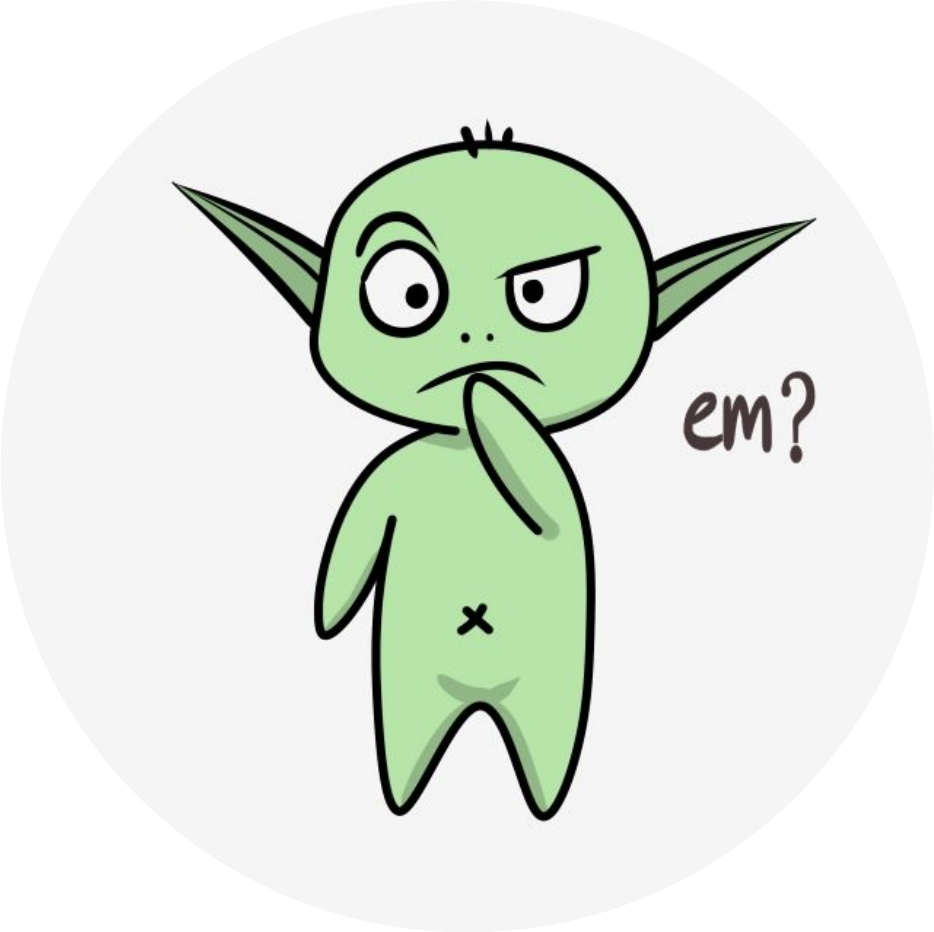
$$\begin{aligned}
&\Pr\{A_t, S_{t+1}, A_{t+1}, \ldots, S_T \mid S_t, A_{t:T-1} \sim \pi\} \\
&\quad = \pi(A_t|S_t)p(S_{t+1}|S_t, A_t)\pi(A_{t+1}|S_{t+1}) \cdots p(S_T|S_{T-1}, A_{T-1}) \\
&\quad = \prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k),
\end{aligned}$$

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

We require *coverage*:
$b(a|s) > 0$ when $\pi(a|s) > 0$

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

The IS ratio does not depend on the MDP, that is, on $p(s', r \mid s, a)$!



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

Last class: 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:

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

Incremental update (Weighted IS)

We want to form the estimate

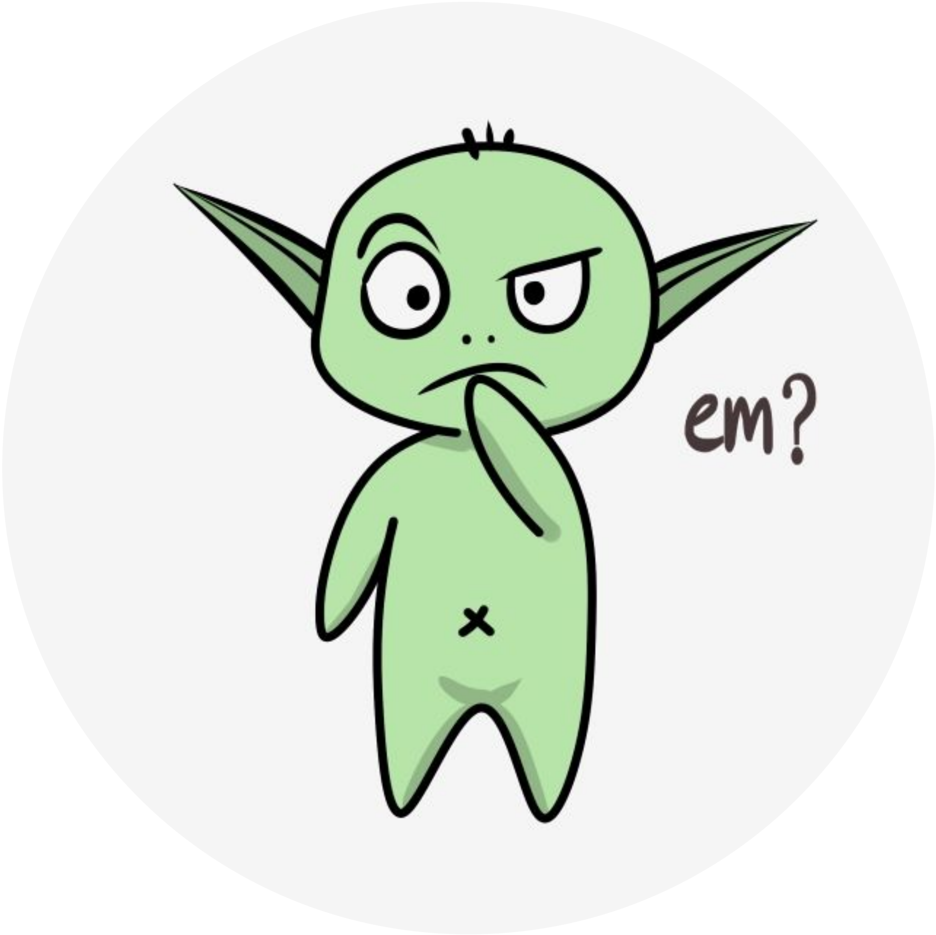
$$V_n \doteq \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}, \qquad n \geq 2. \qquad\qquad W_i = \rho_{t_i:T(t_i)-1}$$

The update rule for V_n is

$$V_{n+1} \doteq V_n + \frac{W_n}{C_n}\Big[G_n - V_n\Big], \qquad n \geq 1,$$

and

$$C_{n+1} \doteq C_n + W_{n+1}$$



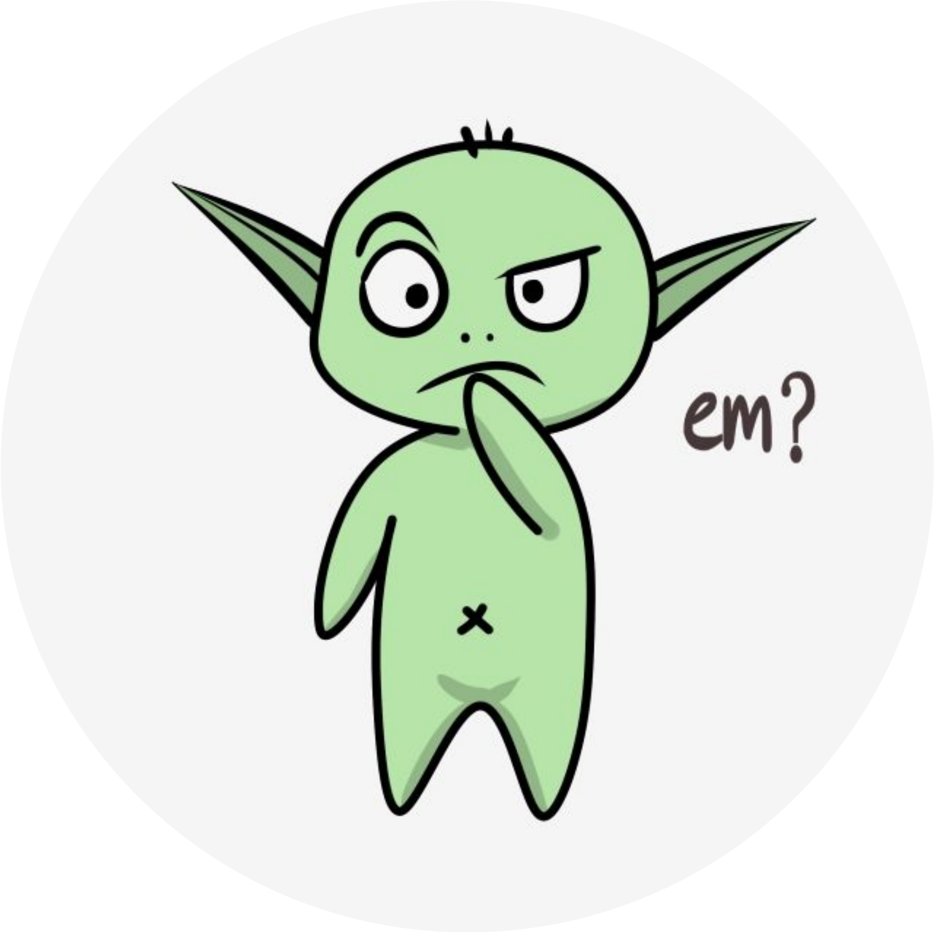
https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Off-policy MC prediction for estimating q_π

Off-policy MC prediction (policy evaluation) for estimating $Q \approx q_\pi$

Input: an arbitrary target policy π
Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s, a) \in \mathbb{R}$ (arbitrarily)
    $C(s, a) \leftarrow 0$

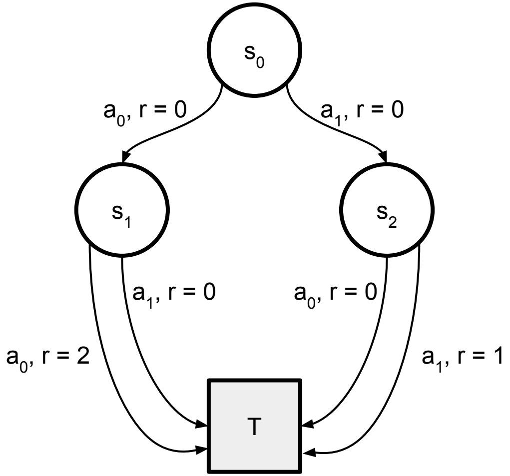
Loop forever (for each episode):
    $b \leftarrow$ any policy with coverage of π
    Generate an episode following b: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
    $G \leftarrow 0$
    $W \leftarrow 1$
    Loop for each step of episode, $t = T-1, T-2, \ldots, 0$, while $W \neq 0$:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
        $W \leftarrow W \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$



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

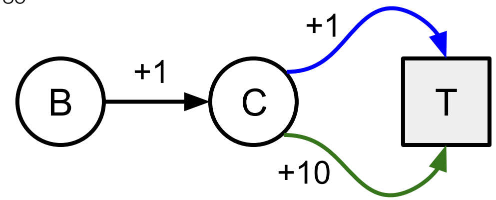
Practice Exercise 1

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

Practice Exercise 2



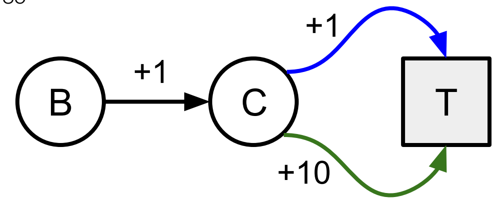
Off-policy Monte Carlo Prediction allows us to use sample trajectories to estimate the value function for a policy that may be different than the one used to generate the data. Consider the following MDP, with two states, B and C, with 1 action in state B and two actions in state C, with γ = 1.0. In 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 π(A = 1 | C) = 0.9 and π(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.

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

Practice Exercise 3

Let $\rho_t = \dfrac{\pi(A_t \mid S_t)}{b(A_t \mid S_t)}$.

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

Hint: $r(s, a) = \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a] = \Sigma_r\, r\, \Sigma_{s'}\, p(s', r \mid s, a)$.