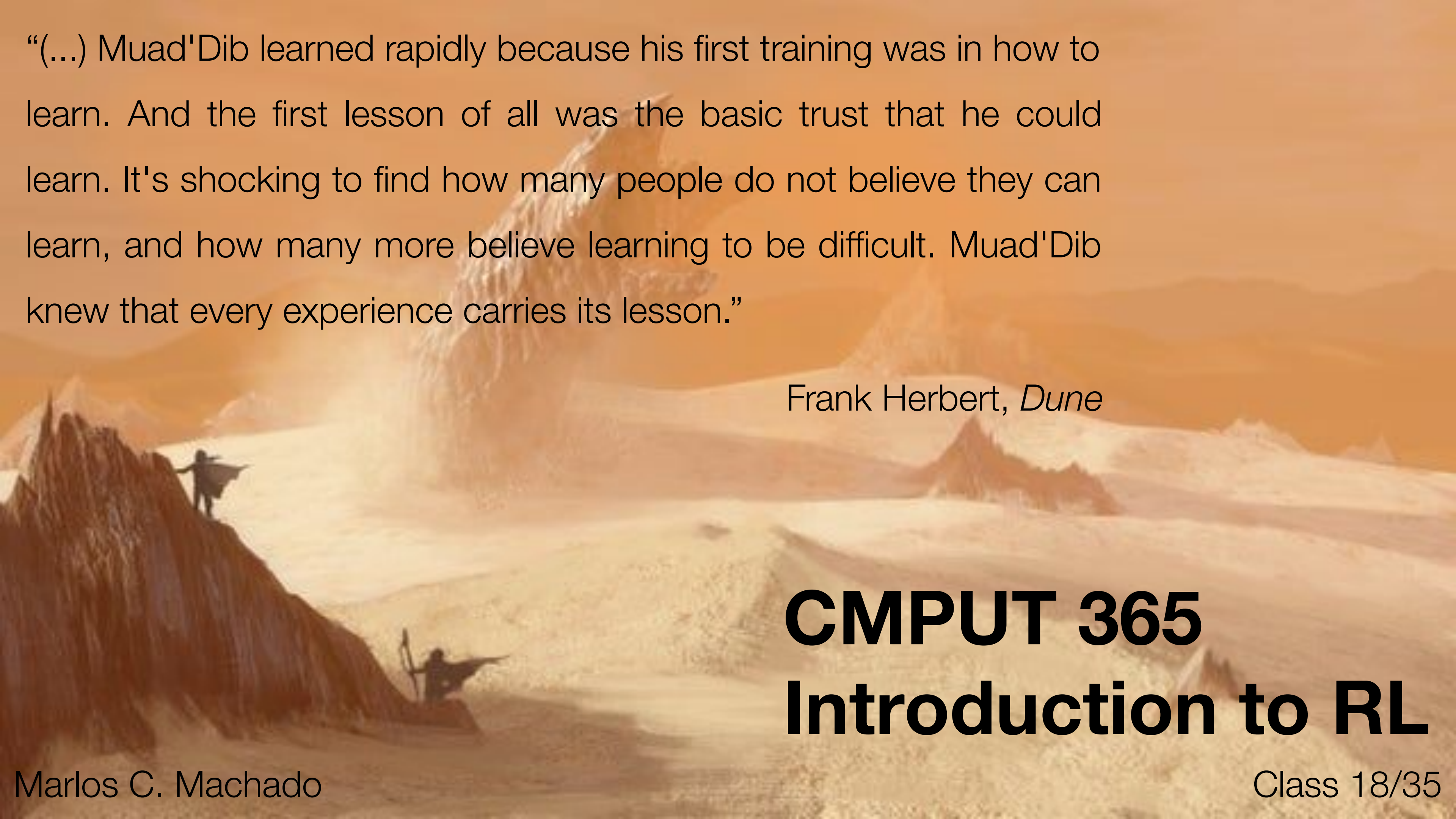


“(...) Muad'Dib learned rapidly because his first training was in how to learn. And the first lesson of all was the basic trust that he could learn. It's shocking to find how many people do not believe they can learn, and how many more believe learning to be difficult. Muad'Dib knew that every experience carries its lesson.”

Frank Herbert, *Dune*

CMPUT 365 Introduction to RL

Marlos C. Machado

Class 18/35

Coursera 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.

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`.

Reminders and Notes

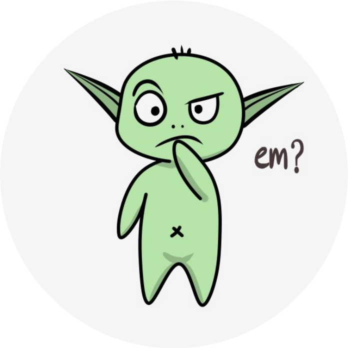
- What I plan to do today:
  - Exercises and Examples
- Useful information for you:
  - For the next class, read the rest of Chapter 6
  - Practice Quiz (Expected Sarsa) is due on Wednesday

SPOT: Mid-term Course Evaluation

Overall, it was quite positive 😅

… but

Please, interrupt me at any time!



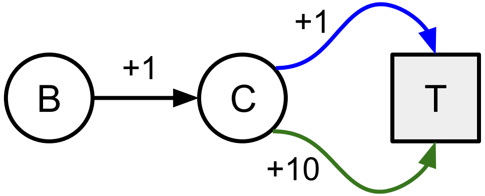
https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Chapter 6

Temporal-Difference Learning

Prediction

MC Methods: 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.

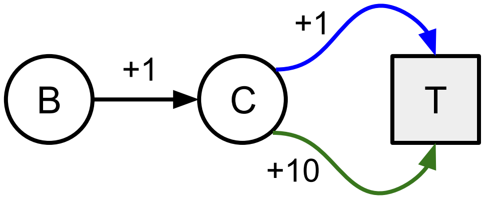
a) What are the true values v_π?

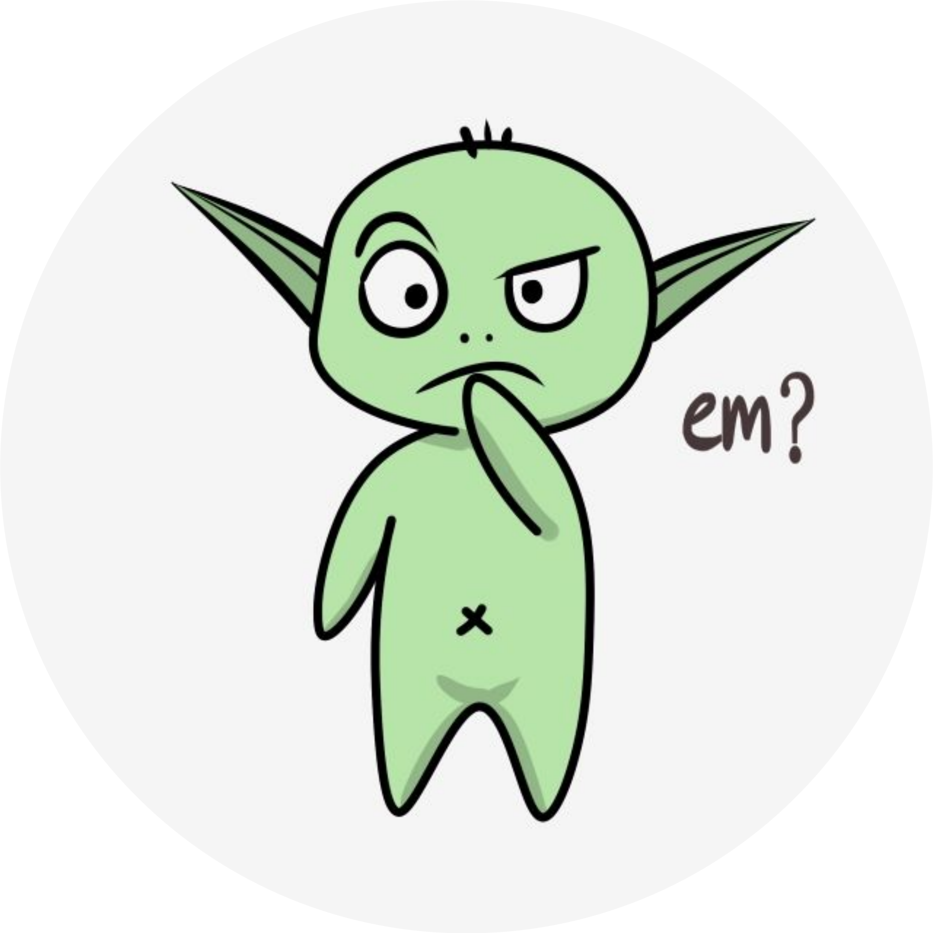
b) Imagine you got to execute *π* in the environment for one episode, and observed the episode trajectory S_0= B, A_0= 1, R_1= 1, S_1= C, A_1= 1, R_2= 1. What is the return for B for this episode? Additionally, what are the value estimates V_π, using this one episode with Monte Carlo updates?

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 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 estimate 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?

MC Methods: Practice Exercise 2



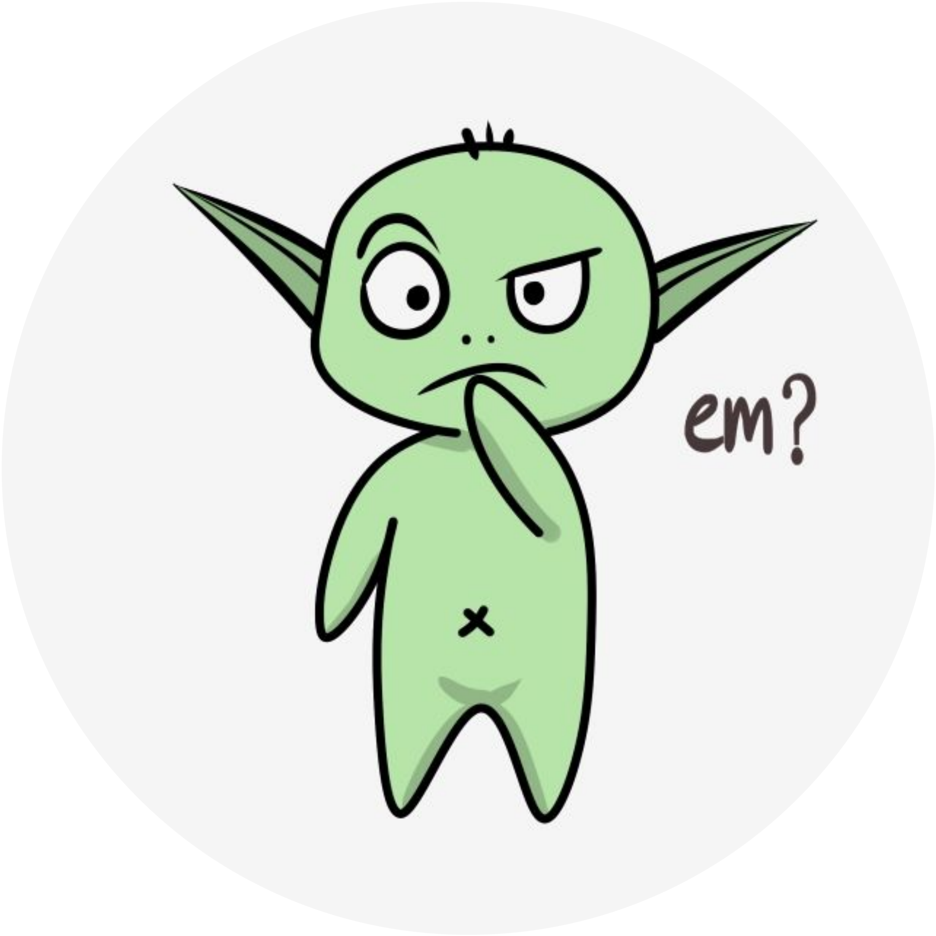


https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

A note/clarification on the use of importance sampling

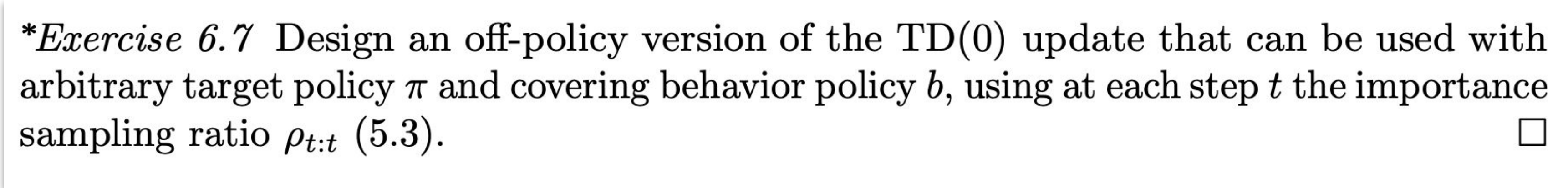
$$V(s) \doteq \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{|\mathcal{T}(s)|}$$

Note to students who are just seeing this slide but missed class:
This is important. If you missed this class, ask a classmate about this.

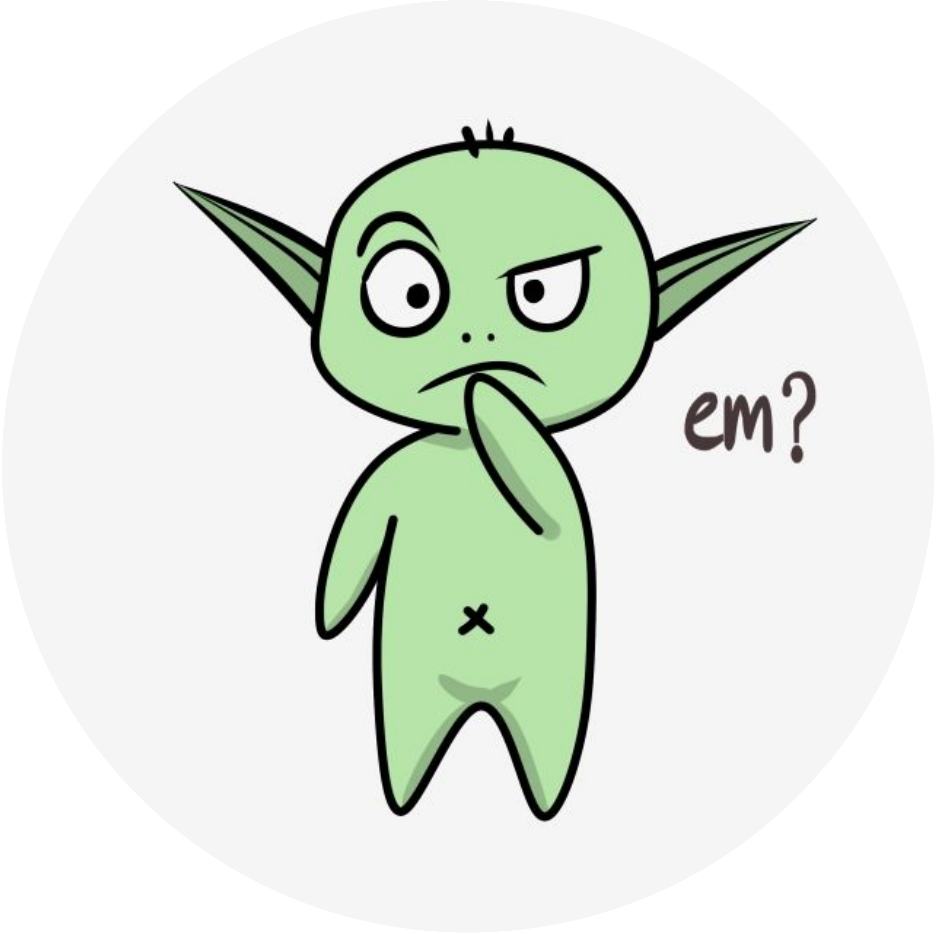


https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Exercise 6.7

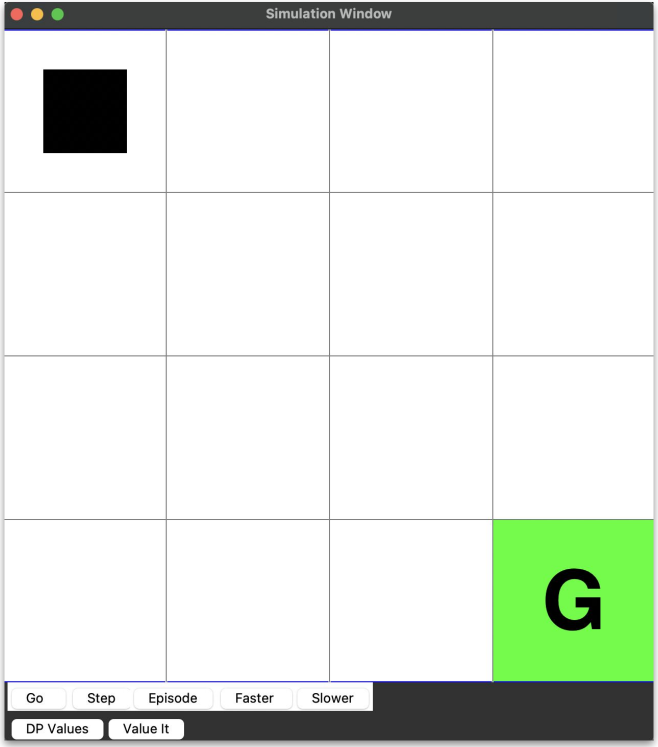


**Exercise 6.7* Design an off-policy version of the TD(0) update that can be used with arbitrary target policy π and covering behavior policy b, using at each step t the importance sampling ratio $\rho_{t:t}$ (5.3). □



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

Demo


Simulation Window
G
Go
Step
Episode
Faster
Slower
DP Values
Value It