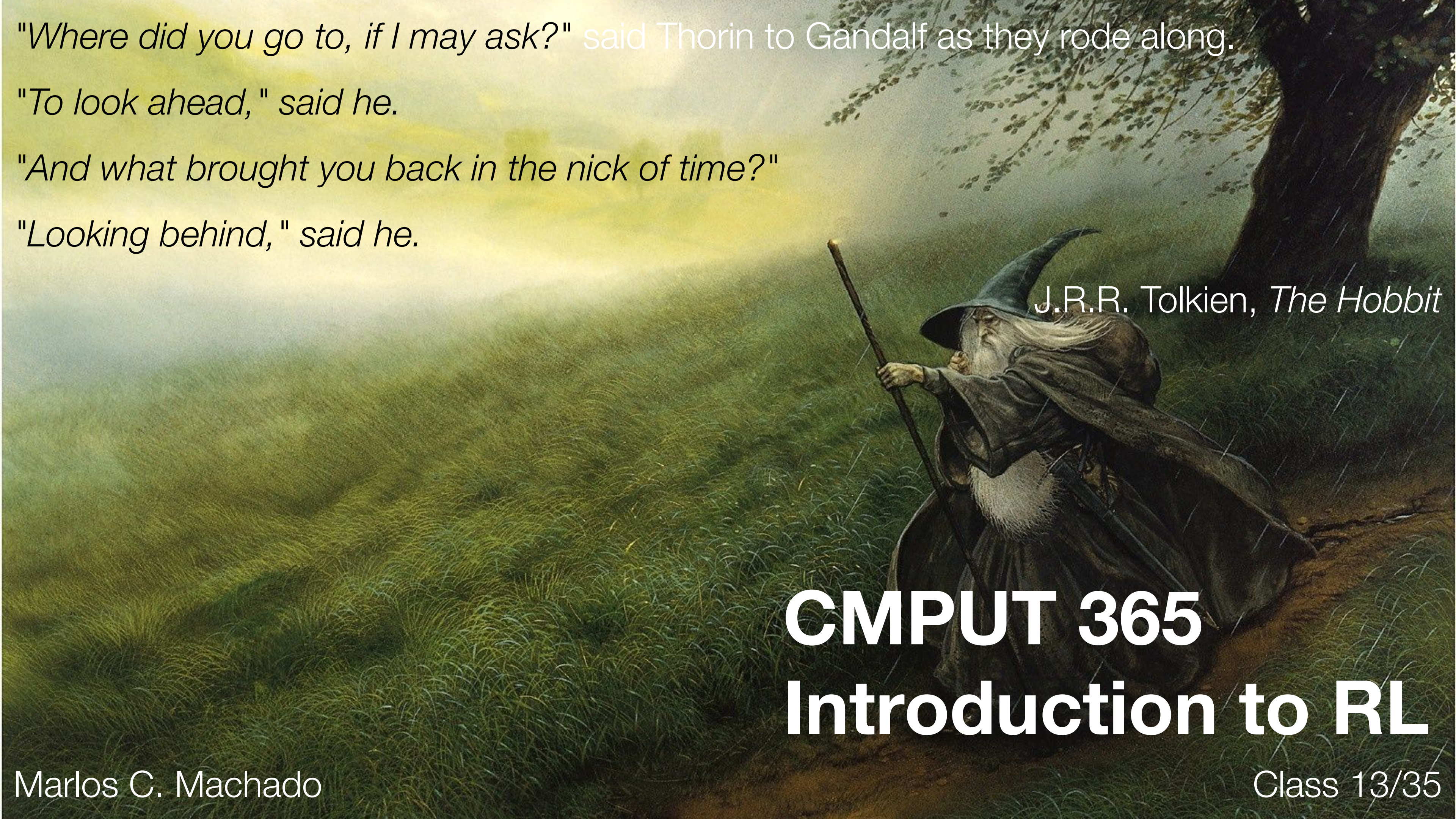


"Where did you go to, if I may ask?" said Thorin to Gandalf as they rode along.

"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 13/35

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.

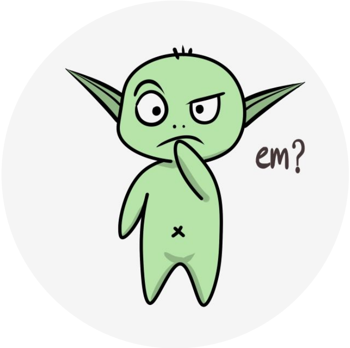
There were **20 pending invitations** last time I checked!

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

Reminders and Notes

- On the midterm:
  - It is marked, there are only a few left. You should have the marks by Wednesday.
- What **I** plan to do today:
  - Where are we?
  - Overview of Monte Carlo Methods for Prediction & Control (Chapter 5 of the textbook).
- What I recommend **YOU** to do for next class:
  - Read Chapter 5 up to Section 5.5.
  - Graded Quiz (Off-policy Monte Carlo).
  - Programming Assignment is not graded this week.

Please, interrupt me at any time!



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

Interlude

An overview

- Main features of a reinforcement learning problem:
  - Trial-and-error learning
  - Exploration
  - Delayed credit assignment

An overview

- Main features of a reinforcement learning problem:
  - Trial-and-error learning
  - Exploration
  - Delayed credit assignment

A flavour of RL: Bandits (Chapter 2)

An overview

- Main features of a reinforcement learning problem:
  - Trial-and-error learning
  - Exploration
  - Delayed credit assignment ←

But what does that mean?
What is this sequential decision-making problem we are trying to solve?
What does solution mean here?

A problem formulation: MDPs (Chapter 3)

An overview

- Main features of a reinforcement learning problem:
  - Trial-and-error learning
  - Exploration
  - Delayed credit assignment

- What about the solution?

A first solution: Dynamic Programming (Chapter 4)

An overview

- Main features of a reinforcement learning problem:
    - Trial-and-error learning
    - Exploration
    - Delayed credit assignment

- What about the solution?
    - Dynamic programming! ← We need to know $p(s', r \mid s, a)$ and it can be computationally expensive to solve the system of linear equations.

Our first learning algorithm: Monte Carlo Methods (Chapter 5)

Chapter 5

Monte Carlo Methods

Monte Carlo Methods – Why?

- This is our **first learning** method.
- We do not assume complete knowledge of the environment.
- "Monte Carlo methods **require only experience** — sample sequences of states, actions, and rewards from actual or simulated interaction with an environment." 🤯
- It works! And different variations are used everywhere in the field (n-step returns, TD(λ), MCTS–AlphaGo/AlphaZero–, etc).
- … but we still need a model, albeit only a sample model.

MC Methods are ways of solving the RL problem based on avg. sample returns (similar to bandits, but instead of rewards we are sampling returns).

Monte Carlo Prediction

First-visit MC prediction, for estimating $V \approx v_\pi$

Input: a policy π to be evaluated

Initialize:

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

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

Loop forever (for each episode):

    Generate an episode following π: $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$

    $G \leftarrow 0$

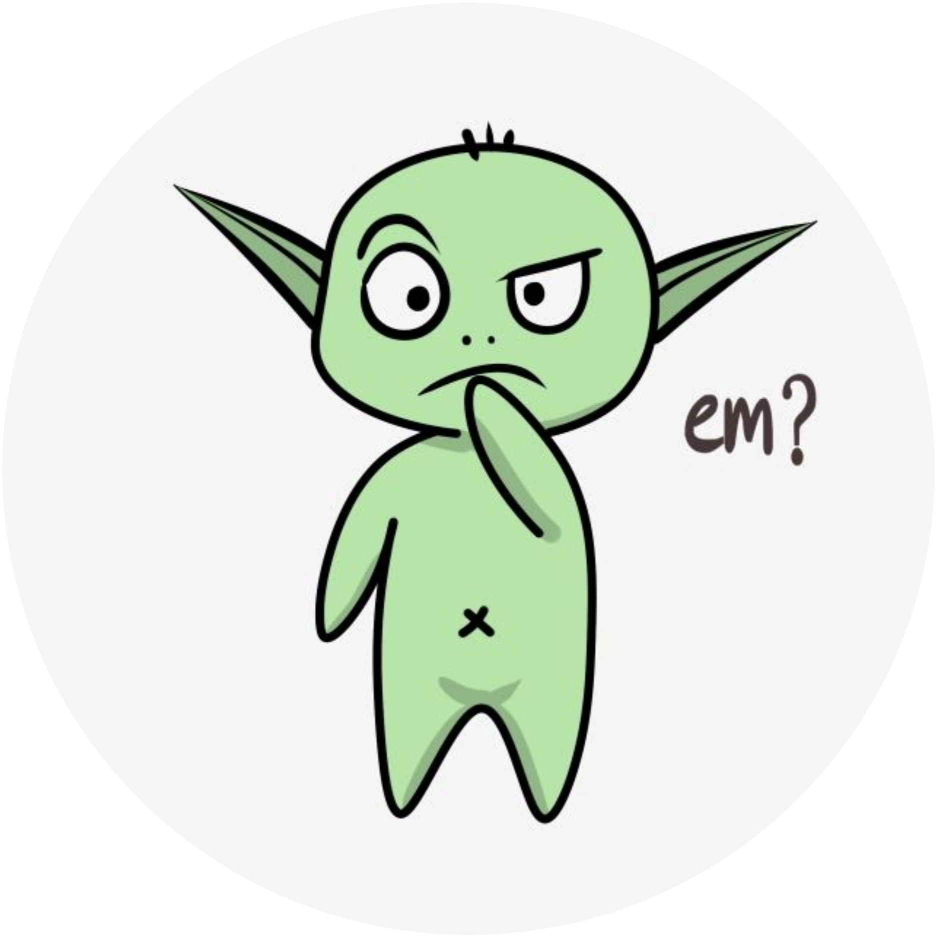
    Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

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

        Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:

            Append G to $Returns(S_t)$

            $V(S_t) \leftarrow \text{average}(Returns(S_t))$



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