“The rotten tree-trunk, until the very moment when the storm-blast breaks it in two, has all the appearance of might it ever had.”

Isaac Asimov, *Foundation*

CMPUT 365 Introduction to RL

Marlos C. Machado

Class 7/ 35

Plan

- Value Functions and Bellman Equations
  - A roadmap to the course
  - Non–comprehensive overview
  - We are still not talking about solution methods, we are only formalizing things

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.

Some students who are enrolled in Coursera **haven't submitted any quizzes or assignments** in the private session, and that's all I can see.

The deadlines in the public session **do not align** with the deadlines in Coursera.

Plan

- Value Functions and Bellman Equations
  - Non–comprehensive overview

Please, interrupt me at any time!

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

Why? Where are we?! We need a roadmap.

- Reinforcement learning is about solving sequential decision-making problems from interactions with the environment.
  - Key features:
    - Trial-and-error
    - Exploration-exploitation trade-off
    - Delayed credit-assignment

Why? Where are we?! We need a roadmap.

- Reinforcement learning is about solving sequential decision-making problems from interactions with the environment.
  - Key features:
    - Trial-and-error
    - Exploration-exploitation trade-off
    - Delayed credit-assignment
- That's too abstract! Can we be more concrete and start from a simple example?

Why? Where are we?! We need a roadmap.

- Reinforcement learning is about solving sequential decision-making problems from interactions with the environment.
  - Key features:
    - Trial-and-error
    - Exploration-exploitation trade-off
    - Delayed credit-assignment
- That's too abstract! Can we be more concrete and start from a simple example?
  - Yes! Bandits.

Chapter 2 of the textbook
Week 1 of *Fundamentals of RL*

Why? Where are we?! We need a roadmap.

- Reinforcement learning is about solving sequential decision-making problems from interactions with the environment.
    - Key features:
        - Trial-and-error
        - Exploration-exploitation trade-off
        - Delayed credit-assignment
- That's too abstract! Can we be more concrete and start from a simple example?
    - Yes! Bandits.
- What if actions have consequences? What's a sequential decision-making problem? What does "solving" a sequential decision-making problem means?

Why? Where are we?! We need a roadmap.

- Reinforcement learning is about solving sequential decision-making problems from interactions with the environment.
  - Key features:
    - Trial-and-error
    - Exploration-exploitation trade-off
    - Delayed credit-assignment
- That's too abstract! Can we be more concrete and start from a simple example?
  - Yes! Bandits.
- What if actions have consequences? What's a sequential decision-making problem? What does "solving" a sequential decision-making problem means?
  - We need a formal language for that: MDPs.

Chapter 3 of the textbook
Weeks 2 & 3 of *Fundamentals of RL*

Why? Where are we?! We need a roadmap.

- How can we do that?

Why? Where are we?! We need a roadmap.

- How can we do that?
  - We can leverage Bellman equations and do Dynamic Programming.

Chapter 4 of the textbook
Week 4 of *Fundamentals of RL*

Why? Where are we?! We need a roadmap.

- How can we do that?
  - We can leverage Bellman equations and do Dynamic Programming.
- But what if you don't know how the world works (you don't know $p(s', r \mid s, a)$)?

Why? Where are we?! We need a roadmap.

- How can we do that?
  - We can leverage Bellman equations and do Dynamic Programming.
- But what if you don't know how the world works (you don't know $p(s', r \mid s, a)$)?
  - Well, we can use Monte Carlo methods.

Chapter 5 of the textbook
Week 2 of *Sample-based Learning Methods*

Why? Where are we?! We need a roadmap.

- How can we do that?
  - We can leverage Bellman equations and do Dynamic Programming.
- But what if you don't know how the world works (you don't know $p(s', r \mid s, a)$)?
  - Well, we can use Monte Carlo methods.
- Do we really need to wait until episodes are over to learn something? What about continuing tasks?

Why? Where are we?! We need a roadmap.

- How can we do that?
  - We can leverage Bellman equations and do Dynamic Programming.
- But what if you don't know how the world works (you don't know $p(s', r \mid s, a)$?
  - Well, we can use Monte Carlo methods.
- Do we really need to wait until episodes are over to learn something? What about continuing tasks?
  - Nope! Temporal-difference learning.

Chapter 6 of the textbook
Weeks 3 & 4 of *Sample-based Learning Methods*

Why? Where are we?! We need a roadmap.

- How can we do that?
  - We can leverage Bellman equations and do Dynamic Programming.
- But what if you don't know how the world works (you don't know $p(s', r \mid s, a)$?
  - Well, we can use Monte Carlo methods.
- Do we really need to wait until episodes are over to learn something? What about continuing tasks?
  - Nope! Temporal-difference learning.
- Can't we learn more efficiently? Can we only learn from interactions with the environment?

Why? Where are we?! We need a roadmap.

- How can we do that?
  - We can leverage Bellman equations and do Dynamic Programming.
- But what if you don't know how the world works (you don't know $p(s', r \mid s, a)$?
  - Well, we can use Monte Carlo methods.
- Do we really need to wait until episodes are over to learn something? What about continuing tasks?
  - Nope! Temporal-difference learning.
- Can't we learn more efficiently? Can we only learn from interactions with the environment?
  - We can be more efficient, we can do planning alongside learning.

Chapter 8 of the textbook
Week 5 of *Sample-based Learning Methods*

Why? Where are we?! We need a roadmap.

- But what if we have many (maybe infinite) states? This doesn't scale!

Why? Where are we?! We need a roadmap.

- But what if we have many (maybe infinite) states? This doesn't scale!
  - We then do function approximation.

Chapters 9 & 10 of the textbook
Weeks 1, 2, & 3 of *Prediction and Control with Function Approximation*

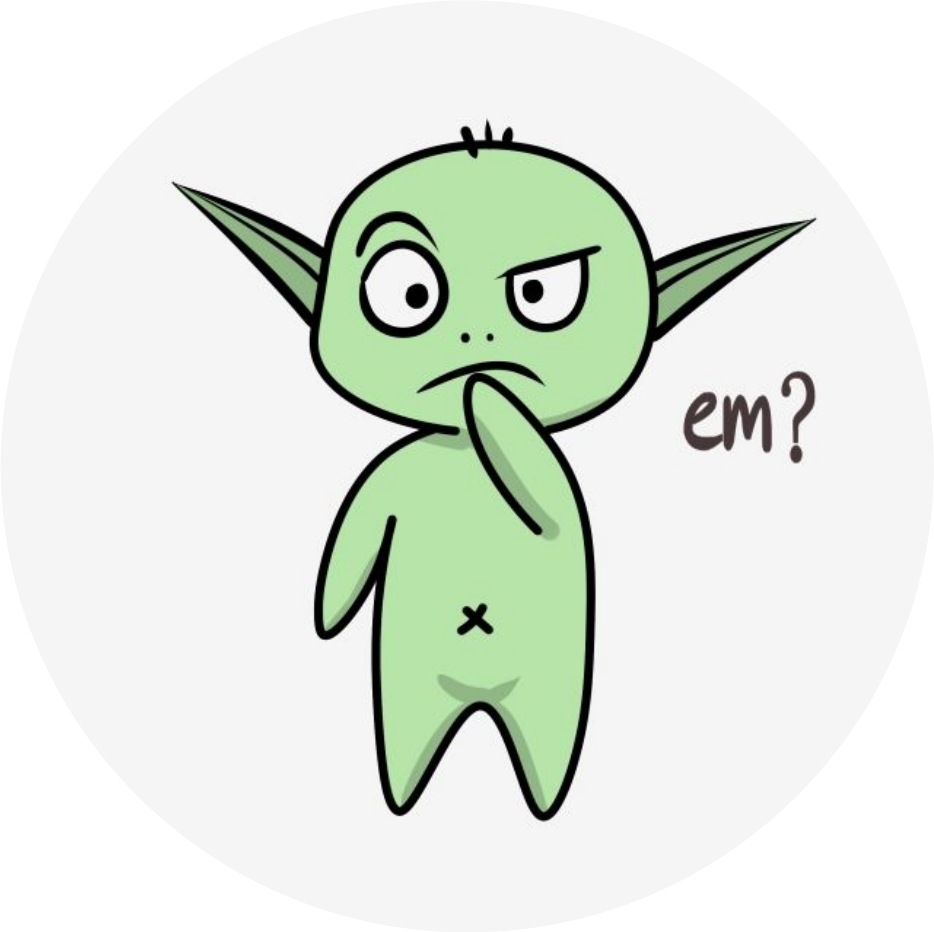
Why? Where are we?! We need a roadmap.

- But what if we have many (maybe infinite) states? This doesn't scale!
  - We then do function approximation.
- What about many (maybe infinite) actions?

Why? Where are we?! We need a roadmap.

- But what if we have many (maybe infinite) states? This doesn't scale!
  - We then do function approximation.
- What about many (maybe infinite) actions?
  - A way to tackle this problem is with policy gradient methods.

Chapter 13 of the textbook
Week 4 of *Prediction and Control with Function Approximation*



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

Value Functions and Policies

- *Value functions are "functions of states (or state-action pairs) that estimate how good it is for the agent to be in a given state".*
- "How good" means expected return.
- Expected returns depend on how the agent behaves, that is, its *policy*.

Policy

- A policy is a mapping from states to probabilities of selecting each possible action:

$$\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$$

in other words, $\pi(a|s)$ is the probability that $A_t = a$ if $S_t = s$.

Exercise 3.11 If the current state is S_t, and actions are selected according to a stochastic policy π, then what is the expectation of R_{t+1} in terms of π and the four-argument function p (3.2)? $\square$

Value Function

- The value function of a state s under a policy π, denoted $v_\pi(s)$ is the expected return when starting in s and following π thereafter.

state-value function for policy π

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right]$$

$$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$$

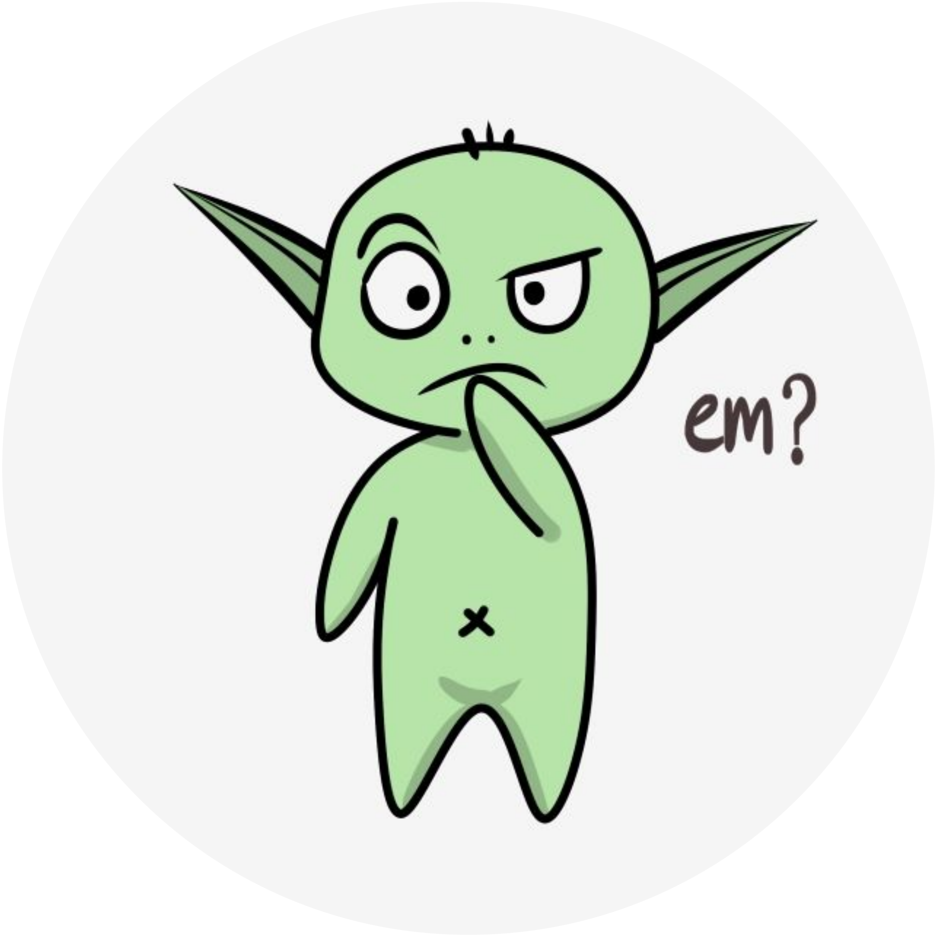
action-value function for policy π

Why is this difference important?

Exercises from the Textbook

Exercise 3.12 Give an equation for v_π in terms of q_π and π. $\square$

Exercise 3.13 Give an equation for q_π in terms of v_π and the four-argument p. $\square$



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

Next class

- What I plan to do:
  - Exercises and Examples

- What I recommend YOU to do for next class:
  - Submit Graded Quiz for Fundamental of RL: Value functions & Bellman equations (Week 3).