“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 11/ 35

Plan

- Value Functions and Bellman Equations
  - 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.

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

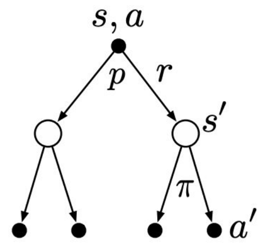
If you have any questions or concerns, **talk with the TAs** or email us `cmput365@ualberta.ca`.

Please, interrupt me at any time!

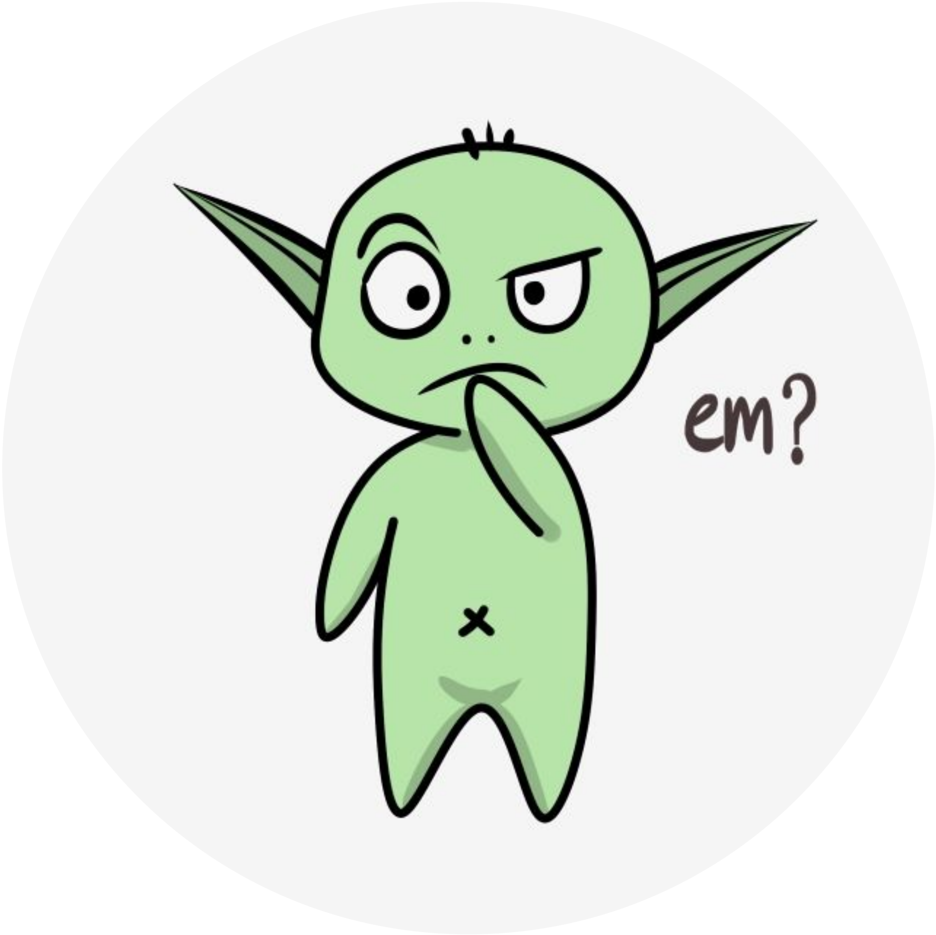
https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

State-Action Value Functions Satisfy Recursive Relationships

Exercise 3.17 What is the Bellman equation for action values, that is, for q_π? It must give the action value $q_\pi(s, a)$ in terms of the action values, $q_\pi(s', a')$, of possible successors to the state–action pair (s, a). Hint: The backup diagram to the right corresponds to this equation. Show the sequence of equations analogous to (3.14), but for action values. □



q_π backup diagram



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

Optimal Policies and Optimal Value Functions

- Value functions define a partial ordering over policies.
  - $\pi \geq \pi'$ iff $v_\pi(s) \geq v_{\pi'}(s)$ for all $s \in \mathcal{S}$.
  - There is always at least one policy that is better than or equal to all other policies. The *optimal policy*.

$$v_*(s) \doteq \max_\pi v_\pi(s)$$

$$q_*(s,a) \doteq \max_\pi q_\pi(s,a)$$

$$q_*(s,a) = \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a]$$

Optimal Policies and Optimal Value Functions

- Because v_* is the value function for a policy, it must satisfy the self-consistency condition given by the Bellman equation for state values.

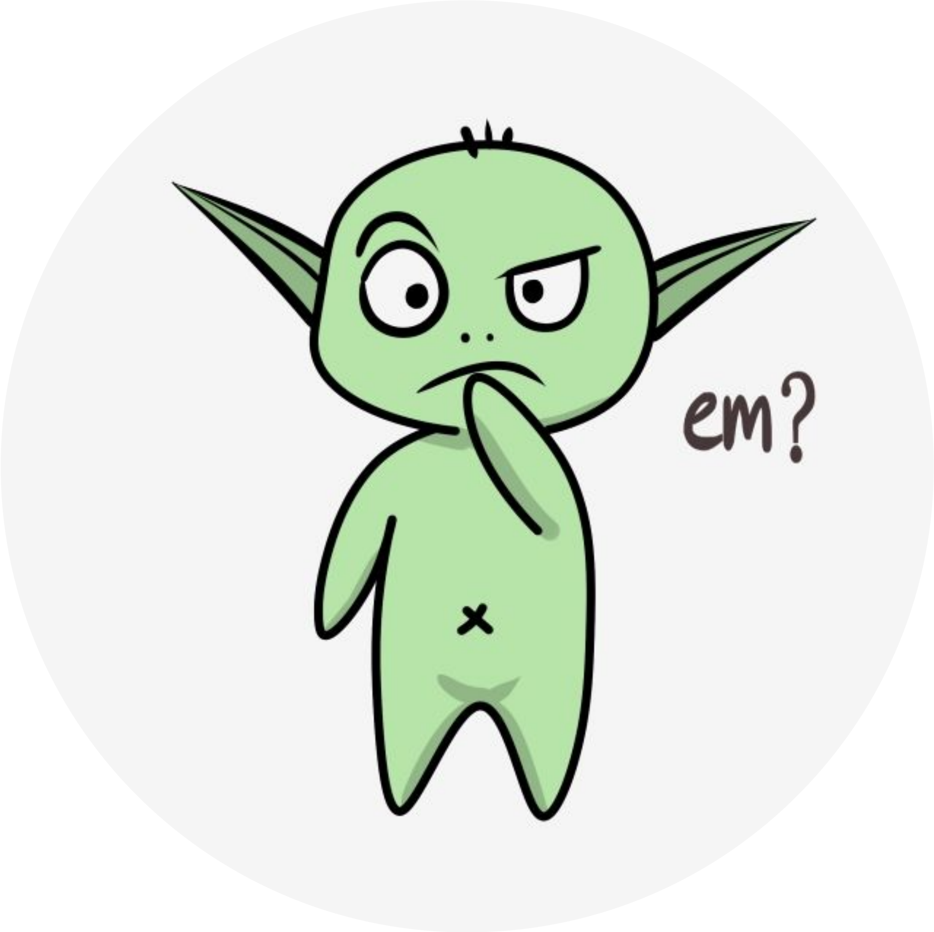
$$v_*(s) = \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s, a)$$

Optimal Policies and Optimal Value Functions

- Because v_* is the value function for a policy, it must satisfy the self-consistency condition given by the Bellman equation for state values.

$$\begin{aligned} v_*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a) \\ &= \max_a \mathbb{E}_{\pi_*}[G_t \mid S_t = s, A_t = a] \\ &= \max_a \mathbb{E}_{\pi_*}[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a] \\ &= \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a] \\ &= \max_a \sum_{s',r} p(s',r|s,a)\big[r + \gamma v_*(s')\big]. \end{aligned}$$

$$\begin{aligned} q_*(s,a) &= \mathbb{E}\Big[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1},a') \Bigm| S_t = s, A_t = a\Big] \\ &= \sum_{s',r} p(s',r|s,a)\Big[r + \gamma \max_{a'} q_*(s',a')\Big]. \end{aligned}$$



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

Reinforcement learning is very related to search algorithms

“Heuristic search methods can be viewed as expanding the right-hand side of the equation below several times, up to some depth, forming a “tree” of possibilities, and then using a heuristic evaluation function to approximate v_*, at the “leaf” nodes.”

$$v_*(s) = \max_a \sum_{s',r} p(s', r | s, a) \big[r + \gamma v_*(s') \big].$$

Yay! We solved sequential decision-making problems

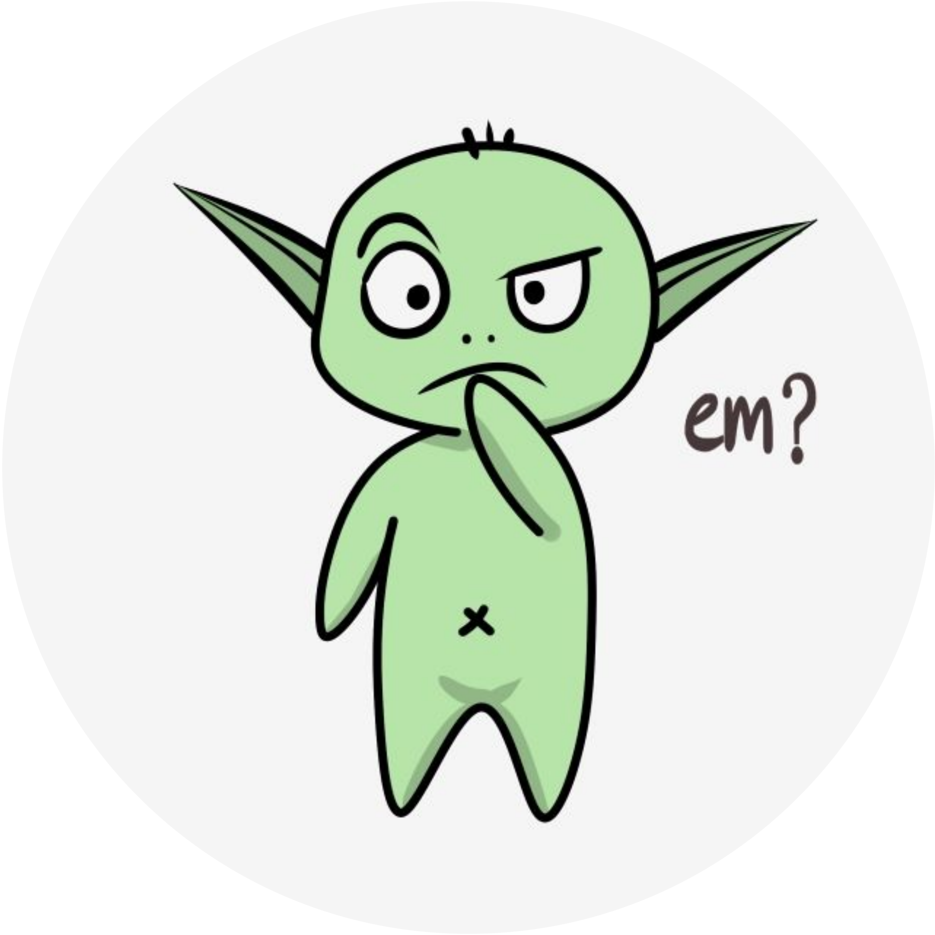
Except…

1.
2.
3.

Yay! We solved sequential decision-making problems

Except…

1. we need to know the dynamics of the environment
2. we have enough computational resources to solve the system of linear eq.
3. the Markov property



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

Exercises from the Textbook

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)? □

Exercise 3.12 Give an equation for v_π in terms of q_π and π. □

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

Next class

- What I plan to do:
  - Non-comprehensive overview of Dynamic Programming (Chapter 4 of the textbook)

- What I recommend YOU to do for next class:
  - Read (most of) Chapter 4 of the textbook.
  - Submit Practice Quiz for Fundamental of RL: Dynamic Programming (Week 4).

Marlos C. Machado