“All have their worth,” said Yavanna,

“and each contributes to the worth of the others”.

J.R.R. Tolkien, *The Silmarillion*

CMPUT 365
Introduction to RL

Marlos C. Machado

Class 14/35

Plan

- Dynamic programming
  - Value iteration and Generalized Policy Improvement
-

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.

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`.

Reminder II

- Midterm is next Wednesday, after Thanksgiving.
  - There will be an attendance sheet.
  - You need to have your OneCard with you and show it to us.
- If you are not enrolled in Coursera's private session, Sample-based Learning Methods, please enroll!

Please, interrupt me at any time!

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

Last class: Policy Iteration

Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

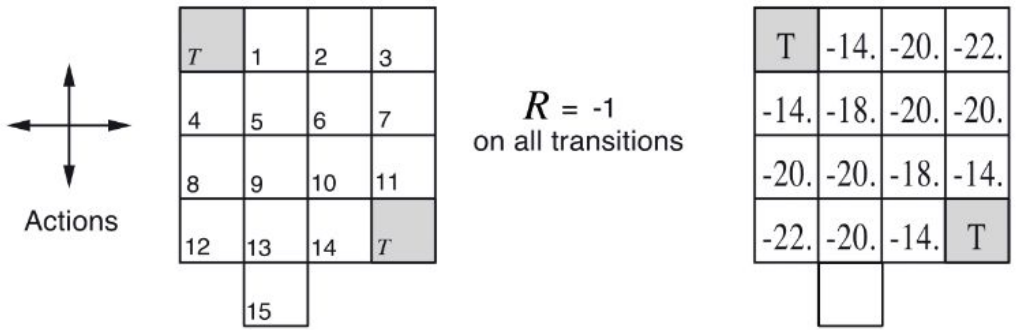
1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$; $V(\textit{terminal}) \doteq 0$

2. Policy Evaluation
   Loop:
   $\Delta \leftarrow 0$
   Loop for each $s \in \mathcal{S}$:
   $v \leftarrow V(s)$
   $V(s) \leftarrow \sum_{s',r} p(s', r \mid s, \pi(s)) \big[r + \gamma V(s')\big]$
   $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
   old-action $\leftarrow \pi(s)$
   $\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s', r \mid s, a) \big[r + \gamma V(s')\big]$
   If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

Quiz Question – Week 4

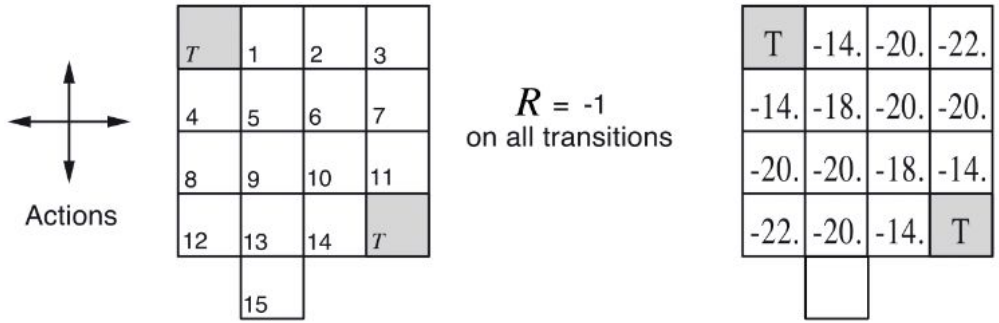
9. Consider the undiscounted, episodic MDP below. There are four actions possible in each state, A = {up, down, right, left}, which deterministically cause the corresponding state transitions, except that actions that would take the agent off the grid in fact leave the state unchanged. The right half of the figure shows the value of each state under the equiprobable random policy. If π is the equiprobable random policy, what is $q(11, \text{down})$?



- ○ $q(11, \text{down}) = -1$
- ○ $q(11, \text{down}) = -15$
- ○ $q(11, \text{down}) = -14$
- ○ $q(11, \text{down}) = 0$

Quiz Question – Week 4

9. Consider the undiscounted, episodic MDP below. There are four actions possible in each state, A = {up, down, right, left}, which deterministically cause the corresponding state transitions, except that actions that would take the agent off the grid in fact leave the state unchanged. The right half of the figure shows the value of each state under the equiprobable random policy. If π is the equiprobable random policy, what is $q(7, \text{down})$?



- $q(7, \text{down}) = -14$
- $q(7, \text{down}) = -20$
- $q(7, \text{down}) = -21$
- $q(7, \text{down}) = -15$

Value Iteration

Value Iteration, for estimating $\pi \approx \pi_*$

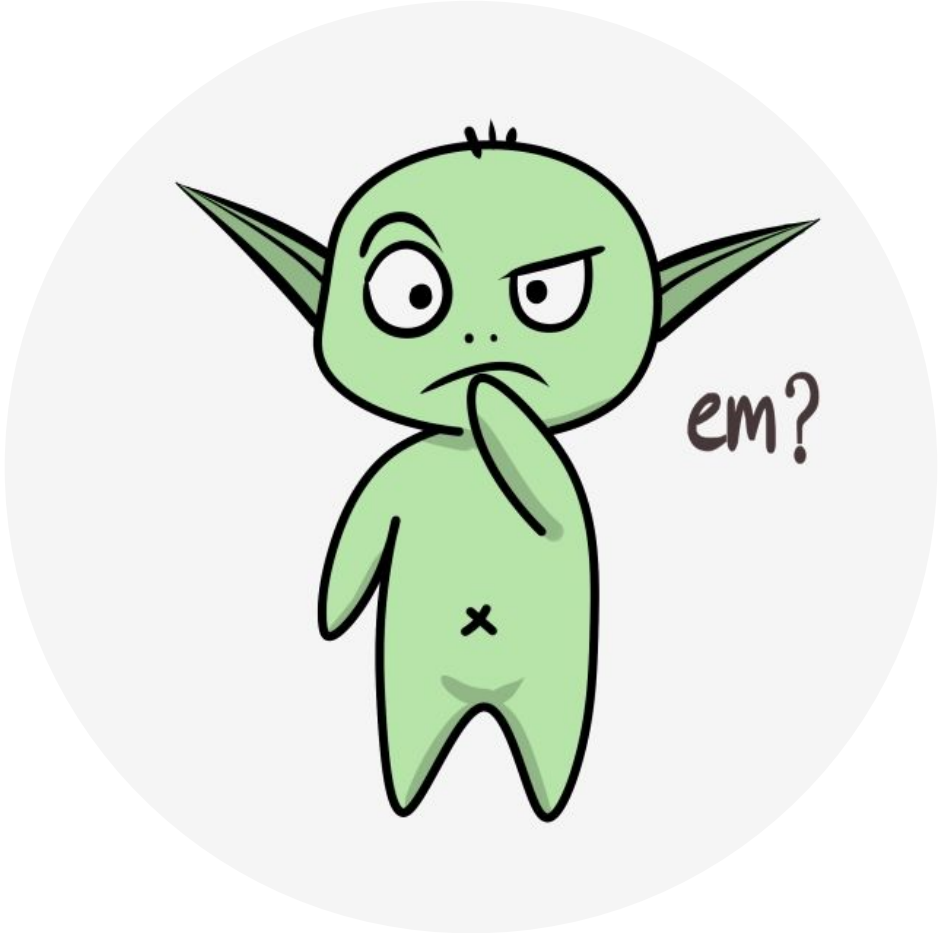
Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
| $\Delta \leftarrow 0$
| Loop for each $s \in \mathcal{S}$:
| | $v \leftarrow V(s)$
| | $V(s) \leftarrow \max_a \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma V(s')\right]$
| | $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
$\pi(s) = \arg\max_a \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma V(s')\right]$

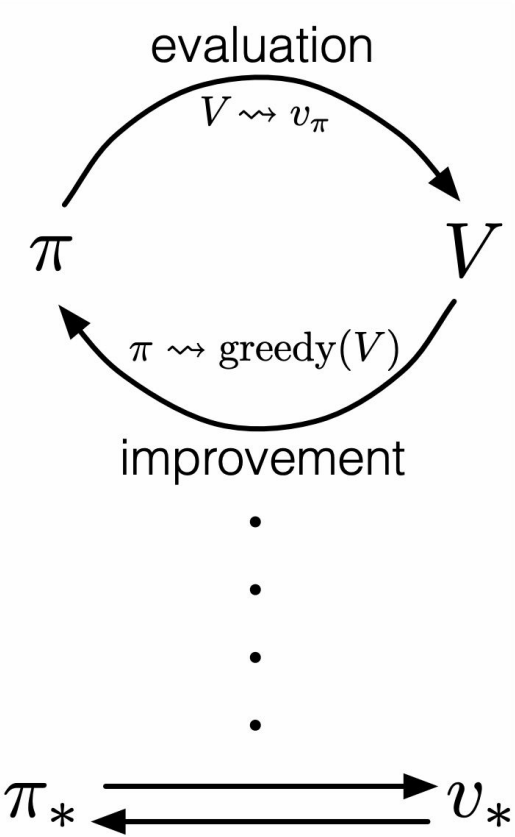
It doesn't need to be so synchronous

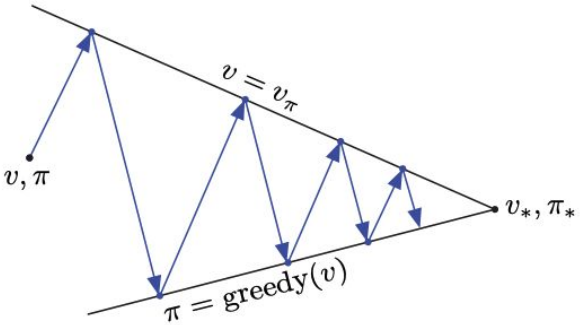
We just turned the Bellman optimality equation into an update rule!

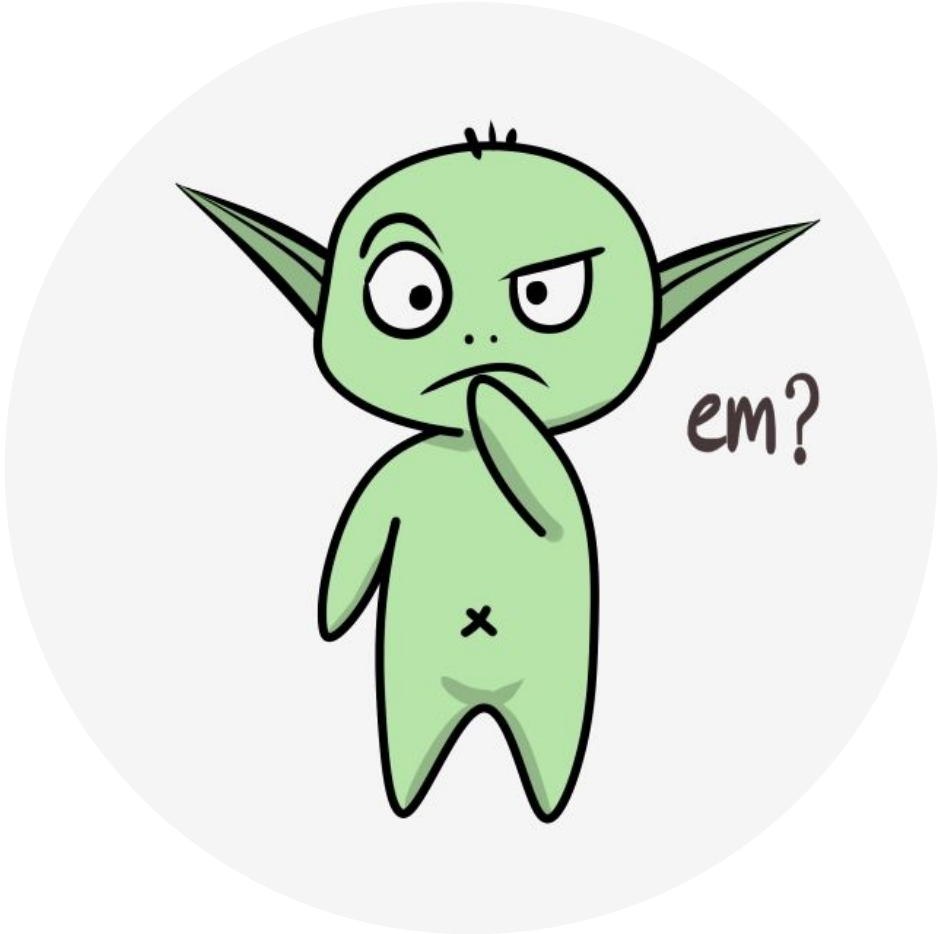


https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Generalized Policy Iteration







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

Quiz Question – Week 4

4. What is the relationship between value iteration and policy iteration? [Select all that apply]

- ☐ Policy iteration is a special case of value iteration.
- ☐ Value iteration and policy iteration are both special cases of generalized policy iteration.
- ☐ Value iteration is a special case of policy iteration.

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. □

Exercises from the Textbook

Exercise 3.25 Give an equation for v_* in terms of q_*. □

Exercise 3.26 Give an equation for q_* in terms of v_* and the four-argument p. □

Exercises from the Textbook

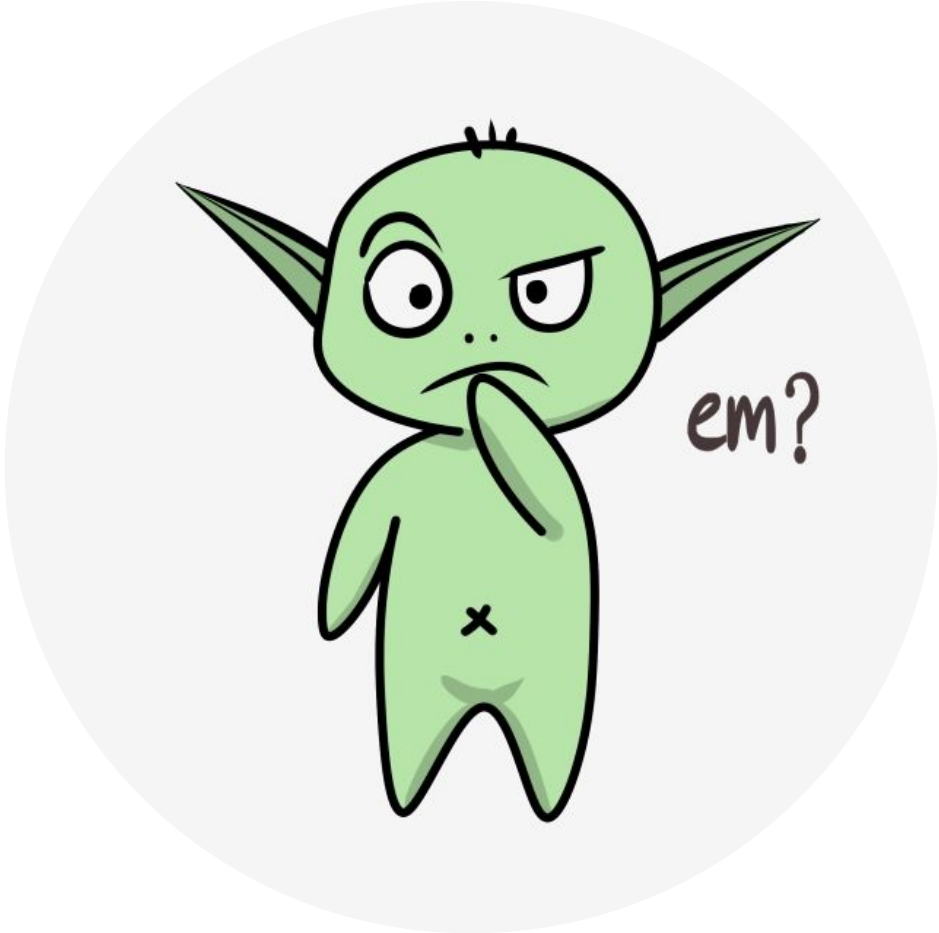
Exercise 3.27 Give an equation for π_* in terms of q_*. □

Exercise 3.28 Give an equation for π_* in terms of v_* and the four-argument p. □

Exercise 3.29 Rewrite the four Bellman equations for the four value functions (v_π, v_*, q_π, and q_*) in terms of the three argument function p (3.4) and the two-argument function r (3.5). □

$$p(s'|s,a) \doteq \Pr\{S_t=s' \mid S_{t-1}=s, A_{t-1}=a\} = \sum_{r\in\mathcal{R}} p(s',r|s,a). \tag{3.4}$$

$$r(s,a) \doteq \mathbb{E}[R_t \mid S_{t-1}=s, A_{t-1}=a] = \sum_{r\in\mathcal{R}} r \sum_{s'\in\mathcal{S}} p(s',r|s,a), \tag{3.5}$$



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