

“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

Classes 6 & 7 / 35

Plan

- Overview of Markov decision processes
 - This is about the problem, not the solution!

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!



Markov Decision Processes – Why?

- “MDPs are a classical formalization of sequential decision making, where actions influence not just immediate rewards, but also subsequent situations, or states, and through those future rewards.”
- “Thus MDPs involve delayed reward and the need to trade off immediate and delayed reward.”
- “Whereas in bandit problems we estimated the value $q_*(a)$ of each action a , in MDPs we estimate the value $q_*(s,a)$ of each action a in each state s , or we estimate the value $v_*(s)$ of each state given optimal action selections.”
- MDPs are a mathematically idealized form of the reinforcement learning problem for which precise theoretical statements can be made.

Markov Decision Processes – Why?

- “MDPs are a classical formalization of sequential decision making, where actions influence not just immediate rewards, but also subsequent situations, or states, and through those future rewards.”

“In this chapter we introduce the formal problem of finite Markov decision processes, or finite MDPs, which we try to solve in the rest of the book.”

MDPs we estimate the value $q_*(s,a)$ of each action a in each state s , or we estimate the value $v_*(s)$ of each state given optimal action selections.”

- MDPs are a mathematically idealized form of the reinforcement learning problem for which precise theoretical statements can be made.

The Agent-Environment Interface

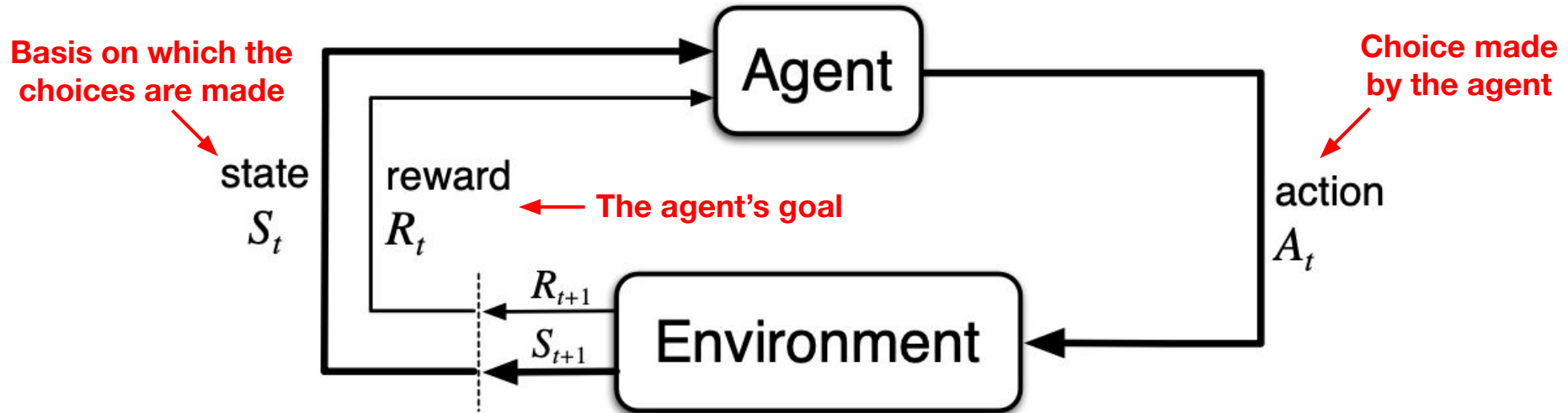


Figure 3.1: The agent–enviro

$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \dots$

Example 1: Navigating a maze

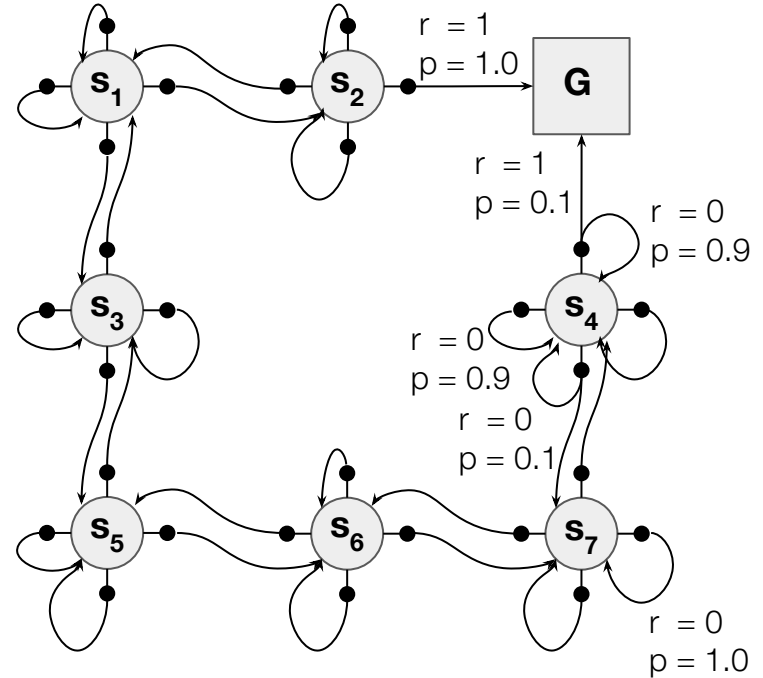
s_1	s_2	G
s_3		s_4
s_5	s_6	s_7

States: cell #

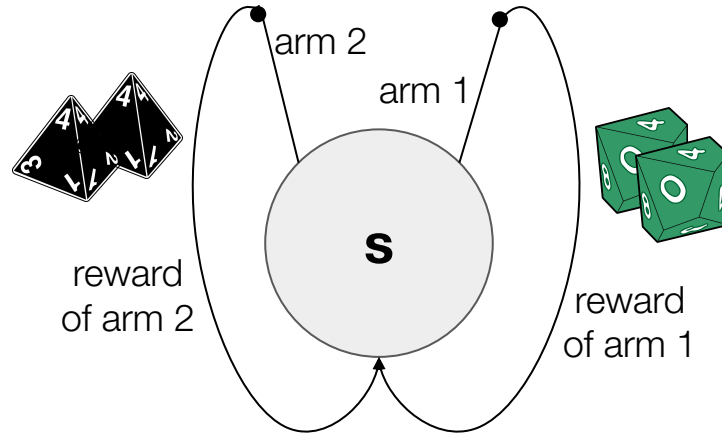
Actions: [up, down, left, right]

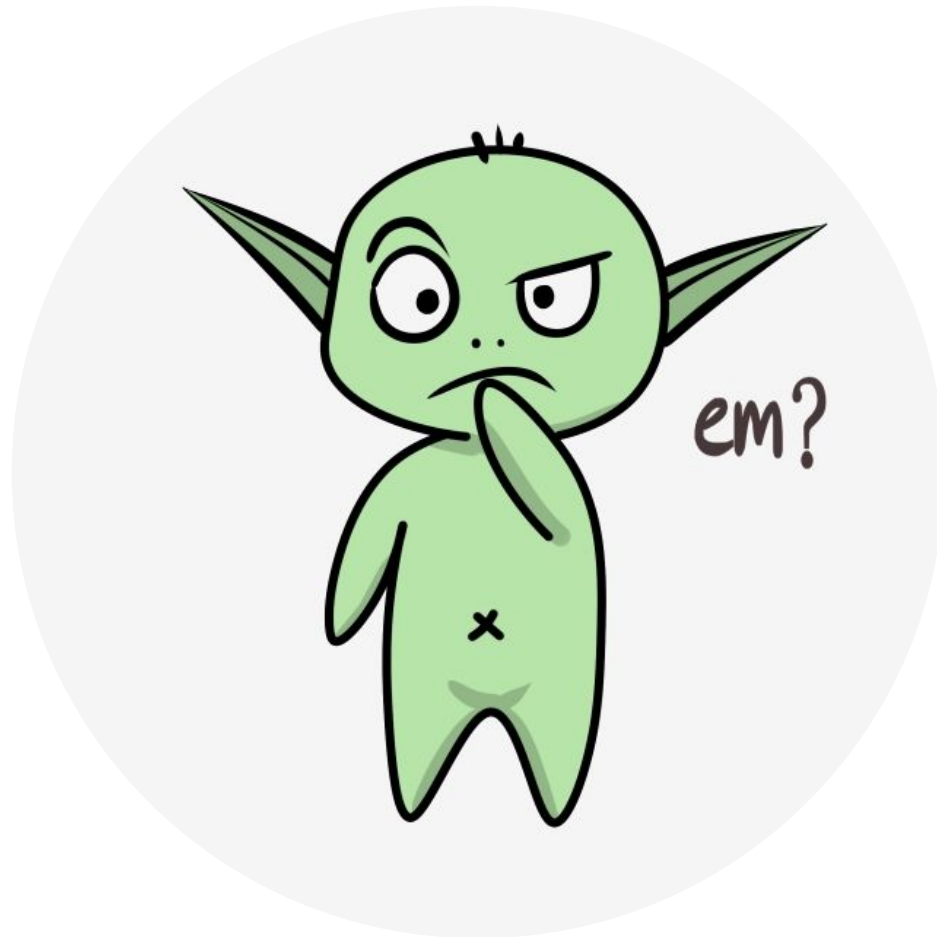
Reward: +1 upon arrival to G
0 otherwise

Dynamics: deterministic outside mud puddle
at the mud puddle you can get stuck
with probability 0.9.



Example 2: Bandits



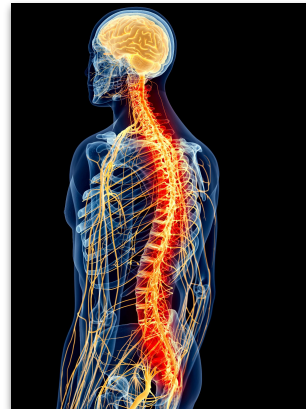


Where's the boundary between agent and environment?

It depends!

And it is often much closer than you think!

“The agent-environment boundary represents the limit of the agent’s *absolute control*, not of its knowledge.”



Formalizing the Agent-Environment Interface


$$p(s', r | s, a) \doteq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}$$

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r | s, a) = 1, \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}(s)$$

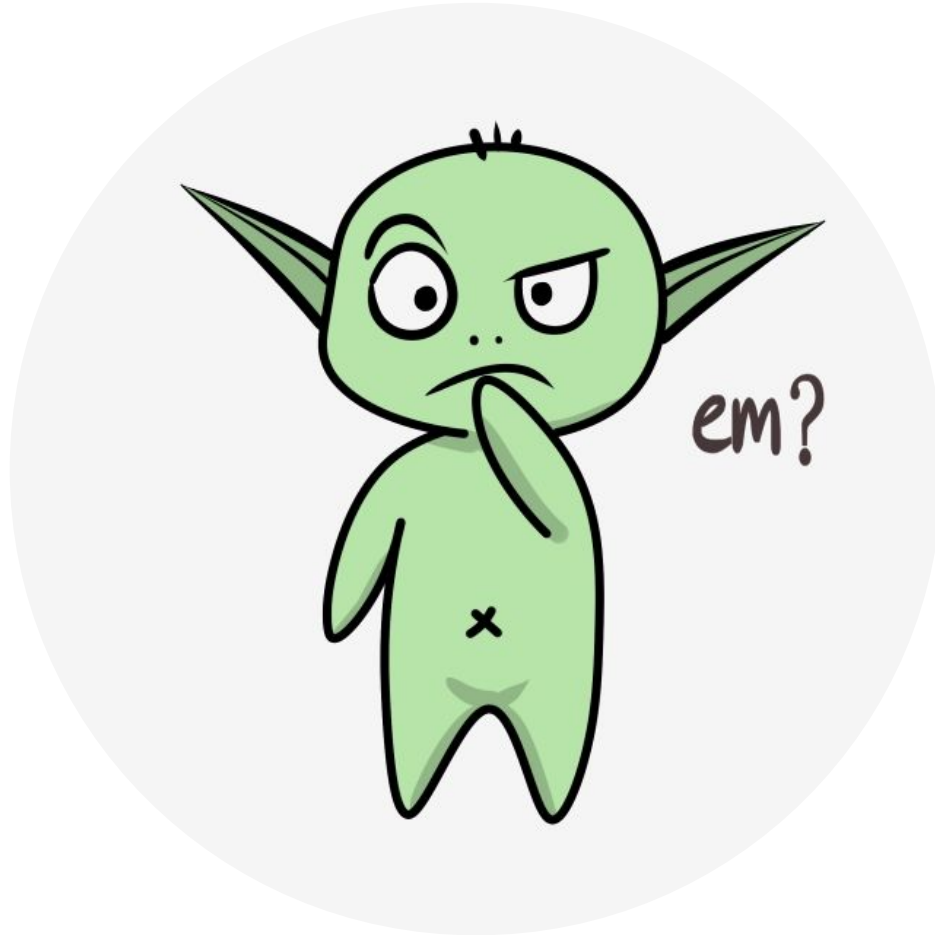
$$p(s' | s, a) \doteq \Pr\{S_t = s' \mid S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in \mathcal{R}} p(s', r | s, a)$$

Formalizing the Agent-Environment Interface

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

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


Can you show this?



The Markov Property

“The future is independent of the past given the present”

$$\mathbf{Pr}(S_{t+1}|S_t) = \mathbf{Pr}(S_{t+1} \mid S_1, \dots, S_t]$$

This should probably be seen as a restriction on the state, not on the decision process.

The Markov Property

Definition: We say that $(S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, \dots)$ has the *Markov property* if for any $t \geq 0$, $s_0, s_1, \dots, s_{t+1} \in \mathcal{S}$, $a_0, a_1, \dots, a_t \in \mathcal{A}$, and $r_1, r_2, \dots, r_{t+1} \in \mathcal{R}$, it holds that

$$\Pr(R_{t+1} = r_{t+1}, S_{t+1} = s_{t+1} \mid S_0 = s_0, A_0 = a_0, R_1 = r_1, \dots, R_t = r_t, S_t = s_t, A_t = a_t)$$

only depends on the values of s_t , a_t , s_{t+1} and r_{t+1} . In particular, none of the other past values matter when calculating probabilities of the form above. That is:

$$\begin{aligned} &\Pr(R_{t+1} = r_{t+1}, S_{t+1} = s_{t+1} \mid S_0 = s_0, A_0 = a_0, R_1 = r_1, \dots, R_t = r_t, S_t = s_t, A_t = a_t) \\ &= \Pr(R_{t+1} = r_{t+1}, S_{t+1} = s_{t+1} \mid S_t = s_t, A_t = a_t). \end{aligned}$$

The Markov Property

Definition: We say that $(S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, \dots)$ has the *Markov property* if for any $t \geq 0$, $s_0, s_1, \dots, s_{t+1} \in \mathcal{S}$, $a_0, a_1, \dots, a_t \in \mathcal{A}$, and $r_1, r_2, \dots, r_{t+1} \in \mathcal{R}$, it holds that

The Markov property does not mean that the state representation tells all that would be useful to know, only that it has not forgotten anything that would be useful to know.

past

$$\begin{aligned} \Pr(R_{t+1} = r_{t+1}, S_{t+1} = s_{t+1} \mid S_0 = s_0, A_0 = a_0, R_1 = r_1, \dots, R_t = r_t, S_t = s_t, A_t = a_t) \\ = \Pr(R_{t+1} = r_{t+1}, S_{t+1} = s_{t+1} \mid S_t = s_t, A_t = a_t). \end{aligned}$$



Reward Hypothesis

“That all of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (called reward).”

The ultimate goal: Maximize Returns

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

End of an episode

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

Continuing task

$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \\ &= R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned}$$

Unifying Notation

$$G_t \doteq \sum_{k=0}^T R_{t+k+1}$$

$$G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

- We can't use the same notation for episodic and continuing tasks because:
 - We are not specifying the episodes in the indices of an episodic task, we should actually have $R_{t,i}$.
 - In continuing tasks we have a sum over infinite numbers and in episodic tasks we sum over finite numbers.

Unifying Notation

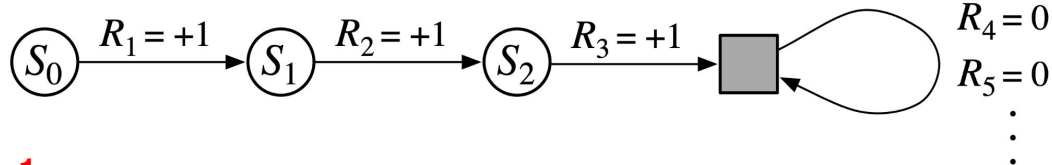
$$G_t \doteq \sum_{k=0}^T R_{t+k+1}$$

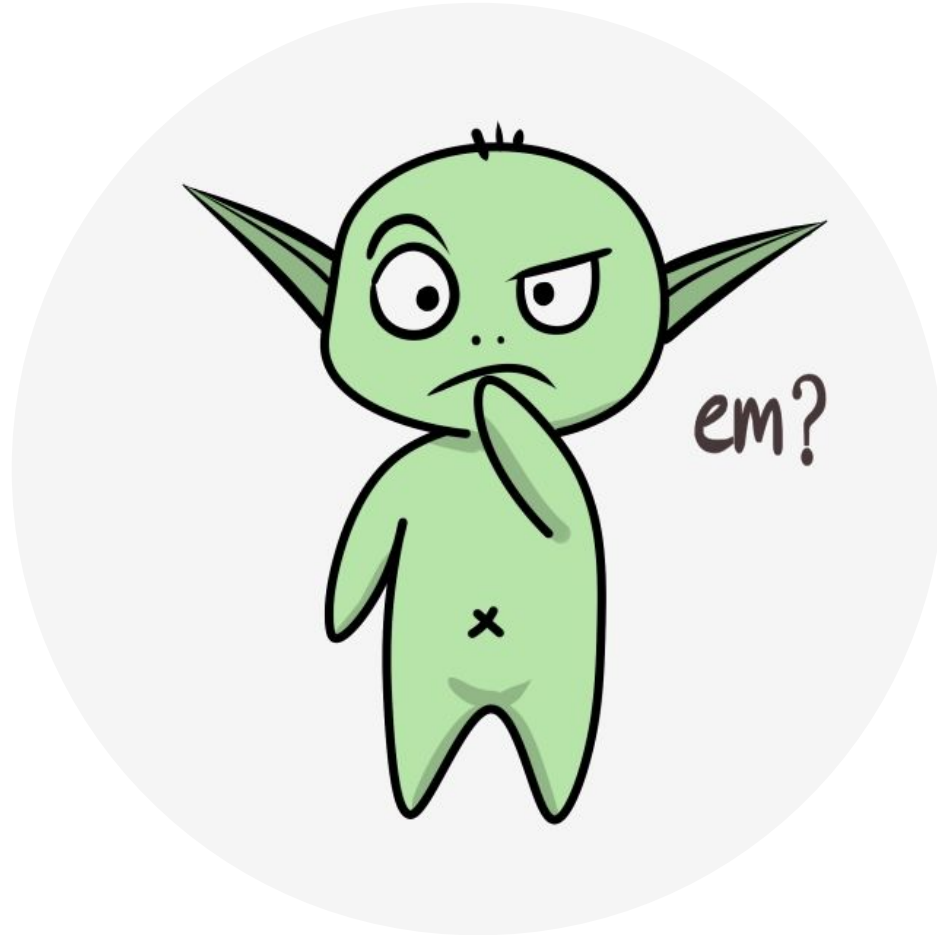
$$G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

- We can't use the same notation for episodic and continuing tasks because:
 - We are not specifying the episodes in the indices of an episodic task, we should actually have $R_{t,i}$.
 - In continuing tasks we have a sum over infinite numbers and in episodic tasks we sum over finite numbers.
- Solution:
 - It is mostly fine to drop the episode number.
 - We create an absorbing state!

$$G_t \doteq \sum_{k=t+1}^T \gamma^{k-t-1} R_k$$

**$T = \infty$ or $\gamma = 1$
(but not both)**





Next class

- What **I** plan to do: Wrap up this first half of MDPs and, time permitting, solve some exercises with you.
- What I recommend **YOU** to do for next class:
 - Complete the assigned reading.
 - Take a look at the second worksheet.
 - Make sure your grade for the first Coursera activity is correct on Canvas.
 - **Submit practice quiz by Wednesday. It is due at midnight.**