
“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 5/ 35

Plan

- Non-comprehensive 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!

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

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

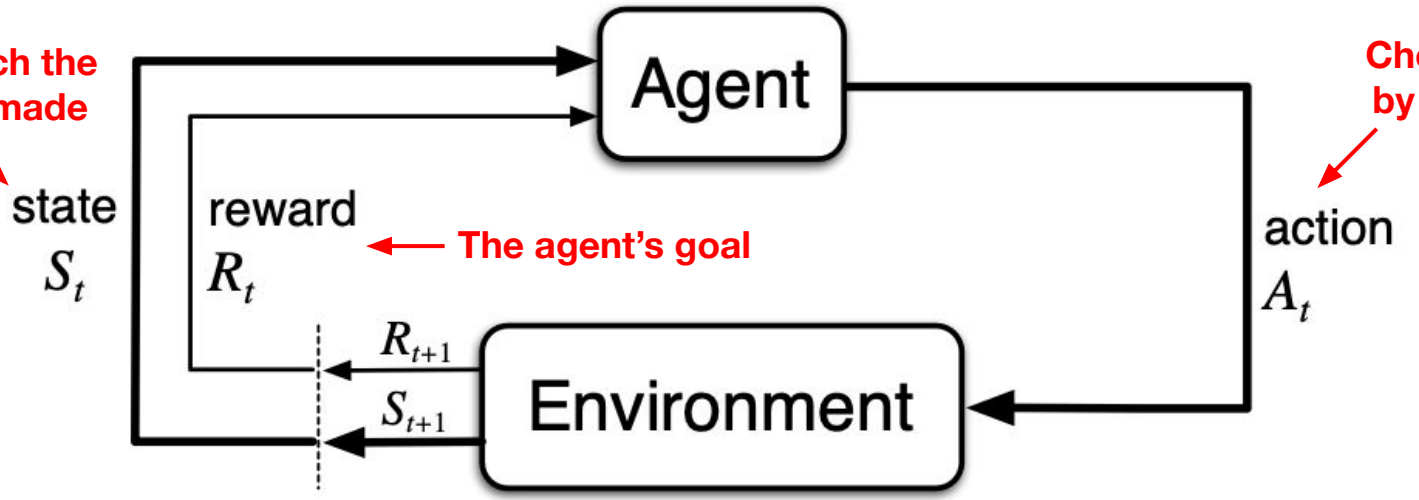
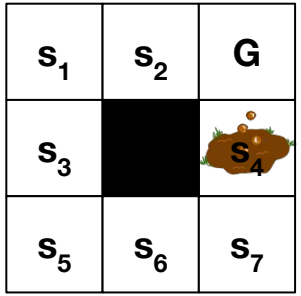


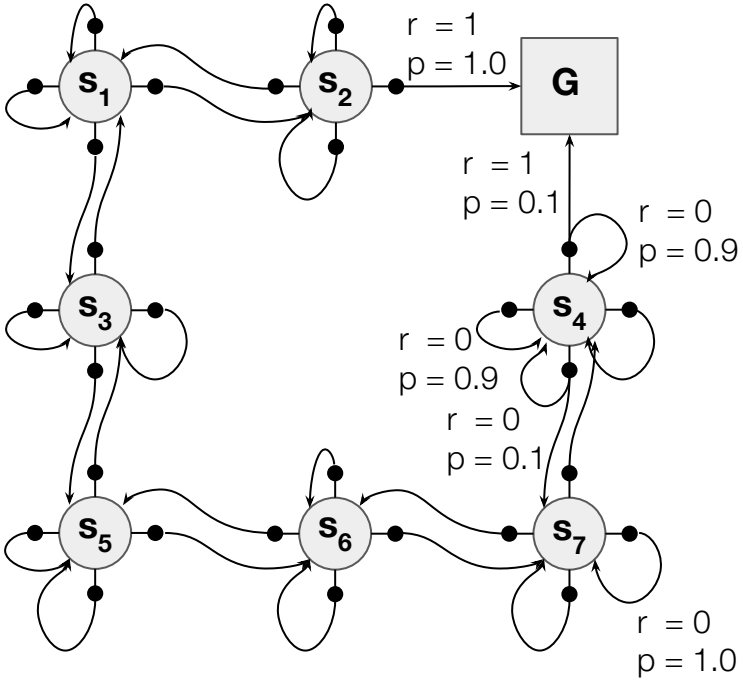
Figure 3.1: The agent–enviro

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

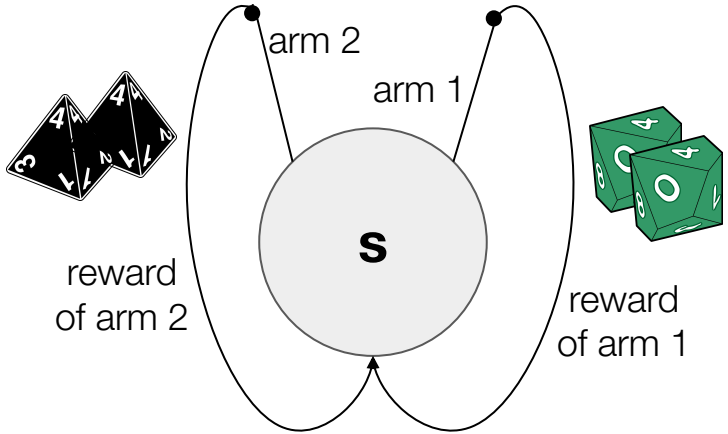
Example 1: Navigating a maze

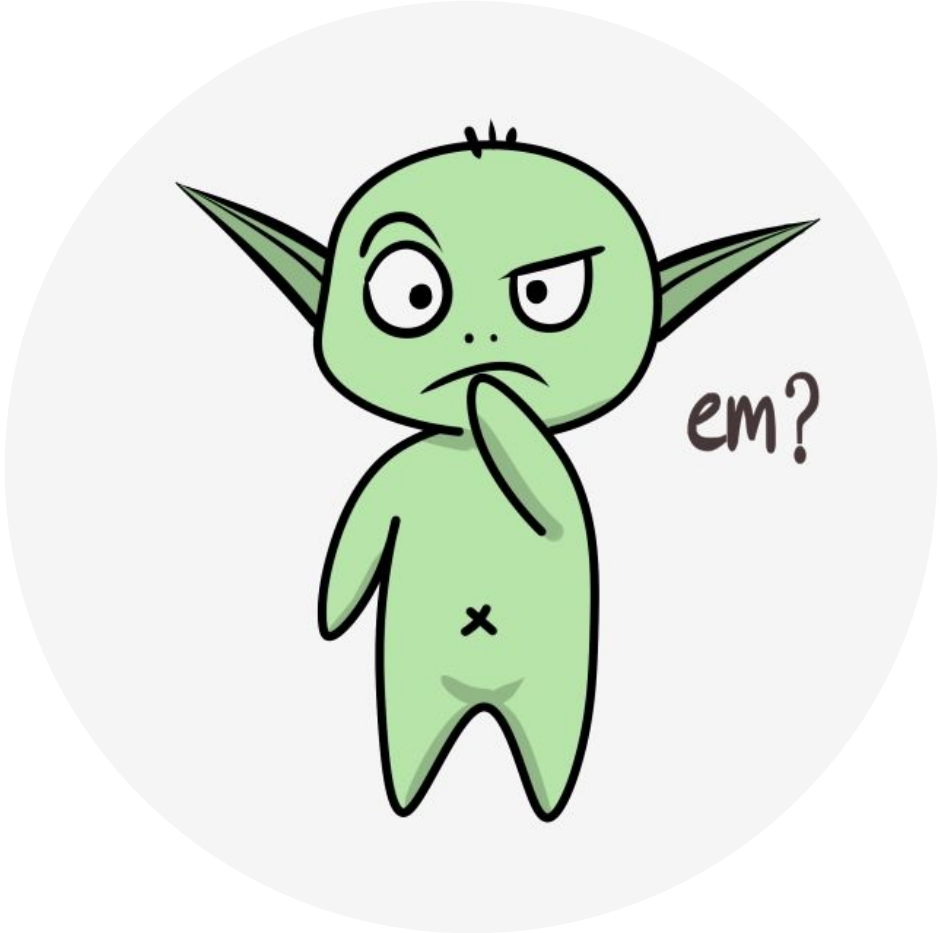


States: `cell #`
Actions: `[up, down, left, right]`
Reward: `+1` upon arrival to G
`0` otherwise
Dynamics: `deterministic` outside mud puddle
at the mud puddle you can get `stuck` `with probability 0.9.`



Example 2: Bandits





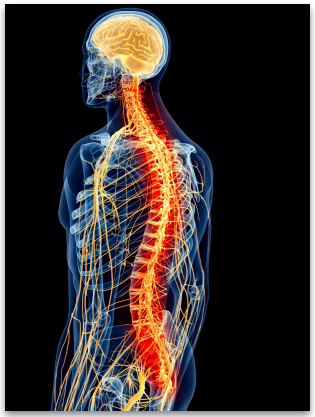
https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

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.”



https://media.wired.com/photos/645c372a88479249e0cc28b6/master/w_2560%2Cc_limit/Driverless-Cars-D
-Diet-Business-1247366812.jpg
https://peakperformancechiropractic.com/wp-content/uploads/2020/06/lighted-nervous-system.

Formalizing the Agent-Environment Interface

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

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

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

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?

Next class

- What **I** plan to do: Answer questions and solve exercises on MDPs.

- What I recommend **YOU** to do for next class:
  - **Submit practice quiz today. It is due at midnight.**