“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

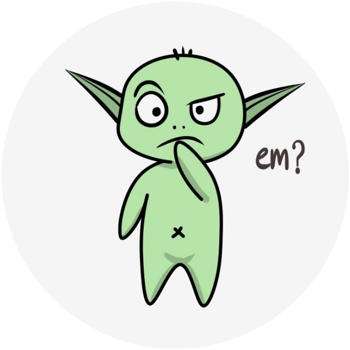
Marlos C. Machado

Class 6/ 35

Plan

- Non-comprehensive overview of Markov decision processes
  - This is about the problem, not the solution!

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 $\boxed{q_*(a)}$ of each action a, in MDPs we estimate the value $\boxed{q_*(s,a)}$ of each action a in each state s, or we estimate the value $\boxed{v_*(s)}$ 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

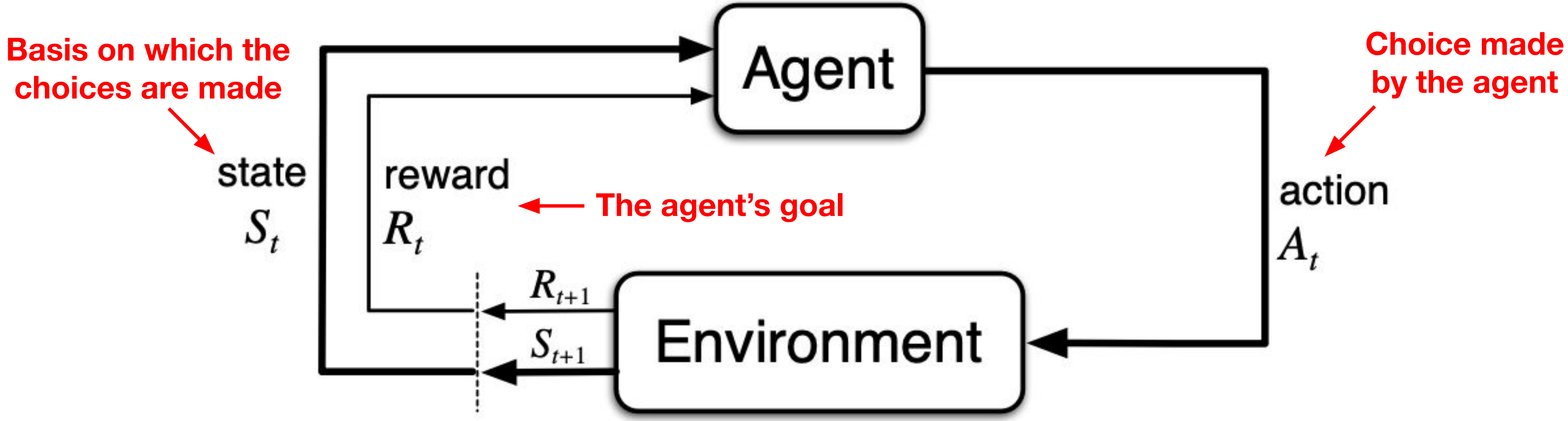
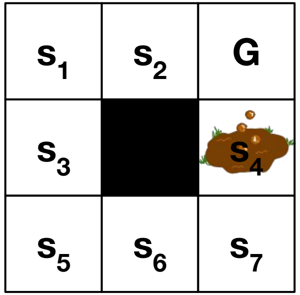
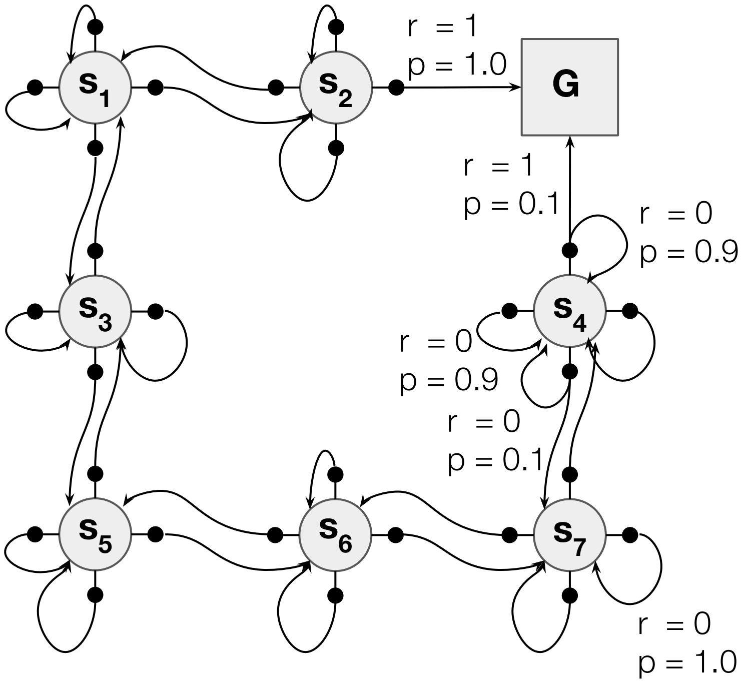


Figure 3.1: The agent–envir

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

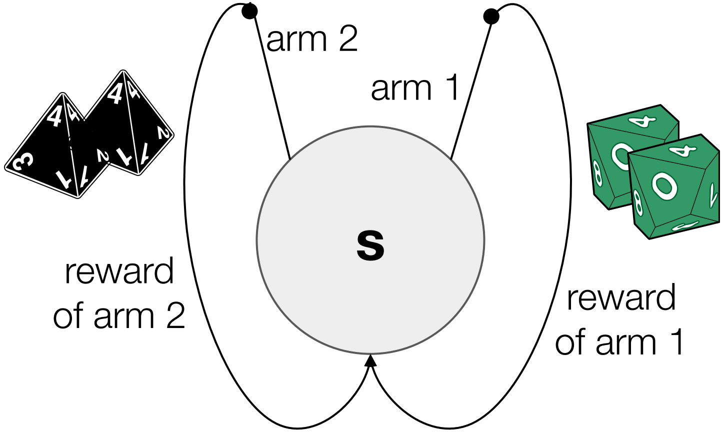
Example 1: Navigating a maze

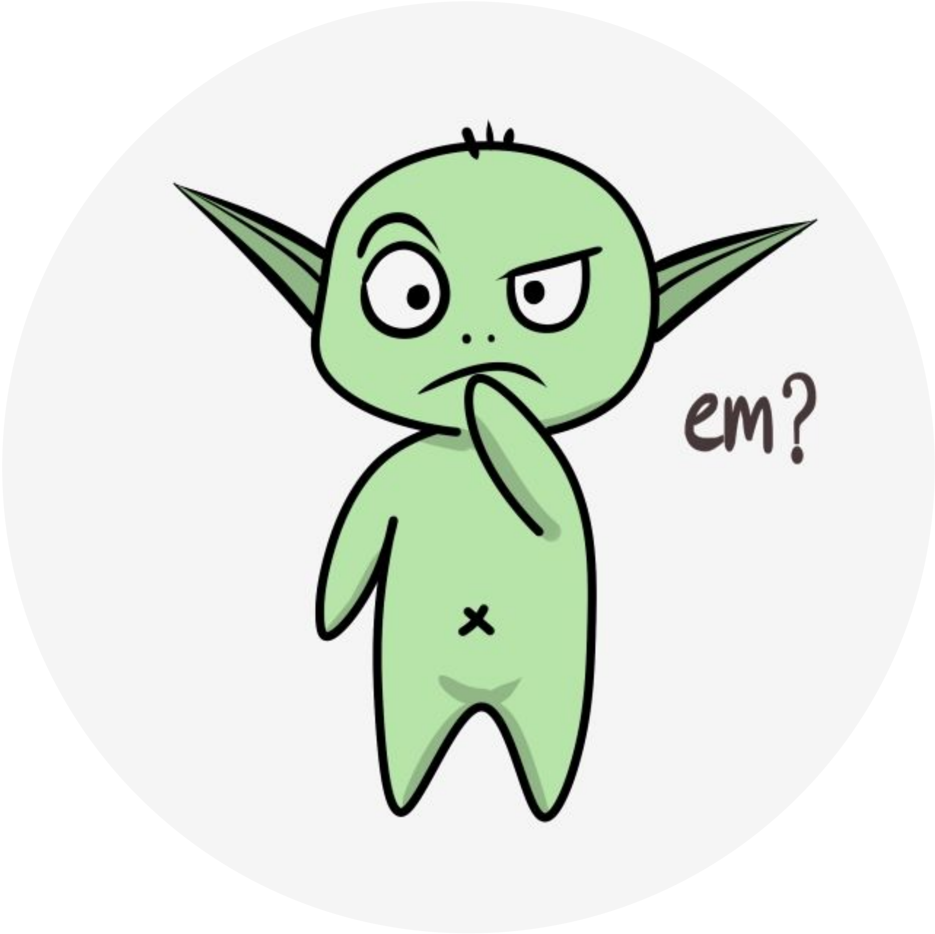




States: `cell #`
Actions: `[up, down, left, right]`
Reward: `+1` upon arrival to G
`0` otherwise
Dynamics: `deterministic` outside mud puddle
at the mud puddle you can get `stuck with probability 0.9.`

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

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

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

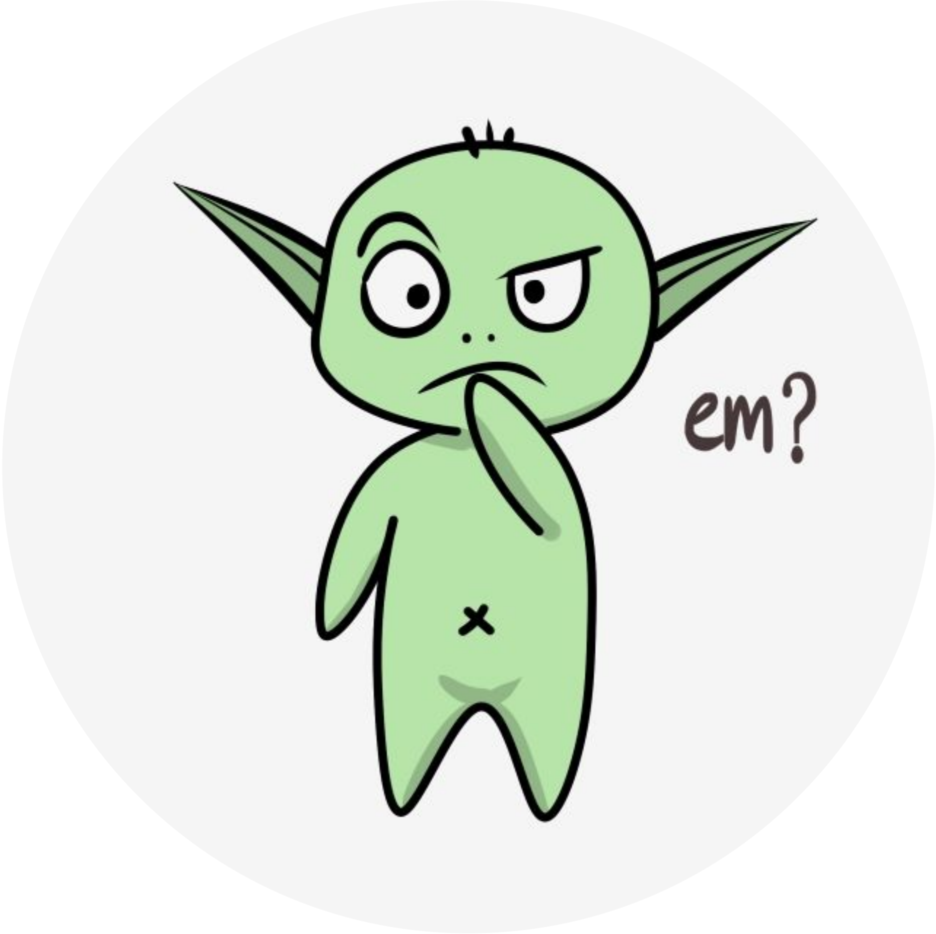
Can you show this?

The Markov Property

"The future is independent of the past given the present"

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

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



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

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

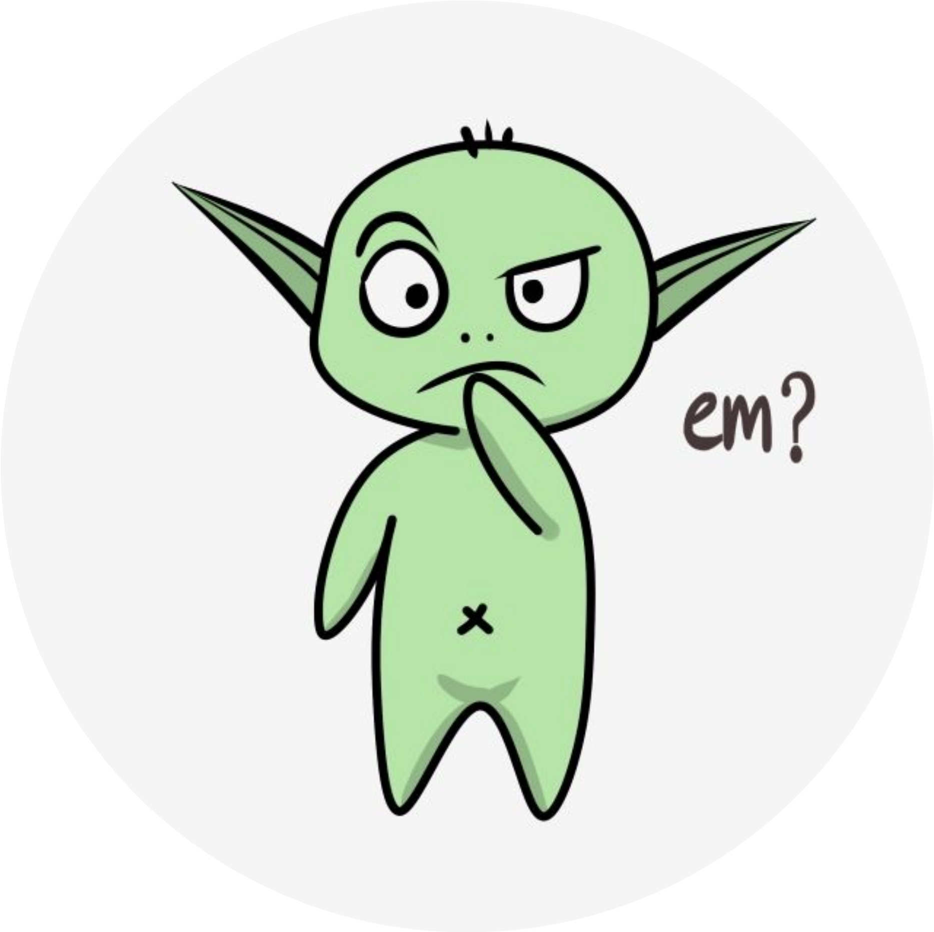
End of an episode

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

Continuing task

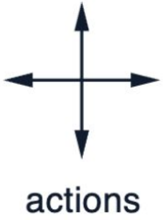
$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

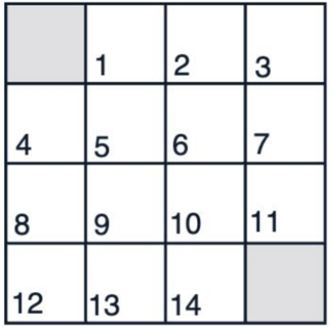
$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \\ &= R_{t+1} + \gamma \left(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots \right) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned}$$



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

Practice Exercise





$R_t = -1$
on all transitions

$p(6, -1 \,|\, 5, \texttt{right}) =$

$p(10, r \,|\, 5, \texttt{right}) =$

$p(7, -1 \,|\, 7, \texttt{right}) =$

Practice Exercise – Modeling

Assume you have a bandit problem with 4 actions, where the agent can see rewards from the set $\mathcal{R} = \{-3.0, -0.1, 0, 4.2\}$. Assume you have the probabilities for rewards for each action: $p(r|a)$ for $a \in \{1, 2, 3, 4\}$ and $r \in \{-3.0, -0.1, 0, 4.2\}$. How can you write this problem as an MDP? Remember that an MDP consists of $(\mathcal{S}, \mathcal{A}, \mathcal{R}, P, \gamma)$.

More abstractly, recall that a Bandit problem consists of a given action space $\mathcal{A} = \{1, ..., k\}$ (the k arms) and the distribution over rewards $p(r|a)$ for each action $a \in \mathcal{A}$. Specify an MDP that corresponds to this Bandit problem.

Marlos C. Machado

Practice Exercise – Modeling

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.**