
“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 5 & 6/ 35

Plan

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

Notes

- The sheet with partial answers for Worksheet 1 is now available on Canvas.
- Your grades for quiz and assignment 1 are now available on Canvas.
  - Please, please, email cmput365@ualberta.ca, not me.

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

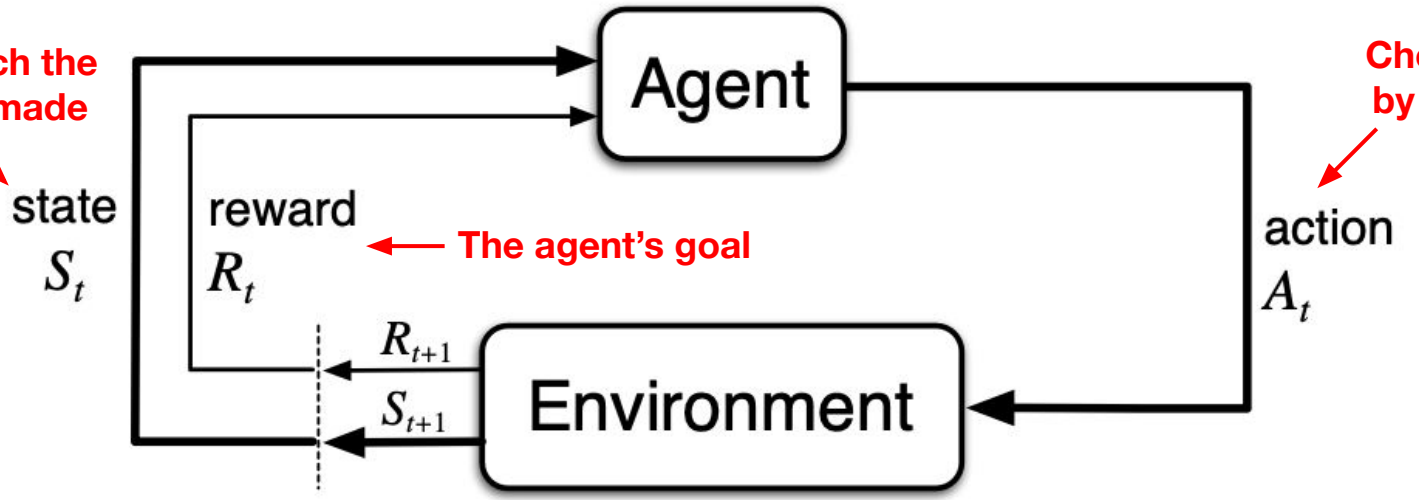
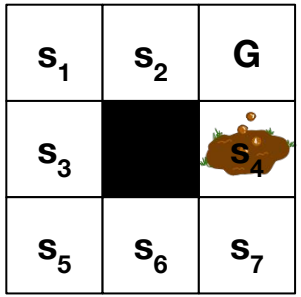


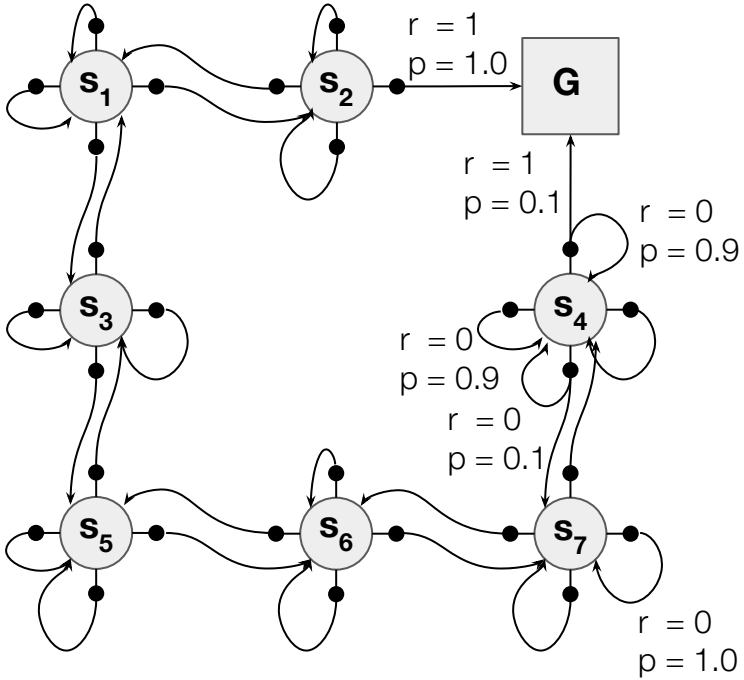
Figure 3.1: The agent–enviro

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

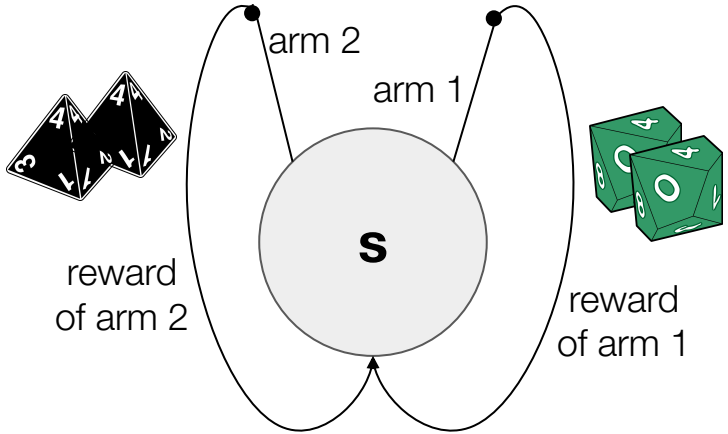
Example 1: Navigating a maze

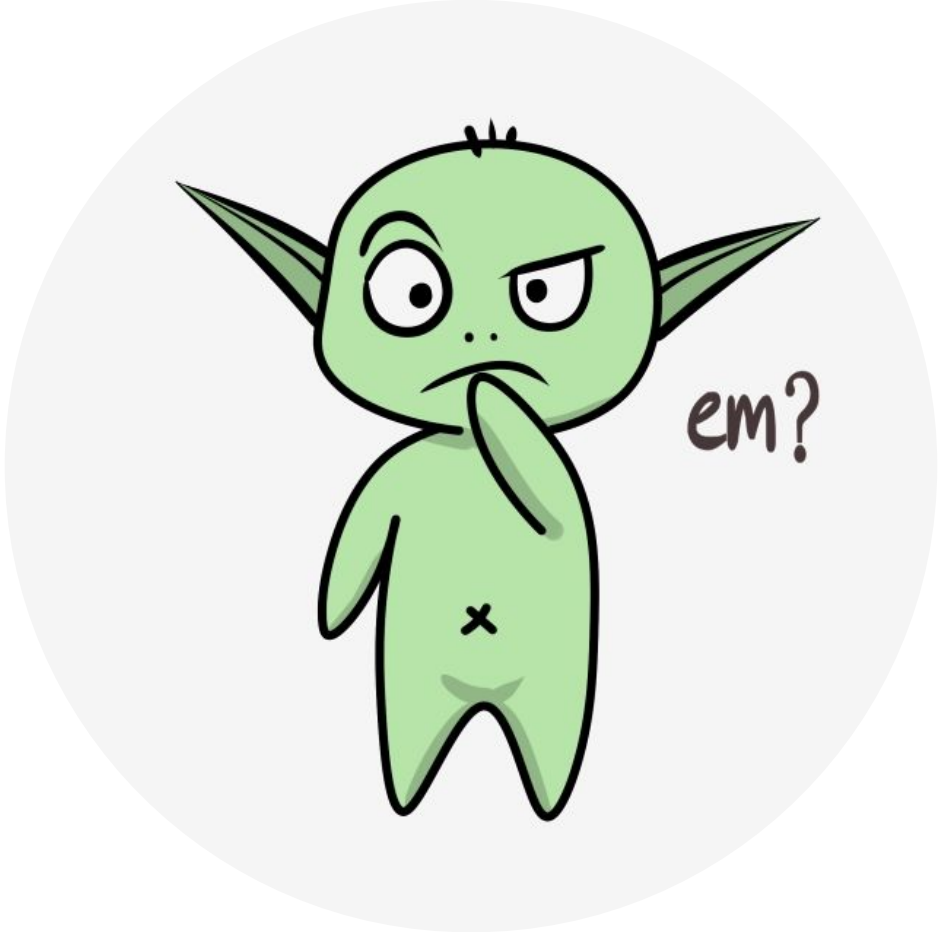


States: `cell #`
Actions: `[up, down, left, right]`
Reward: `+1` upon arrival to G
`0` otherwise
Dynamics: `deterministic` outside mud puddle
at the mud puddle you can get `stuck` `with probability 0.9.`



Example 2: Bandits





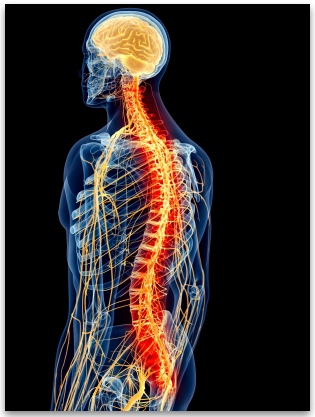
https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

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

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

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

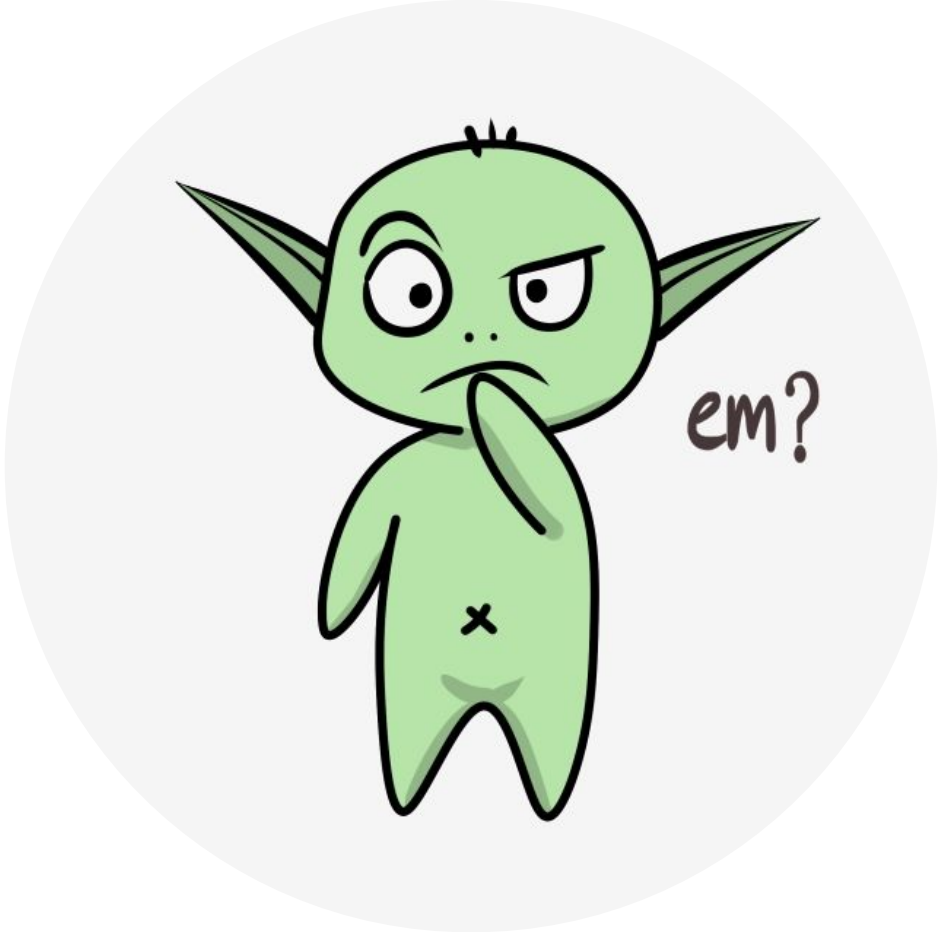
$$p(s' \mid s, a) \doteq \Pr\{S_t = s' \mid S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in \mathcal{R}} p(s', r \mid s, a)$$

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?



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

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.

The Markov Property

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

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

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:

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

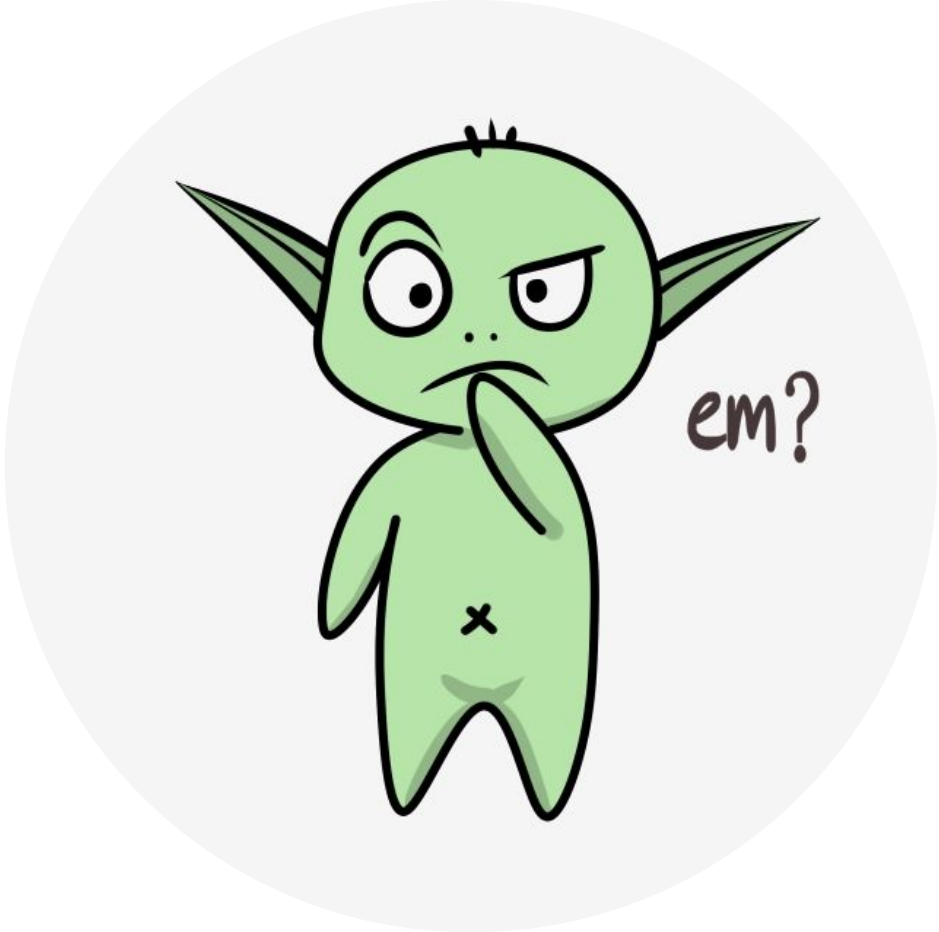
The Markov Property

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

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



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

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

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

Unifying Notation

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

- 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

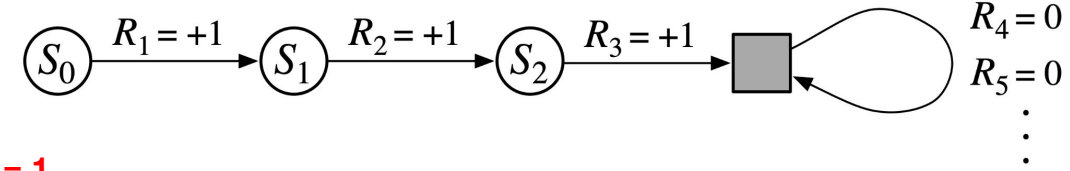
$$G_t \doteq \sum_{k=0}^{T} R_{t+k+1} \qquad G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

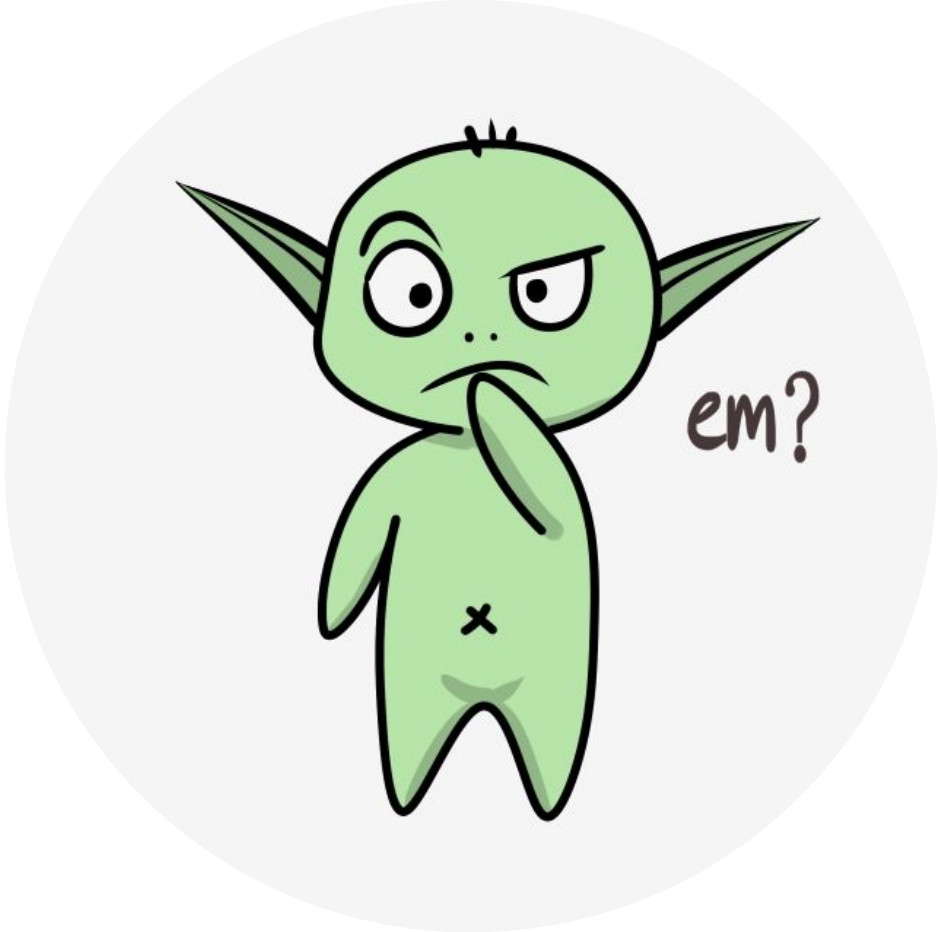
- We can't use the same notation for episodic and continuing tasks because:
  - We are not specifying the episodes in the indices of an episodic task, we should actually have $R_{t,i}$.
  - In continuing tasks we have a sum over infinite numbers and in episodic tasks we sum over finite numbers.

- 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 = ∞ or γ = 1 (but not both)





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

Next class

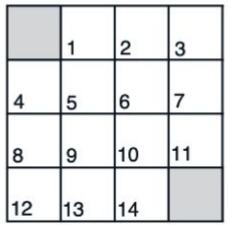
- What **I** plan to do: Wrap up this first half of MDPs and, time permitting, solve some exercises with you.

- 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.**

Question 1. 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.

Question 2. Consider the 4×4 gridworld shown below.



$R_t = -1$
on all transitions

The nonterminal states are $\mathcal{S} = \{1, 2, \ldots, 14\}$. There are four actions possible in each state, $\mathcal{A} = \{$ `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. In this context, what are the values for the expressions below?

(a) $p(6, -1|5,\texttt{right}) =$ (b) $p(7, -1|7,\texttt{right}) =$ (c) $p(10, r|5,\texttt{right}) =$

Question 3. Prove that the discounted sum of rewards is always finite, if the rewards are bounded: $|R_{t+1}| \leq R_{\max}$ for all t for some finite $R_{\max} > 0$.

$$\left| \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \right| \qquad \text{for } \gamma \in [0, 1).$$

Hint: Recall that $|a + b| < |a| + |b|$.

Exercise 3.6 Suppose you treated pole-balancing as an episodic task but also used discounting, with all rewards zero except for -1 upon failure. What then would the return be at each time? How does this return differ from that in the discounted, continuing formulation of this task? □

Exercise 3.7 Imagine that you are designing a robot to run a maze. You decide to give it a reward of +1 for escaping from the maze and a reward of zero at all other times. The task seems to break down naturally into episodes—the successive runs through the maze—so you decide to treat it as an episodic task, where the goal is to maximize expected total reward (3.7). After running the learning agent for a while, you find that it is showing no improvement in escaping from the maze. What is going wrong? Have you effectively communicated to the agent what you want it to achieve? □

Exercise 3.8 Suppose $\gamma = 0.5$ and the following sequence of rewards is received $R_1 = -1$, $R_2 = 2$, $R_3 = 6$, $R_4 = 3$, and $R_5 = 2$, with $T = 5$. What are G_0, G_1, ..., G_5? Hint: Work backwards. □

Exercise 3.9 Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2$ followed by an infinite sequence of 7s. What are G_1 and G_0? □

Exercise 3.9 Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2$ followed by an infinite sequence of 7s. What are G_1 and G_0? □

Exercise 3.10 Prove the second equality in (3.10). □

$$G_t = \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}.$$

Next class

- What **I** plan to do: Start the 2nd half of MDPs: Value Functions & Bellman Eqs.

- What I recommend **YOU** to do for next class:
  - Complete the assigned reading: Chapter 3, §3.5-§3.8 (pp. 58-69).
  - Make sure your grade for the first Coursera activity is correct on Canvas.
  - **Submit practice quiz today (there's only one activity). It is due at midnight.**