
“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

- Finish Non-comprehensive overview of MDPs
  - Returns and Episodes
- Go over common errors in Coursera
- More exercises

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

Last class: MDPs

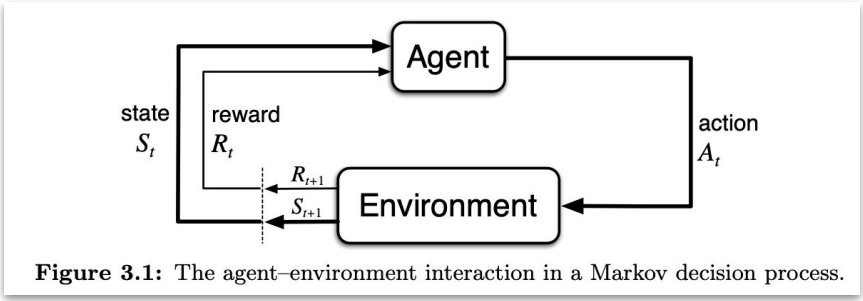
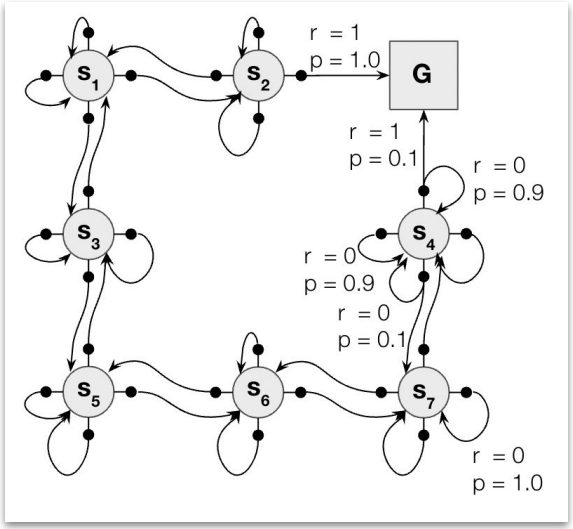


Figure 3.1: The agent–environment interaction in a Markov decision process.



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

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

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

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

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:

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

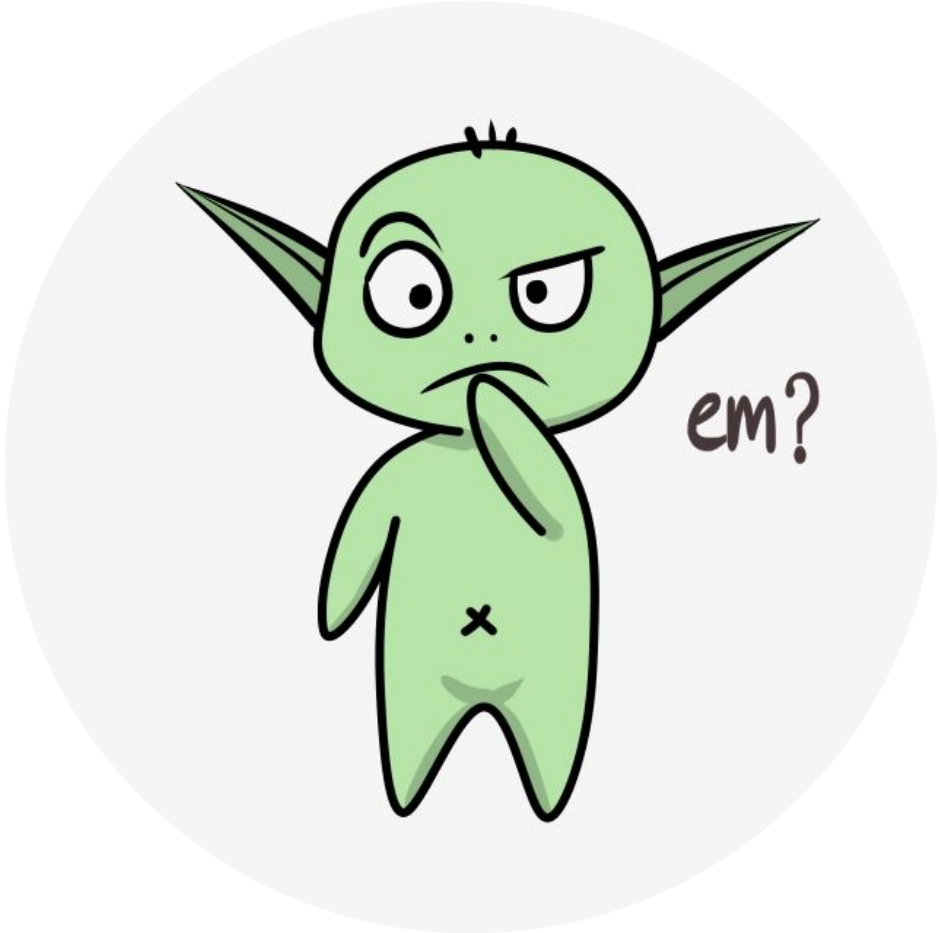
The Markov Property

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.

er

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



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

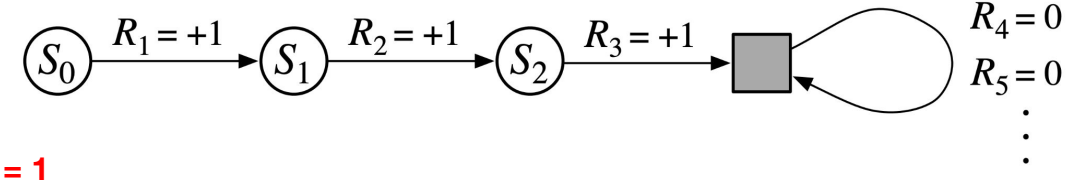
$$G_t \doteq \sum_{k=0}^{T} R_{t+k+1} \qquad G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

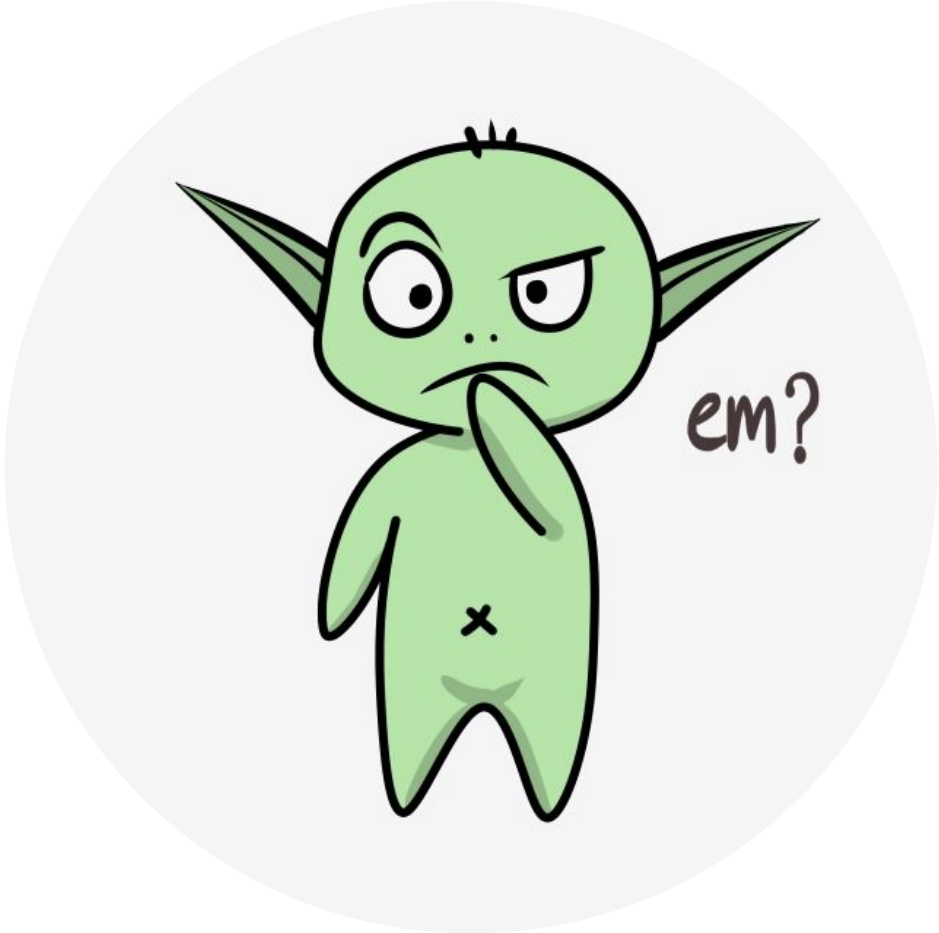
- We can't use the same notation for episodic and continuing tasks because:
  - We are not specifying the episodes in the indices of an episodic task, we should actually have $R_{t,i}$.
  - In continuing tasks we have a sum over infinite numbers and in episodic tasks we sum over finite numbers.

- 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

Practice Exercise

actions

	1	2	3
4	5	6	7
8	9	10	11
12	13	14	

$R_t = -1$
on all transitions

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

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

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

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.

Practice Exercise – Modeling

Exercise 6 from Quiz on Coursera

6. Suppose $\gamma = 0.8$ and the reward sequence is $R_1 = 5$ followed by an infinite sequence of 10s. What is G_0?

- ○ 55
- ○ 45
- ○ 15

Exercise 6 from Quiz on Coursera

6. Suppose $\gamma = 0.8$ and we observe the following sequence of rewards: $R_1 = -3, R_2 = 5, R_3 = 2, R_4 = 7$, and $R_5 = 1$, with $T = 5$. What is G_0? Hint: Work Backwards and recall that $G_t = R_{t+1} + \gamma G_{t+1}$.

- ○ 6.2736
- ○ 8.24
- ○ 11.592
- ○ -3
- ○ 12

Exercise 3.8 of the Textbook

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. □

Solution Exercise 3.8 of the Textbook

Exercise 3.7 of the Textbook

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? □

Solution Exercise 3.7 of the Textbook

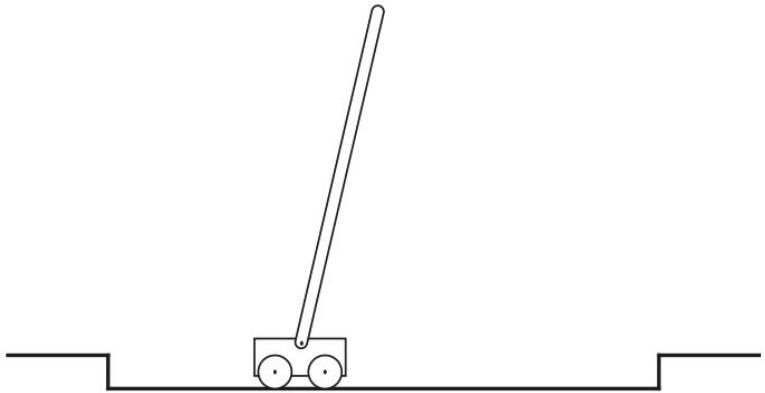
Exercise 10 from Quiz on Coursera

10. Imagine, an agent is in a maze-like gridworld. You would like the agent to find the goal, as quickly as possible. You give the agent a reward of +1 when it reaches the goal and the discount rate is 1.0, because this is an episodic task. When you run the agent its finds the goal, but does not seem to care how long it takes to complete each episode. How could you fix this? (**Select all that apply**)

- ☐ Give the agent a reward of 0 at every time step so it wants to leave.
- ☐ Set a discount rate less than 1 and greater than 0, like 0.9.
- ☐ Give the agent a reward of +1 at every time step.
- ☐ Give the agent -1 at each time step.

Example

Example 3.4: Pole-Balancing
The objective in this task is to apply forces to a cart moving along a track so as to keep a pole hinged to the cart from falling over: A failure is said to occur if the pole falls past a given angle from vertical or if the cart runs off the track. The pole is reset to vertical after each failure. This task could be treated as episodic, where the natural episodes are the repeated attempts to balance the pole. The reward in this case could be $+1$ for every time step on which failure did not occur, so that the return at each time would be the number of steps until failure. In this case, successful balancing forever would mean a return of infinity. Alternatively, we could treat pole-balancing as a continuing task, using discounting. In this case the reward would be -1 on each failure and zero at all other times. The return at each time would then be related to $-\gamma^{K-1}$, where K is the number of time steps before failure (as well as to the times of later failures). In either case, the return is maximized by keeping the pole balanced for as long as possible. ■



Exercise 3.6 of the Textbook

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? □

Solution Exercise 3.6 of the Textbook

Exercise 3.10 of the Textbook

Exercise 3.10 Prove the second equality in (3.10).

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

Solution Exercise 3.10 of the Textbook

Practice Exercise

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 $R_{max} > 0$.

$|\sum_{i=0}^{\infty} \gamma^i R_{t+1+i}| < \infty$ for $\gamma \in [0, 1)$. Hint: Recall that $|a + b| < |a| + |b|$.

Solution Practice Exercise

Next class

- What **I** plan to do:
  - Start “new” week, still on Chapter 3: Value Functions & Bellman Equations

- What I recommend **YOU** to do for next class:
  - Read Chapter 3, §3.4–§3.8 (pp. 57–69).
  - Submit Practice Quiz for Fundamental of RL: Value functions & Bellman equations (Week 4)
  - Start Graded Quiz for Fundamental of RL: Value functions & Bellman equations (Week 4)