Mini-Course 1, Module 2
Markov Decision Processes

CMPUT 365
Fall 2021

Reminders: Sept 13, 2021

- Schedule with deadlines on github pages
  - https://docs.google.com/spreadsheets/d/1ooFqttGCklw7rsst9xwL77_SA84LszLvZwpWo06Ltas
- **We are making best 10 of 11 for Graded Assignments (one freebie)**
- Graded Assessment for Course 1, Module 2 (3 MDPs) due **this Friday Noon**
- **Peer-review for Course 1, Module 2 (3 MDPs) due this Sunday**
- Any questions about admin?

Review of Course 1, Module 2

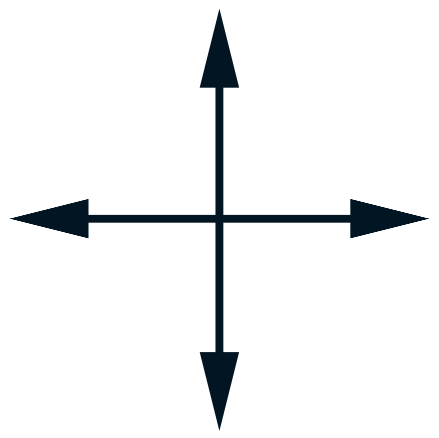
Video 1: Markov Decision Processes

- Discussed the MDP formalism: states, actions, time steps, rewards, agents, environments
- Goals:
  - Understand **Markov Decision Processes**, or **MDPs**; and
  - describe how the **dynamics of an MDP** are defined
- *What are some of the key differences here between RL problems and supervised learning?*

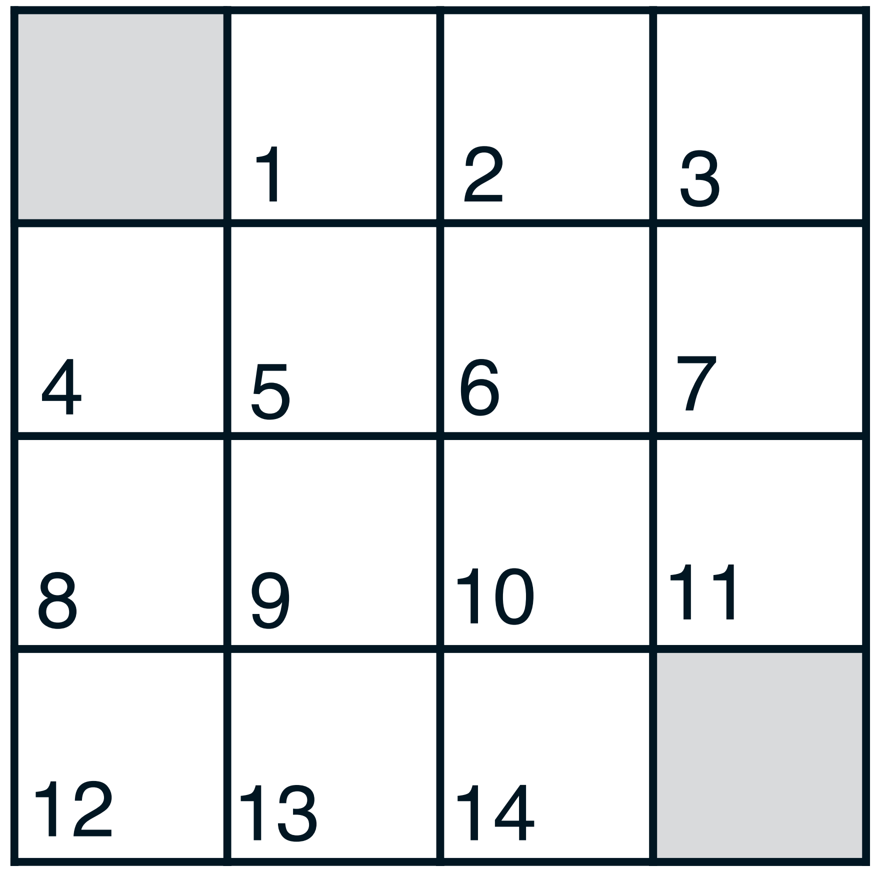
Video 2: Examples of MDPs

- Discussed several sample problems and how they can be expressed in the language of MDPs
- Goals:
  - Gain experience **formalizing** decision-making problems as MDPs
  - Appreciate the **flexibility** of the MDP formalism
- *Aren't MDPs too limited? Can you think of problems that cannot be formulated as MDPs?*

Practice Questions



actions



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

$R_t = -1$

on all transitions

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

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

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

Video 3: The Goal of Reinforcement Learning

- Discussed the goal of an RL agent, and how that relates to future reward
- Goals:
  - Describe how **rewards** relate to the **goal of an agent**, and
  - Identify **episodic tasks**
- *Why not just formulate the goal of RL to be reaching goal states?*

The 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)."

- Can you think of counter-examples of this hypothesis?

Video 4: Continuing Tasks

- Discussed why continuing tasks are special and how to define the return for continuing tasks
- Goals:
    - Differentiate between **episodic** and **continuing** tasks
    - Formulate **returns** for continuing tasks using **discounting**; and
    - Describe how **returns at successive** time steps are related to each other.

Video 5: Examples of Episodic Tasks and Continuing Tasks

- Discussed several examples of continuing tasks, and how to formulate them as MDPs.
- **Goal**: Understand when to formalize a task as episodic or continuing
- *Which do you think is more common episodic or continuing?*

Where are the solution methods?

- This chapter is only about defining the problem setting of RL: solving finite MDPs
- Every chapter after ch3 will discuss solution methods
- We must first carefully understand the problem before solving it!

Separating problem from solutions

- “Is it always true that we are able to design a reward beforehand for all problems? if not, is it possible to learn the reward function from the environment? if yes, how?”
    - Think of the task of sorting a list. We define it as: given a list of integers arrange them from smallest to largest. We don’t let our sorting algorithm decide “lets only sort the numbers if they are <100, or only sort half the list. It would not make sense to allow the algorithm/the-solution to change the problem definition.
        - **Such an algorithm is by definition failing to solve the task**
    - The MDP, including the reward function and gamma are the problem definition. The agent that we design is a solution method. The solution cannot change the problem definition
- *Chapter 3 is all about defining the problem. We must never ever ever mix problem definition and solution strategy*

Practice Quiz Review

- https://www.coursera.org/learn/fundamentals-of-reinforcement-learning/quiz/Kqiyt/mdps

Your questions from Discord

- Talk about continuing and episodic:
  - Satellite adjustment vs Halo
- “For continuing case, Gt is finite mathematically. But my question is if the gamma = 1, in practical case (programming) how can we compute the Gt as the time is infinite? how to make sure that the reward sequence is bounded?”
- The return is defined on future rewards the agent hasn’t seen yet. That seems impossible to deal with?
  - You are wondering about the solution method. This is the problem statement
  - We will get to solution methods next chapter …

Your questions from Discord

- More on dealing with future unseen rewards …
  - Remember when I asked you about probability of dice?
  - Imagine I asked about the expected value of an unfair dice (that depends on things we could never know—the true probabilities)
  - We can talk about the this question, without worrying how to approximate it…**same with an RL agent and return**
  - We have a strategy of rolling the dice over and over and counting the outcomes we observe
    - Rolling once, get some data, roll a few more times, get some more data and a better estimate. **Our agents will do the same with rewards!**

Worksheet Question 1

Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2, R_2 = -2, R_3 = 0$ followed by an infinite sequence of 7s. What are G_1 and G_0?