Mini-Course 1, Module 2
Markov Decision Processes

CMPUT 397
Fall 2020

Reminders: Sept 14, 2019

- Schedule with deadlines on github pages (https://marthawhite.github.io/rlcourse/schedule.html)

- **Some confusion the first week.** Slido for Participation Questions, naming C1M1 for Bandits

- **We are making best 10 of 11 for Graded Assignments (one freebie)**

- Graded Assessment for Course 1, Module 2 (3 MDPs) due **this Friday**

- **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

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

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**

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)."

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

Question and Answer

- Let’s discuss a few questions from Slido
  - I will post the question I am answering in Zoom chat, labeled [Slido Q]
- Also feel free to post any questions in the Zoom chat
- I will answer these using a Whiteboard (my iPad)

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?

Worksheet Question 2

(Exercise 2.2 from S&B 2nd edition) Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = 1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = 2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

Worksheet Question 4

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 $R_{\max} > 0$.

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

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