Course 1, Module 3
Markov Decision Processes

CMPUT 397
Fall 2019

Weekly Schedule

- Sunday: Discussion question due, deadline for completing practice quiz
- Monday: **Review** of module, **Q&A session** about content. Finish with class exercise question
- Wednesday: **In-class Discussion** based on your submitted discussion topics
- Thursday: Graded Assessment (usually python notebook) due
- Friday: Finish discussion if needed. More in class **exercise questions from worksheet**

Grading structure

- Assignments (graded on Coursera): 30%
- Project: 10%
- In-class Participation: 10% (discussion question and practice quiz)
- Midterm Exam: 20%
- Final Exam: 30%

Reminders: Sept 16, 2019

- Schedule with deadlines on github pages (https://marthawhite.github.io/rlcourse/schedule.html)
- Graded Assessment for Course 1, Module 3 (3 MDPs) due **this Thursday**
- **Peer-review for Course 1, Module 3 (3 MDPs) due this Sunday**
- Any questions about admin?

Review of Course 1, Module 3

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 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

Worksheet Question 4

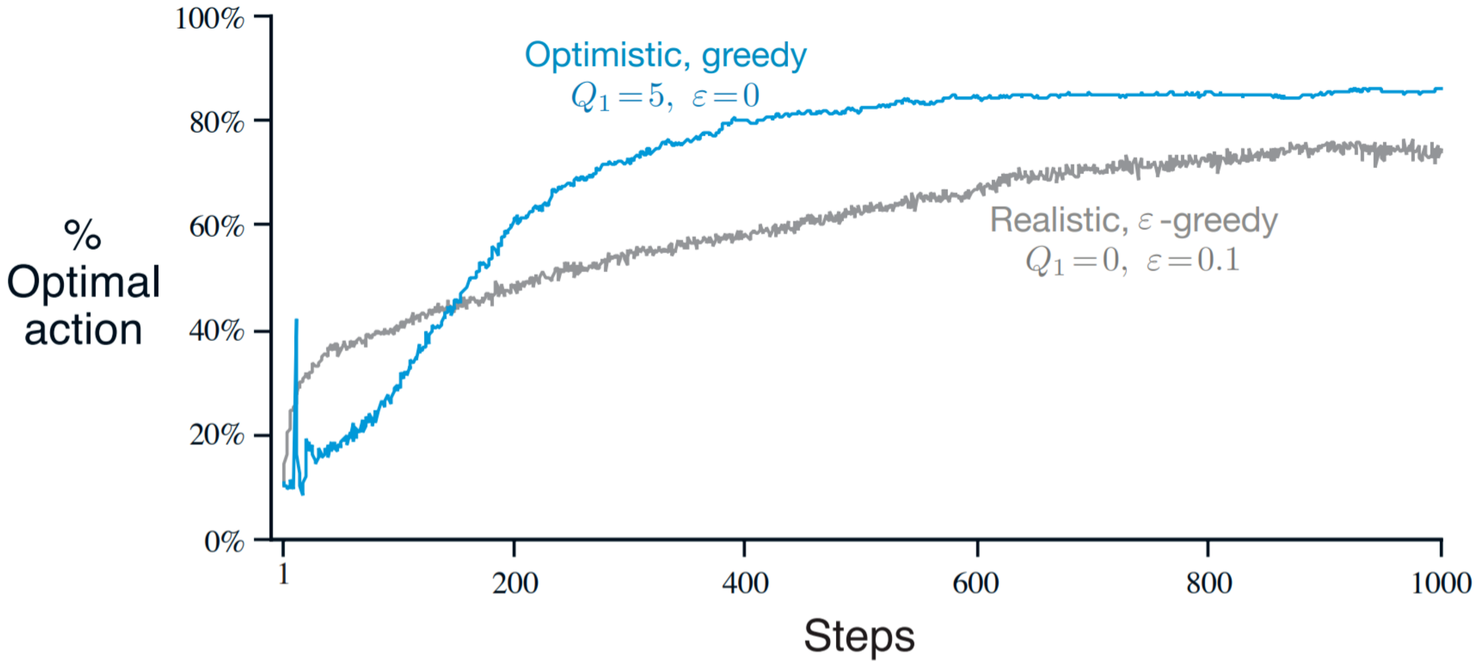
Prove that the discounted sum of rewards is always finite, if the rewards are bounded: $|R_{t+1}| \leq R_{\max}$ for all t for some finite $R_{\max} > 0$.

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

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

Worksheet Challenge Question

(Exercise 2.6 from S&B 2nd edition) The results shown in Figure 2.3 should be quite reliable because they are averages over 2000 individual, randomly chosen 10-armed bandit tasks. Why, then, are there oscillations and spikes in the early part of the curve for the optimistic method? In other words, what might make this method perform particularly better or worse, on average, on particular early steps?



alpha = 0.1