“And always, he fought the temptation to choose a clear, safe course, warning “That path leads ever down into stagnation.””

Frank Herbert, *Dune*

CMPUT 365
Introduction to Sequential-Decision Making

Marlos C. Machado

Class 3/ 35

Plan

- Motivation
- *Non-comprehensive* overview of Intro to Sequential-Decision Making in Coursera (Bandits, Chapter 2 of the textbook)

Reminder

You **should be enr** for CMPUT 365.

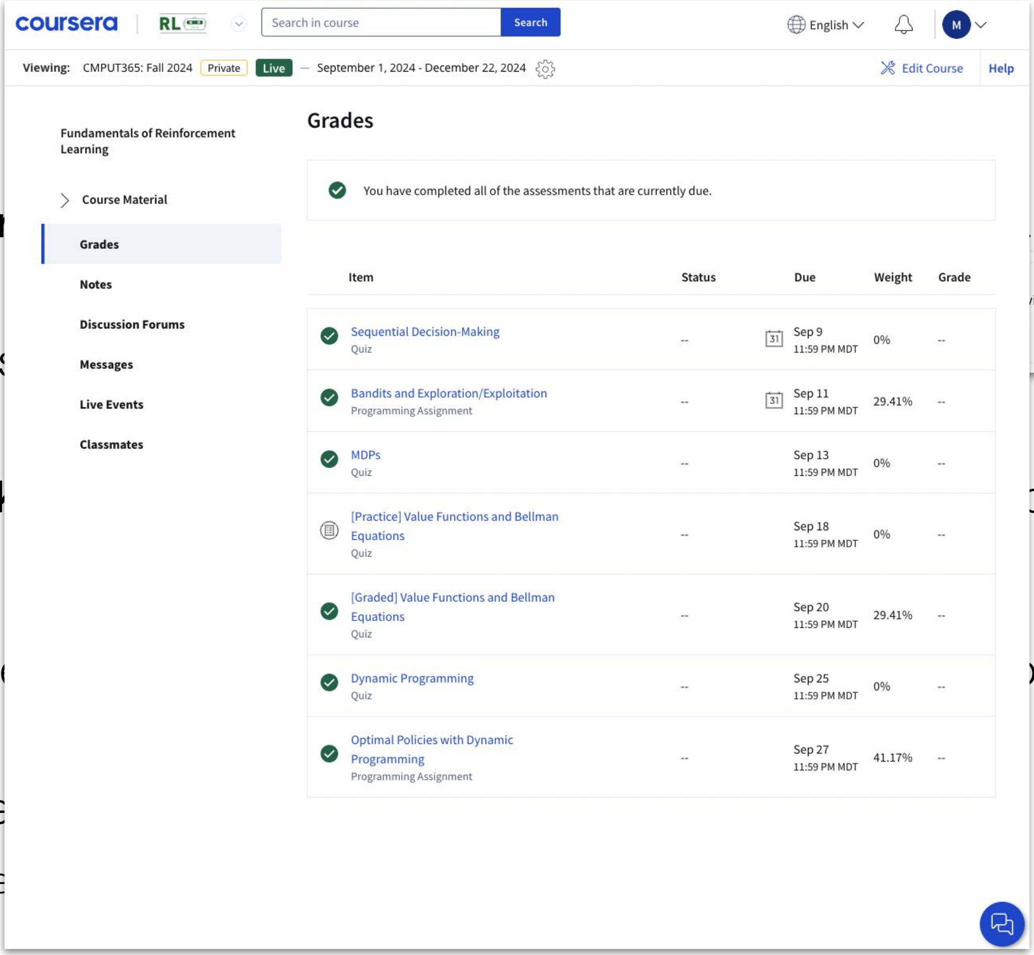
I **cannot** use marks

You **need** to **check** ou are submitting quizzes and assign

The deadlines in the ursera.

If you have any que

`cmput365@ualbe`



Viewing: CMPUT365: Fall 2024 Private Live — September 1, 2024 - December 22, 2024

Fundamentals of Reinforcement Learning

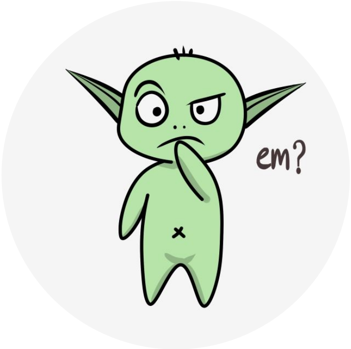
Grades

You have completed all of the assessments that are currently due.

Item	Status	Due	Weight	Grade
✓ Sequential Decision-Making Quiz	--	Sep 9 11:59 PM MDT	0%	--
✓ Bandits and Exploration/Exploitation Programming Assignment	--	Sep 11 11:59 PM MDT	29.41%	--
✓ MDPs Quiz	--	Sep 13 11:59 PM MDT	0%	--
[Practice] Value Functions and Bellman Equations Quiz	--	Sep 18 11:59 PM MDT	0%	--
✓ [Graded] Value Functions and Bellman Equations Quiz	--	Sep 20 11:59 PM MDT	29.41%	--
✓ Dynamic Programming Quiz	--	Sep 25 11:59 PM MDT	0%	--
✓ Optimal Policies with Dynamic Programming Programming Assignment	--	Sep 27 11:59 PM MDT	41.17%	--

Current Enrollments ⓘ
135

Please, interrupt me at any time!



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

Let’s play a game!

https://m.media-amazon.com/images/I/91q+3X0sUZL._AC_SL1500_.jp

Bandits

Arm 1	Arm 2	Arm 3
9, 7, 11, 12, 7, 6	6, 4, 5	8, 9, 9, 10

Reinforcement learning (RL)

- RL is about learning from *evaluative* feedback (an evaluation of the taken actions) rather than *instructive* feedback (being given the correct actions).
    - Exploration is essential in reinforcement learning.

- It is not necessarily about online learning, as said in the videos, but more generally about sequential decision-making.

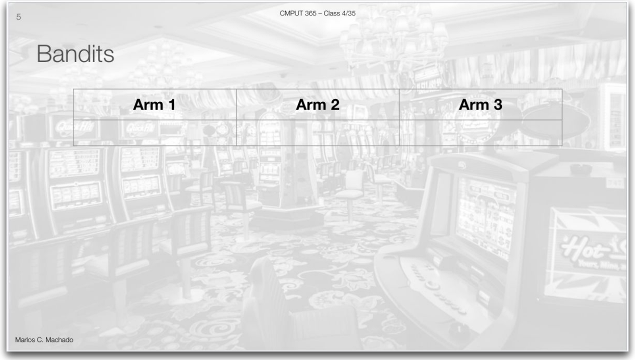
- Reinforcement learning potentially allows for continual learning but in practice, quite often we deploy our systems.

Why study bandits?

- Bandits are the simplest possible reinforcement learning problem.
  - Actions have no delayed consequences.

- Bandits are deployed in so many places! [Source: Csaba's slides]
  - Recommender systems (Microsoft paper):
    - News,
    - Videos,
    - …
  - Targeted COVID-19 border testing (Deployed in Greece, paper).
  - Adapting audits (Being deployed at IRS in the USA, paper).
  - Customer support bots (Microsoft paper).
  - … and more.

Why study bandits?



5

CMPUT 365 – Class 4/35

Bandits

Arm 1	Arm 2	Arm 3

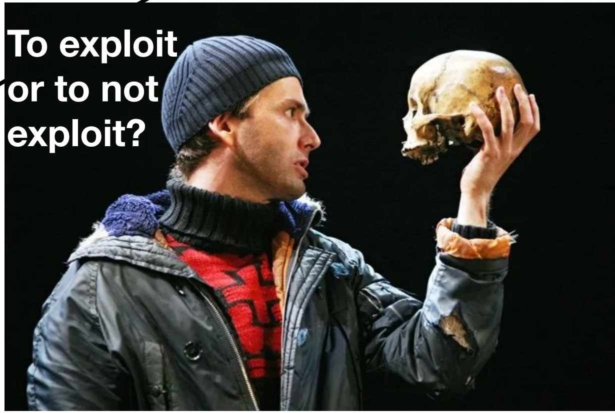
Marlos C. Machado

$$q^*(a) \doteq \mathbb{E}[R_t \mid A_t = a]$$

$$A_t \doteq \operatorname{argmax}_a Q_t(a)$$

Greedy action

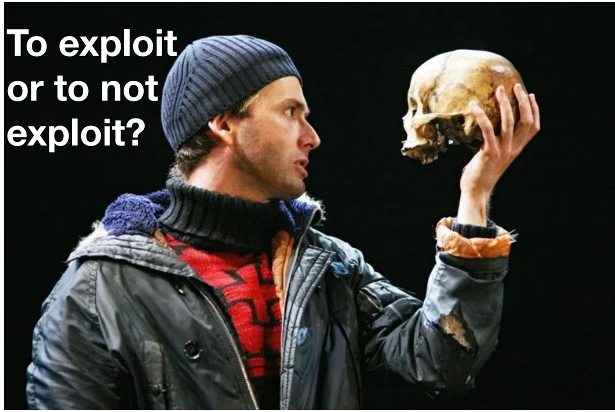
We don't really know q*, so we use an estimate of it, Q_t



https://shakespeareanstudent.files.wordpress.com/2019/10/fb89ff1a-6f63-11e6-acba-85f5c900fc1a1735507721165554264.

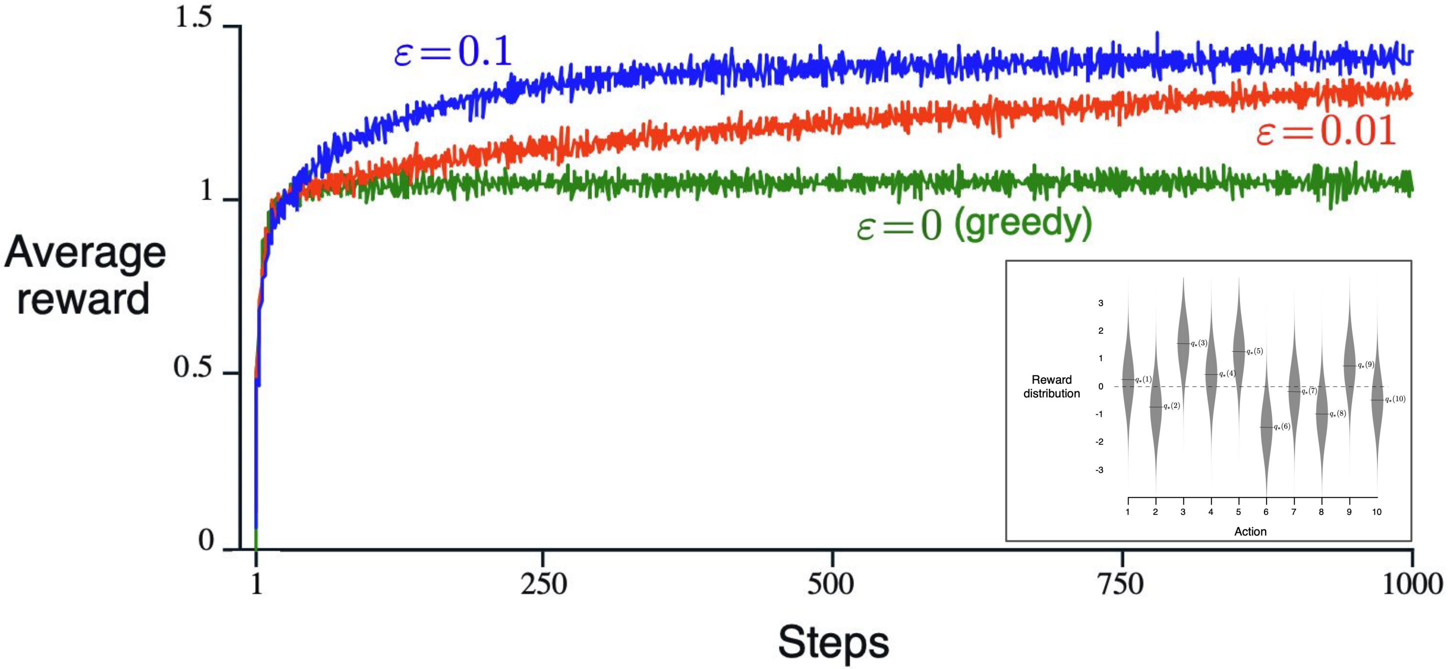
Exploration

- Exploration is the opposite of exploitation.
- It is a whole, very active area of research, despite the textbook not focusing on it.
- How can we explore?
    - Randomly (∈-greedy)
    - Optimism in the face of uncertainty
    - Uncertainty
    - Novelty / Boredom / Surprise
    - Temporally-extended exploration
    - …



https://shakespeareanstudent.files.wordpress.com/2019/10/fb89ff1a-6f63-11e6-acba-85f5c900fc1a1735507721165554264.

Exploration matters



https://shakespeareanstudent.files.wordpress.com/2019/10/fb89ff1a-6f63-11e6-acba-85f5c900fc1a1735507721165554264.

Incremental updates to estimate q_*

$$Q_{n+1} = \frac{1}{n}\sum_{i=1}^{n} R_i$$

Incremental updates to estimate q$_*$

$$\begin{aligned}
Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\
&= \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\left(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\Big(R_n + (n-1)Q_n\Big) \\
&= \frac{1}{n}\Big(R_n + nQ_n - Q_n\Big) \\
&= Q_n + \frac{1}{n}\Big[R_n - Q_n\Big]
\end{aligned}$$

Next class

Reminder: Practice Quiz for Coursera's Fundamentals of RL: Sequential decision-making is due today at midnight.

Programming Assignment for Coursera's Fundamentals of RL: Sequential decision-making is due on Wednesday.

- What **I** plan to do: Wrap up Fundamentals of RL: An introduction to sequential decision-making (Bandits)
  - Go over some of your questions from Slack and eClass.
  - Time permitting, we'll work on some exercises in the classroom.