CS394R
Reinforcement Learning:
Theory and Practice

Scott Niekum and Peter Stone

Department of Computer Science
The University of Texas at Austin

Good Morning Colleagues

Good Morning Colleagues

- Are there any questions?

Logistics

Logistics

- Registering for the course

Logistics

- Registering for the course
- If you missed Tuesday . . .

Logistics

- Registering for the course
- If you missed Tuesday . . .
  - Watch intro lecture video
  - Read webpage carefully
- Email both instructors and TAs

Logistics

- Registering for the course
- If you missed Tuesday . . .
  - Watch intro lecture video
  - Read webpage carefully
- Email both instructors and TAs
- Nice responses!

Logistics

- Registering for the course
- If you missed Tuesday . . .
  - Watch intro lecture video
  - Read webpage carefully
- Email both instructors and TAs
- Nice responses!
  - Length and content mostly good

Logistics

- Registering for the course
- If you missed Tuesday ...
  - Watch intro lecture video
  - Read webpage carefully
- Email both instructors and TAs
- Nice responses!
  - Length and content mostly good
    - Be clear and specific

Logistics

- Registering for the course
- If you missed Tuesday . . .
  - Watch intro lecture video
  - Read webpage carefully
- Email both instructors and TAs
- Nice responses!
  - Length and content mostly good
    - Be clear and specific
    - Short and focussed is fine

Logistics

- Registering for the course
- If you missed Tuesday . . .
  - Watch intro lecture video
  - Read webpage carefully
- Email both instructors and TAs
- Nice responses!
  - Length and content mostly good
    - Be clear and specific
    - Short and focussed is fine
    - Help us help you

Logistics

- Registering for the course
- If you missed Tuesday . . .
  - Watch intro lecture video
  - Read webpage carefully
- Email both instructors and TAs
- Nice responses!
  - Length and content mostly good
    - Be clear and specific
    - Short and focussed is fine
    - Help us help you
    - Also ask in class or on discussion board

More Logistics

More Logistics

- Next readings:

More Logistics

- Next readings:
  - **2nd edition!!**

More Logistics

- Next readings:
  - **2nd edition!!**
  - MDPs and Dynamic Programming

More Logistics

- Next readings:
  - **2nd edition!!**
  - MDPs and Dynamic Programming
  - Budget a good amount of time!

More Logistics

- Next readings:
  - **2nd edition!!**
  - MDPs and Dynamic Programming
  - Budget a good amount of time!
  - Mostly chapter 3 Tuesday, then chapter 4

More Logistics

- Next readings:
  - **2nd edition!!**
  - MDPs and Dynamic Programming
  - Budget a good amount of time!
  - Mostly chapter 3 Tuesday, then chapter 4
  - Single written response to cover both.

More Logistics

- Next readings:
  - **2nd edition!!**
  - MDPs and Dynamic Programming
  - Budget a good amount of time!
  - Mostly chapter 3 Tuesday, then chapter 4
  - Single written response to cover both.
- Do the first exercises and programming assignment

More Logistics

- Next readings:
  - **2nd edition!!**
  - MDPs and Dynamic Programming
  - Budget a good amount of time!
  - Mostly chapter 3 Tuesday, then chapter 4
  - Single written response to cover both.
- Do the first exercises and programming assignment
- Look at resources page

Our Role

Our Role

- Our role isn't to teach RL

Our Role

- Our role isn't to teach RL
- It's to help **you** learn RL

Our Role

- Our role isn't to teach RL
- It's to help **you** learn RL
  - provide context
  - guide your learning (assign readings, exercises, activities)
  - clarify misconceptions

Our Role

- Our role isn't to teach RL
- It's to help **you** learn RL
  - provide context
  - guide your learning (assign readings, exercises, activities)
  - clarify misconceptions
- You have to do the learning

Our Role

- Our role isn't to teach RL
- It's to help **you** learn RL
  - provide context
  - guide your learning (assign readings, exercises, activities)
  - clarify misconceptions
- You have to do the learning
- Read, write, ask, answer, program (investigate)

Let’s Play!

Let's Play!

- I'm a 2-armed bandit

Let's Play!

- I'm a 2-armed bandit
- As a class, you choose which arm

Let's Play!

- I'm a 2-armed bandit
- As a class, you choose which arm
- Maximize your payoff.

Let's Play!

- I'm a 2-armed bandit
- As a class, you choose which arm
- Maximize your payoff.
- The answer:

Let's Play!

- I'm a 2-armed bandit
- As a class, you choose which arm
- Maximize your payoff.
- The answer:

```
(defun l () (+ 5 (random 7)))        expectation: 8

(defun r ()
   (let ((x (random 4)))
     (case x
       (0 20) (1 0) (2 0)
       (3 (+ 7 (random 11))))))      expectation: 8.5
```

Let's Play!

- I'm a 2-armed bandit
- As a class, you choose which arm
- Maximize your payoff.
- The answer:

```
(defun l () (+ 5 (random 7)))          expectation: 8


(defun r ()
   (let ((x (random 4)))
     (case x
       (0 20) (1 0) (2 0)
       (3 (+ 7 (random 11))))))        expectation: 8.5
```

- What about minimizing risk?

N-armed bandit in practice?

N-armed bandit in practice?

- Choosing mechanics
- Choosing a barber/hairdresser

N-armed bandit in practice?

- Choosing mechanics
- Choosing a barber/hairdresser

stationary or non-stationary?

Common Questions

- How to initialize hyperparameters?
- Theoretical guarantees about exploration vs exploitation

Common Questions

- How to initialize hyperparameters?
- Theoretical guarantees about exploration vs exploitation
- Do dynamic epsilon value strategies exist in the field of RL? Are they effective?

Common Questions

- How to initialize hyperparameters?
- Theoretical guarantees about exploration vs exploitation
- Do dynamic epsilon value strategies exist in the field of RL? Are they effective?
- How do we determine the convergence of RL algorithms?
- How to deal with local minima in RL algorithms?

Common Questions

- How to initialize hyperparameters?
- Theoretical guarantees about exploration vs exploitation
- Do dynamic epsilon value strategies exist in the field of RL? Are they effective?
- How do we determine the convergence of RL algorithms?
- How to deal with local minima in RL algorithms?
- How do gradient bandit approaches work?

Shivaram's Slides

Shivaram's Slides

- Steven Callahan: Why are they called "bandit" algorithms?
- Nikos Mouzakis: What changes if we dont have infinite attempts at the bandits, but a limited amount. How should we weight exploration vs exploitation then?
- Natasha Frumkin: Why do we even care about theoretical bounds if they don't hold in practice?

RL Questions

- Neha Akode: How differentiate between an optimization and a reinforcement learning problem?

RL Questions

- Neha Akode: How differentiate between an optimization and a reinforcement learning problem?
- Yigit Ege Bayiz: Bandit problems often use regret as a performance measure, is there a way to extend the notion of regret to RL problems as well?

RL Questions

- Neha Akode: How differentiate between an optimization and a reinforcement learning problem?
- Yigit Ege Bayiz: Bandit problems often use regret as a performance measure, is there a way to extend the notion of regret to RL problems as well?
- Sharachchandra Bhat: If two RL agents are trained against each other would both the policies learnt be the minimax solution?

Non-stationary problems

- Sravan Ankireddy: How do we expect the estimated reward to converge when the true reward is non-stationary?

Non-stationary problems

- Sravan Ankireddy: How do we expect the estimated reward to converge when the true reward is non-stationary?

- Hasan Burhan Beytur: Why is the step-size is kept constant?

Incremental implementation

- Nathaniel Sauerberg: In section 2.4, I was confused by the claim that the incremental implementation for tracking the sample-mean of an arm requires only constant memory. Doesn't it need to keep track of how many times the arm has been pulled (n), which should take log(# times steps) space? The claim only makes sense if this number of times steps is constant, in which case the super naive method is also constant space.

Bandit vs. RL

- Alec Mehra: One good example of the K-armed bandit problem might be driving from your home to work. Here the situation is the same but the driver may have many possible routes to get to work. Of course every time they drive to work the traffic may be slightly different leading to varying actual driving times. The driver should explore for alternative routes but also exploit those routes to find the true average time. We could also apply upper confidence bound selection because we can estimate the total distance of a path and speed limits that would constrain the minimum time required. This may show us that certain paths are highly non optimal and should not be chosen

Gradient Bandits

Assignments

- Monitor and contribute to discussion forums!

Assignments

- Monitor and contribute to discussion forums!
- 1st exercises and programming assignment

Assignments

- Monitor and contribute to discussion forums!
- 1st exercises and programming assignment
- Read Chapters 3 and 4

Assignments

- Monitor and contribute to discussion forums!
- 1st exercises and programming assignment
- Read Chapters 3 and 4
- Submit a reading response by 5pm Monday