CS394R
Reinforcement Learning: Theory and Practice

Peter Stone

Department of Computer Science
The University of Texas at Austin
Good Morning Colleagues
Good Morning Colleagues

- Are there any questions?
Good Morning Colleagues

- Are there any questions?
Logistics

• Responses
Logistics

- Responses
  - All were good
Logistics

• Responses
  – All were good
    – Especially if you summarize, it’s helpful if you flag your questions clearly - especially most “important” ones.
Logistics

- Responses
  - All were good
    - Especially if you summarize, it’s helpful if you flag your questions clearly - especially most “important” ones.
  - Won’t always be able to answer
Logistics

• Responses
  – All were good
    – Especially if you summarize, it’s helpful if you flag your questions clearly - especially most “important” ones.
    – Won’t always be able to answer
  – Also bring them up in class.
Logistics

- Responses
  - All were good
    - Especially if you summarize, it’s helpful if you flag your questions clearly - especially most “important” ones.
    - Won’t always be able to answer
    - Also bring them up in class.
  - Look for programming assignment opportunities!
Logistics

- Responses
  - All were good
    - Especially if you summarize, it’s helpful if you flag your questions clearly - especially most “important” ones.
    - Won’t always be able to answer
    - Also bring them up in class.
  - Look for programming assignment opportunities!
    - OK to reproduce graphs, but then explore variations
Logistics

• Responses

  – All were good
    – Especially if you summarize, it’s helpful if you flag your questions clearly - especially most “important” ones.
    – Won’t always be able to answer
    – Also bring them up in class.

  – Look for programming assignment opportunities!
    – OK to reproduce graphs, but then explore variations
    – First example: Wesley Tansey on self-play TTT
Logistics

● Responses
  – All were good
    – Especially if you summarize, it’s helpful if you flag your questions clearly - especially most “important” ones.
    – Won’t always be able to answer
    – Also bring them up in class.
  – Look for programming assignment opportunities!
    – OK to reproduce graphs, but then explore variations
    – First example: Wesley Tansey on self-play TTT
  – Need a volunteer to present next week.
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: 3 times around.
Let’s Play!

• I’m a 2-armed bandit
• As a class, you choose which arm: 3 times around.
• Maximize your payoff.
Let’s Play!

- I’m a 2-armed bandit
- As a class, you choose which arm: 3 times around.
- Maximize your payoff.
- The answer:
Let’s Play!

- I’m a 2-armed bandit
- As a class, you choose which arm: 3 times around.
- Maximize your payoff.
- The answer:

```lisp
(defun l () (+ 5 (random 7)))
(defun r ()
    (let ((x (random 3)))
        (case x
            (0 20)
            (1 0)
            (2 (+ 7 (random 11)))
        )))
```

- What about minimizing risk?
N-armed bandit in practice?
N-armed bandit in practice?

- Choosing mechanics
- Choosing a barber/hairdresser
Student-led Discussion

- Elad Liebman on how to judge policy performance
What’s Happened Since?

• Interval estimation
What’s Happened Since?

- Interval estimation
- Shivaram’s slides
Chapter 3

- Defines the problem
Chapter 3

• Defines the problem

• Introduces some important notation and concepts.
Chapter 3

• Defines the problem

• Introduces some important notation and concepts.
  – Returns
  – Markov property
  – State/action value functions
  – Bellman equations
Chapter 3

- Defines the problem

- Introduces some important notation and concepts.
  - Returns
  - Markov property
  - State/action value functions
  - Bellman equations
  - Get comfortable with them!
Chapter 3

● Defines the problem

● Introduces some important notation and concepts.
  – Returns
  – Markov property
  – State/action value functions
  – Bellman equations
  – Get comfortable with them!

● Solution methods come next
Chapter 3

- Defines the problem
- Introduces some important notation and concepts.
  - Returns
  - Markov property
  - State/action value functions
  - Bellman equations
  - Get comfortable with them!
- Solution methods come next
  - What does it mean to solve an RL problem?
Formulating the RL problem

- Art more than science
- States, actions, rewards
- Rewards: no hints on how to solve the problem
Formulating the RL problem

• Art more than science

• States, actions, rewards

• Rewards: no hints on how to solve the problem
  – Dependent on next state (p. 66)
Formulating the RL problem

- Art more than science
- States, actions, rewards
- Rewards: no hints on how to solve the problem
  - Dependent on next state (p. 66)
- Discounted vs. non-discounted
Formulating the RL problem

- Art more than science
- States, actions, rewards
- Rewards: no hints on how to solve the problem
  - Dependent on next state (p. 66)
- Discounted vs. non-discounted
- Episodic vs. continuing
Formulating the RL problem

- Art more than science
- States, actions, rewards
- Rewards: no hints on how to solve the problem
  - Dependent on next state (p. 66)
- Discounted vs. non-discounted
- Episodic vs. continuing
- Exercises 3.4, 3.5 (p. 59)
Value functions

- Consider the week 0 environment
Value functions

- Consider the week 0 environment
- For some $s$, what is $V(s)$?
Value functions

- Consider the week 0 environment
- For some $s$, what is $V(s)$?
- OK - consider the policy we ended with
- Now, for some $s$, what is $V(s)$?
Value functions

- Consider the week 0 environment
- For some \( s \), what is \( V(s) \)?
- OK - consider the policy we ended with
- Now, for some \( s \), what is \( V(s) \)?
- Construct \( V \) in undiscounted, episodic case
Value functions

• Consider the week 0 environment
• For some \( s \), what is \( V(s) \)?
• OK - consider the policy we ended with
• Now, for some \( s \), what is \( V(s) \)?
• Construct \( V \) in undiscounted, episodic case
• Construct \( Q \) in undiscounted, episodic case
Value functions

- Consider the week 0 environment
- For some $s$, what is $V(s)$?
- OK - consider the policy we ended with
- Now, for some $s$, what is $V(s)$?
- Construct $V$ in undiscounted, episodic case
- Construct $Q$ in undiscounted, episodic case
- What if it’s discounted?
Value functions

- Consider the week 0 environment
- For some $s$, what is $V(s)$?
- OK - consider the policy we ended with
- Now, for some $s$, what is $V(s)$?
- Construct $V$ in undiscounted, episodic case
- Construct $Q$ in undiscounted, episodic case
- What if it’s discounted?
- What if it’s continuing?
Value functions

- Consider the week 0 environment
- For some $s$, what is $V(s)$?
- OK - consider the policy we ended with
- Now, for some $s$, what is $V(s)$?
- Construct $V$ in undiscounted, episodic case
- Construct $Q$ in undiscounted, episodic case
- What if it's discounted?
- What if it's continuing?
- Continuing tasks without discounting?
Value functions

• Consider the week 0 environment
• For some \( s \), what is \( V(s) \)?
• OK - consider the policy we ended with
• Now, for some \( s \), what is \( V(s) \)?
• Construct \( V \) in undiscounted, episodic case
• Construct \( Q \) in undiscounted, episodic case
• What if it’s discounted?
• What if it’s continuing?
• Continuing tasks without discounting?
• Exercises 3.10, 3.11, 3.17
Markov property

- What is it?
Markov property

- What is it?
- Does it hold in the real world?
Markov property

- What is it?

- Does it hold in the real world?
  - Are any systems "fundamentally" non-Markovian?
Markov property

• What is it?

• Does it hold in the real world?
  – Are any systems "fundamentally" non-Markovian?
  – What if there’s a time horizon?
Markov property

- What is it?

- Does it hold in the real world?
  - Are any systems "fundamentally" non-Markovian?
  - What if there’s a time horizon?

- It’s an ideal
  - Will allow us to prove properties of algorithms
Markov property

- What is it?

- Does it hold in the real world?
  - Are any systems "fundamentally" non-Markovian?
  - What if there’s a time horizon?

- It’s an ideal
  - Will allow us to prove properties of algorithms
  - Algorithms may still work when not provably correct
Markov property

• What is it?

• Does it hold in the real world?
  – Are any systems "fundamentally" non-Markovian?
  – What if there’s a time horizon?

• It’s an ideal
  – Will allow us to prove properties of algorithms
  – Algorithms may still work when not provably correct
  – Could you compensate? Do algorithms change?
Markov property

- What is it?

- Does it hold in the real world?
  - Are any systems "fundamentally" non-Markovian?
  - What if there’s a time horizon?

- It’s an ideal
  - Will allow us to prove properties of algorithms
  - Algorithms may still work when not provably correct
  - Could you compensate? Do algorithms change?
  - If not, you may want different algorithms (Monte Carlo)
Markov property

- What is it?

- Does it hold in the real world?
  - Are any systems "fundamentally" non-Markovian?
  - What if there’s a time horizon?

- It’s an ideal
  - Will allow us to prove properties of algorithms
  - Algorithms may still work when not provably correct
  - Could you compensate? Do algorithms change?
  - If not, you may want different algorithms (Monte Carlo)

- Exercise 3.6