CS394R
Reinforcement Learning:
Theory and Practice

Scott Niekum and Peter Stone

Department of Computer Science
The University of Texas at Austin

Good Morning Colleagues

- Are there any questions?

Logistics

- Midterm Thursday or Friday - 3.5 hours timed

Logistics

- Midterm Thursday or Friday - 3.5 hours timed
- No class Thursday

Logistics

- Midterm Thursday or Friday - 3.5 hours timed
- No class Thursday
  - Forum for AI talk on GT Sophy - Friday 11am

Logistics

- Midterm Thursday or Friday - 3.5 hours timed
- No class Thursday
  - Forum for AI talk on GT Sophy - Friday 11am
- Feedback on final project proposals coming

Logistics

- Midterm Thursday or Friday - 3.5 hours timed
- No class Thursday
  - Forum for AI talk on GT Sophy - Friday 11am
- Feedback on final project proposals coming
- Next step: literature surveys

Logistics

- Midterm Thursday or Friday - 3.5 hours timed
- No class Thursday
  - Forum for AI talk on GT Sophy - Friday 11am
- Feedback on final project proposals coming
- Next step: literature surveys
  - Build on proposal

Logistics

- Midterm Thursday or Friday - 3.5 hours timed
- No class Thursday
  - Forum for AI talk on GT Sophy - Friday 11am
- Feedback on final project proposals coming
- Next step: literature surveys
  - Build on proposal
- Next week's readings

Logistics

- Midterm Thursday or Friday - 3.5 hours timed
- No class Thursday
  - Forum for AI talk on GT Sophy - Friday 11am
- Feedback on final project proposals coming
- Next step: literature surveys
  - Build on proposal
- Next week's readings
  - Options and hierarchy

Logistics

- Midterm Thursday or Friday - 3.5 hours timed
- No class Thursday
  - Forum for AI talk on GT Sophy - Friday 11am
- Feedback on final project proposals coming
- Next step: literature surveys
  - Build on proposal
- Next week's readings
  - Options and hierarchy
  - No longer a textbook

Ch.16: Applications and Case Studies

- Many more applications on resources page

Ch. 16: Applications and Case Studies

- Many more applications on resources page
- Skipped connections to:
  - Ch. 14 psychology
  - Ch. 15 neuroscience

Ch. 16: Applications and Case Studies

- Many more applications on resources page
- Skipped connections to:
  - Ch. 14 psychology
  - Ch. 15 neuroscience
- Ch. 17 summarizes much of what's to come

Context

- Srinivas Bangalore Seshadri: Elaborate on the quote *Applications of reinforcement learning are still far from routine and typically require as much art as science. Making applications easier and more straightforward is one of the goals of current research in reinforcement learning.*

Common Questions

- Doesn't DQN meet conditions of deadly triad?

Common Questions

- Doesn't DQN meet conditions of deadly triad?
  - Yes!

Common Questions

- Doesn't DQN meet conditions of deadly triad?
  - Yes!
  - So how is it stable?

Common Questions

- Doesn't DQN meet conditions of deadly triad?
  - Yes!
  - So how is it stable?
- Clips TD-error to (-1,1)

Common Questions

- Doesn't DQN meet conditions of deadly triad?
  - Yes!
  - So how is it stable?
- Clips TD-error to (-1,1)
- What is *experience replay*? Why use it?

Common Questions

- Doesn't DQN meet conditions of deadly triad?
  - Yes!
  - So how is it stable?
- Clips TD-error to (-1,1)
- What is *experience replay*? Why use it?
  - like Dyna
  - allows the samples not to be strongly correlated
- DQN: How does using a target network help?
  - Avoids chasing a moving target

Other Common Questions

- How does stacking frames make Atari "more Markovian"?

Other Common Questions

- How does stacking frames make Atari "more Markovian"?
- More detail on AlphaGo
  - How does MCTS improve policy in AlphaGo Zero?

Other Common Questions

- How does stacking frames make Atari "more Markovian"?
- More detail on AlphaGo
  - How does MCTS improve policy in AlphaGo Zero?
- Can you transfer real-world data to simulators?

Other Interesting Questions

- Jordi Ramos Chen: Does Samuel's checker-playing program use a typical RL algorithm?

Other Interesting Questions

- Jordi Ramos Chen: Does Samuel's checker-playing program use a typical RL algorithm?

- Yancheng Du: How does a computer program play Jeopardy? Can't it have a big database of Q/A pairs and/or look up the answer on Google?

Other Interesting Questions

- Jordi Ramos Chen: Does Samuel's checker-playing program use a typical RL algorithm?
- Yancheng Du: How does a computer program play Jeopardy? Can't it have a big database of Q/A pairs and/or look up the answer on Google?
- Caroline Wang: In self play, since the network knows what side it's playing, why doesn't it learn losing moves for one side?

Other Interesting Questions

- Jordi Ramos Chen: Does Samuel's checker-playing program use a typical RL algorithm?
- Yancheng Du: How does a computer program play Jeopardy? Can't it have a big database of Q/A pairs and/or look up the answer on Google?
- Caroline Wang: In self play, since the network knows what side it's playing, why doesn't it learn losing moves for one side?
- Zifan Xu: In self play, if there are many winning strategies, how does it not get into a cycle?

Other Interesting Questions

- Yang Hu: AlphaGo requires supervised learning to initialize the policy network, while AlphaGo Zero just uses random weights to initialize the policy network. Intuitively, supervised learning based on human knowledge should be more helpful than random weighting. But the truth is that AlphaGo Zero performs much better than AlphaGo. Is it meaning that human knowledge on Go is actually not correct at all?

Other Interesting Questions

- Yang Hu: AlphaGo requires supervised learning to initialize the policy network, while AlphaGo Zero just uses random weights to initialize the policy network. Intuitively, supervised learning based on human knowledge should be more helpful than random weighting. But the truth is that AlphaGo Zero performs much better than AlphaGo. Is it meaning that human knowledge on Go is actually not correct at all?
  - Sutton's "The Bitter Lesson"