CS395T
Reinforcement Learning: Theory and Practice
Fall 2004

Peter Stone

Department or Computer Sciences
The University of Texas at Austin

Week11a: Tuesday, November 16th

Good Afternoon Colleagues

- Are there any questions?

- (I won't be able to answer them all)

Logistics

- How are the final projects coming?

Helicopter Control

- State: position, orientation, velocity, angular vels
- Actions: Settings of the 4 or 5 controls
- Goal: Hover

Helicopter Control

- State: position, orientation, velocity, angular vels
- Actions: Settings of the 4 or 5 controls
- Goal: Hover
- How would you formulate the problem "by the book"?

Helicopter Control

- State: position, orientation, velocity, angular vels
- Actions: Settings of the 4 or 5 controls
- Goal: Hover
- How would you formulate the problem “by the book”?
- Could you implement that? Why or why not?

Helicopter Control

- State: position, orientation, velocity, angular vels
- Actions: Settings of the 4 or 5 controls
- Goal: Hover
- How would you formulate the problem “by the book”?
- Could you implement that? Why or why not?
- At a high level, what do they do instead?

Helicopter Control

- State: position, orientation, velocity, angular vels
- Actions: Settings of the 4 or 5 controls
- Goal: Hover
- How would you formulate the problem "by the book"?
- Could you implement that? Why or why not?
- At a high level, what do they do instead?
  - Collect a small amount of human expert data
  - Use that to train a **1-step** model (simulator)
  - Determine the optimal policy in the simulator
  - Fly it!

Some meta issues

- Why do the papers use different terminologies?

Some meta issues

- Why do the papers use different terminologies?
- How else do they differ?

Some meta issues

- Why do the papers use different terminologies?
- How else do they differ?
- Why was the Ng paper more understandable?

Ng paper

- Why hover upside down?

Ng paper

- Why hover upside down?
- Why quadratic reward (p. 6)?

Ng paper

- Why hover upside down?
- Why quadratic reward (p. 6)? Or why not (Bagnell eq. 8)

Ng paper

- Why hover upside down?
- Why quadratic reward (p. 6)? Or why not (Bagnell eq. 8)
- PEGASUS - how does it help policy evaluation?

Ng paper

- Why hover upside down?
- Why quadratic reward (p. 6)? Or why not (Bagnell eq. 8)
- PEGASUS - how does it help policy evaluation?
  - General question: is policy good or lucky?

Ng paper

- Why hover upside down?
- Why quadratic reward (p. 6)? Or why not (Bagnell eq. 8)
- PEGASUS - how does it help policy evaluation?
  - General question: is policy good or lucky?
  - Use same random samples for evaluation of each policy

Ng paper

- Why hover upside down?
- Why quadratic reward (p. 6)? Or why not (Bagnell eq. 8)
- PEGASUS - how does it help policy evaluation?
  - General question: is policy good or lucky?
  - Use same random samples for evaluation of each policy
- How does he do policy optimization?

Ng paper

- Why hover upside down?
- Why quadratic reward (p. 6)? Or why not (Bagnell eq. 8)
- PEGASUS - how does it help policy evaluation?
  - General question: is policy good or lucky?
  - Use same random samples for evaluation of each policy
- How does he do policy optimization?
  - greedy hillclimbing over few parameters (the NNs)!
- Could the approach be used to invert the helicopter? Or is it easier just to hover?
- Can it generalize to adverse conditions?
- Where's the power? Is it an easy problem or a powerful approach?

Bagnell paper

- Equation 1 is our standard setup

Bagnell paper

- Equation 1 is our standard setup
- Equation 2 is just a definition of near-optimal

Bagnell paper

- Equation 1 is our standard setup
- Equation 2 is just a definition of near-optimal
- Can get near-optimal with few samples

Bagnell paper

- Equation 1 is our standard setup
- Equation 2 is just a definition of near-optimal
- Can get near-optimal with few samples
  - VC dimension: shatter space of possible "correct action labelings"
- Can't get all the way to optimal
- Equation 8 is reward function: note penalty for leaving known region