CS394R
Reinforcement Learning: Theory and Practice
Fall 2007

Peter Stone

Department of Computer Sciences
The University of Texas at Austin

Good Afternoon Colleagues

- Are there any questions?

Logistics

- Start thinking about final projects

Logistics

- Start thinking about final projects
  - Think of a **domain**
  - Proposals week 8 **or earlier**
  - RL competition entry?

Logistics

- Start thinking about final projects
  - Think of a **domain**
  - Proposals week 8 **or earlier**
  - RL competition entry?
  - Work in pairs?

Logistics

- Start thinking about final projects
  - Think of a **domain**
  - Proposals week 8 **or earlier**
  - RL competition entry?
  - Work in pairs?
- Do your programming assignments!

Monte Carlo on week 0 task

- Episodic, undiscounted
- Equiprobable random action in start state, then prefer right

Monte Carlo on week 0 task

- Episodic, undiscounted
- Equiprobable random action in start state, then prefer right
- State values

Monte Carlo on week 0 task

- Episodic, undiscounted
- Equiprobable random action in start state, then prefer right
- State values
- Action values

Monte Carlo on week 0 task

- Episodic, undiscounted
- Equiprobable random action in start state, then prefer right
- State values
- Action values
  - Why action values preferable?

Monte Carlo on week 0 task

- Episodic, undiscounted
- Equiprobable random action in start state, then prefer right
- State values
- Action values
  - Why action values preferable?
- Relationship to n-armed bandit?

Monte Carlo on week 0 task

- Episodic, undiscounted
- Equiprobable random action in start state, then prefer right
- State values
- Action values
  - Why action values preferable?
- Relationship to n-armed bandit?
  - multiple situations (associative)
  - nonstationary
- (book slides)

Relationship to DP

Relationship to DP

- MC doesn't need a (full) model
  - Can learn from actual or simulated experience

Relationship to DP

- MC doesn't need a (full) model
  - Can learn from actual or simulated experience
- DP takes advantage of a full model
  - Doesn't need **any** experience

Relationship to DP

- MC doesn't need a (full) model
  - Can learn from actual or simulated experience
- DP takes advantage of a full model
  - Doesn't need **any** experience
- MC expense independent of number of states

Relationship to DP

- MC doesn't need a (full) model
  - Can learn from actual or simulated experience
- DP takes advantage of a full model
  - Doesn't need **any** experience
- MC expense independent of number of states
- No bootstrapping in MC

Relationship to DP

- MC doesn't need a (full) model
  - Can learn from actual or simulated experience
- DP takes advantage of a full model
  - Doesn't need **any** experience
- MC expense independent of number of states
- No bootstrapping in MC
  - Not harmed by Markov violations

First/Every Visit

- Why is every visit trickier to analyze?

First/Every Visit

- Why is every visit trickier to analyze?
- Every visit still converges to V^π
  - Singh and Sutton '96 paper
  - Revisited in Chapter 7 (replacing traces)

Blackjack

- Fig. 5.2 (114): Why values mainly independent of dealer showing?
- As true in Fig. 5.5? (121)
- Possible explanation for notch in usable ace policy?
- Why not just use DP?

Control

- Q more useful than V without a model
- But to get it need to explore
- Exploring starts vs. stochastic policies
  - Does ES converge?

Control

- Q more useful than V without a model
- But to get it need to explore
- Exploring starts vs. stochastic policies
  - Does ES converge? Tsitsiklis paper:
    We settle the above mentioned open problem, for the case of a discounted cost criterion, under the assumption that every state-action pair is used to initialize the observed trajectories with the same frequency.

Control

- Q more useful than V without a model
- But to get it need to explore
- Exploring starts vs. stochastic policies
  - Does ES converge? Tsitsiklis paper:

    We settle the above mentioned open problem, for the case of a discounted cost criterion, under the assumption that every state-action pair is used to initialize the observed trajectories with the same frequency.
  - Epsilon-soft vs. epsilon-greedy (122)
  - Why consider off-policy methods?

Learning off policy

- Off policy equations (5.3 and next 2: 125)
- Change week 0 policy from equiprobable in start state to 50/25/25

Learning off policy

- Off policy equations (5.3 and next 2: 125)
- Change week 0 policy from equiprobable in start state to 50/25/25
- Why only learn from tail in Fig. 5.7?

Learning off policy

- Off policy equations (5.3 and next 2: 125)
- Change week 0 policy from equiprobable in start state to 50/25/25
- Why only learn from tail in Fig. 5.7?