CS394R
Reinforcement Learning: Theory and Practice

Amy Zhang and Peter Stone

Departments of ECE and CS
The University of Texas at Austin

Good Morning Colleagues

- Are there any (logistics) questions?

Logistics

- Do programming assignments!

Logistics

- Do programming assignments!
- Next week's readings

Logistics

- Do programming assignments!
- Next week's readings
  - Multi-step bootstrapping

Logistics

- Do programming assignments!
- Next week's readings
  - Multi-step bootstrapping
  - "Planning" and learning (tabular models)

MC vs. DP

MC vs. DP

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

MC vs. 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 vs. 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

MC vs. 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

MC vs. 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 12 (replacing traces)

Control

- Q more useful than V without a model

Control

- Q more useful than V without a model
- But to get it need to explore

Control

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

Learning off policy

- Importance sampling

TD on week 0 task

- Equiprobable random policy
  - Values initialized to 0
  - 3 trajectories

TD on week 0 task

- Equiprobable random policy
  - Values initialized to 0
  - 3 trajectories
- Compare with MC

SARSA vs. Q

- Week 0 example
  - (Remember no access to real model)
  - $\alpha = .1$, ϵ-greedy $\epsilon = .75$, break ties in favor of $\rightarrow$

SARSA vs. Q

- Week 0 example
  - (Remember no access to real model)
  - $\alpha = .1$, ϵ-greedy $\epsilon = .75$, break ties in favor of $\rightarrow$
  - Where did policy change?

SARSA vs. Q

- Week 0 example
  - (Remember no access to real model)
  - $\alpha = .1$, ϵ-greedy $\epsilon = .75$, break ties in favor of $\rightarrow$
  - Where did policy change?
- How do their convergence guarantees differ?

SARSA vs. Q

- Week 0 example
  - (Remember no access to real model)
  - $\alpha = .1$, ϵ-greedy $\epsilon = .75$, break ties in favor of $\rightarrow$
  - Where did policy change?

- How do their convergence guarantees differ?
  - Sarsa depends on policy's dependence on Q:
  - Policy must converge to greedy

SARSA vs. Q

- Week 0 example
  - (Remember no access to real model)
  - $\alpha = .1$, ϵ-greedy $\epsilon = .75$, break ties in favor of $\rightarrow$
  - Where did policy change?
- How do their convergence guarantees differ?
  - Sarsa depends on policy's dependence on Q:
  - Policy must converge to greedy
  - Q-learning value function converges to Q^*
  - As long as all state-action pairs visited infinitely
  - And step-size satisfies stochastic convergence equations

More SARSA vs. Q

- Why does Q-learning learn to hug the cliff? (p. 132)