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

- Class survey — due Thursday

Logistics

- Class survey — due Thursday
- Programming assignments, final project

TD on week 0 task

- Equiprobable random policy

TD on week 0 task

- Equiprobable random policy
- Compare with MC

TD on week 0 task

- Equiprobable random policy
- Compare with MC
- (book slides)

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' 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' 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 (2.8)

R-learning

- Average reward, continuing task

- Ergodic: non-zero probability of reaching any state

R-learning

- Average reward, continuing task
- Ergodic: non-zero probability of reaching any state
- Consider 2-state example

R-learning

- Average reward, continuing task
- Ergodic: non-zero probability of reaching any state
- Consider 2-state example
- Can be Off-policy

R-learning

- Average reward, continuing task
- Ergodic: non-zero probability of reaching any state
- Consider 2-state example
- Can be Off-policy
- R-learning: why negative in 6.17?

R-learning

- Average reward, continuing task
- Ergodic: non-zero probability of reaching any state
- Consider 2-state example
- Can be Off-policy
- R-learning: why negative in 6.17?
- (Afterstates)