CS394R
Reinforcement Learning: Theory and Practice

Peter Stone

Department of Computer Science
The University of Texas at Austin

Good Morning Colleagues

- Are there any questions?

Logistics

- Signup schedule

Logistics

- Signup schedule
- Move to new building?

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

TD on week 0 task

- Equiprobable random policy
  - Values initialized to 0
  - 3 trajectories
- 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'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 (2.8)

More SARSA vs. Q

- Why does Q-learning learn to hug the cliff?
- Ex. 6.10, p. 149 (Exp. Q)

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

Actor-Critic

- How can actor learn continuous actions?

BREAK TIME!

BREAK TIME!

- Bon appetit!

Eligibility Traces

- N-step return in week 0 task (on-line)

Eligibility Traces

- N-step return in week 0 task (on-line)
- On-line vs. off-line in week 0 task

Eligibility Traces

- N-step return in week 0 task (on-line)
- On-line vs. off-line in week 0 task
- TD(λ) on week 0 task