CS395T
Reinforcement Learning:
Theory and Practice
Fall 2004

Peter Stone

Department or Computer Sciences
The University of Texas at Austin

Week6b: Thursday, October 7th

Good Afternoon Colleagues

- Are there any questions?

Good Afternoon Colleagues

- Are there any questions?
- Pending questions:
  - How can actor learn continuous actions?
  - Can knowing actions help the critic?

Good Afternoon Colleagues

- Are there any questions?
- Pending questions:
  - How can actor learn continuous actions?
  - Can knowing actions help the critic?
  - Windy grid - why not MC?
    - Can't we guarantee convergence? (147)

Good Afternoon Colleagues

- Are there any questions?
- Pending questions:
  - How can actor learn continuous actions?
  - Can knowing actions help the critic?
  - Windy grid - why not MC?
    - Can't we guarantee convergence? (147)
  - Afterstates vs. state values?

Logistics

- Fill out survey by 12:30pm tomorrow

Logistics

- Fill out survey by 12:30pm tomorrow
- Chapter 7 important and a bit tricky

Random walks

- Exercises 6.2, 6.4 (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)

Actor-Critic

- Mazda's discussion

Actor-Critic

- Mazda's discussion
- How can actor learn coninuous actions?
- Can knowing actions help the critic?

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 sum converges?

R-learning

- Average reward, continuing task
- Ergodic: non-zero probability of reaching any state
- Consider 2-state example
- Can be Off-policy
- R-learning sum converges?
- 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 sum converges?
- R-learning: why negative in 6.17?
- R-learning better than Q? Converges to optimal? (David)

R-learning

- Average reward, continuing task
- Ergodic: non-zero probability of reaching any state
- Consider 2-state example
- Can be Off-policy
- R-learning sum converges?
- R-learning: why negative in 6.17?
- R-learning better than Q? Converges to optimal? (David)
- (Afterstates)