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

- Do programming assignments!
Logistics

• Do programming assignments!

• Not into piazza?
Logistics

- Do programming assignments!
- Not into piazza?
- Next week’s readings
Logistics

- Do programming assignments!
- Not into piazza?
- Next week’s readings
  - Multi-step bootstrapping
Logistics

- Do programming assignments!
- Not into piazza?
- Next week’s readings
  - Multi-step bootstrapping
  - “Planning” and learning (tabular models)
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?
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 $V^\pi$
  – 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
Control

• $Q$ more useful than $V$ without a model
• But to get it need to explore
• Exploring starts vs. stochastic policies
  – $\pi^*$ always deterministic? (if not, why ES?)
Control

- Q more useful than V without a model
- But to get it need to explore
- Exploring starts vs. stochastic policies
  - $\pi^*$ always deterministic? (if not, why ES?)
  - Does ES converge?
Control

- $Q$ more useful than $V$ without a model
- But to get it need to explore
- Exploring starts vs. stochastic policies
  - $\pi^*$ always deterministic? (if not, why ES?)
  - Does ES converge? Tsitsiklis:
    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
  - $\pi^*$ always deterministic? (if not, why ES?)
  - Does ES converge? Tsitsiklis:
    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.
  - Why consider off-policy methods?
Learning off policy

- Importance sampling slides
Learning off policy

- Importance sampling slides
- Change week 0 policy from equiprobable in start state to 50/25/25
Learning off policy

- Importance sampling slides
- Change week 0 policy from equiprobable in start state to 50/25/25
- Why only learn from tail on p. 115?
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)
  – $\alpha = .1$, $\epsilon$-greedy $\epsilon = .75$, break ties in favor of →
SARSA vs. Q

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

- How do their convergence guarantees differ?
SARSA vs. Q

- Week 0 example
  - (Remember no access to real model)
  - $\alpha = .1$, $\epsilon$-greedy $\epsilon = .75$, break ties in favor of →
  - 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)
  – $\alpha = .1$, $\epsilon$-greedy $\epsilon = .75$, break ties in favor of →
  – 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. 139)