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?
- Understand every step of the math

Logistics

- Do programming assignments!
- Not into piazza?
- Understand every step of the math
  - Go back to sections 3.7 and 3.8 if need be

Chapter 4

- Solution methods **given a model**

Chapter 4

- Solution methods **given a model**
  - So no exploration vs. exploitation

Chapter 4

- Solution methods **given a model**
  - So no exploration vs. exploitation
- Why is it called dynamic programming?

Student Discussion

- Ali on policy iteration

Policy Evaluation

- V^π exists and is unique if $\gamma < 1$ or termination guaranteed for all states under policy π. (p. 90)

Policy Evaluation

- V^π exists and is unique if $\gamma < 1$ or termination guaranteed for all states under policy π. (p. 90)

- Policy evaluation converges under the same conditions (p. 91)

Policy Evaluation

- V^{π} exists and is unique if $\gamma < 1$ or termination guaranteed for all states under policy π. (p. 90)
- Policy evaluation converges under the same conditions (p. 91)
- Policy evaluation on the week 0 problem
  - undiscounted, episodic

Policy Evaluation

- V^{π} exists and is unique if $\gamma < 1$ or termination guaranteed for all states under policy π. (p. 90)
- Policy evaluation converges under the same conditions (p. 91)
- Policy evaluation on the week 0 problem
  - undiscounted, episodic
  - Are the conditions met?

Policy Evaluation

- V^{π} exists and is unique if $\gamma < 1$ or termination guaranteed for all states under policy π. (p. 90)

- Policy evaluation converges under the same conditions (p. 91)

- Policy evaluation on the week 0 problem
  - undiscounted, episodic
  - Are the conditions met?
  - (book slides)

Policy Evaluation

- V^π exists and is unique if $\gamma < 1$ or termination guaranteed for all states under policy π. (p. 90)
- Policy evaluation converges under the same conditions (p. 91)
- Policy evaluation on the week 0 problem
  - undiscounted, episodic
  - Are the conditions met?
  - (book slides)
- Exercises 4.1, 4.2

Policy Improvement

- Policy improvement theorem:
  $\forall s, Q^{\pi}(s, \pi'(s)) \geq V^{\pi}(s) \Rightarrow \forall s, V^{\pi'}(s) \geq V^{\pi}(s)$

Policy Improvement

- Policy improvement theorem:
  $\forall s, Q^{\pi}(s, \pi'(s)) \geq V^{\pi}(s) \Rightarrow \forall s, V^{\pi'}(s) \geq V^{\pi}(s)$

- (book slides)

Policy Improvement

- Policy improvement theorem:
  $\forall s, Q^{\pi}(s, \pi'(s)) \geq V^{\pi}(s) \Rightarrow \forall s, V^{\pi'}(s) \geq V^{\pi}(s)$
- (book slides)
- Polynomial time convergence (in number of states and actions) even though m^n policies.
  - Ignoring effect of γ and bits to represent rewards/transitions

Policy Improvement

- Policy improvement theorem:
  $\forall s, Q^{\pi}(s, \pi'(s)) \geq V^{\pi}(s) \Rightarrow \forall s, V^{\pi'}(s) \geq V^{\pi}(s)$
- (book slides)
- Polynomial time convergence (in number of states and actions) even though m^n policies.
  - Ignoring effect of γ and bits to represent rewards/transitions
- What if non-Markov?

Value Iteration on Week 0 problem

- Show the new policy at each step
  - Not actually to compute policy

Value Iteration on Week 0 problem

- Show the new policy at each step
  - Not actually to compute policy
  - Break policy ties with equiprobable actions

Value Iteration on Week 0 problem

- Show the new policy at each step
  - Not actually to compute policy
  - Break policy ties with equiprobable actions
  - No stochastic transitions

Value Iteration on Week 0 problem

- Show the new policy at each step
  - Not actually to compute policy
  - Break policy ties with equiprobable actions
  - No stochastic transitions

- How would policy iteration proceed in comparison?
  - More or fewer policy updates?

Value Iteration on Week 0 problem

- Show the new policy at each step
  - Not actually to compute policy
  - Break policy ties with equiprobable actions
  - No stochastic transitions
- How would policy iteration proceed in comparison?
  - More or fewer policy updates?
  - True in general?

Chapter 4 Summary

- Chapter 4 treats **bootstrapping** with a model

Chapter 4 Summary

- Chapter 4 treats **bootstrapping** with a model
  – Next: no model and no bootstrapping

Chapter 4 Summary

- Chapter 4 treats **bootstrapping** with a model
  - Next: no model and no bootstrapping
  - Then: no model, but bootstrapping

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?

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?
  - multiple situations (associative)
  - nonstationary
- (book slides)

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^{π}
  - Singh and Sutton '96 paper
  - Revisited in Chapter 7 (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
  - π^* 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
  - π^* 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
  - π^* 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
  - π^* 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.
  - ϵ-soft on-policy: converge to π^*?

Control

- Q more useful than V without a model
- But to get it need to explore
- Exploring starts vs. stochastic policies
  - π^* 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.
  - ϵ-soft on-policy: converge to π^*? what if ϵ changes?

Control

- Q more useful than V without a model
- But to get it need to explore
- Exploring starts vs. stochastic policies
  - π^* 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.
  - ϵ-soft on-policy: converge to π^*? what if ϵ changes?
  - Why consider off-policy methods?

Control

- Q more useful than V without a model
- But to get it need to explore
- Exploring starts vs. stochastic policies
  – π^* 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.
  – ϵ-soft on-policy: converge to π^*? what if ϵ changes?
  – Why consider off-policy methods?

Learning off policy

- Change week 0 policy from equiprobable in start state to 50/25/25

Learning off policy

- Change week 0 policy from equiprobable in start state to 50/25/25

- Off policy equations (5.3 and next 2: 125)

Learning off policy

- Change week 0 policy from equiprobable in start state to 50/25/25
- Off policy equations (5.3 and next 2: 125)
- Why only learn from tail in Fig. 5.7?

Learning off policy

- Change week 0 policy from equiprobable in start state to 50/25/25
- Off policy equations (5.3 and next 2: 125)
- Why only learn from tail in Fig. 5.7?
- How choose behavior policy?

Learning off policy

- Change week 0 policy from equiprobable in start state to 50/25/25
- Off policy equations (5.3 and next 2: 125)
- Why only learn from tail in Fig. 5.7?
- How choose behavior policy?