CS394R
Reinforcement Learning:
Theory and Practice

Scott Niekum and 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!
- Start thinking about final project

Logistics

- Do programming assignments!
- Start thinking about final project
- Next week's readings

Logistics

- Do programming assignments!
- Start thinking about final project
- Next week's readings
  - On-policy prediction with approximation

Bridging Methods

- n-step methods bridge TD and MC
  - TD(0) $\longrightarrow$ MC

Bridging Methods

- n-step methods bridge TD and MC
  - TD(0) $\longrightarrow$ MC
  - All online (model-free)

Bridging Methods

- n-step methods bridge TD and MC
  - TD(0) $\longrightarrow$ MC
  - All online (model-free)

- Today we talk about bridging to DP (model-based)
  - TD,MC $\longrightarrow$ DP (e.g. VI)

Bridging Methods

- n-step methods bridge TD and MC
  - TD(0) $\longrightarrow$ MC
  - All online (model-free)
- Today we talk about bridging to DP (model-based)
  - TD,MC $\longrightarrow$ DP (e.g. VI)
  - Also called learning vs. planning

Bridging Methods

- n-step methods bridge TD and MC
  - TD(0) $\longrightarrow$ MC
  - All online (model-free)
- Today we talk about bridging to DP (model-based)
  - TD,MC $\longrightarrow$ DP (e.g. VI)
  - Also called learning vs. planning
  - Model-based RL does both

Bridging Methods

- n-step methods bridge TD and MC
  - TD(0) $\longrightarrow$ MC
  - All online (model-free)
- Today we talk about bridging to DP (model-based)
  - TD,MC $\longrightarrow$ DP (e.g. VI)
  - Also called learning vs. planning
  - Model-based RL does both
  - computational efficiency vs. sample efficiency

Bridging Methods

- n-step methods bridge TD and MC
  - TD(0) $\longrightarrow$ MC
  - All online (model-free)
- Today we talk about bridging to DP (model-based)
  - TD,MC $\longrightarrow$ DP (e.g. VI)
  - Also called learning vs. planning
  - Model-based RL does both
  - computational efficiency vs. sample efficiency
- Recall TD(0) converges to certainty equivalence model

Bridging Methods

- n-step methods bridge TD and MC
  - TD(0) $\longrightarrow$ MC
  - All online (model-free)
- Today we talk about bridging to DP (model-based)
  - TD,MC $\longrightarrow$ DP (e.g. VI)
  - Also called learning vs. planning
  - Model-based RL does both
  - computational efficiency vs. sample efficiency
- Recall TD(0) converges to certainty equivalence model
  - So does Dyna

2 distinct types of planning

- Model-based learning
  - e.g. Dyna

2 distinct types of planning

- Model-based learning
  - e.g. Dyna
- Lookahead search
  - e.g. Monte Carlo Tree Search (MCTS)

Heuristic Search

- Good Old Fashioned AI (GOFAI)

Heuristic Search

- Good Old Fashioned AI (GOFAI)
- Rich area (even though the book minimizes it)

Heuristic Search

- Good Old Fashioned AI (GOFAI)
- Rich area (even though the book minimizes it)
- Generally searches for a single path, not a policy

Heuristic Search

- Good Old Fashioned AI (GOFAI)
- Rich area (even though the book minimizes it)
- Generally searches for a single path, not a policy
- Uses a relational representation

Heuristic Search

- Good Old Fashioned AI (GOFAI)
- Rich area (even though the book minimizes it)
- Generally searches for a single path, not a policy
- Uses a relational representation
- Main conference: ICAPS

Heuristic Search

- Good Old Fashioned AI (GOFAI)
- Rich area (even though the book minimizes it)
- Generally searches for a single path, not a policy
- Uses a relational representation
- Main conference: ICAPS
- **Not** same as evolutionary search, black-box optimization

Interesting Questions

- Shubhankar Agarwal: Is Planning always better than learning if we have an accurate dynamics model?

Interesting Questions

- Shubhankar Agarwal: Is Planning always better than learning if we have an accurate dynamics model?

- Imran Khan: When would planning and learning give us the exact same policy? Would it only be in the case where we have a perfect model?

Interesting Questions

- Shubhankar Agarwal: Is Planning always better than learning if we have an accurate dynamics model?
- Imran Khan: When would planning and learning give us the exact same policy? Would it only be in the case where we have a perfect model?
- Hsing-Huan Chung: Why would we want to use model-free learning methods if we have a full MDP model? Wouldn't dynamic programming solve the problem easily?

Interesting Questions

- Sharachchandra Bhat: For Dyna-Q, what happens when the transition function has never seen the particular state-action pair?

Interesting Questions

- Sharachchandra Bhat: For Dyna-Q, what happens when the transition function has never seen the particular state-action pair?

- Shenghui Chen: In Dyna-Q, the learned model is used to simulate experience n times; how do we choose n?

Interesting Questions

- Chima: Is there a reason to use vanilla DP over RTDP?

Interesting Questions

- Chima: Is there a reason to use vanilla DP over RTDP?
  - Trajectories may only start from a certain states

Interesting Questions

- Chima: Is there a reason to use vanilla DP over RTDP?
  - Trajectories may only start from a certain states
- Humza Qavi: Why is RTDP guaranteed to find an optimal policy without visiting some of the states?

Summary So Far

- Bandits

Summary So Far

- Bandits
- Markov Decision Processes (MDPs)

Summary So Far

- Bandits
- Markov Decision Processes (MDPs)
- Dynamic Programming (DP)

Summary So Far

- Bandits
- Markov Decision Processes (MDPs)
- Dynamic Programming (DP)
- Monte Carlo (MC)

Summary So Far

- Bandits
- Markov Decision Processes (MDPs)
- Dynamic Programming (DP)
- Monte Carlo (MC)
- Temporal Difference (TD)

Summary So Far

- Bandits
- Markov Decision Processes (MDPs)
- Dynamic Programming (DP)
- Monte Carlo (MC)
- Temporal Difference (TD)
- n-step bootstrapping: TD $\longrightarrow$ MC

Summary So Far

- Bandits
- Markov Decision Processes (MDPs)
- Dynamic Programming (DP)
- Monte Carlo (MC)
- Temporal Difference (TD)
- n-step bootstrapping: TD $\longrightarrow$ MC
- Planning and learning: TD,MC $\longrightarrow$ DP

Summary So Far

- Bandits
- Markov Decision Processes (MDPs)
- Dynamic Programming (DP)
- Monte Carlo (MC)
- Temporal Difference (TD)
- n-step bootstrapping: TD $\longrightarrow$ MC
- Planning and learning: TD,MC $\longrightarrow$ DP
- Next: value function approximation