Review and Wrap Up

Professor Emma Brunskill

CS234 RL

Winter 2025

- Today the 3rd part of the lecture includes some slides from David Silver's introduction to RL slides or modifications of those slides

Where We Are In The Course & Reminders

- Last time: Quiz
- Today: Review and Looking Forward
- Thursday Poster Session. 1:30pm
  - Location: AT&T Patio (Green space behind Computer Science Gates Building).
  - Reminder: Poster should also be uploaded before session. No late days.
  - Note: If the weather is rainy, we may move indoors. We will email by the end of Wed night if the poster session location is changing.
- Final report due: Tuesday March 18 at 6pm. No late days.

Today's Plan

- Quiz Recap
- Review and looking forward

Quiz

-

Today's Plan

- Quiz Recap
- Review and looking forward

Reinforcement Learning

Learning through experience/data to make good decisions under uncertainty

High Level Learning Goals[1]

- Define the key features of RL
- Given an application problem know how (and whether) to use RL for it
- Implement (in code) common RL algorithms
- Describe (list and define) multiple criteria for analyzing RL algorithms and evaluate algorithms on these metrics: e.g. regret, sample complexity, computational complexity, empirical performance, convergence, etc.
- Describe the exploration vs exploitation challenge and compare and contrast at least two approaches for addressing this challenge (in terms of performance, scalability, complexity of implementation, and theoretical guarantees)

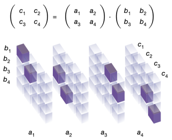
[1]For more detailed descriptions, see website

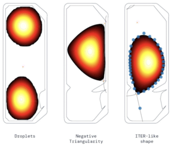
Revisiting Motivating Domains from First Lecture

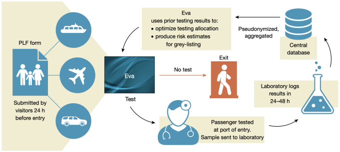
AlphaTensor

Plasma

covid testing

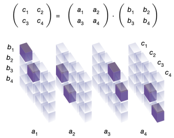




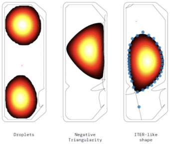


CYU: Answer For One of These Domains

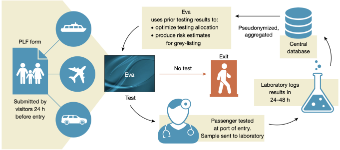
Alpha Tensor



Plasma



Covid testing

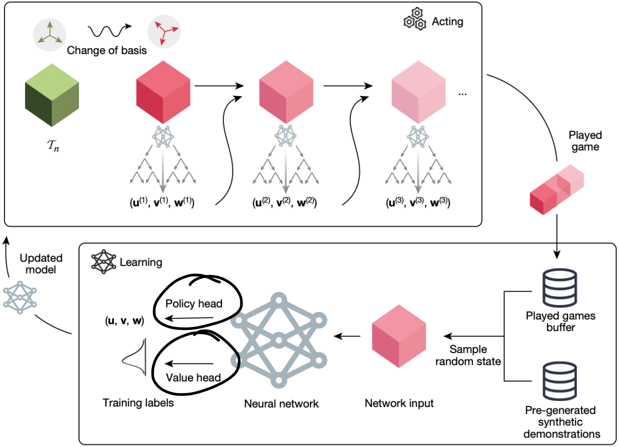


- Which domain are you choosing?
- Is this problem a bandit? A multi-step RL problem?
- Is the problem online / offline or some combination?
- What might the state / action / rewards be?
- What algorithms might be useful here?

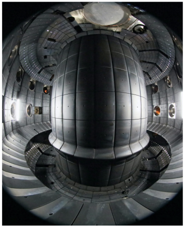
AlphaTensor. Fawzi et al. 2022

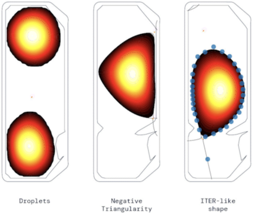
multistep RL

MCTS



Revisiting: Learning Plasma Control for Fusion Science[2]





multistep RL

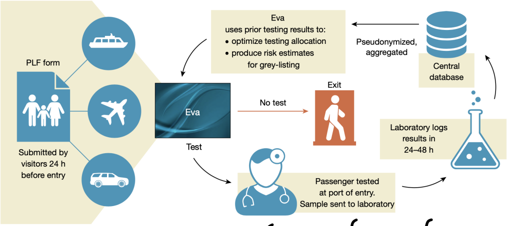
actor critic
simple complex

simulator

penalties in the reward to avoid inaccuracies in the simulator or unsafe unaccuracies

[2]Image credits: left Alain Herzog / EPFL, right DeepMind & SPC/EPFL. Degrave et al. Nature 2022 https://www.nature.com/articles/s41586-021-04301-9

Revisiting: Efficient and targeted COVID-19 border testing via RL [3]

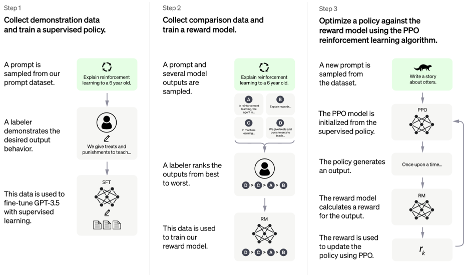


batch bandit w/ delayed outcomes w/ constraints

nonstationarity

[3] Bastani et al. Nature 2021

Revisiting: ChatGPT (https://openai.com/blog/chatgpt/)



Reinforcement Learning

- Learn a policy $\pi(a|s)$ from data to optimize future expected reward
- Optimization, delayed consequences, exploration, generalization
- Actions impact data distribution: rewards observed and states reached

Reinforcement Learning: Standard Settings

- State dependence
  - Bandits: next state independent of prior state and action
  - General decision process: next state depends on prior states and actions
- Online/Offline
  - Offline / batch: Learn from historical data only
  - Online: Agent / algorithm can actively gather its own data

Reinforcement Learning: Core Ideas[4]

- Function approximation + Offpolicy learning is a key challenge
  - New policy introduces new distribution over (s,a,r)
  - Important because want data efficient RL in complex domains
  - PPO: Control with clipping
  - DAGGER: mitigate by obtaining more expert labels
  - Pessimistic Q Learning / CQL / MOPO: introduce pessimism into offline RL

[4] These align closely with many of the core points of Chelsea Finn's Deep RL course summary slides

Reinforcement Learning: Core Ideas[5]

- Function approximation + Offpolicy learning is a key challenge
  - New policy introduces new distribution over (s,a,r)
  - Important because want data efficient RL in complex domains
  - PPO: Control with clipping
  - DAGGER: mitigate by obtaining more expert labels
  - Pessimistic Q Learning / CQL / MOPO: introduce pessimism into offline RL
- Models, values and policies
  - Models: easier to represent uncertainty (why?), useful for MCTS
  - Q function: summarizes performance of policy & and implies policy
  - Policies: the main target of most RL applications
- Computational vs Data Efficiency
  - Data efficient techniques often very computationally intensive
  - In some domains, data = computation (e.g. simulated settings)

[5] These align closely with many of the core points of Chelsea Finn's Deep RL course summary slides

Open Challenges

- Practical, robust RL
  - Robust/stable: Need for automatic hyperparameter tuning, model selection, and generally robust methods for off-the-shelf RL
  - Efficiency: Need for data and computationally efficient methods
  - Hybrid offline-online:
- Framing the problem
  - Alternate formulations to Markov decision processes?
  - Multi-task vs single task?
  - Alternate forms of feedback?
  - Stochastic vs adversarial vs cooperative decision processes?
  - Continuous learning + planning vs system identification then planning?
- Advancing data-driven decision making in domains that could benefit

Learning More

- CS224R Deep RL (Chelsea Finn)
- CS238 Decision Making under Uncertainty (Mykel Kochenderfer)
- CS239 / CS332 Advanced Decision Making Under Uncertainty / RL
- Ben Van Roy often offers an advanced class on bandits or RL

Thanks!

- Thanks for being part of the course!
- We look forward to your posters!