Deep Reinforcement Learning
Review & Frontiers

CS 224R
Course Reminders

- Project poster session on Wednesday! (see Ed post for logistics)
- Final project report due next Monday.

Plan for Today

From me:
- The course in review
- Open challenges

From our guests:
- Research lightning talks!
Connecting the Pieces

Reinforcement learning problem statement

Learn behavior $\pi(a \mid s)$. - from experience, indirect feedback - data not i.i.d.: actions $a$ affect the future observations.

Core solutions

Learning from expert data
- Direct imitation learning
- Learn reward functions

Learning from experience & reward feedback

Online RL
- On-policy
  - Policy gradient
- Off-policy
  - Q-learning

Model-based RL

Offline RL
- Explicit and implicit pessimism (CQL, IQL)
Connecting the Pieces

**Addressing sample inefficiency through transfer**
- Across tasks
- Multi-task RL
- Goal-conditioned RL
- Meta-RL
- From sim to real world
- Aligning dynamics
- Domain randomization
- Fast adaptation

**Addressing limited human supervision**
- **Autonomy**: Learning without environment resets
- **Skill discovery**: Learning useful behaviors without rewards

**Applications**
- Robotics
- Language models
- Education
- Chip design
Some Recurring Themes

Efficient learning requires controlling distribution shift.

- Imitation learning: gather data with DAgger to mitigate shift
- Online/offline RL: limit deviation from current policy / behavior policy

Learned functions can be exploited when optimized against.

- Occurs in Q-learning, model-based RL, reward learning, offline RL.
- Various tools: regularization, ensembles, pessimism
- When applicable: online data collection

Trade-off between computational and data efficiency.

- Data efficient methods often the most computationally heavy (e.g. MBRL).
- Use different methods if in cheap simulator vs. expensive real world.
Open Challenges

**Challenges with core algorithms**

- **Data/computational efficiency**: How long does it take to get a good policy?
- **Stability**: How sensitive is it to hyper parameters, random seed, environment config?
- **Offline workflow**: How to select policies, checkpoints?

**Challenges with assumptions**

- Formulating the problem in the context of MDPs.
- Are MDPs even the right problem formulation?
- What is the source and form of supervision?

You are well-equipped to start to answer some of these questions!
Research Lightning Talks
Thank you!

Thank you for bearing with us as we design a new course!

Thank you for all of your engagement and your feedback!
(Excited to revisit design choices, improve upon the course next Spring!)

We encourage you to fill out the course evaluations.