Overview, Markov Decision Processes

Tengyang Xie

CS839: Mathematical Principles of Reinforcement Learning

What’s this course about?

- A PhD-level seminar course on **RL Theory**
  - When and why RL works
  - Fundamental insights behind RL algorithms
- with focus on sample complexity analyses
- all about proofs, some perspectives, **0 implementation**
- No textbook. See course website for reference materials
- Lectures: slides (outline) + **whiteboard** (details) hybrid

Who should take this course?

This course will be a good fit for you if you either

a) have exposure to RL + comfortable with long mathematical derivations + interested in understanding RL from a purely theoretical perspective

b) have solid grasp in a related theory field (e.g., theoretical computer science or learning theory) and are comfortable with highly abstract description of concepts / models / algorithms

Prerequisites

- Calculus, Linear Algebra, Probability & Statistics
- Optional: stochastic processes, numerical analysis
- Useful
  - ML: sample complexity analysis for supervised learning (PAC)
  - OPT: Convex (linear) optimization, e.g., gradient descent for convex functions
  - STAT: basics of concentration (e.g., Hoeffding’s), tricks such as union bound

Coursework

3~4 assignments ~50%

Course Project ~50%

- Including **proposal**, **report** and **final presentation**
- options: 1) reproduce theoretical analysis in existing papers; 2) work on novel research questions; 3) something in between
- Details provided ~next week

Contents of the courses

- Impossible to cover all the important topics in RL
- Central theme of this course: mathematical principles behind the popular RL ideas/algorithms
- Coverage (tentative)
    1) Fundamentals (MDP, VI, PI)
    2) Classic RL theory (FQI, error propagation, IS, PG, etc)
    3) Exploration and exploitation (optimism, pessimism)
    4) Modern topics

Logistics

- Course website: https://mdp.sh/teaching/2026spring_cs839
  - logistics, links to slides/notes, and all resources
- Canvas for announcements: https://canvas.wisc.edu/courses/500353
- Piazza for Q&A: https://piazza.com/wisc/spring2026/cs839005/home
- Gradescope for HW (and project, TBD): https://www.gradescope.com/courses/1240195
- Office hours: After class + on-demand
- Time: Mon 2:25-5:25pm
- Two Lectures each class (~75 mins each); ~15 min break between

AI Policy

- Prohibited ❌
  - Homework Assignments
  - Course Project
- Allowed ✅
  - Understanding **lecture content**
- When in Doubt
  - Ask me

Motivation: Why RL theory?

RL Success Stories

- **Game Playing:** TD-Gammon (1995), AlphaGo (2016), AlphaZero (2017), OpenAI Five (2018)
- **Robotics:** Manipulation, locomotion, dexterous control
- **Language Models:** ChatGPT, Claude, AI agent
- **Real-World:** Data center cooling, chip design, recommendations

Why Study RL Theory?

- **Sample Complexity:** How many interactions needed to learn a good policy?
- **Convergence Guarantees:** Does the algorithm converge? To what? Under what conditions?
- **Algorithm Design:** Theory guides design of new algorithms
- **Debugging:** When RL fails, theory helps diagnose why

RL vs Supervised Learning

Supervised Learning

- Data: i.i.d. samples (x, y)
- Goal: Learn $f(x) \approx y$
- Feedback: Immediate, complete
- No sequential decisions
- Distribution is fixed

Reinforcement Learning

- Data: Sequential interactions
- Goal: Maximize cumulative reward
- Feedback: Delayed, partial
- Sequential decision making
- Distribution depends on policy!

Major Challenges in RL

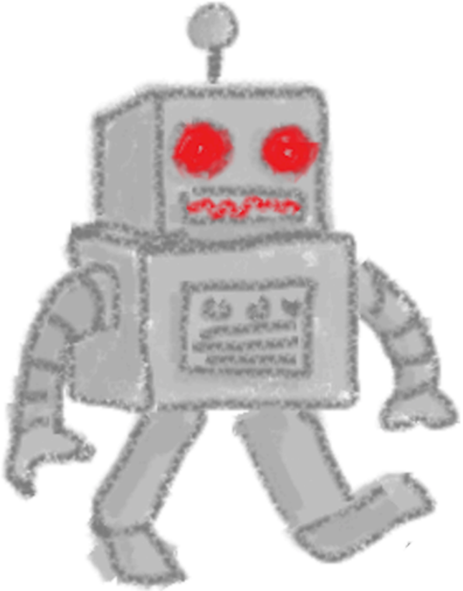
RL solves **sequential decision-making** problems

- **Exploration vs. Exploitation:** Try new actions to learn, or exploit what we know?
- **Generalization:** Decision-making in unseen states.
- **Credit Assignment:** Which past actions led to current reward? Rewards may be delayed.

Introduction to MDPs

Markov Decision Process

Learning Agent



$$a \sim \pi(s)$$

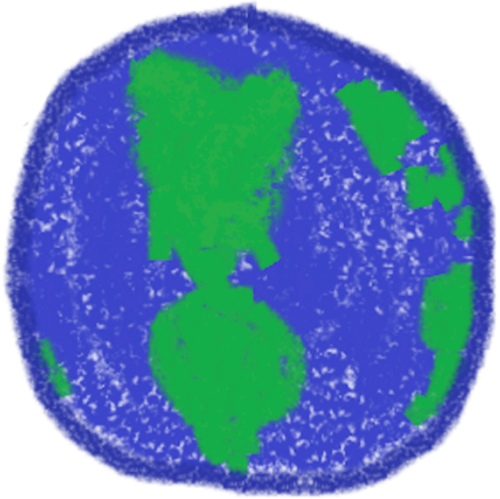
Policy: determine **action** based on **state**

Multiple Steps

Send **reward** and **next state** from a Markovian transition dynamics

$$r(s, a), s' \sim P(\cdot \mid s, a)$$

Environment



Based on slides from Wen Sun