# Reinforcement Learning
CS280 Plan

Zhangir Azerbayev

Spring 2023

**Supervisor:** Prof. Dragomir Radev

**Meetings:** Weekly on Wednesdays

## 1 Course Description

The subject of this course is reinforcement learning (RL) and its applications. A strong background in machine learning, especially deep learning, is assumed. This course will have two phases: the first will cover classical topics in the theory of RL, such as Markov decision processes, dynamic programming, Monte Carlo methods, temporal-difference learning, and policy gradient methods. The second phase of the course will study recent papers in deep reinforcement learning on topics such as game playing, applications in discrete optimization, and RL from human feedback.

## 2 Qualifications and Academics Goals

The student is prepared for this study due to his background in algorithms (CPSC366), machine learning (CPSC481, MATH322), and deep learning (from research in the LILY lab). The goal of this course is to prepare the student for graduate study in machine learning.

## 3 Course Aims

- Understand the mathematical foundations of RL.
- Understand the taxonomy of canonical RL algorithms and their strengths and weaknesses.
- Be able to implement classical RL algorithms (e.g dynamic programming).
- Be prepared for research in deep RL.

## 4 Syllabus

**Textbook:** *Reinforcement Learning: An Introduction*, Richard Sutton and Andrew Barto (2nd edition)

**Final Project:** Student will either replicate the main findings of a recent research paper in deep RL, explore an improvement to a deep RL algorithm, or implement a novel application of deep RL.

**Topics:**

*Part 1: Classical Topics*

Week 1: Multi-Armed Bandits

- Sutton and Barto, Chs. 1, 2

Week 2. MDPs and Dynamic Programming

- Sutton and Barto, Chs. 3, 4

Week 3. TD Learning

- Sutton and Barto, Chs. 5 (skim), 6

Week 4. On-policy prediction

- Sutton and Barto, Ch. 9

Week 5. On-policy control

- Sutton and Barto, Ch. 10

Week 6. RL and the Brain.

- Sutton and Barto, Chs. 14, 15
- [Niv09]

Week 7. RL and Game Theory

- TBD

*Part 2: Research Topics*

Week 8. Deep RL Basics

- Policy Gradient, PPO, SAC, DQN: [Ach18]

Weeks 9 and 10. Game Playing

- Board games: [SHS+17]
- Poker: [BLGS18]
- Atari: [BPK+20]
- real-time strategy: [BBC+19], [VBC+19]

Week 11. RL for discrete optimization

- [Wag21, RRK+21, FBH+22]

Week 12. RL from Human Feedback (RLHF)

- RLHF algorithm: [CLB+17]
- NLP applications: [SOW+20, OWJ+22]

# References


[Ach18] Joshua Achiam. Spinning Up in Deep Reinforcement Learning, 2018.

[BBC+19] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Jozefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique P. d. O. Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with large scale deep reinforcement learning, 2019.

[BLGS18] Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization, 2018.

[BPK+20] Adrià Puigdomenech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, and Charles Blundell. Agent57: Outperforming the atari human benchmark, 2020.

[CLB+17] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2017.

[FBH+22] Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning, 2022.

[Niv09] Yael Niv. Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139–154, 2009.

[OWJ+22] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[RRK+21] Rajarshi Roy, Jonathan Raiman, Neel Kant, Ilyas Elkin, Robert Kirby, Michael Siu, Stuart Oberman, Saad Godil, and Bryan Catanzaro. Prefixrl: Optimization of parallel prefix circuits using deep reinforcement learning. In *2021 58th ACM/IEEE Design Automation Conference (DAC)*, pages 853–858. IEEE, 2021.

[SHS+17] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering chess and shogi by self-play with a general reinforcement learning algorithm, 2017.

[SOW+20] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2020.

[VBC+19] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michael Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[Wag21] Adam Zsolt Wagner. Constructions in combinatorics via neural networks, 2021.