CSE 542: Statistical Reinforcement Learning

Final Project Assignment Spring 2026

University of Washington Instructor: Kevin Jamieson

Overview

The final project is your opportunity to dive into a sub-area of reinforcement learning beyond what we cover in lecture and to communicate that research to your peers. You will (i) read several[1] papers on a focused topic, (ii) write a short literature review, and (iii) produce a **5-minute YouTube video** explaining the area to someone who has taken this course but does not know the specific topic. The list of topics in Section is a starting point. Each topic comes with 2–3 starter references that you should be able to use to begin reading. You are welcome to propose your own topic — come to office hours and we will help you scope it.

Learning goals

- Practice reading research papers in RL theory and synthesizing them.
- Practice communicating technical research clearly and concisely.
- Build a working understanding of one frontier of RL beyond what you saw in class.

Deliverables and timeline

There are two deliverables.

1. **Literature review (2-3 pages).** A self-contained writeup that introduces the problem, surveys the chosen papers, and identifies what is open. The literature review must be typeset in LaTeX and submitted as a single PDF on Gradescope. Think of this as the references and companion resource to the video.
2. **YouTube video (5 minutes).** A self-contained explanation of the topic targeted at your CSE 542 classmates. Hosted on YouTube. You do not need to make it public–it is perfectly acceptable to keep it as "Unlisted" so that it is unsearchable and only accessible by people with the link.

Ed Posting You will make a post on Ed in the *Project Links* area with a title "[Name]: [Project Title]". You will attach your writeup and a link to your YouTube video. You will submit the writeup via Gradescope.

Group size: You must work alone–one project per person. No group work is allowed.

[1] Some areas may not have several papers but most will.

Video specifications

- **Length: 5 minutes** (allow ±30 seconds). The 5-minute constraint is real — a major part of the exercise is learning to compress a research area down to its essentials.
- **Audience:** a fellow CSE 542 student who knows MDPs, Bellman equations, UCB, UCB-VI, MLE plug-in, and pessimism, but is unfamiliar with your specific topic.
- **Content:** (i) set up the problem and how it relates to what is covered in lecture, (ii) state a main result or two precisely (theorem statement, key assumption), (iii) give the one or two ideas that make the proof or algorithm work, (iv) point at open questions and what to read next. Do not try to fit too much in.
- **Format:** slides + voiceover, talking-head + whiteboard, animated explainer, or any combination. Production quality is not graded; clarity and accuracy are. If using a Mac, Quicktime has a convenient screen recording feature.
- **Citations:** include a credits slide at the end listing the papers you discussed.

Grading rubric (out of 100)

- **Literature review — (40 pts).** Theorems are stated correctly; assumptions are clear; the connection between papers is explained, not just listed.
- **Video — (60 pts).** Audible, well-paced, on-time. The 5-minute window is met. At least one big idea is clearly conveyed getting the main point across.

Topic catalogue

The list below organizes topics into ten themes. Each topic comes with 2–3 starter references you can use to orient yourself. The list is not exhaustive — if you have a topic in mind that is not here, come talk.

A. Exploration beyond what we covered

A1. Tabular infinite-horizon / average-reward MDPs. UCRL2 is the canonical model-based-optimism algorithm for the average-reward setting; modern model-free analogues replace the model with optimistic Q-learning.

- Jaksch, Ortner, and Auer (2010). *Near-Optimal Regret Bounds for Reinforcement Learning.* Journal of Machine Learning Research.
- Wei, Jafarnia-Jahromi, Chen, Jain, and Jiang (2020). *Model-Free Reinforcement Learning in Infinite-Horizon Average-Reward Markov Decision Processes.* ICML.

A2. Reward-free exploration. Decouple data collection from reward; given enough exploration data, plan post-hoc for any reward function.

- Jin, Krishnamurthy, Simchowitz, and Yu (2020). *Reward-Free Exploration for Reinforcement Learning.* ICML.
- Kaufmann, Ménard, Domingues, Jonsson, Leurent, and Valko (2021). *Adaptive Reward-Free Exploration.* ALT.

A3. Posterior sampling for RL (PSRL / Thompson sampling). Bayesian alternative to UCB-style optimism; matches the same regret rates and is often empirically better.

- Osband, Russo, and Van Roy (2013). *(More) Efficient Reinforcement Learning via Posterior Sampling.* NeurIPS.
- Russo and Van Roy (2014). *Learning to Optimize via Posterior Sampling.* Mathematics of Operations Research.

A4. Adversarial rewards in tabular MDPs. Replaces the i.i.d.-reward assumption with an adversary; needs online-learning machinery (mirror descent / FTRL) on top of MDP exploration.

- Even-Dar, Kakade, and Mansour (2009). *Online Markov Decision Processes.* Mathematics of Operations Research.
- Rosenberg and Mansour (2019). *Online Convex Optimization in Adversarial Markov Decision Processes.* ICML.
- Jin, Jin, Luo, Sidford, and Sun (2020). *Learning Adversarial Markov Decision Processes with Bandit Feedback and Unknown Transition.* ICML.

A5. Lower bounds for online RL. The natural counterpart to UCB-VI's upper bounds. Le Cam / Fano arguments showing that $\Omega(H\sqrt{SAK})$ is unimprovable.

- Domingues, Ménard, Kaufmann, and Valko (2021). *Episodic Reinforcement Learning in Finite MDPs: Minimax Lower Bounds Revisited.* ALT.
- Jin, Allen-Zhu, Bubeck, and Jordan (2018). *Is Q-Learning Provably Efficient?* NeurIPS. (See the lower-bound section.)

B. Sharper rates

B1. Horizon-free RL when the total return is bounded. The key punchline: when $\sum_h r_h \le 1$, RL has the same leading-order sample complexity as contextual bandits, i.e., the $\mathrm{poly}(H)$ factors disappear.

- Wang, Du, Wang, and Yang (2020). *Is Long Horizon RL More Difficult than Short Horizon RL?* NeurIPS.
- Zhang, Ji, and Du (2021). *Is Reinforcement Learning More Difficult Than Bandits? A Near-Optimal Algorithm Escaping the Curse of Horizon.* COLT.

B2. Variance-aware / Bernstein bonuses. Replacing $H\sqrt{\iota/n}$ by $\sqrt{\sigma^2\iota/n}$ where σ^2 is the variance of $V^\star$ under the dynamics.

- Azar, Osband, and Munos (2017). *Minimax Regret Bounds for Reinforcement Learning.* ICML.
- Zanette and Brunskill (2019). *Tighter Problem-Dependent Regret Bounds in Reinforcement Learning Without Domain Knowledge Using Value Function Bounds.* ICML.

B3. Stochastic shortest path (minimax regret). Goal-reaching problems with no a-priori horizon bound; a natural bridge between finite and infinite horizon.

- Tarbouriech, Garcelon, Valko, Pirotta, and Lazaric (2020). *No-Regret Exploration in Goal-Oriented Reinforcement Learning.* ICML.
- Cohen, Efroni, Mansour, and Rosenberg (2021). *Minimax Regret for Stochastic Shortest Path.* NeurIPS.

C. Function approximation

C1. Linear MDPs: reward-free and instance-dependent (extends the lecture treatment of LSVI-UCB).

- Wagenmaker, Chen, Simchowitz, Du, and Jamieson (2022). *Reward-Free RL is No Harder Than Reward-Aware RL in Linear Markov Decision Processes.* ICML.
- Wagenmaker, Simchowitz, and Jamieson (2022). *Beyond No Regret: Instance-Dependent PAC Reinforcement Learning.* COLT.

C2. Bellman rank and Bellman-eluder dimension. Complexity measures characterizing which function-approximation problems are sample-efficiently learnable.

- Jiang, Krishnamurthy, Agarwal, Langford, and Schapire (2017). *Contextual Decision Processes with Low Bellman Rank are PAC-Learnable.* ICML.
- Jin, Liu, and Miryoosefi (2021). *Bellman Eluder Dimension: New Rich Classes of RL Problems, and Sample-Efficient Algorithms.* NeurIPS.

C3. Decision-Estimation Coefficient (DEC). The unifying complexity measure across bandits, RL, and partial monitoring.

- Foster, Kakade, Qian, and Rakhlin (2021). *The Statistical Complexity of Interactive Decision Making.* arXiv:2112.13487.
- Foster, Golowich, and Han (2023). *Tight Guarantees for Interactive Decision Making with the Decision-Estimation Coefficient.* COLT.

C4. Block MDPs and latent dynamics. Representation learning meets RL: high-dimensional observations that are noisy renderings of a small latent state.

- Du, Krishnamurthy, Jiang, Agarwal, Dudík, and Langford (2019). *Provably Efficient RL with Rich Observations via Latent State Decoding.* ICML.
- Misra, Henaff, Krishnamurthy, and Langford (2020). *Kinematic State Abstraction and Provably Efficient Rich-Observation Reinforcement Learning.* ICML.

D. Policy optimization

D1. Policy gradient: convergence theory. Where lecture 4 ends and policy-based methods begin. Agarwal et al. is the modern theoretical reference for global convergence.

- Sutton, McAllester, Singh, and Mansour (2000). *Policy Gradient Methods for Reinforcement Learning with Function Approximation.* NeurIPS.
- Agarwal, Kakade, Lee, and Mahajan (2021). *On the Theory of Policy Gradient Methods: Optimality, Approximation, and Distribution Shift.* JMLR.
- Mei, Xiao, Szepesvári, and Schuurmans (2020). *On the Global Convergence Rates of Softmax Policy Gradient Methods.* ICML.

D2. Natural policy gradient and policy mirror descent. NPG = mirror descent on Π; linear convergence under entropy regularization.

- Kakade (2001). *A Natural Policy Gradient.* NeurIPS.
- Lan (2023). *Policy Mirror Descent for Reinforcement Learning: Linear Convergence, New Sampling Complexity, and Generalized Problem Classes.* Mathematical Programming.

D3. TRPO and PPO: heuristics and theory. The most-deployed RL algorithms in practice; a project here can review the original heuristic motivation and recent global-optimality analyses.

- Schulman, Levine, Moritz, Jordan, and Abbeel (2015). *Trust Region Policy Optimization.* ICML.
- Schulman, Wolski, Dhariwal, Radford, and Klimov (2017). *Proximal Policy Optimization Algorithms.* arXiv:1707.06347.

E. Partial observability

E1. POMDPs: tractable subclasses. General POMDPs are computationally hard, but undercomplete / observable / decodable subclasses admit efficient algorithms.

- Jin, Kakade, Krishnamurthy, and Liu (2020). *Sample-Efficient Reinforcement Learning of Undercomplete POMDPs.* NeurIPS.
- Liu, Netrapalli, Szepesvári, and Jin (2023). *Optimistic MLE: A Generic Model-Based Algorithm for Partially Observable Sequential Decision Making.* STOC.
- Golowich, Moitra, and Rohatgi (2023). *Planning in Observable POMDPs in Quasipolynomial Time.* STOC.

E2. Predictive state representations (PSRs). A constructive alternative to belief-state DP that sidesteps the hidden-state inference problem.

- Littman, Sutton, and Singh (2002). *Predictive Representations of State.* NeurIPS.
- Singh, James, and Rudary (2004). *Predictive State Representations: A New Theory for Modeling Dynamical Systems.* UAI.

F. Offline RL extensions

F1. Practical offline algorithms (BCQ, CQL, IQL). Industrial descendants of pessimism. Levine et al. is a thorough survey for orientation.

- Fujimoto, Meger, and Precup (2019). *Off-Policy Deep Reinforcement Learning Without Exploration.* ICML. (BCQ)
- Kumar, Zhou, Tucker, and Levine (2020). *Conservative Q-Learning for Offline Reinforcement Learning.* NeurIPS. (CQL)
- Kostrikov, Nair, and Levine (2022). *Offline Reinforcement Learning with Implicit Q-Learning.* ICLR. (IQL)
- Levine, Kumar, Tucker, and Fu (2020). *Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems.* arXiv:2005.01643. (Survey.)

F2. Off-policy evaluation. Estimate V^π from logged data; complementary to PEVI's optimization story.

- Jiang and Li (2016). *Doubly Robust Off-policy Value Evaluation for Reinforcement Learning.* ICML.
- Liu, Li, Tang, and Zhou (2018). *Breaking the Curse of Horizon: Infinite-Horizon Off-Policy Estimation.* NeurIPS.
- Thomas and Brunskill (2016). *Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning.* ICML.

F3. Single-policy concentrability variants. PEVI as covered in lecture assumes coverage of $\pi^\star$; recent work generalizes to any comparator policy and to function-class concentrability.

- Liu, Swaminathan, Agarwal, and Brunskill (2020). *Provably Good Batch Reinforcement Learning Without Great Exploration.* NeurIPS.
- Xie, Cheng, Jiang, Mineiro, and Agarwal (2021). *Bellman-Consistent Pessimism for Offline Reinforcement Learning.* NeurIPS.
- Uehara and Sun (2022). *Pessimistic Model-Based Offline Reinforcement Learning under Partial Coverage.* ICLR.

G. Imitation, RLHF, preference learning

G1. Behavior cloning vs. offline RL. When does pessimism actually help over plain BC?

- Rajaraman, Yang, Jiao, and Ramchandran (2020). *Toward the Fundamental Limits of Imitation Learning.* NeurIPS.
- Spencer, Choudhury, Venkatraman, Ziebart, and Bagnell (2021). *Feedback in Imitation Learning: The Three Regimes of Covariate Shift.* arXiv:2102.02872.

G2. DAgger and interactive imitation. Reduction from imitation learning to no-regret online learning.

- Ross and Bagnell (2010). *Efficient Reductions for Imitation Learning.* AISTATS.
- Ross, Gordon, and Bagnell (2011). *A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning.* AISTATS. (DAgger)

G3. Theory of RLHF / preference-based RL. How does the online–offline split apply when reward is itself learned from preferences?

- Christiano, Leike, Brown, Martic, Legg, and Amodei (2017). *Deep Reinforcement Learning from Human Preferences.* NeurIPS.
- Zhu, Jordan, and Jiao (2023). *Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons.* ICML.
- Rafailov, Sharma, Mitchell, Ermon, Manning, and Finn (2023). *Direct Preference Optimization: Your Language Model is Secretly a Reward Model.* NeurIPS.

H. Multi-agent

H1. Markov games and self-play. Two-player zero-sum analogue of MDPs; sample complexity of Nash equilibria.

- Littman (1994). *Markov Games as a Framework for Multi-Agent Reinforcement Learning.* ICML.
- Bai and Jin (2020). *Provable Self-Play Algorithms for Competitive Reinforcement Learning.* ICML.
- Liu, Yu, Bai, and Jin (2021). *A Sharp Analysis of Model-Based Reinforcement Learning with Self-Play.* ICML.

I. Constrained / risk-sensitive

I1. Constrained MDPs. Maximize reward subject to safety constraints; primal–dual algorithms with regret guarantees in both objective and constraint violation.

- Altman (1999). *Constrained Markov Decision Processes.* Chapman & Hall/CRC. (Foundational text.)
- Achiam, Held, Tamar, and Abbeel (2017). *Constrained Policy Optimization.* ICML.
- Efroni, Mannor, and Pirotta (2020). *Exploration-Exploitation in Constrained MDPs.* arXiv:2003.02189.

I2. Risk-sensitive RL (CVaR / entropic risk). Replace expectation with a tail measure; affects both planning and exploration.

- Tamar, Glassner, and Mannor (2015). *Optimizing the CVaR via Sampling.* AAAI.
- Fei, Yang, and Wang (2021). *Risk-Sensitive Reinforcement Learning with Function Approximation: A Debiasing Approach.* ICML.

J. Foundations / lower bounds

J1. Lower-bound techniques (Le Cam, Fano, coupling). The other side of every upper bound proved in class. The techniques are clean and reusable.

- Domingues, Ménard, Kaufmann, and Valko (2021). *Episodic Reinforcement Learning in Finite MDPs: Minimax Lower Bounds Revisited.* ALT.
- Mannor and Tsitsiklis (2004). *The Sample Complexity of Exploration in the Multi-Armed Bandit Problem.* JMLR.
- Tsybakov (2009). *Introduction to Nonparametric Estimation.* Springer. (Methods reference; see chapters 2–3.)

Practical advice

- **Pick a topic that extends a tool you have already seen.** The video is much easier to record when you can say "in lecture we did X; this paper relaxes assumption Y and pays Z."
- **Read the abstract and introduction of all your starter papers before committing to a topic.** Confirm you can articulate the technical contribution in your own words. If you cannot, switch topics now rather than later.
- **Avoid surveys as your only references.** Pick original source papers.
- **Multiple students can tackle the same project.**
- **Ask for help.** Office hours or on Ed.