CSC415 Course Project Topics

Technical Introductions & Reading Lists

January 1, 2026

Overview

Students may choose one of the following topics for their final project. Each topic includes a comprehensive technical introduction, a list of historical landmark papers, and a selection of recent state-of-the-art (SOTA) papers to guide your research.

---

1 Exploration in Reinforcement Learning

Technical Introduction:
Exploration remains one of the most critical challenges in Reinforcement Learning. It addresses the fundamental trade-off between exploiting current knowledge to maximize reward and exploring the environment to discover potentially superior strategies. In environments with sparse or deceptive rewards (e.g., Montezuma's Revenge), simple dithering strategies like ϵ-greedy fail completely. Research in this domain focuses on designing intrinsic reward signals that guide the agent in the absence of extrinsic feedback. Key approaches include:

- **Count-Based Exploration:** Generalizing visitation counts to continuous state spaces using density models or pseudo-counts.
- **Intrinsic Motivation:** Formulating curiosity based on prediction error (the inability to predict the next state), novelty, or information gain.
- **Maximum Entropy RL:** Encouraging stochastic policies that cover the state space more broadly.
- **Goal-Oriented Exploration:** Automatically generating and pursuing sub-goals to traverse the state space.

Landmark Papers (Foundations):

- Bellemare et al. (2016) - Unifying Count-Based Exploration and Intrinsic Motivation (NeurIPS) [4].
  [Introduces pseudo-counts for continuous exploration].
- Burda et al. (2018) - Exploration by Random Network Distillation (RND) (ICLR) [5].
  [A standard baseline for intrinsic motivation using prediction error of a fixed random network].
- Pathak et al. (2017) - Curiosity-driven Exploration by Self-supervised Prediction (ICML) [29].
  [Introduces the Intrinsic Curiosity Module (ICM) based on inverse dynamics models].
- Ecoffet et al. (2019) - Go-Explore: a New Approach for Hard-Exploration Problems (Nature) [11].
  [Demonstrates the power of returning to promising states rather than just random walking].

State-of-the-Art Papers (2022-2024):

- Henaff et al. (2022) - Exploration via Elliptical Episodic Bonuses (E3B) (NeurIPS) [17].
[Proposes a computationally efficient exploration bonus based on the elliptical potential of the visited state embeddings].

- Zhang et al. (2023) - Novgrid: A Flexible Grid World for Evaluating Agent Creativity and Exploration (AAAI/arXiv) [40].
[While a benchmark paper, it introduces novel metrics and baselines for evaluating open-ended exploration in procedurally generated worlds].

2 Regularization and Representation Learning in RL

Technical Introduction:
Deep RL agents often struggle to generalize beyond their training environments due to overfitting to specific visual cues or dynamics (e.g., background colors, friction). Representation learning aims to map high-dimensional observations (pixels) into a compact, informative latent space that is robust to task-irrelevant distractions. Key techniques include:

- **Data Augmentation:** Applying random crops, color jitters, or rotations to observations to force the agent to learn invariant features.
- **Contrastive Learning:** Using self-supervised losses (like InfoNCE) to learn state embeddings that distinguish between temporally close and distant states.
- **Auxiliary Tasks:** Training the network to predict depth, reward, or future states alongside the policy, enriching the learned representation.
- **Bisimulation Metrics:** Learning state representations that are mathematically equivalent only if they lead to identical future reward sequences.

Landmark Papers (Foundations):

- Laskin et al. (2020) - Reinforcement Learning with Augmented Data (RAD) (NeurIPS) [21].
[Shows that simple image augmentations can match or beat complex model-based methods].

- Srinivas et al. (2020) - CURL: Contrastive Unsupervised Representations for Reinforcement Learning (ICML) [35].
[Applies contrastive losses to learn state representations without reconstruction].

- Jaderberg et al. (2016) - Reinforcement Learning with Unsupervised Auxiliary Tasks (ICLR) [18].
[Introduces the UNREAL agent which learns faster by solving side-tasks].

- Cobbe et al. (2019) - Quantifying Generalization in Reinforcement Learning (ICML) [8].
[Highlights the overfitting problem in RL and introduces the CoinRun benchmark].

State-of-the-Art Papers (2022-2024):

- Yarats et al. (2022) - Mastering Visual Continuous Control: Improved Data-Augmented Reinforcement Learning (DrQ-v2) (ICLR) [38].
[The current gold standard for sample-efficient model-free RL from pixels].

- Schwarzer et al. (2023) - Bigger, Better, Faster: Human-level Atari with human-level efficiency (BBF) (ICML) [33].
[Demonstrates that scaling networks and using SSL objectives allows model-free agents to reach human-level sample efficiency].

- Zhang et al. (2020) - Learning invariant representations for reinforcement learning without reconstruction (ICLR) [39].
  [Achieves robustness to visual distractions by learning bisimulation metrics rather than pixel reconstruction].
- Mondal et al. (2022) - EqR: Equivariant Representations for Data-Efficient Reinforcement Learning (ICML) [26].
  [Leverages geometric symmetries in the environment to improve sample efficiency through equivariant representations].
- Dunion et al. (2023) - Conditional Mutual Information for Disentangled Representations in Reinforcement Learning (NeurIPS) [10].
  [Proposes a new objective based on conditional mutual information to disentangle state representations for better generalization].

3 RL for Robotics

Technical Introduction:
Transferring RL from simulation to the real world ("Sim-to-Real") is the central bottleneck in robotic learning. Real-world robots are expensive, fragile, and data-inefficient compared to simulated agents. Research topics include:

- **Sim-to-Real Transfer:** Using Domain Randomization (varying physics/visuals in sim) so the real world looks like just another variation.
- **Sample Efficiency:** Algorithms like Soft Actor-Critic (SAC) that can learn stable policies with minimal interaction.
- **Offline RL:** Learning from pre-recorded datasets of human or robot demonstrations without online interaction.
- **Residual Learning:** Learning a corrective "delta" policy on top of a classical controller.

Landmark Papers (Foundations):

- Haarnoja et al. (2018) - Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor (ICML) [13].
  [The gold standard algorithm for continuous control robotics].
- Tobin et al. (2017) - Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World (IROS) [36].
  [Key technique for Sim-to-Real transfer].
- Andrychowicz et al. (2020) - Learning Dexterous In-Hand Manipulation (IJRR/OpenAI) [28].
  [Demonstrates solving the Rubik's cube with a robot hand using RL].
- Johannink et al. (2019) - Residual Reinforcement Learning for Robot Control (ICRA) [20].
  [Combines deep RL with conventional feedback control to solve complex manipulation tasks].

State-of-the-Art Papers (2022-2024):

- Zhao et al. (2020) - Domain Generalization via Entropy Regularization (NeurIPS) [41].
  [Introduces entropy regularization techniques to improve the generalization of policies across different domains].
- Seo et al. (2023) & Dunion et al. (2024) - Multi-View Robotic Fusion [9, 34].
  [Approaches for fusing multiple camera streams to handle occlusions and improve state estimation in robotics].

4 Sequence Modeling & Action Chunking (Generative RL)

Technical Introduction:
This emerging paradigm reframes Reinforcement Learning not as a dynamic programming problem (Bellman updates), but as a sequence modeling problem. Inspired by the success of Large Language Models (LLMs), these approaches treat a trajectory of (state, action, reward) as a sequence of tokens.

- **Decision Transformers:** Predicting the next action given the history and a target return (conditioning on the desired outcome).
- **Action Chunking:** Instead of outputting a single action a_t at every high-frequency timestep (which is hard for long horizons), the policy predicts a "chunk" of actions $a_{t:t+k}$ to execute open-loop. This is crucial for high-frequency robotic control.
- **Diffusion Models:** Generating entire trajectories of optimal behavior using diffusion probabilistic models, allowing for flexible constraints and "inpainting" of behaviors.

Landmark Papers (Foundations):

- Chen et al. (2021) - Decision Transformer: Reinforcement Learning via Sequence Modeling (NeurIPS) [6].
  [Treats RL as a sequence modeling problem using Transformers, predicting actions from states and returns].
- Janner et al. (2022) - Planning with Diffusion for Flexible Behavior Synthesis (Diffuser) (ICML) [19].
  [Uses diffusion models to generate entire trajectories of states and actions].
- Zhao et al. (2023) - Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware (ACT) (RSS/CoRL) [42].
  [Introduces Action Chunking with Transformers (ACT) explicitly for robotics imitation learning].
- Ajay et al. (2022) - Is Conditional Generative Modeling all you need for Decision-Making? (ICLR) [2].
  [Argues for generative models over standard Bellman updates].

State-of-the-Art Papers (2022-2024):

- Li et al. (2025) - Action Chunking [22].
  [Uses Transformers to predict "chunks" of future actions, smoothing control and overcoming latency].
- Chi et al. (2023) - Diffusion Policy for RL [7, 31].
  [Models the robot's policy as a conditional diffusion process, enabling the generation of multimodal behavior].

5 Model-Based RL and World Models

Technical Introduction:
Model-Based RL (MBRL) addresses sample inefficiency by learning a model of the environment's dynamics ($P(s'|s,a)$) and rewards. The agent can then "dream" or plan inside this learned model to update its policy without interacting with the real world. Key areas of innovation:

- **Latent Dynamics:** Learning the model in a compressed latent space rather than pixel space to avoid the difficulty of predicting high-dimensional images.
- **Uncertainty Estimation:** Using ensembles of models to estimate epistemic uncertainty and avoid model exploitation (planning in regions where the model is wrong).
- **Planning Algorithms:** Integrating tree search (MCTS) or trajectory optimization (CEM) with learned value functions.

Landmark Papers (Foundations):

- Ha & Schmidhuber (2018) - World Models (NeurIPS) [12].
  [Separates the "vision" model (VAE) from the "memory" model (RNN) to learn inside a dream].
- Hafner et al. (2019/2020) - Dreamer: Scalable Reinforcement Learning Using World Models (ICLR) [14].
  [State-of-the-art model-based agent that learns behaviors purely from latent imagination].
- Chua et al. (2018) - Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models (PETS) (NeurIPS).
  [Uses ensembles of probabilistic networks to handle model uncertainty].
- Schrittwieser et al. (2020) - Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model (MuZero) (Nature) [32].
  [Learns a value-equivalent model for planning without reconstructing observations].

State-of-the-Art Papers (2022-2024):

- Hafner et al. (2023) - Mastering Diverse Domains through World Models (DreamerV3) (arXiv/ICLR) [15].
  [The first algorithm to solve tasks across disparate domains (Atari, Minecraft, Robotics) with fixed hyperparameters].
- Micheli et al. (2023) - Transformers are Sample Efficient World Models (IRIS) (ICLR) [25].
  [Demonstrates that replacing RNNs with Transformers in world models leads to superior sample efficiency in Atari].
- Hansen et al. (2022) - Temporal Difference Learning for Model Predictive Control (TD-MPC) (ICML) [16].
  [Combines model-based planning with model-free value estimation for robust continuous control].

Advanced Topics

A. Hierarchical RL (HRL)

Technical Introduction:
HRL tackles the "curse of dimensionality" in time. For tasks with long horizons (thousands of steps), standard RL suffers from vanishing gradients and credit assignment issues. HRL decomposes the problem into layers of abstraction:

- **High-Level Controller (Manager):** Selects abstract "goals" or "options" (temporally extended macro-actions).
- **Low-Level Controller (Worker):** Executes primitive actions to achieve the goal set by the Manager.

Current research focuses on Goal Representation (how to define sub-goals?), Skill Discovery (learning useful sub-policies without supervision), and Off-policy correction (handling non-stationary transitions between levels).

Landmark Papers (Foundations):

- Vezhnevets et al. (2017) - FeUdal Networks for Hierarchical Reinforcement Learning (ICML) [37].
[A Manager-Worker architecture where the Manager sets abstract goals for the Worker].
- Nachum et al. (2018) - Data-Efficient Hierarchical Reinforcement Learning (HIRO) (NeurIPS) [27].
[Introduces off-policy correction to allow high-level policies to learn from old low-level transitions].
- Bacon et al. (2017) - The Option-Critic Architecture (AAAI) [3].
[Extends actor-critic methods to learn "Options" (temporally extended actions) end-to-end].
- Eysenbach et al. (2019) - Search on the Replay Buffer: Bridging Planning and Reinforcement Learning (NeurIPS).
[Uses HRL to plan using waypoints from the replay buffer].

State-of-the-Art Papers (2022-2024):

- Li et al. (2023) - Hierarchical Reinforcement Learning with Timed Subgoals (NeurIPS) [23].
[Addresses the issue of temporal abstraction by explicitly learning the duration of subgoals].

B. Safe Reinforcement Learning

Technical Introduction:
Deploying RL in the real world (e.g., autonomous driving, healthcare, industrial control) requires strict safety guarantees. Standard RL maximizes expected return, which may involve unacceptable risks during exploration. Safe RL reformulates the problem as a Constrained MDP (CMDP): maximize reward subject to cost constraints (e.g., "drive fast but keep collision probability $< 0.1\%$"). Techniques include:

- **Lagrangian Relaxation:** Converting constraints into adaptive penalty terms in the reward function.
- **Lyapunov Functions:** Creating stability certificates that guarantee the agent stays within a safe region of the state space.

- **Safety Layers:** Analytical shields that override the RL agent's actions if they violate safety rules.

Landmark Papers (Foundations):

- Achiam et al. (2017) - Constrained Policy Optimization (CPO) (ICML) [1].
  [The standard for guaranteeing constraint satisfaction during policy updates].
- Ray et al. (2019) - Benchmarking Safe Exploration in Deep Reinforcement Learning (OpenAI) [30].
  [Provides the Safety Gym environment and benchmarks for safe RL].
- Chow et al. (2018) - A Lyapunov-based Approach to Safe Reinforcement Learning (NeurIPS).
  [Uses Lyapunov functions to guarantee stability and safety].
- Dulac-Arnold et al. (2019) - Challenges of Real-World Reinforcement Learning (ICML).
  [Comprehensive overview of safety and other friction points in applied RL].

State-of-the-Art Papers (2022-2024):

- Liu et al. (2022) - Constrained Variational Policy Optimization for Safe Reinforcement Learning (ICML) [24].
  [Improves upon CPO by using variational inference for better stability].

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pages 22–31. PMLR, 2017.

[2] Anurag Ajay, Yilun Du, Pieter Abbeel, Sergey Levine, Ali Farhadi, and Mohit Mohammad. Is conditional generative modeling all you need for decision-making? In *International Conference on Learning Representations*, 2022.

[3] Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[4] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[5] Yuri Burda, Harrison Edwards, Amos Storkey, and Deepak Pathak. Exploration by random network distillation. In *International Conference on Learning Representations*, 2018.

[6] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems*, volume 34, pages 15084–15097, 2021.

[7] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023.

[8] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *International conference on machine learning*, pages 1282–1289. PMLR, 2019.

[9] Mhairi Dunion and Stefano V Albrecht. Multi-view disentanglement for reinforcement learning with multiple cameras. *arXiv preprint arXiv:2404.14064*, 2024.

[10] Mhairi Dunion, Trevor McInroe, Kevin Luck, Josiah Hanna, and Stefano Albrecht. Conditional mutual information for disentangled representations in reinforcement learning. *Advances in neural information processing Systems*, 36:80111–80129, 2023.

[11] Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. First return, then explore. *Nature*, 590(7847):580–586, 2021.

[12] David Ha and Jürgen Schmidhuber. World models. In *Advances in Neural Information Processing Systems*, pages 3078–3087, 2018.

[13] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870. PMLR, 2018.

[14] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

[15] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. In *arXiv preprint arXiv:2301.04104*, 2023.

[16] Nicklas Hansen, Haojun Wang, and Linxi Fan. Temporal difference learning for model predictive control. In *International Conference on Machine Learning*, pages 8387–8406. PMLR, 2022.

[17] Mikael Henaff, Roberta Raileanu, Minqi Jiang Ferris, and Tim Rocktäschel. Exploration via elliptical episodic bonuses. In *Advances in Neural Information Processing Systems*, volume 35, pages 37631–37646, 2022.

[18] Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*, 2016.

[19] Michael Janner, Yilun Fu, Amy Zhang, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, pages 9902–9915. PMLR, 2022.

[20] Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *2019 international conference on robotics and automation (ICRA)*, pages 6023–6029. IEEE, 2019.

[21] Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. *Advances in neural information processing systems*, 33:19884–19895, 2020.

[22] Qiyang Li, Zhiyuan Zhou, and Sergey Levine. Reinforcement learning with action chunking. *arXiv preprint arXiv:2507.07969*, 2025.

[23] Siyuan Li et al. Hierarchical reinforcement learning with timed subgoals. In *Advances in Neural Information Processing Systems*, 2023.

[24] Zuxin Liu et al. Constrained variational policy optimization for safe reinforcement learning. In *International Conference on Machine Learning*, 2022.

[25] Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *International Conference on Learning Representations*, 2023.

[26] Arnab Kumar Mondal, Vineet Jain, Kaleem Siddiqi, and Siamak Ravanbakhsh. Eqr: Equivariant representations for data-efficient reinforcement learning. In *International Conference on Machine Learning*, pages 15908–15926. PMLR, 2022.

[27] Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[28] OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Pintineau, Matthias Plappert, Glenn Powell, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

[29] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, pages 2771–2780. PMLR, 2017.

[30] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *OpenAI Spinning Up*, 2019.

[31] Allen Z. Ren, Justin Lidard, Lars L. Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy optimization. In *arXiv preprint arXiv:2409.00588*, 2024.

[32] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Koplowitz, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588 (7839):604–609, 2020.

[33] Max Schwarzer, Johan Obando Ceron, Aaron Courville, Etienne Marcotte, Mohammad Mahdi Ghadiri, Ankesh Anand, Rishabh Agarwal, and Pablo Samuel Castro. Bigger, better, faster: Human-level atari with human-level efficiency. In *International Conference on Machine Learning*, 2023.

[34] Younggyo Seo, Junsu Kim, Stephen James, Kimin Lee, Jinwoo Shin, and Pieter Abbeel. Multi-view masked world models for visual robotic manipulation. In *International Conference on Machine Learning*, pages 30613–30632. PMLR, 2023.

[35] Aravind Srinivas, Michael Laskin, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pages 9049–9058. PMLR, 2020.

[36] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30, 2017.

[37] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. In *International Conference on Machine Learning*, pages 3540–3549. PMLR, 2017.

[38] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=_SJ-_yyes8.

[39] Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2020.

[40] Ben Zhang, Kyriacos Voudouris, Tianye Wang, Matthew Crosby, and Murray Shanahan. Novgrid: A flexible grid world for evaluating agent creativity and exploration. *arXiv preprint arXiv:2303.01158*, 2023.

[41] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. *Advances in neural information processing systems*, 33:16096–16107, 2020.

[42] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, 2023.