Project Proposal for ECE595: RL Theory and Algorithms
”Counterfactual Data Augmentation for Sample-Efficient RL”

Sajan K
kumar836@purdue.edu

Shilpa N
snoushad@purdue.edu

Pratyush U
puppulur@purdue.edu

1. Objective

This project aims to implement and validate a counterfactual-based data augmentation approach, as proposed in the CTRL framework [3], to investigate its effectiveness in addressing two core challenges in practical reinforcement learning — sample efficiency and exploration under limited data regimes. Through empirical studies on benchmark tasks (*MountainCar*, *LunarLander*), the goal is to demonstrate how causal counterfactual reasoning can serve as a viable mechanism for safe and data-efficient policy improvement.

2. Motivation

Reinforcement Learning (RL) has shown remarkable success in simulated settings, but its real-world adoption in domains such as healthcare, robotics, and industrial control remains limited by costly data collection and unsafe exploration. Improving **sample efficiency and exploration** has thus become a central goal in practical RL, motivating approaches that learn effectively from small, fixed datasets while preserving theoretical convergence guarantees.

The CTRL framework (*Sample-Efficient Reinforcement Learning via Counterfactual-Based Data Augmentation*, NeurIPS 2020) introduces a **Structural Causal Model (SCM)** to represent environment dynamics as

$$S_{t+1} = f(S_t, A_t, U_{t+1}), \tag{1}$$

where U_{t+1} denotes exogenous noise. By inferring this latent variable for each observed transition and reusing it to compute **counterfactual next states** $S'_{t+1} = f(S_t, a', U_{t+1})$ for alternate actions, CTRL generates additional, causally consistent experiences without new interactions—enabling **sample-efficient “imaginative exploration”** that broadens data coverage by asking *“what if the agent had taken action a' instead of a?”* The paper shows that if f is monotone in U, counterfactual outcomes are identifiable (Theorem 1 in [3]) and that Q-learning trained on this augmented dataset converges to the optimal value function Q^* (Theorem 2 in [3]), aligning with theoretical guarantees on convergence emphasized in the course. Together, these results make CTRL particularly compelling—uniting causal validity and reinforcement learning optimality within a single, data-efficient framework.

To learn this causal mechanism, CTRL employs a **Bidirectional Conditional GAN (BiCoGAN)[2]** that jointly trains an encoder, generator, and discriminator, ensuring consistency between causal and inverse mappings, enabling inference of latent noi/??//se and reconstruction of next states. With its theoretical grounding and simplicity, CTRL offers a practical, scalable framework to evaluate sample efficiency and exploration on benchmark tasks like *MountainCar* and *LunarLander* [1].

3. Plan

Three-Week Timeline

Week 1: SCM Modeling and Data Preparation

- Implement core SCM modules and collect small offline datasets for *MountainCar* and *LunarLander*.
- Train initial MLP-based SCMs for next-state reconstruction and noise inference.

Week 2: Counterfactual Augmentation and RL Integration

- Generate counterfactual transitions for alternate actions and build an augmented replay buffer.
- Integrate augmented data into DQN/D3QN training and evaluate performance across dataset sizes.

Week 3: Architecture Exploration and Evaluation

- Test alternative SCM architectures and compare sample efficiency and stability.
- Produce evaluation plots and compile the final report on counterfactual augmentation benefits.

Deliverables

- Modular PyTorch implementation (`scm/`, `cf_augment/`, `agents/`) with configurable architectures.
- Experimental results comparing baseline and counterfactual-augmented RL across *MountainCar* and *LunarLander*.
- Final report analyzing sample efficiency, exploration coverage, and architectural effects on performance.

References

[1] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. https://arxiv.org/abs/1606.01540, 2016. arXiv:1606.01540. 1

[2] Ayush Jaiswal, Wael AbdAlmageed, Yue Wu, and Premkumar Natarajan. Bidirectional conditional generative adversarial networks. In *Computer Vision – ACCV 2018*, pages 216–232, Cham, 2019. Springer International Publishing. 1

[3] Jun Zhang, Dragomir Radev, Le Song, Devendra Subramanian, Haoran Xu, Wenlong Liao, and Jimeng Sun. Sample-efficient reinforcement learning via counterfactual-based data augmentation. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Workshop on Causal Discovery and Causality-Inspired Machine Learning*, 2020. arXiv:2012.09092. 1