Master Thesis
Prüfungsnr.: xxxxxx
Prüfer: Prof. Dr. Marius Kloft

AN EMPIRICAL STUDY OF SPECTRAL REGULARIZATION FOR MITIGATING SPURIOUS CORRELATIONS IN REINFORCEMENT LEARNING

November 2nd, 2025

Zahra Khodabakhshian
Mtrk.nr.: 426198
Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau (RPTU)
jov98mam@rptu.de

Contents

1. Introduction

Reinforcement learning (RL) has shown strong results across a wide range of control and decision-making tasks. However, ensuring that learned policies generalize beyond their training environments remains a major challenge. In particular, when training environments contain correlations that do not reflect the true structure of the task, agents may learn to rely on shortcuts. These spurious correlations can lead to policies that perform well during training but fail once the environment changes. This issue is especially common in realistic RL settings, where observations are often influenced by latent or unobserved factors that introduce non-causal correlations between state components [1].

To address this problem, recent work in robust reinforcement learning has proposed explicit mitigation strategies, often motivated by causal reasoning. A notable example is Seeing is not Believing, which models spurious correlations as the result of unobserved confounders and introduces robustness through state perturbations and counterfactual transition generation [2]. While such approaches have demonstrated promising empirical performance, they typically intervene at the level of the environment or training data and require additional modeling assumptions.

At the same time, research in representation learning suggests a complementary perspective. Studies in self-supervised learning have shown that spurious correlations can also appear in the structure of learned representations, where a small number of dominant directions capture most of the variance in high-dimensional feature spaces. Spectral regularization has been proposed as a way to counteract this effect by encouraging a more balanced eigenspectrum, leading to improved robustness and transfer performance [3]. Motivated by this representation-level view, this thesis investigates whether applying spectral regularization to the high-dimensional internal representations learned by Soft Actor-Critic can improve robustness to spurious correlations, without relying on explicit state-level interventions.

1.1. Background

This section introduces the core concepts that motivate and support this thesis. It covers the notion of spurious correlations in reinforcement learning, the Soft Actor-Critic algorithm used in this work, and spectral regularization methods from representation learning that inspire the proposed approach.

1.1.1. Spurious Correlations in Reinforcement Learning

In reinforcement learning, spurious correlations occur when parts of the observed state are correlated due to latent or unobserved factors rather than causal relationships. Agents may exploit such correlations during training because they provide an easy path to reward maximization. However, when these correlations change at test time, the learned policies can fail. This issue has been studied in recent work that frames spurious correlations in RL through unobserved confounders and investigates robustness under distribution shifts [1].

1.1.2. Soft Actor-Critic (SAC)

Soft Actor-Critic (SAC) is an off-policy actor–critic algorithm for continuous control that combines reward maximization with an entropy-based exploration objective. Due to its stability and widespread use, SAC provides a suitable baseline for studying robustness and representation-level effects in reinforcement learning.

1.1.3. Spectral Regularization in Representation Learning

Spectral regularization methods aim to control how variance is distributed across learned feature dimensions. In self-supervised learning, such methods have been shown to reduce reliance on dominant feature directions and improve robustness by encouraging more balanced representations [2]. Whether similar representation-level effects exist in reinforcement learning, where representations are learned implicitly, remains an open question.

1.2. Relevance

Robustness to spurious correlations is essential for applying reinforcement learning in real-world settings. Existing approaches often intervene at the level of data or environment dynamics, which can increase complexity and require additional assumptions [1].

This thesis explores a complementary direction by focusing on the structure of learned representations. Its relevance lies in:

- **Simplicity:** The thesis investigates a representation-level alternative to explicit state perturbation or causal modeling.
- **Insight:** By analyzing spectral properties of RL representations, this work contributes to a better understanding of how spurious correlations affect policy learning.
- **Connection:** The thesis connects ideas from self-supervised representation learning to robustness in reinforcement learning.

1.3. Research Question

The central research question of this thesis is:

Can spectral regularization of learned representations improve the robustness of Soft Actor-Critic policies in reinforcement learning environments that exhibit spurious correlations?

To address this question, the following sub-questions are investigated:

- Q1: Do SAC representations show spectral imbalance under spurious correlations?
- Q2: How does spectral regularization affect the representation structure?
- Q3: Does this lead to better performance under distribution shift?

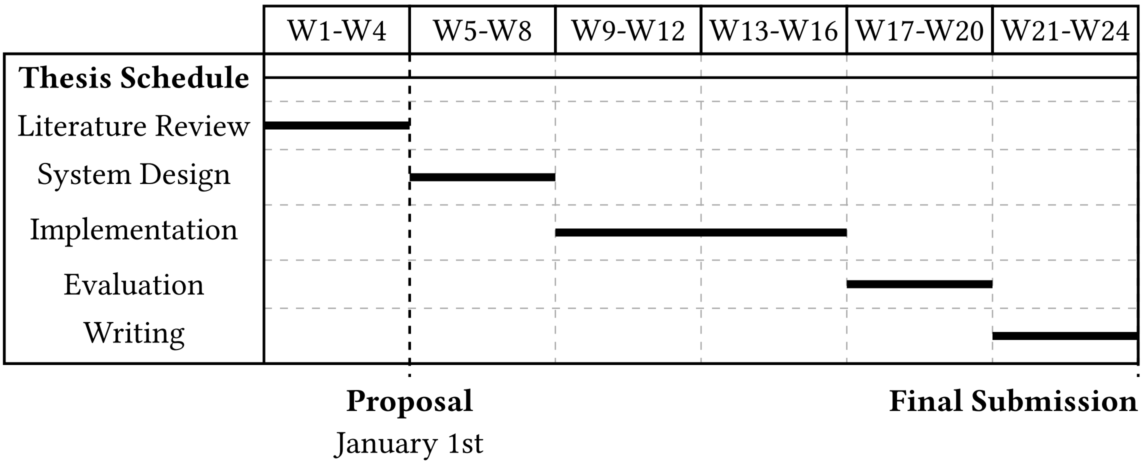
1.4. Approach

1.5. Methods

1.6. Evaluation

2. Outline

3. Schedule



4. Supervisor

Who is your supervisor? (Naghmeh ghanoni) Have you discussed your proposal with them? What do you still need to clear up?

References

[1] Wenhao Ding and Laixi Shi and Yuejie Chi and Ding Zhao, “Seeing is not Believing: Robust Reinforcement Learning against Spurious Correlation.”

[2] Naghmeh Ghanoni and Marius Kloft, “Spectral Regularization for Self-Supervised Representation Learning.”