Spectral Robustness Analysis of Deep Imitation Learning

Ezgi Korkmaz
ezgikorkmazmail@gmail.com

Abstract

Deep reinforcement learning algorithms enabled learning functioning policies in MDPs with complex state representations. Following these advancements deep reinforcement learning polices have been deployed in many diverse settings. However, a line of research argued that in certain settings building a reward function can be more complicated than learning it. Hence, several studies proposed different methods to learn a reward function by observing trajectories of a functioning policy (i.e. inverse reinforcement learning). Following this line of research several studies proposed to directly learn a functioning policy by solely observing trajectories of an expert (i.e. imitation learning). In this paper, we propose a novel method to analyze the spectral robustness of deep neural policies. We conduct several experiments in the Arcade Learning Environment, and demonstrate that simple vanilla trained deep reinforcement learning policies are more robust than deep imitation learning policies. We believe that our method provides a comprehensive analysis on the policy robustness and can help in understanding the fundamental properties of different training techniques.

1 Introduction

The capabilities achieved via interacting with a given environment solely based on observations and receiving rewards upon taking actions in high-dimensional state observation MDPs gained substantial acceleration with the recent advancements in deep reinforcement learning research Mnih et al. (2016). Currently, from robotics objectives to solving complex games, several different fields from pharmaceuticals to finance benefited from the advancements achieved in deep neural policies.

While deep reinforcement learning progressed towards solving more complicated tasks, a line of research focused on the questions arising from learning functioning policies without the presence of a reward function. In this line of research, initially some proposed to learn the reward function via observing the trajectories of a functioning policies (i.e. inverse reinforcement learning). Others proposed to learn a functioning policy via observing expert demonstrations (i.e. imitation learning).

The adversarial non-robustness of deep neural networks has been extensively discussed in the machine learning community starting from the pioneering study of Goodfellow et al. (2015). Following this, adversarial vulnerabilities have also been discussed in deep reinforcement learning policies Huang et al. (2017); Korkmaz (2022a, 2021e). While some of these studies focus on optimization of the adversarial directions Korkmaz (2020), in others these directions are used as a tool to highlight the vulnerability types and variations across different training techniques Korkmaz (2021e). While the robustness of deep reinforcement learning policies has been extensively studied we focus on the robustness of the state-of-the-art deep inverse reinforcement learning and deep imitation learning policies and in our paper we want to answer the following questions:

- *How does learning from expert demonstrations affect the robustness of deep imitation learning policies?*

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

- *How can we analyze and quantify the robustness of deep imitation learning policies in the frequency spectrum?*

Hence, to answer these questions in this paper we focus on analyzing the spectral properties of deep inverse reinforcement learning policies, and make the following contributions:

- We propose a novel method to analyze deep inverse reinforcement learning and deep imitation learning policy robustness in the frequency spectrum.
- We conduct experiments in the Arcade Learning Environment (ALE) and we compare the frequency vulnerabilities of the state-of-the-art imitation learning policy to the vanilla deep reinforcement learning algorithm for high dimensional state representation environments.
- Our method reveals the spectral contrast between the vanilla deep reinforcement learning policies and the state-of-the-art imitation learning policies.

2 Background and Related Work

2.1 Reinforcement Learning

A Markov Decision Process (MDP) is represented as a tuple $\langle S, A, \mathcal{P}, r, \gamma, \tau_0 \rangle$ of a set of states S, a set of actions A, transition probability distribution $\mathcal{P}(s_{t+1}|s_t, a_t)$, and a reward function $r : S \times A \to \mathbb{R}$, discount factor γ, and initial state distribution τ_0. The objective in reinforcement learning is to learn a policy that will maximize the expected discounted cumulative rewards obtained by the policy $\pi : S \to \mathcal{P}(A)$. This objective can be achieved via Q-learning that essentially learns a Q function $Q : S \times A \to \mathbb{R}$ that will assign values to each state-action (s, a) pair to reveal what would be the expected cumulative discounted rewards obtained if the action a is taken in state s. The Q-function is learnt via iterative Bellman update

$$Q(s_t, a_t) = r(s_t, a_t, s_{t+1}) + \gamma \sum_{s_t} \mathcal{P}(s_{t+1}|s_t, a_t) \max_a Q(s_{t+1}, a). \tag{1}$$

Upon the construction of the state-action value function the policy would execute the action that maximizes the state-action value function.

$$a^* = \operatorname*{argmax}_{a \in A} Q(s, a) \quad \text{and} \quad V(s) = \max_a Q(s, a) \tag{2}$$

2.2 Deep Inverse Reinforcement Learning

For a given setting where the reward function is not present, inverse reinforcement learning was proposed to learn a reward function by observing the trajectories of a functioning policy. The initial study that proposed inverse reinforcement learning achieves this objective via linear programming Ng & Russell (2000).

$$\begin{aligned} \text{maximize} \sum_{s \in S_\rho} \min_{a \in A} \{ p(\mathbb{E}_{s' \sim \mathcal{P}(s, a_1|\cdot)} V^\pi(s') - \mathbb{E}_{s' \sim \mathcal{P}(s, a|\cdot)} V^\pi(s')) \} \\ \text{s.t. } |\alpha_i| \leq 1 \,, \; i = 1, 2, \ldots, d \end{aligned} \tag{3}$$

While some studies focused on learning the reward function itself others focused on directly learning a policy from demonstrations Kostrikov et al. (2020). Quite recently, Garg et al. (2021) focused on learning a state-action value function via solely observing the trajectories of a functioning policy (inverse Q-learning). Inverse Q-learning is the first algorithm that can achieve the ability to learn policies in high dimensional state observations. Most importantly, the authors of this study argue that once the state-action value function is learnt, the reward function can be reconstructed from this information. Furthermore, note that the inverse Q-learning algorithm can learn a functioning policy and a reward function simultaneously; hence, throughout the paper the inverse Q-learning algorithm will be referred to as an imitation learning and inverse reinforcement learning algorithm interchangeably.

Algorithm 1 SRA: Spectral Robustness Analysis

```
Input: Actions $a \in \mathcal{A}$, states $s \in \mathcal{S}$, policy $\pi(s,a)$, $\delta$ blocked frequencies, $d$ dimension of the state
Output: Performance Drop
for $\delta = 0$ to $d/2$ do
    for $s = s_0$ to $s_T$ do
        $\mathcal{F}_s(i,j) = \frac{1}{\mathcal{MN}} \sum_{m=0}^{\mathcal{M}-1} \sum_{n=0}^{\mathcal{N}-1} s(m,n)e^{-j2\pi(um/\mathcal{M}+vn/\mathcal{N})}$
        $\mathcal{F}_s[\delta, -\delta : \delta] = \mathcal{F}_s[-\delta, -\delta : \delta] = \mathcal{F}_s[-\delta : \delta, \delta] = \mathcal{F}_s[-\delta : \delta, -\delta] = 0$
        $s_{\mathcal{F}}(m,n) = \sum_{i=0}^{\mathcal{M}-1} \sum_{j=0}^{\mathcal{N}-1} \mathcal{F}(i,j)e^{j2\pi(um/\mathcal{M}+vn/\mathcal{N})}$
        $\pi(s_{\mathcal{F}}, a) = \text{softmax}(Q(s_{\mathcal{F}}, a))$
    end for
end for
Return: Impact $\mathcal{I}$
```

2.3 Robustness in Reinforcement Learning

The adversarial vulnerabilities of deep reinforcement learning policies were initially discussed in Huang et al. (2017). This study essentially introduces fast gradient sign method produced adversarial perturbations Goodfellow et al. (2015) in to the observation system of the deep reinforcement learning policy. In this line of research some studies tried to further optimize adversarial directions Korkmaz (2020), while others focused on contrasting the differences between adversarial and natural directions Korkmaz (2023, 2021d). Targeting the adversarial vulnerabilities of deep reinforcement learning policies, some studies proposed to train with an adversary to gain robustness to adversarial directions Gleave et al. (2020); Pinto et al. (2017). However, some recent studies discussed the problems with adversarial training starting from learning a different set of non-robust features Korkmaz (2021e)[1], to losing generalization capacity Korkmaz (2023) and learning inaccurate and inconsistent state-action value functions Korkmaz (2021c). Some further argued that there are underlying shared adversarial directions that are learnt by many different policies independent from the algorithm, and that are shared across MDPs Korkmaz (2022a).

3 Spectral Robustness Analysis of Deep Imitation Learning and Deep Inverse Reinforcement Learning

In this section we will describe the main proposal of the paper. In particular, Spectral Robustness Analysis (SRA) is based on systematically blocking frequencies and analyzing the effects of these frequencies on the deep reinforcement learning policy performance. In particular, for a state $s \in S$ the discrete Fourier transform of the state s is

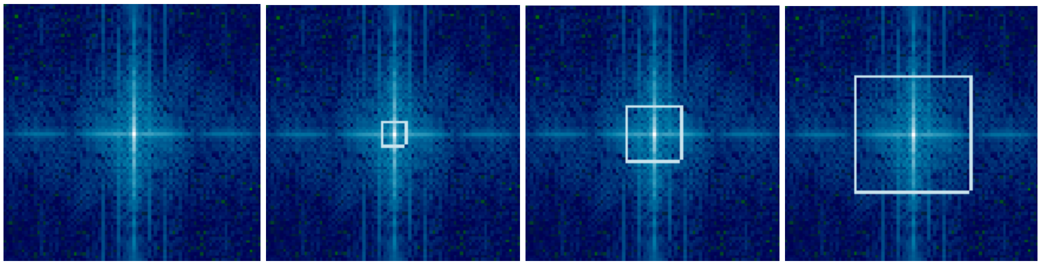
$$\mathcal{F}_s(i,j) = \frac{1}{\mathcal{MN}} \sum_{m=0}^{\mathcal{M}-1} \sum_{n=0}^{\mathcal{N}-1} s(m,n)e^{-j2\pi(um/\mathcal{M}+vn/\mathcal{N})} \tag{4}$$

Upon the blocking the δ-frequencies the discrete Fourier transform is inverted and the observation of the deep neural policy consists of $s_{\mathcal{F}}$

$$s_{\mathcal{F}}(m,n) = \sum_{i=0}^{\mathcal{M}-1} \sum_{j=0}^{\mathcal{N}-1} \mathcal{F}_s(i,j)e^{j2\pi(um/\mathcal{M}+vn/\mathcal{N})}. \tag{5}$$

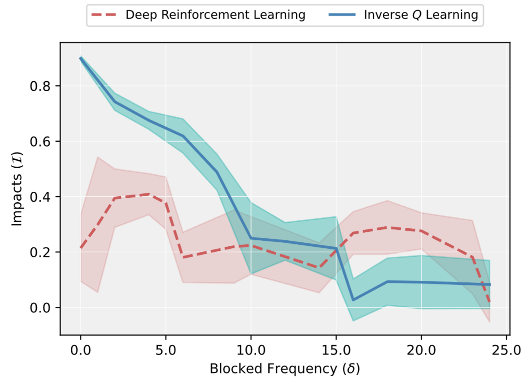
Algorithm 1 provides the pseudocode for the Spectral Robustness Analysis (SRA) algorithm.

[1] The adversarially trained deep reinforcement learning policies are shown to be more vulnerable towards lower frequency adversarial directions Korkmaz (2021b). Furthermore, it is demonstrated that the non-robust features learnt by the adversarially trained policies form different patterns than the vanilla trained deep reinforcement learning policies. These non-robust features are not eliminated by the adversarially training techniques, but rather shifted Korkmaz (2021a), Korkmaz (2021e). Some recent studies further utilized the techniques proposed in Korkmaz (2021e) to demonstrate the non-robust features learnt by deep inverse reinforcement learning algorithms Korkmaz (2022b).



Baseline $\delta = 5$ $\delta = 10$ $\delta = 20$

Figure 1: Left: Spectral Robustness Analysis (SRA) with variations of blocked frequencies δ.



Spectral Robustness Analysis

Figure 2: Spectral Robustness Analysis (SRA) results for the deep reinforcement learning policy and the state-of-the-art deep inverse reinforcement learning policy in the Seaquest game.

4 Experimental Analysis

Our experiments are conducted in the Arcade Learning Environment Bellemare et al. (2013). The deep reinforcement learning policy is trained via double-Q learning Hasselt et al. (2016). The experiments are conducted with 10 random runs. We included standard error of the mean in all of the results presented in the paper. The impact on the policy performance is measured by

$$\mathcal{I} = \frac{\text{Score}_{\text{baseline}} - \text{Score}_{\mathcal{F}_s}}{\text{Score}_{\text{baseline}}}. \tag{6}$$

Figure 1 reports the steps of the spectral robustness analysis (SRA) with variations of blocked frequencies. Figure 2 provides results on the spectral robustness analysis of the deep reinforcement learning policy and the deep inverse reinforcement learning policy as we block the δ-frequencies that have been described in Algorithm 1. The results reported in Figure 2 demonstrate that the vanilla trained deep reinforcement learning policies are more robust than the policies trained via deep inverse reinforcement learning. In particular, there is a high increase in the sensitivities towards lower frequencies for the deep inverse reinforcement learning policy.

5 Conclusion

This paper aims to answer the following questions: (i) How can we analyze the deep neural policies that are trained in MDPs with complex state representations in the spectral domain?, (ii) Is there a fundamental difference between learning via exploration vs learning via observing functioning policies in terms of their robustness?, and (iii) How can we unveil the fundamental robustness differences between deep reinforcement learning policies and deep inverse reinforcement learning policies? To be able to address these questions we propose a novel method that provides a comprehensive analysis of the spectral robustness of deep neural policies. We conduct extensive experiments in the Arcade

Learning Environment and demonstrate that the deep reinforcement learning policies are more robust than the deep inverse reinforcement learning policies.

References

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research.*, pp. 253–279, 2013.

Garg, D., Chakraborty, S., Cundy, C., Song, J., and Ermon, S. Iq-learn: Inverse soft-q learning for imitation. *Neural Information Processing Systems (NeurIPS) [Spotlight Presentation]*, 2021.

Gleave, A., Dennis, M., Wild, C., Neel, K., Levine, S., and Russell, S. Adversarial policies: Attacking deep reinforcement learning. *International Conference on Learning Representations ICLR*, 2020.

Goodfellow, I., Shelens, J., and Szegedy, C. Explaning and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

Hasselt, H. v., Guez, A., and Silver, D. Deep reinforcement learning with double q-learning. *In Thirtieth AAAI conference on artificial intelligence*, 2016.

Huang, S., Papernot, N., Goodfellow, Ian an Duan, Y., and Abbeel, P. Adversarial attacks on neural network policies. *Workshop Track of the 5th International Conference on Learning Representations*, 2017.

Korkmaz, E. Nesterov momentum adversarial perturbations in the deep reinforcement learning domain. *International Conference on Machine Learning, ICML 2020, Inductive Biases, Invariances and Generalization in Reinforcement Learning Workshop.*, 2020.

Korkmaz, E. Non-robust feature mapping in deep reinforcement learning. *International Conference on Machine Learning, ICML Adversarial Machine Learning Workshop*, 2021a.

Korkmaz, E. Adversarially trained neural policies in fourier domain. *International Conference on Machine Learning, ICML Adversarial Machine Learning Workshop*, 2021b.

Korkmaz, E. Inaccuracy of state-action value function for non-optimal actions in adversarially trained deep neural policies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2323–2327, June 2021c.

Korkmaz, E. Adversarial training blocks generalization in neural policies. *International Conference on Learning Representation (ICLR) Robust and Reliable Machine Learning in the Real World Workshop*, 2021d.

Korkmaz, E. Investigating vulnerabilities of deep neural policies. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021*, volume 161 of *Proceedings of Machine Learning Research (PMLR)*, pp. 1661–1670. AUAI Press, 2021e.

Korkmaz, E. Deep reinforcement learning policies learn shared adversarial features across MDPs. *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7229–7238, 2022a.

Korkmaz, E. The robustness of inverse reinforcement learning. *International Conference on Machine Learning (ICML) Artificial Intelligence for Agent Based Modelling Workshop*, 2022b.

Korkmaz, E. Adversarial robust deep reinforcement learning requires redefining robustness. *AAAI Conference on Artificial Intelligence*, 2023.

Kostrikov, I., Nachum, O., and Tompson, J. Imitation learning via off-policy distribution matching. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Mnih, V., Puigdomenech, A. B., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. *In International Conference on Machine Learning*, pp. 1928–1937, 2016.

Ng, A. Y. and Russell, S. J. Algorithms for inverse reinforcement learning. In Langley, P. (ed.), *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pp. 663–670, 2000.

Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A. Robust adversarial reinforcement learning. *International Conference on Learning Representations ICLR*, 2017.