H_∞ Model-free Reinforcement Learning with Robust Stability Guarantee

Minghao Han
Department of Control Science and Engineering
Harbin Institute of Technology
mhhan@hit.edu.cn

Lixian Zhang
Department of Control Science and Engineering
Harbin Institute of Technology
lixianzhang@hit.edu.cn

Yuan Tian
Chair of Intelligent Maintenance Systems
ETH Zurich
yutian@ethz.ch

Jun Wang
Department of Computer Science
University College London
jun.wang@cs.ucl.ac.uk

Wei Pan*
Department of Cognitive Robotics
Delft University of Technology
wei.pan@tudelft.nl

Abstract

Reinforcement learning is showing great potentials in robotics applications, including autonomous driving, robot manipulation and locomotion. However, with complex uncertainties in the real-world environment, it is difficult to guarantee the successful generalization and sim-to-real transfer of learned policies theoretically. In this paper, we introduce and extend the idea of robust stability and H_∞ control to design policies with both stability and robustness guarantee. Specifically, a sample-based approach for analyzing the Lyapunov stability and performance robustness of a learning-based control system is proposed. Based on the theoretical results, a maximum entropy algorithm is developed for searching Lyapunov function and designing a policy with provable robust stability guarantee. Without any specific domain knowledge, our method can find a policy that is robust to various uncertainties and generalizes well to different test environments. In our experiments, we show that our method achieves better robustness to both large impulsive disturbances and parametric variations in the environment than the state-of-art results in both robust and generic RL, as well as classic control. Anonymous code is available to reproduce the experimental results at https://github.com/RobustStabilityGuaranteeRL/RobustStabilityGuaranteeRL.

1 Introduction

As a powerful learning control paradigm, reinforcement learning is extremely suitable for finding the optimal policy in tasks where the dynamics are either unknown or affected by severe uncertainty [Buşoniu et al., 2018]. Its combination with the deep neural network has boosted applications in autonomous driving [Sallab et al., 2017], complicated robot locomotion Hwangbo et al. [2019], and skilful games like Atari [Mnih et al., 2015] and Go [Silver et al., 2017]. However, overparameterized policies are prone to become overfitted to the specific training environment, limiting its generalization

$*$ for correspondence.

NeurIPS 2019 Workshop on Robot Learning: Control and Interaction in the Real World, Vancouver, Canada

to the various scenarios Pinto et al. [2017]. Additionally, RL agents trained in simulation, though cheap to obtain, but likely suffer from the *reality gap* problem Koos et al. [2010] when transferred from virtual to the real world. To overcome these drawbacks, various efforts are made to enhance the robustness of the policy Jakobi et al. [1995], Tobin et al. [2017], Mordatch et al. [2015], since a robust policy has a greater chance of successful generalization and transfer.

Contribution In this paper, we propose a unified framework of designing policies with both stability and robust performance guarantee against the various uncertainties in the environment. Without any specific domain knowledge, our method is able to find policy that is robust to large exogenous disturbances and generalizes well to different test environment. First, a novel model-free method for analyzing the Lyapunov stability and H_∞ performance of the closed-loop system is developed. Based on the theoretical results, we propose the Robust Lyapunov-based Actor-Critic (RLAC) algorithm to simultaneously find the Lyapunov function and policy that can guarantee the robust stability of the closed-loop system. We evaluate RLAC on a simulated cartpole in the OpenAI gym [Brockman et al., 2016] environment and show that our approach is robust to: **i) Large impulsive disturbance:** The trained agent is able to recover when disturbed by adversary impulses 4-6 times of the maximum control input, while other baselines fail almost surely. **ii) Parametric Uncertainty:** The learned policy generalizes better than the baselines to different test environment settings (e.g. different mass and structural values).

2 Preliminaries

2.1 Markov Decision Process and Reinforcement Learning

A Markov decision process (MDP) is a tuple, (S, A, c, P, ρ), where S is the set of states, A is the set of actions, $c(s,a) \in [0,\infty)$ is the cost function, $P(s'|s,a)$ is the transition probability function, and $\rho(s)$ is the starting state distribution. $\pi(a|s)$ is a stationary policy denoting the probability of selecting action a in state s. In addition, the cost function under stationary policy is defined as $c_\pi(s) \doteq \mathbb{E}_{a\sim\pi} c(s,a)$.

In this paper, we divide the state s into two vectors, s^1 and s^2, where s^1 is composed of elements of s that are aimed at tracking the reference signal r while s^2 contains the rest. The cost function c is defined as $\|s^1 - r\|$, where $\|\cdot\|$ denotes the Euclidean norm.

2.2 Robust Control Against Environment Uncertainty

Definition 1. *The stochastic system is said to be stable in mean cost if* $\lim_{t\to\infty} \mathbb{E}_{s_t} c_\pi(s_t) = 0$ *holds for any initial condition* $s_0 \in \{s_0 | c_\pi(s_0) \le b\}$. *If* b *is arbitrarily large then the stochastic system is globally stable in mean cost.*

To address the performance of the agent in the presence of uncertainty, the following definition is needed.

Definition 2. *The system is said to be stable in mean cost with an* l_2 *gain less or equal than* η*, if the system is MSS when* $w = 0$*, and the following holds for all* $w \in l_2[0, +\infty)$*,*

$$\sum_{t=0}^{\infty} \mathbb{E}_{s_t} c_\pi(s_t) \le \sum_{t=0}^{\infty} \mathbb{E}_{s_t} \eta^2 \|w(s_t)\|_2 \tag{1}$$

where $\eta \in \mathbb{R}_+$. $w(s_t)$ *is the uncertainty, which is composed of both environmental disturbance and modelling error.*

The robust performance guarantee (1) holds for all w, is equivalent to guaranteeing the inequality for the worst case induced by w, i.e.,

$$\sup_{w} \sum_{t=0}^{\infty} \mathbb{E}_{s_t} c_\pi(s_t) - \eta^2 \|w(s_t)\|_2 \le 0 \tag{2}$$

3 Main Results

3.1 Lyapunov-based H_∞ Learning Control

In this section, we propose the main assumptions and a new theorem.

Assumption 1. *The stationary distribution of state* $q_\pi(s) \triangleq \lim_{t\to\infty} P(s|\rho,\pi,t)$ *exists.*

Assumption 2. *There exists a positive constant* b *such that* $\rho(s) > 0, \forall s \in \{s|c_\pi(s) \leq b\}$*.*

The core theoretical results on analyzing the stability and robust performance of the closed-loop system with the help of Lyapunov function and sampled data are presented. The Lyapunov function is a class of continuously differentiable semi-positive definite functions $L : \mathcal{S} \to \mathbb{R}_+$. The general idea of exploiting Lyapunov function is to ensure that the derivative of Lyapunov function along the state trajectory is semi-negative definite so that the state goes in the direction of decreasing the value of Lyapunov function and eventually converges to the set or point where the value is zero.

Theorem 1. *If there exists a continuous differentiable function* $L : \mathcal{S} \to \mathbb{R}_+$ *and positive constants* α_1*,* α_2*,* α_3, η, k_1, k_2 *such that*

$$\alpha_1 c_\pi(s) \leq L(s) \leq \alpha_2 c_\pi(s) \tag{3}$$

$$\mathbb{E}_{\beta(s)}(\mathbb{E}_{s'\sim P_\pi} L(s') - L(s)) < \mathbb{E}_{\beta(s)}[\eta\|w(s)\| - \alpha_3 c_\pi(s)] \tag{4}$$

holds for all $\{\rho|\mathbb{E}_{s_0\sim\rho} c_\pi(s_0) \leq k_1\}$ *and* $\{w|\|w\| \leq k_2\}$*.* $\beta_\pi(s) \triangleq \lim_{N\to\infty}\frac{1}{N}\sum_{t=0}^{N} P(s_t = s|\rho,\pi,t)$ *is the sampling distribution. Then the system is mean square stable and has* l_2 *gain no greater than* η/α_3*. If the above holds for* $\forall k_1, k_2 \in \mathbb{R}_+$*, then the system is globally mean square stable with finite* l_2 *gain.*

Proof of Theorem 1 is given in Appendix A.

3.2 Learning the Adversarial Disturber

In our setting, in addition to the control policy π, a disturber policy $\mu(w|s)$ is introduced to actively select the worst disturbance for a given state. More specifically, the adversarial disturber seeks to find the disturbance input w over which the system has the greatest l_2 gain, i.e. maximizing the following cost function,

$$\max_{\theta_\mu} J(\mu) = \mathbb{E}_{\beta(s),\mu(w|s)}(c_\pi(s) - \eta^2\|w\|) \tag{5}$$

where θ_μ is the parameter of the disturber policy μ.

4 Algorithm

In this section, based on the theoretical results in Section 3, we propose an actor-critic style algorithm with robust stability guarantee (RLAC).

In this algorithm, we include a critic Lyapunov function L_c to provide the policy gradient, which satisfies $L(s) = \mathbb{E}_{a\sim\pi} L_c(s,a)$. Through Lagrangian method, the objective function for π is obtained as follow,

$$\begin{aligned} J(\pi) &= \mathbb{E}_{(s,a,w,c,s')\sim\mathcal{D}}\left[\beta\log(\pi(f_{\theta_\pi}(\epsilon,s)|s)) + \lambda\Delta L(s,a,w,c,s')\right] \\ \Delta L(s,a,w,c,s') &= \left(L_c(s', f_{\theta_\pi}(\epsilon,s')) - L_c(s,a) + \alpha_3 c - \eta^2\|w\|_2\right) \end{aligned} \tag{6}$$

where π is parameterized by a neural network f_{θ_π} and ϵ is an input vector consisted of Gaussian noise. In the above objective, ν and λ are the positive Lagrangian multipliers, of which the values are adjusted automatically. The gradient of (6) with respect to the policy parameter θ_π is approximated by

$$\nabla_\theta J(\pi) = \mathbb{E}_\mathcal{D}\left[\nabla_\theta\nu\log(\pi_\theta(a|s)) + \nabla_a\nu\log(\pi_\theta(a|s))\nabla_\theta f_\theta(\epsilon,s) + \lambda\nabla_{a'}L_c(s',a')\nabla_\theta f_\theta(\epsilon,s')\right] \tag{7}$$

The Lyapunov function is updated through minimizing the following objective

$$J(L_c) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\frac{1}{2}(L_c(s,a) - L_{\text{target}}(s,a))^2\right] \tag{8}$$

We use the sum of cost over a finite time horizon N as the Lyapunov candidate, i.e.

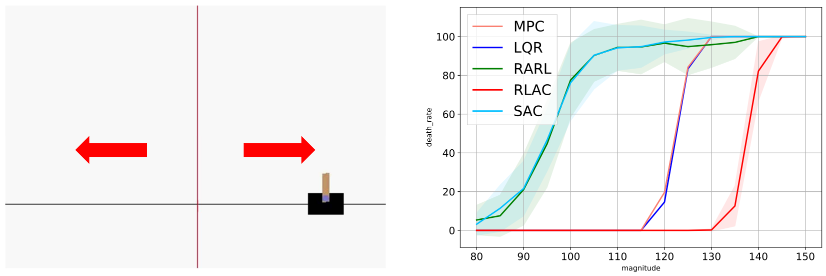
$$L_{\text{target}}(s,a) = \sum_{t=t_0}^{N} c(s_t,a_t) \tag{9}$$

which has long been exploited as the Lyapunov function in establishing the stability criteria for model predictive control (MPC) [Mayne et al., 2000]. The pseudo-code of RLAC is presented in Algorithm 1.

5 Experimental Results

In this section, we evaluate the robustness of RLAC against i) large impulsive disturbances; ii) parametric uncertainty. Setup of the experiment is referred to Appendix C.

5.1 Robustness to Impulsive Disturbances



(a) Visualization of the disturbance (b) Death rate of the cartpole

Figure 1: (a) Direction of the disturbance applied on the cartpole, which is dependent on the relative position of cart concerning the origin. (b) The death rate of agents trained by RLAC, RARL, SAC, MPC and LQR in the presence of impulsive force F with different magnitudes. The trained policies are initialized by 10 random seeds. The policies with different initializations are evaluated equally for 500 episodes. The line indicates the average death rate of these policies and the shadowed region shows the 1-SD confidence interval.

We evaluate the robustness of the agents trained by RLAC and baselines against unseen exogenous disturbance. We measure the robust performance via the death rate, i.e., the probability of pole falling after impulsive disturbance. As observed in the figure, RLAC gives the most robust policy against the impulsive force. It maintains the lowest death rate throughout the experiment, far more superior than SAC and RARL. Moreover, RLAC performs even better than MPC and LQR, which possess the full information of the model and are available.

5.2 Robustness to Parametric Uncertainty

In this experiment, we evaluate the trained policies in environments with different parameter settings. In the training environment, the parameter *length of pole* $l = 0.5$ and *mass of cart* $m_c = 1$, while during evaluation l and m_c are selected in a 2-D grid with $l \in [0.2, 2.0]$ and $m_c \in [0.4, 2.0]$.

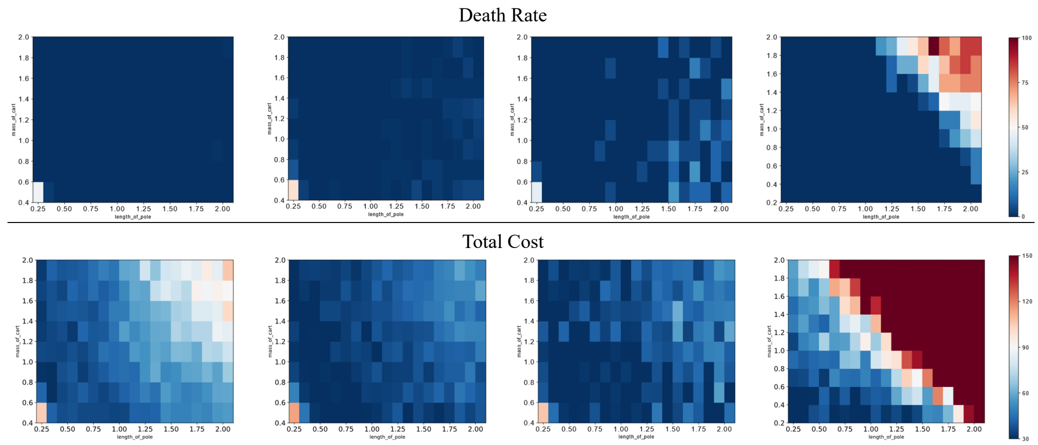
As shown in the heat maps in Figure 2, RLAC achieves the lowest death rate (zero for the majority of the parameter settings) and obtains reasonable total cost (lower than 100). The total cost of RLAC is slightly higher than SAC and RARL since the agents hardly die and sustain longer episodes. Compared to SAC, RARL achieves lower death rate and comparable total cost performance. LQR performs well in the region where parameters are close to the nominal model but deteriorates soon as parameters vary. All of the model-free methods outperform LQR in terms of robustness to parametric uncertainty, except for the case of low l and m_c (left bottom of the grid). This is potentially due to the overparameterized policy does not generalize well to the model where dynamic is more sensitive to input than the one used for training.

References

Lucian Buşoniu, Tim de Bruin, Domagoj Tolić, Jens Kober, and Ivana Palunko. Reinforcement learning for control: Performance, stability, and deep approximators. *Annual Reviews in Control*, 2018.

Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017(19):70–76, 2017.

Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26):eaau5872, 2019.



(a) RLAC (b) RARL (c) SAC (d) LQR

Figure 2: Death rate and total costs of agents trained by RLAC, RARL, SAC and LQR in the presence of different parametric uncertainty which are *unseen* during training and different from dynamic randomization. l (X-axis) and m_c (Y-axis) vary with the step size of 0.1 and 0.2 respectively. At each point of the parameter grid, the results are averaged between the agents with different initializations over 100 episodes.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2817–2826. JMLR. org, 2017.

Sylvain Koos, Jean-Baptiste Mouret, and Stéphane Doncieux. Crossing the reality gap in evolutionary robotics by promoting transferable controllers. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, pages 119–126. ACM, 2010.

Nick Jakobi, Phil Husbands, and Inman Harvey. Noise and the reality gap: The use of simulation in evolutionary robotics. In *European Conference on Artificial Life*, pages 704–720. Springer, 1995.

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 *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30. IEEE, 2017.

Igor Mordatch, Kendall Lowrey, and Emanuel Todorov. Ensemble-cio: Full-body dynamic motion planning that transfers to physical humanoids. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5307–5314. IEEE, 2015.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

David Q Mayne, James B Rawlings, Christopher V Rao, and Pierre OM Scokaert. Constrained model predictive control: Stability and optimality. *Automatica*, 36(6):789–814, 2000.

Halsey Lawrence Royden. *Real analysis*. Krishna Prakashan Media, 1968.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.