End-to-end safe reinforcement learning using differentiable simulation

Background

Reinforcement learning (RL) has demonstrated remarkable success in solving complex control tasks, such as robotic manipulation and autonomous driving. However, many real-world control scenarios impose safety constraints that vanilla RL algorithms struggle to satisfy. Guaranteeing constraint satisfaction in RL is an active field of research. Most safeguarding approaches, such as predictive safety filters, rely on a (potentially simplified) analytical model of the system under control [2]. However, this model is treated as a black box from the perspective of the RL agent. The central idea of this thesis is to incorporate the model knowledge used in safeguarding into the training process. By using a differentiable simulation as well as a fully differentiable safeguarding approach, we can obtain the gradient of the reward w.r.t. the agent's actions. This promises to improve sample efficiency and speed up training, which is advantageous since the safeguarding is computationally expensive. We aim to combine previous work on policy learning with fully differentiable simulation [3] with a differentiable action projection safety shield that can be integrated into the RL agent's policy. Your goal is to evaluate whether this approach can improve sample efficiency and wall clock time during training compared to model-free RL algorithms with non-differentiable safety layers.

Tasks

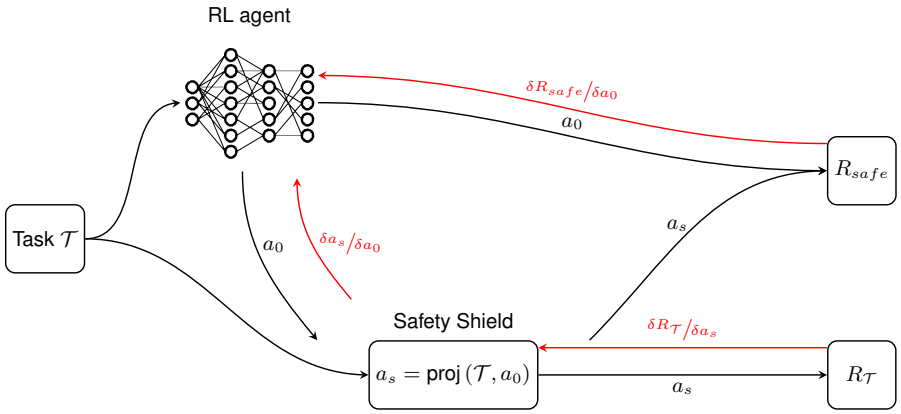

Figure 1: The RL agent is trained on two rewards: A task reward $R_{\mathcal{T}}$ and a reward R_{safe} that punishes unsafe actions.

The goal of this thesis is to evaluate whether a fully differentiable simulation and safety shield can improve sample efficiency and runtime over model-free state-of-the-art RL algorithms with non-differentiable safety shields. This includes the following steps:

- Set up a differentiable simulation of the CartPole environment
- Replicate the results from [3] for the CartPole swing-up task without safeguarding
- Set up a non-differentiable safety shield for PPO for the CartPole
- Integrate a differentiable safety shield into SHAC, e.g., using CVXPyLayers [1]
- Benchmark PPO with safety layer against SHAC with differentiable safety layer
- Repeat this process for other environments

What We Offer

- State-of-the-art research in machine learning,
- Weekly meetings with your advisors,
- Flexible start and schedule for the thesis project, and
- Thesis topics that will be tailored to your interest

Technical University of Munich

School of Computation, Information and Technology

Chair of Robotics, Artificial Intelligence and Real-time Systems

Supervisor:
Prof. Dr.-Ing. Matthias Althoff

Advisor:
Hannah Markgraf, Jonathan Külz

Type:
MA

Research area:
Safe reinforcement learning, optimization, differentiable simulation, implicit learning

Programming language:
Python

Required skills:
Solid understanding of reinforcement learning, practical experience with PyTorch

Language:
English

For more information please contact us:

Phone:

E-Mail:
hannah.markgraf@tum.de

Website:
www.ce.cit.tum.de/cps/home/

References

[1] A. Agrawal, B. Amos, S. Barratt, S. Boyd, S. Diamond, and Z. Kolter. Differentiable convex optimization layers. In *Advances in Neural Information Processing Systems*, 2019.

[2] Hanna Krasowski, Jakob Thumm, Marlon Müller, Lukas Schäfer, Xiao Wang, and Matthias Althoff. Provably safe reinforcement learning: Conceptual analysis, survey, and benchmarking. *Transactions on Machine Learning Research*, 2023.

[3] Jie Xu, Viktor Makoviychuk, Yashraj Narang, Fabio Ramos, Wojciech Matusik, Animesh Garg, and Miles Macklin. Accelerated policy learning with parallel differentiable simulation. In *International Conference on Learning Representations*, 2022.

Technical University of Munich

School of Computation, Information and Technology

Chair of Robotics, Artificial Intelligence and Real-time Systems