Learning-based Model Predictive Control for Safe Reinforcement Learning

Torsten Koller
University of Freiburg
Freiburg, Germany
kollert@informatik.uni-freiburg.de

Felix Berkenkamp, Matteo Turchetta,
Andreas Krause
ETH Zurich
Zurich, Switzerland
{befelix,matteotu,krausea}@inf.ethz.ch

Joschka Boedecker
University of Freiburg
Freiburg, Germany
jboedeck@informatik.uni-freiburg.de

***Abstract*—Reinforcement learning has been successfully used to solve difficult tasks in complex unknown environments. However, these methods typically do not provide any safety guarantees, which prevents their use in safety-critical, real-world applications. In this paper, we attempt to bridge the gap between learning-based techniques that are scalable and highly autonomous but often unsafe and robust control techniques, which have a solid theoretical foundation that guarantees safety but often require extensive expert knowledge to identify the system and estimate disturbance sets. We combine a provably safe learning-based MPC scheme that allows for input-dependent uncertainties with techniques from model-based RL to solve tasks with only limited prior knowledge. We evaluate the resulting algorithm to solve a reinforcement learning task in a simulated cart-pole dynamical system with safety constraints.**

I. Introduction

In model-based reinforcement learning (RL,[8]), we aim to learn the dynamics of an unknown system from data and use it to derive a policy that optimizes the long-term behavior of the system. In order to be successful, these methods need to *explore* regions of the state-space that are unknown to collect observations that improve the model. In real-world safety-critical systems, we require exploratory actions to be safe to perform by satisfying state and control constraints at all times. In contrast, current approaches often use exploration strategies that could lead to unsafe behavior and, hence, have limited applicability to real-world systems [11]. Learning-based approaches that guarantee safety, on the other hand, often rely on conservative assumptions, such as fixed, *state-independent* disturbance sets, or require extensive computation [1, 7].
In this paper, we extend a previously proposed learning-based MPC algorithm with safety guarantees to solve RL tasks [6]. We design an objective function that allows the MPC scheme to find controls that maximize the performance of the system while remaining safe at all times. This is illustrated in Figure 1.

II. Problem Statement

We consider a nonlinear, continuously differentiable, discrete-time dynamical system

$$x_{t+1} = f(x_t, u_t) = \underbrace{h(x_t, u_t)}_{\text{prior model}} + \underbrace{g(x_t, u_t)}_{\text{unknown error}}, \qquad (1)$$

where $x_t \in \mathbb{R}^p$ is the state and $u_t \in \mathbb{R}^q$ is the control input to the system at time step $t \in \mathbb{N}$. We encode all the prior knowledge that we have about our system in a prior model h, while the model errors and disturbances are given by the *a priori* unknown function g. We want to *learn* the unknown model-error g by actively collecting observations of our system. To this end, we use a statistical model with mean $\mu_n(x_t, u_t)$ and corresponding *input-dependent* uncertainty estimates $\sigma_n(x_t, u_t)$, where n is the number of observations we collected so far. We now make the assumption that our statistical model reliably estimates the uncertainty about our system. That is, we assume that for every $\delta > 0$ there exists a $\beta > 0$ such that with probability at least $1-\delta$, jointly $\forall n \in N$ we have for all $1 \leq j \leq p$, $z \in \mathcal{X} \times \mathcal{U}$ that

$$|\mu_{n-1,j}(z) - g_j(z)| \leq \beta \cdot \sigma_{n-1,j}(z), \qquad (2)$$

where both, the set of admissible states $\mathcal{X} \subset \mathbb{R}^p$ and controls $\mathcal{U} \subset \mathbb{R}^q$ are assumed to be polytopic. Under certain technical assumptions on g and the choice of our statistical model, a β can be computed in closed-form that guarantees (2) [2]. In a more practical sense, we can regard the scaling factor β as a parameter that controls our confidence in the statistical model.

Further, we rely on the assumption that, typically around the origin, we often have a good understanding of the behavior of a system. This often allows us to find a controller that satisfies our constraints in a bounded region of the state space. Hence, we assume the existence of a safety controller π_{safe} and a polytopic safety region $\mathcal{X}_{\text{safe}}$ such that for arbitrary $k \in \mathbb{N}$, we have $x_k \in \mathcal{X}_{\text{safe}} \Rightarrow f(x_t, \pi_{\text{safe}}(x_t)) \in \mathcal{X}, \quad \forall t \geq k$ [6, 10]. As the safety controller can only be used in this possibly small region, we want to design a controller π that can be used

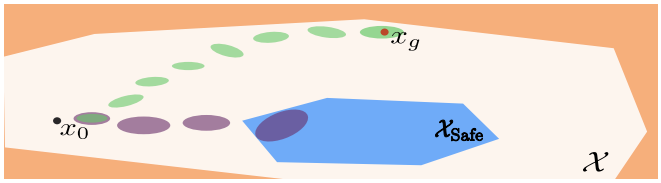

Fig. 1. Simultaneous planning of a performance trajectory (green ellipsoids) using an approximate uncertainty propagation technique and a safety trajectory (purple ellipsoids) based on a robust multi-step ahead prediction technique [6]. While the performance trajectory optimizes the expected long-term utility (e.g. distance to goal state x_g) of applying a control input at x_0, the safety trajectory guarantees that the same input could return the system to the safe set $\mathcal{X}_{\text{Safe}}$ without violating the safety constraints $\mathcal{X}$ and control constraints $\mathcal{U}$.

outside of $\mathcal{X}_{\text{safe}}$. Given the limited knowledge of our system, we can only require *probabilistic* worst-case safety guarantees,

$$\Pr\left[\forall t \in \mathbb{N} : f(x_t, \pi(x_t)) \in \mathcal{X},\, \pi(x_t) \in \mathcal{U}\right] \geq 1 - \delta, \quad (3)$$

i.e. a high-probability safety guarantee over the whole operation time. Given a mission objective to reach and remain at certain set point x_g of our system, we want to find a controller that solves this RL task while guaranteeing that (3) holds.

III. Safe reinforcement learning

We design a MPC scheme that can solve a given RL task under safety constraints. In order to do so, we need to reliably propagate the uncertainty of our system, design appropriate constraints and find a suitable objective function.

A. Uncertainty propagation and constraints

Given control inputs $\mathbf{u} = \{u_0, .., u_{T-1}\}$ and an ellipsoidal state estimate $\mathcal{R}_0$ known to contain the true system state, an ellipsoidal uncertainty propagation technique of the form

$$\mathcal{R}_{t+1} = \tilde{m}(\mathcal{R}_t, u_t), t = 0, .., T-1, \quad (4)$$

is derived in [6] using our statistical model and Lipschitz information of the prior model h and the statistical model (μ_n, σ_n). Independently of T and how often this technique is applied, the system will be contained in the corresponding sequences of ellipsoids jointly with high probability. The system constraints along the trajectory, $R_i \subset \mathcal{X}$, $t = 1, .., T$, $\mathbf{u} \subset \mathcal{U}$, can now be verified analytically [9].

B. Reinforcement learning objective and MPC scheme

We require an objective function that jointly encourages exploration and finding a good control strategy based on our current statistical model. Since (4) provides a worst-case outer approximation of our system, it does not reflect the probabilistic nature of our statistical model. This prevents us from accurately estimating the *expected* performance of a sequence of control inputs. We hence employ a second, probabilistic uncertainty propagation technique using our statistical model, $s_{t+1} = m_{perf}(s_t, v_t) \sim \mathcal{N}(m_t, S_t)$, $t = 0, .., H-1$, with $s_0 = p_0$, the center of the ellipsoid R_0, providing us with a *peformance trajectory* $s_0, .., s_T$ of Gaussian distributed states under inputs $\mathbf{v} = \{v_0, .., v_{H-1}\}$. We then compute the approximated expected long-term cost of applying the controls in closed-form using the *saturating cost* function c_{sat}, i.e.

$$J^{\mathbf{v}}(s_0) = \sum_{t=0}^{T} \mathbb{E}[c_{sat}(s_t, x_g)], \quad (5)$$

providing an efficient trade-off between exploration and exploitation [3]. We can now formulate the MPC problem

$$\begin{aligned}
\underset{\mathbf{u}\subset\mathcal{U},\mathbf{v}\subset\mathcal{U}}{\text{minimize}} \quad & J^{\mathbf{v}}(s_0) && (6a)\\
\text{subject to} \quad & \mathcal{R}_{t+1} = \tilde{m}(\mathcal{R}_t, u_t),\, t = 0, .., T-1 && (6b)\\
& \mathcal{R}_t \subset \mathcal{X}, u_t \in \mathcal{U},\, t = 1, .., T-1 && (6c)\\
& \mathcal{R}_T \subset \mathcal{X}_{\text{safe}} && (6d)\\
& v_0 = u_0, u_0 \in \mathcal{U}, && (6e)
\end{aligned}$$

where (6c), (6d) guarantees that there exists a collision-free return path to $\mathcal{X}_{\text{safe}}$. With $v_0 = u_0$, we ensure that the first control input both optimizes the long-term cost (5) and can recover the system to the safe set if needed. In case of infeasibility of (6) in the next time step, we execute the safety controls $\mathbf{u}$ open-loop until there is a new feasible solution or we switch to π_{safe} after $T-1$ consecutively infeasible solutions. This guarantees that the system remains safe with high probability throughout operation [6, Theorem 2].

TABLE I

Ratio of failed rollouts of all episodes and cumulative final episode cost (averaged over successful rollouts) for varying lengths $H \in \{5, 10, 15\}$ of the performance trajectory. Lower is better for both benchmarks.

	Cautious MPC ($T = 0$)		SafeMPC ($T = 2$)	
H	Failures[%]	C_{ep}	Failures[%]	C_{ep}
5	87.5	281.88	0.0	> 1000
10	10.4	164.26	0.0	661.04
15	18.7	153.16	0.0	163.42

IV. Experiments

We evaluate the proposed MPC scheme to solve a RL task in an underactuated cart-pole system. The state of the system is given by the position x_{cart} and velocity $\dot{x}_{cart}$ of the cart as well as the pendulum angle θ and the corresponding velocity $\dot{\theta}$. Initialized in an upright position at $x_{cart} = -2$, the task is to control the system to $x_{cart} = 2.6$. We limit the length of the rail by $x_{cart} \in [-10, 3.0]$ and simulate a floor, i.e. $\theta \in [-90, 90]$. The known part of our system h is given by a linearized and discretized wrongly identified system. We use Gaussian process (GP) regression to estimate the unknown model-error g and the performance trajectory is computed using the uncertainty propagation technique proposed in [4] with varying length $H \in \{5, 10, 15\}$. We compare a cautious MPC setting with $T = 0$ and chance constraints on the performance trajectory, similar to [5], with our proposed algorithm using $T = 2$ and $\beta = 2$. We run our MPC algorithm in an episode setting over 50 time steps, reset the system afterwards and update the GP model with the observed noisy transition and repeat. We report the cumulative episode cost $C_{ep} = \sum_{t=0}^{n_{steps}} 0.1(x_t^{cart} - x_g)^2$ and the percentage of rollouts that violate the constraints after eight episodes. We average over ten repetitions of the experiment. The results show that our approach can safely solve the task, while approaches without explicit safety strategies may violate the safety constraints.

V. Conclusion

We extended a provably safe learning-based MPC algorithm to solve a RL task under safety constraints. By combining the safety features of the learning based MPC framework with techniques from model-based RL, we can guarantee the safety of the system while learning a given task. We experimentally showed that our proposed RL algorithm is capable of learning a task in a simulated cart-pole system without violating safety constraints.

References

[1] Anil Aswani, Humberto Gonzalez, S. Shankar Sastry, and Claire Tomlin. Provably safe and robust learning-based model predictive control. *Automatica*, 49(5):1216–1226, May 2013.

[2] Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. *In Proc. of Neural Information Processing Systems (NIPS)*, May 2017.

[3] Marc Peter Deisenroth and Carl Edward Rasmussen. PILCO: A model-based and data-efficient approach to policy search. In *In Proceedings of the International Conference on Machine Learning*, pages 465–472, 2011.

[4] A. Girard, C. E. Rasmussen, J. Quiñonero-Candela, R. Murray-Smith, Becker, S, S. Thrun, and K. Obermayer. Multiple-step ahead prediction for non linear dynamic systems: A Gaussian Process treatment with propagation of the uncertainty. In *Sixteenth Annual Conference on Neural Information Processing Systems (NIPS 2002)*, pages 529–536. MIT Press, October 2003.

[5] Lukas Hewing and Melanie N. Zeilinger. Cautious model predictive control using Gaussian process regression. *arXiv preprint arXiv:1705.10702*, 2017.

[6] Torsten Koller, Felix Berkenkamp, Matteo Turchetta, and Andreas Krause. Learning-based Model Predictive Control for Safe Exploration. *In Proc. of the IEEE Conference on Decision and Control (CDC)*, December 2018.

[7] Raffaele Soloperto, Matthias A. Müller, Sebastian Trimpe, and Frank Allgöwer. Learning-Based Robust Model Predictive Control with State-Dependent Uncertainty. *IFAC-PapersOnLine*, 51(20):442–447, January 2018.

[8] R. S. Sutton and A. G. Barto. Reinforcement Learning: An Introduction. *IEEE Transactions on Neural Networks*, 9(5):1054–1054, September 1998.

[9] D. H. van Hessem and O. H. Bosgra. Closed-loop stochastic dynamic process optimization under input and state constraints. In *In Proc. of the American Control Conference (ACC)*, volume 3, pages 2023–2028, May 2002.

[10] Kim P. Wabersich and Melanie N. Zeilinger. Safe exploration of nonlinear dynamical systems: A predictive safety filter for reinforcement learning. *arXiv:1812.05506 [cs]*, December 2018.

[11] C. Xie, S. Patil, T. Moldovan, S. Levine, and P. Abbeel. Model-based reinforcement learning with parametrized physical models and optimism-driven exploration. In *In Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 504–511, May 2016.