SFU AI

Safe Reinforcement Learning using Robust Tube-Based Model Predictive Control

Abolfazl Eskandarpour, Hossein Sheikhi Darani, Mohammadhadi Mohandes

Introduction

- RL agents typically do not guarantee safety constraints during learning stages [1]
- Even if MBRL ensure safety, they may suffer from providing a desirable closed-loop performance [2]
- MPC techniques can be used to address safety and constraint satisfaction
- Safe learning is a mutually beneficial cooperation for RL and MPC
- Safely finding an efficient trade-off between exploitation and exploration is tricky [3]
- Stochastic and Tube-Based MPC are computationally burdensome techniques for complex uncertain systems exposed to disturbances [4]
- Wisely decoupling the optimization problem from its constraint satisfaction criteria
- We propose a safety filter by utilizing a Tube-Based MPC for a Model-based RL to generate a safe backup trajectory

Proposed Approach

Problem :

- Safe exploration and exploitation using RL in practice for an uncertain system
- Complexity in solving Stochastic and Robust Tube-based MPC

Solution and Contribution:

- Combining MBRL and MPC in which MPC works as a Safety Filter
- RL solved the optimization problem while the MPC watching constraint violation
- The optimization problem (3) is solved in a short horizon
- Nonlinear Robust MPC-based RL and Tube-based Linear MPC-based RL are designed for the systems
- PILCO algorithm is modified to be utilized for the RL part
- Closed-loop and Open-loop MPC-based RL are provided to investigate the uncertainty
- MPC-based algorithm not only does improve the performance of the system, but also provides a safety exploration and exploitation

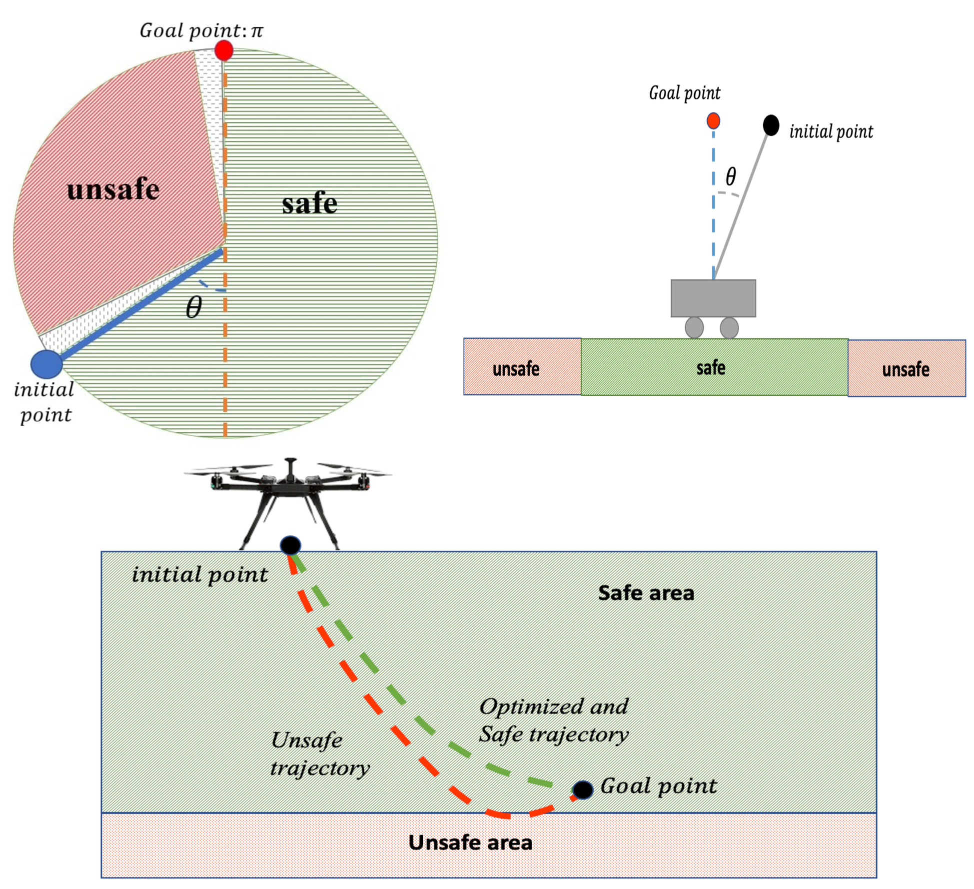

Fig. 1. Safe RL Examples

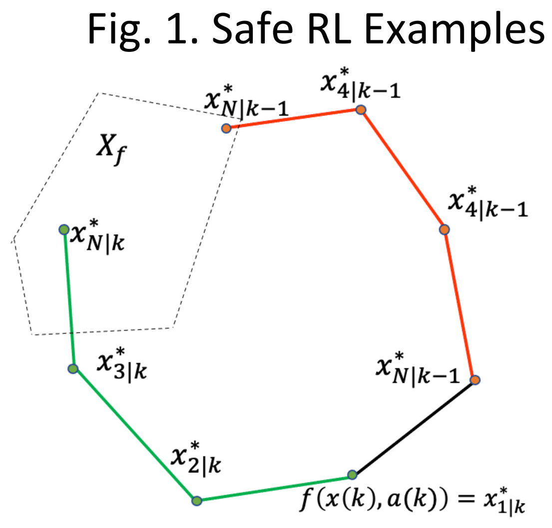

Fig. 2. Backup Trajectory

Proposed System structure

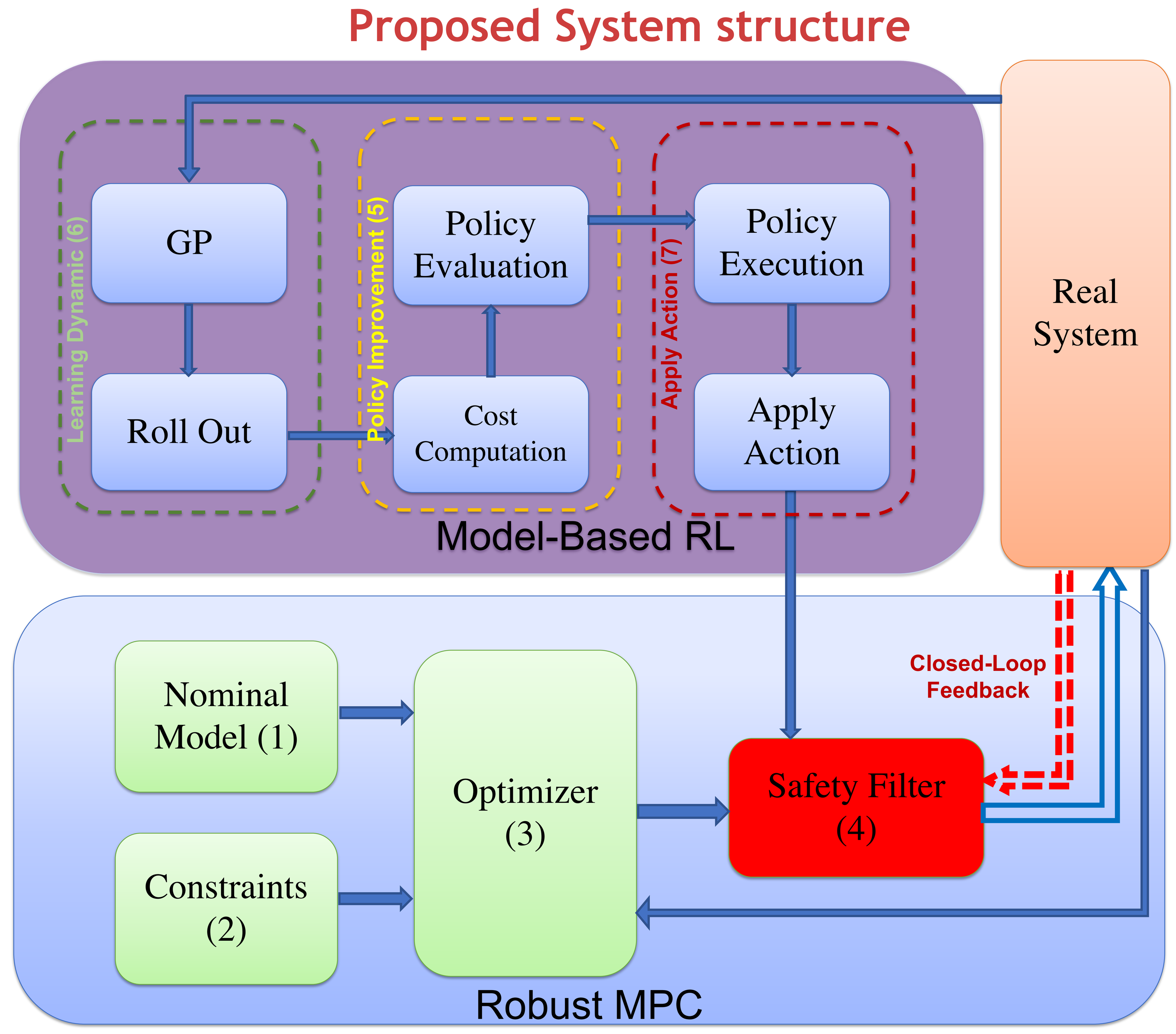

Fig. 3. Safe Reinforcement Learning using Robust MPC

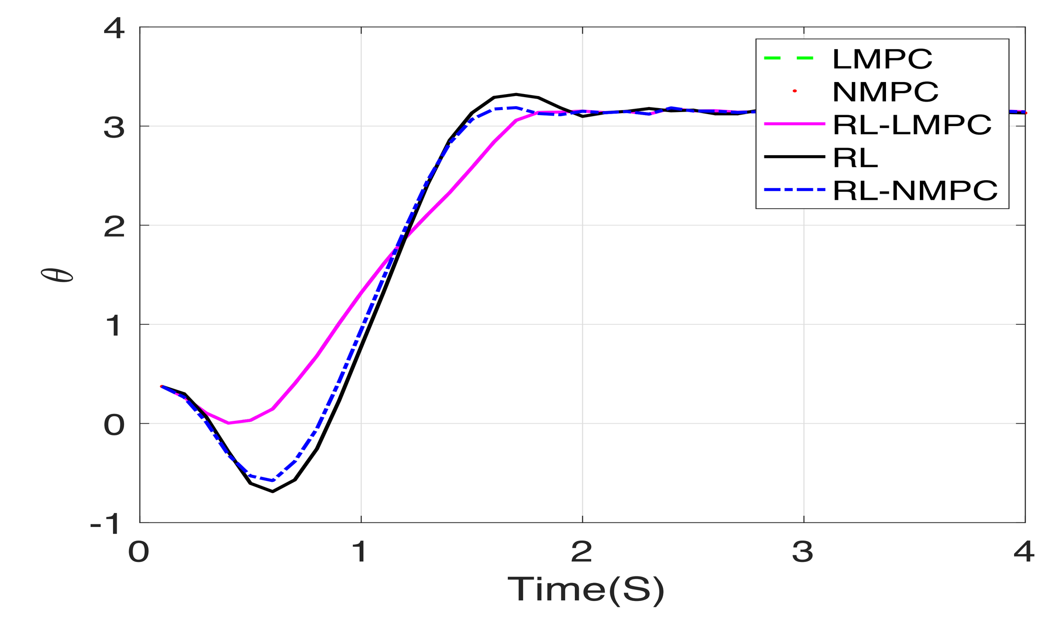

Fig. 4. Performance Comparison

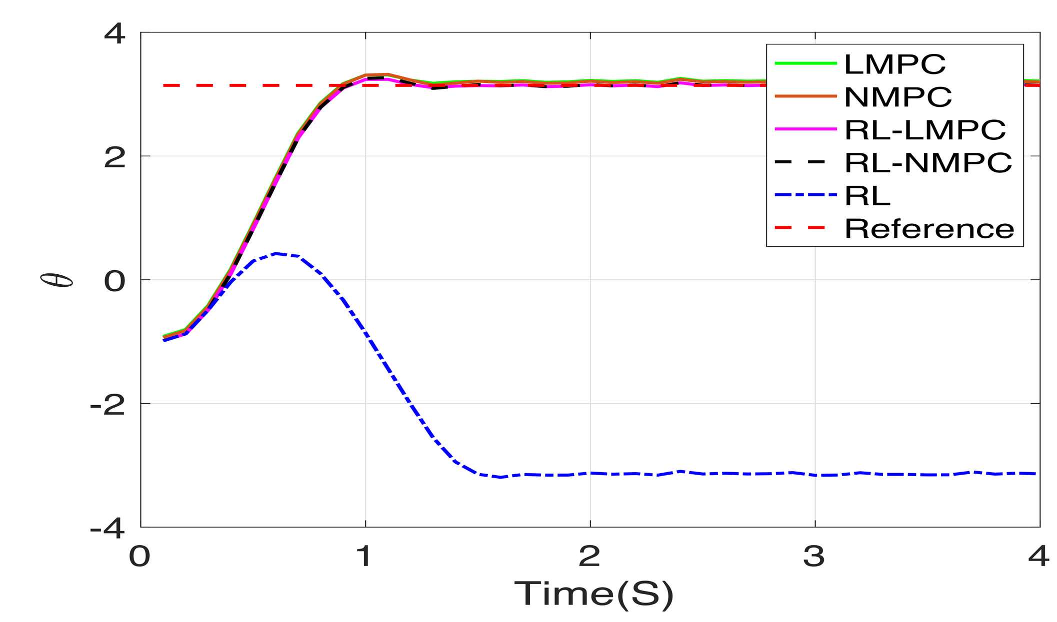

Fig. 5. Safety Comparison

Problem Formulation

1. System Dynamic:

- A discrete-time deterministic dynamical system:

$$x(k+1) = f\big(x(k), u(k)\big) = \underbrace{h\big(x(k), u(k)\big)}_{Nominal\ Model} + \underbrace{g\big(x(k), u(k)\big)}_{Uncertainty} \tag{1}$$

- Subject to the polytopic state and control constraints:

$$\begin{aligned} \mathcal{X} &= \{x \in \mathbb{R}^p | H_x x \le h_x, h_x \in \mathbb{R}^{m_x}\}, \\ \mathcal{U} &= \{u \in \mathbb{R}^p | H_u u \le h_u, h_u \in \mathbb{R}^{m_u}\}. \end{aligned} \tag{2}$$

2. Model Predictive Control:

- Assuming that Terminal set $\mathcal{X}_f \in \mathcal{X}$ is a Robust Positive Invariant set, optimization problem is:

$$\min_{\Delta U} \sum_{i=1}^{N_p} \hat{X}(k+i|k)^T Q\, X(k+i|k) + \sum_{j=1}^{N_p} \Delta\hat{U}(k+j|k)^T R\, \Delta\hat{U}(k+j|k) \tag{3}$$

$$subject\ to: \Delta\hat{U}(k|k) \in \mathcal{U}, \hat{X}(k|k) \in \mathcal{X}, \hat{X}(k+N|k) \in \mathcal{X}_f, System\ (1)$$

- The optimal control signal is sued as a Backup controller when RL actions violate:

$$u_{Backup} = u^*(k|k) \tag{4}$$

3. Model-Based RL

- *Objective* is to find a deterministic *policy* π that minimizes the expected return:

$$J^{\pi}(\theta) = \sum_{t=0}^{T} \mathbb{E}_{x_t}[c(x_t)]\,, \qquad x_0 \sim \mathcal{N}(\mu_0, \Sigma_0) \tag{5}$$

- *Dynamic Model Learning* is implemented as a GP that yields one-step predictions:

$$P(x_t | x_{t-1}, u_{t-1}) = \mathcal{N}(x_t | \mu_t, \Sigma_t) \tag{6}$$

$$\mu_t = x_{t-1} + \mathbb{E}_f[\Delta_t]\,, \Sigma_t = var_f[\Delta_t]$$

- *Policy Evaluation:* evaluating and minimizing J^{π} requires long term predictions of states $p(x_1), \ldots, p(x_T)$ which are obtained by utilizing moment matching algorithm.
- *Policy Improvement:* PILCO derives equations to analytically compute the gradients of the expected return by using gradient based methods, e.g. L-BFGS. Policy is implemented as a nonlinear RBF network, i.e.:

$$\pi(x, \theta) = \sum_{i=1}^{n} w_i \phi_i(x)\ , \phi_i(x) = \exp(-0.5(x-\mu_i)^{\top}\Lambda^{-1}(x-\mu_i)) \tag{7}$$

Result and Comparison

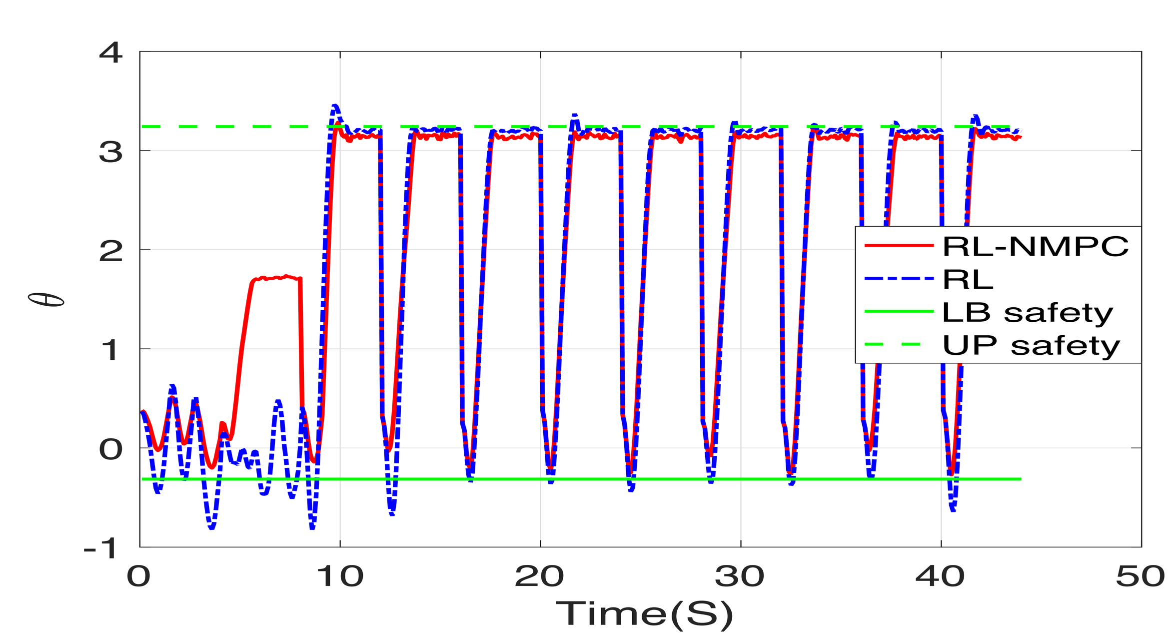

Fig. 6. Learning Episodes Safety Comparison

Conclusion and Future Work

Conclusion

- The proposed algorithms provides a safe exploration and exploitation
- Performance improvement of MPC-based RL compared to merely using of RL and MPC
- Closed-loop MPC-based RL provides a better performance than Open-loop algorithm

Future work

- Using stochastic MPC instead of robust MPC and Linear MPC
- Investigating the consecutive effect of the predicted trajectory using safety filter by changing the applied predicted trajectory length
- Utilize a dynamic matrix weight for the MPC to improve the performance
- Employ on a more complex dynamics like quadrotor
- Nominal Model improvement for MPC while the RL explore the real dynamic model
- Improving the dynamic learning using bootstrapping and incremental neural networks

References

[1] L. Hewing, K. P. Wabersich, M. Menner, and M. N. Zeilinger, “LearningBased Model Predictive Control: Toward Safe Learning in Control,” Annual Review of Control, Robotics, and Autonomous Systems, vol. 3, no. 1, pp. 269–296, 2020. _eprint: https://doi.org/10.1146/annurevcontrol

[2] T. Koller, F. Berkenkamp, M. Turchetta, J. Boedecker, and A. Krause, “Learning-based Model Predictive Control for Safe Exploration and Reinforcement Learning,” *ArXiv190612189 Cs Eess*, Jun. 2019, Accessed: Sep. 30, 2020. [Online]. Available: http://arxiv.org/abs/1906.12189.

[3] M. Zanon and S. Gros, “Safe Reinforcement Learning Using Robust MPC,” *IEEE Trans. Autom. Control*, pp. 1–1, 2020, doi:10.1109/TAC.2020.3024161.

[4] K. P. Wabersich, L. Hewing, A. Carron, and M. N. Zeilinger, “Probabilistic model predictive safety certification for learning-based control,” *ArXiv190610417 Cs Eess*, Jan. 2021, Accessed: Apr. 12, 2021. [Online]. Available: http://arxiv.org/abs/1906.10417.