Robust Adaptive Multi-Step Predictive Shielding (Student Abstract)

Tanmay Ambadkar[1], **Darshan Chudiwal**[1], **Greg Anderson**[2], **Abhinav Verma**[1]

[1]The Pennsylvania State University, 201 Old Main, University Park, PA 16802, USA
[2]Reed College, 3203 SE Woodstock Blvd, Portland, OR 97202, USA
{ambadkar, dsc5635, verma}@psu.edu, grega@reed.edu

Abstract

Ensuring safety in deep reinforcement learning is challenging, as formal methods that provide strong guarantees often fail to scale to complex, high-dimensional systems. We introduce RAMPS, a scalable shielding framework that pairs a general-purpose, learned linear dynamics model with a robust, multi-step Control Barrier Function (CBF) for real-time safety interventions. Experiments show RAMPS significantly reduces safety violations in high-dimensional environments compared to state-of-the-art methods, without sacrificing task performance.

Introduction

Deep reinforcement learning (RL) has achieved remarkable success in solving complex control problems, yet its deployment in safety-critical applications like autonomous vehicles and robotics remains a significant challenge. A core requirement is ensuring safety throughout the entire learning process, a problem known as safe exploration. Model-predictive shielding has emerged as a promising paradigm to address this, where a safety module oversees and corrects an agent's actions to prevent violations .

Existing shielding frameworks, however, present a difficult trade-off. Neural shields learn safety critics from data, but often require extensive experience and fail to prevent violations during early training (Bharadhwaj et al. 2021). Conversely, symbolic shields provide formal guarantees by analyzing an environment model (Anderson et al. 2020) but suffer from a critical limitation: they typically rely on partitioning the state space into a patchwork of local linear models. This approach faces the curse of dimensionality, rendering it computationally intractable for the complex, high-dimensional environments where modern deep RL excels (Anderson, Chaudhuri, and Dillig 2023).

This paper introduces RAMPS (Robust Adaptive Multi-Step Predictive Shielding), a framework that bridges this critical gap. RAMPS is built upon a **single, globally linear representation** of the system's dynamics. This model is general-purpose, and can range from a simple linear regression to a more complex approach like a Deep Koopman Operator operating in a learned feature space (Shi and Meng 2022). This unified representation enables a robust, multi-step Control Barrier Function (CBF) that provides sound, real-time safety interventions, making formal shielding practical for high-dimensional nonlinear systems.

The RAMPS Framework

Our framework is composed of two core components: (1) a general-purpose learner for a global linear dynamics model, and (2) a robust, multi-step safety shield that uses this model to perform real-time interventions.

Learning an Affine Dynamics Model Our framework requires a learned affine dynamics model to approximate the system's evolution. The model must learn parameters A, B, and a constant vector c to represent the next state prediction, $\hat{z}_{k+1}$, in the form: $\hat{z}_{k+1} = Az_k + Bu_k + c$ Here, z_k is the current state and u_k is the control input. After training, we compute a worst-case, one-step error bound ϵ on a hold-out dataset, such that the true next state is bounded by $z_{k+1} = \hat{z}_{k+1} + w_k$, where the model error w_k satisfies $\|w_k\|_\infty \leq \epsilon$.

Multi-Step Robust Shielding The learned linear model enables the use of a Robust Control Barrier Function (CBF)(Ames et al. 2017) to certify safety. Standard one-step discrete CBFs often fail in systems where control inputs have a delayed effect on safety constraints (i.e., a relative degree greater than one). To overcome this, our shield enforces safety over a variable prediction horizon H, drawing on principles from High-Order CBFs (Tan, Cortez, and Dimarogonas 2022). For a polyhedral safe set defined by $\mathcal{C} = \{z \mid p_i^T z + b_i \leq 0, \forall i\}$, we formulate a set of multi-step constraints for each timestep $j \in [1, H]$ and each face i:

$$p_i^T z_j(\mathbf{u}) + b_i \leq \lambda^j (p_i^T z_k + b_i) - \mathcal{E}_j(p_i) \qquad (1)$$

where $z_j(\mathbf{u})$ is the predicted state at step j given the control sequence $\mathbf{u} = (u_k, \ldots, u_{k+H-1})$, $\lambda \in (0, 1]$ is a decay hyperparameter, and $\mathcal{E}_j(p_i) = \sum_{l=0}^{j-1} \epsilon \|p_i^T A^l\|_1$ is a robust tightening term that accounts for accumulated model error. At each timestep, we solve a Quadratic Program (QP) to find a safe control sequence $\mathbf{u}^*$ that minimizes the deviation from the RL agent's proposed action, $\|u_k^* - a_\pi\|_2^2$, subject to the constraints in Eq. 1. Only the first action, u_k^*, is applied to the system.

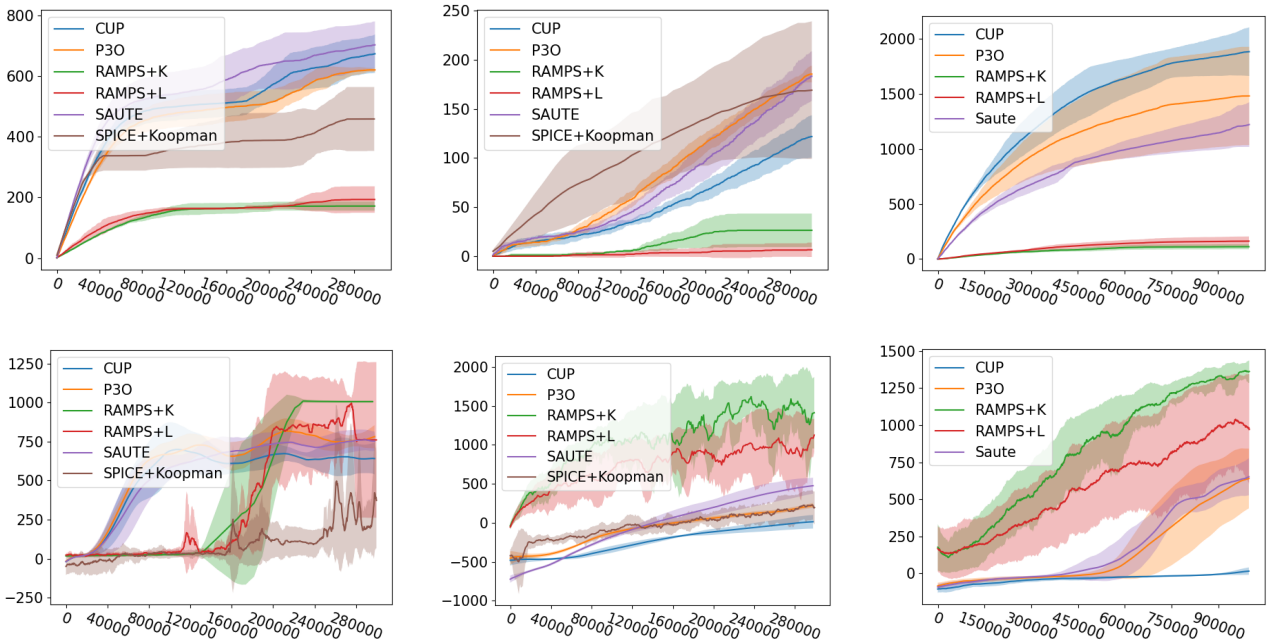

Figure 1: Cumulative safety violations (Top Row) and reward curves (Bottom Row) for SafeHopper (Left), SafeCheetah (Middle) and SafeAnt (Right).

Environment	SauteRL	CUP	P3O	SPICE +K	RAMPS + L	RAMPS + K
SafeHopper	703 ± 78	673 ± 63	620 ± 6	459 ± 105	193 ± 44	$\mathbf{172 \pm 15}$
SafeCheetah	183 ± 25	122 ± 22	185 ± 8	169 ± 70	$\mathbf{7 \pm 7}$	26 ± 17
SafeAnt	1221 ± 203	1883 ± 221	1481 ± 446	Failed	162 ± 42	$\mathbf{111 \pm 23}$

Table 1: Cumulative safety violations during training. Failed indicates that the training process quit or the agent never completed a single safe episode. L stands for our linear regression, and K stands for our learned Koopman Dynamics.

We employ an iterative process where the dynamics model and policy co-evolve. An initial model, calibrated on seed data, shields the RL agent during training. Newly collected experiences are used to retrain the model and recalibrate error bounds, progressively mitigating distribution shift and reducing shield conservatism.

Experimental Evaluation

We evaluate RAMPS on a suite of challenging control tasks to answer two key questions: (1) Does RAMPS reduce safety violations more effectively than state-of-the-art baselines? (2) Does it scale to high-dimensional environments where other formal shielding methods fail?

Results. As shown in Figure 1, RAMPS consistently and significantly outperforms all baselines in terms of safety. In the high-dimensional `SafeHopper`, `SafeCheetah`, and `SafeAnt` environments, RAMPS reduces total safety violations by over 75%, 87%, and 92% respectively, compared to the best-performing CMDP baseline. We have also replaced the piecewise linear model in **SPICE** with a Koopman model to learn accurate dynamics for a fair comparison and to show that the Koopman model is not the main contributing factor. While the baselines continue to accumulate violations throughout training (learning no safe behaviour), the violation curve for RAMPS flattens early on, indicating that the agent learns to operate safely under the shield's protection. This dramatic improvement in safety is achieved while achieving higher task rewards, demonstrating that our shield is minimally invasive and allows the agent to learn a high-performance policy.

Conclusion

We presented RAMPS, a scalable shielding framework that bridges the gap between the strong guarantees of formal methods and the scalability required for modern deep RL. By integrating a general-purpose, learned affine dynamics model with a robust, multi-step Control Barrier Function, RAMPS significantly reduces safety violations without sacrificing performance. Our framework succeeds in high-dimensional domains where prior symbolic methods are computationally intractable, representing a key step towards deploying RL agents in real-world, safety-critical systems.

References

Ames, A. D.; Xu, X.; Grizzle, J. W.; and Tabuada, P. 2017. Control Barrier Function Based Quadratic Programs for Safety Critical Systems. *IEEE Transactions on Automatic Control*, 62(8): 3861–3876.

Anderson, G.; Chaudhuri, S.; and Dillig, I. 2023. Guiding Safe Exploration with Weakest Preconditions. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Anderson, G.; Verma, A.; Dillig, I.; and Chaudhuri, S. 2020. Neurosymbolic Reinforcement Learning with Formally Verified Exploration. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 6172–6183. Curran Associates, Inc.

Bharadhwaj, H.; Kumar, A.; Rhinehart, N.; Levine, S.; Shkurti, F.; and Garg, A. 2021. Conservative Safety Critics for Exploration. In *International Conference on Learning Representations*.

Shi, H.; and Meng, M. Q. H. 2022. Deep Koopman Operator with Control for Nonlinear Systems. arXiv:2202.08004.

Tan, X.; Cortez, W. S.; and Dimarogonas, D. V. 2022. High-Order Barrier Functions: Robustness, Safety, and Performance-Critical Control. *IEEE Transactions on Automatic Control*, 67(6): 3021–3028.