ICRA2026 Late Breaking Results Poster presented at
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Beyond Domain Randomization: Safety Certificates for Reinforcement Learning

Paula Stocco[1], Francesco Micheli[2], Niklas Schmid[1], John Lygeros[1], Efe C. Balta[1,3]
[1]Automatic Control Laboratory, ETH Zürich, Switzerland [2]Amazon, Milan, Italy [3]inspire AG, Zürich, Switzerland

***Abstract*—With the growing acceptance of robotics in daily life there is a growing need for certifiably safe control policies. While simulation provides a safe training environment, policies often fail in sim-to-real transfer. We propose a data-driven certification framework for reinforcement learning based on Pick-to-Learn (P2L), a meta-algorithm that uses data preference ordering to compute probabilistic bounds on the satisfaction of application dependent properties of interest. Our results demonstrate that using P2L maintains high performance while distinguishing between policies that appear similar under domain randomization alone. This work offers a practical method for preparing safe reinforcement learning policies by providing formal safety guarantees prior to hardware deployment.**

I. Introduction

Reinforcement Learning (RL) has demonstrated remarkable success using large, black box neural networks. The behavior of the resulting controller is difficult to anticipate on unseen samples and thus difficult to certify. This work contributes to the urgent question: *How can reinforcement learning be safely deployed in real-world systems?* Two considerations are: learning policies safely, and safety complexity, i.e., how much and what data guarantees safety. Domain randomization is widely adopted to address sim-to-real transfer, but relies on sufficient diversity for the real system to appear as a variation within the training set and thus does not guarantee safety under distributional shift [1]. Rather than precise uncertainty modeling or exhaustive scenario coverage, we extend data-driven certification to RL using Pick-to-Learn (P2L) [2], which jointly synthesizes a controller and certifies its safety risk, using compression based guarantees pre-deployment. While previously applied to various control tasks, to our knowledge this is the first use of P2L for RL. Our results show that this wraparound, data-based certification framework learns policies with competitive performance while providing out-of-sample safety guarantees.

II. Methodology

A. Reinforcement Learning Baseline

We consider a constrained reinforcement learning problem under uncertainty, with system dynamics depending on unknown system parameters. Modeling the system as a Constrained Markov Decision Process (CMDP), an optimal solution would be a safe policy, π, that maximizes expected reward (performance) while maintaining the expected cost below a predefined budget. To handle system uncertainty, domain randomization samples a set of environments from some probability distribution $\{z_i\}_{i=1}^{N} \overset{\text{i.i.d.}}{\sim} \mu$, assumed to be a strong prior for the real system parameters. The CMDP solver then optimizes over simulated environments defined by a vector of parameters to solve for a stationary policy, developed in [3]: $\max_{\pi\in\Pi_s} \mathbb{E}_{z\sim\mu}[J_z(\pi)]$ s.t. $\mathbb{E}_{z\sim\mu}[C_z(\pi)] \leq d$. While domain randomization is scalable and effective, it enforces constraints only in expectation over the environment distributions. A policy may have high expected return and satisfy cost limits while exhibiting unsafe behavior for a subset of environments.

B. Pick to Learn Framework

To address these limitations, we wrap the P2L [2] framework around the baseline policy learning algorithm to provide safety guarantees. We assume access to a dataset of i.i.d. draws of environments, $\mathcal{D} = \{z_1, ..., z_N\}$, with safety defined as a binary function, $\mathbb{1}_{C_z(\pi)>d}$. Statistical risk, defined as $R(\mathcal{D}) = \mathbb{P}_z\big[z : C_z(\pi)\big] \leq d]$, measures safety as the probability that a policy constructed on $\mathcal{D}$ will not violate constraints on new data from the same underlying distribution. P2L constructs a compressed subset, $T \subseteq \mathcal{D}$, during training.

In the P2L framework, the internal learning algorithm is optimized in a loop, where at each iteration, after a set number of policy updates, the policy is evaluated on all environments, the worst-violating samples are added to T, and the policy continues to train on the augmented set. The training loop continues until a desired compression or performance is reached, when it then returns a policy and the cardinality of the training set, $|T|$. With this set, a probabilistic bound for the risk can be calculated with confidence $1-\delta$ as a function of $|T|, \delta, N$ following [2].

III. Results

A. Hardware Experiments

For our hardware experiments, we compare two instances of P2L and domain randomization on the Quanser linear cartpole using PPO-SauteRL for the same seeds. As a baseline, we adopt domain randomization with CMDP solver Proximal Policy Optimization (PPO) with SauteRL following [3]. For the chosen parameters, the simulation closely matches the true hardware dynamics (Figure 3). The cartpole must stay within ± 0.25 m of center (allowing up to 10 violations), with reward a function of upright posture, control

This work was supported as a part of NCCR Automation, a National Centre of Competence in Research, funded by the Swiss National Science Foundation (grant number 51NF40_225155)

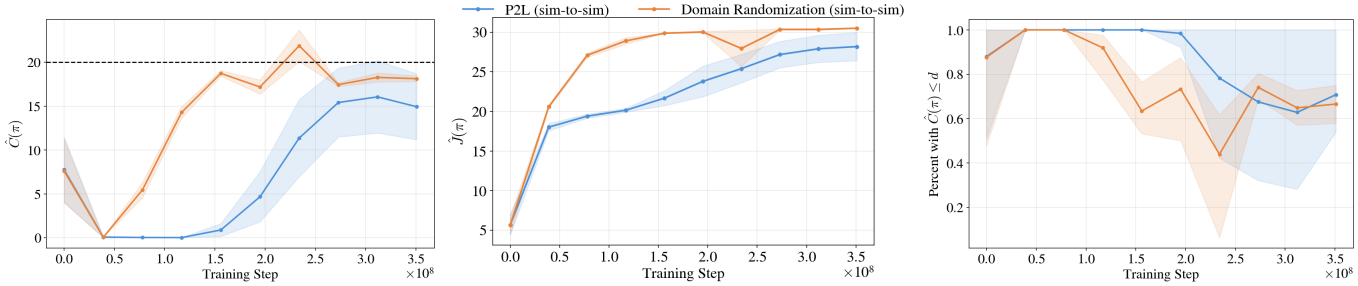

Fig. 1: Simulation results for sim-to-sim transfer on the Unitree Go1 using domain randomization and P2L: cost (left) and reward (middle) plots show the empirical mean and standard error across evaluation environments (constraint ≤ 20). Safety (right) plots show the fraction of safe episodes, along with min–max ranges across seeds.

effort, pole velocity, and cart position. The same training parameters are used for both P2L and domain randomization, with identical P2L-specific settings across P2L runs. For each method, policies labeled A and B correspond to two independent runs differing only in their random seeds. Figure 2 shows five trials on hardware for each policy, each trial running for 3000 steps at 500 Hz. P2L maintains a competitive reward achieving swingup on the real system while preserving safety and providing formal out-of-sample safety guarantees. By contrast, the two randomization policies behaved riskily on the real system, with higher velocity at the bounds. As a consequence, domain randomization B saw two out of five runs violate constraints, one with 34 and one with 28 violations.

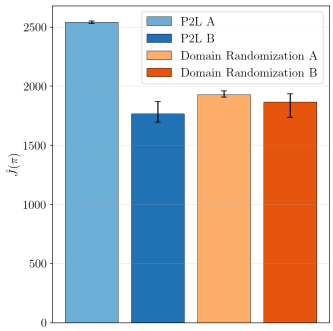

Fig. 2: Real-system reward over five trials (whiskers are standard error); P2L risk bound for both A&B is 0.269 (trained with $66/128$ environments).[1]

B. Sim-to-Sim

We further evaluate P2L on the Unitree Go1 joystick to test its practicality on high dimension systems. We adopt PPO with a Lagrangian penalty as in [3] for the domain randomization benchmark; for full details on parameter ranges, constraints, and rewards, we refer to [3]. Figure 1 shows results for five seeds of domain randomization and P2L, and Table I shows the calculated risk measure for those policies. As with the hardware experiments, P2L provides competitive reward with a safety certification. Notably, domain randomization achieves $\hat{C}$ below the 20.0 constraint, yet has many unsafe runs with high rewards. At a similar level of expected cost, P2L achieves a higher percent of safe runs, up to 100%. The wide variability between maximum and minimum achieved safety shows the effect of different seeds. These influence how the policy is trained and therefore which and when environments are added to training, demonstrating the importance of the data both selected and not in training.

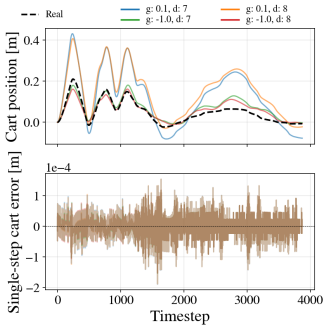

Fig. 3: Simulated vs real cart position outputs using control inputs from a P2L A trial (top) and corresponding single-step error (bottom) (gear ratio: "g", slide dampening: "d")

IV. Conclusion

P2L provides a practical method for training and certifies policies in safety-critical settings, where domain randomization can still fail for small sim-to-real gap without warning. The reward reduction in some configurations of P2L reflects the cost of robustness, as policies are constrained to regions that remain safe under the worst case. This is exacerbated under greater simulator mismatch, where expanding randomized parameter ranges improves coverage but complicates learning. Future work will explore learning tighter domain ensembles thereby also increasing confidence that the certification covers the real system.

TABLE I: P2L certification results on the Unitree Go1

Step	min	max
0	0.126	0.631
78M	0.482	0.661
156M	0.309	0.661
234M	0.143	0.661
312M	0.000	0.661

V. Acknowledgements

In line with the IEEE-RAS AI use policies, Claude was used to assist in the generation of plots and code for this research.

References

[1] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. 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)*, Sep. 2017, pp. 23–30, iSSN: 2153-0866.

[2] D. Paccagnan, M. Campi, and S. Garatti, "The Pick-to-Learn Algorithm: Empowering Compression for Tight Generalization Bounds and Improved Post-training Performance," *Advances in Neural Information Processing Systems*, vol. 36, pp. 18 165–18 185, Dec. 2023.

[3] Y. As, C. Qu, B. Unger, D. Kang, M. van der Hart, L. Shi, S. Coros, A. Wierman, and A. Krause, "SPiDR: A simple approach for zero-shot safety in sim-to-real transfer," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.