Safe Reinforcement Learning via Probabilistic Logic Shields

Wen-Chi Yang[1], Giuseppe Marra[1], Gavin Rens[2] and Luc De Raedt[1,3]

[1]*Leuven AI, KU Leuven, Belgium*
[2]*Stellenbosch University, South Africa*
[3]*Centre for Applied Autonomous Sensor Systems, Örebro University, Sweden*

Keywords
probabilistic shields, statistical relational AI, probabilistic logic programming, safe reinforcement learning

Extended Abstract

Safe Reinforcement learning (Safe RL) aims at learning optimal policies while ensuring that the agent stays safe. A popular solution to Safe RL is shielding, which uses a logical safety specification to prevent an RL agent from taking unsafe actions. Traditional rejection-based shielding techniques provide rigorous safety guarantees, however, they are difficult to integrate with continuous, end-to-end deep RL methods for three reasons.

1. Previous shielding approaches have been limited to symbolic state spaces.
2. Rejection-based shields are deterministic, assuming that any action is either safe or unsafe in a particular state. However, this is an unrealistic assumption as the world is inherently uncertain, and safety is often a matter of degree rather than an absolute concept.
3. Even when given perfect safety information, rejection-based shields may result in a sub-optimal policy.

To address uncertainty, some methods incorporate randomization, e.g. simulating future states in an emulator to estimate risk, using ϵ-greedy exploration that permits unsafe actions, or randomizing the policy based on the current belief state. However, these methods rely on sampling and do not have a clear connection to uncertainty present in the environment. We will exploit such uncertainty by applying probabilistic logic programming principles.

We introduce **probabilistic logic shields (PLS)** as an alternative to deterministic rejection-based shields. PLS is a model-based, neural symbolic, Safe RL technique, encoding logical safety constraints, noisy sensors, the neural policy and their interactions in a probabilistic logic

NeSy 2023, 17th International Workshop on Neural-Symbolic Learning and Reasoning, Certosa di Pontignano, Siena, Italy
wenchi.yang@kuleuven.be (W. Yang); giuseppe.marra@kuleuven.be (G. Marra); gavinrens@sun.ac.za (G. Rens); luc.deradet@kuleuven.be (L. De Raedt)
https://wenchiyang.github.io/ (W. Yang)
0000-0002-5340-3829 (W. Yang); 0000-0001-5940-9562 (G. Marra); 0000-0003-2950-9962 (G. Rens); 000-0002-6860-6303 (L. De Raedt)

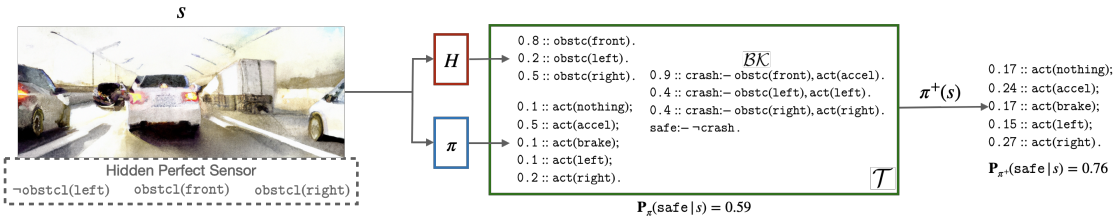

Figure 1: A motivating example of Probabilistic Logic Shields. We encode the interaction between the base policy π, the noisy sensors H and the safety specification $\mathscr{BK}$ using a ProbLog program $\mathscr{T}$. This provides a uniform language to express many aspects of the shielding process. The shielded policy $\pi^+(s)$ decreases the probability of unsafe actions, e.g. acceleration, and increases the likelihood of being safe.

program, as shown in Fig. 1. To ensure safety, we consider a set of noisy sensors H surrounding the agent. These sensor are represented in the probabilistic logic program as neural predicates, which take an image as input and generate probabilities indicating the presence of obstacles or potential dangers. The program can then be automatically compiled into a differentiable structure, allowing for the optimization of a single loss function through the shield, enforcing safety directly in the neural policy. Therefore, PLS can be seamlessly applied to any policy gradient algorithm while still providing the same convergence guarantees. Overall, probabilistic logic shields have the following benefits compared to rejection-based shields.

- *Realistic Safety Function.* PLSs allow for risk control by using probabilistic safety measure.
- *Simpler Model.* PLSs use a simpler safety model that only represents internal safety-related properties, which is less demanding than many model-based approaches that require the full MDP (e.g. [1]).
- *End-to-end Deep RL.* PLSs are differentiable and can be seamlessly applied to any model-free RL agent such as PPO, TRPO, A2C, etc.
- *Convergence.* PLSs in deep RL provide convergence guarantees.

Our experiments show that applying PLS to policy gradients leads to safer and more rewarding policies compared to other state-of-the-art shielding techniques in different discrete and continuous Atari domains. The sources are available on https://github.com/wenchiyang/pls.

This work has been accepted at IJCAI 2023 Main Track. A preprint can be found on https://arxiv.org/abs/2303.03226.

References

[1] N. Hunt, N. Fulton, S. Magliacane, T. N. Hoang, S. Das, A. Solar-Lezama, Verifiably safe exploration for end-to-end reinforcement learning, in: Proceedings of the 24th International Conference on Hybrid Systems: Computation and Control, HSCC '21, 2021. URL: https://doi.org/10.1145/3447928.3456653. doi:10.1145/3447928.3456653.