Towards Probabilistic Safety Guarantees for Model-Free Reinforcement Learning

1st Felippe Schmoeller Roza
Fraunhofer IKS
Munich, Germany

2nd Karsten Roscher
Fraunhofer IKS
Munich, Germany

3rd Stephan Günnemann
Technical University of Munich
Munich, Germany

***Abstract*—Improving safety in model-free Reinforcement Learning is necessary if we expect to deploy such systems in safety-critical scenarios. However, most of the existing constrained Reinforcement Learning methods have no formal guarantees for their constraint satisfaction properties. In this paper, we show the theoretical formulation for a safety layer that encapsulates model epistemic uncertainty over a distribution of constraint model approximations and can provide probabilistic guarantees of constraint satisfaction.**

***Index Terms*—Reinforcement Learning, Safe AI, CMDP**

I. Introduction

With the recent advancements of Deep Learning, Reinforcement Learning (RL) has resurfaced in the field of Artificial Intelligence (AI) and achieved remarkable accomplishments in challenging tasks such as playing complex video games and controlling robotic systems. RL enables agents to learn optimal behaviors through interactions with dynamic environments, without requiring explicit supervision or predefined rules and holds the promise of revolutionizing decision-making and control systems across a wide range of domains. However, to unlock this potential, safety considerations become paramount for instilling confidence and trust in RL-based systems, especially in applications where the impact of failures or incorrect decisions can have catastrophic consequences. Safety in this paper refers to the assurance of reliable and secure operation to protect individuals, the environment, and assets from harm or damage. The need for safety in RL is not only necessary to protect human lives and the surrounding environment but also to fulfill the regulatory and ethical requirements that demand responsible and accountable deployment of AI technologies.

There are different approaches for tackling safety within the RL landscape. *Safe exploration* refers to the process of actively exploring the environment while avoiding actions that pose a risk of significant harm. *Risk awareness* contemplates algorithms that can evaluate potential risks associated with its actions, taking measures to minimize the likelihood and severity of adverse outcomes. *Adversarial robustness* aims at designing agents able to defend against adversarial attacks that could otherwise compromise the agent's performance or integrity. *Online monitoring and verification* involves the integration of real-time monitoring and verification mechanisms to assess the agent's behavior during runtime, allowing for timely intervention or corrective measures to prevent potential harm.

We will rather consider the *constraint satisfaction* paradigm, which focuses on RL systems able to adhere to safety specifications in the form of constraints. Safety constraints can represent physical limits, legal and ethical considerations, operational constraints, resource limitations, etc. More specifically, we consider environments that are modeled as a constrained Markov Decision Process (CMDP) [1], defined as a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \gamma, R, P, c)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}$ is the reward function, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ is the transition probability function, $\gamma \in (0, 1)$ is the discount factor, and $c : \mathcal{S} \mapsto \mathbb{R}$ is the safety cost function. To simplify the notation, $c(s_t)$ will be represented as an immediate cost c_t. The safety cost dynamics is given by the function $f : \mathbb{R} \times \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$, with

$$c_{t+1} = f(c_t, s_t, a_t). \tag{1}$$

The RL goal in a CMDP is to find the policy $\pi : \mathcal{S} \mapsto \mathcal{A}$ that optimizes the long-term cumulative reward while keeping the cost bounded by a safe threshold $h \in \mathbb{R}$, as shown in eq. (2).

$$\max_{\pi \in \Pi} \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R\left(s_t, \pi\left(s_t\right)\right)\right] \text{ s.t. } c_t \leq h,\ \forall t \geq 0. \tag{2}$$

II. CMDPs and RL - Related Work

When considering constrained RL methods that can solve continuous control problems, a popular approach is to use the Lagrangian operator and transform the problem into an unconstrained optimization problem, as done in [2] and [3]. Constrained Policy Optimization (CPO) is another popular algorithm and was the first policy gradient method to solve the CMDP problem [4]. [5] and [6] look at the problem from a different perspective, which consists in integrating a *safety layer* that projects potentially unsafe actions produced by a Deep Neural Network (DNN) into a safe set. The safety layer is composed of linear approximations to the constraint models, with the safe actions calculated by solving a constrained least squares problem.

It is important to notice that, as mentioned in [7], much of the work available in this direction is limited to simplistic simulated tasks, indicating that enabling RL to be applied in real-world constrained systems is not trivial and remains a challenge to overcome. Additionally, [7] defines three levels of

safety in control systems[1] and show how existing constrained-RL algorithms are only able to tackle the most basic level of safety, with stronger safety guarantees only being possible when prior knowledge about the system dynamics is embedded into the controller.

III. Deterministic Safety Layer

Among the previous attempts to solve the CMDP problem using RL, we build on top of the safety filter approaches from [5] and [6] to achieve better constraint satisfaction properties. The underlying constraint dynamics from eq. (1) can be approximated by the first-order Taylor expansion, shown in eq. (3).

$$c_{t+1} \approx c_t + g_\phi(s_t)^\top a_t, \tag{3}$$

where $g_\phi(s_t)$ is a DNN, parametrized by weights ϕ, that approximates the system's constraint dynamics.

A safe action ${a_t}^*$ can be obtained through the optimization problem shown below.

$$\begin{aligned} {a_t}^* = \ & \arg\min_x \frac{1}{2}\|x - a_t\|^2 \\ & \text{s.t. } c_t + g_\phi(s_t)^\top x \leq h. \end{aligned} \tag{4}$$

IV. Probabilistic Safety Layer

One limitation of eq. (4) is its dependency on the accuracy of the approximated dynamics model $g_\phi(s_t)$. To address this, as solution we propose replacing the deterministic constraint model with a distribution over models or trained weights, allowing to achieve robustness in the face of epistemic uncertainty. The new constraint criterion becomes guaranteeing that the predicted safety signal c_{t+1} for a model $\psi(\cdot)$ sampled from the distribution stays below the given threshold with probability p:

$$Prob_{\psi(\cdot)\sim\mathcal{N}_k(\mu,\Sigma)}\left[c_t + \psi(s_t)^\top a_t \leq h\right] \geq p, \tag{5}$$

where $\mathcal{N}_k(\mu, \Sigma)$ represents the multivariate normal distribution of a k-dimensional random vector, the same dimension as the action vector a_t, parameterized by the mean vector $\mu \in \mathbb{R}^k$ and the covariance matrix $\Sigma \in \mathbb{R}^{k\times k}$. Since the random vector is linearly independent, Σ is a diagonal matrix, i.e., $\Sigma = \begin{bmatrix} \sigma_1 & ... & 0 \\ ... & ... & ... \\ 0 & ... & \sigma_k \end{bmatrix}^{k\times k}$.

The dot product $\psi(s_t)^\top a_t = \sum_{i=1}^k \psi_i a_i$, which is a sum of normal distributions, can be substituted by an univariate random variable z, resulting in the following equivalent distribution[2]:

$$Prob_{z\sim\mathcal{N}\left(\sum_{i=0}^k \mu_i a_i, \sum_{i=0}^k \sigma_i^2 a_i^2\right)}\left[c_t + z \leq h\right] \geq p. \tag{6}$$

[1] *Safety Level I* comprises systems that encourage constraint satisfaction with no formal guarantee, *Safety Level II* considers algorithms that satisfy constraints with probability p, and *Safety Level III* represents controllers able to guarantee hard constraint satisfaction.

[2] consider that the sum of n independent random variables $Y = c_1X_1 + ...+c_nX_n$ with means $\mu_1...\mu_n$ and variances $\sigma_1^2...\sigma_n^2$ is a normal distribution with mean $c_1\mu_1 + ... + c_n\mu_n$ and variance $c_1^2\sigma_1^2 + ... + c_n^2\sigma_n^2$.

Now converting eq. (6) to a standard normal distribution:

$$Prob_{z\sim\mathcal{N}(0,1)}\left[c_t + z\sqrt{\sum_{i=0}^k \sigma_i^2 a_i^2} + \sum_{i=0}^k \mu_i a_i \leq h\right] \geq p. \tag{7}$$

This probability can be calculated with the standard normal cumulative distribution function (CDF), $\Phi(\cdot)$. Rearranging the terms in eq. (7) results in

$$\Phi\left(\frac{h - c_t - \sum_{i=0}^k \mu_i a_i}{\sqrt{\sum_{i=0}^k \sigma_i^2 a_i^2}}\right) \geq p. \tag{8}$$

Finally, after applying the inverse CDF operator and with the appropriate manipulation, the probabilistic constraint criterion is obtained:

$$c_t + \Phi^{-1}(p)\sqrt{\sum_{i=0}^k \sigma_i^2 a_i^2} + \sum_{i=0}^k \mu_i a_i \leq h. \tag{9}$$

V. Conclusion

The lack of safety guarantees prevents existing RL systems from becoming a viable alternative for controlling complex control systems. Existing safe RL approaches primarily focus on encouraging safe policies but lack robust evidence for building a solid safety case. In this work, we introduced a novel safety layer formulation able to solve CMDPs with probability p that, in our view, can help to overcome this limitation by providing probabilistic guarantees to model-free constrained-RL.

The theoretical findings presented in this paper must be backed by empirical results obtained through experimentation. The outlined next steps involve: (i) consider the benefits and limitations of different approaches to obtain distributions over constraint models, (ii) test the proposed method's performance in existing simulation benchmarks, and (iii) compare it to existing constrained-RL methods.

References

[1] E. Altman, *Constrained Markov decision processes*, vol. 7. CRC press, 1999.

[2] Y. Kadota, M. Kurano, and M. Yasuda, "Discounted markov decision processes with utility constraints," *Computers & Mathematics with Applications*, vol. 51, no. 2, pp. 279–284, 2006.

[3] G. Kahn, A. Villaflor, V. Pong, P. Abbeel, and S. Levine, "Uncertainty-aware reinforcement learning for collision avoidance," *arXiv preprint arXiv:1702.01182*, 2017.

[4] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *International conference on machine learning*, pp. 22–31, PMLR, 2017.

[5] G. Dalal, K. Dvijotham, M. Vecerik, T. Hester, C. Paduraru, and Y. Tassa, "Safe exploration in continuous action spaces," *arXiv preprint arXiv:1801.08757*, 2018.

[6] T.-H. Pham, G. De Magistris, and R. Tachibana, "Optlayer-practical constrained optimization for deep reinforcement learning in the real world," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6236–6243, IEEE, 2018.

[7] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, "Safe learning in robotics: From learning-based control to safe reinforcement learning," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 411–444, 2022.