AlwaysSafe: Reinforcement Learning without Safety Constraint Violations during Training (Extended Abstract)

Thiago D. Simão,[1] **Nils Jansen,**[2] **Matthijs T. J. Spaan**[1]
[1] Delft University of Technology, The Netherlands
[2] Radboud University, Nijmegen, The Netherlands
t.diassimao@tudelft.nl, n.jansen@science.ru.nl , m.t.j.spaan@tudelft.nl

Abstract

Deploying reinforcement learning (RL) involves major concerns around safety. Engineering a reward signal that allows the agent to maximize its performance while remaining safe is not trivial. Safe RL studies how to mitigate such problems. For instance, we can decouple safety from reward using constrained Markov decision processes (CMDPs), where an independent signal models the safety aspects. In this setting, an RL agent can autonomously find tradeoffs between performance and safety. Unfortunately, most RL agents designed for CMDPs only guarantee safety after the learning phase, which might prevent their direct deployment. In this work, we investigate settings where a concise abstract model of the safety aspects is given, a reasonable assumption since a thorough understanding of safety-related matters is a prerequisite for deploying RL in typical applications. Factored CMDPs provide such compact models when a small subset of features describe the dynamics relevant for the safety constraints. We propose an RL algorithm that uses this abstract model to learn policies for CMDPs safely, that is without violating the constraints. During the training process, this algorithm can seamlessly switch from a conservative policy to a greedy policy without violating the safety constraints. We prove that this algorithm is safe under the given assumptions. Empirically, we show that even if safety and reward signals are contradictory, this algorithm always operates safely and, when they are aligned, this approach also improves the agent's performance.

Publication. This is an extended abstract of a paper published at AAMAS-21 (Simão, Jansen, and Spaan 2021).

Introduction

Despite the astonishing successes in Reinforcement Learning (RL) (Sutton and Barto 2018), unsafe exploration still prevents its deployment to real-world tasks (Amodei et al. 2016). This issue has motivated the study of constrained RL to ensure safety (Dulac-Arnold, Mankowitz, and Hester 2019). In this framework, an agent interacts with an environment modeled as a Constrained Markov Decision Process (CMDP) (Altman 1999) without knowledge about the transition, reward, and cost functions. In safe RL (García and Fernández 2015), the cost function is used as a proxy to distinguish between safe and unsafe behaviors. Therefore, the agent must find a policy with maximum expected reward among the safe policies, those with expected cost smaller than a safety threshold.

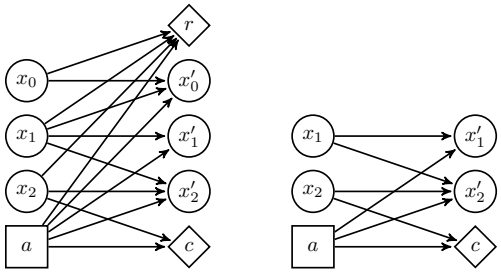

Figure 1: Left: Factored CMDP $\mathcal{M}$ with a safety signal c. Right: Abstraction $\bar{\mathcal{M}}$ ignoring features irrelevant for safety.

Constrained RL algorithms often focus the settings where the agent trains in an assumed perfect simulator and only cares about safety later, when deployed in the real environment. We focus on a setting where the agent interacts directly with the environment and is not allowed to violate the safety constraints while learning.

To provide safety guarantees, one must make some assumptions (Wachi and Sui 2020). We observe that often most of the state description is only relevant for the reward signal and does not influence the safety of the agent. Figure 1 shows an example of such situation, where the feature x_0 does not influence the cost function. In this setting, it can be easy for an expert to define the dynamics relevant for safety. Such constraints may be represented in a compact model and are a prerequisite for deploying RL. Hence, we assume that this compact model is known and is represented by an abstract CMDP $\bar{\mathcal{M}}$. This assumption allows the agent to explore, but **always** within the set of **safe** policies.

Our contribution is four-fold: (*i*) we study the kind of abstraction sufficient to concisely describe and distill safety dynamics. Using factored MDPs (Boutilier, Dearden, and Goldszmidt 1995), (*ii*) we devise an example of such abstract model. Assuming such model is given, (*iii*) we propose the `AlwaysSafe` algorithm, that learns an optimal policy for the CMDP without violating the constraints. Finally, (*iv*) we show that this algorithm is always safe and has no regrets in terms of constraint violation

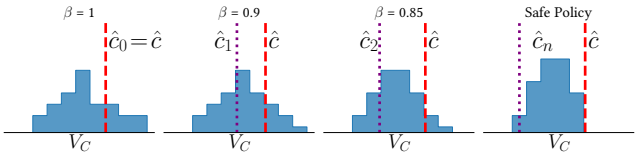

Figure 2: Search for a ground policy that respects the constraints in all CMDPs from the uncertainty set Ξ. The x-axis indicates the expected cost and the y-axis the frequency we can find a CMDP in Ξ for which the policy computed has that expected cost.

Background

Different algorithms have been proposed for constrained RL, one example is the `OptCMDP` algorithm (Efroni, Mannor, and Pirotta 2020). This is a model-based algorithm following the *optimism in the face of uncertainty* principle, so it computes a policy optimistically with respect to a uncertainty set Ξ that contains the true underlying CMDP with high probability. This algorithm has bounded performance regret as well as bounded safety regret.

Unfortunately, the `OptCMDP` algorithm may still violate the constraints during the early episodes, since it lets the agent explore unknown parts of the environment, making its deployment to real-world tasks infeasible. Next, we propose a RL algorithm that can learn without violating the constraints, so it can be used in the *true* RL setting.

Method

This work has a novel perspective on the use of abstractions. While prior work usually focuses on computing a policy on $\bar{\mathcal{M}}$ that maximizes the expected return in the ground CMDP $\mathcal{M}$. Our approach, uses $\bar{\mathcal{M}}$ to compute an abstract policy π_A that is safe for deployment in $\mathcal{M}$. Unfortunately this policy may be suboptimal, since $\bar{\mathcal{M}}$ may ignore features relevant for the reward function. Therefore, we still need a ground policy to eventually achieve optimal behavior.

To find a safe ground policy, we propose to dynamically tight the safety constraints, resulting in increasingly conservative policies. Initially, we compute a ground policy π_G and test if it is safe for all CMDPs on the uncertainty set Ξ. If π_G is not safe, we reduce the safety threshold and compute a new ground policy that is more conservative. This process is repeated until a ground policy that is safe in all CMDPs in Ξ is found, or until the problem becomes infeasible, in which case use deploy the abstract policy π_A to collect more data. This reduces the size of the uncertainty set, which allows us to eventually find a feasible ground policy.

Figure 2 demonstrates a successful search for a safe ground policy using this strategy. The first three plots shows how the cost bound $\hat{c}_i$ changes over the iterations and the last plot shows one of the stopping conditions of the algorithm, when the policy computed according to $\hat{c}_n$ respects the original constraints in all CMDPs in Ξ.

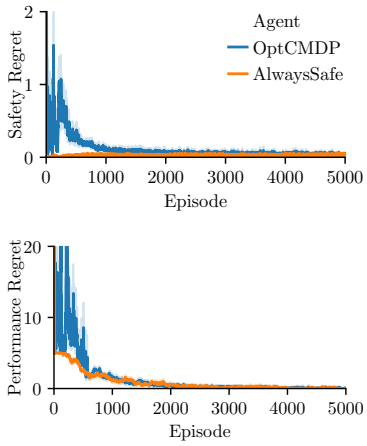

Figure 3: Safety and performance regret of the policy executed in each episode.

Empirical Analysis

The empirical analysis (Figure 3) considers the cliff walking environment with a cost for walking close to the cliff (Lee et al. 2017). It showcases the capabilities of the `AlwaysSafe` algorithm: (*i*) it respects the constraints during training, showing no safety regret; (*ii*) it eventually achieves optimal performance since at the end of training it has no performance regret; and (*iii*) when the cost function is aligned with the reward it reduces the performance regret.

Conclusion

This work considers settings where safety-relevant dynamics are given. We proposed the AlwaysSafe algorithm, that can be optimistic with respect to the reward, while ensuring safety at all times.

In particular, we used an abstract version of the safety-relevant dynamics to compute an abstract policy that is always safe and a ground policy that can achieve high performance. We showed how to switch between these two policies to find an algorithm that is safe and eventually converges to the optimal policy. This method not only enforces the agent to always act safely, but can also prune underperforming actions, improving the training efficiency when the cost function is aligned with the reward function.

In summary, the proposed algorithm is always safe during learning, eventually reaches the optimal policy; and decouples exploration from safety issues in RL.

Acknowledgments

This research is funded by the Netherlands Organisation for Scientific Research (NWO), as part of the Energy System Integration: planning, operations, and societal embedding program and the grants NWO OCENW.KLEIN.187: “Provably Correct Policies for Uncertain Partially Observable Markov Decision Processes” and NWA.1160.18.238: “PrimaVera”.

References

Altman, E. 1999. *Constrained Markov decision processes*, volume 7. CRC Press.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P. F.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety. arXiv:1606.06565.

Boutilier, C.; Dearden, R.; and Goldszmidt, M. 1995. Exploiting Structure in Policy Construction. In *Proc. Int. Joint Conf. on Artificial Intelligence*, 1104–1113.

Dulac-Arnold, G.; Mankowitz, D. J.; and Hester, T. 2019. Challenges of Real-World Reinforcement Learning. In *ICML Workshop RL4RealLife*. arXiv:1904.12901.

Efroni, Y.; Mannor, S.; and Pirotta, M. 2020. Exploration-Exploitation in Constrained MDPs. In *ICML Workshop on Theoretical Foundations of Reinforcement Learning*. arXiv:2003.02189.

García, J.; and Fernández, F. 2015. A Comprehensive Survey on Safe Reinforcement Learning. *Journal of Machine Learning Research* 16: 1437–1480.

Lee, J.; Jang, Y.; Poupart, P.; and Kim, K. 2017. Constrained Bayesian Reinforcement Learning via Approximate Linear Programming. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2088–2095. Melbourne, Australia: ijcai.org.

Simão, T. D.; Jansen, N.; and Spaan, M. T. J. 2021. AlwaysSafe: Reinforcement Learning Without Safety Constraint Violations During Training. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 1226–1235. IFAAMAS.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. MIT press, 2 edition.

Wachi, A.; and Sui, Y. 2020. Safe Reinforcement Learning in Constrained Markov Decision Processes. In *Proceedings of the 37th International Conference on Machine Learning*, 9797–9806. PMLR.