Pure-Past Action Masking

Giovanni Varricchione[1], Natasha Alechina[1,2⋆], Mehdi Dastani[1], Giuseppe De Giacomo[3],
Brian Logan[1,5], and Giuseppe Perelli[4]

[1] Utrecht University, Utrecht, The Netherlands
{n.a.alechina, m.m.dastani, b.s.logan, g.varricchione}@uu.nl
[2] Open University of the Netherlands, Heerlen, The Netherlands
[3] University of Oxford, Oxford, United Kingdom
giuseppe.degiacomo@cs.ox.ac.uk
[4] Sapienza University of Rome, Rome, Italy
perelli@di.uniroma1.it
[5] University of Aberdeen, Aberdeen, United Kingdom

As reinforcement learning (RL) is seeing more applications in real-life scenarios, it is crucial to provide approaches that can enforce safety constraints on the learnt behaviour. In particular, it is desirable to devise approaches that are "*provably safe*" [7], i.e., have mathematical guarantees that agents trained with such approaches do not violate given safety constraints.

In [10], we proposed to use *pure-past linear-time temporal logic* (PPLTL) to define constraints on action selection both during training and execution. Each action a is associated with a PPLTL formula φ_a that should be true in order for the action a to be available to the agent. The formula is evaluated on the history of the system so far and specifies when the action is safe. For example, an action of opening a valve in a water tank is safe if the valve has been closed for two time steps. In this case action `open` will be available to the agent if the formula $\varphi_{\texttt{open}}$ saying that the valve is closed and was closed at the preceding time step. We call a set of PPLTL formulas indexed by the set of agent's actions a "*pure-past action mask*" (PPAM) because unsafe actions are hidden from the agent during training and execution (masked). We compare PPAMs with shields [1,5], another approach from the provably safe RL literature, and show that each safety constraint used to specify a shield has a corresponding PPAM, so PPAMS are as expressive as shields. We analyse thoretical complexity bounds on using PPAMs in RL, and provide experimental evaluations showcasing how PPAMs can be used in practice. This is an extended abstract of the paper published at AAAI 2024; we refer the reader to the full paper for formal definitions and a thorough presentation of the results.

PPLTL [4] is a variant of the well-known linear-time temporal logic. In PPLTL, formulas express properties of histories: for example, the formula "$\mathsf{Y}\varphi$" (read as "*Yesterday* φ") is true given at some timestep i (> 0) of a trace if and only if φ is true at timestep $i-1$ of the same trace. The logic also contains operators H ("*Historically* φ", or "always in the past") and $\varphi\mathsf{Since}\psi$.

⋆ Corresponding author

A useful property of PPLTL shown in [2] is that it is possible to evaluate the truth value of a PPLTL formula at a given time step in a trace by just knowing the current truth values of the propositional symbols and the truth values of the formula's subformulas at the previous time step. This is crucial because it means that we do not need information from the entire history to evaluate a PPLTL formula, but just the values of its subformulas at the previous timestep. We expand each state of the MDP in which the agent learns with the set of subformulas of PPAM formulas. In the worst case this leads to a single exponential blowup of the state space. Compared to shields [1,5], which incur a double exponential blowup in the size of their safety specification, we provide an exponential improvement. Moreover, thanks to a result from [11], we were also able to prove that for every shield, there is a PPAM that enforces the same safety constraints. The full proof of this result is present in the full paper.

We provide two evaluations. Plots of experiments can be found in the full paper.

In the first, we trained agents in the CocktailParty [3] environment. In it, the agent has to serve customers precisely one snack and one drink, and must not serve alcoholic drinks to underage customers. We compare against restraining bolts, a tool introduced in the same paper of this environment. Unlike PPAMs, restraining bolts provide no safety guarantees, allowing the agent to violate the safety constraints but punishing it when this occurs. In our experiments, we observed that the agent trained with the PPAM never violated the safety constraint, whereas the agent trained with the restraining bolt did. Moreover, the agent trained with the PPAM converged faster than the one trained with the restraining bolt. This is in line with a hypothesis claiming that constraining action selection improves sample efficiency [6].

In the second, we showed how PPAMs can be used to address the issue of "*reward gaming*" in the BoatRace environment [8]. Reward gaming [9] is a phenomen that occurs when the agent learns a behaviour that exploits the reward function to obtain a high reward instead of learning the intended task. In BoatRace, the agent has to learn how to navigate around a circular course by passing through specific checkpoints in a clockwise manner. Whenever the agent passes through a checkpoint from the correct clockwise direction, it receives a reward. Left unchecked, the agent learns to step back and forth on a single checkpoint to garner reward. Instead, by using a PPAM, we could easily train an agent to achieve the intended behaviour.

In conclusion, we have presented pure-past action masking, a provably safe RL approach that uses PPLTL to constrain action selection based on the current history (and not just the current state). We have related our approach to shields, which similarly belongs to the category of provably safe RL approaches. After that, we have presented practical applications of PPAMs, showing how they can be used to enforce safety constraints and to solve the issue of reward gaming in a case scenario. In future work, we plan to extend PPAMs to continuous action spaces, as has been done with action masking in other work, and to multi-agent domains, as has been done with shields [5].

References

1. Alshiekh, M., Bloem, R., Ehlers, R., Könighofer, B., Niekum, S., Topcu, U.: Safe reinforcement learning via shielding. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, (AAAI 2018). pp. 2669–2678. AAAI Press (2018)
2. De Giacomo, G., Favorito, M., Fuggitti, F.: Planning for temporally extended goals in pure-past linear temporal logic: A polynomial reduction to standard planning. CoRR **abs/2204.09960** (2022). https://doi.org/10.48550/arXiv.2204.09960, https://doi.org/10.48550/arXiv.2204.09960
3. De Giacomo, G., Iocchi, L., Favorito, M., Patrizi, F.: Foundations for restraining bolts: Reinforcement learning with ltlf/ldlf restraining specifications. In: Proceedings of the 29th International Conference on Automated Planning and Scheduling (ICAPS 2019). pp. 128–136. AAAI Press (2019)
4. De Giacomo, G., Stasio, A.D., Fuggitti, F., Rubin, S.: Pure-past linear temporal and dynamic logic on finite traces. In: Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI 2020). pp. 4959–4965. International Joint Conferences on Artificial Intelligence Organization (2020)
5. ElSayed-Aly, I., Bharadwaj, S., Amato, C., Ehlers, R., Topcu, U., Feng, L.: Safe multi-agent reinforcement learning via shielding. In: Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (IJCAI 2021). p. 483–491. International Foundation for Autonomous Agents and Multiagent Systems (2021)
6. Huang, S., Ontañón, S.: A closer look at invalid action masking in policy gradient algorithms. In: The International FLAIRS Conference Proceedings, 35 (FLAIRS-35). Florida Online Journals (2022)
7. Krasowski, H., Thumm, J., Müller, M., Schäfer, L., Wang, X., Althoff, M.: Provably safe reinforcement learning: Conceptual analysis, survey, and benchmarking (2023)
8. Leike, J., Martic, M., Krakovna, V., Ortega, P.A., Everitt, T., Lefrancq, A., Orseau, L., Legg, S.: AI safety gridworlds. CoRR **abs/1711.09883** (2017), http://arxiv.org/abs/1711.09883
9. Skalse, J., Howe, N., Krasheninnikov, D., Krueger, D.: Defining and characterizing reward gaming. In: Advances in Neural Information Processing Systems 35 (NeurIPS 2022). pp. 9460–9471. Curran Associates, Inc. (2022)
10. Varricchione, G., Alechina, N., Dastani, M., De Giacomo, G., Logan, B., Perelli, G.: Pure-past action masking. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI 2024). pp. 21646–21655. AAAI Press (2024)
11. Zuck, L.: Past temporal logic. The Weizmann Institute of Science (1986)