Pure-Past Action Masking

G. Varricchione[1] N. A. Alechina[2,1] M. M. Dastani[1]
G. De Giacomo[3] B. S. Logan[4,1] G. Perelli[5]

[1] : Utrecht University [2] : Open University [3] : University of Oxford
[4] : University of Aberdeen [5] : Sapienza University of Rome

Highlights

Objective

- Devise an approach that formally guarantees a RL agent chooses only actions that do not violate (non-Markovian) safety constraints

Methodology

- For each action, provide a Pure Past Linear-time Temporal Logic (PPLTL) formula
- Allow the agent to choose an action if and only if its corresponding formula is true

Provably Safe RL

- In [1], a taxonomy of safe RL approaches is proposed, dividing approaches in *provably* and *non-provably* safe
- Provably safe approaches provide theoretical guarantees to satisfy the safety constraints on which they are defined
- Amongst these, *action masking* techniques limit agents by not allowing them to take *unsafe* actions
- While the literature in action masking is rich, **few approaches can enforce non-Markovian safety constraints**, i.e., constraints that depend on the entire history

(Preemptive) Shields

- Preemptive shields [2] are Mealy machines that can enforce non-Markovian safety constraints via action masking
- Shields are synthesised by solving a safety game that involves an abstraction of the MDP and a safety LTL specification
- At each timestep, preemptive shields restrict the set of actions available to the agent, by specifying the set of valid ones through their output function
- While shields guarantee satisfaction of the input LTL specification, both the synthesis and the shield's size can be **double exponential** in the size of the LTL specification

Pure Past Linear-time Temporal Logic

In PPLTL modalities span only over the past:

$$\varphi ::= p \in \mathit{Prop} \mid \neg\varphi \mid \varphi \vee \varphi \mid \ominus\varphi \mid \varphi \mathcal{S} \varphi$$

Given a PPLTL formula φ, if two traces τ, τ' of length n, n' are such that:

- The same propositional symbols are true at τ_n and $\tau'_{n'}$;
- For each subformula $\psi \in \mathit{Subf}(\varphi)$, $\tau_{n-1} \models \psi \Leftrightarrow \tau'_{n'-1} \models \psi$;

Then $\tau \models \varphi \Leftrightarrow \tau' \models \varphi$. This implies that any PPLTL formula φ can be evaluated over a trace τ in time linear in $|\varphi|$ and constant in $|\tau|$, given the truth values of $\mathit{Subf}(\varphi)$ at the penultimate timestep of τ.

Pure-Past Action Masks

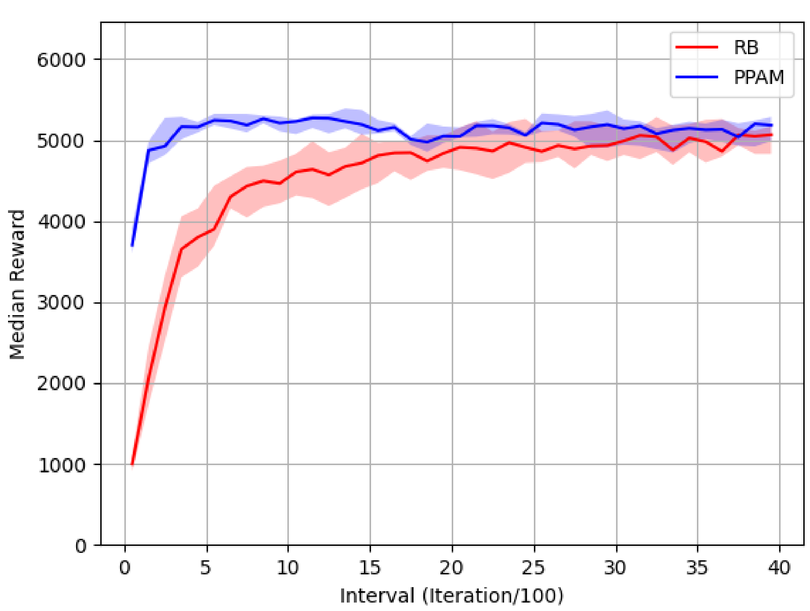
- A pure past action mask (PPAM) is a pair $(\mathcal{L}, \{\varphi_a : a \in \mathit{Act}\})$, where $\mathcal{L}$ is the set of features of the PPAM and φ_a the PPLTL formula associated to action a
- Given the history τ, the agent can perform action a if and only if $\tau \models \varphi_a$
- We expand MDP states to include the PPAM's suformulas true at the previous timestep, if it exists, given the current history - we will use these to easily evaluate the PPAM's formulas at the current timestep
- By restricting the actions the agent can take in the MDP using a PPAM, we formally guarantee that the agent will **never** violate the PPAM's constraints, neither during nor after training

Experimental Results

- In CocktailParty [3] the agent learns how to serve snacks and drinks to customers, with the constraints that it must not serve the same customer twice, and must not serve alcoholic drinks to minors
- We can constrain the action to serve a drink with a PPAM as follows:

$$\varphi_{\texttt{serve\_drink}} = \neg\,(\top \mathcal{S} \mathit{served\_drink}) \wedge (\neg \mathit{minor} \vee \neg \mathit{holding\_alcohol})$$

- We compare the performance of an agent constrained by a PPAM to follow these constraints against that of an agent trained with a restraining bolt, a tool introduced in [3] to easily define non-Markovian rewards that however does not provide safety guarantees:



Comparison with Shields

- Using results from [4] and [5], we can show that for every shield there is a PPAM that masks actions in the same way
- As an example, we consider the WaterTank environment, as presented in the original shields work [2]
- In it, the agent can open the valve if the water level is lower than 93 liters and, in case it is closed in the current timestep, then it has been for at least the last three timesteps
- We can easily model this constraint by using the following PPLTL formula:

$$\mathit{level} \leq 93 \wedge (\mathit{close} \rightarrow \ominus \mathit{close} \wedge \ominus\ominus \mathit{close})$$

Conclusions

- By using PPLTL, we constrain actions given the history
- We memorize (in the MDP state) and update which PPLTL subformulas are true, so that we do not need the whole history to evaluate the PPAM's formulas
- Thus, we "only" incur a **single exponential** blowup (in the size of the PPAM's formulas) to **guarantee** constraint satisfaction

References

[1] Hanna Krasowski, Jakob Thumm, Marlon Müller, Lukas Schäfer, Xiao Wang, and Matthias Althoff. Provably safe reinforcement learning: Conceptual analysis, survey, and benchmarking. *CoRR*, 2023.

[2] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, (AAAI-18)*.

[3] Giuseppe De Giacomo, Luca Iocchi, Marco Favorito, and Fabio Patrizi. 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)*.

[4] Lenore Zuck. *Past temporal logic.* PhD thesis, Weizmann Institute of Science, 1987.

[5] Zohar Manna and Amir Pnueli. A hierarchy of temporal properties. In *Proceedings of the Ninth Annual ACM Symposium on Principles of Distributed Computing*, 1990.