Leveraging Approximate Model-based Shielding for Probabilistic Safety Guarantees in Continuous Environments

Extended Abstract

Alexander W. Goodall
Imperial College London
London, United Kingdom
a.goodall22@imperial.ac.uk

Francesco Belardinelli
Imperial College London
London, United Kingdom
francesco.belardinelli@imperial.ac.uk

ABSTRACT

Shielding is a popular technique for achieving safe reinforcement learning (RL). However, classical shielding approaches come with quite restrictive assumptions making them difficult to deploy in complex environments, particularly those with continuous state or action spaces. In this paper we extend the more versatile *approximate model-based shielding* (AMBS) framework to the continuous setting. In particular we use *Safety Gym* as our test-bed, allowing for a more direct comparison of AMBS with popular constrained RL algorithms. We also provide strong probabilistic safety guarantees for the continuous setting. In addition, we propose two novel penalty techniques that directly modify the policy gradient, which empirically provide more stable convergence in our experiments.

KEYWORDS

Reinforcement Learning; Shielding; Continuous Environments

ACM Reference Format:
Alexander W. Goodall and Francesco Belardinelli. 2024. Leveraging Approximate Model-based Shielding for Probabilistic Safety Guarantees in Continuous Environments: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Derived from formal methods, *shielding* [2] is a powerful approach for guaranteeing the safety of reinforcement learning (RL) systems during training and deployment. In this paper, we leverage *Approximate Model-based Shielding* (AMBS) [8], which can be applied in more realistic settings, where the safety-relevant dynamics of the system are not known in advance. AMBS is a model-based RL algorithm and a general framework for shielding learned RL policies by simulating possible futures in the latent space of a learned dynamics model or *world model* [11], in particular we use *DreamerV3* [12] as the stand-in dynamics model. AMBS is an approximate method that relies on learned components and Monte Carlo sampling, as such it cannot provide the same safety guarantees that classical shielding can [2]. That being said, strong probabilistic guarantees have been established for AMBS in the tabular case [8].

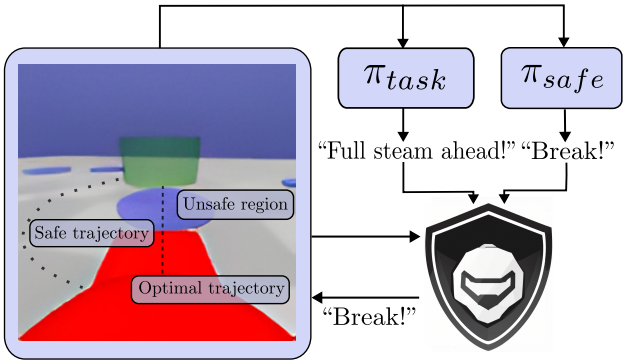

Figure 1: A simple example in Safety Gym [16]. The reward policy proposes actions along the optimal trajectory. However, this trajectory enters an unsafe region, so the shield overrides these actions with *"Break!"* actions proposed by the safe policy. As a result, the safe trajectory is not recovered and the two policies continuously fight for control.

Naïvely applying AMBS to Safety Gym [16], results in a high-variance return distribution. We hypothesise that this phenomena can be explained by the reward policy fighting with the shield to gain control of the system, see Fig. 1. In our paper we alleviate this problem by providing the underlying (unshielded) policy with some safety information by using a simple penalty critic (abbrv. PENL). We also introduce two more sophisticated penalty methods, the first based on *Probabilistic Logic Shielding* [18] (abbrv. PLPG), and the second loosely inspired by counter-examples (abbrv. COPT), a familiar concept from model checking [5] and verification [7].

Contributions. We summarise our contributions: (1) we extend and apply AMBS [8] to the continuous setting, specifically we use Safety Gym [16] to obtain a meaningful comparison with other model-based and safety-aware algorithms. (2) We subtly build on the probabilistic safety guarantees of AMBS, by establishing the same sample complexity bounds for continuous state and action spaces. (3) We demonstrate that extending AMBS with safety information is crucial for stable convergence to a safe policy. In addition, we introduce two novel penalty methods that empirically improve the stability of the learned policy in the later stages of training. (4) In our experiments we demonstrate that our extended version of AMBS dramatically reduces the total number of safety-violations, compared to other safety-aware RL algorithms, while maintaining good convergence properties and performance w.r.t. episode return. The full version of our paper is available here [10].

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

Table 1: Episode return and cumulative violations at the end of training.

		AMBS + PENL	AMBS + PLPG	AMBS + COPT	DreamerV3 + LAG
PointGoal1	Episode Return ↑	17.32 ± 3.29	17.76 ± 2.18	18.19 ± 1.72	**19.15 ± 0.92**
(1M)	# Violations ↓	**9354 ± 3734**	13937 ± 1722	10766 ± 2877	24996 ± 6627
PointGoal2	Episode Return ↑	10.64 ± 2.61	10.30 ± 3.44	9.16 ± 4.07	**15.78 ± 1.84**
(1.5M)	# Violations ↓	**29720 ± 3850**	30673 ± 1800	30839 ± 4647	52157 ± 6151
CarGoal1	Episode Return ↑	8.87 ± 2.95	5.86 ± 2.30	5.96 ± 4.15	**11.23 ± 4.10**
(1M)	# Violations ↓	**11423 ± 1479**	13236 ± 3294	14500 ± 4675	28639 ± 4644
PointGoal1	Episode Return ↑	16.60 ± 2.23	19.45 ± 1.62	18.66 ± 2.15	**19.74 ± 1.43**
(10M)	# Violations ↓	19039 ± 2339	**17049 ± 1321**	18320 ± 3080	46153 ± 4637

2 PROBLEM SETUP AND PRELIMINARIES

To obtain a meaningful comparison with key prior works [4, 13], we opt for vision based input. As such we model the environment as a *partially observable Markov decision process* (POMDP) [15]. In addition, we introduce a set of *atomic propositions AP* and a *labelling function* $L : S \to 2^{AP}$ [5]. The full POMDP tuple is defined as follows $\mathcal{M} = \langle S, A, p, \iota_{init}, R, \gamma, \Omega, O, AP, L\rangle$. In addition to maximising reward (i.e. $\pi^* = \arg\max_\pi \mathbb{E}[\sum_{t=1}^{\infty} \gamma^{t-1} \cdot r_t]$), we are given a propositional safety-formula Ψ, and we seek a policy that also achieves the minimum rate of violations of the safety-formula Ψ.

At each timestep AMBS [8] checks the PCTL [5] property Δ-bounded safety [8, 9]. A given state $s \in S$ satisfies Δ-bounded safety, or formally, $s \models \mathbb{P}_{\geq 1-\Delta}(\square^{\leq n}\Psi)$, iff,

$$\mu_s(\{\tau \mid \tau[0] = s, \forall i\ 0 \leq i \leq n, \tau[i] \models \Psi\}) \in [1-\Delta, 1] \tag{1}$$

where μ_s is a well defined probability measure on the set of traces $\tau = s_0 \to s_1 \to \cdots \to s_n \to \ldots$, with $\tau[i] = s_i$, from the state s. We let $\mu_{s\models\phi}$ be short-hand for this measure.

3 SAFETY GUARANTEES

Let $\mathcal{T}$ denote the true transition system $\mathcal{T} : S \times S \to [0, 1]$ induced by a fixed policy π in the MDP (or POMDP), and let $\widehat{\mathcal{T}}$ denote the approximate transition system $\widehat{\mathcal{T}} : S \times S \to [0, 1]$ induced by the same fixed policy π in a learned approximation of the dynamics of the MDP. We state the following main result.

Theorem 1. *Suppose that for all $s \in S$, the Kullback-Leibler (KL) divergence between the distributions $\mathcal{T}(\cdot \mid s)$ and $\widehat{\mathcal{T}}(\cdot \mid s)$ is upper-bounded by some $\alpha \leq \epsilon^2/(2n^2)$. That is,*

$$D_{KL}\Big(\mathcal{T}(\cdot \mid s) \,\Big\|\, \widehat{\mathcal{T}}(\cdot \mid s)\Big) \leq \alpha\ \forall s \in S \tag{2}$$

Now fix an $s \in S$ and let $\epsilon > 0$, $\delta > 0$ be given. With probability $1 - \delta$ we can obtain an ϵ-approximate estimate of the measure $\mu_{s\models\phi}$, by sampling m traces $\tau \sim \widehat{\mathcal{T}}$, provided that,

$$m \geq \frac{2}{\epsilon^2} \log\left(\frac{2}{\delta}\right) \tag{3}$$

In the partially observable setting, world models [11], such as DreamerV3 [12] have already been theoretically motivated [8] and the extension of these results to the continuous setting is simple.

4 EXPERIMENTAL RESULTS

Setup. We evaluate AMBS [8] with each of the three penalty techniques separately (i.e. PENL, PLPG and COPT) , as a baseline with use a version of DreamerV3 [12] that implements the Augmented Lagrangian [4, 17]. We use the following three vision based environments from Safety Gym [16]: *PointGoal1*, *PointGoal2* and *CarGoal1*. Observations correspond to $3 \times 64 \times 64$ dimensional tensors and the action space is $[-1, +1]^2$ for all environments. The implementation details can be found here at: https://github.com/sacktock/AMBS.

Discussion. We see that across all environments our methods out perform the baseline w.r.t. the cumulative number of violations, see Tab. 1. In terms of episode return our methods do exhibit slower convergence – this is a trade-off we would expect in these environments. In the PointGoal1 environment all algorithms appear to have converged within 1M frames. However, for longer training runs (10M frames) it appears that without the more principled penalty techniques, i.e. PLPG and COPT, then AMBS diverges. More work is required to understand the convergence properties of PLPG and COPT, although empirically they maintain more stable convergence than the simple penalty critic (PENL).

5 CONCLUSION

In our paper we successfully extended AMBS [8] to the continuous setting and we proposed three penalty techniques that are crucial for the convergence of AMBS in environments that have an inherent trade-off between safety and reward, e.g. those in Safety Gym [16].

Compared with CMDP [3], shielding approaches for safe RL are policy agnostic, i.e. the safety of the system depends on entirely on the shield, rather than the convergence of the learned policy. We stress the importance of results such as Thm. 1, since it allows practitioners to have good statistical confidence in their systems. While Thm. 1 does make non-trivial assumptions, for particular settings we know when these assumptions are satisfied from previously established sample complexity bounds [1, 6, 14].

Important future work includes, a more thorough investigation into the convergence properties of PLPG [18] and COPT, and the key components of AMBS [8]. It may also be interesting to establish stronger theoretical claims for common continuous settings, for which sample complexity bounds already exist, and present it in a way that is compatible with AMBS.

ACKNOWLEDGMENTS

This work was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org).

REFERENCES

[1] Yasin Abbasi-Yadkori and Csaba Szepesvári. 2011. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*. JMLR Workshop and Conference Proceedings, 1–26.

[2] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. 2018. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[3] Eitan Altman. 1999. *Constrained Markov decision processes: stochastic modeling*. Routledge.

[4] Yarden As, Ilnura Usmanova, Sebastian Curi, and Andreas Krause. 2022. Constrained policy optimization via bayesian world models. *arXiv preprint arXiv:2201.09802* (2022).

[5] Christel Baier and Joost-Pieter Katoen. 2008. *Principles of model checking*. MIT press.

[6] Emma Brunskill, Bethany R Leffler, Lihong Li, Michael L Littman, and Nicholas Roy. 2009. Provably efficient learning with typed parametric models. (2009).

[7] Edmund Clarke, Orna Grumberg, Somesh Jha, Yuan Lu, and Helmut Veith. 2000. Counterexample-Guided Abstraction Refinement. In *Computer Aided Verification*, E. Allen Emerson and Aravinda Prasad Sistla (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 154–169.

[8] Alexander W Goodall and Francesco Belardinelli. 2023. Approximate Model-Based Shielding for Safe Reinforcement Learning. In *ECAI 2023*. IOS Press, 883–890.

[9] Alexander W Goodall and Francesco Belardinelli. 2023. Approximate Shielding of Atari Agents for Safe Exploration. *arXiv preprint arXiv:2304.11104* (2023).

[10] Alexander W. Goodall and Francesco Belardinelli. 2024. Leveraging Approximate Model-based Shielding for Probabilistic Safety Guarantees in Continuous Environments. arXiv:2402.00816

[11] David Ha and Jürgen Schmidhuber. 2018. Recurrent World Models Facilitate Policy Evolution. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/2de5d16682c3c35007e4e92982f1a2ba-Paper.pdf

[12] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2023. Mastering Diverse Domains through World Models. *arXiv preprint arXiv:2301.04104* (2023).

[13] Weidong Huang, Jiaming Ji, Borong Zhang, Chunhe Xia, and Yaodong Yang. 2023. Safe DreamerV3: Safe Reinforcement Learning with World Models. *arXiv preprint arXiv:2307.07176* (2023).

[14] Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. 2020. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems* 33 (2020), 15312–15325.

[15] Martin L Puterman. 1990. Markov decision processes. *Handbooks in operations research and management science* 2 (1990), 331–434.

[16] Alex Ray, Joshua Achiam, and Dario Amodei. 2019. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708* 7, 1 (2019), 2.

[17] Jorge Nocedal Stephen J Wright. 2006. Numerical optimization.

[18] Wen-Chi Yang, Giuseppe Marra, Gavin Rens, and Luc De Raedt. 2023. Safe Reinforcement Learning via Probabilistic Logic Shields. *arXiv preprint arXiv:2303.03226* (2023).