ProSh: Probabilistic Shielding for Model-free Reinforcement Learning

Edwin Hamel-de le Court†
Imperial College
London, United Kingdom
e.hamel-de-le-court@imperial.ac.uk

Gaspard Ohlmann†
Independent Researcher
Mulhouse, France
gaspard.ohlmann@outlook.com

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

ABSTRACT

Safety is a major concern in reinforcement learning (RL): we aim at developing RL systems that not only perform optimally, but are also safe to deploy by providing formal guarantees about their safety. To this end, we introduce Probabilistic Shielding via Risk Augmentation (ProSh), a model-free algorithm for safe reinforcement learning under cost constraints. ProSh augments the Constrained MDP state space with a risk budget and enforces safety by applying a shield to the agent's policy distribution using a learned cost critic. The shield ensures that all sampled actions remain safe in expectation. We also show that optimality is preserved when the environment is deterministic. Since ProSh is model-free, safety during training depends on the knowledge we have acquired about the environment. We provide a tight upper-bound on the cost in expectation, depending only on the backup-critic accuracy, that is *always* satisfied during training. Under mild, practically achievable assumptions, ProSh guarantees safety even at training time, as shown in the experiments.

KEYWORDS

Safe Reinforcement Learning, Formal Methods, Shielding, Probabilistic Temporal Logic, Stochastic Systems

ACM Reference Format:
Edwin Hamel-de le Court†, Gaspard Ohlmann†, and Francesco Belardinelli. 2026. ProSh: Probabilistic Shielding for Model-free Reinforcement Learning. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/KCVZ6904

† *These authors contributed equally to this work.*

1 INTRODUCTION

A key challenge in AI is designing agents that learn to act optimally in unknown environments [14]. Reinforcement Learning (RL) – particularly when combined with deep neural networks – has made impressive strides in domains such as games [9], robotics [5], and autonomous driving [4]. However, ensuring safety during both training and deployment remains a critical barrier to real-world adoption. This paper focuses on safe learning in Constrained Markov Decision Processes (CMDPs), where agents maximize the expected discounted reward, while keeping the cumulative expected cost below a given threshold. Traditional methods, such as Lagrangian approaches [2, 10, 12], allow to converge towards a safe policy, but do not provide formal guarantees on safety, neither during training nor on the provided policy.

We introduce ProSh: a novel probabilistic shielding method that ensures safety also at training time, in model-free RL. Unlike classic shielding methods that restrict unsafe actions based on a model of the environment, ProSh operates on distributions, augments the CMDP with a risk budget, and uses a learned cost critic to guide safe exploration. Crucially, ProSh does not assume access to a simulator or environment abstraction, and is compatible with continuous environments. We provide a formal bound for the safety of ProSh. In addition, we show that ProSh is optimal in the deterministic setting and that it is safe in both cases up to a controllable and vanishing coefficient. The full version of this paper is available at [7].

2 PROSH: THE AUGMENTATION METHOD

Our approach uses *risk-augmentation*, similar to [6], whereby each state is paired with a risk budget that represents the expected cost the agent is allowed to incur from that point onward. To each state–action pair (s, a) with $a \in A(s)$, we associate an estimate of the achievable minimum expected cost, denoted by the critic $Q_b(s, a)$, and define $Q_b(s) = \min_{a \in A(s)} Q_b(s, a)$.

Definition 1 (Risk-augmented CMDP). *For any CMDP $\mathcal{M} = (S, A, s_i, P, R, C, d, \gamma_r, \gamma_c)$, we augment the set of states and actions, $\overline{S} = S \times \mathbb{R}$ and $\overline{A}(s, x) = A \times \mathbb{R}$. The Transition Probability Function is defined as $\overline{P}((s', x')|(a, y_a), (s, x)) = P(s'|s, a)$ if $\gamma_c x' = y - Q_b(s, a) + \gamma_c Q_b(s')$ and 0 otherwise.*

The transition function $\overline{P}$ ensures that the corresponding risk is spread coherently to each state in the non-deterministic case.

We consider policies that keep the risk budget synchronized with the CMDP dynamics. For these policies, the risk budget corresponds, up to a controllable factor, to an upper bound of the expected cost. We call these the Q_b-shielded policies. These policies are safe, see [7] for the complete theoretical framework.

Definition 2 (Q_b-Shielded policies). *We define the **backup policy** π_b on $\mathcal{M}$ as for any $s \in S$, $\pi_b(s) = \delta_{a_0}$, where $Q_b(s, a_0) = \min_{a \in A(s)} Q_b(s, a)$.*

*A policy $\overline{\pi}$ is said to be Q_b-**shielded** if for any $(s, x) \in \bar{S}$, if $x \geq Q_b(s)$, there exist $(P_a)_{a \in A(s)}$ and $(y_a)_{a \in A(s)}$, with $y_a \geq Q_b(s, a)$ for all $a \in A(s)$, such that for all $(s, x) \in \overline{S}$,*

$$\overline{\pi}(s, x) = \sum_{a \in A(s)} P_a \delta_{(a, y_a)}, \; x \geq \sum_{a \in A(s)} P_a y_a,$$

and $\overline{\pi}(s, x) = \delta_{(\pi_b(s), x)}$ when $x < Q_b(s)$.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/KCVZ6904

The policies are shielded when they always allow a budget for the actions larger than the estimated minimal cost of taking this action. Moreover, the total budget allowed for the actions is, in expectancy, smaller than the current budget. For a shielded policy $\overline{\pi}$, there exists a policy π simulating $\overline{\pi}$ in $\mathcal{M}$ that we call the projection of $\overline{\pi}$ onto $\mathcal{M}$. We now present the Q_b-**shield**, denoted Ξ_{Q_b}, that takes an augmented policy and outputs a Q_b-shielded policy.

Theorem 1. *Let $\overline{\pi}_b$ be the policy that chooses $(\pi_b(s), Q_b(s,a))$ in any augmented state (s,x), and V be the set of augmented policies $\bar{\pi}$ such that for any (s,x), there exist $(P_a)_{a\in A(s)}$ and $(y_a)_{a\in A(s)}$ with $y_a \geq Q_b(s,a)$ satisfying $\bar{\pi}(s,x) = \sum_a P_a \delta_{(a,y_a)}$. For any $\overline{\pi} \in V$, the policy $\Xi_{Q_b}(\overline{\pi})$ defined as*

$$\Xi_{Q_b}(\overline{\pi})(s,x) = \begin{cases} \delta_{(\pi_b(s),x)}, \text{ for } x < Q_b(s), \\ \text{the mixture } (1-\lambda)\overline{\pi} + \lambda\overline{\pi}_b, \ x \leq Q_b(s), \end{cases}$$

*where $\lambda = (x - \sum_a P_a y_a)/(Q_b(s,a) - \sum_a P_a y_a)$, is called the Q_b-**shield** of $\overline{\pi}$ and is Q_b-shielded.*

Learning Algorithm. We present a minimal version of the training loop for the main actor in Algorithm 1. A complete implementation is proposed in [7]. The shield acts on distribution, rather than actions. It uses the values provided by the critic Q_b and combines it with the backup action to obtain a safe distribution.

Algorithm 1 ProSh: Main Actor only

Input: cost budget d, margin δ, cost discount γ_c
Initialize main actor π_r^ψ, backup critic Q_b^θ
for each episode **do**
  $s \leftarrow \bar{s}_i$, $x \leftarrow d - \delta$
  while not done **do**
    Shield: $\mu_{\text{safe}} \leftarrow \Xi_{Q_b}(\bar{\pi}_r^\psi)(s,x)$
    Sample $(a,y) \sim \mu_{\text{safe}}$
    Execute (a,y), observe (s',x'), reward r, cost c
    Store transition $((s,x),(a,y),r,c,(s',r'))$
    $s \leftarrow s'$, $r \leftarrow r'$
  end while
  Update:
  - Update π_θ using shielded distributions from memory
end for
return projected policy π_r^ψ

The shield precedes sampling, so the exploration is budget-safe. The choice of the RL update for θ is open (e.g., PG, PPO, A2C). Additionally, the backup critic Q_b can be learned, either in parallel, off-policy, or even pre-computed. The following theorem holds.

Theorem 2 (Safety and near-Optimality). (i) *At any step of the algorithm, the output policy $\bar{\pi}$ satisfies, with Q_b^* the minimal cost and $x_0 \geq Q_b(s_0)$,*

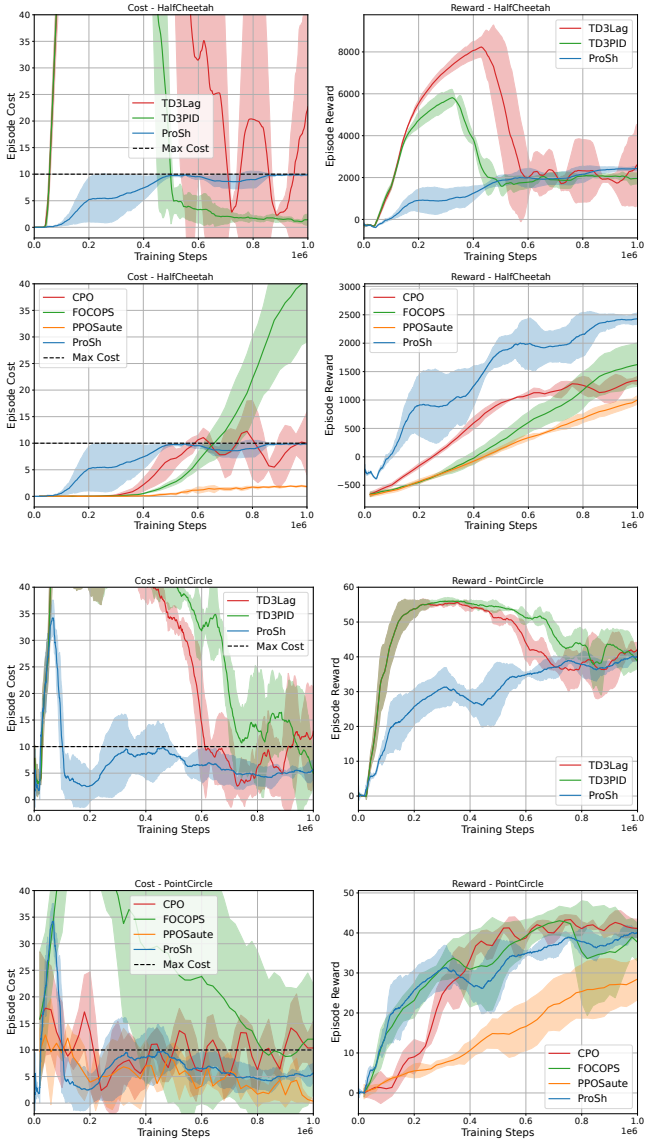
$$C_{\bar{\mathcal{M}}(s_0,x_0)}(\bar{\pi}) \leq x_0 + \frac{2\Delta_b}{1-\gamma_c}, \quad \Delta_b = ||Q_b - Q_b^*||_\infty$$

(ii) *The algorithm is asymptotically optimal in deterministic environments as $\Delta_b \to 0$. More precisely, with Π the set of all policies in $\mathcal{M}$,there exists a shielded policy $\overline{\pi}$ with cost at most $d+2\mathcal{E}$ for $\mathcal{E} = \frac{2\Delta_b}{1-\gamma_c}$, such that $\mathcal{R}(\bar{\pi}) \geq \max_{\pi\in\Pi,\, C(\pi)\leq d} \mathcal{R}(\pi)$.*

Remark 1 (On assumption $\Delta_b \to 0$). *Under several assumptions (e.g. **tabular case**, **Over-parameterized neural networks**,**Batch (fitted) Q-evaluation**), we have $\Delta_b \to 0$ [3, 8, 15, 16].*

3 EXPERIMENTAL EVALUATION

We compared ProSh with PPO-Saute [11], TD3-Lagrangian [10], PID-TD3 [13], FOCOPS [17] and CPO [1]. Please refer to [7] for the complete benchmark. ProSh maintains a high level of safety during training. PPO-Saute is the only other algorithm that achieves a comparable level of safety, although with reduced performance. The results suggest that ProSh provides the best trade-off between safety and performance in the evaluated environments, especially when ensuring safety is critical.



Acknowledgments. The research presented in this paper was supported by the EPSRC grant number EP/X015823/1, "An abstraction-based technique for Safe Reinforcement Learning".

REFERENCES

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained Policy Optimization. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 22–31. https://proceedings.mlr.press/v70/achiam17a.html

[2] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2019. Benchmarking Safe Exploration in Deep Reinforcement Learning. *arXiv preprint arXiv:1805.05800* (2019).

[3] Alekh Agarwal, Sham Kakade, Nan Jiang, and Wen Sun. 2020. Flambe: Structural Complexity and Representation Learning of Low Rank MDPs. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[4] Alex Kendall, Jeffrey Hawke, David Janz, Przemyslaw Mazur, Daniele Reda, John-Mark Allen, Vinh-Dieu Lam, Alex Bewley, and Amar Shah. 2019. Learning to Drive in a Day. In *International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019*. IEEE, 8248–8254. https://doi.org/10.1109/ICRA.2019.8793742

[5] Jens Kober, J. Bagnell, and Jan Peters. 2013. Reinforcement Learning in Robotics: A Survey. *The International Journal of Robotics Research* 32 (09 2013), 1238–1274. https://doi.org/10.1177/0278364913495721

[6] Edwin Hamel-De le Court, Francesco Belardinelli, and Alexander W. Goodall. 2025. Probabilistic Shielding for Safe Reinforcement Learning. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, Toby Walsh, Julie Shah, and Zico Kolter (Eds.). AAAI Press, 16091–16099. https://doi.org/10.1609/AAAI.V39I15.33767

[7] Edwin Hamel-De le Court, Gaspard Ohlmann, and Francesco Belardinelli. 2025. ProSh: Probabilistic Shielding for Model-free Reinforcement Learning. arXiv:2510.15720 [cs.LG] https://arxiv.org/abs/2510.15720

[8] Xingtu Liu. 2024. Information-Theoretic Generalization Bounds for Batch Reinforcement Learning. *Entropy* 26, 11 (2024), 995.

[9] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. 2013. Playing Atari with Deep Reinforcement Learning. *CoRR* abs/1312.5602 (2013). arXiv:1312.5602 http://arxiv.org/abs/1312.5602

[10] Alex Ray, Joshua Achiam, and Dario Amodei. 2019. Benchmarking Safe Exploration in Deep Reinforcement Learning. In *arXiv preprint arXiv:1910.01708*.

[11] Aivar Sootla, Alexander I. Cowen-Rivers, Taher Jafferjee, Ziyan Wang, David Mguni, Jun Wang, and Haitham Bou-Ammar. 2022. Saute RL: Almost Surely Safe Reinforcement Learning Using State Augmentation. arXiv:2202.06558 [cs.LG] https://arxiv.org/abs/2202.06558

[12] Adam Stooke, Joshua Achiam, and Pieter Abbeel. 2020. Responsive Safety in Reinforcement Learning by PID Lagrangian Methods. In *ICML*.

[13] Adam Stooke, Joshua Achiam, and Pieter Abbeel. 2020. Responsive Safety in Reinforcement Learning by PID Lagrangian Methods. arXiv:2007.03964 [math.OC] https://arxiv.org/abs/2007.03964

[14] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement Learning: An Introduction* (2nd ed.). MIT press, Cambridge, MA.

[15] Christopher J. C. H. Watkins and Peter Dayan. 1992. Q-Learning. *Machine Learning* 8, 3–4 (1992), 279–292.

[16] Zhihan Xie, Qi Cai, Ethan Zhou, Alexander Risteski, and Alexander G. Gray. 2021. Bellman Error is a Good Proxy for Value Error. In *International Conference on Machine Learning (ICML)*.

[17] Yiming Zhang, Quan Vuong, and Keith W. Ross. 2020. First Order Constrained Optimization in Policy Space. arXiv:2002.06506 [cs.LG] https://arxiv.org/abs/2002.06506