Scenario Generation for Risk-Aware Reinforcement Learning

Extended Abstract

Mohit Prashant
Nanyang Technological University
Singapore
mohit010@e.ntu.edu.sg

Arvind Easwaran
Nanyang Technological University
Singapore
arvinde@ntu.edu.sg

ABSTRACT

Recent works in reinforcement learning (RL) safety aim to synthesize guarantees via stochastic policy verification — constructing probabilistic barrier-certificates by sampling policy trajectories with respect to safety constraints, thereby demarcating known safe behaviour from unknown behaviour. However, the quality — i.e. tightness — of the bounds on constraint violation is subject to transition uncertainty and adverse perturbation, placing the agent in insufficiently explored states. To address this, we approximate the distribution of the encountered state-space and construct upper and lower-bound barrier-certificates using latent characteristics of states, optimizing for regions of known, safe behaviour with high confidence. We frame this in our work as a dual optimization, where the lower-bound barrier-certificate presents a more conservative estimate of the safe region than the upper-bound barrier-certificate. Sampling states that lie within the set difference of the two during training, i.e. the non-robust region, tightens the bounds and sharpens the quality of the solution.

KEYWORDS

PAC Guarantees, Reinforcement Learning, Safety

ACM Reference Format:
Mohit Prashant and Arvind Easwaran. 2026. Scenario Generation for Risk-Aware Reinforcement Learning: Extended Abstract. 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/PBUD3638

1 INTRODUCTION AND RELATED WORK

Ensuring that reinforcement learning (RL) agents satisfy safety constraints is critical in high-stakes domains such as aviation, healthcare and robotics. Classical safety-verification methods provide rigorous guarantees but rely heavily on model-checking and known transition dynamics [1–3, 6, 10, 11, 14, 16, 24, 33], which struggle to scale in complex, model-free environments.

Recent model-free approaches address this limitation by placing probably approximately correct (PAC) guarantees on constraint violations [20, 30, 36]. Framing safety as a chance-constrained program (CCP) enables empirical evaluation via sampled trajectories, yielding upper and lower bounds $\epsilon_1, \epsilon_2 \in (0, 1)$ on violation probability with confidence $\delta \in (0, 1)$ [5]. While barrier-certificates are traditionally used to certify these safe regions [6, 17, 19, 23, 26], they frequently require pre-defined unsafe states or domain-specific knowledge.

Furthermore, the informativeness of PAC guarantees depends entirely on the tightness of the bounds — i.e. $|\epsilon_1 - \epsilon_2|$. In practice, these bounds remain loose due to biased exploration, transition perturbations and train-deployment distributional shifts [4, 7, 13, 18, 25, 34]. Simultaneously, while curriculum learning (CRL) frameworks and adversarial scenario generation have proven effective at incrementally improving policy robustness [8, 9, 12, 21, 22, 29, 31], they lack an integrated, formal safety-verification mechanism.

To bridge this gap, we extend the PAC barrier-certificate (PABC) framework [30] by integrating it with generative latent spaces to actively guide exploration. Using upper and lower barrier-certificates derived from a dual CCP, we define a *tentatively unsafe* exploration region as the symmetric difference between their corresponding safe sets. By training a variational autoencoder (VAE) on the encountered state distribution, we sequentially produce targeted scenarios to refine the policy. This second learning phase actively reduces $|\epsilon_1 - \epsilon_2|$.

2 PROBLEM FORMULATION

2.1 Markov Decision Processes and VAEs

We model the environment as a continuous, model-free Markov Decision Process (MDP) defined by the tuple (S, A, T, R, γ, H). A policy Π_t generates a trajectory $\tau = \{s_0, s_1, \dots\}$ of length H, starting from an initial state $s_0 \in S^{init}$. Because transition dynamics T and rewards R are unknown to the learner, we rely entirely on empirical trajectory sampling.

To model the state-space generatively, we employ a Variational Autoencoder (VAE) [15]. For an observation $s \in S$, we map it to a latent variable $z \in Z \subset \mathbb{R}^n$ using a Gaussian-approximated posterior $Q_\theta(z|s)$. A state trajectory τ thus maps to an encoded latent trajectory $\zeta \sim Q_\theta(\zeta|\tau)$.

2.2 Barrier-Certificates and Safety Assumptions

Instead of assigning safe/unsafe labels to individual states [30, 36], we evaluate safety over entire trajectories using an empirical indicator function $U(\tau) \in \{-1, 1\}$, where -1 denotes a safety-constraint violation. And to certify safety without strict dynamics, we utilize probably approximate barrier-certificates (PABCs) [30]. A barrier function $B : S \to \mathbb{R}$ classifies a state s as safe if $B(s) \geq 0$. We construct lower and upper-bound PABCs, $B_1(s)$ and $B_2(s)$ and their latent space counterparts $B_{1\phi}, B_{2\phi}$, tolerating constraint violations up to bounds ϵ_1 and ϵ_2 respectively, with user-defined confidence $\delta \in (0, 1)$.

We make the following standard assumptions [30]:

This work is licensed under a Creative Commons Attribution International 4.0 License.

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/PBUD3638

(1) Initial states $s_0 \in S^{init}$ exhibit safe behavior: $B_1(s_0) \geq 0$ and $B_2(s_0) \geq 0$ with probability 1.
(2) Stricter safety bounds yield tighter safe regions. If $\epsilon_1 < \epsilon_2$, then $S^{B_1} \subset S^{B_2} \subset S$.

2.3 Problem Formalization

Our objective is to compute the safety bounds ϵ_1, ϵ_2 alongside their corresponding latent barrier-certificates $B_{1\phi}, B_{2\phi}$ to guarantee safety at time t for a policy Π_t over trajectory distribution $\mathcal{T}_t$.

Problem 1. *Given confidence $\delta \in (0,1)$ and policy Π_t, find $\epsilon_1 \in (0,1)$ and $B_{1\phi}$ such that trajectories $\tau \sim \mathcal{T}_t$ and encoded trajectories $\zeta \sim Q_\theta(\zeta|\tau)$ violating safety constraints with $> \epsilon_1$ probability are correctly bounded:*

$$\mathbb{P}\left(B_{1\phi}(z) > 0 \mid \mathbb{P}(U(\tau) < 0) > \epsilon_1\right) < \delta\,,\ \forall z \in \zeta \tag{1}$$

Problem 2. *Similarly, find an upper-bound $\epsilon_2 \in (0,1)$ and $B_{2\phi}$ such that with confidence δ:*

$$\mathbb{P}\left(B_{2\phi}(z) > 0 \mid \mathbb{P}(U(\tau) < 0) < \epsilon_2\right) < \delta\,,\ \ \forall z \in \zeta \tag{2}$$

3 PROBABLY APPROXIMATE SAFE BARRIER-CERTIFICATES

To solve Problems 1 and 2, we formulate primal-dual chance-constrained programs (CCPs) to define latent-space regions where trajectories remain safe with high confidence. We parameterize convex barrier-certificates as linear functions, $B_{\phi j}(z) = \phi_{wj} \cdot z + \phi_{bj}$ (for $j \in \{1,2\}$).

3.1 Lower and Upper Bounds via Scenario Optimization

The lower-bound ($j = 1$) objective grows a conservative safe region $Z^{B_{\phi 1}}$ outward, while the upper-bound dual objective ($j = 2$) shrinks a generous estimate $Z^{B_{\phi 2}}$ inward. Utilizing convex scenario optimization [5, 30], we relax the computationally expensive CCPs into deterministic programs evaluated over N sampled trajectories:

$$\begin{aligned}
&\max_{\phi 1} \min_{z \in Z^{B_{\phi 1}}, z' \in \zeta} B_{\phi 1}(z') \quad \text{s.t.} \quad U(\tau_i) > 0 \\
&\min_{\phi 2} \max_{z \in Z^{B_{\phi 2}}, z' \in \zeta} B_{\phi 2}(z') \quad \text{s.t.} \quad U(\tau_i) < 0 \\
&\forall z \in \zeta \sim Q_\theta(\zeta|\tau_i),\ \ \forall i \in \{1, \ldots, N\}
\end{aligned} \tag{3}$$

To avoid bias, both barrier-certificates are constructed simultaneously using identical scenarios. Because these deterministic formulations can be overly conservative, we relax the problem further by permitting up to k_j constraint violations.

Relying on the theoretical guarantees for convex scenario optimization established by [5], and the associated corollaries within [30, 35], we can bound the N-fold probability of a violation exceeding ϵ_j with confidence $\delta \in (0,1)$. We explicitly restate the relevant corollary results below:

Corollary 3. *[30, 35] For $|\phi|$ optimization parameters, N scenarios and k observed violations, the empirical bound ϵ is:*

$$\epsilon \geq \frac{1}{N}\left(k + |\phi| - 1 + \sqrt{2k \ln\frac{1}{\delta} + 2(|\phi| - 1)\ln\frac{1}{\delta}}\right) \tag{4}$$

	Phase 1		*Phase 2*	
	Ant			
Algorithm	(ϵ_1, ϵ_2)	Av. Reward	$(\epsilon_1, \epsilon_2)^*$	Av. Reward
*Our Work**	0.11, 0.41	0.62	0.09, 0.18	0.94
Vanilla PPO	-	-	0.20, 0.45	0.81
Epsilon Greedy PPO	-	-	0.22, 0.48	0.77
Perturbed Exploration	-	-	0.12, 0.51	0.74
CoDE	-	-	0.08, 0.29	0.93
Genetic Curriculum	-	-	0.14, 0.45	0.90
Curriculum Adv. Training	-	-	0.09, 0.24	0.96
	CartPole			
*Our Work**	0.21, 0.48	0.80	0.10, 0.15	0.96
Vanilla PPO	-	-	0.21, 0.43	0.91
Epsilon Greedy PPO	-	-	0.19, 0.52	0.83
Perturbed Exploration	-	-	0.12, 0.51	0.62
CoDE	-	-	0.15, 0.29	0.97
Genetic Curriculum	-	-	0.06, 0.24	0.93
Curriculum Adv. Training	-	-	0.10, 0.19	0.95

Table 1: Comparing the PAC Guarantees on Policy Safety Between Algorithms

By substituting the respective recorded violations (k_1 for the lower-bound, k_2 for the upper-bound), Corollary 3 directly yields the bounds ϵ_1 and ϵ_2.

4 BARRIER-CERTIFICATE AIDED LEARNING

To reduce the width of the safety bound, we exploit the generative representation of the state-space provided by the VAE alongside the constructed barrier-certificates. Because the conservative safe region is strictly contained within the less conservative region ($Z^{B_{\phi 1}} \subset Z^{B_{\phi 2}}$), we must actively expose the agent to the tentatively unsafe states occupying the set difference between them. We propose a two-phase learning approach where the first phase learns the initial policy and the second phase refines it using these boundary states.

5 EXPERIMENTS AND CONCLUSIONS

We evaluate our two-phase methodology on the **Ant** and **CartPole** environments [32]. Phase 1 trains a Proximal Policy Optimization (PPO) policy [27, 28] alongside a VAE [15] to reconstruct pixel observations. After establishing safety error bounds (ϵ_1, ϵ_2) via barrier-certificates, Phase 2 refines the policy using our targeted, tentatively unsafe boundary states. We compare this refinement against baselines including vanilla PPO, ϵ-greedy exploration, transition noise, CoDE [12], Genetic Curriculum Learning [31] and Curriculum Adversarial Training [29]. Our method yields consistently tighter safety bounds across both environments. While Curriculum Adversarial Training [29] matches our bound tightness on the simpler CartPole task, our generative approach scales significantly better to the complex, higher-dimensional dynamics of Ant. Crucially, our model maintains comparable normalized reward performance to top baselines [12, 29], demonstrating enhanced rigorous policy robustness and safety verification without sacrificing overall learning efficiency.

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-017). This research is also supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme through the programme DesCartes.

REFERENCES

[1] Edoardo Bacci and David Parker. 2020. Probabilistic Guarantees for Safe Deep Reinforcement Learning. In *Formal Modeling and Analysis of Timed Systems: 18th International Conference, FORMATS 2020, Vienna, Austria, September 1–3, 2020, Proceedings* (Vienna, Austria). Springer-Verlag, Berlin, Heidelberg, 231–248. https://doi.org/10.1007/978-3-030-57628-8_14

[2] Osbert Bastani and Shuo Li. 2021. Safe Reinforcement Learning via Statistical Model Predictive Shielding. https://doi.org/10.15607/RSS.2021.XVII.026

[3] Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. 2017. Safe Model-based Reinforcement Learning with Stability Guarantees. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/766ebcd59621e305170616ba3d3dac32-Paper.pdf

[4] Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. 2022. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems* 5 (2022), 411–444.

[5] Marco C. Campi, Simone Garatti, and Maria Prandini. 2008. The Scenario Approach for Systems and Control Design. *IFAC Proceedings Volumes* 41, 2 (2008), 381–389. https://doi.org/10.3182/20080706-5-KR-1001.00065 17th IFAC World Congress.

[6] Richard Cheng, Gábor Orosz, Richard Murray, and Joel Burdick. 2019. End-to-End Safe Reinforcement Learning through Barrier Functions for Safety-Critical Continuous Control Tasks.

[7] William R Clements, Bastien Van Delft, Benoît-Marie Robaglia, Reda Bahi Slaoui, and Sébastien Toth. 2019. Estimating risk and uncertainty in deep reinforcement learning. *arXiv preprint arXiv:1905.09638* (2019).

[8] Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. 2017. Reverse curriculum generation for reinforcement learning. In *Conference on robot learning*. PMLR, 482–495.

[9] Pierre Fournier, Olivier Sigaud, Mohamed Chetouani, and Pierre-Yves Oudeyer. 2018. Accuracy-based curriculum learning in deep reinforcement learning. *arXiv preprint arXiv:1806.09614* (2018).

[10] Nathan Fulton and André Platzer. 2018. Safe Reinforcement Learning via Formal Methods: Toward Safe Control Through Proof and Learning.

[11] Akshita Gupta and Inseok Hwang. 2020. Safety Verification of Model Based Reinforcement Learning Controllers. arXiv:2010.10740 [cs.LG] https://arxiv.org/abs/2010.10740

[12] Izzeddin Gur, Natasha Jaques, Yingjie Miao, Jongwook Choi, Manoj Tiwari, Honglak Lee, and Aleksandra Faust. 2022. Environment Generation for Zero-Shot Compositional Reinforcement Learning. arXiv:2201.08896 [cs.LG] https://arxiv.org/abs/2201.08896

[13] Tom Haider, Felippe Schmoeller Roza, Dirk Eilers, Karsten Roscher, and Stephan Günnemann. 2021. Domain Shifts in Reinforcement Learning: Identifying Disturbances in Environments.. In *AISafety@ IJCAI*.

[14] John Jackson, Luca Laurenti, Eric Frew, and Morteza Lahijanian. 2020. Safety Verification of Unknown Dynamical Systems via Gaussian Process Regression. In *2020 59th IEEE Conference on Decision and Control (CDC)* (Jeju Island, Korea (South)). IEEE Press, 860–866. https://doi.org/10.1109/CDC42340.2020.9303814

[15] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[16] Morteza Lahijanian, Sean B. Andersson, and Calin Belta. 2015. Formal Verification and Synthesis for Discrete-Time Stochastic Systems. *IEEE Trans. Automat. Control* 60, 8 (2015), 2031–2045. https://doi.org/10.1109/TAC.2015.2398883

[17] Matthew Landers and Afsaneh Doryab. 2023. Deep Reinforcement Learning Verification: A Survey. *ACM Comput. Surv.* 55, 14s, Article 330 (July 2023), 31 pages. https://doi.org/10.1145/3596444

[18] Owen Lockwood and Mei Si. 2022. A review of uncertainty for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 18. 155–162.

[19] Yuping Luo and Tengyu Ma. 2022. Learning Barrier Certificates: Towards Safe Reinforcement Learning with Zero Training-time Violations. arXiv:2108.01846 [cs.LG] https://arxiv.org/abs/2108.01846

[20] Amir Modares, Nasser Sadati, Babak Esmaeili, Farnaz Adib Yaghmaie, and Hamidreza Modares. 2024. Safe Reinforcement Learning via a Model-Free Safety Certifier. *IEEE Transactions on Neural Networks and Learning Systems* 35, 3 (2024), 3302–3311. https://doi.org/10.1109/TNNLS.2023.3264815

[21] Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. 2020. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research* 21, 181 (2020), 1–50.

[22] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. 2017. Robust Adversarial Reinforcement Learning. arXiv:1703.02702 [cs.LG] https://arxiv.org/abs/1703.02702

[23] Stephen Prajna, Ali Jadbabaie, and George Pappas. 2005. Stochastic Safety Verification Using Barrier Certificates. *Proceedings of the IEEE Conference on Decision and Control* 1, 929 – 934 Vol.1. https://doi.org/10.1109/CDC.2004.1428804

[24] Mohit Prashant and Arvind Easwaran. 2022. PAC-Based Formal Verification for Out-of-Distribution Data Detection. In *2022 6th International Conference on System Reliability and Safety (ICSRS)*. 300–309. https://doi.org/10.1109/ICSRS56243.2022.10067329

[25] Mohit Prashant and Arvind Easwaran. 2022. PAC-Based Formal Verification for Out-of-Distribution Data Detection. In *2022 6th International Conference on System Reliability and Safety (ICSRS)*. IEEE, 300–309.

[26] Zengyi Qin, Kaiqing Zhang, Yuxiao Chen, Jingkai Chen, and Chuchu Fan. 2021. Learning Safe Multi-Agent Control with Decentralized Neural Barrier Certificates. arXiv:2101.05436 [cs.MA] https://arxiv.org/abs/2101.05436

[27] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. 2021. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research* 22, 268 (2021), 1–8. http://jmlr.org/papers/v22/20-1364.html

[28] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[29] Junru Sheng, Peng Zhai, Zhiyan Dong, Xiaoyang Kang, Chixiao Chen, and Lihua Zhang. 2022. Curriculum adversarial training for robust reinforcement learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[30] Arambam James Singh and Arvind Easwaran. 2024. PAS: Probably Approximate Safety Verification of Reinforcement Learning Policy Using Scenario Optimization. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (Auckland, New Zealand) *(AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1745–1753.

[31] Yeeho Song and Jeff Schneider. 2022. Robust reinforcement learning via genetic curriculum. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 5560–5566.

[32] Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. 2024. Gymnasium: A Standard Interface for Reinforcement Learning Environments. *arXiv preprint arXiv:2407.17032* (2024).

[33] Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. 2019. Programmatically Interpretable Reinforcement Learning. arXiv:1804.02477 [cs.LG] https://arxiv.org/abs/1804.02477

[34] Jingda Wu, Zhiyu Huang, and Chen Lv. 2022. Uncertainty-aware model-based reinforcement learning: Methodology and application in autonomous driving. *IEEE Transactions on Intelligent Vehicles* 8, 1 (2022), 194–203.

[35] Bai Xue, Martin Fränzle, Hengjun Zhao, Naijun Zhan, and Arvind Easwaran. 2019. Probably approximate safety verification of hybrid dynamical systems. In *Formal Methods and Software Engineering: 21st International Conference on Formal Engineering Methods, ICFEM 2019, Shenzhen, China, November 5–9, 2019, Proceedings 21*. Springer, 236–252.

[36] Linrui Zhang, Qin Zhang, Li Shen, Bo Yuan, Xueqian Wang, and Dacheng Tao. 2022. Evaluating Model-free Reinforcement Learning toward Safety-critical Tasks. arXiv:2212.05727 [cs.LG] https://arxiv.org/abs/2212.05727