# Supplementary material for PAS: Probably Approximate Safety Verification of Reinforcement Learning Policy Using Scenario Optimization

A Preprint

# Appendix

## 1 Experimental Setup

We performed all of our experiments on a Linux system with Ubuntu OS version 2022, a 20-core CPU with 64 GB memory. The Algorithm1 in the main text is implemented in Python3. We implement the barrier certificate using a linear function as $\mathcal{B}_\phi(s) = \phi_w \cdot s + \phi_b$, where $\phi = \{\phi_w, \phi_b\}$. We use the linear layer module of Pytorch to implement the linear function. Regarding the training of unsafe and safe policies for all three RL domains, we use stable-baseline3[Raffin et al.(2021)] implementation of PPO[Schulman et al.(2017)] algorithm.

| Parameters | Value | Description |
|---|---|---|
| $\delta$ | 1e-2 | Learning rate for gradient decent for barrier certificate parameter $\phi$. |
| lr | 1e-2 | Learning rate used in PPO for training *unsafe* and *safe* policies . |
| $H^{\text{safe-nav}}$ | 100 | Episode length of safe navigation environment. |
| $T^{\text{safe-nav}}_{\text{unsafe}}$ | 100k | Total number of training steps for unsafe policy in the safe navigation environment. |
| $T^{\text{safe-nav}}_{\text{safe}}$ | 500k | Total number of training steps for safe policy in the safe navigation environment. |
| $H^{\text{safe-mcar}}$ | 200 | Episode length of safe mountain car environment. |
| $T^{\text{safe-mcar}}_{\text{unsafe}}$ | 1M | Total number of training steps for unsafe policy in the safe mountain car environment. |
| $T^{\text{safe-mcar}}_{\text{safe}}$ | 5M | Total number of training steps for safe policy in the safe mountain car environment. |
| $H^{\text{safe-cpole}}$ | 200 | Episode length of safe cartpole environment. |
| $T^{\text{safe-cpole}}_{\text{unsafe}}$ | 300k | Total number of training steps for unsafe policy in the safe cartpole environment. |
| $T^{\text{safe-cpole}}_{\text{safe}}$ | 1M | Total number of training steps for safe policy in safe cartpole environment. |
| $m = \|\phi\|$ | 5 | Number of parameters of barrier certificate for safe navigation environment. |
| $m = \|\phi\|$ | 2 | Number of parameters of barrier certificate for safe mountain car environment. |
| $m = \|\phi\|$ | 2 | Number of parameters of barrier certificate for safe cartpole environment. |
| $\gamma$ | 0.95 | Discount factor used in training *unsafe* and *safe* policies. |
| $M$ | 300 | Total number of iterations in Algorithm 1. |

## References

[Raffin et al.(2021)] 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`

[Schulman et al.(2017)] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR* abs/1707.06347 (2017).