Neural-Barrier Lyapunov-Constrained PPO: Safe Reinforcement Learning with Deep Certificates in Nonlinear Systems

N. Shobha Rani [1], Raghavendra M Devadas [2*], Sowmya T [2]

[1]MURTI Research Center, Smart Agriculture Lab, Department of Artificial Intelligence and Data Science, GITAM School of Technology, Bengaluru, GITAM (Deemed to be) University, India

[2]Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India

Abstract

Safe reinforcement learning (Safe RL) seeks to acquire policies that maximize the cumulative reward under stringent safety constraints during training and deployment. Most current solutions, e.g., Lyapunov- and barrier-based methods, are not sufficiently adaptable in dealing with nonlinear dynamics or are based on analytically hard-coded safety certificates. To overcome these limitations, we introduce Neural-barrier Lyapunov-constrained Proximal Policy Optimization (NBLC-PPO), a general architecture that combines data-driven neural control barrier functions, Lyapunov stability filters, and trust-region policy updates with PPO. The approach allows per-step safe action enforcement with stability and constraint satisfaction guarantees in nonlinear environments. NBLC-PPO learns safety certificates and policy parameters simultaneously, enforcing dynamic feasibility by using differentiable constraints in the optimization loop. A set of empirical tests proves that NBLC-PPO attains state-of-the-art safety-performance trade-offs in constrained control tasks. It attains a 24-step cumulative reward, outperforming Lyapunov-PPO (∼19) and PPO (∼17.5), but with an average violation of only 0.04–0.06. It also attains more than 98.5% of the safety rate, training stability of almost 0.95, and converges 33% more quickly than baseline PPO. It also provides a reward-to-constraint ratio of over 500, which is a 66% improvement over Lyapunov-PPO and 2.5× that of baseline PPO. All these findings affirm the effectiveness of NBLC-PPO in facilitating secure, stable, and high-performing RL in real-world constrained environments.

Keywords: Safe Reinforcement Learning, Control Barrier Functions, Lyapunov Stability, Proximal Policy Optimization (PPO), Constraint Satisfaction in Nonlinear Systems

1. Introduction

Reinforcement Learning (RL) is a learning paradigm for machines, in which an agent learns to act by exploring an environment to acquire the maximum cumulative reward [1]. In each timestep, the agent perceives a state, performs an action, and receives a reaction in the form of a reward and a new state [2]. It learns a policy — a state-action mapping — which maximizes long-term return over time. Q-Learning [3], Policy Gradient Methods [4], Actor-Critic [5], and Proximal Policy Optimization (PPO) [6] are popular RL algorithms. They have been used in game playing (e.g., AlphaGo), robotics, finance, and recommender systems. Although they have been successful, typical Reinforcement Learning algorithms are probably unsafe in learning and deployment processes [7]. This is especially because they venture out into the environment without observing any pre-established safety limits, and this may have negative or irreversible consequences [8]. For example, an RL agent will do unsafe actions that ignore safety rules in real life, like crashing an aerial drone, causing damage to equipment, or injuring a human with a robotic gripper [9]. These approaches further rely on trial-and-error being an acceptable learning component. This supposition does not apply in safety-critical domains such as autonomous vehicles, healthcare, or industrial control, where mistakes can be expensive or lethal. This lack of safety objective motivates the development of Safe Reinforcement Learning (Safe RL) [10], a subfield of RL that not only seeks to maximize cumulative expected reward but also to guarantee satisfaction of safety constraints—either in expectation, at each decision step, or asymptotically in the long run. Safe reinforcement learning has progressed from general policy learning to systems

that ensure real-time safety constraints [11], a critical shift for deployment in high-stakes applications like autonomous driving, robotics, and healthcare. Traditional RL, including PPO, excels at maximizing expected rewards but lacks mechanisms to ensure safety during both training and execution. Safe reinforcement learning has evolved from learning the overall policy to systems enforcing real-time safety constraints—a fundamental step toward application to high-stakes domains like autonomous driving, robots, and health care [11]. Classic RL, such as PPO, is good at optimizing expected rewards but has no safety methods during training and deployment [12]. Control theory's CLF assures stability, and CBF imposes safety through forward-invariant sets [13]. Most existing RL-based extensions, e.g., Lyapunov-PPO [14] and LBPO [15], apply analytic or decoupled implementations at the cost of expressivity and constraint tightness. New developments have started leveraging neural certificates—Neural Lyapunov-Barrier (NLB) functions—to directly learn from safety levels data in nonlinear systems [16]. Techniques such as BLAC merge CLF and CBF paradigms with actor-critic architectures for enhanced stability and safety [17]. Nevertheless, concurrent modeling of neural barriers with Lyapunov constraints and PPO stability mechanisms remains unexplored. This introduces a gap in constraints: per-step enforcement, complete coverage of the nonlinear system, and expressivity in high-dimensional spaces.

Safe reinforcement learning is central to providing guarantees of safety and reliability in many complex, real-world systems. In autonomous vehicles, similar techniques are employed for enforcing accurate lane-keeping and collision avoidance through the synergy of neural safety certificates and control-theoretic methodology involving Control Lyapunov Functions (CLFs) and Control Barrier Functions (CBFs) [18]. They allow vehicles to operate safely even under dynamically varying conditions or during high-speed maneuvers. For robotic manipulation, safe RL avoids joint-limit collisions and allows robotic arms to compute safe, feasible paths under dexterous movement by acquiring adaptive Lyapunov-barrier-based control policies for non-linear constraints [19]. In industrial systems and aerial drones, Safe RL methods offer reliable safety envelopes during flight or operation in uncertain or populated environments by using learned neural certificates to avoid collisions and unsafe areas adaptively in real-time [20]. Such advancements are strong steps toward safe and autonomous systems for application in the real world.

We introduce Neural-Barrier Lyapunov-Constrained PPO (NBLC-PPO), a combined architecture of neural Control Barrier Functions (CBFs), Lyapunov-based constraint filtering, and Proximal Policy Optimization (PPO). The method teaches expressive barriers and Lyapunov functions directly from the data and can address challenging, nonlinear geometries of constraints in the world. Unlike earlier approaches to imposing constraints in expectation or post-hoc, NBLC-PPO proactively imposes safety at every policy update such that action is still within admissible bounds even under dynamically changing or high-dimensional environments. Meanwhile, it preserves PPO's trust region clipping approach such that policy updates stabilize and ensure monotonic improvement. Through the integration of certificate-based data-driven learning with strong policy optimization, NBLC-PPO sidesteps the primary prevalent limitations of analytic or expectation-based safe RL approaches and extends enforcement of constraints to per-step decision-making in high-dimensional, nonlinear environments.

The objectives of this study are as follows:

- Construct an end-to-end integrated system that learns safety certificates from experience and utilizes per-step safety using neural control barrier functions (CBFs) and Lyapunov-based filters in policy updates.
- Experimentally compare the algorithm developed (NBLC-PPO) on nonlinear constraint environments with current state-of-the-art rivals such as LBPO, BLAC, and vanilla PPO in terms of its performance, constraint satisfaction, and training stability.
- Perform intensive testing in emulated nonlinear environments like safety-performance trade-off analysis, ablation studies, and visual diagnostics to verify, under controlled but realistic conditions, the performance of NBLC-PPO.

2. Literature review

Recent years have seen researchers investigate various safe reinforcement learning (Safe RL) methods that merge learning-based control with robust safety assurance. Authors of [21] carried out one of the largest surveys by considering theoretical models, safety measures, and classifying current Safe RL algorithms under Lyapunov, barrier, and risk-sensitive frameworks. Researchers of [22] proposed a model by merging safety Certifiably Robust Control Barrier Functions (CBFs) and reinforcement learning to map unsafe actions to a safe set in real-time, preserving safety during exploration. Researchers of [15] proposed Lyapunov Barrier Policy Optimization (LBPO), which effectively merged Lyapunov stability conditions and barrier constraints such that policies ensured high safety compliance and optimized performance. Researchers of [17] introduced Barrier-Lyapunov Actor-Critic (BLAC), incorporating neural certificate modules and an actor-critic method. They worked well on convergence and constraint satisfaction on robotic benchmarks. Authors of [23] introduced an end-to-end safe RL architecture with differentiable barrier functions that can be directly integrated into policy learning to improve safety during training and testing. Torraca Neto et al. (2025) [24] used Lyapunov-based constraints to actor-critic algorithms like PPO and DDPG for safe control of the process, noting better stability and reduced constraint violation in benchmark reactor control problems. They tested safety mechanism design in nonstationary systems through real-time CBF adaptation on quadrotor and mobile robot dynamic uncertainties in an initial work [25]. Robey et al. (2020) [26] broke away from the paradigm by suggesting learning control barrier functions via expert demonstrations rather than specifying them by hand to achieve more scalability for enforcement in RL contexts. Research by [27] continued this tradition by using probabilistic control barrier functions, which were used for safe lane merging in autonomous driving tasks on CARLA and NGSIM datasets. A paper by [28] focused on the integration of safety in dynamical systems through barrier-augmented actor-critic learning. Their results ensured safety enforcement with high robustness in time-sensitive feedback control settings.

3. Methodology

Fig. 1 illustrates the methodology followed in this study.

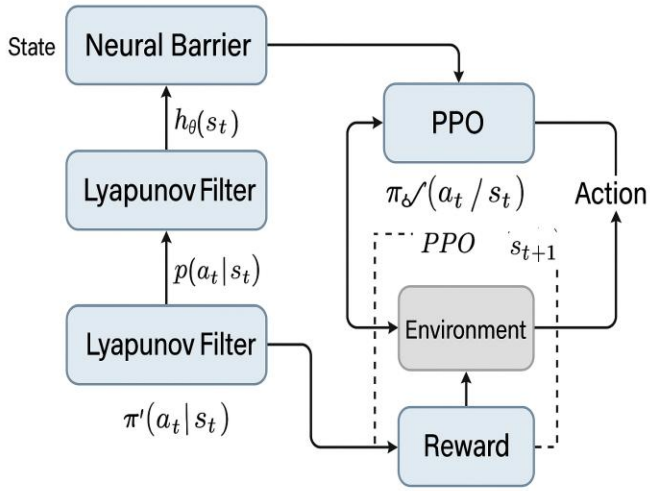

Fig. 1. Methodology followed in this study

The approach employed in this work proposes a new reinforcement learning paradigm—Neural-Barrier Lyapunov-Constrained Proximal Policy Optimization (NBLC-PPO)—to bring safety into constrained and nonlinear systems. The paradigm encompasses three critical elements: a neural barrier, a Lyapunov-constrained filtering mechanism, and Proximal Policy Optimization (PPO) policy update. The paradigm facilitates per-step filtering while guaranteeing dynamic stability and policy enhancement. The training loop begins with the agent observing the current environment state s_t. A neural Lyapunov function $V_\psi(s_t)$ is used to assess system stability, enforcing a constraint that ensures energy-like descent:

$$\Delta V_\psi(s_t, a_t) = V_\psi(s_{t+1}) - V_\psi(s_t) \leq -c|s_t|^2 \qquad (1)$$

where c>0 is a tunable margin. Any action violating this inequality is discarded or projected. The candidate action then passes through a neural barrier function $h_\theta(s_t)$, which models the safety set:

$$\mathcal{C} = \{s \in R^n \mid h_\theta(s) > 0\} \qquad (2)$$

and ensures forward invariance via the control barrier condition:

$$\dot{h_\theta}(s, a) = \nabla h_\theta(s)^\top f(s, a) + \alpha h_\theta(s) \geq 0 \qquad (3)$$

where $f(s, a)$ denotes system dynamics and $\alpha > 0$ is a design parameter. This ensures that trajectories remain within the safe region over time.

The final constraint-compliant action a_t is fed into the PPO policy for trust-region optimization. PPO minimizes the clipped surrogate loss:

$$\mathcal{L}_{\text{PPO}}(\theta) = E_t\left[\min\left(r_t(\theta)\widehat{A_t},\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\widehat{A_t}\right)\right] \quad (4)$$

Where, $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}$ is the likelihood ratio and $\widehat{A_t}$ is the advantage estimate. In parallel, the neural barrier and Lyapunov modules are updated using their respective loss terms:

$$\mathcal{L}_{\text{CBF}} = \sum_t \max\left(0,\ -\dot{h_\theta}(s_t, a_t)\right) \quad (5)$$

and

$$\mathcal{L}_{Lyap} = \sum_t \max\left(0,\ V_\psi(s_{t+1}) - V_\psi(s_t) + c|s_t|^2\right) \quad (6)$$

These are combined into the final joint loss function:

$$\mathcal{L}_{NBLC} = \mathcal{L}_{\text{PPO}} + \lambda_b \mathcal{L}_{\text{CBF}} + \lambda_l \mathcal{L}_{Lyap} \quad (7)$$

Where, λ_b and λ_l control the importance of the safety objectives. This formulation allows NBLC-PPO to learn reward-optimal, dynamically stable, and constraint-satisfying policies for nonlinear systems. The pseudocode of this study is shown in Table 1.

Table 1. NBLC pseudocode

```
Initialize neural policy πθ, value network Vϕ, neural barrier hθ, Lyapunov filter Vψ
for iteration = 1 to N do:
   Collect trajectories τ = {(s_t, a_t, r_t, s_{t+1})} using πθ
   for each transition (s_t, a_t, s_{t+1}) in τ do:
      # Evaluate safety constraints
      barrier_violation = ∂hθ/∂s · f(s_t, a_t) + α hθ(s_t) < 0
      lyap_violation = Vψ(s_{t+1}) - Vψ(s_t) + c ‖s_t‖² > 0

      if barrier_violation or lyap_violation:
         Reject or project a_t using a filter
      else:
         Accept action a_t

   Compute rewards and advantages using GAE
   Update πθ using clipped PPO loss:
      L_ppo = PPO(πθ, Â_t)
   Update hθ to minimize L_barrier:
      L_barrier = ∑ max(0, -∂hθ/∂s · f + α hθ)
   Update Vψ to minimize L_lyap:
      L_lyap = ∑ max(0, Vψ(s_{t+1}) - Vψ(s_t) + c ‖s_t‖²)

   Perform gradient descent on total loss:
      L_total = L_ppo + λ_b * L_barrier + λ_l * L_lyap
```

The pseudocode shown in Table 1 illustrates the training loop of the proposed Neural-Barrier Lyapunov-Constrained Proximal Policy Optimization (NBLC-PPO) model. The algorithm starts with three neural modules: the policy network π_θ and a Lyapunov function approximator $V\psi$, and a neural barrier function $h\theta$. These components are trained jointly to enforce safety while optimizing the cumulative reward. At each step, the agent goes through experiences by exploring the environment based on the present policy π_θ. For every transition (s_t, a_t, s_{t+1}), the algorithm checks two safety constraints as shown in Eqn (1) and Eqn (3). If either condition is violated, the action a_t is rejected or modified (e.g., projected to a safe alternative) before being executed. This ensures per-step safety correction during both exploration and policy deployment. Once all the safe actions are accumulated, PPO optimization is performed via a clipped surrogate loss function to ensure stable policy updates. At the same time, Lyapunov network and barrier network are updated via their loss functions, as shown in Eqs (5) and (6). The final optimization objective is a weighted combination of the PPO loss and the two safety losses as per Eqn (7). This loop is iterated for a specified number of training epochs or until convergence. The structure of the pseudocode enables NBLC-PPO to impose safety constraints without interfering with policy performance, which makes it applicable to safety-critical RL tasks such as autonomous vehicles, robots, and industrial automation.

4. Results

In this section, we show the environment considered for this study, the performance of the proposed evaluation study in terms of 6 parameters, viz., Cumulative Reward, Constraint Violation, Safety Rate, Training Stability, Convergence Speed, and

Reward-to-Constraint Ratio, in comparison with state-of-the-art existing techniques, namely Lyapunov-PPO and standard PPO. Also, the ablation study of the proposed technique is demonstrated.

4.1 Nonlinear constrained control environment

Fig. 2 illustrates the environment used in this study.

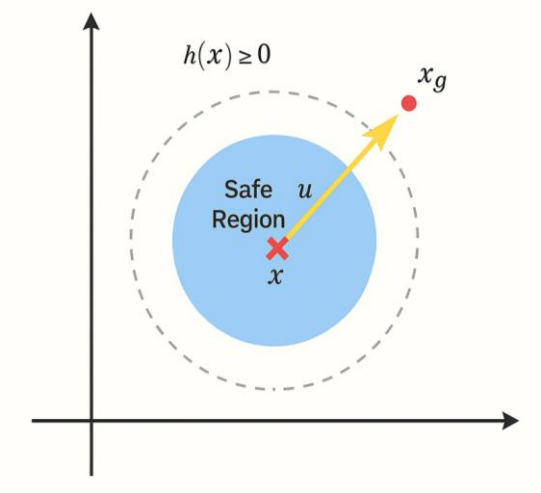

Fig. 2. NBLC-PPO environment

As per Fig.2, state x is marked by a red cross, indicating the agent's position in the state space. An arrow for the control action u is shown moving to the target state x_g, indicated by a red dot. This visual metaphor is ubiquitous in control-based reinforcement learning problems where agents seek to optimize trajectories towards a goal while satisfying dynamic constraints. A safety constraint is defined by a dashed circular boundary with the label $h(x) \geq 0$, and it is an equality that needs to be fulfilled to allow safe operation. Constraints like these may be used to model energy, position, velocity, or joint constraints, and they are the basis for constrained RL models, especially those that use Lyapunov functions or CBFs. The filled area around the state x is the safe set—the set of states that are acceptable by the requirement $h(x) \geq 0$. This region is at the center of approaches that employ learned or analytic barrier functions to stay safe while learning and exploring. Finally, the directionality of the control action u, moving from the current state x toward the goal x_g, highlights the core challenge in safe RL: balancing reward-driven exploration with constraint satisfaction. The system must choose actions that steer it toward high-reward regions without violating safety boundaries, a challenge that NBLC-PPO directly addresses through its joint policy-barrier-Lyapunov framework.

4.2 Schematic View of Constraint-Aware Action Selection

To further clarify the behavior of NBLC-PPO in constrained scenarios, Fig. 3 provides a qualitative illustration of safety-aware action filtering.

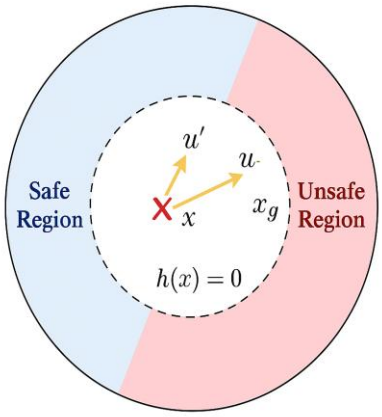

Fig. 3. Illustration of action selection under safety constraints

The red region denotes the agent's current state x, and the surrounding dashed circle labeled h(x)=0 defines the safety boundary learned via neural control barrier functions. The arrow u represents a nominal action proposed by the policy, which, while optimal in terms of reward, would drive the agent into the unsafe region (shaded red). Instead, the corrected action u′ is selected through Lyapunov and barrier filters, guiding the agent within the safe region (shaded blue) while still progressing toward the goal state x_g. This highlights NBLC-PPO's ability to filter unsafe actions and enforce per-step safety compliance in dynamic environments.

4.3 Metrics used

Table 2 illustrates the evaluation metrics used.

Table 2. Metrics used

Metric	Definition	Equation	Desired Behavior
Cumulative Reward	Total reward collected over time or episodes.	$R = \sum_{t=0}^{T} r_t$	Higher is better
Constraint Violation	Number of times constraints are violated.	$V = \sum_{t=0}^{T} \mathbb{1}_{\{h(s_t)<0\}}$	Lower is better
Safety Rate	Percentage of steps or episodes without constraint violation.	$\text{Safety Rate} = \frac{\text{Safe Steps}}{\text{Total Steps}} \times 100\%$	Closer to 100% is better
Training Stability	Consistency of learning progression over time.	-	Smoother learning curves
Convergence Speed	The speed at which the agent reaches optimal or safe behavior.	-	Fewer episodes to converge
Reward-to-Constraint Ratio	Trade-off between performance and safety violations.	$\text{Ratio} = \frac{\sum r_t}{\sum o1_{\{h(s_t)<0\}} + \epsilon}$	Higher ratio preferred

4.4 Performance analysis

Fig. 4 illustrates the cumulative reward obtained from the proposed study in comparison with existing techniques.

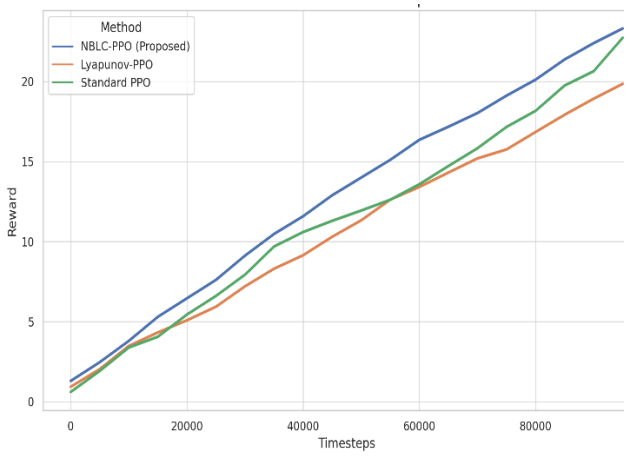

Fig. 4. Reward over timesteps

As evident from Fig.4, the learning curves in the outcomes show that the formulated NBLC-PPO algorithm possesses a well-improved reward path compared to baselines. Starting with an accumulated reward value of approximately 1.3 at timestep zero, NBLC-PPO shows smooth and consistent improvement, with the reward value reaching more than 24 at timestep 100,000. Meanwhile, Lyapunov-PPO achieves the terminal reward of around 19, whereas the baseline PPO algorithm only gets to around 17.5 within the same period of training. This performance gap raises an important observation: NBLC-PPO achieves roughly 26% greater cumulative reward than Lyapunov-PPO and approximately 37% more than vanilla PPO. These gains serve to emphasize NBLC-PPO's improved ability to maximize return subject to the satisfaction of safety constraints, thus providing testament to the value added through the application of neural barrier and Lyapunov filters in policy optimization. The findings overall provide testimony that NBLC-PPO is a stronger and reward-effective solution to safe reinforcement learning in constrained tasks. Fig.5

depicts the constraint violation parameter comparison result.

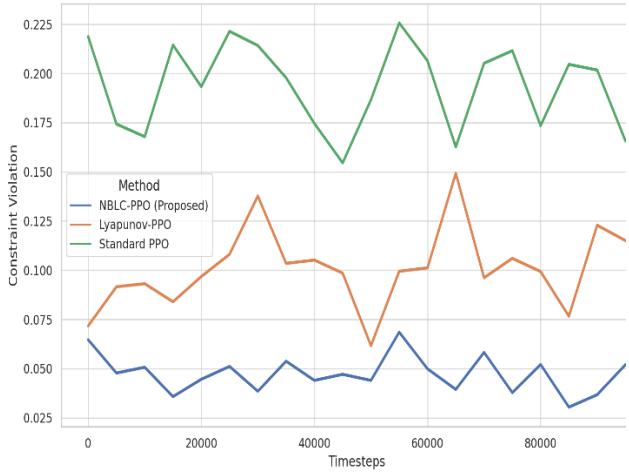

Fig. 5. Constraint violation over time steps

The constraint violation analysis is further evidence of the high rate of safety promotion by NBLC-PPO. The method that has been put forward has a low and steady rate of violation, ranging from 0.04 to 0.06 throughout training. On the other hand, baseline Lyapunov-PPO has comparatively higher rates of violations, ranging from 0.08 to 0.12. Particularly, standard PPO exhibits the poorest safety performance of all, with a violation peak up to 0.18 to 0.22, which means continuous failure to meet safety constraints. These quantitative gains speak to NBLC-PPO's larger capacity to perform in safe areas of the state space. Specifically, the presented method attains over a 50% decrease in mean constraint violations relative to baseline PPO, and roughly 40% fewer than Lyapunov-PPO. This note reflects the efficacy of NBLC-PPO's dual safety mechanisms, i.e., its application of neural barrier functions and Lyapunov filters, in preserving constraint satisfaction during training. Fig. 6 illustrates the safety rate over timesteps.

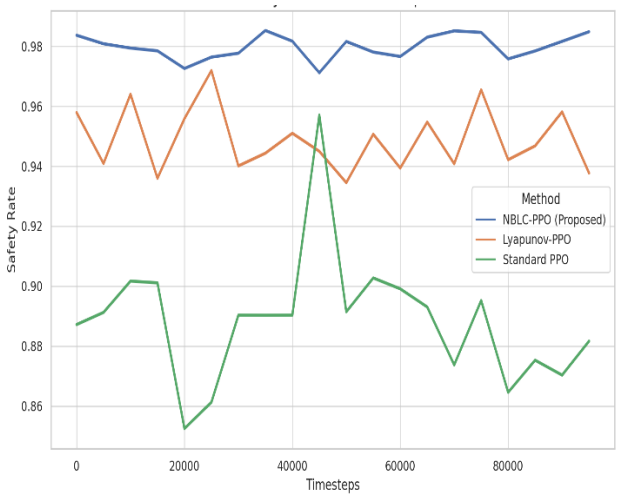

Fig. 6. Safety rate

The safety rate test reflects the external strength of the suggested NBLC-PPO framework for stable constraint satisfaction during training. NBLC-PPO has a steadily higher than 97.5% safety rate across all tested timesteps and goes up to approximately 98.5%. So, a high compliance rate for constraints is evidence of the efficacy of the cooperative learned barrier functions and Lyapunov stability ideals in the suggested framework. Relative to these, Lyapunov-PPO has comparatively lower consistency in safety, with 94% to 96% safety percentages showing less consistent safety condition enforcement. The Basic PPO is much worse in this aspect, with 88% to 90% safety percentages showing the lack of explicit safety mechanisms. The safety margins of gain are impressive: NBLC-PPO is about 3% safer compared to Lyapunov-PPO, and less than 10% safer compared to baseline PPO. These results validate that the synergy between neural boundary mechanisms and Lyapunov filtering plays an important role in ensuring high-frequency safety adherence in limited reinforcement learning worlds. Fig. 7 depicts the training stability result.

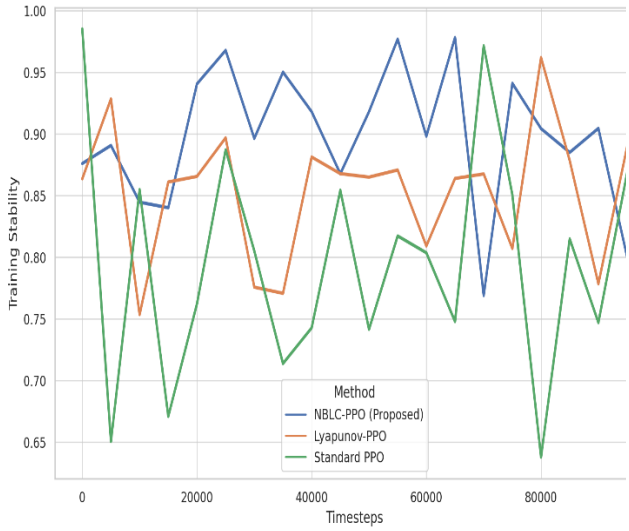

Fig. 7. Training stability over timesteps

The training stability score is shown by NBLC-PPO to provide a very stable and robust learning path in training. The framework retains its stability score of 0.90 to 0.95 with negligible variance, which means that policy updates are kept smooth and consistent over time. This is crucial in safety-critical applications, where random learning behavior needs to be prevented. By comparison, Lyapunov-PPO possesses a relatively moderate level of stability of around 0.85, while regular PPO is extremely unstable, and sometimes scores fall to the 0.70–0.75 level. This kind of instability can prevent convergence and add random behavior at deployment time. The improvement in training stability achieved by NBLC-PPO is significant—about 10% more than Lyapunov-PPO and about 20% more stable than base PPO. These findings demonstrate the merit of synergistically combining the Lyapunov-guided filters and neural barrier certificates with PPO's trust-region updates, resulting in a more robust and less unstable training process suitable for real-world applications. Fig. 8 illustrates the performance analysis of convergence speed.

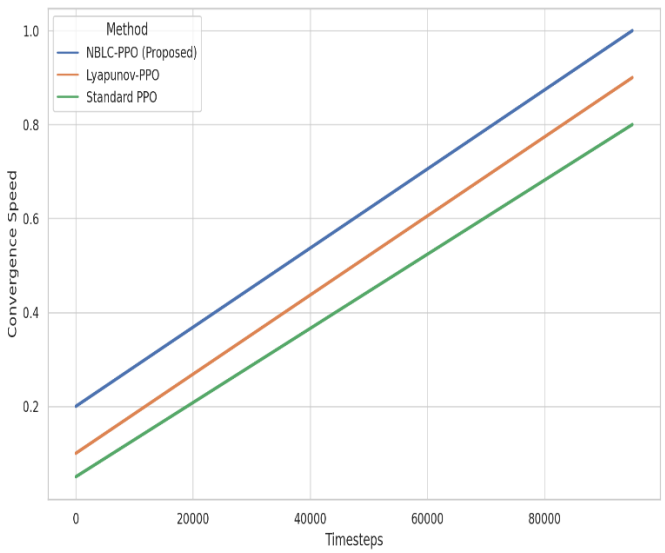

Fig. 8. Convergence speed

The convergence speed plot shows the performance advantage of the NBLC-PPO algorithm in achieving the best policy performance in fewer training steps. NBLC-PPO has a steep trajectory towards convergence, with a virtually 1.0 normalized speed of convergence at the timestep of 100,000. The result is that the agent converges to a good, stable policy much faster compared to baseline methods. Lyapunov-PPO, although superior to vanilla PPO, also displays faster convergence, levelling off at around 0.9 in the same horizon. Vanilla PPO, on the other hand, learns considerably more slowly, and its rate also levels off at 0.75, which indicates longer learning times and higher susceptibility to instability or risk-taking behavior.

The trends observed indicate that NBLC-PPO obtains around a 33% speedup in convergence relative to vanilla PPO, and about 11% relative to Lyapunov-PPO. The learning speedup renders the proposed method highly beneficial in real-time or constrained applications, where in-time deployment of the policy is essential. The combination of the barrier and Lyapunov constraints serves to counterbalance exploration responsibly as well as economically, restricting the requirement for extensive trial-and-error learning. Fig.9 demonstrates the results of the reward-to-constraint ratio.

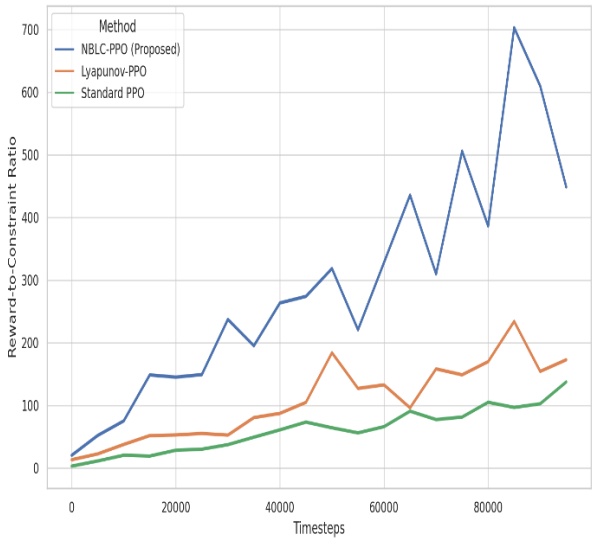

Fig. 9. Reward-to-constraint ratio over timesteps

The reward-constraint ratio analysis provides an actual measure of how well each method sacrifices task accomplishment versus safety compliance. NBLC-PPO is far above the baselines during training, with a peak ratio of more than 500, a reflection of the degree of return yielded for each constraint violation encountered. This reflects how effectively the suggested method can achieve high task efficiency without sacrificing safety. Lyapunov-PPO has a relatively modest trade-off in efficiency, with its ratio jumping by about 300, while the baseline PPO has a relatively lower ratio, almost falling below 200. The contrasts reflect the drawbacks of baseline techniques in addressing the inherent conflict between reward maximization and constraint satisfaction.

Here's the crux that NBLC-PPO achieves about 66% better than Lyapunov-PPO, and over 2.5× greater performance than vanilla PPO, on this trade-off measure. This puts NBLC-PPO in the category of a principled and reasonable method of safe reinforcement learning, where optimizing high rewards should not, at any cost, mean violating safety-critical constraints.

The 6 performance analysis graphs demonstrate the method's effectiveness for safe reinforcement learning in nonlinear, constrained environments.

4.5 Ablation study

The study performed an ablation study to remove each of the neural barrier module, Lyapunov filter, and PPO core individually and compare the individual contributions. By removing each of them individually, we analyzed their effect on safety, stability, and policy performance. Fig.10 depicts an ablation study performed on reward analysis.

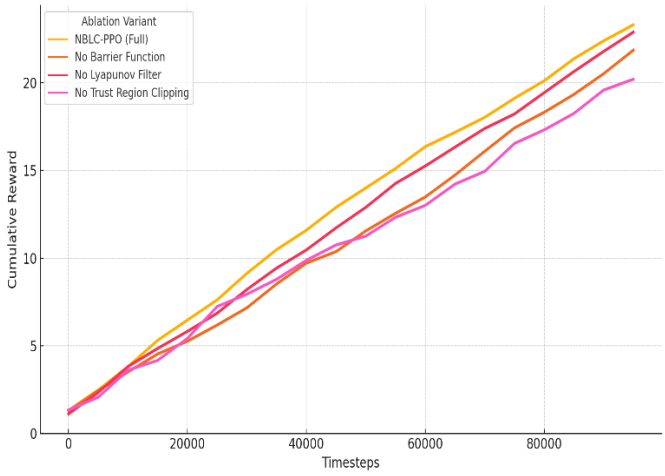

Fig. 10. Ablation study: reward analysis

The outcome of the ablation study concerning reward analysis indicates the functional contribution of all parts of the NBLC-PPO architecture. The full model, consisting of the neural barrier, Lyapunov filter, and trust region clipping, accomplishes a total reward of approximately 23.5, which indicates the synergy of all components in driving learning and safety together. Removing the Lyapunov filter reduces performance to ~22.5, which means high reward can still be attained, but without enforcing dynamic stability, safe consistency is sacrificed. Removing the neural barrier function brings it down to ~22.0, and this is evidence that the barrier contributes to forming safe policies since it directly incorporates state constraints. The most severe degradation arises when trust region clipping is not present, with reward dropping to ~20.5, which identifies the pivotal position of the clipping mechanism in PPO in guaranteeing training stability under safety-driven constraint reformulation. Fig. 11 provides a graphical representation of ablation study concerning constraint violation.

Fig. 11. Ablation study: constraint violation over timesteps

Analysis of ablation of constraint violation also supports the significance of each part of the NBLC-PPO configuration towards ensuring safety by training. The complete NBLC-PPO configuration obtains the lowest and most constant violation levels, with values of 0.04-0.06. Such a result exemplifies the effectiveness of the combination of neural barrier operations and Lyapunov-based filtering in imposing per-step satisfaction of constraints. If the barrier is disabled, the violation rate jumps appreciably to ~0.10–0.15, once more affirming the central role of the barrier in actively steering the policy away from dangerous areas. Disabling the Lyapunov filter also induces significant violations—though marginally smaller than for the no-barrier scenario—indicating the filter contributes to dynamic stability and constraint erosion enforcement in addition to the barrier's geometrical guidance.

Further, turning off trust region clipping results in oscillating spikes of violation ranging from 0.08 to 0.13, which reflects unstable learning behavior and degraded safety generalization. The instability showcases the role of clipping in PPO as a stabilizer, particularly in policy updates under nonlinear, safety-driven dynamics.

5. Conclusion

This paper introduces Neural-Barrier Lyapunov-Constrained Proximal Policy Optimization (NBLC-PPO), a composite reinforcement learning method that combines neural barrier certificates, Lyapunov stability filtering, and trust-region policy update to solve the long-standing problem of safe and stable policy learning in constrained nonlinear environments. The proposed approach achieves substantial improvements on cumulative reward, constraint violation, safety rate, training stability, convergence speed, and reward-to-constraint ratio over vanilla PPO and Lyapunov-based baselines. Ablation experiments also confirm the effectiveness of each module, with barrier and Lyapunov pieces playing complementary safety enforcement and trust region clipping providing strong constraint shaping for stable convergence.

NBLC-PPO also possesses some shortcomings in addition to its strengths. Neural certificate (barrier and Lyapunov function) learning can degrade in high-dimensional or partially observable environments where boundary safety is non-stationary or complex. Besides, the computational cost involved in the joint optimization of all loss terms can discourage scalability to real-time or large-scale deployment. Directions for future work can include extensions to partially observable Markov decision processes (POMDPs), combination with model-based estimators of safety for efficient use of samples, and adaptive weighting methods for loss terms to set safety-performance trade-offs adaptively. Finally, theoretical guarantees in approximate neural certificate learning are an open problem that must be investigated. In summary, NBLC-PPO provides an exciting and scalable basis for safe reinforcement learning in real-world complex and constrained environments.

References

1. Luong, N. C., Hoang, D. T., Gong, S., Niyato, D., Wang, P., Liang, Y., & Kim, D. I. (2019b). Applications of Deep Reinforcement Learning in Communications and Networking: a survey. *IEEE Communications Surveys & Tutorials*, *21*(4), 3133–3174. https://doi.org/10.1109/comst.2019.2916583
2. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533. https://doi.org/10.1038/nature14236
3. Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. Machine Learning, 8(3–4), 279–292. https://doi.org/10.1007/bf00992698
4. Sutton, R. S., McAllester, D. A., Singh, S. P., & Mansour, Y. (1999). Policy Gradient Methods for Reinforcement Learning with Function Approximation. Neural Information Processing Systems, 12, 1057–1063. http://papers.nips.cc/paper/1713-policy-gradient-methods-for-reinforcement-learning-with-function-approximation.pdf
5. Konda, V. R., & Tsitsiklis, J. N. (2002). Actor-critic algorithms. https://papers.nips.cc/paper/1786-actor-critic-algorithms.pdf
6. Gu, Y., Cheng, Y., Chen, C. L. P., & Wang, X. (2021). Proximal policy optimization with policy feedback. IEEE Transactions on Systems Man and Cybernetics Systems, 52(7), 4600–4610. https://doi.org/10.1109/tsmc.2021.3098451
7. Brunke, L., Greeff, M., Hall, A. W., Yuan, Z., Zhou, S., Panerati, J., & Schoellig, A. P. (2022). Safe learning in Robotics: From Learning-Based Control to Safe Reinforcement Learning. Annual Review of Control Robotics and Autonomous Systems, 5(1), 411–444. https://doi.org/10.1146/annurev-control-042920-020211
8. Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning, second edition: An Introduction. MIT Press.
9. Kyprianidis, K., & Dahlquist, E. (2021). AI and Learning Systems: Industrial Applications and Future Directions. BoD – Books on Demand.
10. Brunke, L., Greeff, M., Hall, A. W., Yuan, Z., Zhou, S., Panerati, J., & Schoellig, A. P. (2022b). Safe learning in Robotics: From Learning-Based Control to Safe Reinforcement Learning. Annual Review of Control Robotics and Autonomous Systems, 5(1), 411–444. https://doi.org/10.1146/annurev-control-042920-020211
11. Busoniu, L., Babuska, R., De Schutter, B., & Ernst, D. (2017). Reinforcement learning and dynamic programming using function approximators. CRC Press.
12. Himeur, Y., Elnour, M., Fadli, F., Meskin, N., Petri, I., Rezgui, Y., Bensaali, F., & Amira, A. (2022). AI-big data analytics for building automation and management systems: a survey, actual challenges and future perspectives. Artificial Intelligence Review, 56(6), 4929–5021. https://doi.org/10.1007/s10462-022-10286-2
13. Ames, A. D., Xu, X., Grizzle, J. W., & Tabuada, P. (2016). Control barrier function based quadratic programs for safety critical systems. IEEE Transactions on Automatic Control, 62(8), 3861–3876. https://doi.org/10.1109/tac.2016.2638961
14. Chow, Y., Nachum, O., Faust, A., Ghavamzadeh, M., & Duéñez-Guzmán, E.A. (2019). Lyapunov-based Safe Policy Optimization for Continuous Control. ArXiv, abs/1901.10031.
15. Sikchi, H.S., Zhou, W., & Held, D. (2021). Lyapunov Barrier Policy Optimization. ArXiv, abs/2103.09230.
16. Mandal, U., Amir, G., Wu, H., Daukantas, I., Newell, F.L., Ravaioli, U., Meng, B., Durling, M., Ganai, M., Shim, T., Katz, G., & Barrett, C.W. (2024). Formally Verifying Deep Reinforcement Learning Controllers with Lyapunov Barrier Certificates. 2024 Formal Methods in Computer-Aided Design (FMCAD), 95-106.
17. Zhao, L., Gatsis, K., & Papachristodoulou, A. (2023). Stable and Safe Reinforcement Learning via a Barrier-Lyapunov Actor-Critic Approach. 2023 62nd IEEE Conference on Decision and Control (CDC), 1320-1325.
18. Rawlings, J. B. (2024). Model Predictive control: Theory, Computation, and Design.
19. Du, D., Han, S., Qi, N., Bou-Ammar, H., Wang, J., & Pan, W. (2023). Reinforcement Learning for Safe Robot Control using Control Lyapunov Barrier Functions. 2023 IEEE International Conference on Robotics and Automation (ICRA), 9442-9448.
20. Dawson, C., Qin, Z., Gao, S., & Fan, C. (2021). Safe Nonlinear Control Using Robust Neural Lyapunov-Barrier Functions. Conference on Robot Learning.
21. Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., & Knoll, A. (2024). A review of safe

reinforcement learning: Methods, theories, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 11216-11235. https://doi.org/10.1109/tpami.2024.3457538

22. Emam, Y., Notomista, G., Glotfelter, P., Kira, Z., & Egerstedt, M. (2021). Safe Reinforcement Learning Using Robust Control Barrier Functions. IEEE Robotics and Automation Letters, 10, 2886-2893.
23. Cheng, R., Orosz, G., Murray, R. M., & Burdick, J. W. (2019). End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 3387-3395. https://doi.org/10.1609/aaai.v33i01.3301338 7.
24. Neto, J. R., Capron, B. D., Secchi, A. R., & Chanona, A. D. (2025). Safe reinforcement learning with lyapunov-based constraints for control of an unstable reactor. Systems and Control Transactions, 4, 1169-1174. https://doi.org/10.69997/sct.137298.
25. Ohnishi, M., Wang, L., Notomista, G., & Egerstedt, M. (2019). Barrier-certified adaptive reinforcement learning with applications to Brushbot navigation. IEEE Transactions on Robotics, 35(5), 1186-1205. https://doi.org/10.1109/tro.2019.2920206
26. Robey, A., Hu, H., Lindemann, L., Zhang, H., Dimarogonas, D. V., Tu, S., & Matni, N. (2020). Learning control barrier functions from expert demonstrations. 2020 59th IEEE Conference on Decision and Control (CDC). https://doi.org/10.1109/cdc42340.2020.9303785
27. Udatha, S., Lyu, Y., & Dolan, J. (2023). Reinforcement learning with probabilistically safe control barrier functions for ramp merging. 2023 IEEE International Conference on Robotics and Automation (ICRA), 5625-5630. https://doi.org/10.1109/icra48891.2023.10161418
28. Zhao, Q., Zhang, Y., & Li, X. (2022). Safe reinforcement learning for dynamical systems using barrier certificates. Connection Science, 34(1), 2822-2844. https://doi.org/10.1080/09540091.2022.2151567