Multi-Constraint Safe Reinforcement Learning via Closed-form Solution for Log-Sum-Exp Approximation of Control Barrier Functions

Chenggang Wang   CGWANG-AUV@SJTU.EDU.CN
Shanghai Jiao Tong University, Shanghai, China

Xinyi Wang   XINYWA@UMICH.EDU
University of Michigan, Ann Arbor, MI, USA

Yutong Dong   755467293@SJTU.EDU.CN
Shanghai Jiao Tong University, Shanghai, China

Lei Song   SONGLEI_24@SJTU.EDU.CN
Shanghai Jiao Tong University, Shanghai, China

Xinping Guan   XPGUAN@SJTU.EDU.CN
Shanghai Jiao Tong University, Shanghai, China

Abstract

The safety of training task policies and their subsequent application using reinforcement learning (RL) methods has become a focal point in the field of safe RL. A central challenge in this area remains the establishment of theoretical guarantees for safety during both the learning and deployment processes. Given the successful implementation of Control Barrier Function (CBF)-based safety strategies in a range of control-affine robotic systems, CBF-based safe RL demonstrates significant promise for practical applications in real-world scenarios. However, integrating these two approaches presents several challenges. First, embedding safety optimization within the RL training pipeline requires that the optimization outputs be differentiable with respect to the input parameters, a condition commonly referred to as differentiable optimization, which is non-trivial to solve. Second, the differentiable optimization framework confronts significant efficiency issues, especially when dealing with multi-constraint problems. To address these challenges, this paper presents a CBF-based safe RL architecture that effectively mitigates the issues outlined above. The proposed approach constructs a continuous AND logic approximation for the multiple constraints using a single composite CBF. By leveraging this approximation, a close-form solution of the quadratic programming is derived for the policy network in RL, thereby circumventing the need for differentiable optimization within the end-to-end safe RL pipeline. This strategy significantly reduces computational complexity because of the closed-form solution while maintaining safety guarantees. Simulation results demonstrate that, in comparison to existing approaches relying on differentiable optimization, the proposed method significantly reduces training computational costs while ensuring provable safety throughout the training process. This advancement opens up promising potential for applications in large-scale optimization problems.

Keywords: Safe reinforcement learning, composite control barrier functions, closed-form solution

1. Introduction

The safety of reinforcement learning (RL) during both training and deployment phases has garnered increasing attention Lavanakul et al. (2024); Vaskov et al. (2024); Buerger et al. (2024), particularly due to the safety-critical nature of many robotic systems. A core challenge lies in ensuring provable safety throughout these phases. Traditional RL methods commonly address safety by penalizing unsafe behaviors, which inevitably leads to the exploration of unsafe actions during training and fails to guarantee the safety of the learned policy during deployment. Recent solutions can be divided into two categories: constrained optimization-based methods and safety filter-based methods. For constrained optimization involving multiple safety constraints, Lagrangian-based safe RL methods Xu et al. (2021); Yao et al. (2024) are proposed to improve training efficiency with constraint satisfaction. Safety filter based methods typically rely on certificate functions such as Control Barrier Functions (CBFs), or Hamilton-Jacobi Reachability value functions. However, there remains a lack of efficient and generalizable approaches to ensure safety across all phases of the RL process.

The CBF-based approach Wang et al. (2017); Ames et al. (2019); Agrawal and Panagou (2021); Wang et al. (2023); Xiao and Belta (2022) theoretically ensures safety for control strategies and has been widely applied to various control-affine robotic systems, such as autonomous vehicles Wang et al. (2023), bipedal robots Csomay-Shanklin et al. (2021) and etc. The core idea involves formulating safety constraints for the control strategy, defining a safe set through these constraints, and deriving forward invariance conditions for the safe set to impose decision-variable constraints that ensure safety. These constraints are then integrated into an optimization problem to generate safe strategies. Typically, the safety optimization is based on system models and nominal controllers derived from control theory. Building on this framework, learning-based methods can replace nominal controllers, leveraging the powerful approximation capabilities and superior task performance of learning techniques Cheng et al. (2019).

Integrating safety optimization into the pipeline of control policy learning can be framed as a decision-focused learning paradigm Shah et al. (2022). In this framework, the prediction phase is handled by a RL policy network, followed by downstream safety optimization to generate the final safe strategy and evaluate its performance. This end-to-end approach requires the safety optimization process to be differentiable, which is often challenging due to issues like solution discontinuity Ferber et al. (2020) and gradient approximation Wilder et al. (2019). Recent works in decision-focused learning address these challenges through various methods: using surrogates to replace the original optimization problem and learning loss functions Wilder et al. (2019) or constructing differentiable optimization tools Amos and Kolter (2017); Pineda et al. (2022); Agrawal et al. (2019). For control-affine systems, safety behavior optimization benefits from linear relaxation of decision variables via Nagumo's theorem Ames et al. (2019), which avoids the complexity of differentiable nonlinear programming Pineda et al. (2022) or mixed-integer programming Ferber et al. (2020). This allows the use of differentiable Quadratic Programming (QP) solvers Amos and Kolter (2017); Agrawal et al. (2019).

Recent research on differentiable QP-based safe control Emam et al. (2022); Ma et al. (2022); Amos et al. (2018); Romero et al. (2024) primarily focuses on three aspects: (1) addressing the impact of constraint parameters, such as environmental changes on safety strategies, e.g., Ma et al. (2022), by adjusting the class-$\mathcal{K}$ functions within safety constraints via differentiable QP; (2) constructing linear MPC problems Amos et al. (2018) and tuning receding horizon parameters during optimization through differentiable QP to enhance task performance Romero et al. (2024); and (3)

imitating safe behaviors Xiao et al. (2023) or integrating safe QP as the final layer of RL policy networks Emam et al. (2022) to generate safe optimization strategies. Differentiable QP frameworks offer several notable advantages. First, they enable a decoupled design of task policy learning and safety correction, thereby facilitating the seamless integration of various learning methodologies. Second, the end-to-end learning of optimized strategies often yields superior task performance compared to hierarchical learning frameworks that incorporate safety corrections post-policy training. Despite these benefits, differentiable QP frameworks are not without limitations. Differentiable optimization is inherently complex, particularly for problems involving discrete decision variables. Furthermore, each gradient update requires solving an optimization problem and subsequently differentiating through it, which can result in significant computational cost.

Given these observations, this work focuses on safe RL with CBF-based optimization and addresses the computational complexity associated with differentiable optimization. Given that safety-critical applications often involve multiple constraints within the optimization problem, a continuous Log-Sum-Exp approximation is employed to transform multiple constraints into a single composite constraint. Utilizing this composite constraint, the closed-form solution of the corresponding QP is derived and integrated into the final layer of the RL policy network, which enables an end-to-end training pipeline with analytical computation, effectively serving as a surrogate for the differentiable QP. The proposed framework significantly reduces the computational cost associated with computing the derivatives of the differentiable QP output with respect to its input parameters Amos and Kolter (2017); Agrawal et al. (2019), offering an efficient and scalable solution for training in large-scale optimization problems.

2. Preliminaries

2.1. Safe policy via CBF-based QP

Consider a control-affine system

$$\dot{x} = f(x) + g(x)u, \tag{1}$$

where $x \in \mathbb{R}^n$ denotes system state, $u \in \mathbb{R}^m$ denotes control input (policy). In this paper, we consider $f : \mathbb{R}^n \to \mathbb{R}^n, g : \mathbb{R}^n \to \mathbb{R}^{n \times m}$ are bounded Lipschitz continuous vector fields and f, g are known for safety guarantee. This consideration is common, since most of the mechanical systems can be formulated as the control-affine form including manipulators, autonomous vehicles, drones, bipedal robots, and etc. For the safety of control-affine systems at the dynamical level, CBFs have successful applications. The safety is related to the desired safe sets which can be defined by continuous differentiable functions $h_i(x) : \mathbb{R}^n \to \mathbb{R}, i = 1, \dots, I$:

$$\mathcal{C}_i \triangleq \{x \in \mathbb{R}^n : h_i(x) \geq 0\}, \tag{2}$$

$$\partial\mathcal{C}_i \triangleq \{x \in \mathbb{R}^n : h_i(x) = 0\}, \tag{3}$$

$$\mathrm{Int}\mathcal{C}_i \triangleq \{x \in \mathbb{R}^n : h_i(x) > 0\}. \tag{4}$$

The set $\mathcal{C}_i$ is forward invariant if for any initial state $x(0) \in \mathcal{C}_i, x(t) \in \mathcal{C}_i, \forall t \in [0, \infty)$. The system is safe if all $\mathcal{C}_i, i = 1, \dots, I$ are forward invariant.

Given the dynamics in (1) and safety requirement, the forward invariance condition based on CBF is formulated as follows: Let $\mathcal{C}_i$ be the 0-superlevel set of a continuously differentiable function

$h_i : \mathbb{R}^n \rightarrow \mathbb{R}$. The function h_i is a CBF for (1) w.r.t. $\mathcal{C}_i$ if there exists extended class $\mathcal{K}$ function α and $u \in \mathbb{R}^m$ such that

$$L_f h_i(x) + L_g h_i(x)u \geq -\alpha(h_i(x)), \tag{5}$$

where $L_f h_i(x) = \frac{\partial h_i(x)}{\partial x} f(x), L_g h_i(x) = \frac{\partial h_i(x)}{\partial x} g(x)$ w.r.t f, g.

Since all the Ith safe constraints are linear on u, the control optimization based on QP can be formulated as

$$\begin{aligned} u_s = \arg \min_{u \in \mathbb{R}^m} \ &\frac{1}{2} \|u - \bar{u}\|_2^2 \\ \text{s.t. } L_f h_i(x) &+ L_g h_i(x)u \geq -\alpha(h_i(x)), i = 1, \ldots, I, \end{aligned} \tag{6}$$

where $\bar{u}$ denotes the nominal controller designed for the original task objective. The optimization (6) minimally corrects the nominal controller when $\bar{u}$ violates the safety constraints, resulting in the safe policy u_s. More details are referred to Lemma 2 and 3 in Breeden and Panagou (2023) or Theorem 1 in Aali and Liu (2022), with an assumption for nonempty feasible set of u_s.

2.2. Soft Actor-Critic

As an off-policy RL algorithm, the Soft Actor-Critic (SAC) Haarnoja et al. (2018) leverages its high sample efficiency and entropy regularization features to offer performance advantages in RL methods for continuous action spaces. The entropy objective is optimized by

$$\pi^* = \arg \max \sum_t \mathbb{E}_{(x_t, u_t^\phi) \sim \rho_\pi} \left[r(x_t, u_t^\phi) + \alpha_e H(\pi(\cdot|x_t)) \right], \tag{7}$$

The SAC algorithm utilizes an AC approach, where the critic is represented by a Q-function parameterized by θ, and the actor is represented by a policy π parameterized by ϕ. The critic loss $J_Q(\theta)$ aims to minimize the difference between the Q-values generated by the critic and the sum of the rewards plus the expected value of the next state's value function:

$$\begin{aligned} J_Q(\theta) = \mathbb{E}_{(x_t, u_t^\phi) \sim D_r} \Big[\frac{1}{2} \Big(Q_\theta(x_t, u_t) - \Big(r(x_t, u_t^\phi) \\ + \gamma \mathbb{E}_{x_{t+1} \sim p} \left[V_{\hat{\theta}}(x_{t+1}) \right] \big) \big)^2 \Big], \end{aligned} \tag{8}$$

where D_r is the replay buffer, and $\hat{\theta}$ represents the target Q-network parameters. The replay buffer D_r provides a diverse set of experiences, enabling the critic to learn from a broad range of past states and actions, which enhances sample efficiency. The target Q-network parameters $\hat{\theta}$ ensure stable updates by serving as a slowly updating reference.

The entropy term is included to promote exploration and prevent premature convergence to suboptimal policies, which is given by:

$$H(\pi(\cdot|x_t)) = -\log \pi_\phi \left(u_t^\phi | x_t \right). \tag{9}$$

The policy loss encourages actions that maximize both the expected reward and the entropy, leading to an effective balance between performance and exploration, which is given by

$$J_\pi(\phi) = \mathbb{E}_{x_t \sim \mathcal{D}_r} \left[\mathbb{E}_{u_t^\phi \sim \pi_\phi} [\alpha_e \log \pi_\phi(u_t^\phi | x_t) - Q_\theta(x_t, u_t^\phi)] \right]. \tag{10}$$

One of the primary advantages of constructing a differentiable optimization framework is the ability to decouple the design of the safety layer, enabling seamless integration into the policy network of actor-critic (AC)-based RL methods. Therefore, the SAC serves as a candidate when implementing a differentiable QP layer. The policy loss is given by

$$J_\pi(\phi) = \mathbb{E}_{x_t \sim \mathcal{D}_r} \\ \left[\mathbb{E}_{u_t^\phi \sim \pi_\phi} [\alpha_e \log \pi_\phi(u_t^\phi | x_t) - Q_\theta(x_t, u_t^\phi + u_t^C)] \right], \tag{11}$$

where u_t^C is the compensation term computed by differentiable QP layer.

3. Main results

3.1. Composite CBF for multiple constraints

To solve the CBF-based optimization under multiple constraints, the constraints are regarded as the intersection of safe sets defined by these CBFs. Each safe constraint h_i is defined by a 0-superlevel set and their intersection is defined as

$$\bigcap_{i=1,...,I} C_i = \{x \in \mathbb{R}^n : h_i(x) \geq 0\}, \tag{12}$$

where I denotes the number of the safety constraints.

In other words, the intersection of sets captures the logical AND relationship between multiple safety constraints, which is denoted as

$$x \in \bigcap_{i=1,...,I} C_i \iff x \in C_1 \text{ AND } x \in C_2 \cdots \text{ AND } x \in C_I. \tag{13}$$

When there are multiple constraints, the complexity of the QP problem increases, generally making it impossible to derive a closed-form solution, thus requiring numerical optimization methods such as active set or interior point methods. However, inspired by existing literature Molnar and Ames (2023) solving complex safety specifications, this paper employs a Log-Sum-Exp approximation technique to transform multiple constraints into a single constraint, thereby enabling a closed-form solution for the safe QP.

The approximated composite single CBF is constructed as:

$$h(x) = -\frac{1}{\kappa} \ln \left(\sum_{i=1}^{I} e^{-\kappa h_i(x)} \right), \tag{14}$$

whose Lie derivatives are expressed by:

$$L_f h(x) = \sum_{i=1}^{I} \lambda_i(x) L_f h_i(x), \quad L_g h(x) = \sum_{i=1}^{I} \lambda_i(x) L_g h_i(x), \tag{15}$$

where

$$\lambda_i(x) = e^{-\kappa(h_i(x) - h(x))}, \tag{16}$$

with $\sum_{i \in I} \lambda_i(x) = 1$ and $\kappa > 0$.

Since an equivalent substitution for the constraints of optimization problem (6) is $\min h_i(x) \geq 0, i = 1, \cdots, I$. The composite CBF in (14) shares the following property.

Lemma 1: Molnar and Ames (2023) Consider sets C_i in (2) and their intersection in (12). Continuous function $h(x)$ in (14) under approximates $\min_{i=1,\cdots,I} h_i(x) \geq 0$ with bounds:

$$\min_{i=1,\cdots,I} h_i(x) - \frac{\ln I}{\kappa} \leq h(x) \leq \min_{i=1,\cdots,I} h_i(x) \quad \forall x \in \mathbb{R}^n, \tag{17}$$

such that $\lim_{\kappa \to \infty} h(x) = \min_{i=1,\cdots,I} h_i(x)$. The corresponding set $C = \{x \in \mathbb{R}^n : h(x) \geq 0\}$ lies inside the intersection, $C \subseteq \bigcap_{i=1,\cdots,I} C_i$, such that $\lim_{\kappa \to \infty} C = \bigcap_{i=1,\cdots,I} C_i$.

See Proof of Theorem 4 in Molnar and Ames (2023). $h(x) \geq 0$ guarantees $\min_{i=1,\cdots,I} h_i(x) \geq 0$, indicating all constraints $h_i \geq 0, i = 1, \cdots, I$ are satisfied.

3.2. Closed-form solution for CBF-based QP

The safety-oriented framework offers a QP-based optimization approach to modify a nominal policy to ensure safety. The nominal policy $\bar{u}$, typically designed to achieve a specific task objective, can be derived from model-based control or generated through RL. Based on the established composite CBF $h(x)$ in (14), the optimization problem ensuring system safety can be formulated as the following QP:

$$u_s(x) = \arg\min_{u \in \mathbb{R}^m} \frac{1}{2}\|u - \bar{u}(x)\|_2^2 \tag{18}$$

subject to

$$L_f h(x) + L_g h(x)u \geq -\alpha(h(x)). \tag{19}$$

When the nominal policy satisfies the safety constraint, the constraint (19) is inactive, and the safe policy aligns with the nominal policy. However, when the nominal policy violates the safety constraint, the QP seeks a safe policy that satisfies the constraints while deviating minimally from the nominal policy. The purpose of transforming multiple constraints into a composite CBF is to derive a closed-form solution for the safe policy of the optimization (18). The closed-form solution can be obtained by referring to the following theorem.

Theorem 1: Let C be the 0-superlevel set of a continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}$, and let $\bar{u}(x) : \mathbb{R}^n \to \mathbb{R}^m$ be a nominal controller. If h is a composite CBF for (1) on the set $C \subseteq \bigcap_{i=1,\cdots,I} C_i$ with the corresponding function $\alpha \in \mathcal{K}_\infty^e$, then the optimization problem in (18) is feasible for any $x \in \mathbb{R}^n$ and has a closed-form solution given by

$$u_s(x) = \bar{u}(x) + \max\{0, \eta(x)\} L_g h(x)^\top \tag{20}$$

where the function $\eta : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$\eta(x) = \begin{cases} -\frac{L_f h(x) + L_g h(x)\bar{u}(x) + \alpha(h(x))}{\|L_g h(x)\|_2^2} & \text{if } L_g h(x) \neq 0, \\ 0 & \text{if } L_g h(x) = 0. \end{cases} \tag{21}$$

See proof of Theorem 2 in Alan et al. (2023). Theorem 2 provides a sufficient but not necessary condition for a safe solution, offering an analytical form for solving the QP associated with a single constraint. This formulation eliminates the need to invoke a QP solver, significantly reducing the computational cost. Therefore, this advantage motivates its integration with the RL framework.

Furthermore, the closed-form solution provides the significant advantage of circumventing the requirement for differentiable optimizations within the RL framework, thereby substantially simplifying the gradient computation in the safe policy generation and alleviating the complexity associated with gradient-based optimization, as will be elaborated in the next subsection.
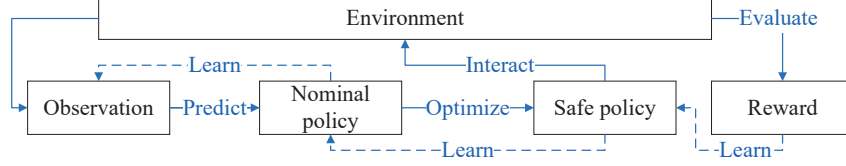

Figure 1: An illustration of an end-to-end training safe RL framework.

3.3. Safety layer via closed-form solution in RL framework

In conventional RL architectures, the final layer of the control policy network typically consists of a fully connected layer, particularly for continuous control actions in affine systems. The output is bounded by the final activation function, such as the hyperbolic tangent, to ensure bounded action outputs. For safe policy generation, an intuitive approach is to correct the RL-derived control policy by adjusting it through safety-oriented mechanisms, such as correcting the control policy to a safe policy using a CBF-based QP Cheng et al. (2019). However, in this approach, the reward from the safe output cannot backpropagate to the RL network, due to the absence of a gradient pathway connecting the safe policy to the RL policy. Recently, decision-focused learning have proposed architectures based on differentiable optimization, embedding optimizable and differentiable structures to achieve an end-to-end learning pipeline, thus enabling CBF-QP-based safe learning, as illustrated in Figure 1. This raises a critical challenge: each training step requires solving a batch of QP problems for the policy loss function, along with calculating the gradient of each QP output concerning the QP parameters, making it computationally expensive and challenging to large-scale multi-constraint problems. To address this issue, we integrate a closed-form solution for the safe policy directly into the RL policy generation pipeline. This approach leverages a composite single-constraint approximation to handle multi-constraint scenarios, alongside explicit QP solutions to circumvent forward optimization and its gradient backpropagation. We replace the final layer in RL policy generation with an analytically computed "safety layer", which, due to its analytical properties, can be integrated into any actor-critic RL method. An illustration of safe policy networks in actor-critic framework is shown in Figure 2 with different safety layers. The proposed framework is demonstrated in Figure 2(a), where the closed-form solution (20) and (21) are integrated into the final layer before safety policy generation. As a comparison, Figure 2(b) demonstrates that taking nominal policy as the input, the differentiable QP layer compute the forward solution with multiple constraints, which is potentially infeasible and computationally expensive.

We illustrate the proposed approach using the SAC method, where the loss functions in this framework are given by

$$J_Q(\theta) = \mathbb{E}_{(x_t, u_s^\phi) \sim \mathcal{D}_R} \left[\frac{1}{2} \left(Q_\theta(x_t, u_s^\phi) - \left(r(x_t, u_s^\phi) + \gamma \mathbb{E}_{x_{t+1} \sim p} \left[V_{\bar{\theta}}(x_{t+1}) \right] \right) \right)^2 \right], \qquad (22)$$
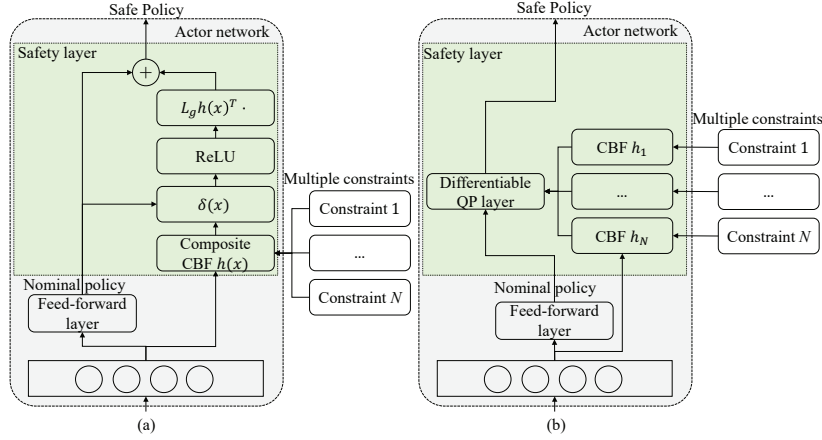
Figure 2: An illustration of safe policy networks with different safety layers. Subfigure (a) demonstrates the proposed framework where N constraints are composited to $h(x)$ using a continuous Log-Sum-Exp approximation. The safety layer is analytical based on the closed-form solution of the composite CBF-based optimization. Subfigure (b) demonstrates the existing framework with differentiable QP layer. The safety layer solves the forward CBF-based optimization and computes the gradient during backpropagation.

$$V_{\bar{\theta}}(x_t) = \mathbb{E}_{u_s^\phi \sim \pi_\phi} \left[Q_{\bar{\theta}}(x_t, u_s^\phi) - \alpha_e \log \pi_\phi(u_s^\phi | x_t) \right], \tag{23}$$

$$J_\pi(\phi) = \mathbb{E}_{x_t \sim \mathcal{D}_r} \left[\mathbb{E}_{u_s^\phi \sim \pi_\phi} [\alpha_e \log \pi_\phi(u_s^\phi | x_t) - Q_\theta(x_t, u_s^\phi)] \right], \tag{24}$$

where π_ϕ denotes the policy generated by the entire policy network, including both the fully connected layers and the QP-based adjustment. In this case, $u_s^\phi \sim \pi_\phi$ would mean that the sample u_s^ϕ is drawn from the distribution defined by the entire policy network, which inherently includes the safety layer for the QP adjustment.

4. Experiment

In this section, we aim to validate the capability of the proposed method to ensure safety during training while achieving faster training efficiency compared to the differentiable QP-based approach. The testing environment is designed as a reachability task, where the agent's objective is to reach the goal position while avoiding obstacles. To illustrate the incorporation of multiple constraints in the safety-oriented optimization, the collision-free constraints are defined as follows:

$$h_i(x) = \|p - p_{i,\text{obs}}\|^2 - r_{\text{safe}}^2 \geq 0, \quad i = 1, \ldots, I, \tag{25}$$

where $p = [p_x, p_y]^\top$ denotes the position of the agent, $p_{i,\text{obs}}$ denotes the position of the ith obstacle, and r_{safe} is the safe radius to avoid collision. Furthermore, within the safety constraints based on CBFs, the selection of the class-$\mathcal{K}$ function α significantly affects the conservativeness of safety enforcement and κ affects the the approximation error of $\min$ operation. In this study, we adopt a trade-off value to balance safety and performance with $\alpha = 5(\cdot), \kappa = 2$.

To demonstrate the safety of the proposed method, we present the following metrics during training: $\min h_i, i = 1, \ldots, I$ and the composite h for each episode with $I = 3$. Moreover, the $\min h_i, i \in I$ and composite h across all steps of training episodes, and the trajectories for the deployment phase after training are also demonstrated. The results are illustrated in the Figure 3 and Figure 4.
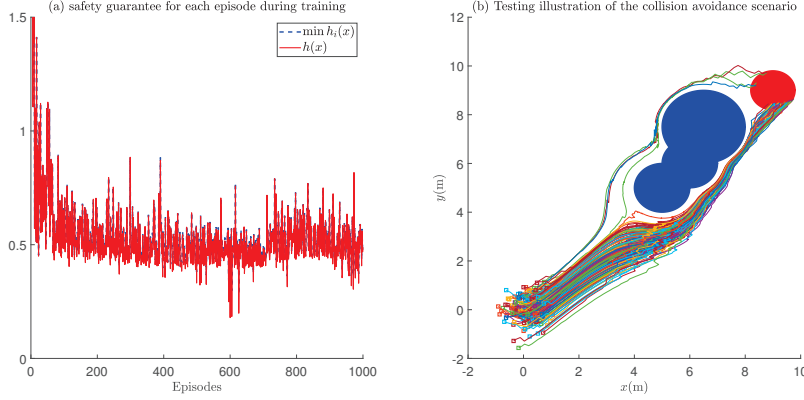

Figure 3: Performance of safe RL training and testing with the proposed method. Subfigure (a) illustrates $\min h_i, i = 1, \ldots, I$ and the composite h for each episode during training. The composite $h(x)$ under approximates $\min h_i$ and maintains positive in each episode. Subfigure (b) illustrates the successful trajectories during testing. The blue area in Subfigure (b) contains three obstacles, each with a different safe radius size. The colored squares denote the initial positions. The red circle represents the target area, while the colored lines indicate the testing trajectories, each starting from different initial positions near the origin.

As shown in Figure 3(a), the minimum value of h_i remains consistently greater than 0 throughout the entire training process, indicating that the system successfully achieves safe training and policy learning. Furthermore, the red line $h(x)$ serves as an under-approximation of the blue dashed line, which is consistent with the conclusion of Theorem 1. Figure 3(b) illustrates the trajectories reaching the target area, where safety is ensured across 200 trials.

A more detailed perspective is provided in Figure 4, which illustrates the evolution of h_i and h across all steps in each episode (different color of curves) over 1000 training iterations (with a maximum step limit of 200). The time intervals during which the h_i curves near the safety set boundary exhibit "flattening" correction, depending on the different positions of obstacles. During these intervals, the condition $L_f h(x) + L_g h(x)\bar{u}(x) + \alpha(h(x)) < 0$ holds. Therefore, the safe policy is actively filtered and corrected based on (20) and (21), ensuring that h remains positive for all time.

Table 1: Comparison of different approaches

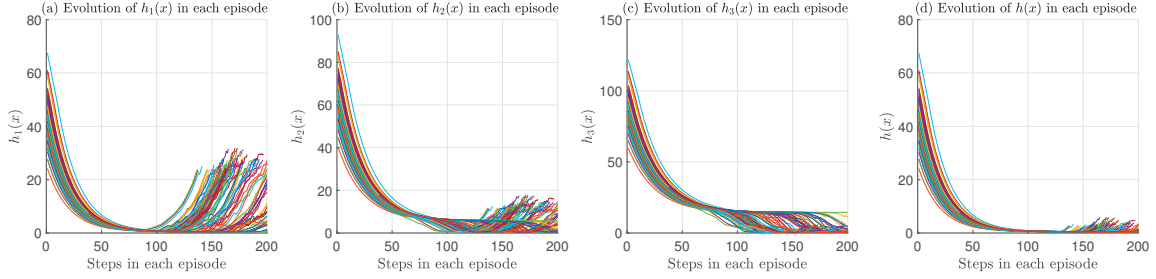| Method	ATTS(s) with $I = 3$	ATTS(s) with $I = 10$	ATTS(s) with $I = 30$
Closed-form solution	0.018	0.024	0.043
CBF Batch QP	0.13	0.25	0.40
CBF CVXPYlayer	0.84	1.45	2.26

Figure 4: Evolution of h_1, h_2, h_3 and the composite h over 1000 episodes. The safe learning during training is guaranteed.

As previously noted, the closed-form solution eliminates the need for differentiable QP solvers, thereby reducing computational costs. This represents another significant advantage of the proposed method. In the same scenario for collision avoidance of multiple obstacles, we compared the proposed method with the differentiable QP solver Batched-QPFunction Amos and Kolter (2017) and the CVXPYLayer Agrawal et al. (2019) in terms of computational performance, which are commonly used in similar works within CBF-based safe learning Emam et al. (2022); Ma et al. (2022); Jiang et al. (2024); Romero et al. (2024). The comparative results are summarized in Table 1, where the performance metric is the average solving time per time step (ATTS) during the RL training process. In addition, scenarios with 10 and 30 constraints are also tested to validate the scalability of the proposed method in solving larger-scale safe RL problems.

As demonstrated in Table 1, the proposed method exhibits a computational speed advantage of at least one order of magnitude because of the close-form nature. The CVXPYlayer-based method, while supporting disciplined parametrized programming, exhibits the lowest solving efficiency due to the lack of support for batch solving and the requirement for gradient computation in QP. The advantage in training time makes it potentially effective for optimization in large-scale safe RL problems.

5. Conclusion

This paper addresses the challenges of ensuring multiple safety constraints and improving training efficiency in safe RL. We propose a safe RL framework based on the closed-form solution of composite CBF. The framework constructs a composite CBF by using the Log-Sum-Exp approximation of the min function to integrate multiple safety constraints in the optimization problem. It also inherits the safety guarantees based on the composite CBF defining the safe set. By serving as a surrogate of the differentiable QP architecture with a closed-form solution, the proposed method significantly enhances training efficiency. Comparative experiments demonstrate that the proposed method is up to 7 times faster than the current state-of-the-art differentiable batch QP solvers, and at least 46 times faster than the differentiable convex optimization layers CVXPYlayer, showcasing its potential for solving optimization in large-scale safe RL problems. Future work will further investigate the composite CBF-QP under explicit input constraints, with a focus on guaranteeing feasibility and improving efficiency within the framework of differentiable optimization.

Acknowledgments

References

Mohammad Aali and Jun Liu. Multiple control barrier functions: An application to reactive obstacle avoidance for a multi-steering tractor-trailer system. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 6993–6998. IEEE, 2022.

Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J Zico Kolter. Differentiable convex optimization layers. *Advances in neural information processing systems*, 32, 2019.

Devansh R. Agrawal and Dimitra Panagou. Safe control synthesis via input constrained control barrier functions. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 6113–6118, 2021.

Anil Alan, Andrew J. Taylor, Chaozhe R. He, Aaron D. Ames, and Gábor Orosz. Control barrier functions and input-to-state safety with application to automated vehicles. *IEEE Transactions on Control Systems Technology*, 31(6):2744–2759, 2023.

Aaron D. Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control barrier functions: Theory and applications. In *2019 18th European Control Conference (ECC)*, pages 3420–3431, 2019.

Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International conference on machine learning*, pages 136–145. PMLR, 2017.

Brandon Amos, Ivan Jimenez, Jacob Sacks, Byron Boots, and J Zico Kolter. Differentiable mpc for end-to-end planning and control. *Advances in neural information processing systems*, 31, 2018.

Joseph Breeden and Dimitra Panagou. Compositions of multiple control barrier functions under input constraints. In *2023 American Control Conference (ACC)*, pages 3688–3695. IEEE, 2023.

Johannes Buerger, Mark Cannon, and Martin Doff-Sotta. Safe learning in nonlinear model predictive control. In Alessandro Abate, Mark Cannon, Kostas Margellos, and Antonis Papachristodoulou, editors, *Proceedings of the 6th Annual Learning for Dynamics and Control Conference*, volume 242 of *Proceedings of Machine Learning Research*, pages 603–614. PMLR, 15–17 Jul 2024.

Richard Cheng, Gábor Orosz, Richard M. Murray, and Joel W. Burdick. 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, Jul. 2019.

Noel Csomay-Shanklin, Ryan K Cosner, Min Dai, Andrew J Taylor, and Aaron D Ames. Episodic learning for safe bipedal locomotion with control barrier functions and projection-to-state safety. In *Learning for Dynamics and Control*, pages 1041–1053. PMLR, 2021.

Yousef Emam, Gennaro Notomista, Paul Glotfelter, Zsolt Kira, and Magnus Egerstedt. Safe reinforcement learning using robust control barrier functions. *IEEE Robotics and Automation Letters*, pages 1–8, 2022.

Aaron Ferber, Bryan Wilder, Bistra Dilkina, and Milind Tambe. Mipaal: Mixed integer program as a layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1504–1511, 2020.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *CoRR*, abs/1801.01290, 2018.

Yongchao Jiang, Chenggang Wang, Ziqi He, and Lei Song. A differentiable qp-based learning framework for safety-critical control of fully actuated auvs. In *2024 3rd Conference on Fully Actuated System Theory and Applications (FASTA)*, pages 259–264. IEEE, 2024.

Will Lavanakul, Jason Choi, Koushil Sreenath, and Claire Tomlin. Safety filters for black-box dynamical systems by learning discriminating hyperplanes. In *6th Annual Learning for Dynamics & Control Conference*, pages 1278–1291. PMLR, 2024.

Hengbo Ma, Bike Zhang, Masayoshi Tomizuka, and Koushil Sreenath. Learning differentiable safety-critical control using control barrier functions for generalization to novel environments. In *2022 European Control Conference (ECC)*, pages 1301–1308, 2022.

Tamas G. Molnar and Aaron D. Ames. Composing control barrier functions for complex safety specifications. *IEEE Control Systems Letters*, 7:3615–3620, 2023.

Luis Pineda, Taosha Fan, Maurizio Monge, Shobha Venkataraman, Paloma Sodhi, Ricky T. Q. Chen, Joseph Ortiz, Daniel DeTone, Austin Wang, Stuart Anderson, Jing Dong, Brandon Amos, and Mustafa Mukadam. Theseus: A library for differentiable nonlinear optimization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 3801–3818. Curran Associates, Inc., 2022.

Angel Romero, Yunlong Song, and Davide Scaramuzza. Actor-critic model predictive control. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14777–14784. IEEE, 2024.

Sanket Shah, Kai Wang, Bryan Wilder, Andrew Perrault, and Milind Tambe. Decision-focused learning without decision-making: Learning locally optimized decision losses. *Advances in Neural Information Processing Systems*, 35:1320–1332, 2022.

Sean Vaskov, Wilko Schwarting, and Chris Baker. Do no harm: A counterfactual approach to safe reinforcement learning. In Alessandro Abate, Mark Cannon, Kostas Margellos, and Antonis Papachristodoulou, editors, *Proceedings of the 6th Annual Learning for Dynamics and Control Con-*

ference, volume 242 of *Proceedings of Machine Learning Research*, pages 1675–1687. PMLR, 15–17 Jul 2024.

Chenggang Wang, Shanying Zhu, Bochen Li, Lei Song, and Xinping Guan. Time-varying constraint-driven optimal task execution for multiple autonomous underwater vehicles. *IEEE Robotics and Automation Letters*, 8(2):712–719, 2023.

Li Wang, Aaron D. Ames, and Magnus Egerstedt. Safety barrier certificates for collisions-free multirobot systems. *IEEE Transactions on Robotics*, 33(3):661–674, 2017.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1658–1665, 2019.

Wei Xiao and Calin Belta. High-order control barrier functions. *IEEE Transactions on Automatic Control*, 67(7):3655–3662, 2022.

Wei Xiao, Tsun-Hsuan Wang, Ramin Hasani, Makram Chahine, Alexander Amini, Xiao Li, and Daniela Rus. Barriernet: Differentiable control barrier functions for learning of safe robot control. *IEEE Transactions on Robotics*, 39(3):2289–2307, 2023.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pages 11480–11491. PMLR, 2021.

Yihang Yao, Zuxin Liu, Zhepeng Cen, Peide Huang, Tingnan Zhang, Wenhao Yu, and Ding Zhao. Gradient shaping for multi-constraint safe reinforcement learning. In Alessandro Abate, Mark Cannon, Kostas Margellos, and Antonis Papachristodoulou, editors, *Proceedings of the 6th Annual Learning for Dynamics and Control Conference*, volume 242 of *Proceedings of Machine Learning Research*, pages 25–39. PMLR, 15–17 Jul 2024.