多约束安全强化学习通过控制障碍函数对数和指数近似的封闭形式解方法

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

Editors: N. Ozay, L. Balzano, D. Panagou, A. Abate

Abstract

训练任务策略的安全性及其在使用强化学习（RL）方法的应用已经成为安全 RL 领域的焦点。该领域中仍然存在一个中心挑战，即在学习和部署过程中建立安全性的理论保证。鉴于控制障碍函数（CBF）为基础的安全策略成功应用于一系列仿射控制系统，基于 CBF 的安全 RL 在实际应用场景中显示出巨大的潜力。然而，将这两种方法结合起来面临着若干挑战。首先，在 RL 训练管道中嵌入安全性优化要求该优化输出对输入参数可微分，这通常被称为可微分优化，解决这一问题并不简单。其次，可微分优化框架遇到显著的效率问题，特别是在处理多约束问题时。为了解决这些挑战，本文提出了一种基于 CBF 的安全 RL 架构，有效缓解了上述问题。所提方法使用单一复合控制障碍函数（CBF）构建多个约束条件的连续 AND 逻辑近似值。通过利用这一近似值，在强化学习中推导出策略网络二次规划的封闭解，从而避免在端到端安全 RL 管道中进行可微分优化的需求。这种策略由于闭合形式解决方案而显著降低了计算复杂性，同时保持安全性保证。仿真结果表明，与依赖于可微分优化的现有方法相比，所提方法在确保训练过程中的证明安全性的同时，显著减少了训练计算成本。这一进展为大规模优化问题的应用开启了有前途的可能性。

Keywords: 安全强化学习，复合控制障碍函数，闭式解法

1. 介绍

强化学习（RL）在训练和部署阶段的安全性已引起越来越多的关注 Lavanakul et al. (2024); Vaskov et al. (2024); Buerger et al. (2024)，特别是由于许多机器人系统具有安全关键性质。核心挑战在于确保在整个这些阶段中的可证明安全性。传统的 RL 方法通常通过惩罚不安全行为来处理安全性问题，这不可避免地会导致在训练过程中探索不安全的动作，并且无法保证部署期间学习策略的安全性。最近的解决方案可以分为两类：基于约束优化的方法和基于安全过滤

器的方法。对于涉及多个安全约束的约束优化问题，提出了基于拉格朗日的安全 RL 方法 Xu et al. (2021); Yao et al. (2024)，以提高满足约束条件下的训练效率。基于安全过滤器的方法通常依赖于证书函数，例如控制障碍函数（CBFs）或哈密顿-雅可比可达性值函数。然而，仍然缺乏在 RL 过程的所有阶段确保安全性的高效且通用方法。

基于 CBF 的方法 Wang et al. (2017); Ames et al. (2019); Agrawal and Panagou (2021); Wang et al. (2023); Xiao and Belta (2022) 从理论上确保了控制策略的安全性，并广泛应用于各种控制仿型机器人系统，如自主车辆 Wang et al. (2023)、双足机器人 Csomay-Shanklin et al. (2021) 等。核心思想是为控制策略制定安全约束条件，通过这些约束定义一个安全集，并推导出该安全集的前向不变性条件，以施加决策变量约束来确保安全性。然后将这些约束整合到优化问题中生成安全策略。通常，安全优化基于系统模型和从控制理论派生的名义控制器。在此框架基础上，学习方法可以替代名义控制器，利用学习技术的强大近似能力和优越的任务性能 Cheng et al. (2019)。

将安全性优化集成到控制策略学习的流水线中可以被定义为一个以决策为中心的学习范式 Shah et al. (2022)。在这个框架中，预测阶段由 RL 策略网络处理，随后进行下游的安全性优化，生成最终的安全策略并评估其性能。这种端到端的方法需要安全性优化过程具有可微分性，这通常由于诸如解的不连续 Ferber et al. (2020) 和梯度近似 Wilder et al. (2019) 等问题而具有挑战性。近期在以决策为中心的学习中的研究通过各种方法解决了这些挑战：使用替代模型来替换原始优化问题和学习损失函数 Wilder et al. (2019)，或者构建可微分的优化工具 Amos and Kolter (2017); Pineda et al. (2022); Agrawal et al. (2019)。对于仿射控制系统，安全行为优化得益于通过 Nagumo 定理 Ames et al. (2019) 对决策变量进行线性松弛，这避免了可微非线性规划 Pineda et al. (2022) 或混合整数规划 Ferber et al. (2020) 的复杂性。这允许使用可微二次规划（QP）求解器 Amos and Kolter (2017); Agrawal et al. (2019)。

最近关于可微分 QP 基础安全控制 Emam et al. (2022); Ma et al. (2022); Amos et al. (2018); Romero et al. (2024) 的研究主要集中在三个方面：(1) 处理约束参数的影响，例如通过可微分 QP 调整安全约束内的类 $\mathcal{K}$ 函数来应对环境变化对安全策略（如 Ma et al. (2022)）的影响；(2) 构建线性 MPC 问题 Amos et al. (2018) 并在优化过程中通过可微分 QP 调整滚动时域参数以增强任务性能 Romero et al. (2024)；(3) 模仿安全行为 Xiao et al. (2023) 或将安全 QP 作为 RL 策略网络的最后一层 Emam et al. (2022) 来生成安全优化策略。可微分 QP 框架具有几个显著的优点。首先，它们使任务策略学习与安全性校正的解耦设计成为可能，从而促进了各种学习方法的无缝集成。其次，优化策略的端到端学习通常比在策略训练后纳入安全校正的层次化学习框架表现出更优的任务性能。尽管有这些优点，可微分 QP 框架也不是没有局限性。可微分优化本身是复杂的，尤其是对于涉及离散决策变量的问题而言。此外，每次梯度更新都需要解决一个优化问题并随后对其进行求导，这可能会导致显著的计算成本。

鉴于这些观察，本研究专注于基于 CBF 优化的安全 RL，并解决与可微分优化相关的计算复杂性。由于安全关键应用通常涉及优化问题中的多个约束条件，因此采用连续的 Log-Sum-Exp 近似方法将多个约束转换为单一复合约束。利用该复合约束，推导出相应的 QP 的封闭形式解并将其集成到 RL 策略网络的最后一层中，这使得可以进行具有解析计算的端到端训练流

程，有效地作为可微分 QP 的替代方案。所提出的框架显著降低了与计算可微分 QP 输出相对于输入参数 Amos and Kolter (2017); Agrawal et al. (2019) 的导数相关的计算成本，为大规模优化问题的训练提供了高效且可扩展的解决方案。

2. 预备知识

2.1. 基于 CBF 的 QP 的安全策略

考虑一个仿射控制系统

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

其中 $x \in \mathbb{R}^n$ 表示系统状态，$u \in \mathbb{R}^m$ 表示控制输入（策略）。在本文中，我们考虑 $f : \mathbb{R}^n \to \mathbb{R}^n, g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ 是有界的 Lipschitz 连续向量场，并且已知 f, g 以保证安全。这种考虑是常见的，因为大多数机械系统可以被表述为仿射控制形式，包括操作器、自动驾驶车辆、无人机、双足机器人等等。对于仿射控制系统的动态层面的安全性，CBFs 已有成功的应用。安全与期望的安全集相关，这些集合可以通过连续可微函数 $h_i(x) : \mathbb{R}^n \to \mathbb{R}, i = 1, \ldots, I$ 定义：

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

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

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

集合 $\mathcal{C}_i$ 是正向不变的，如果对于任何初始状态 $x(0) \in \mathcal{C}_i, x(t) \in \mathcal{C}_i, \forall t \in [0, \infty)$。系统是安全的，如果所有 $\mathcal{C}_i, i = 1, \ldots, I$ 都是正向不变的。

给定动态系统在 (1) 和安全要求下，基于 CBF 的向前不变性条件可表述如下：设 $\mathcal{C}_i$ 是一个连续可微函数 $h_i : \mathbb{R}^n \to \mathbb{R}$ 的 0 超水平集。函数 h_i 是关于 (1) 的 CBF。$\mathcal{C}_i$ 如果存在扩展类 $\mathcal{K}$ 函数 α 和 $u \in \mathbb{R}^m$ 满足

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

其中 $L_f h_i(x) = \frac{\partial h_i(x)}{\partial x} f(x), L_g h_i(x) = \frac{\partial h_i(x)}{\partial x} g(x)$ 关于 f, g。

由于所有的第 I 个安全约束都是关于 u 的线性关系，基于 QP 的控制优化可以被表述为

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

其中 $\bar{u}$ 表示为原始任务目标设计的名义控制器。优化 (6) 最小化地修正了名义控制器当 $\bar{u}$ 违反安全约束时，导致生成安全策略 u_s。更多细节参见引理 2 和 3 在 Breeden and Panagou (2023) 中，或定理 1 在 Aali and Liu (2022) 中，在 u_s 的可行集非空的假设下。

2.2. 软演员批评

作为离策略 RL 算法，软演员评论家 (SAC)Haarnoja et al. (2018) 利用其高样本效率和熵正则化特性，在连续动作空间的 RL 方法中提供性能优势。熵目标通过优化

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

SAC 算法利用了 AC 方法，其中批评者由一个参数为 θ 的 Q 函数表示，而行动者则由一个参数为 ϕ 的策略 π 表示。批评者的损失函数 $J_Q(\theta)$ 目标是最小化批评者生成的 Q 值与奖励之和加上下一个状态的价值函数期望值之间的差异：

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

其中 D_r 是回放缓冲区，而 $\hat{\theta}$ 表示目标 Q 网络参数。回放缓冲区 D_r 提供了一组多样化的经验，使评论家能够从广泛的过去状态和行动中学习，从而提高了样本效率。目标 Q 网络参数 $\hat{\theta}$ 通过充当缓慢更新的参考来确保稳定的更新。

熵项被包含以促进探索并防止过早收敛到次优策略，其由以下给出：

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

策略损失鼓励最大化预期奖励和熵的动作，从而在性能和探索之间实现有效的平衡，这由以下公式给出

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

构建可微优化框架的一个主要优势是可以将安全层的设计解耦，从而无缝集成到基于演员-评论家（AC）的强化学习方法的策略网络中。因此，在实现可微 QP 层时，SAC 可以作为一个候选方案。策略损失由下式给出

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

其中 u_t^C 是由可微 QP 层计算的补偿项。

3. 主要结果

3.1. 多约束条件下的组合 CBF

为了解决基于 CBF 的多约束优化问题，将这些约束视为由这些 CBFs 定义的安全集的交集。每个安全约束 h_i 由一个 0 超水平集定义，它们的交集被定义为

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

其中 I 表示安全约束的数量。

换句话说，集合的交集捕捉了多个安全约束之间的逻辑与关系，这表示为

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

当存在多个约束条件时，QP 问题的复杂性增加，通常使得无法得出闭式解，因此需要使用数值优化方法如主动集或内点法。然而，受到现有文献 Molnar and Ames (2023) 解决复杂安全规范的启发，本文采用 Log-Sum-Exp 近似技术将多个约束条件转换为单个约束条件，从而使得安全 QP 问题能够得到闭式解。

近似的复合单一 CBF 构建为：

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

其李导数表示为：

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

其中

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

带有 $\sum_{i \in I} \lambda_i(x) = 1$ 和 $\kappa > 0$。

由于优化问题 (6) 的约束有一个等价替换是 $\min h_i(x) \geq 0, i = 1, \cdots, I$。复合 CBF 在 (14) 中具有以下性质。

引理 1： Molnar and Ames (2023) 考虑集合 C_i 在 (2) 中及其在 (12) 中的交集。连续函数 $h(x)$ 在 (14) 下近似于 $\min_{i=1,\cdots,I} h_i(x) \geq 0$，其边界为：

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

使得 $\lim_{\kappa \to \infty} h(x) = \min_{i=1,\cdots,I} h_i(x)$。相应的集合 $C = \{x \in \mathbb{R}^n : h(x) \geq 0\}$ 位于交集 $C \subseteq \bigcap_{i=1,\cdots,I} C_i$ 内，使得 $\lim_{\kappa \to \infty} C = \bigcap_{i=1,\cdots,I} C_i$。

参见定理 4 的证明在 Molnar and Ames (2023) 中。$h(x) \geq 0$ 保证了 $\min_{i=1,\cdots,I} h_i(x) \geq 0$，表明所有约束条件 $h_i \geq 0, i = 1, \cdots, I$ 都得到满足。

3.2. 基于 CBF 的 QP 的闭式解

安全导向框架提供了一种基于 QP 的优化方法来修改名义策略以确保安全性。名义策略 $\bar{u}$，通常设计用于实现特定任务目标，可以从模型预测控制或通过强化学习生成。基于已建立的复合 CBF$h(x)$ 在 (14) 中的基础上，可以将保证系统安全的优化问题表述为以下 QP：

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

受约束于

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

当名义策略满足安全约束时，约束 (19) 不起作用，并且安全策略与名义策略一致。然而，当名义策略违反安全约束时，QP 寻求一个在尽量不偏离名义策略的同时满足约束的安全策略。将多个约束转换为复合 CBF 的目的在于推导出优化 (18) 的安全策略的封闭形式解。封闭形式解可以通过参考以下定理获得。

定理 1： 令 C 为连续可微函数 $h : \mathbb{R}^n \to \mathbb{R}$ 的 0 超水平集，并令 $\bar{u}(x) : \mathbb{R}^n \to \mathbb{R}^m$ 为名义控制器。如果 h 是集合 $C \subseteq \bigcap_{i=1,\cdots,I} C_i$ 上关于 (1) 的复合 CBF，并且对应的函数是 $\alpha \in \mathcal{K}^e_\infty$，那么在 (18) 中的优化问题对于任意 $x \in \mathbb{R}^n$ 都是可行的，并且其封闭形式解由

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

给出其中函数 $\eta : \mathbb{R}^n \to \mathbb{R}$ 定义为

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

见定理 2 在 Alan et al. (2023) 中的证明。定理 2 提供了安全解的一个充分但非必要条件，给出了与单个约束相关的 QP 问题的解析形式。这种表述消除了调用 QP 求解器的需求，显著降低了计算成本。因此，这一优势促使将其整合到 RL 框架中。此外，闭式解法避免了在 RL 框架内对可微优化的要求，从而大大简化了安全策略生成中的梯度计算，并减轻了基于梯度的优化复杂性，这将在下一小节详细阐述。

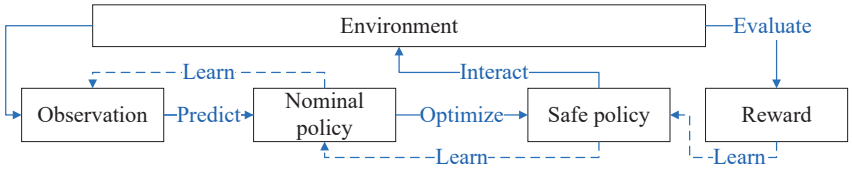

图 1: 一个端到端训练安全强化学习框架的示例。

3.3. 通过闭式解在 RL 框架中的安全层

在传统的 RL 架构中，控制策略网络的最后一层通常由全连接层组成，特别是在仿射系统的连续控制动作中。输出受到最后的激活函数（如双曲正切）的限制，以确保行动输出有界。为了生成安全的策略，一种直观的方法是通过调整基于安全机制的方式来校正从 RL 导出的控制策略，例如使用 CBF 基础的 QPCheng et al. (2019) 将控制策略校正为安全策略。然而，在这种方法中，由于没有连接安全策略与 RL 策略的梯度路径，来自安全输出的奖励无法反向传播到 RL 网络。最近，决策导向学习提出了基于可微优化的不同架构，嵌入可优化和可微结构以实现端到端的学习流程，从而实现 CBF-QP 基础的安全学习，如图 1 所示。这引发了一个关键

挑战：每个训练步骤都需要解决一系列 QP 问题来计算策略损失函数，并且需要计算每个 QP 输出关于 QP 参数的梯度，这使得计算成本高昂并且对大规模多约束问题具有挑战性。为了解决这个问题，我们直接将安全策略的闭式解集成到 RL 策略生成流程中。这种方法利用了复合单约束近似处理多约束场景，以及明确的 QP 解决方案来绕过前向优化及其梯度反向传播。我们在 RL 策略生成的最后一层用一个分析计算得出的 “安全层” 替换，由于其分析特性，它可以整合到任何演员-评论员 RL 方法中。图 2 展示了在演员-评论员框架下具有不同安全层的安全策略网络示例。所提出的框架如图 2（a）所示，在生成安全策略前的最后一层集成了闭式解 (20) 和 (21)。作为比较，图 2（b）展示了将名义政策作为输入时，可微 QP 层计算具有多个约束的前向解，这可能是不可行且计算成本高昂的。

我们使用 SAC 方法来说明所提出的方案，在这个框架中的损失函数由以下给出

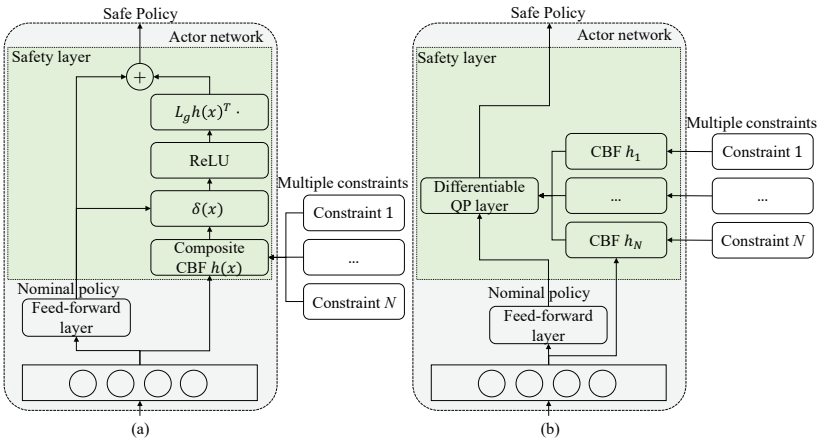

图 2: 具有不同安全层的安全策略网络的示例。子图 (a) 展示了所提出的框架，其中 N 约束通过连续 Log-Sum-Exp 近似组合成 $h(x)$。该安全层基于复合 CBF 优化的闭式解进行分析。子图 (b) 展示了具有可微 QP 层的现有框架。该安全层解决了前向 CBF 优化，并在反向传播过程中计算梯度。

$$J_Q(\theta) = \mathbb{E}_{(x_t, u_s^\phi)\sim\mathcal{D}_R}\left[\frac{1}{2}\left(Q_\theta(x_t, u_s^\phi) - \left(r(x_t, u_s^\phi) + \gamma\mathbb{E}_{x_{t+1}\sim p}\left[V_{\bar{\theta}}(x_{t+1})\right]\right)\right)^2\right], \tag{22}$$

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

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

其中 π_ϕ 表示由整个策略网络生成的策略，包括全连接层和基于 QP 的调整。在这种情况下，$u_s^\phi \sim \pi_\phi$ 意味着样本 u_s^ϕ 是从由整个策略网络定义的分布中抽取的，这其中包括了用于 QP 调整的安全层。

4. 实验

在本节中，我们旨在验证所提出的方法在确保训练过程中安全的同时，与可微分 QP 基方法相比实现更快训练效率的能力。测试环境被设计为一个可达性任务，其中代理的目标是在避开障碍物的情况下到达目标位置。为了说明在以安全为导向的优化中如何结合多个约束条件，无碰撞约束定义如下：

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

其中 $p = [p_x, p_y]^\top$ 表示代理的位置，$p_{i,\text{obs}}$ 表示第 i 个障碍物的位置，r_{safe} 是避免碰撞的安全半径。此外，在基于 CBF 的安全约束内，选择类-$\mathcal{K}$ 函数 α 显著影响安全执行的保守性，并且 κ 影响 min 操作的近似误差。在本研究中，我们采用一个权衡值来平衡安全性和性能与 $\alpha = 5(\cdot)$，$\kappa = 2$。

为了展示所提出方法的安全性，我们在训练过程中展示了以下指标：每个回合的 $\min h_i, i = 1, \ldots, I$ 和复合 h，以及 $I = 3$。此外，还展示了所有训练回合步骤中的 $\min h_i, i \in I$ 和复合 h，以及训练后的部署阶段的轨迹。结果如图 3 和图 4 所示。

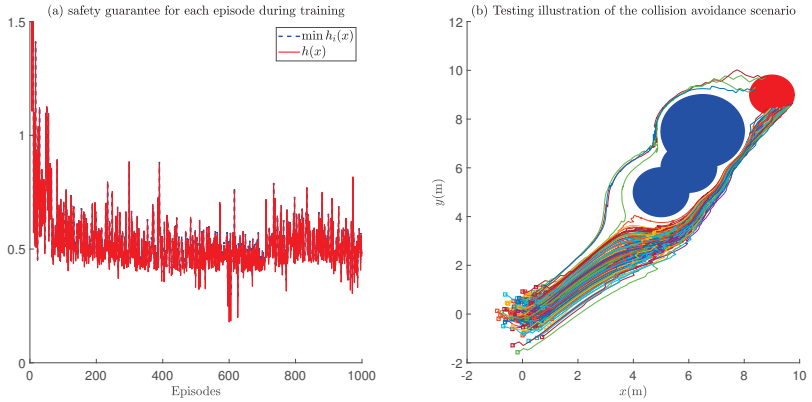

图 3: 安全强化学习训练和测试的性能。子图（a）展示了每个训练阶段中的 $\min h_i, i = 1, \ldots, I$ 以及复合 h。在每个训练阶段中，复合 $h(x)$ 低估了 $\min h_i$ 并保持正值。子图（b）展示了测试过程中的成功轨迹。子图（b）中的蓝色区域包含三个障碍物，各自具有不同的安全半径大小。彩色方块表示初始位置。红色圆圈代表目标区域，而彩色线条指示从原点附近的不同初始位置开始的测试轨迹。

如图 3(a) 所示，整个训练过程中 h_i 的最小值始终大于 0，这表明系统成功实现了安全训练和策略学习。此外，红线 $h(x)$ 作为虚线蓝色线条的下界近似，这与定理 1 的结论一致。图 3(b) 展示了达到目标区域的轨迹，在 200 次试验中确保了安全。

图 4 提供了更详细的视角，展示了每个片段中所有步骤（不同颜色的曲线）下 h_i 和 h 的演变过程（经过 1000 次训练迭代，最大步数限制为 200）。当接近安全集边界时，h_i 曲线在某些时间间隔内表现出“平坦化”校正，这取决于障碍物的不同位置。在此期间，条件 $L_f h(x) +$

$L_g h(x)\bar{u}(x) + \alpha(h(x)) < 0$ 成立。因此，安全策略是基于 (20) 和 (21) 积极过滤和校正的，确保 h 在所有时间都保持正值。

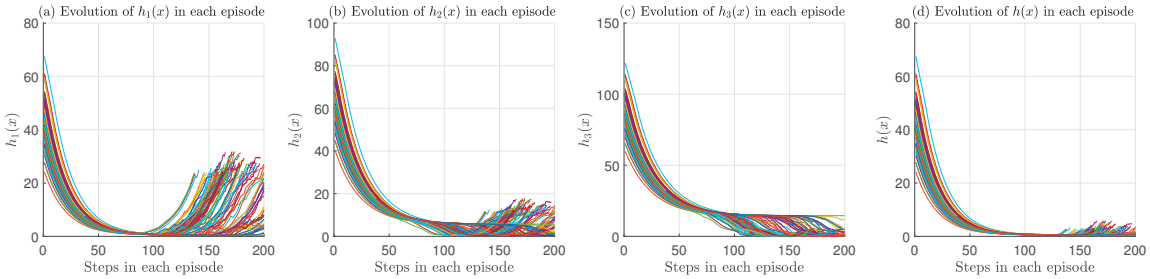

图 4: h_1, h_2, h_3 和复合 h 在 1000 个回合中的演化。训练期间的安全学习得到了保证。

表 1: 不同方法的比较

方法	ATTS(s) **具有** $I = 3$	**带有** (s) **的** ATTS $I = 10$	ATT(s) **与** $I = 30$
闭式解	0.018	0.024	0.043
CBF 批量 QP	0.13	0.25	0.40
CBF CVXPY 层	0.84	1.45	2.26

如前所述，封闭形式的解消除了对可微 QP 求解器的需求，从而降低了计算成本。这代表了所提出方法的另一个显著优势。在同一场景中进行多障碍物避碰时，我们比较了所提出的方法与常用在基于 CBF 的安全学习 Emam et al. (2022); Ma et al. (2022); Jiang et al. (2024); Romero et al. (2024) 中的可微 QP 求解器 Batched-QPFunctionAmos and Kolter (2017) 和 CVXPYLayerAgrawal et al. (2019) 的计算性能。比较结果总结于表 1 中，其中性能指标是在 RL 训练过程中的每步平均求解时间（ATTS）。此外，还测试了 10 个和 30 个约束条件的情景以验证所提出方法在解决更大规模安全 RL 问题时的可扩展性。

如表 1 所示，所提出的方法由于具有闭式性质，计算速度优势至少高出一个数量级。基于 CVXPYlayer 的方法虽然支持纪律化的参数化编程，但由于缺乏批处理求解的支持以及需要进行 QP 中的梯度计算，表现出最低的求解效率。训练时间上的优势使其可能在大规模安全强化学习问题中的优化中有效。

5. 结论

本文解决了确保多重安全约束并提高训练效率在安全 RL 中的挑战。我们提出了一种基于复合 CBF 的闭式解的安全 RL 框架。该框架通过使用 min 函数的 Log-Sum-Exp 近似来构建一个复合 CBF，以将多个安全约束整合到优化问题中。它还继承了基于定义安全集的复合 CBF 提供的安全性保证。作为具有闭式解的可微 QP 架构的替代方案，所提出的方法显著提升了训练效率。对比实验表明，所提出的方法比当前最先进的可微批量 QP 求解器快达 7 倍，并且至

少比可微凸优化层 CVXPYlayer 快 46 倍，展示了其在解决大规模安全 RL 问题中的潜力。未来工作将进一步研究显式输入约束下的复合 CBF-QP，重点在于保证框架内不同 iable 优化的可行性并提高效率。

Acknowledgments

该项工作得到国家自然科学基金（批准号 62303316）的支持，部分经费来自国家自然科学基金科学中心项目（批准号 62188101），部分经费来自中国博士后创新人才支持计划资助（批准号 BX20240224），以及上海交通大学海洋交叉学科项目（项目编号 SL2022MS010）的支持。

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 Conference, 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.