Safe and Efficient Reinforcement Learning Using Disturbance-Observer-Based Control Barrier Functions

Yikun Cheng[†1] YIKUN2@ILLINOIS.EDU
Pan Zhao[†1] PANZHAO2@ILLINOIS.EDU
Naira Hovakimyan[†] NHOVAKIM@ILLINOIS.EDU
[†] *University of Illinois at Urbana-Champaign*
[1] *Yikun Cheng and Pan Zhao contributed equally to this work.*

Editors: N. Matni, M. Morari, G. J. Pappas

Abstract

Safe reinforcement learning (RL) with assured satisfaction of hard state constraints during training has recently received a lot of attention. Safety filters, e.g., based on control barrier functions (CBFs), provide a promising way for safe RL via modifying the unsafe actions of an RL agent on the fly. Existing safety filter-based approaches typically involve learning of uncertain dynamics and quantifying the learned model error, which leads to conservative filters before a large amount of data is collected to learn a good model, thereby preventing efficient exploration. This paper presents a method for safe and efficient RL using disturbance observers (DOBs) and control barrier functions (CBFs). Unlike most existing safe RL methods that deal with hard state constraints, our method does not involve model learning, and leverages DOBs to accurately estimate the pointwise value of the uncertainty, which is then incorporated into a robust CBF condition to generate safe actions. The DOB-based CBF can be used as a safety filter with model-free RL algorithms by minimally modifying the actions of an RL agent whenever necessary to ensure safety throughout the learning process. Simulation results on a unicycle and a 2D quadrotor demonstrate that the proposed method outperforms a state-of-the-art safe RL algorithm using CBFs and Gaussian processes-based model learning, in terms of safety violation rate, and sample and computational efficiency.

Keywords: Reinforcement learning, robot safety, robust control, uncertainty estimation

1. Introduction

Reinforcement learning (RL) has demonstrated impressive performance in robotic control in recent years. Many real-world systems are subject to safety constraints. As a result, safe RL has recently received a lot of attention, although there are different definitions of "safety" Garcıa and Fernández (2015); Brunke et al. (2022). We limit our discussion to safe RL that aims to ensure satisfaction of *hard state constraints* all the time during both training and deployment.

Among different safe RL paradigms, a commonly used one is to leverage *safety filters* (SFs) to constrain the actions of RL agents and modify them whenever necessary to ensure satisfaction of safety constraints. The advantages of this paradigm mainly lie in its flexibility, i.e., a safety filter can often work with many existing RL algorithms without (many) modifications to the RL algorithms. Along this line, researchers have proposed different safety filters based on shielding Alshiekh et al. (2018), control barrier functions (CBFs) Cheng et al. (2019); Ohnishi et al. (2019); Emam et al. (2021), Hamilton-Jacobi reachability (HJR) Fisac et al. (2018), and model predictive safety certification (MPSC) Wabersich et al. (2021). Among these different SFs, the shielding SF of Alshiekh

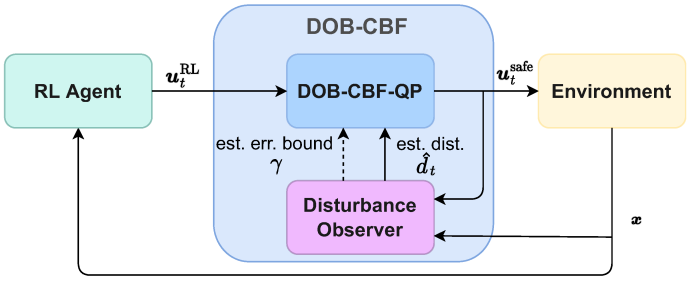

Figure 1: Proposed safe RL framework via using DOB-CBF. At each time step t, the RL agent action u_t^{RL} potentially violates the predefined safety constraints. A DOB-CBF based safety filter will render a safe action u_t^{safe} based on u_t^{RL}, a precomputed estimation error bound γ and an estimated disturbance $\hat{d}_t$. Then, u_t^{safe} is applied to interact with the environment to enforce safety during and after the policy training.

et al. (2018) only works for discrete state and action spaces. All other SFs are model-based, and hinge on Gaussian process regression (GPR) to learn the uncertain dynamics together with quantifiable prediction error for robust safety assurance. When applying these SFs to model-free RL, model learning is still needed. More importantly, due to the reliance on model learning, when the learned model is poor due to insufficient data, existing model-based SFs will be overly conservative, preventing efficient exploration of RL agents. Additionally, it is well known that GPR is computationally demanding (standard GPR model training involves computing the inverse of an $N \times N$ covariance matrix, where N is the number of data points). As a result, GPR is probably not scalable for disturbance estimation in high-dimensional systems, particularly compared to alternative methods such as disturbance observers.

This paper presents a safe RL approach using disturbance observer (DOB) based robust CBFs that were introduced in Zhao et al. (2020), and later extended in Daş and Murray (2022). As illustrated in Figure 1, our approach leverages a DOB to accurately estimate the value of the lumped disturbance at each time step with a pre-computable estimation error bound (EER). The estimated disturbance together with the EEB is incorporated into a quadratic programming (QP) module with robust CBF conditions that generates safe actions at each step by minimally modifying the RL actions. Compared to existing CBF-based safe RL approaches, e.g., Cheng et al. (2019); Ohnishi et al. (2019); Emam et al. (2021), our approach does not need model learning of the uncertain dynamics (although a nominal model is needed).

This article is organized as follows. Section 2 includes preliminaries related to DOBs, CBFs, RL. Section 3 presents the proposed safe RL framework, while Section 4 includes the simulation results for verifying the proposed framework using a unicycle and a 2D quadrotor.

2. Preliminaries

We consider a nonlinear control-affine system

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

where $x(t) \in \mathcal{X} \subset \mathbb{R}^n$ denotes the state vector, $u(t) \in \mathcal{U} \subset \mathbb{R}^m$ is the input vector, $\mathcal{X}$ and $\mathcal{U}$ are compact sets, $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are known and locally Lipschitz-continuous functions, and $d : \mathbb{R}^n \to \mathbb{R}^n$ is an unknown function that captures uncertain dynamics.

Assumption 1 *There exist positive constants l_d and b_d such that for any $x, y \in \mathcal{X}$, the following inequalities hold:*

$$\|d(x) - d(y)\| \leq l_d \|x - y\|, \tag{2}$$

$$\|d(0)\| \leq b_d. \tag{3}$$

Moreover, the constants l_d and b_d are known.

Remark 1 *Assumption 1 is a fairly standard assumption in nonlinear systems stating that the uncertain function $d(x)$ is locally Lipschitz continuous with a known bound on the Lipschitz constant in the compact set $\mathcal{X}$ and is bounded by a known constant at the origin. The Lipschitz constant and uniform bounds will affect the estimation error bound when using a DOB to estimate the uncertainty $d(x)$, as characterized in Lemma 4.*

2.1. Reinforcement Learning

Reinforcement learning aims to find an optimal policy π^* in an environment which can be formulated as a Markov decision process (MDP). In this work, an MDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, p, r)$, where the state space $\mathcal{S}$ and the action space $\mathcal{A}$ are continuous. Given the current state $s_t \in \mathcal{S}$ and action $a_t \in \mathcal{A}$, the transition function $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, \infty)$ represents the probability density of the succeeding state $s_{t+1} \in \mathcal{S}$. The reward function $r : \mathcal{S} \times \mathcal{A} \to [r_{\min}, r_{\max}]$ determines a bounded reward for each transition.

Our proposed safe RL scheme can work with any model-free RL algorithm. For illustration and experimental demonstration in Section 4, we choose soft actor-critic (SAC) Haarnoja et al. (2018), a state-of-the-art model-free RL algorithm. SAC uses an off-policy formulation that reuses historical data to improve sample efficiency and utilizes entropy maximization to improve the stability of the training process. In general, SAC aims to find a policy to maximize an entropy objective which is formed as $\sum_{t=0}^{T} \mathbb{E}_{(x_t,a_t)\sim\rho_\pi} \left[r\left(x_t, a_t\right) + \alpha \mathcal{H}\left(\pi\left(\cdot \mid x_t\right)\right)\right]$, where $\mathcal{H}(\cdot)$ is the entropy term that incentivizes exploration, α is a positive parameter to determine the relative importance of the entropy term against the reward, ρ_π denotes the states and actions distribution induced by the policy π, and T is the termination time.

2.2. Control Barrier Function

The CBFs are introduced in Ames et al. (2016) to synthesize control laws to ensure forward invariance of some sets (often related to safety) for nonlinear control-affine systems. They are often used as safety filters to modify a baseline control law to ensure that the system stays in a safety set. Consider a set

$$\mathcal{C} := \{x \in \mathbb{R}^n : h(x) \geq 0\} \subseteq \mathcal{X}, \tag{4}$$

where $h(x)$ is a continuously differentiable function h. A function $\beta : (-b, a) \to (-\infty, \infty)$ is said to belong to extended class $\mathcal{K}$ for some $a, b > 0$ if it is strictly increasing and $\beta(0) = 0$.

Definition 2 *(CBF Ames et al. (2016)). Given a set $\mathcal{C}$ defined using $h(x)$ via (4), $h(x)$ is a control barrier function for (1) if there exists an extended class $\mathcal{K}$ function β such that $\forall x \in \mathcal{C}$*

$$\sup_{u \in \mathcal{U}} \{L_f h(x) + L_g h(x) u + h_x(x) d(x)\} \geq -\beta(h(x)), \tag{5}$$

where $h_x(x) \triangleq \frac{\partial h(x)}{\partial x}$, $L_f h(x) \triangleq \frac{\partial h(x)}{\partial x} f(x)$ and $L_g h(x) \triangleq \frac{\partial h(x)}{\partial x} g(x)$.

We define the input relative degree (disturbance relative degree) of a differentiable function $h : \mathbb{R}^n \to \mathbb{R}$ with respect to (1) as the number of times we need to differentiate it along (1) until the input u (the disturbance d) explicitly shows up. Condition (5) works only for constraints with input relative degrees (IRDs) of one. To handle constraints with higher IRDs, high-order CBFs are introduced in Xiao and Belta (2021). Before introducing high-order CBFs, we make the following assumption, which indicates that the input u and the disturbance d show up together when differentiating h.

Assumption 2 *The disturbance relative degree is equal to the input relative degree.*

Define a sequence of functions $\phi_i : \mathbb{R}^n \to \mathbb{R}, i \in \{1, ..., m\}$ as:

$$\phi_i(x) = \dot{\phi}_{i-1}(x) + \beta_i(\phi_{i-1}(x)), \phi_0 = h(x). \tag{6}$$

Furthermore, define an associate sequence of sets as:

$$\mathcal{C}_i = \{x \in \mathbb{R}^n : \phi_{i-1}(x) \geq 0\} \subseteq \mathcal{X}, \ i \in \{1, \ldots, m\}. \tag{7}$$

Definition 3 *(High-Order CBF under Perturbed System Dynamics). Consider a sequential function $\phi_i(x)$ defined in (6) and a sequential set $\mathcal{C}_i, i \in \{1, ..., m\}$ defined in (7). Under Assumption 2, an m^{th}-order differentiable function $h : \mathbb{R}^n \to \mathbb{R}$ is a high-order CBF of IRD m for (1) if there exist extended differentiable class $\mathcal{K}$ functions $\beta_i, i \in \{1, ..., m\}$, such that $\forall x \in \mathcal{C}_1 \cap ... \cap \mathcal{C}_m$*

$$\sup_{u \in \mathcal{U}} \mathcal{L}_f^m h(x) + \mathcal{L}_g \mathcal{L}_f^{m-1} h(x) u + [\mathcal{L}_f^{m-1} h(x)]_x d(x) + O(h(x)) + \beta_m \left(\phi_{m-1}(x)\right) \geq 0, \tag{8}$$

where $L_f^m h(x) = \frac{\partial L_f^{m-1} h(x)}{\partial x} f(x)$, $\mathcal{L}_g \mathcal{L}_f^{m-1} h(x) = \frac{\partial L_f^{m-1} h(x)}{\partial x} g(x)$ and $[\mathcal{L}_f^{m-1} h(x)]_x = \frac{\partial L_f^{m-1} h(x)}{\partial x}$, and $O(h(x)) = \sum_{i=1}^{m-1} \mathcal{L}_f^i \left(\beta_{m-i} \circ \phi_{m-i-1}\right)(x)$.

The true uncertainty d in Definitions 2 and 3 is not accessible in practice. Therefore, it is impossible to evaluate whether a function $h(x)$ obeys the constraints in (5) or (8). One solution is to derive a sufficient condition for (5) or (8) using a uniform bound for the uncertainty $d(x)$, as adopted in Zhao et al. (2020); Nguyen and Sreenath (2016). In the following, we will derive an alternative sufficient condition to define the so-called DOB-CBFs.

2.3. Disturbance Observer (DOB) with a Precomputable Estimation Error Bound

Disturbance observers have been widely used in control of uncertain systems Chen et al. (2015). All different types of DOBs share a common idea, i.e., lumping all the uncertainties (that may consist of unknown parameters, unmodeled dynamics and external disturbances) together as a total disturbance and estimate its value at each time instant. In this work, we leverage the DOB presented in Zhao et al. (2020), which is inspired by the piecewise-constant (PC) adaptive law Wang et al. (2017) and (Hovakimyan and Cao, 2010, Section 3.3), and was used in $\mathcal{L}_1$ adaptive control of manned aircraft Ackerman et al. (2017, 2019), and in learning-enabled control Gahlawat et al. (2020); Cheng et al. (2022b). The DOB contains two components, i.e., a state predictor and a PC estimation law to estimate the disturbance. For the disturbed system (1), the state predictor is given by

$$\dot{\hat{x}}(t) = f(x) + g(x)u + \hat{d}(t) - a\tilde{x}, \tag{9}$$

where $\tilde{x} = \hat{x} - x$ denotes the prediction error, $a > 0$ is a constant, and $\hat{d}(t)$ is the estimated disturbance. The disturbance estimation is updated according to

$$\begin{cases} \hat{d}(t) = \hat{d}(iT), \quad t \in [iT, (i+1)T), \\ \hat{d}(iT) = -\dfrac{a}{e^{aT}-1}\tilde{x}(iT), i = 0, 1, ..., \end{cases} \tag{10}$$

where T is the estimation sampling time. The estimation error bound associated with the DOB is given next.

Lemma 4 *(Estimation Error Bound Zhao et al. (2020)) Given the uncertain system (1) subject to Assumption 1, and the DOB defined via (9) and (10), the estimation error can be bounded as*

$$\|\hat{d}(t) - d(x(t))\| \leq \delta(t) \triangleq \begin{cases} \theta \triangleq l_d \max_{x\in\mathcal{X}} \|x\| + b_d, & \forall 0 \leq t < T, \\ \gamma(T) \triangleq 2\sqrt{n}\eta T + \sqrt{n}\left(1 - e^{-aT}\right)\theta, & \forall t \geq T, \end{cases} \tag{11}$$

where $\eta \triangleq l_d(\max_{x\in\mathcal{X}, u\in\mathcal{U}} \|f(x) + g(x)u\| + \theta)$. Moreover, $\lim_{T\to 0} \gamma(T) = 0$.

Remark 5 *Lemma 4 implies that the estimated disturbance can be made arbitrarily accurate for $t \geq T$, by reducing T, the latter only subject to hardware limitations, including measurement noises.*

Remark 6 *As seen from Lemma 4, larger values of the Lipschitz bound l_d and the uniform bound b_d (introduced in Assumption 1) lead to larger estimation error bounds. On the other hand, we found that the estimation error bound in (11) could be quite conservative in most cases, potentially due to the frequent use of Lipschitz continuity properties and triangular inequalities. Therefore, for practical implementation one could leverage some empirical study, e.g., doing simulations under a few user-selected functions of $d(x)$ and determining a tighter bound than that defined in (11). The conservativeness also indicates that even if the constants l_d and b_d violate the conditions (3) to a certain extent, the error bound in (11) most likely still holds in practice.*

3. Main Approach

In this section, we first introduce high-order DOB-based CBFs (DOB-CBFs), by extending the result in Zhao et al. (2020), which only considers constraints with IRDs of one. The work in Daş and Murray (2022) considers high IRD constraints using exponential CBFs in the presence of matched uncertainties (which are injected to the system through the same channel as control inputs). Compared to Daş and Murray (2022), we do not constrain the uncertainties to be matched, and leverage high-order CBFs Xiao and Belta (2021) which are generalizations of exponential CBFs. Then, we introduce our DOB-CBF based safe RL scheme (DOB-CBF-RL).

3.1. High-Order DOB-Based Control Barrier Function (DOB-CBF)

Given the disturbance bound in Assumption 1 and the estimation error bound in Lemma 4, we have $[\mathcal{L}_f^{m-1}h(x)]_x d(x) \leq \|[\mathcal{L}_f^{m-1}h(x)]_x\|(\theta + 2\delta(t))$. The detailed proof could be found in the extended version of this paper Cheng et al. (2022a). Therefore, we have the following definition.

Definition 7 *(DOB-CBF) Consider the system in (1), the DOB defined via (9) and (10) with an estimation error bound given by (11), and a sequential function $\phi_i(x)$ defined in (6) and a sequential set $\mathcal{C}_i, i \in \{1, ..., m\}$ defined in (7). Under Assumption 2, an m^{th}-order differentiable function $h : \mathbb{R}^n \to \mathbb{R}$ is a high-order DOB-based control barrier function of relative degree m for (1) if there exist extended class $\mathcal{K}$ functions $\beta_i, i \in \{1, ..., m\}$, such that*

$$\sup_{u\in\mathcal{U}} \mathcal{L}_f^m h(x)+\mathcal{L}_g\mathcal{L}_f^{m-1}h(x)u-\|[\mathcal{L}_f^{m-1}h(x)]_x\|(\theta+2\max_{t\in[0,\infty)}\delta(t))+O(h(x))+\beta_m\left(\phi_{m-1}(x)\right) \geq 0, \tag{12}$$

for all $x \in \mathcal{C}_1 \cap ..., \cap \mathcal{C}_m$.

It is obvious that if a control input u is a solution for (12), it also satisfies (8). We next define

$$\mathcal{K}(t, x, u) \triangleq L_f^m h(x)+L_g L_f^{m-1}h(x)u+O(h(x))+[\mathcal{L}_f^{m-1}h(x)]_x\hat{d}(t)-\|[\mathcal{L}_f^{m-1}h(x)]_x\|\delta(t). \tag{13}$$

Then, the main theorem of the proposed approach is introduced as follows.

Theorem 8 *Suppose the condition (12) holds. Then, the condition*

$$\sup_{u\in\mathcal{U}} \mathcal{K}(t, x, u) \geq -\beta_m\left(\phi_{m-1}(x)\right) \tag{14}$$

is a sufficient condition for (8). It is also a necessary condition for (8) for any $t \geq T$ when $T \to 0$.

The proof of Theorem 8 could be found in Cheng et al. (2022a).

3.2. Safe RL Policy Training with DOB-CBFs

Using the condition in (14) that depends on the estimated disturbance $\hat{d}(t)$, we can compute the safe control inputs via solving a quadratic programming (QP) problem defined as

$$\begin{aligned} u_{\text{safe}} = \underset{u\in\mathbb{R}^m}{\operatorname{argmin}} \frac{1}{2}(u - u_{\text{RL}})^T P(u - u_{\text{RL}}) \ &(\textbf{DOB-CBF-QP}) \\ \text{s.t. } \mathcal{K}(t, x, u) + \beta_m\left(\phi_{m-1}(x)\right) \geq 0, \\ u \in \mathcal{U}, \end{aligned} \tag{15}$$

where P is a positive-definite weighting matrix, u_{RL} is the action of RL policy and u_{safe} is the final control input applied to the system (1) during both policy training and policy deployment. In case u_{RL}satisfies the constraints of (15) and is therefore safe, we have u_{safe}=u_{RL}; otherwise, (15) produces safe inputs that are mostly close to u_{RL}. With the DOB-CBF-QP in (15), our proposed DOB-CBF-RL scheme is summarized in Algorithm 1. At each step during training, the vanilla RL policy determines a potentially unsafe action. This action is then modified by the DOB-CBF-QP in (15) to produce a safe action u_{safe} that is applied to the environment. It is worth mentioning that the tuple $(x_t, u_t^{\text{safe}}, x_{t+1}, r_t)$ involving the safe action is added to the reply buffer $\mathcal{D}$ and used to update the policy. Using safe actions for policy training will promote the agent to learn a safe and optimal policy (although not guaranteed), which the DOB-CBF as a safety filter does not need to (frequently) intervene with.

Algorithm 1: DOB-CBF based safe RL

Input: Initial SAC policy π_θ, number of episodes N, number of steps per episode M, number of policy update G, nominal dynamics $\dot{x} = f(x) + g(x)u$, DOB defined via (9) and (10) with estimation error bound γ,

for $i = 1, ..., N$ **do**
  for $t = 1, ..., M$ **do**
    Obtain action u_t^{RL} from policy π_θ
    Obtain disturbance estimation $\hat{d}_t$ from the DOB defined via (9) and (10)
    Obtain safe action u_t^{safe} from DOB-CBF-QP defined in (15), using u_t^{RL}, γ and $\hat{d}_t$
    Take action u_t^{safe} in the environment
    Add transition $(x_t, u_t^{\text{safe}}, x_{t+1}, r_t)$ to replay buffer $\mathcal{D}$
    for $j = 1, ..., G$ **do**
      Sample mini-batch $\mathcal{B}$ from $\mathcal{D}$
      Update policy π_θ using $\mathcal{B}$
    end
  end
end

4. Simulation

We use a unicycle and a 2D quadrotor to validate the efficacy of the proposed DOB-CBF-RL method. For comparison, we also implemented a state-of-the-art safe RL method based on CBFs and GPR-based model learning (denoted as GP-CBF-RL) in Cheng et al. (2019). The policy training was performed on a desktop with an INTEL i9-9980XE and an NVIDIA 3090Ti GPU and 64GB RAM. The codes are available at https://github.com/Adriancyk/Safe-RL-with-Disturbance-Observer.

4.1. Unicycle

A unicycle model was borrowed from Emam et al. (2021) and adapted. The state $x = [p_x, p_y, \theta]^T$, where p_x and p_y denote the robot position along the x-axis and y-axis, respectively, and θ is the counterclockwise angle between the positive direction of the x-axis and the head direction of the robot. The control inputs are the linear velocity v and angular velocity ω of the system. The goal is to navigate the unicycle from the red dot to the yellow dot without colliding with any obstacles, as shown in Figure 2 (Right). The vehicle's equations of motion are given by

$$\begin{aligned} v_x &= \cos\theta(v + d_m), \\ v_y &= \sin\theta(v + d_m), \\ \dot{\theta} &= \omega, \end{aligned} \tag{16}$$

where the uncertainty d_m is introduced to mimic the slippery ground that causes the vehicle to lose partial control efficiency. We train DOB-CBF-RL policy and the GP-CBF-RL policy separately and define $h_i(x) = \frac{1}{2}((\|p_{i,\text{obs}}\| - \sqrt{p_x^2 + p_y^2})^2 - r_{i,\text{obs}}^2)$, where $i = 1, 2, 3$, $p_{i,\text{obs}}$ denotes the location of ith obstacle in xy-plane, and $r_{i,\text{obs}}$ is the radius of the ith obstacle. The parameters of the system dynamics and the details of the RL training are included in Cheng et al. (2022a).

We can see from Figure 2 (Left) that the DOB-CBF-RL policy consistently converges within 60 episodes. In comparison, it takes at least 90 episodes for the GP-CBF-RL to find an equally good

policy. Considering the use of 200 episodes to train a policy, a 30-episode gap is a considerable improvement in training efficiency. It is worth noting that compared with DOB-CBF-RL, there are more variations in the training performance across different trials under GP-CBF-RL. It indicates that DOB-CBF-RL can further improve the training stability by providing accurate disturbance estimation. Figure 2 (Right) compares the navigation performance. DOB-CBF-RL provides a more aggressive way to approach the target. In comparison, GP-CBF-RL chooses a more conservative trajectory and fails to reach the target, although the deviation is negligible. The safety violation rate for each 50 training episodes during training is listed in Table 1. One can see that DOB-CBF-RL achieves zero-violation rates during the entire training process, while GP-CBF has a higher violation rate at the initial training stage. With the estimation accuracy of GP model increasing, the violation rate gradually decreases. It is worth mentioning that in Cheng et al. (2019), the estimation error bound in GPR-based uncertainty estimation is simply determined by a constant k_δ selected purely according to the desired confidence level, $1-\delta$. The computation of the error bound in this way is incorrect, and could potentially give an underestimation of the true error bound, especially when data is limited. The underestimated error bounds lead to high safety violation rates at the initial learning stage. A more rigorous approach for error bound determination is given in Lederer et al. (2019).

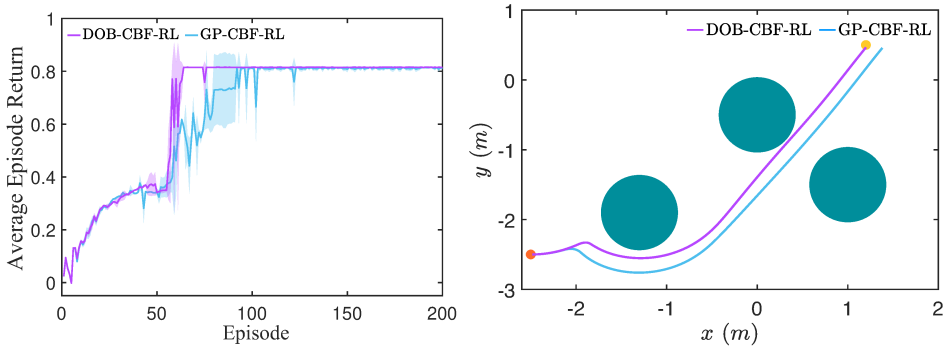

Figure 2: (Left) Unicycle training curves for DOB-CBF-RL and GP-CBF-RL. The solid lines and shaded areas denote the mean and standard deviation over five trials. (Right) Navigation performance for DOB-CBF-RL policy and GP-CBF-RL policy trained in 200 episodes.

Table 1: Safety violation rate during training for the unicycle

Training Episode	1∼50	51∼100	101∼150	151∼200
DOB-CBF	0.0%	0.0%	0.0%	0.0%
GP-CBF	12.0%	6.0%	2.0%	0.0%

4.2. 2D Quadrotor

The state of the quadrotor is $x = [p_x, v_x, p_z, v_z, \theta, \dot{\theta}]^T$, where $[p_x, p_z]$ and $[v_x, v_z]$ are the position and velocity of the quadrotor in the xz-plane, respectively, and $[\theta, \dot{\theta}]$ are the pitch angle, i.e., the angle between x direction of the quadrotor body frame and the x direction of the inertia frame, and its angular velocity, respectively. The dynamics are given as follows:

$$\begin{aligned} \dot{v_x} &= -\sin\theta\,(u_1+u_2+d_{u_1}+d_{u_2})\,/m + d_x, \\ \dot{v_z} &= \cos\theta\,(u_1+u_2+d_{u_1}+d_{u_2})\,/m - g + d_z, \\ \ddot{\theta} &= (u_2-u_1-d_{u_1}+d_{u_2})\,L/I_{yy}, \end{aligned} \tag{17}$$

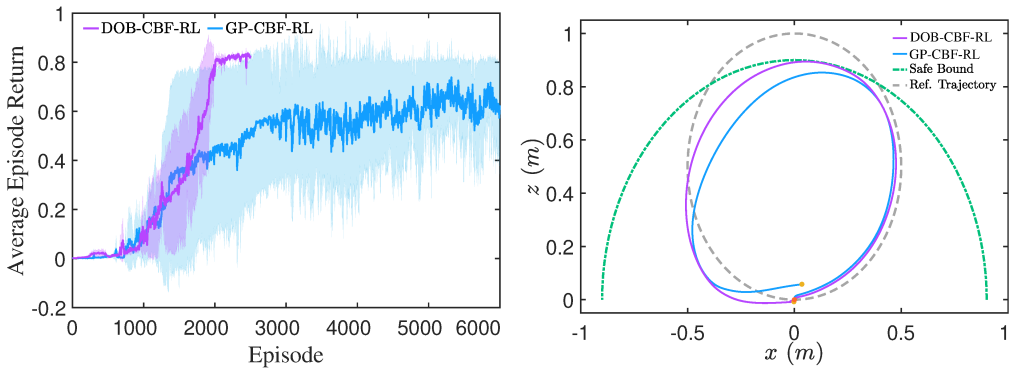

Figure 3: (Left) Quadrotor training curves for DOB-CBF-RL and GP-CBF-RL. The solid lines and shaded areas denote the mean and standard deviation over five trials. The cumulative reward is normalized. (Right) Trajectory tracking performance for DOB-CBF-RL policy trained in 2500 episodes and GP-CBF-RL policy trained in 5000 episodes.

where g is the gravitational acceleration, m denotes the total mass of the quadrotor, L is the effective moment arm, I_{yy} is the moment of inertia around y-axis, d_{u_1} and d_{u_2} are uncertainties to mimic the rotational friction of motors, and d_x and d_z denote the air resistance along each axis. To be realistic, we impose the constrained control input $u_i \in [0, u_{\max}]$ for $i = 1, 2$, where $u_{\max} = 2$ N is the maximum thrust force generated by each rotor. The objective is to control the quadrotor to track a reference trajectory (denoted by the gray line in Fig. 3) while staying within a circle boundary with a radius $r_{\text{bnd}} = 0.85$m. For DOB-CBF design, we first chose a function $h(x) = \frac{1}{2}(r_{\text{bnd}}^2 - (p_x^2 + p_y^2))$ as a candidate high-order CBF function for the nominal (i.e., uncertainty-free) system in the absence of control limits. The details about the dynamics, RL training and simulation settings can be found in Cheng et al. (2022a). Figure 3 (Left) shows that the DOB-CBF-RL method can significantly improve the training efficiency, allowing the SAC policy to converge in less than two-thousand episodes. In any case, the GP-CBF-RL method failed to find an equally good policy in 6000 episodes in most trials. One can see from Figure 3 (Right) that DOB-CBF-RL enables the agent to generate a more aggressive trajectory. Knowing the accurate disturbance estimation, the policy trained with DOB-CBF-RL pushes the agent to finish the task as perfectly as possible, while still enforces the safety of the quadrotor.

Figure 4 (Left) shows the disturbance estimation result at different training steps. DOB shows a relatively stable and decent estimation performance starting from the beginning and consistently yields an estimation error that is smaller than 5%, while the GP model gradually decreases the error and yields larger estimation error even at 6×10^5 steps. It is well known that GP model training involves computing a $N \times N$ covariance matrix Σ, where N is the number of data points, which is computationally expensive when N is large. Figure 4 (Right) shows the average computation time per one thousand training steps, from which the computation time of GP-CBF-RL is at least about three times longer than the computation time of DOB-CBF-RL. To better validate the safe exploration feature, we compute the safety violation rates for every 500 episodes during training and summarize the results in Table 2. The "w/ pre-training" means we first trained a vanilla policy using nominal dynamics and used the pre-trained policy as the starting point for GP-CBF-RL. We can see from Table 2, that DOB-CBF-RL shows an overwhelming advantage over the GP-CBF-RL method. Without pre-training, GP-CBF-RL shows significant safety guarantee performance at the initial training stage. In "w/ pre-training" case, GP-CBF still doesn't demonstrate evenly matched performance as DOB-CBF while its violation rates have been significantly lowered by introducing

a pretrained policy. Theoretically, DOB-CBF-RL is supposed to guarantee zero safety violations by leveraging a DOB-CBF function defined in Theorem 7. However, verifying whether a given function is a DOB-CBF, especially in the presence of control limits, is still a challenging problem. In other words, the intuitively selected function h may not be a DOB-CBF in the presence of the uncertainties and control limits. As a result, the rigorous safety guarantee provided by our DOB-CBF-RL framework is lost. However, compared to GP-CBF-RL, our DOB-CBF framework still achieves much lower constraint violation rate throughout the learning phase.

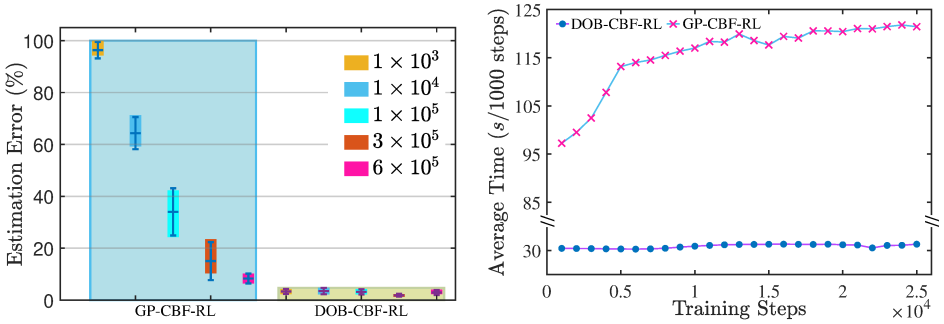

Figure 4: (Left) Disturbance estimation percentage error yielded by DOB and GP models at different steps during training. Five trials were performed and the mean with the standard deviation is shown as an error bar at each test step. The color box attached to each error bar denotes the worst and best estimation error at each shown step. For better illustration, we categorize all cases where the estimation percentage error is greater than 100% as instances of 100% percentage error cases. (Right) Average computation time per 1000 steps. Solid lines with markers plot the average computation time per 1000 steps at nth training step.

Table 2: Safety violation rate during training for 2D quadrotor

Training Episode	1∼500	501∼1000	1001∼1500	1501∼2000
DOB-CBF	15.8%	2.6%	0.2%	0.0%
GP-CBF	91.4%	79.6%	59.8%	40.4%
GP-CBF(w/ pre-training)	30.8%	22.8%	20.0%	7.8%

5. Conclusion and Future Work

This paper presents a safe and efficient reinforcement learning (RL) scheme based on disturbance observers (DOBs) and control barrier functions (CBFs). Our approach leverages a DOB that can accurately estimate the pointwise value of the uncertainty, and a quadratic programming (QP) module with a robust CBF condition, to generate safe actions by minimally modifying the (potentially unsafe) actions generated by the RL policy. Unlike existing safe RL approaches based on CBFs, which often rely on model learning of the uncertain dynamics, our approach completely removes the need for model learning and facilitates more sample- and computationally-efficient policy training. The efficacy of our proposed scheme is validated in simulated environments, in comparison with an existing CBF-based safe RL approach.

Our future work includes experimental validation of the proposed DOB-CBF-RL framework on a real robot, and extension of the framework to model-based RL settings.

6. Acknowledgments

This work is supported in part by the AFOSR through Grant FA9550-21-1-0411, in part by the NASA through the ULI Grant 80NSSC17M0051, in part by NSF through the RI Grant 2133656 and the CMMI Grant 2135925, and in part by the HUJI/UIUC Joint Research and Innovation Seed Grant.

References

Kasey Ackerman, Javier Puig-Navarro, Naira Hovakimyan, M. Christopher Cotting, Dustin J. Duke, Miguel J. Carrera, Nathan C. McCaskey, Dario Esposito, Jessica M. Peterson, and Jonathan R. Tellefsen. Recovery of desired flying characteristics with an $\mathcal{L}_1$ adaptive control law: Flight test results on Calspan's VSS Learjet. In *AIAA SciTech 2019 Forum*, San Diego, California, January 2019. AIAA 2019-1084.

Kasey A. Ackerman, Enric Xargay, Ronald Choe, Naira Hovakimyan, M. Christopher Cotting, Robert B. Jeffrey, Margaret P. Blackstun, T. Paul Fulkerson, Timothy R. Lau, and Shawn S. Stephens. Evaluation of an $\mathcal{L}_1$ adaptive flight control law on Calspan's variable-stability Learjet. *AIAA Journal of Guidance, Control, and Dynamics*, 40(4):1051–1060, 2017.

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 2669–2678, 2018.

Aaron D Ames, Xiangru Xu, Jessy W Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Trans. Automat. Contr.*, 62(8):3861–3876, 2016.

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

Wen-Hua Chen, Jun Yang, Lei Guo, and Shihua Li. Disturbance-observer-based control and related methods—-An overview. *IEEE Transactions on Industrial Electronics*, 63(2):1083–1095, 2015.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3387–3395, 2019.

Yikun Cheng, Pan Zhao, and Naira Hovakimyan. Safe and efficient reinforcement learning using disturbance-observer-based control barrier functions. *arXiv preprint arXiv:2211.17250*, 2022a.

Yikun Cheng, Pan Zhao, Fanxin Wang, Jerome Daniel Block, and Naira Hovakimyan. Improving the robustness of reinforcement learning policies with $\mathcal{L}_1$ adaptive control. *IEEE Robotics and Automation Letters*, 7(3):6574–6581, 2022b.

Ersin Daş and Richard M Murray. Robust safe control synthesis with disturbance observer-based control barrier functions. In *61st IEEE Conference on Decision and Control (CDC)*, pages 5566–5573, 2022.

Yousef Emam, Paul Glotfelter, Zsolt Kira, and Magnus Egerstedt. Safe model-based reinforcement learning using robust control barrier functions. *arXiv preprint arXiv:2110.05415*, 2021.

Jaime F Fisac, Anayo K Akametalu, Melanie N Zeilinger, Shahab Kaynama, Jeremy Gillula, and Claire J Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2018.

Aditya Gahlawat, Pan Zhao, Andrew Patterson, Naira Hovakimyan, and Evangelos A Theodorou. $\mathcal{L}_1$-$\mathcal{GP}$: $\mathcal{L}_1$ adaptive control with Bayesian learning. In *Conference on Learning for Dynamics and Control*, volume 120, pages 1–12, 2020.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 1861–1870, 2018.

Naira Hovakimyan and Chengyu Cao. *$\mathcal{L}_1$ Adaptive Control Theory: Guaranteed Robustness with Fast Adaptation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2010.

Armin Lederer, Jonas Umlauft, and Sandra Hirche. Uniform error bounds for Gaussian process regression with application to safe control. In *Advances in Neural Information Processing Systems*, volume 32, pages 657–667, 2019.

Quan Nguyen and Koushil Sreenath. Optimal robust control for constrained nonlinear hybrid systems with application to bipedal locomotion. In *American Control Conference*, pages 4807–4813, 2016.

Motoya Ohnishi, Li Wang, Gennaro Notomista, and Magnus Egerstedt. Barrier-certified adaptive reinforcement learning with applications to brushbot navigation. *IEEE Transactions on Robotics*, 35(5):1186–1205, 2019.

Kim Peter Wabersich, Lukas Hewing, Andrea Carron, and Melanie N Zeilinger. Probabilistic model predictive safety certification for learning-based control. *IEEE Transactions on Automatic Control*, 67(1):176–188, 2021.

Xiaofeng Wang, Lixing Yang, Yu Sun, and Kun Deng. Adaptive model predictive control of nonlinear systems with state-dependent uncertainties. *Int. J. Robust Nonlinear Control*, 27(17):4138–4153, 2017.

Wei Xiao and Calin Belta. High order control barrier functions. *IEEE Transactions on Automatic Control*, 67(7):3655–3662, 2021.

Pan Zhao, Yanbing Mao, Chuyuan Tao, Naira Hovakimyan, and Xiaofeng Wang. Adaptive robust quadratic programs using control Lyapunov and barrier functions. In *59th IEEE Conference on Decision and Control*, pages 3353–3358, 2020.