Neural Network-assisted Interval Reachability for Systems with Control Barrier Function-Based Safe Controllers

Damola Ajeyemi* DAJEYEMI@BU.EDU
Division of Systems Engineering, Boston University

Saber Jafarpour* SABER.JAFARPOUR@COLORADO.EDU
Department of Electrical, Computer, and Energy Engineering, University of Colorado Boulder

Emiliano Dall'Anese EDALLANE@BU.EDU
Department of Electrical and Computer Engineering and Division of Systems Engineering, Boston University

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

Abstract

Control Barrier Functions (CBFs) have been widely utilized in the design of optimization-based controllers and filters for dynamical systems to ensure forward invariance of a given set of safe states. While CBF-based controllers offer safety guarantees, they can compromise the performance of the system, leading to undesirable behaviors such as unbounded trajectories and emergence of locally stable spurious equilibria. Computing reachable sets for systems with CBF-based controllers is an effective approach for runtime performance and stability verification, and can potentially serve as a tool for trajectory re-planning. In this paper, we propose a computationally efficient interval reachability method for performance verification of systems with optimization-based controllers by: (i) approximating the optimization-based controller by a pre-trained neural network to avoid solving optimization problems repeatedly, and (ii) using mixed monotone theory to construct an embedding system that leverages state-of-the-art neural network verification algorithms for bounding the output of the neural network. Results in terms of closeness of solutions of trajectories of the system with the optimization-based controller and the neural network are derived. Using a single trajectory of the embedding system along with our closeness of solutions result, we obtain an over-approximation of the reachable set of the system with optimization-based controllers. Numerical results are presented to corroborate the technical findings.

Keywords: Optimization-based control; safety filters, reachability analysis; neural networks.

1. Introduction

Control Barrier Functions (CBFs) have emerged as a powerful tool for the design of optimization-based control methods for safety-critical systems. Given a desired set of *safe* states, CBF-based constraints can be embedded into the optimization problem defining the control law to enforce forward invariance of the safe set (see, e.g., Wieland and Allgöwer (2007); Ames et al. (2019); Xiao et al. (2023); Garg et al. (2024)). CBFs have been widely applied in the context of safety filters, where a nominal controller is minimally modified to satisfy CBF constraints. They have also been used alongside control Lyapunov functions to achieve both safety and stability (Ames et al. (2014); Ong and Cortés (2019)), and to address scenarios involving high-relative-degree constraints (Xiao and Belta (2019)). The optimization problem defining the controller can also include input constraints (Agrawal and Panagou (2021); Chen et al. (2024a); Cortez et al. (2021)).

* These authors contributed equally.

While CBF-based controllers and safety filters ensure forward invariance of the safe set, they can impact the performance of the system by introducing undesirable behaviors. For instance, it is well-known that designing CBF filters for systems with stabilizing nominal controllers can lead to emergence of *undesirable* equilibrium points or unbounded trajectories. Moreover, some of these undesirable equilibria may even be locally stable (see, e.g., Reis et al. (2021); Cortez and Dimarogonas (2022); Tan and Dimarogonas (2024); Chen et al. (2024b)), and their stability properties cannot be changed by simply changing the CBF (Chen et al. (2024a)).

Given these potential undesirable behaviors, this paper proposes using feed-forward neural network (FNN) to estimate the optimization-based controllers and leveraging reachability analysis for their performance and stability verification in real time. Continuously estimating the reachable sets during the operation of the system has several benefits: *(i)* it provides predictive information on the system's evolution and performance, identifying potential undesirable behaviors such as convergence to undesirable equilibria or loss of controllability (Reis et al., 2021; Garg et al., 2024); *(ii)* it informs high-level trajectory (re)planners (Matni et al. (2024)), adaptation strategies (Black and Panagou (2023)), or reach-avoid strategies (Fisac et al. (2015); Landry et al. (2018)). Real-time safety *and* performance verification for optimization-based controllers is a challenging problem (Garg et al. (2024)) due to several factors: (a) the existing methods are restrictive and often not applicable to the cases where the output of the optimization-based controller cannot be computed in closed form; (b) generating trajectories of these systems can be computationally intractable as one needs to solve a continuum of optimization problems.

Reachability and CBF-based filters. In the context of safety filters, reachability analysis has been used to refine a CBF that is safe and not overly conservative in, e.g., Tonkens and Herbert (2022); it has also been used in Tonkens et al. (2023) with similar offline design purposes in the context of safe value functions. In both cases, Hamilton-Jacobi (HJ) reachability methods were used (Wabersich et al. (2023)). HJ reachability was also used in Choi et al. (2021) and Gong et al. (2022) to construct a control Lyapunov value function to stabilize a point of interest, and in Kumar et al. (2023) to construct an implicit CBF through HJ reach-avoid differential dynamic programming. Despite offering high accuracy in estimating reachable sets, HJ-based approaches do not scale well with the size of the system and can become computationally heavy for real-time implementation. Other reachability approaches have been used in Abate and Coogan (2020); Srinivasan et al. (2020) and Llanes et al. (2022) for safety assurance using CBFs.

Reachability of FNN-controlled systems. Reachability of systems with neural network controllers have been studied extensively in the literature. Existing approaches for linear systems include NNV (Tran et al., 2020), and simulation-guided interval analysis (Xiang et al., 2021), ReachLP (Everett et al., 2021), Reach-SDP (Hu et al., 2020). For nonlinear systems, the existing approaches include ReachNN (Huang et al., 2019), Sherlock (Dutta et al., 2019), Verisig 2.0 (Ivanov et al., 2021), POLAR (Huang et al., 2022), JuliaReach (Schilling et al., 2022), and mixed integer programming in (Sidrane et al., 2022)

Contributions. We propose a computationally efficient framework for stability and performance verification in systems with optimization-based controllers using reachability analysis. Our approach is based on two main ingredients. First, we approximate the optimal solution map of the optimization problem with a feedforward neural network (FNN) and leverage interval analysis and mixed monotonicity theory to over-approximate the reachable sets of the FNN-approximated system. Second, we present a novel result on the *closeness of trajectories* between the system with the optimization-based controller and the system with the FNN approximation. This result enables us to

over-approximate the reachable sets of the original system with the optimization-based controller. For the reachability analysis of the approximated system with the FNN, we construct an embedding system by integrating state-of-the-art neural network verification methods with suitable inclusion functions for the open-loop system. This construction of the embedding system is inspired by Jafarpour et al. (2023), where interval reachability for safety verification of neural network controlled system was investigated. In this setting, a single trajectory of the embedding system provides a hyper-rectangular over-approximation of the reachable sets for the system with the FNN approximation. We then combine these reachable set estimates with our closeness of trajectories result to obtain an over-approximation of the reachable sets for the original system with optimization-based controller. This over-approximation is expressed as the Minkowski sum of a hyper-rectangle and a ball (defined using an arbitrary norm). Our strategy eliminates the need to repeatedly solve optimization problems to compute the controller's input. Instead, we compute a hyper-rectangular over-approximation of the reachable set for the FNN-approximated system using a single trajectory of the embedding system. Through numerical experiments, we demonstrate that the proposed method is computationally efficient, enabling estimation of multiple reachable sets for the system starting from multiple initial hyper-rectangles, all within a short computation time.

2. Preliminaries and Problem Statement

2.1. Notation and Definitions

We denote the set of real numbers, non-negative real numbers, and natural numbers by $\mathbb{R}$, $\mathbb{R}_{\geq 0}$, and $\mathbb{N}$, respectively. Vectors are represented using lowercase letters $x \in \mathbb{R}^n$, x_i is the ith entry of x, while matrices use uppercase letters (*e.g.*, $A \in \mathbb{R}^{n\times m}$). Given two sets $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^n$, we define the Minkowski sum of the sets $\mathcal{X}$ and $\mathcal{Y}$ by $\mathcal{X} \oplus \mathcal{Y} = \{x + y \mid x \in \mathcal{X}, \quad y \in \mathcal{Y}\}$. Given a norm $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, the induced norm of a matrix A is denoted by $\|A\|$, and the associated norm ball with radius r centered at $z \in \mathbb{R}^n$ is denoted by $\mathcal{B}_{\|\cdot\|}(r, z) = \{x \in \mathbb{R}^n \mid \|x - z\| \leq r\}$. Given a positive definite matrix $P \in \mathbb{R}^{n\times n}$, the weighted ℓ_2 norm is defined as $\|x\|_P = \sqrt{x^\top P x}$. For a scalar function $g : \mathbb{R}^n \to \mathbb{R}$, its gradient is denoted by $\nabla g(x)$, and its Hessian matrix by $\nabla^2 g(x)$. For a vector-valued function $g : \mathbb{R}^n \to \mathbb{R}^m$, the Jacobian matrix is represented by $\frac{\partial g(x)}{\partial x}$. A continuous function $\alpha : [0, a) \to \mathbb{R}_{\geq 0}$ is of class $\mathcal{K}$ if $\alpha(0) = 0$ and it is strictly increasing. A function belongs to class $\mathcal{K}_\infty$ if $a = \infty$ and $\alpha(s) \to \infty$ as $s \to \infty$. A continuous function $\alpha : \mathbb{R} \to \mathbb{R}$ is an extended class-$\mathcal{K}$ function if $\alpha(0) = 0$, it is strictly increasing, and $\alpha(s) \to \infty$ as $s \to \infty$. Given a norm $\|\cdot\| : \mathbb{R} \to \mathbb{R}_{\geq 0}$ on $\mathbb{R}^n$ and a continuous map $F : \mathbb{R}^n \to \mathbb{R}$, we denote as $\mathrm{osLip}(f)$ the (minimal) one-sided Lipschitz constant of f (Bullo, 2024, Sec. 3). Given the vectors $x, z \in \mathbb{R}^n$, $(x, z) \in \mathbb{R}^{2n}$ is their vector concatenation; i.e., $(x, z) = [x^\top, z^\top]^\top$. The partial order $\leq$ on $\mathbb{R}^n$ if defined as $x \leq z$ is and only if $x_i \leq z_i$ for all $i = 1, \ldots, n$. For any $x, z \in \mathbb{R}^n$, we define the interval $[x, z] := \{w \in \mathbb{R}^n : x \leq w \leq z\}$. For $z, w \in \mathbb{R}^n$ and every $i \in \{1, \ldots, n\}$, we define the vector $z_{[i:w]}$ with entries $z_{[i:w],i} = z_j$ if $j \neq i$ and $z_{[i:w],i} = w_j$ if $j = i$. Finally, given a matrix $A \in \mathbb{R}^{n\times n}$, we denote the non-negative part of A as $[A]^+ := \max\{A, 0\}$ and the non-positive part of A by $[A]^- := \min\{A, 0\}$, where the max and min are taken entry-wise.

2.2. Main setup and problem statement

We consider a control-affine dynamical system of the form:

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

where $x \in \mathbb{R}^n$ is state of the system, $u \in \mathbb{R}^m$ is the control input, and the functions $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ are continuously differentiable. Suppose that a locally-Lipschitz nominal controller $\kappa : \mathbb{R}^n \to \mathbb{R}^m$ is designed so that the system $\dot{x} = \tilde{f}(x) := f(x) + g(x)\kappa(x)$ has a unique equilibrium, and it renders this equilibrium globally asymptotically stable. In the remainder of the paper, we assume without loss of generality that such equilibrium is the origin. We consider the case where it is desirable for the system to operate within a given *safe* set; to this end, in the following we introduce the notion of CBF (see, e.g., Ames et al. (2019); Xiao et al. (2023)).

Definition 1 (CBF) Let $\mathcal{S}_i \subset \mathbb{R}^n$ be a subset of $\mathbb{R}^n$. Let $h_i : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function such that $\mathcal{S}_i = \{x \in \mathbb{R}^n : h_i(x) \geq 0\}$ and $\partial\mathcal{S}_i = \{x \in \mathbb{R}^n : h_i(x) = 0\}$. The function h_i is a CBF of $\mathcal{S}_i$ for the system (1) if there exists an extended class $\mathcal{K}_\infty$ function α_i such that, for all $x \in \mathcal{S}_i$, $\exists\, u \in \mathbb{R}^m$ satisfying $\nabla h_i(x)^\top (f(x) + g(x)u) + \alpha_i(h_i(x)) \geq 0$. □

Given N continuously differentiable functions $h_i : \mathbb{R}^n \to \mathbb{R}$ defining the sets $\{\mathcal{S}_i\}_{i=1}^N$ as in Def. 1, we define the safe set for the system (1) as $\mathcal{S} := \bigcap_{i=1}^N \mathcal{S}_i$. With this definition of CBF and safe set, hereafter we refer to the *filtered* system as:

$$\dot{x} = \tilde{f}(x) + g(x)v(x), \tag{2}$$

where the map $x \mapsto v(x)$ is defined as:

$$v(x) := \arg\min_{\theta\in\mathbb{R}^m} \frac{1}{2}\|\theta\|^2, \tag{3a}$$

$$\text{s.to: } \ell_i(\theta, x) \leq 0, \quad i = 1, \ldots, L, \tag{3b}$$

$$\nabla h_i(x)(\tilde{f}(x) + g(x)\theta) + \alpha_i(h_i(x)) \geq 0, \quad \forall\, i = 1, \ldots N \tag{3c}$$

with the constraints (3c) embed the CBFs of the sets $\{\mathcal{S}_i\}_{i=1}^N$ while the constraints $\ell_i(\theta, x) \leq 0$ specific additional performance and operational requirements; for example, (3b) may include Control Lyapunov Function (CLF) constraints, input constraints, etc. We make the following assumptions.

Assumption 1 *For each $i = 1, \ldots, N$, $\nabla h_i(x) \neq 0$ for all $x \in \partial\mathcal{S}_i$. The set $\mathcal{S} = \bigcap_{i=1}^N \mathcal{S}_i$ is non-empty, and the origin is in its interior.* □

Assumption 2 *For any given $x \in \mathbb{R}^n$, each of the functions $\ell_i(\theta, x)$ is convex and continuously differentiable in θ. Additionally, for any given $x \in \mathbb{R}^n$, problem (3) is feasible, and it satisfies the Mangasarian-Fromovitz constraint qualification and the constant-rank condition.* □

In this setup, (3) is a parametric convex program with a strongly convex cost; hence, it has a unique (globally) optimal solution for any given x. If, additionally, the functions $\ell_i(\theta, x)$ are linear in θ, then (3) is a linearly-constrained quadratic program (QP). Assumption 2 ensures that the map $x \mapsto v(x)$ is locally Lipschitz (Liu, 1995, Theorem 3.6); see also the conditions in Chen et al. (2024a) for safety filters designed based on only one obstacle. This, in turn, ensures existence and uniqueness of solutions to (2). Additionally, when Assumptions 1 and 2 hold, it follows from (Ames et al., 2019, Theorem 2) that the filtered system (2) renders the set $\mathcal{S}$ *forward invariant*.

Next, we introduce the notion of *undesirable* equilibrium (also referred to as spurious in, e.g., Reis et al. (2021)). We say that a point $x^*_{\texttt{unde}} \in \mathbb{R}^n$ is an undesirable equilibrium if $\tilde{f}(x^*_{\texttt{unde}}) + g(x^*_{\texttt{unde}})v(x^*_{\texttt{unde}}) = 0$, and $\tilde{f}(x^*_{\texttt{unde}}) \neq 0$; that is, $x^*_{\texttt{unde}}$ us an equilibrium of the filtered system (2) but not of the system under the nominal controller $\dot{x} = \tilde{f}(x)$. We recall that, without loss of generality, the only equilibrium of the system under the nominal controller is the origin.

Let $\phi(t, x_0)$ denote the state of the dynamical system (2) at time $t \geq 0$, when starting from the state $x(0) = x_0 \in \mathbb{R}^n$. Then, given a set of initial states $\mathcal{X}_0 \subset \mathbb{R}^n$, we define the *reachable set* of (2) at time t when starting from $\mathcal{X}_0$ as $\mathcal{R}_{\texttt{fs}}(t, \mathcal{X}_0) := \{\phi(t, x_0) : x_0 \in \mathcal{X}_0\}$, with $t \geq 0$. Since our systems are time-invariant, the notion of reachable set extends to states that are reachable at $t_1 \geq t_0$, when starting from a set of points at time $t_0 \geq 0$.

As explained in Section 1, CBF-based controllers can introduce undesirable behaviors, including the emergence of spurious equilibria. Continuously estimating the reachable set of (2) can provide benefits during the *real-time* system operation, including identifying in real time potential undesirable behaviors. Given this, the problem addressed in this paper can be stated as follows.

Problem 1. For the system (2), develop a computationally efficient method to estimate a reachable set $\mathcal{R}_{\texttt{fs}}(t, \mathcal{X})$ (or multiple reachable sets $\mathcal{R}_{\texttt{fs}}(t, \mathcal{X}_i)$, $i = 1, \ldots, N_s$). □

In particular, we are interested in methods with low computational complexity, so that reachable sets $\mathcal{R}_{\texttt{fs}}(t, \mathcal{X})$ can be estimated efficiently and they can be continuously updated.

3. Reachability via Neural Networks and Mixed-monotonicity

3.1. Main framework

We address Problem 1 by proposing a framework to estimate the reachable sets for the filtered system (2) by: *(i)* approximating $v(x)$ using a pre-trained FNN; *(ii)* leveraging mixed monotonicity to efficiently compute approximations of reachable set of the system when $v(x)$ is approximated by the FNN; *(iii)* leveraging closeness of trajectory results to compute over-approximations of $\mathcal{R}_{\texttt{fs}}(t, \mathcal{X}_0)$.

Neural network approximation. We approximate $v(x)$ using a FNN $\mathcal{N}(x)$, trained offline to replicate optimal solutions to (3). Here, $\mathcal{N}(x)$ is structured as an H-layer FNN, defined as:

$$u = \mathcal{N}(x) := W^{(H)}\varphi^{(H)} + b^{(H)} \tag{4a}$$

$$\varphi^{(i)} = \Phi^{(i)}\left(W^{(i-1)}\varphi^{(i-1)} + b^{(i-1)}\right), \quad i = 1, \ldots, H, \text{ and } \varphi^{(0)} = x \tag{4b}$$

where $W^{(i-1)} \in \mathbb{R}^{n_i \times n_{i-1}}$ and $b^{(i-1)} \in \mathbb{R}^{n_i}$ are the weight matrix and bias vector of the ith layer of the network, n_i is the number of neurons in the ith layer, $\varphi^{(i)} \in \mathbb{R}^{n_i}$ is the ith hidden variable, and $\Phi^{(i)} : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ is the Lipschitz-continuous diagonal activation function of the ith layer. We assume that each element satisfies the inequalities $0 \leq \frac{(\Phi_\ell^{(i)}(x) - \Phi_\ell^{(i)}(y))}{x-y} \leq 1$, for any pair x, y, and for $\ell = 1, \ldots, n_i$. We note that activation functions such as the ReLU, leaky ReLU, sigmoid, and tanh, all satisfy this condition. The dynamics under the FNN approximation of the controller are then given by:

$$\dot{x}_{\texttt{nn}} = \tilde{f}(x_{\texttt{nn}}) + g(x_{\texttt{nn}})\mathcal{N}(x_{\texttt{nn}}) \tag{5}$$

where we recall that $\tilde{f}(x_{\texttt{nn}}) = f(x_{\texttt{nn}}) + g(x_{\texttt{nn}})\kappa(x_{\texttt{nn}})$, and where the subscript $\texttt{nn}$ is used to emphasize that trajectories are generated using an approximation of $v(x)$. Hereafter, similarly to (2), we let $\phi_{\texttt{nn}}(t, x_0)$ denote the state of the dynamical system (5) at time $t \geq 0$, when starting from the state $x(0) = x_0 \in \mathbb{R}^n$. Additionally, we define the reachable set of (5) at time t when starting from $\mathcal{X}_0$ as $\mathcal{R}_{\texttt{nn}}(t, \mathcal{X}_0) := \{\phi_{\texttt{nn}}(t, x_0 : x_0 \in \mathcal{X}_0\}$, with $t \geq 0$.

Embedding system via inclusion functions. In the next step, we review the framework proposed in (Jafarpour et al., 2024) for interval reachability of the approximate system (5). The main

idea is to embed the neural network controlled system (5) into a larger dimensional space (Jafarpour et al., 2024, Theorem 3) and use a single trajectory of this embedding system to over-approximate reachable sets of the original system (5). For a given function $d : \mathbb{R}^n \to \mathbb{R}^m$ and an interval $[\underline{y}, \overline{y}]$, we introduce the notion of inclusion function of d on interval $[\underline{y}, \overline{y}]$ as the map $\mathsf{D}_{[\underline{y},\overline{y}]} = \begin{bmatrix} \underline{\mathsf{D}}_{[\underline{y},\overline{y}]} \\ \overline{\mathsf{D}}_{[\underline{y},\overline{y}]} \end{bmatrix} : \mathbb{R}^{2n} \to \mathbb{R}^{2m}$ satisfying (Jaulin et al., 2001)

$$\underline{\mathsf{D}}_{[\underline{y},\overline{y}]}(\underline{x}, \overline{x}) \leq d(x) \leq \overline{\mathsf{D}}_{[\underline{y},\overline{y}]}(\underline{x}, \overline{x}), \qquad \text{for all } x \in [\underline{x}, \overline{x}] \subseteq [\underline{y}, \overline{y}].$$

For a given function $d : \mathbb{R}^n \to \mathbb{R}^m$, there exists various methods to compute its inclusion function over a given interval including natural compositional approach (Jaulin et al., 2001), Jacobian-based methods, and decomposition-based methods (Coogan and Arcak, 2015). We refer to (Jafarpour et al., 2024) for more details on construction of inclusion functions. For a FNN $\mathcal{N} : \mathbb{R}^n \to \mathbb{R}^m$, one can alternatively obtain the inclusion function for $\mathcal{N}(x)$ using existing neural network verification algorithms including CROWN (Zhang et al., 2018), LipSDP (Fazlyab et al., 2022), and IBP (Gowal et al., 2019). In particular, some neural network verification algorithms can provide *affine* $[y, \widehat{y}]$-localized inclusion functions for $\mathcal{N}$; examples include CROWN and its variants (Zhang et al., 2018). Given an interval $[y, \widehat{y}]$, these algorithms provide a tuple $(C, \underline{d}, \overline{d})$ defining affine upper and lower bounds for the output of the neural network

$$C(y, \widehat{y})x + \underline{d}(y, \widehat{y}) \leq \mathcal{N}(x) \leq C(y, \widehat{y})x + \overline{d}(y, \widehat{y}), \tag{6}$$

for every $x \in [y, \widehat{y}]$. Consider the FFN-controlled system (5) with an inclusion function $\tilde{\mathsf{F}} = \begin{bmatrix} \underline{\tilde{\mathsf{F}}} \\ \overline{\tilde{\mathsf{F}}} \end{bmatrix} : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ for the vector field $\tilde{f}$ and an inclusion function $\mathsf{G} = \begin{bmatrix} \underline{\mathsf{G}} \\ \overline{\mathsf{G}} \end{bmatrix} : \mathbb{R}^{2n} \to \mathbb{R}^{2m}$ for the map g. We can construct the embedding system associated to the FNN-controlled system (5) as follows:

$$\begin{aligned} \dot{\underline{x}}_i &= \underline{\tilde{\mathsf{F}}}(\underline{x}, \overline{x}_{[i:\underline{x}]}) + [\overline{\mathsf{G}}(\underline{x}, \overline{x})]^+(C_{[\underline{x},\overline{x}]}x + \underline{d}_{[\underline{x},\overline{x}]}) + [\overline{\mathsf{G}}(\underline{x}, \overline{x})]^-(C_{[\underline{x},\overline{x}]}x + \overline{d}_{[\underline{x},\overline{x}]}) \\ \dot{\overline{x}}_i &= \overline{\tilde{\mathsf{F}}}(\underline{x}_{[i:\overline{x}]}, \overline{x}) + [\overline{\mathsf{G}}(\underline{x}, \overline{x})]^-(C_{[\underline{x},\overline{x}]}x + \underline{d}_{[\underline{x},\overline{x}]}) + [\overline{\mathsf{G}}(\underline{x}, \overline{x})]^+(C_{[\underline{x},\overline{x}]}x + \overline{d}_{[\underline{x},\overline{x}]}) \,. \end{aligned} \tag{7}$$

Let $\mathcal{X}_0 = [\underline{x}_0, \overline{x}_0]$ and $t \mapsto \begin{bmatrix} \underline{x}(t) \\ \overline{x}(t) \end{bmatrix}$ be a trajectory of the embedding system (7) starting from $\begin{bmatrix} \underline{x}(0) \\ \overline{x}(0) \end{bmatrix} = \begin{bmatrix} \underline{x}_0 \\ \overline{x}_0 \end{bmatrix}$. Then, by (Jafarpour et al., 2024, Theorem 4(i)), it follows that for every $t \geq 0$,

$$\mathcal{R}_{\text{nn}}(t, \mathcal{X}_0) \subseteq [\underline{x}(t), \overline{x}(t)]. \tag{8}$$

3.2. Closeness of solutions and over-approximation of the reachable sets

In this section, we obtain bounds on the distance between the trajectories of the filtered system (2) and its approximation using the FNN-controlled system (5). We begin with the following assumption on the accuracy of the FNN $x \mapsto \mathcal{N}(x)$ in approximating the map $x \mapsto v(x)$ defined in (3).

Assumption 3 (FNN approximation error) *For every norm $\|\cdot\|_{\mathcal{U}}$ on $\mathbb{R}^m$ and every compact set $\mathcal{C} \subseteq \mathbb{R}^n$, there exists a constant $M_{\mathcal{C}} \geq 0$ such that $\|\mathcal{N}(x) - v(x)\|_{\mathcal{U}} \leq M_{\mathcal{C}}$, for every $x \in \mathcal{C}$.* □

This assumption is grounded on the universal approximation capabilities of neural networks on compact sets (see the seminal works of Hornik (1991); Barron (1992) and some recent results in,

for example, Duan et al. (2023); Wang et al. (2024)). One can use Assumption 3 and the notion of incremental state-to-input bounds to obtain upper bounds on the distance between the trajectories of the filtered system (2) and their associated trajectories of the FNN-controlled system (5).

Theorem 1 (Closedness of solutions) *Consider the filtered system (2) and its approximation (5) with the FNN $\mathcal{N}(x)$ satisfying Assumption 3. Suppose that $t \mapsto \begin{bmatrix} \underline{x}_{nn}(t) \\ \overline{x}_{nn}(t) \end{bmatrix}$ is a curve such that $\mathcal{R}_{nn}(t, \mathcal{X}_0) \subseteq [\underline{x}_{nn}(t), \overline{x}_{nn}(t)]$, $\forall\, t \geq 0$, $\mathcal{X}_0 = [\underline{x}_0, \overline{x}_0]$. Let $t \mapsto x(t)$ be a trajectory of the filtered system (2) and $t \mapsto x_{nn}(t)$ be associated the trajectory of the neural network controlled system (5), for $x(0) = x_{nn}(0)$, and $\mathcal{C} \subseteq \mathbb{R}^n$ be a compact set that contains both trajectories. Then, $\forall\, t \geq 0$,*

$$\|x(t) - x_{nn}(t)\| \leq \ell M_{\mathcal{C}} \exp\left(\int_0^t c(\tau)d\tau\right) \int_0^t \exp\left(-\int_0^\tau c(s)ds\right) d\tau \tag{9}$$

where $\ell = \sup_{x\in\mathcal{C}} \|g(x)\|_i$ and $\|\cdot\|_i$ is induced norm from $\|\cdot\|$ and $\|\cdot\|_{\mathcal{U}}$ and $t \mapsto c(t)$ is a curve satisfying

$$\operatorname{osLip}\left(\tilde{f}(x) + C_{[\underline{x}_{nn}(t), \overline{x}_{nn}(t)]}x\right) \leq c(t), \qquad \text{for every } x \in \mathcal{C},\ \ t \geq 0\,. \tag{10}$$

Proof We first note that, for every $t \geq 0$ and every $x \in [\underline{x}_{\mathrm{nn}}(t), \overline{x}_{\mathrm{nn}}(t)]$, we have

$$C_{[\underline{x}_{\mathrm{nn}}(t), \overline{x}_{\mathrm{nn}}(t)]}x + \underline{d}_{[\underline{x}_{\mathrm{nn}}(t), \overline{x}_{\mathrm{nn}}(t)]} \leq \mathcal{N}(x) \leq C_{[\underline{x}_{\mathrm{nn}}(t), \overline{x}_{\mathrm{nn}}(t)]}x + \overline{d}_{[\underline{x}_{\mathrm{nn}}(t), \overline{x}_{\mathrm{nn}}(t)]}$$

This implies that $\mathcal{N}(x) = C_{[\underline{x}_{\mathrm{nn}}(t), \overline{x}_{\mathrm{nn}}(t)]}x + d(t)$, where $d(t) \in [\underline{d}_{[\underline{x}_{\mathrm{nn}}(t), \overline{x}_{\mathrm{nn}}(t)]}, \overline{d}_{[\underline{x}_{\mathrm{nn}}(t), \overline{x}_{\mathrm{nn}}(t)]}]$. Now consider the following control-affine system $\dot{x} = \tilde{f}(x) + g(x)\mathcal{N}(x) + g(x)u$. For $u(t) = 0$, the trajectory of the system starting from $x(0)$ is $t \mapsto x_{\mathrm{nn}}(t)$. For $u(t) = v(x(t)) - \mathcal{N}(x(t))$, the trajectory of the system starting from $x(0)$ is $t \mapsto x(t)$. Letting D^+ denote the Dini upper right-hand derivative, we have that (see, e.g., (Bullo, 2024, Thm. 3.16))

$$\begin{aligned} D^+\|x(t) - x_{\mathrm{nn}}(t)\| &\leq c(t)\|x(t) - x_{\mathrm{nn}}(t)\| + \ell\|v(x(t)) - \mathcal{N}(x(t))\| \\ &\leq c(t)\|x(t) - x_{\mathrm{nn}}(t)\| + \ell M_{\mathcal{C}} \end{aligned}$$

for all $t \geq 0$, where we used Assumption 3 for the second inequality. For notational simplicity, let $\xi(t) := \exp\left(\int_0^t c(s)ds\right)$, and note that $\xi(t) \geq 0$ for any $t \geq 0$. Then, we compute

$$\begin{aligned} D^+\left(\frac{\|x(t) - x_{\mathrm{nn}}(t)\|}{\xi(t)}\right) &= \frac{\left(D^+\|x(t) - x_{\mathrm{nn}}(t)\|\right)\xi(t) - \dot{\xi}(t)\|x(t) - x_{\mathrm{nn}}(t)\|}{\xi^2(t)} \\ &\leq \frac{(c(t)\|x(t) - x_{\mathrm{nn}}(t)\| + \ell M_{\mathcal{C}})\xi(t) - c(t)\xi(t)\|x(t) - x_{\mathrm{nn}}(t)\|}{\xi^2(t)} \leq \frac{\ell M_{\mathcal{C}}}{\xi(t)}\,. \end{aligned}$$

Next, integrating over time, we get $\frac{\|x(t)-x_{\mathrm{nn}}(t)\|}{\xi(t)} \leq \frac{\|x(0)-x_{\mathrm{nn}}(0)\|}{\xi(0)} + \int_0^t \frac{\ell M_{\mathcal{C}}}{\xi(s)}ds$, where we note that $\|x(0) - x_{\mathrm{nn}}(0)\| = 0$ since $x(0) = x_{\mathrm{nn}}(0)$. Therefore,

$$\|x(t) - x_{\mathrm{nn}}(t)\| \leq \ell M_{\mathcal{C}}\xi(t)\int_0^t \tfrac{1}{\xi(s)}ds = \ell M_{\mathcal{C}} \exp\left(\int_0^t c(\tau)d\tau\right) \int_0^t \exp\left(-\int_0^\tau c(s)ds\right) d\tau\,.$$

■

Using this theorem, we can find over-approximation of the reachable set of (2) using the hyper-rectangular over-approximations of reachable sets of the FNN-controlled system (5) given by (8).

Theorem 2 (Minkoswki Sum) *Consider the filtered system (2) and its approximation (5) with the FNN $\mathcal{N}(x)$ satisfying Assumption 3. Suppose that $t \mapsto \begin{bmatrix} \underline{x}_{nn}(t) \\ \overline{x}_{nn}(t) \end{bmatrix}$ is a curve such that $\mathcal{R}_{nn}(t, \mathcal{X}_0) \subseteq [\underline{x}_n(t), \overline{x}_{nn}(t)]$, for every $t \geq 0$, with $\mathcal{X}_0 = [\underline{x}_0, \overline{x}_0]$. Assume that the reachable sets $\mathcal{R}_{fs}(t, \mathcal{X}_0)$ and $\mathcal{R}_{nn}(t, \mathcal{X}_0)$ are bounded and $\mathcal{C} \subseteq \mathbb{R}^n$ is a compact set such that $\mathcal{R}_{fs}(t, \mathcal{X}_0) \cup \mathcal{R}_{nn}(t, \mathcal{X}_0) \subseteq \mathcal{C}$ for all times $t \geq 0$. Then,*

$$\mathcal{R}_{fs}(t, \mathcal{X}_0) \subseteq [\underline{x}_{nn}(t), \overline{x}_{nn}(t)] \oplus \mathcal{B}_{\|\cdot\|}(r_t, 0), \qquad \text{for all } t \geq 0, \tag{11}$$

where $\oplus$ is the Minkowski sum and $r_t = \ell M_{\mathcal{C}} \exp\left(\int_0^t c(\tau)d\tau\right) \int_0^t \exp\left(-\int_0^\tau c(s)ds\right) d\tau$ with $\ell = \sup_{x \in \mathcal{C}} \|g(x)\|_i$ where $\|\cdot\|_i$ is the induced norm from $\|\cdot\|$ and $\|\cdot\|_{\mathcal{U}}$ and $t \mapsto c(t)$ is a curve satisfying the inequality (10)

Proof The result follows by combining Theorem 1 and the hyper-rectangular bound (8) and the definition of Minkowski sum. ■

Computing the Minkowski sum of two arbitrary sets can be computationally complicated. However, when the sets are ellipsoids or polytopes, there exist efficient algorithms for estimating their Minkowski sum (Gritzmann and Sturmfels, 1993; Halder, 2018). In particular, when $\mathcal{B}_{\|\cdot\|}(r_t, 0) = \mathcal{B}_\infty(r_t, 0)$ is ℓ_∞-norm ball, the Minkowski sum (11) can be easily computed as $[\underline{x}_{\text{nn}}(t), \overline{x}_{\text{nn}}(t)] \oplus \mathcal{B}_\infty(r_t, 0) = [\underline{x}_{\text{nn}} - r_t, \overline{x}_{\text{nn}} + r_t]$.

4. Numerical Experiments

We show the efficiency of our proposed reachability method for performance verification of CBF-based controllers with two sets of numerical experiments.

Scenario 1 *(Integrator dynamics with circular obstacle).* Consider the integrator dynamics $\dot{x} = u$, where $x \in \mathbb{R}^2$ represents the state vector and $u \in \mathbb{R}^2$ is the control input. The nominal controller is designed as $\kappa(x) = Kx$, with $K = \begin{bmatrix} -1 & 0 \\ 0 & -5 \end{bmatrix}$, and it stabilizes the origin. We consider a circular obstacle located at $\boldsymbol{o} = [2, 0]^\top$ with a radius of $r = 1$, and define the safe set $\mathcal{S} = \{x \in \mathbb{R}^2 \mid h_1(x) \geq 0\}$, with CBF $h_1(x) = \|x - \boldsymbol{o}\|^2 - r^2$. By analyzing the closed-loop dynamics with the CBF filter, from Chen et al. (2024a) it follows that additional equilibria emerge at $x^*_{\text{unde}} = \left(\frac{5}{2}, \pm\frac{\sqrt{3}}{2}\right)^\top$ and $x^*_{\text{unde}} = [3, 0]^\top$. Evaluating the Jacobian $J_{h_1,\alpha}(x^*_{\text{unde}})$ at these points, it follows that that the equilibrium at $[3, 0]^\top$ is locally asymptotically stable, while the other two are saddle points. To better highlight the stability of the undesirable equilibrium, we introduce the modified CBF (a scaled, equivalent version of h_1) $h_2(x) = \left(\|x - \boldsymbol{o}_2\|^2 + 1\right) h_1(x)$, where $\boldsymbol{o}_2 = [5, 1]^\top$ (Chen et al., 2024a, Sec. 7), and we use $\alpha(s) = s$. The safety filter computes the control input $v(x)$ by using $h_2(x)$ and without additional constraints. We train a FNN offline to approximate the mapping $x \mapsto v(x)$. The neural network is structured as an FNN with architecture [$2 \times 400 \times 400 \times 400 \times 2$, ReLU activations], resulting in the closed-loop approximate system $\dot{x}_{\text{nn}} = \kappa(x_{\text{nn}}) + \mathcal{N}(x_{\text{nn}})$, as per (5). The network was trained using 99,472 data points over the region illustrated in the Figure 1, with points that are equi-spaced on a grid. The training achieved a maximum ℓ_2-norm error of 0.2 and a maximum ℓ_∞-norm error of 0.19. We simulate the system dynamics using Euler integration with a step size of 0.04.

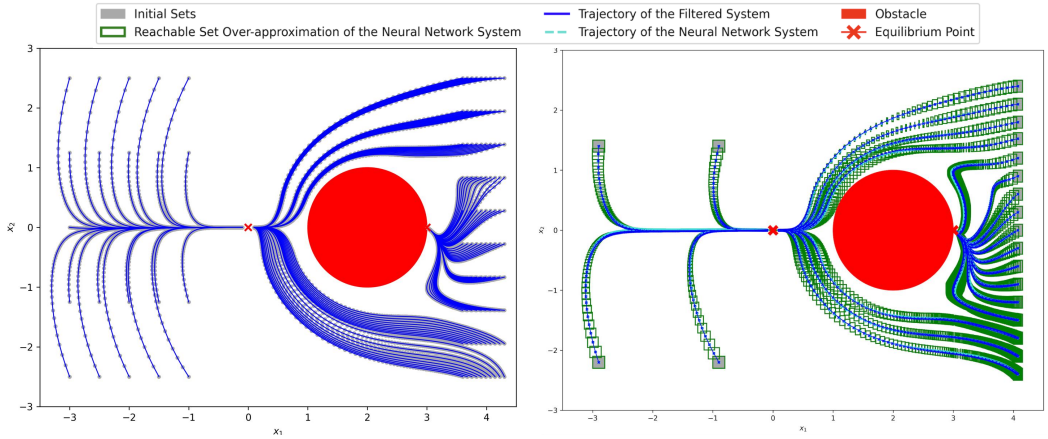

Figure 1: (**Left**) Trajectories of (2) in Scenario 1. (**Right**) System (5) with the FNN approximation, and hyper-rectangles $[\underline{x}_{\mathrm{nn}}(t), \overline{x}_{\mathrm{nn}}(t)]$ starting from several initial sets $[\underline{x}_{\mathrm{nn}}(0), \overline{x}_{\mathrm{nn}}(0)]$.

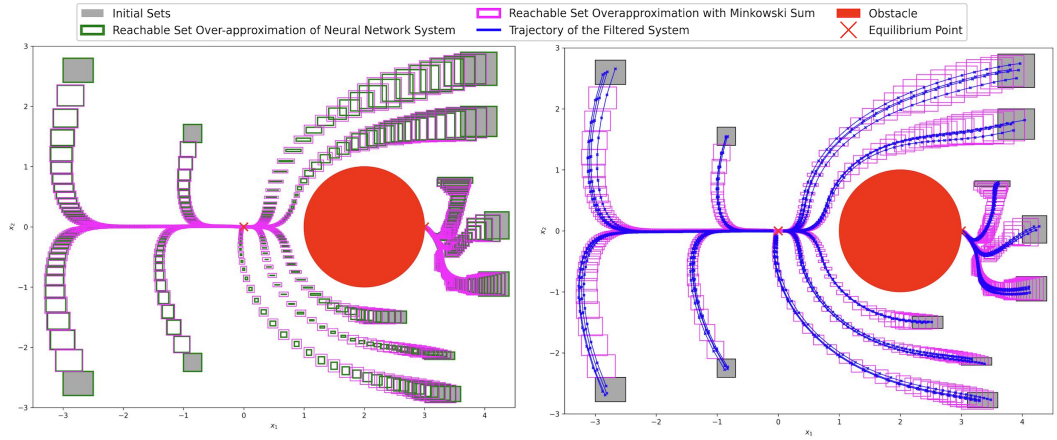

Figure 2: (**Left**) Over-approximations $[\underline{x}_{\mathrm{nn}}(t), \overline{x}_{\mathrm{nn}}(t)]$ and $[\underline{x}_{\mathrm{nn}}(t), \overline{x}_{\mathrm{nn}}(t)] \oplus \mathcal{B}(r_t, 0)$ of the reachable sets, for different initial sets, over $T = 10$ seconds. (**Right**) As expected, trajectories of the filtered system (2) are contained in $[\underline{x}_{\mathrm{nn}}(t), \overline{x}_{\mathrm{nn}}(t)] \oplus \mathcal{B}_\infty(r_t, 0)$.

Figure 1 (left) shows the trajectories of the filtered system for several initial points, over a time horizon of $T = 10$ seconds. Several trajectories converge to the undesirable equilibrium at $[3, 0]^\top$, located on the obstacle boundary. Figure 1 (right) shows trajectories $x_{\mathrm{nn}}(t)$ of the system with the FNN approximation, along with the hyper-rectangles $[\underline{x}_{\mathrm{nn}}(t_k), \overline{x}_{\mathrm{nn}}(t_k)]$, for several times t_k and for several sets of initial states $\mathcal{X}_0 = [\underline{x}_0, \overline{x}_0]$ (gray rectangles), over a time horizon of $T = 10$ seconds. The affine inclusion functions for the FNN are obtained from CROWN and computed using autoLiRPA, and we construct the embedding system as described in Section 3. Figure 2 (left) illustrates the sets $[\underline{x}_{\mathrm{nn}}(t), \overline{x}_{\mathrm{nn}}(t)]$ and $[\underline{x}_{\mathrm{nn}}(t), \overline{x}_{\mathrm{nn}}(t)] \oplus \mathcal{B}_\infty(r_t, 0)$, constructed as described in Section 3, for several initial sets $\mathcal{X}_0 = [\underline{x}_0, \overline{x}_0]$; here, $\mathcal{B}_\infty(r_t, 0)$ is computed with respect to ℓ_∞-norm using Theorem 1 with $\mathcal{C} = [-3.5, 4.5] \times [-3, 3]$ and $M_{\mathcal{C}} = 0.19$. Figure 2 (left) shows that trajectories for the filtered system (2) are contained in the sets $[\underline{x}_{\mathrm{nn}}(t), \overline{x}_{\mathrm{nn}}(t)] \oplus \mathcal{B}_\infty(r_t, 0)$. For the initial sets shown in Figure 2, our reachability framework efficiently and rigorously verifies

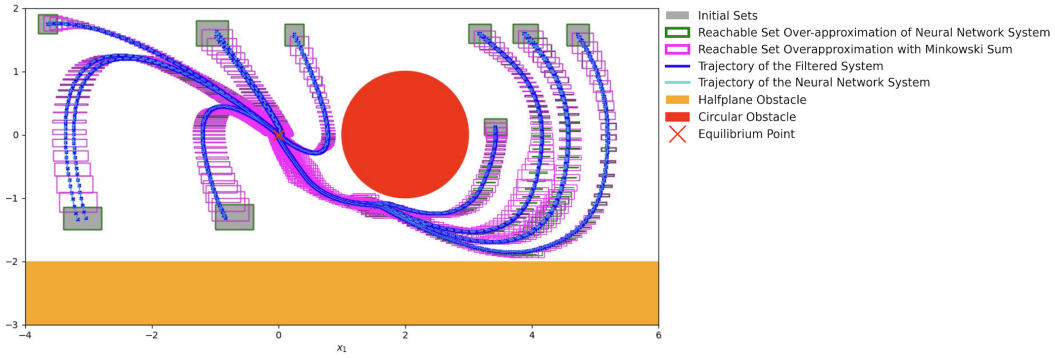

Figure 3: Trajectories of (2) in Scenario 2, hyper-rectangles $[\underline{x}_{\text{nn}}(t), \overline{x}_{\text{nn}}(t)]$, and over-approximations $[\underline{x}_{\text{nn}}(t), \overline{x}_{\text{nn}}(t)] \oplus \mathcal{B}_\infty(r_t, 0)$ of reachable sets, for different initial sets.

whether the system converges to the origin or a spurious equilibrium point. The time to compute $[\underline{x}_{\text{nn}}(t), \overline{x}_{\text{nn}}(t)]$ in Figure 2 from the 12 initial sets across all was approximately 2.15 seconds, and the time to compute the Minkowski sum was approximately 0.02 sec.

Scenario 2 *(Double integrator dynamics with multiple obstacles).* Next, we extend the experimental results to multiple obstacles. We consider a double integrator system with position x_1, velocity x_2, and an acceleration control input $u \in \mathbb{R}$. The system dynamics are expressed as $\dot{x}_1 = x_2$ and $\dot{x}_2 = u$, and the nominal control input is designed as $\kappa(x) = -k^\top x$, with $k = [1, 2]^\top$, stabilizing the closed-loop system to the origin. We define two CBF constraints, associated with two unsafe regions. The first constraint prevents collisions with a circular obstacle located at $\boldsymbol{o} = [2, 0]^\top$ with radius $r_1 = 1$. The corresponding safe set is defined as: $\mathcal{S}_1 = \{x \in \mathbb{R}^2 \mid h_1(x) \geq 0\}, \quad h_1(x) = (x_1 - 2)^2 + x_2^2 - r_1^2$. The second constraint ensures that the system remains in the half-plane $x_2 \geq \bar{x}_2 = -2$; the corresponding safe set is given by: $\mathcal{S}_2 = \{x \in \mathbb{R}^2 \mid h_2(x) \geq 0\}, \quad h_2(x) = x_2 - \bar{x}_2$. These constraints ensure that the overall safe set $\mathcal{S} = \bigcap_{i=1}^2 \mathcal{S}_i$ remains forward invariant. To approximate the optimization-based filter $v(x)$, we use 91,326 data points uniformly distributed across a grid within the region shown in Figure 3 to train a FNN with the architecture [$[2 \times 500 \times 500 \times 500 \times 1]$, ReLU activations]. The network achieved a maximum ℓ_∞-norm error of 0.26. The embedding system (7) is computed using FNN inclusion functions from CROWN and simulated with Euler integration (step size 0.05, simulation time $T = 10$ seconds). We use Theorem 2 where $\mathcal{B}_\infty(r_t, 0)$ is computed with respect to ℓ_∞-norm with $\mathcal{C} = [-4, 6] \times [-2, 2]$ and $M_{\mathcal{C}} = 0.26$. In this scenario, $v(x)$ cannot be computed in closed form, further motivating our approach. For the 9 initial sets shown in Figure 3, our reachability framework rigorously verifies that the system converges to the origin. The time to compute $[\underline{x}_{\text{nn}}(t), \overline{x}_{\text{nn}}(t)]$ from the 9 initial sets was approximately 2.7 sec. on average, and the Minkowski sum computation took approximately 0.015 sec.

5. Conclusions

We proposed a computationally efficient interval reachability method for performance verification of systems with optimization-based controllers. Our method builds on an approximation of the optimization-based controller by a pre-trained FNN, the use of an embedding system, and bounds for the distance of solutions of the system with the optimization-based controller and the FNN approximation. For future research, we plan to extend the proposed approach to systems with disturbance, and investigate reachability-based control design approaches.

Acknowledgments

This work was supported in part by the National Science Foundation award 2448264, by the Air Force Office of Scientific Research award FA9550-23-1-0740, and by the U.S. Department of Energy, Office of Electricity, Advanced Grid Modeling program.

References

M. Abate and S. Coogan. Enforcing safety at runtime for systems with disturbances. In *IEEE Conference on Decision and Control*, pages 2038–2043. IEEE, 2020.

D. Agrawal and D. Panagou. Safe control synthesis via input constrained control barrier functions. In *IEEE Conference on Decision and Control*, pages 6113–6118. IEEE, 2021.

A. Ames, J. Grizzle, and P. Tabuada. Control barrier function based quadratic programs with application to adaptive cruise control. In *IEEE Conference on Decision and Control*, pages 6271–6278. IEEE, 2014.

A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada. Control barrier functions: theory and applications. In *European Control Conference*, pages 3420–3431, Naples, Italy, 2019.

Andrew R Barron. Neural net approximation. In *Proc. 7th Yale workshop on adaptive and learning systems*, volume 1, pages 69–72, 1992.

M. Black and D. Panagou. Adaptation for validation of consolidated control barrier functions. In *IEEE Conference on Decision and Control*, pages 751–757. IEEE, 2023.

F. Bullo. *Contraction Theory for Dynamical Systems*. Kindle Direct Publishing, 1.2 edition, 2024. ISBN 979-8836646806. URL https://fbullo.github.io/ctds.

Y. Chen, P. Mestres, J. Cortes, and E. Dall'Anese. Equilibria and their stability do not depend on the control barrier function in safe optimization-based control. *arXiv preprint arXiv:2409.06808*, 2024a.

Y. Chen, P. Mestres, E. Dall'Anese, and J. Cortes. Characterization of the dynamical properties of safety filters for linear planar systems. In *IEEE Conference on Decision and Control*, 2024b.

J. Choi, D. Lee, K. Sreenath, C. Tomlin, and S. Herbert. Robust control barrier–value functions for safety-critical control. In *IEEE Conference on Decision and Control*, pages 6814–6821. IEEE, 2021.

S. Coogan and M. Arcak. Efficient finite abstraction of mixed monotone systems. In *Proceedings of the 18th International Conference on Hybrid Systems: Computation and Control*, pages 58–67, April 2015. doi: 10.1145/2728606.2728607.

W. S. Cortez and D. V. Dimarogonas. On compatibility and region of attraction for safe, stabilizing control laws. *IEEE Transactions on Automatic Control*, 67(9):7706–7712, 2022.

W. S. Cortez, X. Tan, and D. V Dimarogonas. A robust, multiple control barrier function framework for input constrained systems. *IEEE Control Systems Letters*, 6:1742–1747, 2021.

Yifei Duan, Guanghua Ji, Yongqiang Cai, et al. Minimum width of leaky-relu neural networks for uniform universal approximation. In *International Conference on Machine Learning*, pages 19460–19470. PMLR, 2023.

S. Dutta, X. Chen, and S. Sankaranarayanan. Reachability analysis for neural feedback systems using regressive polynomial rule inference. In *Proceedings of the 22nd ACM International Conference on Hybrid Systems: Computation and Control*, page 157–168, 2019. doi: 10.1145/3302504.3311807.

M. Everett, G. Habibi, C. Sun, and J. P. How. Reachability analysis of neural feedback loops. *IEEE Access*, 9:163938–163953, 2021. doi: 10.1109/ACCESS.2021.3133370.

M. Fazlyab, M. Morari, and G. J. Pappas. Safety verification and robustness analysis of neural networks via quadratic constraints and semidefinite programming. *IEEE Transactions on Automatic Control*, 67(1):1–15, 2022. doi: 10.1109/TAC.2020.3046193.

J. Fisac, M. Chen, C. Tomlin, and S. Sastry. Reach-avoid problems with time-varying dynamics, targets and constraints. In *International Conference on Hybrid Systems: Computation and Control*, pages 11–20, 2015.

K. Garg, J. Usevitch, J. Breeden, M. Black, D. Agrawal, H. Parwana, and D. Panagou. Advances in the theory of control barrier functions: Addressing practical challenges in safe control synthesis for autonomous and robotic systems. *Annual Reviews in Control*, 57:100945, 2024.

Z. Gong, M. Zhao, T. Bewley, and S. Herbert. Constructing control lyapunov-value functions using Hamilton-Jacobi reachability analysis. *IEEE Control Systems Letters*, 7:925–930, 2022.

S. Gowal, K. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, R. Arandjelovic, T. A. Mann, and P. Kohli. Scalable verified training for provably robust image classification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4841–4850, 2019. doi: 10.1109/ICCV.2019.00494.

P. Gritzmann and B. Sturmfels. Minkowski addition of polytopes: Computational complexity and applications to gröbner bases. *SIAM Journal on Discrete Mathematics*, 6(2):246–269, 1993. doi: 10.1137/0406019.

A. Halder. On the parameterized computation of minimum volume outer ellipsoid of minkowski sum of ellipsoids. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 4040–4045, 2018. doi: 10.1109/CDC.2018.8619508.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4 (2):251–257, 1991.

H. Hu, M. Fazlyab, M. Morari, and G. J. Pappas. Reach-SDP: Reachability analysis of closed-loop systems with neural network controllers via semidefinite programming. In *59th IEEE Conference on Decision and Control (CDC)*, pages 5929–5934, 2020. doi: 10.1109/CDC42340.2020.9304296.

C. Huang, J. Fan, W. Li, X. Chen, and Q. Zhu. ReachNN: Reachability analysis of neural-network controlled systems. *ACM Transactions on Embedded Computing Systems (TECS)*, 18(5s):1–22, 2019.

C. Huang, J. Fan, X. Chen, W. Li, and Q. Zhu. POLAR: A polynomial arithmetic framework for verifying neural-network controlled systems. In *Automated Technology for Verification and Analysis*, pages 414–430. Springer International Publishing, 2022.

R. Ivanov, T. Carpenter, J. Weimer, R. Alur, G. Pappas, and I. Lee. Verisig 2.0: Verification of neural network controllers using taylor model preconditioning. In *International Conference on Computer Aided Verification*, pages 249–262. Springer, 2021.

S. Jafarpour, A. Harapanahalli, and S. Coogan. Interval reachability of nonlinear dynamical systems with neural network controllers. In *Learning for Dynamics and Control Conference*, pages 12–25. PMLR, 2023.

S. Jafarpour, A. Harapanahalli, and S. Coogan. Efficient interaction-aware interval analysis of neural network feedback loops. *IEEE Transactions on Automatic Control*, pages 1–16, 2024. doi: 10.1109/TAC.2024.3420968.

L. Jaulin, M. Kieffer, O. Didrit, and É. Walter. *Applied Interval Analysis*. Springer London, 2001. doi: 10.1007/978-1-4471-0249-6.

A. R. Kumar, K.-C. Hsu, P. J Ramadge, and J. F Fisac. Fast, smooth, and safe: implicit control barrier functions through reach-avoid differential dynamic programming. *IEEE Control Systems Letters*, 2023.

B. Landry, M. Chen, S. Hemley, and M. Pavone. Reach-avoid problems via sum-or-squares optimization and dynamic programming. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4325–4332. IEEE, 2018.

J. Liu. Sensitivity analysis in nonlinear programs and variational inequalities via continuous selections. *SIAM Journal on Control and Optimization*, 33(4):1040–1060, 1995.

C. Llanes, M. Abate, and S. Coogan. Safety from fast, in-the-loop reachability with application to UAVs. In *ACM/IEEE 13th International Conference on Cyber-Physical Systems*, pages 127–136. IEEE, 2022.

N. Matni, A. Ames, and J. Doyle. A quantitative framework for layered multirate control: Toward a theory of control architecture. *IEEE Control Systems Magazine*, 44(3):52–94, 2024.

P. Ong and J. Cortés. Universal formula for smooth safe stabilization. In *IEEE Conference on Decision and Control*, pages 2373–2378. IEEE, 2019.

M. F. Reis, A. P. Aguilar, and P. Tabuada. Control barrier function-based quadratic programs introduce undesirable asymptotically stable equilibria. *IEEE Control Systems Letters*, 5(2):731–736, 2021.

C. Schilling, M. Forets, and S. Guadalupe. Verification of neural-network control systems by integrating Taylor models and zonotopes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8169–8177, 2022. doi: 10.1609/aaai.v36i7.20790.

C. Sidrane, A. Maleki, A. Irfan, and M. J. Kochenderfer. OVERT: An algorithm for safety verification of neural network control policies for nonlinear systems. *Journal of Machine Learning Research*, 23(117):1–45, 2022.

M. Srinivasan, C. Santoyo, and S. Coogan. Continuous reachability task transition using control barrier functions. *IFAC-PapersOnLine*, 53(2):9696–9701, 2020.

X. Tan and D. Dimarogonas. On the undesired equilibria induced by control barrier function based quadratic programs. *Automatica*, 159:111359, 2024.

S. Tonkens and S. Herbert. Refining control barrier functions through Hamilton-Jacobi reachability. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 13355–13362. IEEE, 2022.

S. Tonkens, A. Toofanian, Z. Qin, S. Gao, and S. Herbert. Patching neural barrier functions using Hamilton-Jacobi reachability. *arXiv preprint arXiv:2304.09850*, 2023.

H.-D. Tran, X. Yang, D. Manzanas Lopez, P. Musau, L. V. Nguyen, W. Xiang, S. Bak, and T. T. Johnson. NNV: The neural network verification tool for deep neural networks and learning-enabled cyber-physical systems. In *Computer Aided Verification*, pages 3–17. Springer International Publishing, 2020.

K. Wabersich, A. Taylor, J. Choi, K. Sreenath, C. Tomlin, A. Ames, and M. Zeilinger. Data-driven safety filters: Hamilton-jacobi reachability, control barrier functions, and predictive methods for uncertain systems. *IEEE Control Systems Magazine*, 43(5):137–177, 2023.

Jian Wang, Shujun Wu, Huaqing Zhang, Bin Yuan, Caili Dai, and Nikhil R Pal. Universal approximation abilities of a modular differentiable neural network. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

P. Wieland and F. Allgöwer. Constructive safety using control barrier functions. *IFAC Proceedings Volumes*, 40(12):462–467, 2007.

W. Xiang, H-D. Tran, X. Yang, and T. T. Johnson. Reachable set estimation for neural network control systems: A simulation-guided approach. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):1821–1830, 2021. doi: 10.1109/TNNLS.2020.2991090.

W. Xiao and C. Belta. Control barrier functions for systems with high relative degree. In *IEEE Conference on Decision and Control*, pages 474–479. IEEE, 2019.

W. Xiao, C. G Cassandras, and C. Belta. *Safe Autonomy with Control Barrier Functions: Theory and Applications*. Springer, 2023.

H. Zhang, T-W. Weng, P-Y. Chen, C-J. Hsieh, and L. Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems*, volume 31, page 4944–4953, 2018.