

# Backward Reachability Analysis for Neural Feedback Loops

Nicholas Rober, Michael Everett, and Jonathan P. How

**Abstract**—The increasing prevalence of neural networks (NNs) in safety-critical applications calls for methods to certify their behavior and guarantee safety. This paper presents a backward reachability approach for safety verification of neural feedback loops (NFLs), i.e., closed-loop systems with NN control policies. While recent works have focused on forward reachability as a strategy for safety certification of NFLs, backward reachability offers advantages over the forward strategy, particularly in obstacle avoidance scenarios. Prior works have developed techniques for backward reachability analysis for systems without NNs, but the presence of NNs in the feedback loop presents a unique set of problems due to the nonlinearities in their activation functions and because NN models are generally not invertible. To overcome these challenges, we use existing forward NN analysis tools to find affine bounds on the control inputs and solve a series of linear programs (LPs) to efficiently find an approximation of the backprojection (BP) set, i.e., the set of states for which the NN control policy will drive the system to a given target set. We present an algorithm<sup>1</sup> to iteratively find BP set estimates over a given time horizon and demonstrate the ability to reduce conservativeness in the BP set estimates by up to 88% with low additional computational cost. We use numerical results from a double integrator model to verify the efficacy of these algorithms and demonstrate the ability to certify safety for a linearized ground robot model in a collision avoidance scenario where forward reachability fails.

## I. INTRODUCTION

Neural networks (NNs) play an important role in many modern robotic systems. However, despite achieving high performance in nominal scenarios, many works have demonstrated that NNs can be sensitive to small perturbations in the input space [1], [2]. Thus, before applying NNs to safety-critical systems such as self-driving cars [3] and aircraft collision avoidance [4], there is a need for tools that provide safety guarantees, which presents computational challenges due to the high dimensionality and nonlinearities of NNs.

Numerous tools have recently been developed to analyze both NNs in isolation [5]–[12] and neural feedback loops (NFLs), e.g., closed-loop systems with NN control policies, [13]–[20]. While many of these tools focus on forward reachability [13]–[19], which certifies safety by estimating where the NN will drive the system, this work focuses on backward reachability [20], as shown in Fig. 1a. Backward reachability accomplishes safety certification by finding *backprojection* (BP) sets that define parts of the state space for which the NN will drive the system to the target set, which can be chosen to contain an obstacle. Backward reachability offers

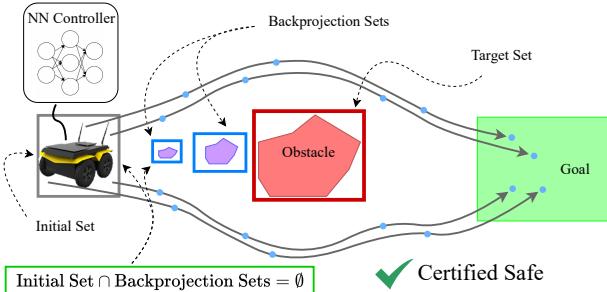
an advantage over forward reachability in scenarios where the possible future trajectories diverge in multiple directions. This phenomenon is demonstrated by the collision-avoidance scenario in Fig. 1b where forward reachability is used and the robot’s position within the initial state set determines whether the vehicle will go above or below the obstacle. When using a single convex representation of reachable sets, forward reachability analysis will be unable to certify safety because the reachable set estimates span the two sets of possible trajectories, thus intersecting with the obstacle. Conversely, as shown in Fig. 1a, backward reachability analysis correctly evaluates the situation as safe because the vehicle starts outside the avoid set’s BP, and thus the vehicle is guaranteed to avoid the obstacle. Moreover, in the ideal case that the NN control policy is always able to avoid an obstacle, the true BP set will be empty, allowing the algorithm to terminate, thereby reducing the computational cost compared to a forward reachability strategy that must calculate reachable sets for the full time horizon.

While forward and backward reachability differ only by a change of variables for systems without NNs [21]–[23], both the nonlinearities and dimensions of the matrices associated with NN controllers lead to fundamental challenges that make propagating sets backward through an NFL complicated. Despite promising prior work [11], [19], [20], there are no existing techniques that efficiently find BP set estimates over multiple timesteps for the general class of linear NFLs considered in this work. Our work addresses this issue by using a series of NN relaxations to constrain a set of linear programs (LPs) that can be used to find BP set approximations that are guaranteed to contain the true BP set. We leverage CROWN [5], an efficient open-loop NN verification tool, to generate affine bounds on the NN output for a given set of inputs. These bounds are used to constrain the system input and solve an LP maximizing the size of the BP set subject to constraints on the dynamics and control limits. The contributions of this work include:

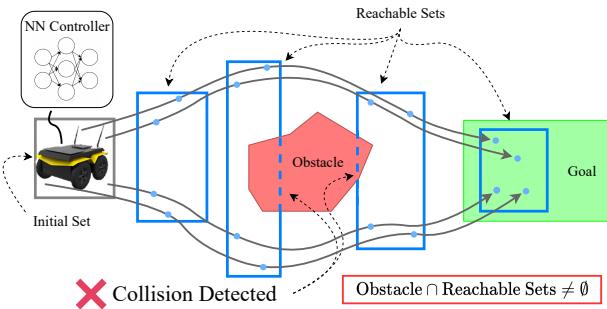
- BReach-LP: an LP-based technique to efficiently find multi-step BP over-approximations for NFLs that can be used to guarantee that the system will avoid collisions,
- ReBReach-LP: an algorithm to refine multiple one-step BP over-approximations, reducing conservativeness in the BP estimate by up to 88%,
- Numerical experiments that exhibit our BP estimation techniques with a double integrator model and a demonstration certifying safety for a linearized ground robot model whereas the forward reachability tools proposed in [18], [19] fail.

Aerospace Controls Laboratory, Massachusetts Institute of Technology, Cambridge, USA. e-mail: {nrober, mfe, jhow}@mit.edu. Research supported by Ford Motor Company.

<sup>1</sup>Code: [https://github.com/mit-acl/nn\\_robustness\\_analysis](https://github.com/mit-acl/nn_robustness_analysis)



(a) Backward reachability strategy for collision avoidance. The BP set estimates define the set of states that will lead to the obstacle, thus if the initial state set does not intersect with any BPs, the situation is safe.



(b) Forward reachability strategy for collision avoidance. The reachable set estimates define the set of possible future states the system will be in, thus any intersection of an obstacle means safety cannot be certified.

Fig. 1: Collision-avoidance scenario where backward reachability is able to correctly guarantee safety whereas forward reachability fails.

## II. RELATED WORK

Reachability analysis can be broadly categorized into three categories by system type: NNs in isolation (i.e., open-loop analysis), closed-loop systems without neural components, and NFLs.

Open-loop NN analysis encompasses techniques that relax the nonlinearities in the NN activation functions to quickly provide relatively conservative bounds on NN outputs [5], [24], and techniques that take more time to provide exact bounds [10], [11]. Many of these open-loop studies are motivated by the goal of guaranteeing robustness to adversarial attacks against perception models [8], [24], but cannot be directly used to guarantee safety for closed-loop systems because they do not consider closed-loop system dynamics.

For closed-loop systems without NNs, reachability analysis is a well established method of providing safety verification. Hamilton-Jacobi methods [21], [22], CORA [25], Flow\* [26], SpaceEx [27], and C2E2 [28] are tools that are commonly used for reachability analysis, but because they do not handle open loop NN analysis, they cannot be used to analyze NFLs.

Forward reachability analysis is the focus of many recent works [13]–[19], but while more traditional approaches to reachability analysis, e.g., Hamilton-Jacobi methods, can easily switch from forward to backward with a change of

variables [21], [22], [29]–[31], backward reachability for NFLs is less straightforward. One challenge with propagating sets backward through a NN is that many activation functions have finite range, meaning that there is not a one-to-one mapping of inputs to outputs (e.g.,  $\text{ReLU}(x) = 0$  corresponds to all values of  $x \leq 0$ ), which can cause large amounts of conservativeness in the BP set estimate as there is an infinite set of possible inputs. Additionally, even if an infinite-range activation function is used, NN weight matrices may be singular or rank deficient and are thus generally not invertible, again causing problems with determining inputs given a set of outputs. While recent works on NN inversion have developed NN architectures that are designed to be invertible [32] and training procedures that regularize for invertibility [33], dependence on these techniques would be a major limitation on the class of systems for which backward reachability analysis could be applied. Our approach avoids the challenges associated with finite-range activation functions and NN-invertibility and can be applied to the same class of NN architectures as CROWN [5], i.e. NNs for which an affine relaxation can be found.

Several recent works have investigated backwards reachability analysis for NFLs. Ref. [11] describes a method for open-loop backward reachability on a NN dynamics model, but this method does not directly apply to this work's problem of interest, namely a NN controller with known dynamics. Alternatively, while [20] analyzes NFLs, they use a quantized state approach [12] that requires an alteration of the original NN through a preprocessing step that can affect its overall behavior. Finally, previous work by the authors [19] derives a closed-form equation that can be used to find under-approximations of the BP set, but this is most useful for goal checking when it is desirable to guarantee that all states in the BP estimate will reach the target set. This work builds off of [19], adapting some of the steps used to find BP under-approximations to instead find BP over-approximations, which are good for obstacle avoidance because they contain all the states that reach the target set.

## III. PRELIMINARIES

### A. System Dynamics

We first assume that the system of interest can be described by the linear discrete-time system,

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \mathbf{c} \\ \mathbf{y}_t &= \mathbf{C}^T \mathbf{x}_t, \end{aligned} \quad (1)$$

where  $\mathbf{x}_t \in \mathbb{R}^{n_x}$ ,  $\mathbf{u}_t \in \mathbb{R}^{n_u}$ ,  $\mathbf{y}_t \in \mathbb{R}^{n_y}$  are state, control, and output vectors,  $\mathbf{A}$ ,  $\mathbf{B}$ ,  $\mathbf{C}$  are known system matrices, and  $\mathbf{c} \in \mathbb{R}^{n_x}$  is a known exogenous input. We assume the control input is constrained by control limits, i.e.,  $\mathbf{u}_t \in \mathcal{U}$ , and is determined by a state-feedback control policy  $\mathbf{u}_t = \pi(\mathbf{x}_t)$  (i.e.,  $\mathbf{C} = \mathbf{I}_{n_x}$ ) where  $\pi(\cdot)$  is an  $m$ -layer feedforward NN. Denote the closed-loop system (1) and control policy  $\pi$  as

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t; \pi). \quad (2)$$

## B. Control Policy Neural Network Structure

Consider a feedforward NN with  $L$  hidden layers and two additional layers for input and output. We denote the number of neurons in each layer as  $n_l \forall l \in [L+1]$  where  $[i]$  denotes the set  $\{0, 1, \dots, i\}$ . The  $l$ -th layer has weight matrix  $\mathbf{W}^l \in \mathbb{R}^{n_{l+1} \times n_l}$ , bias vector  $\mathbf{b}^l \in \mathbb{R}^{n_{l+1}}$ , and activation function  $\sigma^l : \mathbb{R}^{n_{l+1}} \rightarrow \mathbb{R}^{n_{l+1}}$ , where  $\sigma^l$  can be any option handled by CROWN [5], e.g., sigmoid, tanh, ReLU, etc. For an input  $\mathbf{x} \in \mathbb{R}^{n_0}$ , the NN output  $\pi(\mathbf{x})$  is computed as

$$\begin{aligned}\mathbf{x}^0 &= \mathbf{x} \\ \mathbf{z}^l &= \mathbf{W}^l \mathbf{x}^l + \mathbf{b}^l, \forall l \in [L] \\ \mathbf{x}^{l+1} &= \sigma^l(\mathbf{z}^l), \forall l \in [L-1] \\ \pi(\mathbf{x}) &= \mathbf{z}^L.\end{aligned}\quad (3)$$

## C. Neural Network Robustness Verification

To avoid the computational cost associated with calculating exact BP sets, we relax the NN's activation functions to obtain affine bounds on the NN outputs for a known set of inputs. The range of inputs are represented using the  $\ell_p$ -ball

$$\mathcal{B}_p(\dot{\mathbf{x}}, \epsilon) \triangleq \{\mathbf{x} \mid \lim_{\epsilon' \rightarrow \epsilon^+} \|(\mathbf{x} - \dot{\mathbf{x}}) \odot \epsilon'\|_p \leq 1\}, \quad (4)$$

where  $\dot{\mathbf{x}} \in \mathbb{R}^n$  is the center of the ball,  $\epsilon \in \mathbb{R}_{\geq 0}^n$  is a vector whose elements are the radii for the corresponding elements of  $\mathbf{x}$ , and  $\odot$  denotes element-wise division.

**Theorem 3.1 ([5], Convex Relaxation of NN):** Given an  $m$ -layer neural network control policy  $\pi : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_u}$ , there exist two explicit functions  $\pi_j^L : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_u}$  and  $\pi_j^U : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_u}$  such that  $\forall j \in [n_m], \forall \mathbf{x} \in \mathcal{B}_p(\dot{\mathbf{x}}, \epsilon)$ , the inequality  $\pi_j^L(\mathbf{x}) \leq \pi_j(\mathbf{x}) \leq \pi_j^U(\mathbf{x})$  holds true, where

$$\pi_j^U(\mathbf{x}) = \Psi_{j,:} \mathbf{x} + \alpha_j, \quad \pi_j^L(\mathbf{x}) = \Phi_{j,:} \mathbf{x} + \beta_j, \quad (5)$$

where  $\Psi, \Phi \in \mathbb{R}^{n_u \times n_x}$  and  $\alpha, \beta \in \mathbb{R}^{n_u}$  are defined recursively using NN weights, biases, and activations (e.g., ReLU, sigmoid, tanh), as detailed in [5].

## D. Backreachable & Backprojection Sets

The distinction between sets used in this work is shown in Fig. 2. Given a convex target set  $\mathcal{X}_T$  (right), each of the four sets on the left contain states that will reach  $\mathcal{X}_T$  under different conditions on the control input  $\mathbf{u}$ , described below.

First, the one-step backreachable set

$$\mathcal{R}_{-1}(\mathcal{X}_T) \triangleq \{\mathbf{x} \mid \exists \mathbf{u} \in \mathcal{U} \text{ s.t. } \mathbf{Ax} + \mathbf{Bu} + \mathbf{c} \in \mathcal{X}_T\}, \quad (6)$$

contains the set of all states that transition to  $\mathcal{X}_T$  in one timestep given some  $\mathbf{u} \in \mathcal{U}$ . The importance of the backreachable set (or at least an over-approximation to the backreachable set, as will be seen later) is that it only depends on the control limits  $\mathcal{U}$  and not the NN control policy  $\pi$ . Thus, while  $\mathcal{R}_{-1}(\mathcal{X}_T)$  is itself a very conservative over-approximation of the true set of states that will reach  $\mathcal{X}_T$  under  $\pi$ , it provides a region over which we can relax the NN with forward NN analysis tools, thereby avoiding issues with NN invertibility.

Next, we define the one-step true BP set as

$$\mathcal{P}_{-1}(\mathcal{X}_T) \triangleq \{\mathbf{x} \mid \mathbf{Ax} + \mathbf{B}\pi(\mathbf{x}) + \mathbf{c} \in \mathcal{X}_T\}, \quad (7)$$

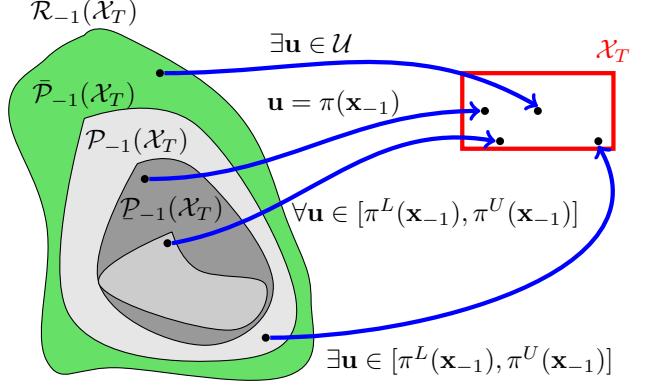


Fig. 2: Backreachable, backprojection, and target sets. Given a target set,  $\mathcal{X}_T$ , the backreachable set  $\mathcal{R}_{-1}(\mathcal{X}_T)$  contains all states for which *some* control exists to move the system to  $\mathcal{X}_T$  in one timestep. BP set  $\mathcal{P}_{-1}(\mathcal{X}_T)$  contains all states for which the NN controller leads the system to  $\mathcal{X}_T$ . BP under-approximation  $\mathcal{P}_{-1}(\mathcal{X}_T)$  and over-approximation  $\bar{\mathcal{P}}_{-1}(\mathcal{X}_T)$  contain states for which *all* and *some*, respectively, controls that the relaxed NN could apply lead the system to  $\mathcal{X}_T$ .

which denotes the set of all states that will reach  $\mathcal{X}_T$  in one timestep given the input from  $\pi$ . As previously noted, calculating  $\mathcal{P}_{-1}(\mathcal{X}_T)$  exactly is computationally intractable, which motivates the development of approximation techniques. Thus, the final two sets shown in Fig. 2 are the BP set over-approximation

$$\begin{aligned}\bar{\mathcal{P}}_{-1}(\mathcal{X}_T) &\triangleq \{\mathbf{x} \mid \exists \mathbf{u} \in [\pi^L(\mathbf{x}), \pi^U(\mathbf{x})] \text{ s.t.} \\ &\quad \mathbf{Ax} + \mathbf{Bu} + \mathbf{c} \in \mathcal{X}_T\},\end{aligned}\quad (8)$$

and BP set under-approximation

$$\begin{aligned}\mathcal{P}_{-1}(\mathcal{X}_T) &\triangleq \{\mathbf{x} \mid \forall \mathbf{u} \in [\pi^L(\mathbf{x}), \pi^U(\mathbf{x})] \text{ s.t.} \\ &\quad \mathbf{Ax} + \mathbf{Bu} + \mathbf{c} \in \mathcal{X}_T\}.\end{aligned}\quad (9)$$

Comparison of the motivation behind over- and under-approximation strategies is given in the next section.

## E. Backprojection Set: Over- vs. Under-Approximations

The need to approximate the BP set naturally leads to the question of whether we should compute over- or under-approximations. Both types of BP set approximations have relevant physical meaning and are valuable for different reasons. An under-approximation is useful if the target set is a goal set, because we aim to find a set of states at the previous timestep that will *certainly* drive the system into the goal set. This leads to a “for all” condition on the relaxed NN (i.e.,  $\forall \mathbf{u} \in [\pi^L(\mathbf{x}), \pi^U(\mathbf{x})]$ ) in (9). Conversely, an over-approximation is useful if the target set is an obstacle/avoid set, because we aim to find all states at the previous timestep for which  $\pi$  *could* drive the system into the avoid set. This leads to an “exists” condition on the relaxed NN (i.e.,  $\exists \mathbf{u} \in [\pi^L(\mathbf{x}), \pi^U(\mathbf{x})]$ ) in (8). Note that in this work we use “over-approximation” and “outer-bound” interchangeably.

Ref. [19] introduced a closed-form equation capable of one-step under-approximations of BP sets. Unfortunately,

this closed form equation hinged on the “for all” condition and thus cannot be used to generate BP set over-approximations, necessitating a different approach.

#### IV. APPROACH

This section first outlines a technique to find one-step BP set over-approximations (Algorithm 1) by solving a series of LPs. We then introduce BREach-LP, which iteratively calls Algorithm 1 to calculate BP set estimates over a desired time horizon. Finally, we propose ReBReach-LP, which further refines the BP set estimates from BREach-LP with another series of LPs, thus reducing conservativeness with some additional computational cost.

##### A. Over-Approximation of 1-Step Backprojection Sets

The proposed approach is as follows:

- 1) Ignoring the NN and using control limits  $\mathcal{U}$ , solve two LPs for each element of the state vector to find the hyper-rectangular bounds  $\bar{\mathcal{R}}_{-1}$  on the backreachable set  $\mathcal{R}_{-1}$  (note that  $\bar{\mathcal{R}}_{-1} \supseteq \mathcal{P}_{-1}$ )
- 2) Find upper/lower affine control bounds  $\pi^U(\mathbf{x}_t)$  and  $\pi^L(\mathbf{x}_t)$  by relaxing the NN controller (we use CROWN [5], but other tools, e.g., [6], [7], could also be used) within the backreachable set
- 3) Solve two LPs for each element of the state vector to compute hyper-rectangular bounds  $\bar{\mathcal{P}}_{-1}$  on the states that will lead to the target set for *some* control effort within the upper/lower bounds calculated in Step 2

The last step gives an over-approximation of the BP set, which is the set of interest.

The following lemma provides hyper-rectangle bounds on  $\bar{\mathcal{P}}_t(\mathcal{X}_T)$  for a single timestep, which is the key component of the recursive algorithm introduced in the next section.

*Lemma 4.1:* Given an  $m$ -layer NN control policy  $\pi : \mathbb{R}^{n_y} \rightarrow \mathbb{R}^{n_u}$ , closed-loop dynamics  $f : \mathbb{R}^{n_x} \times \Pi \rightarrow \mathbb{R}^{n_x}$  as in Eqs. (1) and (2), and target set  $\mathcal{X}_T$ , the set

$$\bar{\mathcal{P}}_{-1}(\mathcal{X}_T) = \{\mathbf{x}_t \mid \underline{\mathbf{x}}_t \leq \mathbf{x}_t \leq \bar{\mathbf{x}}_t\} \quad (10)$$

is a superset of the true BP set  $\mathcal{P}_{-1}(\mathcal{X}_T)$ , where  $\underline{\mathbf{x}}_t$  and  $\bar{\mathbf{x}}_t$  are computed elementwise by solving the LPs in (14).

*Proof:* Given dynamics Eqs. (1) and (2) and constraints

$$\mathcal{F}_{\mathcal{R}} \triangleq \left\{ \mathbf{x}_t, \mathbf{u}_t \mid \begin{array}{l} \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \mathbf{c} \in \mathcal{X}_T, \\ \mathbf{u}_t \in \mathcal{U}, \end{array} \right\}, \quad (11)$$

solve these optimization problems for each state  $k \in [n_x]$ ,

$$\bar{\mathbf{x}}_{t;k} = \min_{\mathbf{x}_t, \mathbf{u}_t \in \mathcal{F}_{\mathcal{R}}} \mathbf{e}_k^\top \mathbf{x}_t, \quad \underline{\mathbf{x}}_{t;k} = \max_{\mathbf{x}_t, \mathbf{u}_t \in \mathcal{F}_{\mathcal{R}}} \mathbf{e}_k^\top \mathbf{x}_t, \quad (12)$$

where the notation  $\mathbf{x}_{t;k}$  denotes the  $k^{\text{th}}$  element of  $\mathbf{x}_t$  and  $\mathbf{e}_k \in \mathbb{R}^{n_x}$  denotes the indicator vector, i.e., the vector with  $k^{\text{th}}$  element equal to one and all other elements equal to zero. Eq. (12) provides a hyper-rectangular outer bound  $\bar{\mathcal{R}}_{-1}(\mathcal{X}_T) \triangleq \{\mathbf{x} \mid \underline{\mathbf{x}}_t \leq \mathbf{x}_t \leq \bar{\mathbf{x}}_t\}$  on the backreachable set. Note that this is a LP for the convex  $\mathcal{X}_T, \mathcal{U}$  used here.

---

#### Algorithm 1 oneStepBackproj

---

**Input:** target state set  $\mathcal{X}_T$ , trained NN control policy  $\pi$ , partition parameter  $\mathbf{r}$

**Output:** BP set approximation  $\bar{\mathcal{P}}_{-1}(\mathcal{X}_T)$

- 1:  $\bar{\mathcal{P}}_{-1}(\mathcal{X}_T) \leftarrow \emptyset$
- 2:  $\bar{\mathcal{R}}_t(\mathcal{X}_T) = [\underline{\mathbf{x}}_t, \bar{\mathbf{x}}_t] \leftarrow \text{backreach}(\mathcal{X}_T, \mathcal{U})$
- 3:  $\mathcal{S} \leftarrow \text{partition}([\underline{\mathbf{x}}_t, \bar{\mathbf{x}}_t], \mathbf{r})$
- 4: **for**  $[\underline{\mathbf{x}}_t, \bar{\mathbf{x}}_t]$  in  $\mathcal{S}$  **do**
- 5:    $\Psi, \Phi, \alpha, \beta \leftarrow \text{CROWN}(\pi, [\underline{\mathbf{x}}_t, \bar{\mathbf{x}}_t])$
- 6:   **for**  $k \in n_x$  **do**
- 7:      $\bar{\mathbf{x}}_{t;k} \leftarrow \text{lpMax}(\mathcal{X}_T, \bar{\mathcal{R}}_t, \Psi, \Phi, \alpha, \beta)$
- 8:      $\underline{\mathbf{x}}_{t;k} \leftarrow \text{lpMin}(\mathcal{X}_T, \bar{\mathcal{R}}_t, \Psi, \Phi, \alpha, \beta)$
- 9:   **end for**
- 10:    $\mathcal{A} \leftarrow \{\mathbf{x} \mid \forall k \in n_x, \underline{\mathbf{x}}_{t;k} \leq \mathbf{x} \leq \bar{\mathbf{x}}_{t;k}\}$
- 11:    $\bar{\mathcal{P}}_{-1}(\mathcal{X}_T) \leftarrow \bar{\mathcal{P}}_{-1}(\mathcal{X}_T) \cup \mathcal{A}$
- 12: **end for**
- 13: **return**  $\bar{\mathcal{P}}_{-1}(\mathcal{X}_T)$

---

Given the bound,  $\underline{\mathbf{x}}_t \leq \mathbf{x}_t \leq \bar{\mathbf{x}}_t$ , Theorem 3.1 provides  $\Psi, \Phi, \alpha, \beta$ . Then, define the set of state and control constraints  $\mathcal{F}_{\bar{\mathcal{P}}}$  as

$$\mathcal{F}_{\bar{\mathcal{P}}} \triangleq \left\{ \mathbf{x}_t, \mathbf{u}_t \mid \begin{array}{l} \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \mathbf{c} \in \mathcal{X}_T, \\ \pi^L(\mathbf{x}_t) \leq \mathbf{u}_t \leq \pi^U(\mathbf{x}_t), \\ \mathbf{x}_t \in \bar{\mathcal{R}}_{-1}, \end{array} \right\} \quad (13)$$

and solve the following optimization problems for each state  $k \in [n_x]$ :

$$\bar{\mathbf{x}}_{t;k} = \min_{\mathbf{x}_t, \mathbf{u}_t \in \mathcal{F}_{\bar{\mathcal{P}}}} \mathbf{e}_k^\top \mathbf{x}_t, \quad \underline{\mathbf{x}}_{t;k} = \max_{\mathbf{x}_t, \mathbf{u}_t \in \mathcal{F}_{\bar{\mathcal{P}}}} \mathbf{e}_k^\top \mathbf{x}_t. \quad (14)$$

The LPs solved in (14) provide a hyper-rectangular outer bound of the BP set, i.e.,

$$\bar{\mathcal{P}}_{-1}(\mathcal{X}_T) = \{\mathbf{x}_t \mid \underline{\mathbf{x}}_t \leq \mathbf{x}_t \leq \bar{\mathbf{x}}_t\} \quad (15)$$

$$\supseteq \{\mathbf{x}_t \mid \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \mathbf{c} \in \mathcal{X}_T, \\ \pi^L(\mathbf{x}_t) \leq \mathbf{u}_t \leq \pi^U(\mathbf{x}_t)\}, \quad (16)$$

$$\supseteq \mathcal{P}_{-1}(\mathcal{X}_T). \quad (17)$$

$\bar{\mathcal{P}}_{-1}(\mathcal{X}_T)$  is the hyper-rectangular outer bound of (16), thus guaranteeing the relationship between (15) and (16). The relation between (16) and (17) holds because  $\pi^L(\mathbf{x}_t) \leq \pi(\mathbf{x}_t) \leq \pi^U(\mathbf{x}_t)$  (via Theorem 3.1). It follows that  $\bar{\mathcal{P}}_{-1}(\mathcal{X}_T) \supseteq \mathcal{P}_{-1}(\mathcal{X}_T)$ . ■

Note that we relaxed the NN over the entire range of  $\bar{\mathcal{R}}_{-1}$ , but if we instead divide it into smaller regions and relax each of them individually, we may find tighter bounds on the control input. While partitioning  $\bar{\mathcal{R}}_{-1}$  in this way is not required, it can be used to reduce the conservativeness of the BP estimate at the cost of increased computation time associated with solving LPs for each partitioned region. Ref. [19] provides a detailed discussion on different partitioning strategies.

##### B. Algorithm for Computing Backprojection Sets

Algorithm 1 follows the procedure outlined in §IV-A and Lemma 4.1 to obtain  $\bar{\mathcal{P}}_{-1}(\mathcal{X}_T)$ , i.e., an over-approximation

---

**Algorithm 2** BReach-LP

---

**Input:** target state set  $\mathcal{X}_T$ , trained NN control policy  $\pi$ , time horizon  $\tau$ , partition parameter  $\mathbf{r}$

**Output:** BP set approximations  $\bar{\mathcal{P}}_{-\tau:0}(\mathcal{X}_T)$ , affine control bound parameters  $\Omega_{-\tau:-1}$

- 1:  $\bar{\mathcal{P}}_0(\mathcal{X}_T) \leftarrow \mathcal{X}_T$
- 2: **for**  $t$  in  $\{-1, -2, \dots, -\tau\}$  **do**
- 3:    $\bar{\mathcal{P}}_{t+1}(\mathcal{X}_T) \leftarrow \text{boundWithRectangle}(\bar{\mathcal{P}}_{t+1}(\mathcal{X}_T))$
- 4:    $\bar{\mathcal{P}}_t(\mathcal{X}_T) \leftarrow \text{oneStepBackproj}(\bar{\mathcal{P}}_{t+1}(\mathcal{X}_T), \pi, \mathbf{r})$
- 5:    $[\mathbf{x}'_t, \bar{\mathbf{x}}'_t] \leftarrow \bar{\mathcal{P}}_t(\mathcal{X}_T)$
- 6:    $\Omega_t = [\Psi, \Phi, \bar{\alpha}, \beta] \leftarrow \text{CROWN}(\pi, [\mathbf{x}'_t, \bar{\mathbf{x}}'_t])$
- 7: **end for**
- 8: **return**  $\bar{\mathcal{P}}_{-\tau:0}(\mathcal{X}_T)$ ,  $\Omega_{-\tau:-1}$

---

of the BP set for a single timestep. The functions `lpMax` and `lpMin` solve the LPs formulated by (14) and `backreach` solves the LPs formulated by (12). The partition parameter  $\mathbf{r} \in \mathbb{R}^{n_x}$  gives the option to uniformly split the backreachable set, e.g.,  $\mathbf{r} = [2, 3]$  will split  $\bar{\mathcal{R}}_{-1}(\mathcal{X}_T)$  into a  $2 \times 3$  grid of cells.

To provide safety guarantees over an extended time horizon  $\tau$ , we extend this idea to iteratively compute BPs at multiple timesteps  $\bar{\mathcal{P}}_{-\tau:0}(\mathcal{X}_T)$ . We first initialize the zeroth BP set as the target set (Line 1). Then we step backward in time (Line 2), recursively using Algorithm 1 (`oneStepBackproj`), and the BP from the previous step to iteratively compute the new BP set (Line 4). This is done  $\tau$  times to give a list of BP set estimates  $\bar{\mathcal{P}}_{-\tau:0}(\mathcal{X}_T)$ . The proposed procedure is summarized in Algorithm 2. Note that from this, we can see that the number of LPs solved  $N_{LP}$  can be written as  $N_{LP} = 2n_x N_r \tau$ , where  $N_r$  is the number of partitions associated with  $\mathbf{r}$  at each step. Thus, the computational complexity is linear with respect to state dimension. However,  $N_r$  can grow quickly with state dimension, therefore necessitating future investigation into efficient partitioning strategies to reduce computation time.

### C. Algorithm for Computing N-Step Backprojection Sets

Notice that by iteratively making over-approximations using the previously calculated BP over-approximation, BReach-LP tends to accrue conservativeness over the time horizon due to the wrapping effect [34]. Thus even if  $\bar{\mathcal{P}}_{-\tau}(\mathcal{X}_T)$  tightly bounds the set of states that reach  $\bar{\mathcal{P}}_{-\tau+1}(\mathcal{X}_T)$ , it may be an overly conservative estimate of the set of states that ultimately end up in  $\mathcal{X}_T$  in  $\tau$  timesteps. To reduce the accrued conservativeness, we present ReBReach-LP (Algorithm 3) that uses BReach-LP to initialize  $\bar{\mathcal{P}}_{-\tau:0}(\mathcal{X}_T)$  and the collected affine control bound parameters  $\Omega_{-\tau:-1}$ . We then use a procedure similar to the one outlined in §IV-A, but now we relax the NN over  $\mathcal{P}_t$  instead of  $\bar{\mathcal{R}}_t$  and include additional constraints that require the future states of the system to progress through the future BP estimates and eventually reach the target set while satisfying the relaxed affine control bounds at each step along the way. The number of LPs calculated in Algorithm 3

is  $N_{LP} = 2n_x N_r (2\tau - 1)$ , which is again linear in state dimension, but with the same issue given by  $N_r$ .

**Lemma 4.2:** Given an  $m$ -layer NN control policy  $\pi : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_u}$ , closed-loop dynamics  $f : \mathbb{R}^{n_x} \times \Pi \rightarrow \mathbb{R}^{n_x}$  as in Eqs. (1) and (2), a set of BP estimates  $\bar{\mathcal{P}}_{-\tau:0}(\mathcal{X}_T)$ , a corresponding set of affine control bounds  $\Omega_{-\tau:-1}$ , and target set  $\mathcal{X}_T$ , the following relations hold:

$$\mathcal{P}_t(\mathcal{X}_T) \subseteq \bar{\mathcal{P}}'_t(\mathcal{X}_T) \subseteq \bar{\mathcal{P}}_t(\mathcal{X}_T), \forall t \in \underbrace{\{-\tau, -\tau + 1, \dots, -1\}}_{\triangleq \mathcal{T}},$$

where  $\bar{\mathcal{P}}'_t(\mathcal{X}_T) \triangleq \{\mathbf{x}_t \mid \mathbf{x}'_t \leq \mathbf{x}_t \leq \bar{\mathbf{x}}'_t\}$  with  $\mathbf{x}'_t$  and  $\bar{\mathbf{x}}'_t$  calculated using the LPs specified by (18).

*Proof:* Given dynamics from Eqs. (1) and (2), solve the following optimization problems for each state  $k \in [n_x]$  and for each  $t \in \mathcal{T}$ ,

$$\bar{\mathbf{x}}_{t,k} = \min_{\mathbf{x}_t, \mathbf{u}_t \in \mathcal{F}_{\bar{\mathcal{P}}'_t}} \mathbf{e}_k^\top \mathbf{x}_t, \quad \mathbf{x}_{t,k} = \max_{\mathbf{x}_t, \mathbf{u}_t \in \mathcal{F}_{\bar{\mathcal{P}}'_t}} \mathbf{e}_k^\top \mathbf{x}_t, \quad (18)$$

where

$$\mathcal{F}_{\bar{\mathcal{P}}'_t} \triangleq \left\{ \mathbf{x}_t, \mathbf{u}_t \middle| \begin{array}{l} \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \mathbf{c} = \mathbf{x}_{t+1}, \\ \mathbf{x}_{t+1} \in \bar{\mathcal{P}}_{t+1}, \\ \pi_t^L(\mathbf{x}_t) \leq \mathbf{u}_t \leq \pi_t^U(\mathbf{x}_t), \\ \mathbf{x}_t \in \bar{\mathcal{P}}_t(\mathcal{X}_T), \end{array} \right\}, \quad (19)$$

with  $\bar{\mathcal{P}}_0(\mathcal{X}_T) = \mathcal{X}_T$ , and  $\pi_t^L$  and  $\pi_t^U$  obtained from  $\Omega_{-\tau:-1}$ .

The final constraint in (19) guarantees  $\bar{\mathcal{P}}'_t(\mathcal{X}_T) \subseteq \bar{\mathcal{P}}_t(\mathcal{X}_T), \forall t \in \mathcal{T}$ . The third constraint ensures that the relations Eqs. (15) to (17) in the proof of Lemma 4.1 holds for all  $t \in \mathcal{T}$ , thus guaranteeing  $\mathcal{P}_t(\mathcal{X}_T) \subseteq \bar{\mathcal{P}}'_t(\mathcal{X}_T)$ . It follows that  $\mathcal{P}_t(\mathcal{X}_T) \subseteq \bar{\mathcal{P}}'_t(\mathcal{X}_T) \subseteq \bar{\mathcal{P}}_t(\mathcal{X}_T), \forall t \in \mathcal{T}$ . ■

Notice that the first two constraints provide the key advantage of the LPs solved by (18) over those given by Lemma 4.1 in that they require the state to trace back through the set of BPs leading to the original target set, thus providing a better approximation of the true BP set.

## V. NUMERICAL RESULTS

In this section we use numerical experiments to verify our algorithms and demonstrate their properties as they relate to each other and to the forward reachability tool proposed in [19]. First we show how ReBReach-LP can be used to reduce the conservativeness gathered by BReach-LP and we quantify the additional computation cost. We then show how BReach-LP can be used in a collision-avoidance scenario that causes Reach-LP [19] to fail.

All numerical results were collected with the LP solver `cvxpy` [35] on a machine running Ubuntu 20.04 with an i7-6700K CPU and 32 GB of RAM.

### A. Double Integrator

Consider the discrete-time double integrator model [17]

$$\mathbf{x}_{t+1} = \underbrace{\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}}_A \mathbf{x}_t + \underbrace{\begin{bmatrix} 0.5 \\ 1 \end{bmatrix}}_B \mathbf{u}_t \quad (20)$$

with  $\mathbf{c} = 0$ ,  $\mathbf{C} = \mathbf{I}_2$ , and discrete sampling time  $t_s = 1s$ . The NN controller (identical to the double integrator

---

**Algorithm 3** ReBReach-LP

---

**Input:** target state set  $\mathcal{X}_T$ , trained NN control policy  $\pi$ , time horizon  $\tau$ , partition parameter  $\mathbf{r}$

**Output:** Refined BP set approximations  $\bar{\mathcal{P}}'_{-\tau:0}(\mathcal{X}_T)$

- 1:  $\bar{\mathcal{P}}'_{-\tau:0}(\mathcal{X}_T) \leftarrow \emptyset$
- 2:  $\bar{\mathcal{P}}_{-\tau:0}(\mathcal{X}_T), \Omega_{-\tau:-1} \leftarrow \text{BReach-LP}(\mathcal{X}_T, \pi, \tau, \mathbf{r})$
- 3:  $\bar{\mathcal{P}}'_{-1:0}(\mathcal{X}_T) \leftarrow \bar{\mathcal{P}}_{-1:0}(\mathcal{X}_T)$
- 4: **for**  $t$  in  $\{-2, \dots, -\tau\}$  **do**
- 5:    $[\underline{\mathbf{x}}_t, \bar{\mathbf{x}}_t] \leftarrow \text{boundWithRectangle}(\bar{\mathcal{P}}_t(\mathcal{X}_T))$
- 6:    $\mathcal{S} \leftarrow \text{partition}([\underline{\mathbf{x}}_t, \bar{\mathbf{x}}_t], \mathbf{r})$
- 7:   **for**  $[\underline{\mathbf{x}}_t, \bar{\mathbf{x}}_t]$  in  $\mathcal{S}$  **do**
- 8:      $\Psi, \Phi, \alpha, \beta \leftarrow \text{CROWN}(\pi, [\underline{\mathbf{x}}_t, \bar{\mathbf{x}}_t])$
- 9:     **for**  $k \in n_x$  **do**
- 10:        $\bar{\mathbf{x}}_{t;k} \leftarrow \text{NStepLpMax}(\bar{\mathcal{P}}_{-\tau:0}, \Omega_{-\tau:-1}, \Psi, \Phi, \alpha, \beta)$
- 11:        $\underline{\mathbf{x}}_{t;k} \leftarrow \text{NStepLpMin}(\bar{\mathcal{P}}_{-\tau:0}, \Omega_{-\tau:-1}, \Psi, \Phi, \alpha, \beta)$
- 12:     **end for**
- 13:      $\mathcal{A} \leftarrow \{\mathbf{x} \mid \forall k \in n_x, \underline{\mathbf{x}}_{t;k} \leq \mathbf{x} \leq \bar{\mathbf{x}}_{t;k}\}$
- 14:      $\bar{\mathcal{P}}'_{-1}(\mathcal{X}_T) \leftarrow \bar{\mathcal{P}}_{-1}(\mathcal{X}_T) \cup \mathcal{A}$
- 15:   **end for**
- 16:    $\bar{\mathcal{P}}'_t(\mathcal{X}_T) \leftarrow \bar{\mathcal{P}}'_{-1}(\mathcal{X}_T)$
- 17: **end for**
- 18: **return**  $\bar{\mathcal{P}}'_{-\tau:0}(\mathcal{X}_T)$

---

TABLE I: Compare error (21) for BReach-LP and ReBReach-LP (reduces conservativeness in final BP set estimate by 88% with  $2.5\times$  computation of BReach-LP).

Algorithm	Runtime [s]	Final Step Error
BReach-LP	$1.349 \pm 0.046$	21.96
ReBReach-LP	$3.383 \pm 0.095$	2.74

controller used in [19]) has  $[5, 5]$  neurons, ReLU activations and was trained with state-action pairs generated by an MPC controller. Fig. 3 compares BReach-LP (orange) and ReBReach-LP (blue). As shown, both algorithms collect some approximation error

$$\text{error} = \frac{A_{\text{true}} - A_{\text{BPE}}}{A_{\text{true}}}, \quad (21)$$

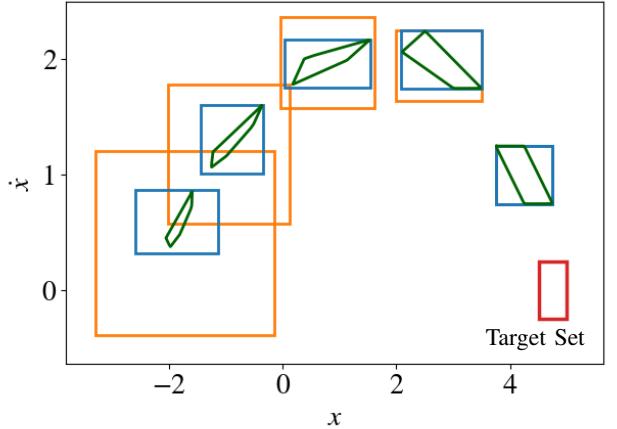
where  $A_{\text{true}}$  denotes the area of the tightest rectangular bound of the true BP set (dark green in Fig. 3a), calculated using Monte Carlo simulations, and  $A_{\text{BPE}}$  denotes the area of the BP estimate. However, because of the additional constraints and partitioning steps included in ReBReach-LP, it is able to reduce conservativeness in the final BP set estimate by 88%, as shown in Table I.

### B. Linearized Ground Robot

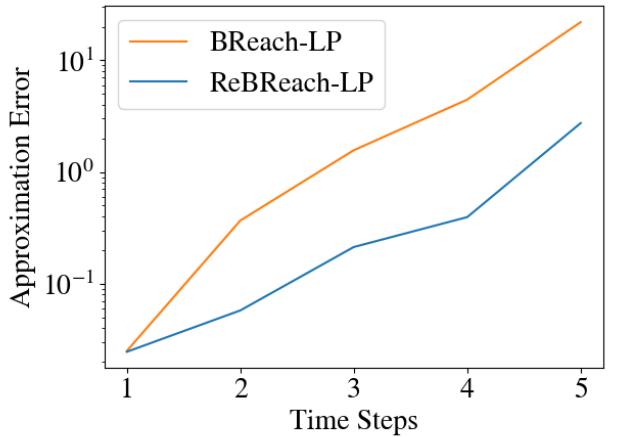
Using the feedback linearization technique proposed in [36], we represent the common unicycle model as a pair of integrators

$$\mathbf{x}_{t+1} = \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{\mathbf{A}} \mathbf{x}_t + \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{\mathbf{B}} \mathbf{u}_t \quad (22)$$

with  $\mathbf{c} = 0$ ,  $\mathbf{C} = \mathbf{I}_2$ , and sampling time  $t_s = 1\text{s}$ . With this system formulation, we can consider  $\mathbf{x}_t = [p_x, p_y]^\top$



(a) BP set estimates and true BP convex hulls (dark green)



(b) Approximation error (21) calculated at each time step

Fig. 3: Compare BP set estimates for a double integrator extending from the target set (red) calculated with BReach-LP (orange) and ReBReach-LP (blue).

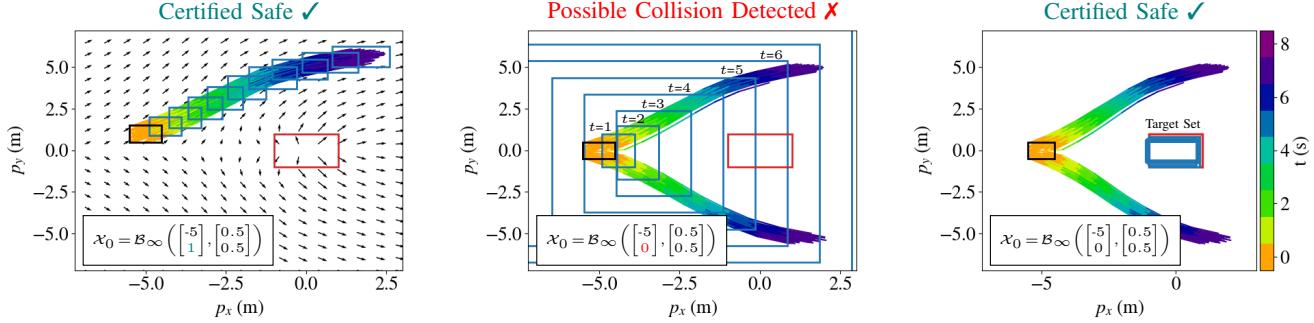
to represent the position of a vehicle in the  $x$ - $y$  plane and  $\mathbf{u}_t = [v_x, v_y]^\top$ .

To emulate the scenarios demonstrated by Fig. 1, we trained a NN with  $[10, 10]$  neurons and ReLU activations to mimic the vector field given by

$$\mathbf{u}(\mathbf{x}) = \begin{bmatrix} \max(\min(1 + \frac{2p_x}{p_x^2 + p_y^2}, 1), -1) \\ \max(\min(\frac{p_y}{p_x^2 + p_y^2} + 2\mathbf{s}(p_y)\frac{e^{-\frac{p_x}{2}+2}}{(1+e^{-\frac{p_x}{2}+2})^2}, 1), -1) \end{bmatrix} \quad (23)$$

where  $\mathbf{s}(\cdot)$  returns the sign of the argument. The vector field (23) is visualized in Fig. 4a and produces trajectories that drive the system away from an obstacle bounded by the target set  $\mathcal{X}_T = \mathcal{B}_\infty([0, 0]^\top, [1, 1]^\top)$  (shown in red). Eq. (23) was used to generate  $10^5$  data points sampled from the state space region  $\mathcal{B}_\infty([0, 0]^\top, [10, 10]^\top)$ , which were then used to train the NN for 20 epochs with a batch size of 32.

First, in Fig. 4a, we demonstrate a typical forward reachability example using the method described in [19]. The blue bounding boxes represent the forward reachable set estimates calculated using Reach-LP [19] with  $\mathbf{r} = [4, 4]$  and the



(a) Nominal forward reachability collision avoidance scenario with vector field representation of control input. No intersection of target set (red) and reachable sets (blue) implies that safety can correctly be certified.

(b) Forward reachability strategy for collision avoidance at decision boundary. Reachable sets explode in response to uncertainty in which set of trajectories will be taken, causing an incorrect assessment of unsafe.

(c) Backward reachability strategy for collision avoidance at decision boundary. Safety can correctly be certified because none of the BP set estimates (blue) intersect with the initial state set (black).

Fig. 4: Collision-avoidance situation that [19] incorrectly labels as dangerous whereas BReach-LP correctly certifies safety.

lines represent the time progression ( $\mathbf{x}_0 \rightarrow \mathbf{x}_\tau$ : orange  $\rightarrow$  purple) of a set of possible trajectories. In this scenario, the system's initial state set  $\mathcal{X}_0$  lies above the  $x$ -axis and the NN control policy uniformly commands the system to go above the obstacle. The resulting reachability analysis works as expected with reachable set estimates that tightly bound the future trajectories.

The scenario shown in Fig. 4b is identical to that of Fig. 4a, except the initial state is now centered on the  $x$ -axis. This scenario demonstrates a breakdown in standard reachability analysis tools due to the uncertainty in which trajectory will be taken by the system. While some other tools, e.g., [18], can be shown to reduce conservativeness in the upper and lower bounds of the reachable set estimates, the authors are not aware of any methods to remove the regions between the trajectories and correctly certify this situation as safe. Also note that while it may initially seem like a good strategy to simply partition the initial set so that each element goes in one of the two directions, this would only work if the initial set could be split perfectly along the decision boundary, which may be difficult. This is also the reason that we solve LPs to find  $\bar{\mathcal{P}}_t$  rather than propagate states from  $\bar{\mathcal{R}}_t$  to  $\mathcal{X}_T$  using forward reachability analysis.

Finally, in Fig. 4c, we demonstrate the same situation as in Fig. 4b, but now use backward reachability analysis as the strategy for safety certification. Rather than propagating forward from the initial state set, we propagate backward from the target set, which was selected to bound the obstacle. Because the control policy was designed to avoid the obstacle, the BP over-approximations, calculated using BReach-LP with  $\mathbf{r} = [4, 4]$ , do not intersect with the initial state set (black), thus implying that safety can be certified over the time horizon. While the individual BP sets are harder to distinguish than the forward sets shown in Figs. 4a and 4b, we use  $\tau = 9$  in each scenario, thereby checking safety over the same time horizon. Note that the reachable sets in Fig. 4b were calculated in 0.5s compared to 2.35s for the BPs in Fig. 4c, but the result from BReach-LP (Fig. 4c) provides more useful information.

Finally, in Fig. 5 we confirm that our algorithms (in this

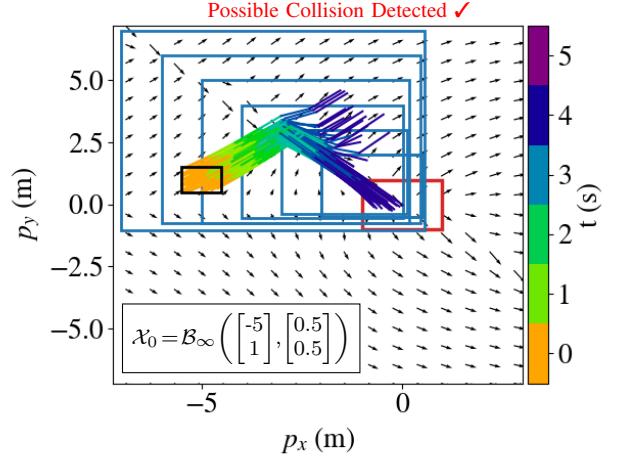


Fig. 5: As expected, ReBReach-LP is unable to certify safety for a faulty NN control policy.

case, ReBReach-LP) are able to detect a possible collision. Here we retrained the policy used in Fig. 4 but simulate a bug in the NN training process by commanding states along the line  $y = -x$  to direct the system towards the obstacle at the origin. The initial state set is the same as that in Fig. 4a, but now we see that the system reaches the target set in 6 seconds. Because the 5th and 6th BP set estimates intersect with  $\mathcal{X}_0$ , ReBReach-LP cannot certify that the system is safe, thus demonstrating the desired behavior for a safety certification algorithm given a faulty controller.

## VI. CONCLUSION

This paper presented two algorithms, BReach-LP and ReBReach-LP, for computing BP set estimates, i.e., sets for which a system will be driven to a designated target set, for linear NFLs over a given time horizon. Because backward reachability analysis is challenging for systems with NN components, this work employs available forward analysis tools in a way that provides over-approximations of BP sets. The key idea is to constrain the possible inputs of the system using typical analysis tools, then solve a set of LPs maximizing the size of the BP set subject to those constraints. This technique is used iteratively by BReach-LP

to find BP set estimates multiple timesteps from the target set. ReBReach-LP builds on BReach-LP to include additional computations that reduce the conservativeness in the over-approximation. Finally, we compared the performance of our two algorithms, demonstrating the trade-off between conservativeness and computation time, on a double integrator model, and compared our strategy to forward reachability in a collision-avoidance scenario with a linearized ground robot model.

Future work includes extending our methods to nonlinear NFLs, allowing us to handle more complex system dynamics. Additionally, making use of symbolic propagation techniques inspired by [18] may allow for additional reductions in conservativeness of the BP set estimates. Computation time may also be reduced by considering more efficient partitioning methods.

## REFERENCES

- [1] A. Kurakin, I. Goodfellow, S. Bengio *et al.*, “Adversarial examples in the physical world,” 2016.
- [2] X. Yuan, P. He, Q. Zhu, and X. Li, “Adversarial examples: Attacks and defenses for deep learning,” *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2805–2824, 2019.
- [3] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, “Deepdriving: Learning affordance for direct perception in autonomous driving,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2722–2730.
- [4] K. D. Julian, M. J. Kochenderfer, and M. P. Owen, “Deep neural network compression for aircraft collision avoidance systems,” *Journal of Guidance, Control, and Dynamics*, vol. 42, no. 3, pp. 598–608, 2019.
- [5] H. Zhang, T.-W. Weng, P.-Y. Chen, C.-J. Hsieh, and L. Daniel, “Efficient neural network robustness certification with general activation functions,” *Advances in neural information processing systems*, vol. 31, 2018.
- [6] L. Weng, H. Zhang, H. Chen, Z. Song, C.-J. Hsieh, L. Daniel, D. Boning, and I. Dhillon, “Towards fast computation of certified robustness for relu networks,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 5276–5285.
- [7] K. Xu, Z. Shi, H. Zhang, Y. Wang, K.-W. Chang, M. Huang, B. Kailkhura, X. Lin, and C.-J. Hsieh, “Automatic perturbation analysis for scalable certified robustness and beyond,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 1129–1141, 2020.
- [8] V. Tjeng, K. Xiao, and R. Tedrake, “Evaluating robustness of neural networks with mixed integer programming,” *arXiv preprint arXiv:1711.07356*, 2017.
- [9] G. Katz, D. A. Huang, D. Ibeling, K. Julian, C. Lazarus, R. Lim, P. Shah, S. Thakoor, H. Wu, A. Zeljić *et al.*, “The marabou framework for verification and analysis of deep neural networks,” in *International Conference on Computer Aided Verification*. Springer, 2019, pp. 443–452.
- [10] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, “Reluplex: An efficient smt solver for verifying deep neural networks,” in *International conference on computer aided verification*. Springer, 2017, pp. 97–117.
- [11] J. A. Vincent and M. Schwager, “Reachable polyhedral marching (rpm): A safety verification algorithm for robotic systems with deep neural network components,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 9029–9035.
- [12] K. Jia and M. Rinard, “Verifying low-dimensional input neural networks via input quantization,” in *International Static Analysis Symposium*. Springer, 2021, pp. 206–214.
- [13] 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*, 2019, pp. 157–168.
- [14] 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)*, vol. 18, no. 5s, pp. 1–22, 2019.
- [15] R. Ivanov, J. Weimer, R. Alur, G. J. Pappas, and I. Lee, “Verisig: verifying safety properties of hybrid systems with neural network controllers,” in *Proceedings of the 22nd ACM International Conference on Hybrid Systems: Computation and Control*, 2019, pp. 169–178.
- [16] J. Fan, C. Huang, X. Chen, W. Li, and Q. Zhu, “Reachnn\*: A tool for reachability analysis of neural-network controlled systems,” in *International Symposium on Automated Technology for Verification and Analysis*. Springer, 2020, pp. 537–542.
- [17] 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 *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 5929–5934.
- [18] C. Sidrane, A. Maleki, A. Irfan, and M. J. Kochenderfer, “Overt: An algorithm for safety verification of neural network control policies for nonlinear systems,” *arXiv preprint arXiv:2108.01220*, 2021.
- [19] M. Everett, G. Habibi, C. Sun, and J. P. How, “Reachability analysis of neural feedback loops,” *IEEE Access*, vol. 9, pp. 163 938–163 953, 2021.
- [20] S. Bak and H.-D. Tran, “Closed-loop acas xu nnce is unsafe: Quantized state backreachability for verification,” *arXiv preprint arXiv:2201.06626*, 2022.
- [21] S. Bansal, M. Chen, S. Herbert, and C. J. Tomlin, “Hamilton-jacobi reachability: A brief overview and recent advances,” in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 2017, pp. 2242–2253.
- [22] L. C. Evans, “Graduate studies in mathematics,” 1998.
- [23] I. M. Mitchell, “Comparing forward and backward reachability as tools for safety analysis,” in *International Workshop on Hybrid Systems: Computation and Control*. Springer, 2007, pp. 428–443.
- [24] A. Raghunathan, J. Steinhardt, and P. S. Liang, “Semidefinite relaxations for certifying robustness to adversarial examples,” *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [25] M. Althoff, “An introduction to cora 2015,” in *Proc. of the workshop on applied verification for continuous and hybrid systems*, 2015, pp. 120–151.
- [26] X. Chen, E. Ábrahám, and S. Sankaranarayanan, “Flow\*: An analyzer for non-linear hybrid systems,” in *International Conference on Computer Aided Verification*. Springer, 2013, pp. 258–263.
- [27] G. Frehse, C. L. Guernic, A. Donzé, S. Cotton, R. Ray, O. Lebeltel, R. Ripado, A. Girard, T. Dang, and O. Maler, “Spaceex: Scalable verification of hybrid systems,” in *International Conference on Computer Aided Verification*. Springer, 2011, pp. 379–395.
- [28] P. S. Duggirala, S. Mitra, M. Viswanathan, and M. Potok, “C2e2: A verification tool for stateflow models,” in *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. Springer, 2015, pp. 68–82.
- [29] B. Xue, Z. She, and A. Easwaran, “Under-approximating backward reachable sets by polytopes,” in *International Conference on Computer Aided Verification*. Springer, 2016, pp. 457–476.
- [30] N. Kochdumper and M. Althoff, “Computing non-convex inner-approximations of reachable sets for nonlinear continuous systems,” in *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 2130–2137.
- [31] L. Yang and N. Ozay, “Scalable zonotopic under-approximation of backward reachable sets for uncertain linear systems,” *IEEE Control Systems Letters*, vol. 6, pp. 1555–1560, 2021.
- [32] L. Ardizzone, J. Kruse, S. Wirkert, D. Rahner, E. W. Pellegrini, R. S. Klessen, L. Maier-Hein, C. Rother, and U. Köthe, “Analyzing inverse problems with invertible neural networks,” *arXiv preprint arXiv:1808.04730*, 2018.
- [33] J. Behrmann, W. Grathwohl, R. T. Chen, D. Duvenaud, and J.-H. Jacobsen, “Invertible residual networks,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 573–582.
- [34] C. Le Guernic, “Reachability analysis of hybrid systems with linear continuous dynamics,” Ph.D. dissertation, Université Joseph-Fourier-Grenoble I, 2009.
- [35] S. Diamond and S. Boyd, “Cvxpy: A python-embedded modeling language for convex optimization,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2909–2913, 2016.
- [36] J. B. Martinez, H. M. Becerra, and D. Gomez-Gutierrez, “Formation tracking control and obstacle avoidance of unicycle-type robots guaranteeing continuous velocities,” *Sensors*, vol. 21, no. 13, p. 4374, 2021.