

Robustness Certificates for Implicit Neural Networks: A Mixed Monotone Contractive Approach

Saber Jafarpour*

Georgia Institute of Technology

SABER@GATECH.EDU

Matthew Abate*

Georgia Institute of Technology

MATT.ABATE@GATECH.EDU

Alexander Davydov*

University of California, Santa Barbara

DAVYDOV@UCSB.EDU

Francesco Bullo

University of California, Santa Barbara

BULLO@UCSB.EDU

Samuel Coogan

Georgia Institute of Technology

SAM.COOGAN@GATECH.EDU

Editors: R. Firoozi, N. Mehr, E. Yel, R. Antonova, J. Bohg, M. Schwager, M. Kochenderfer

Abstract

Implicit neural networks are a general class of learning models that replace the layers in traditional feedforward models with implicit algebraic equations. Compared to traditional learning models, implicit networks offer competitive performance and reduced memory consumption. However, they can remain brittle with respect to input adversarial perturbations.

This paper proposes a theoretical and computational framework for robustness verification of implicit neural networks; our framework blends together mixed monotone systems theory and contraction theory. First, given an implicit neural network, we introduce a related embedded network and show that, given an ℓ_∞ -norm box constraint on the input, the embedded network provides an ℓ_∞ -norm box overapproximation for the output of the original network. Second, using ℓ_∞ -matrix measures, we propose sufficient conditions for well-posedness of both the original and embedded system and design an iterative algorithm to compute the ℓ_∞ -norm box robustness margins for reachability and classification problems. Third, of independent value, we show that employing a suitable relative classifier variable in our analysis will lead to tighter bounds on the certified adversarial robustness in classification problems. Finally, we perform numerical simulations on a Non-Euclidean Monotone Operator Network (NEMON) trained on the MNIST dataset. In these simulations, we compare the accuracy and run time of our mixed monotone contractive approach with the existing robustness verification approaches in the literature for estimating the certified adversarial robustness.

Keywords: Implicit Neural Networks, Robustness Analysis, Verification, Mixed Monotone Systems Theory, Contraction Theory

1. Introduction

Neural networks are increasingly being deployed in real-world applications, including natural language processing, computer vision, and self-driving vehicles. However, they are notoriously vul-

* These authors contributed equally

nerable to adversarial attacks; slight perturbations in the input can lead to large deviations in the output (Szegegy et al., 2014). Understanding this input sensitivity is essential in safety-critical applications, since the consequences of adversarial perturbations can be disastrous. Several different strategies have been proposed in the literature to design neural networks that are robust with respect to adversarial perturbations (Goodfellow et al., 2015; Papernot et al., 2016). Unfortunately, many of these approaches are based on robustness with respect to specific attacks and they do not provide formal robustness guarantees (Madry et al., 2018; Carlini and Wagner, 2017). Recently, there has been a large interest in providing provable robustness guarantees for neural networks. Most existing approaches focus on either the ℓ_2 -norm or ℓ_∞ -norm robustness measures. For neural networks with high-dimensional inputs and subject to dense perturbations, the ℓ_2 -norm robustness measures are known to provide overly conservative estimates of robustness and are less informative than their ℓ_∞ -norm counterparts. Rigorous verification methods generally fall into four different categories (i) Lipschitz bound methods (Fazlyab et al., 2019; Virmaux and Scaman, 2018; Combettes and Pesquet, 2020), (ii) interval bound methods (Mirman et al., 2018; Gowal et al., 2018; Zhang et al., 2020; Wang et al., 2018), (iii) optimization-based methods (Wong and Kolter, 2018; Zhang et al., 2018), and (iv) probabilistic methods (Cohen et al., 2019; Li et al., 2019). However, these methods suffer from several limitations. Regarding the Lipschitz bound approach, the proposed methods are either too conservative (Szegegy et al., 2014), not scalable to large-scale problems (Virmaux and Scaman, 2018; Combettes and Pesquet, 2020), or only applicable to a specific class of activation functions (Wang et al., 2018). Similar concerns apply to interval-bound propagation methods and optimization-based methods. Finally, probabilistic approaches provide some guarantees for ℓ_1 and ℓ_2 -norm robustness but there are theoretical limitations in their applicability for certifying ℓ_∞ -robustness (Blum et al., 2020).

In this paper we study the robustness properties of implicit neural networks, a recently proposed class of learning models with strong scalability properties. Implicit neural networks replace the notion of layer in the traditional neural networks with an implicit fixed-point equation (Bai et al., 2019; El Ghaoui et al., 2021). Compared to layer-based neural networks, implicit neural networks are known to (i) be significantly more memory efficient (Bai et al., 2019), (ii) generalize traditional architectures such as feedforward, convolutional, and residual networks (El Ghaoui et al., 2021), and (iii) enjoy improved training due to fewer vanishing and exploding gradients (Kag et al., 2020). Additionally, preliminary empirical evidence indicates that appropriately-trained implicit neural networks are more robust than traditional feedforward models (Pabbaraju et al., 2021); however this phenomenon is not yet well understood and open questions remain regarding the stability and robustness of implicit models.

We propose a rigorous computationally efficient certification method for implicit neural network robustness. We note that many of the robustness analysis tools for traditional neural networks are either not applicable to implicit neural networks or provide overly-conservative results. Our novel approach is derived from mixed monotone systems theory and contraction theory. Unlike the robustness verification approaches based on Lipschitz bounds, our framework takes into account the propagation of the ℓ_∞ -error bounds through the network and is scalable with the size of the network.

Related works

Implicit learning models. Implicit neural networks are a class of learning algorithms that replace the recursive function evaluations in traditional neural networks with an implicit algebraic equation.

In literature, implicit neural networks have been introduced under different names including deep equilibrium networks (Bai et al., 2019), implicit deep learning models (El Ghaoui et al., 2021), and equilibrated neural networks (Kag et al., 2020). One of the main challenges in studying implicit neural networks is their well-posedness, *i.e.*, existence and uniqueness of solutions for their fixed-point equation. The work (El Ghaoui et al., 2021) proposes a sufficient spectral condition for convergence of the Picard iterations associated with the fixed-point equation. In (Winston and Kolter, 2020; Revay et al., 2020), using monotone operator theory, a suitable parametrization of the weight matrix is proposed which guarantees convergence of suitable fixed-point iterations. Our previous work (Jafarpour et al., 2021b) proposes non-Euclidean contraction theory to design implicit neural networks and study their well-posedness, stability, and robustness with respect to the ℓ_∞ -norm; the general theory is developed in (Davydov et al., 2021) and a short tutorial is given in (Bullo et al., 2021).

Robustness of neural networks. Starting with (Szegedy et al., 2014), there has been a large body of work in machine learning to understand adversarial examples (Athalye et al., 2018). Rigorous verification of feedforward neural networks has been studied using abstract interpretation approaches (Katz et al., 2017), interval bound propagation methods (Wang et al., 2018; Zhang et al., 2018; Mirman et al., 2018; Gowal et al., 2018), and convex-relaxation approaches (Wong and Kolter, 2018; Fazlyab et al., 2020). Regarding implicit neural networks, there are far fewer works on their robustness guarantees. In (El Ghaoui et al., 2021) a sensitivity-based robustness analysis for implicit neural network is proposed. Approximation of the Lipschitz constants of deep equilibrium networks has been studied in (Pabbaraju et al., 2021; Revay et al., 2020). Recently, ellipsoidal methods (Chen et al., 2021) and interval-bound propagation methods (Wei and Kolter, 2022) have been proposed for robustness certification of deep equilibrium networks.

Mixed monotone system theory. Mixed monotone systems theory (Enciso et al., 2006; Angeli et al., 2014; Coogan and Arcak, 2015; Coogan, 2020) provides a generalization of classical monotone systems theory (Smith, 1995; Farina and Rinaldi, 2000; Angeli and Sontag, 2003), applicable to all dynamical systems bearing a locally Lipschitz continuous vector field (Yang and Ozay, 2019; Abate et al., 2021). A dynamical system is mixed monotone when there exists a related decomposition function that separates the system’s vector field or update map into increasing and decreasing components. Such a decomposition then facilitates robustness analysis for the initial mixed monotone system and specifically enables, *e.g.*, the efficient computation of robust reachable sets and invariant sets (Abate and Coogan, 2020).

Contributions

Based on mixed monotone system theory, this paper proposes a theoretical and computational framework to study the robustness of implicit neural networks. Given an implicit neural network, we introduce an associated embedded network with twice as many inputs and outputs as the original system. This embedded implicit network takes an ℓ_∞ -norm box as its input and generates an ℓ_∞ -norm box as its output. Then, we study the connection between the well-posedness of the embedded network and the robustness of the original implicit network. Our main theoretical contribution is as follows: if the ℓ_∞ -matrix measure of the original network’s weight matrix is less than one, then (i) the implicit neural network has a unique fixed-point, (ii) the embedded network has a unique fixed-point which can be computed using a suitable average-iteration, and (iii) for a given ℓ_∞ -norm box constraint on the input of the implicit neural network, the output of the embedded implicit neural

network is an ℓ_∞ -norm box overapproximation of output the original implicit network. In particular, result (iii) above shows how bounds on the network output are obtained directly from bounds on the network input, allowing for efficient reachability analysis for implicit neural networks. However, the output bounds obtained using this approach can lead to conservative robustness estimates in classifications. As a practical contribution, we use a suitable classifier variable together with mixed monotone systems theory to provide sharper robustness estimates in classification. In order to evaluate the robustness guarantees of implicit neural networks, we empirically examine their certified adversarial robustness. We then use (i) estimates of Lipschitz bounds, (ii) the interval bound propagation method, and (iii) our mixed monotone contractive approach to provide lower bounds on certified adversarial robustness. Finally, we compare the certified adversarial robustness of the three approaches mentioned above on a pre-trained implicit neural network. Our simulation results show that the mixed monotone contractive approach significantly outperforms the other two methods.

2. Mathematical preliminaries

Vectors and matrices. Given a matrix $B \in \mathbb{R}^{n \times m}$, we denote the non-negative part of B by $[B]^+ = \max(B, 0)$ and the nonpositive part of B by $[B]^- = \min(B, 0)$. The Metzler and non-Metzler part of square matrix $A \in \mathbb{R}^{n \times n}$ are denoted by $[A]^{\text{Mzl}}$ and $[A]^{\text{Mzl}}$, respectively, where

$$([A]^{\text{Mzl}})_{ij} = \begin{cases} A_{ij} & A_{ij} \geq 0 \text{ or } i = j \\ 0 & \text{otherwise,} \end{cases} \quad [A]^{\text{Mzl}} = A - [A]^{\text{Mzl}}.$$

For matrices $C \in \mathbb{R}^{n \times m}$ and $D \in \mathbb{R}^{p \times q}$, the Kronecker product of C and D is denoted by $C \otimes D$.

Matrix measures. For every $\eta \in \mathbb{R}_{>0}^n$, the diagonally weighted ℓ_∞ -norm is defined by $\|x\|_{\infty, [\eta]^{-1}} = \max_i |x_i|/\eta_i$, the diagonally weighted ℓ_∞ -matrix measure is defined by $\mu_{\infty, [\eta]^{-1}}(A) = \max_{i \in \{1, \dots, n\}} A_{ii} + \sum_{j \neq i} \frac{\eta_j}{\eta_i} |A_{ij}|$.

Lipschitz constants. Let $F : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ be a locally Lipschitz map in the first argument. For every $u \in \mathbb{R}^m$ and every $\alpha \in (0, 1]$, we define the α -average map $F_\alpha : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ by $F_\alpha = (1 - \alpha)I + \alpha F$, where I is the identity map on \mathbb{R}^n . Given a positive vector $\eta \in \mathbb{R}_{>0}^n$, $F(x, u)$ is Lipschitz in x with respect to the norm $\|\cdot\|_{\infty, [\eta]^{-1}}$ with constant $\text{Lip}_{\infty, [\eta]^{-1}}^x(F) \in \mathbb{R}_{\geq 0}$ if, for every $x_1, x_2 \in \mathbb{R}^n$ and every $u \in \mathbb{R}^m$,

$$\|F(x_1, u) - F(x_2, u)\|_{\infty, [\eta]^{-1}} \leq \text{Lip}_{\infty, [\eta]^{-1}}^x(F) \|x_1 - x_2\|_{\infty, [\eta]^{-1}},$$

Mixed monotone mappings. Given a map $F : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ and a Lipschitz function $d : \mathbb{R}^{2n} \times \mathbb{R}^{2m} \rightarrow \mathbb{R}^n$, we say F is *mixed monotone with respect to d* , if for every $i \in \{1, \dots, n\}$,

- (i) $d_i(x, x, u, u) = F_i(x, u)$;
- (ii) $d_i(x, \hat{x}, u, \hat{u}) \leq d_i(y, \hat{y}, u, \hat{u})$, for every $x \leq y$ such that $x_i = y_i$, and every $\hat{y} \leq \hat{x}$;
- (iii) $d_i(x, \hat{x}, u, \hat{u}) \leq d_i(x, \hat{x}, v, \hat{v})$, for every $u \leq v$ and every $\hat{v} \leq \hat{u}$.

Conditions (i)–(iii) are sometimes referred to as the Kamke conditions for mixed monotonicity¹ as developed in (Abate et al., 2021); see also (Coogan, 2020) for an equivalent infinitesimal characterization of mixed monotonicity. Every locally Lipschitz map F is mixed monotone with respect to

1. These are the conditions for ensuring that the continuous-time dynamical system with vector field defined by such a mapping (possibly added to a scaling of identity) is mixed monotone.

some decomposition function (Abate et al., 2021), however, finding a closed form decomposition function is in general challenging. A remarkable property of implicit neural networks, shown below, is that a decomposition function is easily available in closed-form.

3. Implicit neural networks

An implicit neural network is described by the following fixed-point equation:

$$\begin{aligned} x &= \Phi(Ax + Bu + b) := \mathbf{N}(x, u) \\ y &= Cx + c \end{aligned} \quad (1)$$

where $x \in \mathbb{R}^n$ is the hidden variable, $u \in \mathbb{R}^r$ is the input and $y \in \mathbb{R}^q$ is the output. The matrices $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times r}$, and $C \in \mathbb{R}^{q \times n}$ are weight matrices, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}^q$ are bias vectors, and $\Phi(x) = (\phi_1(x_1), \dots, \phi_n(x_n))^\top$ is the diagonal matrix of activation functions, where, for every $i \in \{1, \dots, n\}$, $\phi_i : \mathbb{R} \rightarrow \mathbb{R}$ satisfies $0 \leq \frac{\phi_i(x) - \phi_i(y)}{x - y} \leq 1$, for every $x, y \in \mathbb{R}$. Compared to feedforward neural networks, one of the main challenges in studying implicit neural networks is their well-posedness; a unique solution for the fixed-point equation (1) might not exist. We refer the readers to (Winston and Kolter, 2020; El Ghaoui et al., 2021; Revay et al., 2020; Jafarpour et al., 2021b) for discussions on the well-posedness of implicit networks.

Training implicit neural networks Given an input data $U = [u_1, \dots, u_m] \in \mathbb{R}^{r \times m}$ and its corresponding output data $Y = [y_1, \dots, y_m] \in \mathbb{R}^{q \times m}$, the training optimization problem learns weights and biases which minimizes $\mathcal{L}(Y, CX + c)$ subject to $X = \Phi(AX + BU)$, where $\mathcal{L} : \mathbb{R}^{q \times m} \times \mathbb{R}^{q \times m} \rightarrow \mathbb{R}$ is a suitable cost function. Thus, the training optimization problem is given by

$$\begin{aligned} \min_{A, B, C, b, c, X} \quad & \mathcal{L}(Y, CX + c) \\ & X = \Phi(AX + BU + b). \end{aligned} \quad (2)$$

In order to ensure that the implicit neural network is well-posed, an extra constraint is usually added to this training optimization problem. For instance, in (Winston and Kolter, 2020) the constraint $\mu_2(A) \leq \gamma$, in (El Ghaoui et al., 2021) the constraint $\|A\|_\infty \leq \gamma$, and in (Jafarpour et al., 2021b) the constraint $\mu_{\infty, [\eta]^{-1}}(A) \leq \gamma$ is proposed, for some $\gamma < 1$ and some $\eta \in \mathbb{R}_{>0}^n$.

4. Robustness certificates via a mixed monotone contracting approach

In this section, we use mixed monotone systems theory and contraction theory to study the input-to-output robustness of implicit neural networks.

Robustness of implicit neural networks. We first introduce the embedded implicit neural network associated with (1). Given $\underline{u} \leq \bar{u}$ in \mathbb{R}^r , we define *embedded implicit neural network* by

$$\begin{aligned} \begin{bmatrix} \underline{x} \\ \bar{x} \end{bmatrix} &= \begin{bmatrix} \Phi([A]^{Mz1} \underline{x} + [A]^{Mz1} \bar{x} + [B]^+ \underline{u} + [B]^- \bar{u} + b) \\ \Phi([A]^{Mz1} \bar{x} + [A]^{Mz1} \underline{x} + [B]^+ \bar{u} + [B]^- \underline{u} + b) \end{bmatrix} := \begin{bmatrix} \mathbf{N}^E(\underline{x}, \bar{x}, \underline{u}, \bar{u}) \\ \mathbf{N}^E(\bar{x}, \underline{x}, \bar{u}, \underline{u}) \end{bmatrix}, \\ \begin{bmatrix} \underline{y} \\ \bar{y} \end{bmatrix} &= \begin{bmatrix} [C]^+ & [C]^- \\ [C]^- & [C]^+ \end{bmatrix} \begin{bmatrix} \underline{x} \\ \bar{x} \end{bmatrix} + \begin{bmatrix} c \\ c \end{bmatrix}. \end{aligned} \quad (3)$$

The embedded implicit neural network (3) can be considered as a neural network with the box input $[\underline{u}, \bar{u}]$ and the box output $[\underline{y}, \bar{y}]$ (see Figure 1). Next, we study well-posedness of the embedded implicit neural network (3) and its connection with robustness of the implicit neural network (1).

Theorem 1 (Robustness of implicit neural networks) Consider the implicit neural network (1). The following statement holds:

(i) the map $N(x, u)$ is mixed monotone with respect to the decomposition function N^E ;

Moreover, let $\eta \in \mathbb{R}_{>0}^n$ be such that $\mu_{\infty, [\eta]^{-1}}(A) < 1$. For every $\underline{u} \leq \bar{u}$, every $u \in [\underline{u}, \bar{u}]$, and every $\alpha \in [0, \alpha^* := (1 - \min_{i \in \{1, \dots, n\}} (A_{ii})^-)^{-1}]$,

(ii) the α -average map $(\underline{x}, \bar{x}) \mapsto \begin{bmatrix} N_{\alpha}^E(\underline{x}, \bar{x}, \underline{u}, \bar{u}) \\ N_{\alpha}^E(\bar{x}, \underline{x}, \bar{u}, \underline{u}) \end{bmatrix}$ is a contraction mapping with respect to the norm $\|\cdot\|_{\infty, \mathcal{I}_2 \otimes [\eta]^{-1}}$ with minimum contraction factor $\text{Lip}\left(\begin{bmatrix} N_{\alpha^*}^E \\ N_{\alpha^*}^E \end{bmatrix}\right) = 1 - \frac{1 - \mu_{\infty, [\eta]^{-1}}(A)^+}{1 - \min_{i \in \{1, \dots, n\}} (A_{ii})^-}$;

(iii) the α -average map N_{α} is a contraction mapping with respect to the norm $\|\cdot\|_{\infty, [\eta]^{-1}}$ minimum contraction factor $\text{Lip}(N_{\alpha^*}) = 1 - \frac{1 - \mu_{\infty, [\eta]^{-1}}(A)^+}{1 - \min_{i \in \{1, \dots, n\}} (A_{ii})^-}$;

(iv) the embedded network (3) has a unique fixed point $\begin{bmatrix} \underline{x}^* \\ \bar{x}^* \end{bmatrix}$ such that $\underline{x}^* \leq \bar{x}^*$ and we have $\lim_{k \rightarrow \infty} \begin{bmatrix} \underline{x}^k \\ \bar{x}^k \end{bmatrix} = \begin{bmatrix} \underline{x}^* \\ \bar{x}^* \end{bmatrix}$, where the sequence $\left\{ \begin{bmatrix} \underline{x}^k \\ \bar{x}^k \end{bmatrix} \right\}_{k=1}^{\infty}$ is defined iteratively by

$$\begin{bmatrix} \underline{x}^{k+1} \\ \bar{x}^{k+1} \end{bmatrix} = \begin{bmatrix} N_{\alpha^*}^E(\underline{x}^k, \bar{x}^k, \underline{u}, \bar{u}) \\ N_{\alpha^*}^E(\bar{x}^k, \underline{x}^k, \bar{u}, \underline{u}) \end{bmatrix}, \quad \text{for every } k \in \mathbb{Z}_{\geq 0}, \quad \begin{bmatrix} \underline{x}^0 \\ \bar{x}^0 \end{bmatrix} \in \mathbb{R}^{2n}; \quad (4)$$

(v) the implicit neural network (1) has a unique fixed-point x_u^* such that $x_u^* \in [\underline{x}^*, \bar{x}^*]$ and we have $\lim_{k \rightarrow \infty} x_u^k = x_u^*$ where the sequence $\{x_u^k\}_{k=1}^{\infty}$ is defined iteratively by

$$x_u^{k+1} = N_{\alpha^*}(x_u^k, u), \quad \text{for every } k \in \mathbb{Z}_{\geq 0}, \quad x^0 \in \mathbb{R}^n. \quad (5)$$

Remark 2

- (i) Theorem 1 can be interpreted as a dynamical system approach to study robustness of implicit neural networks. Indeed, it is easy to see that the α -average iteration (4) (resp. (5)) are the forward Euler discretization of the dynamical system $\frac{d}{dt} \begin{bmatrix} \underline{x} \\ \bar{x} \end{bmatrix} = - \begin{bmatrix} \underline{x} \\ \bar{x} \end{bmatrix} + \begin{bmatrix} N^E(\underline{x}, \bar{x}, \underline{u}, \bar{u}) \\ N^E(\bar{x}, \underline{x}, \bar{u}, \underline{u}) \end{bmatrix}$ (resp. $\frac{dx}{dt} = -x + N(x, u)$). We refer to (Jafarpour et al., 2021a) for a proof of Theorem 1.
- (ii) Theorem 1(iv) and (v) show that $\mu_{\infty, [\eta]^{-1}}(A) < 1$ is a sufficient condition for existence and uniqueness of the fixed-point of both the original neural network and embedded neural network. In (Jafarpour et al., 2021b), to ensure well-posedness, the NEMON model is trained by adding the sufficient condition $\mu_{\infty, [\eta]^{-1}}(A) < 1$ to the training problem (2). Therefore, for the NEMON model, the embedded implicit network provides a margin of robustness with respect to any ℓ_{∞} -norm box uncertainty on the input.
- (iii) In terms of evaluation time, computing the ℓ_{∞} -box bounds on the output is equivalent to two forward passes of the original implicit network (see Figure 1).
- (iv) Implicit neural networks contain feedforward neural networks as a special case (El Ghaoui et al., 2021). Indeed, for a feedforward neural network with k layers and n neurons in each layer, there exists an implicit network representation with block upper diagonal weight matrix

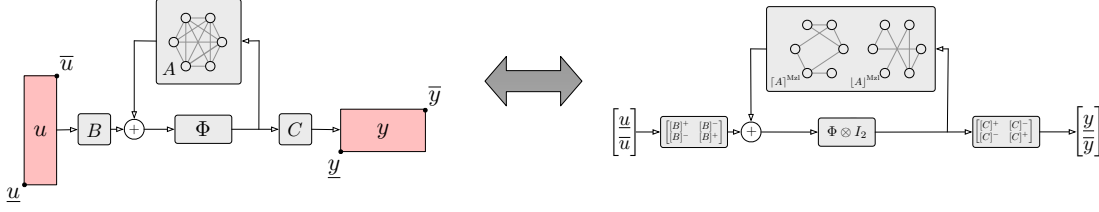


Figure 1: The original implicit neural network and its associated embedding network. The input-output behavior of the embedding system provides a box estimate for robustness of the original network.

$A \in \mathbb{R}^{kn \times kn}$. Since A is block upper diagonal, one can choose $\eta = [\delta, \delta^2, \dots, \delta^k]^\top \in \mathbb{R}_{>0}^k$ with sufficiently small $\delta > 0$ to obtain $\mu_{\infty, [\eta]^{-1} \otimes I_n}(A) < 1$. In this case, the fixed-point of the embedded implicit network (3) is unique, can be computed explicitly using back-substitution, and corresponds exactly to the approach taken in (Gowal et al., 2018).

Robustness verification via relative classifiers. The embedded network output $[y, \bar{y}]$ provides bounds on the elements of the initial implicit network’s output, thus allowing for efficient reachability analysis. However, for classification problems, where the goal is to identify the maximum element of y , these boxes can lead to overly-conservative estimates of robustness. In this section, we propose an alternative approach for certified robustness in classification problem by studying a suitable classifier variable using mixed monotone systems theory. Suppose the input $u \in \mathbb{R}^r$ leads to the output $y(u) \in \mathbb{R}^q$ and the correct label of u is $i \in \{1, \dots, q\}$. We are interested to study the robustness of our classifier with respect to a perturbed set of inputs $[\underline{u}, \bar{u}] \ni u$. For every $v \in [\underline{u}, \bar{u}]$, the *relative classifier variable* $z^u(v) \in \mathbb{R}^{q-1}$ is defined by

$$z^u(v) := y(v)_i \mathbb{1}_{q-1} - y(v)_{-i}, \quad (6)$$

where $y(v)_{-i} = (y(v)_1, \dots, y(v)_{i-1}, y(v)_{i+1}, \dots, y(v)_q) \in \mathbb{R}^{q-1}$ (see (Zhang et al., 2020) for a similar construction). Note that $z^u(v) \geq 0$ only when the perturbed input v retains the correct label i , i.e., the perturbation does not have any effect on the classification. Using (1), we write (6) as

$$z^u(v) = T^u y(v) = T^u C x^* + T^u c, \quad (7)$$

where x^* is the fixed-point of the implicit neural network (1) with input v and $T^u \in \{-1, 0, 1\}^{(q-1) \times q}$ is the linear transformation defined by (6). Now, we construct

$$\underline{z}^u = [T^u C]^+ \underline{x}^* + [T^u C]^- \bar{x}^* + T^u c, \quad (8)$$

where $\underline{x}^*, \bar{x}^*$ solves (3) with \underline{u}, \bar{u} being the above perturbation bounds on the input.

Lemma 3 (Properties of the relative classifier variable) *Let $u \in [\underline{u}, \bar{u}]$ be an input with the correct label $i \in \{1, \dots, q\}$ and $\begin{bmatrix} y \\ \bar{y} \end{bmatrix}$ be the output of the embedded network (3) with input $\begin{bmatrix} u \\ \bar{u} \end{bmatrix}$. Then,*

- (i) $\underline{z}^u \geq 0$ implies that the every perturbed input $v \in [\underline{u}, \bar{u}]$ is given the same label as u , that is, $y_i(v) > y_j(v)$ for all $j \neq i$ and every $v \in [\underline{u}, \bar{u}]$;
- (ii) $\underline{y}_i - \max_{j \neq i} \bar{y}_j \geq 0$ implies that $\underline{z}^u \geq 0$.

Note that the converse of Lemma 3(ii) need not hold in general. Indeed, Lemma 3 shows that using $\underline{z}^u \geq 0$ for classification leads to less conservative robustness certificates compared to using $\underline{y}_i - \max_{j \neq i} \bar{y}_j \geq 0$. We refer to (Jafarpour et al., 2021a) for the proof of Lemma 3.

5. Theoretical and numerical comparisons

In this section, we compare our robustness bounds with the existing bounds in the literature. Before we proceed with the comparison, following (Gowal et al., 2018; Pabbaraju et al., 2021), we introduce the notion of certified adversarial robustness which plays a crucial role in our numerical comparison for classification problems. To this end, we consider a set of labeled test data $\mathcal{U} \subset \mathbb{R}^r$ and we define the *deviation function* $\delta : \mathbb{R}_{\geq 0} \times \mathcal{U} \rightarrow \mathbb{R}$ by

$$\delta(\epsilon, u) = \max_{v \in \mathbb{R}^r} \{y(v)_i - \max_{j \neq i} y(v)_j \mid \|u - v\|_\infty \leq \epsilon, \quad i \text{ is the correct label of } u\}, \quad (9)$$

where $y(u)$ and $y(v)$ are the implicit neural network outputs generated by inputs u and v respectively. We say that the network is *certified adversarially robust* for radius ϵ at input u if $\delta(\epsilon, u) > 0$. Certifying adversarial robustness can be complicated due to the non-convexity of the optimization problem on v for the deviation function. We briefly review the existing methods for robustness verification of implicit neural networks and show how these methods can be used to provide lower bound on the certified adversarial robustness.

Method 1: Lipschitz constants. For implicit neural network, the estimates on the input-output Lipschitz constants are studied for deep equilibrium networks in (Winston and Kolter, 2020; Pabbaraju et al., 2021; Revay et al., 2020), for implicit deep learning models in (El Ghaoui et al., 2021), and for non-Euclidean monotone operator networks in (Jafarpour et al., 2021b). For an implicit neural network (1) with ℓ_∞ input-output Lipschitz constant $\text{Lip}_\infty^{u \rightarrow y} \in \mathbb{R}_{\geq 0}$, the output can be bounded as $\|y(u) - y(v)\|_\infty \leq \text{Lip}_\infty^{u \rightarrow y} \|u - v\|_\infty$. We define $\delta^{\text{Lip}}(\epsilon, u) := (y(u)_i - \max_{j \neq i} y(u)_j) - 2(\text{Lip}_\infty^{u \rightarrow y})\epsilon$. One can see that $\delta^{\text{Lip}}(\epsilon, u) > 0$ is a sufficient condition for certified adversarial robustness.

Method 2: Interval bound propagation. In (Gowal et al., 2018) a framework based on interval bound propagation has been proposed for training robust feedforward neural networks. This method has recently been extended for training deep equilibrium networks in (Wei and Kolter, 2022). Given an implicit neural network (1) with input perturbation $\|u - v\|_\infty \leq \epsilon$, we can adopt the approach in (Gowal et al., 2018) to the implicit framework and propose the following fixed-point equation for estimating the output of the network:

$$\begin{bmatrix} \underline{x} \\ \bar{x} \end{bmatrix} = \begin{bmatrix} \Phi([A]^+ \underline{x} + [A]^- \bar{x} + [B]^+ \underline{u} + [B]^- \bar{u} + b) \\ \Phi([A]^+ \bar{x} + [A]^- \underline{x} + [B]^+ \bar{u} + [B]^- \underline{u} + b) \end{bmatrix}, \quad (10)$$

$$\begin{bmatrix} \underline{y} \\ \bar{y} \end{bmatrix} = \begin{bmatrix} [C]^+ & [C]^- \\ [C]^- & [C]^+ \end{bmatrix} \begin{bmatrix} \underline{x} \\ \bar{x} \end{bmatrix} + \begin{bmatrix} c \\ c \end{bmatrix}, \quad (11)$$

where $\underline{u} = u - \epsilon \mathbb{1}_m$, $\bar{u} = u + \epsilon \mathbb{1}_m$, and (\underline{x}, \bar{x}) are the solutions of the fixed-point equation (10). It is worth mentioning that the condition $\mu_{\infty, [\eta]^{-1}}(A) < 1$ proposed in Theorem 1 does not, in general, ensure well-posedness of the fixed-point equation (10). The output of the neural network then can be bounded by the box $[\underline{y}, \bar{y}]$. We define $\delta^{\text{IBP}}(\epsilon, u) = \underline{y}_i - \max_{j \neq i} \bar{y}_j$. One can see that $\delta^{\text{IBP}}(\epsilon, u) > 0$ is a sufficient condition for certified adversarial robustness.

Method 3: Mixed monotone contractive approach. Given an implicit neural network (1) with input perturbation $\|u - v\|_\infty \leq \epsilon$, we first use Theorem 1 to obtain bounds on the output of the network. Indeed, by Theorem 1(ii), the α -average iteration (4) with $\underline{u} = u - \epsilon \mathbb{1}_m$, $\bar{u} = u + \epsilon \mathbb{1}_m$

converges to (\underline{x}, \bar{x}) and therefore, we have $y(v) \in [\underline{y}, \bar{y}]$. Moreover, we can define $\delta^{\text{MM}}(\epsilon, u) = \underline{y}_i - \max_{j \neq i} \bar{y}_j$. One can see that $\delta^{\text{MM}}(\epsilon, u) > 0$ is a sufficient condition for certified adversarial robustness. Alternatively, we can use Theorem 1 with the output transformation (8) to provide less conservative lower bounds on for certified adversarial robustness of the network. We define $\delta^{\text{MM-C}}(\epsilon, u) = \min_{i \in \{1, \dots, q-1\}} \underline{z}_i^u$, where \underline{z}^u is as defined in equation (8). Then, by Lemma 3, one can obtain the tighter sufficient condition $\delta^{\text{MM-C}}(\epsilon, u) > 0$ for certified adversarial robustness.

5.1. A simple example

In this section, we consider a simple 2-dimensional implicit neural network to compare different approaches for robustness verification. Consider an implicit neural network (1) with $A = \begin{bmatrix} -\frac{1}{4} & -\frac{1}{4} \\ \frac{3}{4} & -\frac{1}{4} \end{bmatrix}$,

$B = \begin{bmatrix} \frac{1}{2} & 1 \\ 1 & \frac{1}{2} \end{bmatrix}$, $C = I_2$, $b = c = \mathbb{0}_2$, and $\Phi(\cdot) = \text{ReLU}(\cdot)$. Suppose that the nominal input is $u = \begin{bmatrix} \frac{1}{3} \\ \frac{1}{2} \end{bmatrix}$ and due to uncertainty, the input is in the box $v \in [\underline{u}, \bar{u}]$, where $\underline{u} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and $\bar{u} = \begin{bmatrix} \frac{1}{3} \\ 2 \end{bmatrix}$.

We compare the robustness bounds obtained using the Lipschitz bound approach, the interval bound propagation method, and our mixed monotone contractive approach. Regarding the Lipschitz bound approach, we use the framework in (Jafarpour et al., 2021b, Corollary 5) to estimate the input-output Lipschitz constant of the networks and thus we get $\|y(u) - y(v)\|_\infty \leq \frac{\|B\|_\infty \|C\|_\infty}{1 - \mu_\infty(A)^+} \|u - v\|_\infty = 3\|u - v\|_\infty$. Regarding the interval bound propagation method, using the iterations in (10), we obtain $y(v) \in \left[\begin{pmatrix} 0.0342 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.7265 \\ 2.1026 \end{pmatrix} \right]$. Finally, regarding the mixed monotone contractive approach,

using the α -average iteration (4) in Theorem 1(iv), we get $y(v) \in \left[\begin{pmatrix} 0.3939 \\ 0.6364 \end{pmatrix}, \begin{pmatrix} 1.6061 \\ 2.0303 \end{pmatrix} \right]$. Figure 2 compares the robustness certificates obtained using these different approaches.

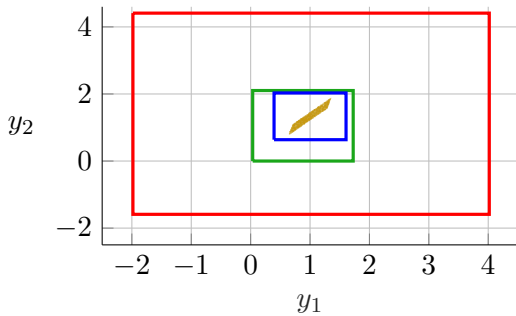


Figure 2: Problem Setting of Section 5.1: Comparing the application of Theorem 1 to existing verification methods for implicit neural networks. The yellow parallelogram shows different value of $y(v)$ for 1000 iid uniformly randomly selected $v = (v_1, v_2)^\top$ satisfying $0 \leq v_1 \leq \frac{1}{3}$ and $1 \leq v_2 \leq 2$. Robustness certificates attained from the Lipschitz bound approach, the interval bound propagation approach, and the application of Theorem 1 are shown as red, green, and blue boxes, respectively.

5.2. MNIST experiment

In this section, we compare the certified adversarial robustness of different approaches on the MNIST handwritten digit dataset². We trained a fully-connected NEMON model with $n = 100$ neurons as in the training problem (2). For well-posedness, we imposed $\mu_{\infty, [\eta]^{-1}}(A) \leq 0$, where we directly parametrize the set of such A as $A = [\eta]^{-1}T[\eta] - \text{diag}(|T|\mathbb{1}_n)$ for unconstrained T (Jafarpour et al., 2021b, Lemma 9). Training data was broken up into batches of 100 and the model

2. Code is available at <https://github.com/davydovalalexander/l4dc-inn-verification>

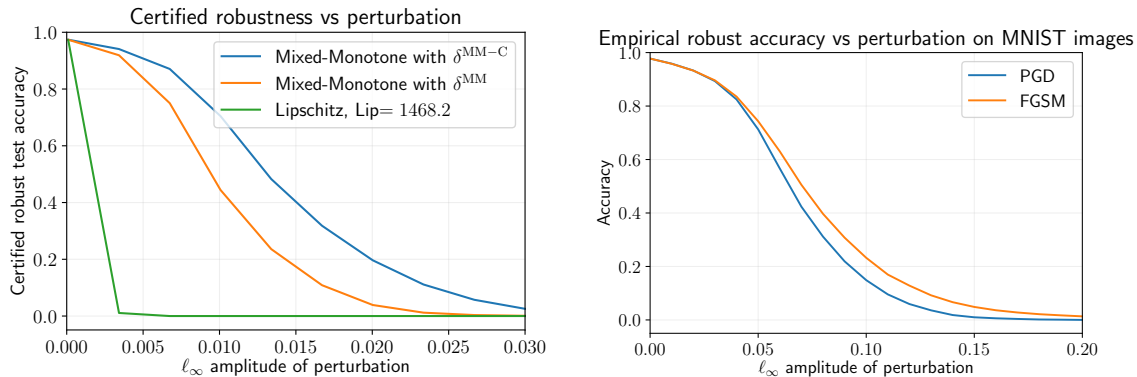


Figure 3: On the left is a plot of the certified adversarial robustness of the trained NEMON model using a Lipschitz method and two mixed monotonicity methods. For fixed ϵ , the fraction of test inputs which are certified robust are plotted. On the right is a plot of the empirical robustness of the same NEMON model subject to PGD and FGSM attacks. Note the difference in scale on the horizontal axis.

was trained for 15 epochs with a learning rate of 10^{-3} . After training, the model was validated on test data using the sufficient conditions for certified adversarial robustness in the previous section. For fixed ϵ and the 10000 test images, over 10 trials, it took, on average, 2.250 seconds to compute $\delta^{\text{Lip}}(\epsilon, u)$, 218.099 seconds to compute $\delta^{\text{IBP}}(\epsilon, u)$, 9.087 seconds to compute $\delta^{\text{MM}}(\epsilon, u)$, and 11.291 seconds to compute $\delta^{\text{MM-C}}(\epsilon, u)$. To provide a conservative upper-bound on the certified adversarial robustness and to observe empirical robustness, the model was additionally attacked using projected gradient descent (PGD) and fast-gradient sign method (FGSM) attacks. Results from these experiments are shown in Figure 3.

Summary evaluation. We draw several conclusions from the experiments. First, the bounds on the certified adversarial robustness provided from the interval-bound propagation are not plotted since they provided a trivial lower bound of zero adversarial robustness for every ϵ tested. Second, we see that the bounds on the certified adversarial robustness provided by the mixed monotonicity approaches are tighter than the bounds provided by the Lipschitz constant. Third, we note the additional tightness in the bounds provided by computing the relative classifier variable z^u . Finally, we observe that although mixed monotonicity approaches provide better bounds than the better-known Lipschitz and interval-bound propagation approaches, the gap between the certified robustness and the empirical robustness remains sizable, especially for larger ℓ_∞ -perturbations.

6. Conclusions

Using mixed monotone systems theory and contraction theory, we developed a framework for studying robustness of implicit neural networks. A key tool in this approach is an embedded network that provides ℓ_∞ -norm box estimates for input-output behavior of the given implicit neural network. Empirical evidence shows that our approach outperforms existing methods. Future work will include (i) applying the mixed monotone contractive approach to train robust implicit neural networks, (ii) designing appropriate state transformations (Abate and Coogan, 2021) to improve the input-output bounds in Theorem 1, and (iii) comparing our approach with the existing formal verification techniques such as Reluplex (Katz et al., 2017) and Neurify (Wang et al., 2018).

Acknowledgments

This work was supported in part by Air Force Office of Scientific Research under grants FA9550-22-1-0059 and FA9550-19-1-0015, the National Science Foundation under grant # 1836932 and the National Science Foundation Graduate Research Fellowship under grant #2139319.

References

- M. Abate and S. Coogan. Computing robustly forward invariant sets for mixed-monotone systems. In *IEEE Conf. on Decision and Control*, pages 4553–4559, 2020. doi:[10.1109/CDC42340.2020.9304461](https://doi.org/10.1109/CDC42340.2020.9304461).
- M. Abate and S. Coogan. Improving the fidelity of mixed-monotone reachable set approximations via state transformations. In *American Control Conference*, pages 4674–4679, 2021. doi:[10.23919/ACC50511.2021.9483264](https://doi.org/10.23919/ACC50511.2021.9483264).
- M. Abate, M. Dutreix, and S. Coogan. Tight decomposition functions for continuous-time mixed-monotone systems with disturbances. *IEEE Control Systems Letters*, 5(1):139–144, 2021. doi:[10.1109/LCSYS.2020.3001085](https://doi.org/10.1109/LCSYS.2020.3001085).
- D. Angeli and E. D. Sontag. Monotone control systems. *IEEE Transactions on Automatic Control*, 48(10):1684–1698, 2003. doi:[10.1109/TAC.2003.817920](https://doi.org/10.1109/TAC.2003.817920).
- D. Angeli, G. A. Enciso, and E. D. Sontag. A small-gain result for orthant-monotone systems under mixed feedback. *Systems & Control Letters*, 68:9–19, 2014. doi:[10.1016/j.sysconle.2014.03.002](https://doi.org/10.1016/j.sysconle.2014.03.002).
- A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. In *International Conference on Machine Learning*, pages 284–293, 2018. URL <https://openreview.net/forum?id=BJDH5M-AW>.
- S. Bai, J. Z. Kolter, and V. Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, 2019. URL <https://arxiv.org/abs/1909.01377>.
- A. Blum, T. Dick, N. Manoj, and H. Zhang. Random smoothing might be unable to certify ℓ_∞ robustness for high-dimensional images. *Journal of Machine Learning Research*, 21(211):1–21, 2020. URL <http://jmlr.org/papers/v21/20-209.html>.
- F. Bullo, P. Cisneros-Velarde, A. Davydov, and S. Jafarpour. From contraction theory to fixed point algorithms on Riemannian and non-Euclidean spaces. In *IEEE Conf. on Decision and Control*, December 2021. doi:[10.1109/CDC45484.2021.9682883](https://doi.org/10.1109/CDC45484.2021.9682883).
- N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017. doi:[10.1145/3128572.3140444](https://doi.org/10.1145/3128572.3140444).
- T. Chen, J. B. Lasserre, V. Magron, and E. Pauwels. Semialgebraic representation of monotone deep equilibrium models and applications to certification. In *Advances in Neural Information Processing Systems*, 2021. URL <https://openreview.net/forum?id=m4rb1Rlfdi>.

- J. Cohen, E. Rosenfeld, and J. Z. Kolter. Certified adversarial robustness via randomized smoothing. In *Int. Conf. on Machine Learning*, pages 1310–1320, 2019. URL <https://arxiv.org/abs/1902.02918>.
- P. L. Combettes and J-C. Pesquet. Lipschitz certificates for layered network structures driven by averaged activation operators. *SIAM Journal on Mathematics of Data Science*, 2(2):529–557, 2020. doi:10.1137/19M1272780.
- S. Coogan. Mixed monotonicity for reachability and safety in dynamical systems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 5074–5085, 2020. doi:10.1109/CDC42340.2020.9304391.
- S. Coogan and M. Arcak. Efficient finite abstraction of mixed monotone systems. In *Hybrid Systems: Computation and Control*, pages 58–67, April 2015. doi:10.1145/2728606.2728607.
- A. Davydov, S. Jafarpour, and F. Bullo. Non-Euclidean contraction theory for robust nonlinear stability. *IEEE Transactions on Automatic Control*, July 2021. URL <https://arxiv.org/abs/2103.12263>. Conditionally accepted as Paper.
- L. El Ghaoui, F. Gu, B. Travacca, A. Askari, and A. Tsai. Implicit deep learning. *SIAM Journal on Mathematics of Data Science*, 3(3):930–958, 2021. doi:10.1137/20M1358517.
- G. A. Enciso, H. L. Smith, and E. D. Sontag. Nonmonotone systems decomposable into monotone systems with negative feedback. *Journal of Differential Equations*, 224(1):205–227, 2006. doi:10.1016/j.jde.2005.05.007.
- L. Farina and S. Rinaldi. *Positive Linear Systems: Theory and Applications*. John Wiley & Sons, 2000. ISBN 0471384569.
- M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. J. Pappas. Efficient and accurate estimation of Lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, 2019. URL <https://arxiv.org/abs/1906.04893>.
- 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*, 2020. doi:10.1109/TAC.2020.3046193.
- I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. URL <https://arxiv.org/abs/1412.6572>.
- S. Gowal, K. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, R. Arandjelovic, T. Mann, and P. Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018. URL <https://arxiv.org/abs/1810.12715>.
- S. Jafarpour, M. Abate, A. Davydov, F. Bullo, and S. Coogan. Robustness certificates for implicit neural networks: A mixed monotone contractive approach. 2021a. URL <https://arxiv.org/abs/2112.05310>. Technical Note.

- S. Jafarpour, A. Davydov, A. V. Proskurnikov, and F. Bullo. Robust implicit networks via non-Euclidean contractions. In *Advances in Neural Information Processing Systems*, 2021b. URL <http://arxiv.org/abs/2106.03194>.
- A. Kag, Z. Zhang, and V. Saligrama. RNNs incrementally evolving on an equilibrium manifold: A panacea for vanishing and exploding gradients? In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=HylpqA4FwS>.
- G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *Computer Aided Verification*, pages 97–117, Cham, 2017. Springer International Publishing. URL <https://arxiv.org/abs/1702.01135>.
- B. Li, C. Chen, W. Wang, and L. Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems*, 2019. URL <https://arxiv.org/abs/1809.03113>.
- A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Machine Learning*, 2018. URL <https://arxiv.org/abs/1706.06083>.
- M. Mirman, T. Gehr, and M. Vechev. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, volume 80, pages 3578–3586, 2018. URL <https://proceedings.mlr.press/v80/mirman18b.html>.
- C. Pabbaraju, E. Winston, and J. Z. Kolter. Estimating Lipschitz constants of monotone deep equilibrium models. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=VcB4QkSfyO>.
- N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy*, pages 582–597, 2016. doi:[10.1109/SP.2016.41](https://doi.org/10.1109/SP.2016.41).
- M. Revay, R. Wang, and I. R. Manchester. Lipschitz bounded equilibrium networks. 2020. URL <https://arxiv.org/abs/2010.01732>.
- H. L. Smith. *Monotone Dynamical Systems: An Introduction to the Theory of Competitive and Cooperative Systems*. American Mathematical Society, 1995. ISBN 082180393X.
- C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL <https://arxiv.org/abs/1312.6199>.
- A. Virmaux and K. Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems*, volume 31, page 3839–3848, 2018. URL <https://arxiv.org/abs/1805.10965>.
- S. Wang, K. Pei, J. Whitehouse, J. Yang, and S. Jana. Efficient formal safety analysis of neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS’18, page 6369–6379, 2018. URL <https://arxiv.org/abs/1809.08098>.

- C. Wei and J. Z. Kolter. Certified robustness for deep equilibrium models via interval bound propagation. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=y1PXylgrXZ>.
- E. Winston and J. Z. Kolter. Monotone operator equilibrium networks. In *Advances in Neural Information Processing Systems*, 2020. URL <https://arxiv.org/abs/2006.08591>.
- E. Wong and J. Z. Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295, 2018. URL <http://proceedings.mlr.press/v80/wong18a.html>.
- L. Yang and N. Ozay. Tight decomposition functions for mixed monotonicity. In *IEEE Conf. on Decision and Control*, pages 5318–5322, 2019. doi:[10.1109/CDC40024.2019.9030065](https://doi.org/10.1109/CDC40024.2019.9030065).
- 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*, page 4944–4953, 2018. URL <https://arxiv.org/abs/1811.00866>.
- H. Zhang, H. Chen, C. Xiao, S. Goyal, R. Stanforth, Bo Li, D. Boning, and C-J. Hsieh. Towards stable and efficient training of verifiably robust neural networks. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=Skxuk1rFwB>.