Proceedings of Machine Learning Research vol 144:1–13, 2021

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

Saber Jafarpour* SABER@GATECH.EDU
Georgia Institute of Technology

Matthew Abate* MATT.ABATE@GATECH.EDU
Georgia Institute of Technology

Alexander Davydov* DAVYDOV@UCSB.EDU
University of California, Santa Barbara

Francesco Bullo BULLO@UCSB.EDU
University of California, Santa Barbara

Samuel Coogan SAM.COOGAN@GATECH.EDU
Georgia Institute of Technology

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 given network. Second, using ℓ_∞-matrix measures, we propose sufficient conditions for well-posedness of both original and embedded system and design an iterative algorithm to compute the ℓ_∞-norm box robustness margins for reachability and classification problems. Third, we propose a novel input-dependent output transformation that leads to more accurate evaluations of certified adversarial robustness in classification problems. Finally, we train a Non-Euclidean Monotone Operator Network (NEMON) on the MNIST dataset. We perform numerical simulations comparing the accuracy and run time of our mixed-monotone approach with the existing robustness verification approaches in the literature for estimating the certified adversarial robustness.

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

* These authors contributed equally

1. Introduction

Neural networks are increasingly deployed in real-world applications, including natural language processing, computer vision, and self-driving vehicles. However, they are notoriously vulnerable to adversarial attacks; slight perturbations in the input can lead to large deviations in the output (Szegedy et al., 2014). Understanding this sensitivity to input is essential in safety-critical applications, since consequences of such adversarial perturbations can be disastrous. A classical approach for studying the robustness of a neural network is to use its input-output Lipschitz constant. While computing the exact Lipschitz constant of a neural network is shown to be an NP-hard problem (Virmaux and Scaman, 2018), several works provide tractable algorithms for estimating the input-output Lipschitz constants (Szegedy et al., 2014; Combettes and Pesquet, 2020; Fazlyab et al., 2019). Nonetheless, these algorithms suffer from two major drawbacks. First, most of these approaches are either too conservative (for instance (Szegedy et al., 2014)) or are not scalable to large-scale neural networks (for instance (Virmaux and Scaman, 2018; Combettes and Pesquet, 2020)). Second, these algorithms provide estimates on the ℓ_2-norm input-output Lipschitz constants of the neural networks. For neural networks with large-scale inputs under widespread adversarial perturbations (examples of such adversarial perturbations can be found in (Szegedy et al., 2014)), the ℓ_2-norm input-output Lipschitz constants are known to provide an overly conservative measure of robustness and are less informative than their ℓ_∞-norm counterparts.

Implicit neural networks are a class of recently-developed learning models that replace the notion of layer with an implicit fixed-point equation (Bai et al., 2019; El Ghaoui et al., 2021). They can be considered as infinite-depth weight-tied neural networks where recursive function evaluation is performed via solving a single implicit algebraic equation. The implicit framework generalizes many classical neural networks including feedforward neural networks, convolutional neural networks, and residual networks (El Ghaoui et al., 2021). Compared to traditional neural networks, implicit neural networks feature competitive accuracy and reduced memory consumption (Bai et al., 2019). Despite these advantages, implicit neural networks can suffer from stability and robustness issues. Indeed, due to the implicit nature of their recursive rules, the robustness verification of implicit neural networks is challenging; many of the classical robustness analysis tools for traditional neural networks are either not applicable to implicit neural networks or will lead to conservative results. In this paper, we present a novel approach, derived from mixed monotone systems theory, for studying robustness of implicit neural network. Unlike the robustness verification approaches based on estimates of Lipschitz constants, our framework takes into account how the ℓ_∞-error bounds propagate through the network and is scalable with the size of the network.

Related works

Implicit learning models. Implicit neural networks have been proposed as a generalization of feedforward neural networks (Bai et al., 2019; El Ghaoui et al., 2021). Deep equilibrium networks (Bai et al., 2019). Implicit deep learning model (El Ghaoui et al., 2021). (Kag et al., 2020) not suffer from vanishing nor exploding gradients. One of the main challenges in studying implicit neural networks is their well-posedness, *i.e.*, existence and unqiueness of solutions for their fixed-point equation. (El Ghaoui et al., 2021) proposes a sufficient spectral condition for convergence of the Picard iterations associated with the fixed-point equation of implicit networks. In (Winston and Kolter, 2020; Revay et al., 2020), using monotone operator theory, a suitable parametrization of the weight matrix is proposed which guarantees the stable convergence of suitable fixed-point iterations.

In (Jafarpour et al., 2021b), non-Euclidean contraction theory with respect to the ℓ_∞-norm has been used to design implicit neural networks and study their well-posedness, stability, and robustness.

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). Several different strategies have been proposed in the literature for designing 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 any formal robustness guarantees (Madry et al., 2018; Carlini and Wagner, 2017). Motivated by application of neural networks in safety-critical system, there has been a recent interest in providing robustness guarantee for them. Several such examples for certified robustness training and analysis include (Wong and Kolter, 2018; Zhang et al., 2018; Gowal et al., 2018; Zhang et al., 2020; Mirman et al., 2018; Cohen et al., 2019). Regarding implicit neural networks, there are 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, the ellipsoid methods based on semi-definite programming (Chen et al., 2021) and the interval-bound propagation method (Anonymous, 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) 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 a mixed monotone when there exists a related decomposition function that separates the system's vector field or update map into increasing and decreasing components (Yang et al., 2019; Coogan, 2020). 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 (Meyer et al., 2019) 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 network's state 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 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 propose a new classifier variable, again based upon mixed-monotone theory, that leads to 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 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 approach significantly outperforms the other two methods.

2. Mathematical preliminaries

Vectors and matrices. The set of $n \times n$ positive-definite matrices is denoted by $\mathbb{S}^n$. 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 $\lceil A \rceil^{\mathrm{Mzl}}$ and $\lfloor A \rfloor^{\mathrm{Mzl}}$, respectively, where

$$(\lceil A \rceil^{\mathrm{Mzl}})_{ij} = \begin{cases} A_{ij} & A_{ij} \geq 0 \text{ or } i = j \\ 0 & \text{otherwise,} \end{cases} \qquad \lfloor A \rfloor^{\mathrm{Mzl}} = A - \lceil A \rceil^{\mathrm{Mzl}}.$$

For every two 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 $C \otimes D$. For every two vectors $v, w \in \mathbb{R}^n$ and every $i \in \{1, \dots, n\}$, the vector obtained by replacing the i^{th} element of v by the i^{th} element of w is denoted by $v_{i:w}$.

Matrix measures. For every $\eta \in \mathbb{R}^n_{>0}$, 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 $\mu_{\infty,[\eta]^{-1}}(A) = \lim_{h \to 0^+} \frac{\|I_n + hA\|_{\infty,[\eta]^{-1}} - 1}{h}$, where $\|\cdot\|_{\infty,[\eta]^{-1}}$ also denotes its induced matrix norm.

Lipschitz constants. Let $\mathsf{F} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ 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* $\mathsf{F}_\alpha : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ by $\mathsf{F}_\alpha = (1 - \alpha)\mathsf{I} + \alpha\mathsf{F}$, where I is the identity map on $\mathbb{R}^n$. Given a positive vector $\eta \in \mathbb{R}^n_{>0}$, $\mathsf{F}(x, u)$ is Lipschitz in x with respect to the norm $\|\cdot\|_{\infty,[\eta]^{-1}}$ with constant $\mathrm{Lip}^x_{\infty,[\eta]^{-1}}(\mathsf{F}) \in \mathbb{R}_{\geq 0}$ if, for every $x_1, x_2 \in \mathbb{R}^n$ and every $u \in \mathbb{R}^m$,

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

and $\mathsf{F}(x, u)$ is Lipschitz in u with respect to the norm $\|\cdot\|_{\infty,[\eta]^{-1}}$ with constant $\mathrm{Lip}^u_{\infty,[\eta]^{-1}}(\mathsf{F}) \in \mathbb{R}_{\geq 0}$ if, for every $u_1, u_2 \in \mathbb{R}^m$ and every $x \in \mathbb{R}^n$,

$$\|\mathsf{F}(x, u_1) - \mathsf{F}(x, u_2)\|_{\infty,[\eta]^{-1}} \leq \mathrm{Lip}^u_{\infty,[\eta]^{-1}}(\mathsf{F}) \|u_1 - u_2\|_{\mathcal{U}}$$

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

(i) $d_i(x, x, u, u) = \mathsf{F}_i(x, u)$;

(ii) $d_i(x, \widehat{x}, u, \widehat{u}) \leq d_i(y, \widehat{y}, u, \widehat{u})$, for every $x \leq y$ such that $x_i = y_i$, and every $\widehat{y} \leq \widehat{x}$;

(iii) $d_i(x, \widehat{x}, u, \widehat{u}) \leq d_i(x, \widehat{x}, v, \widehat{v})$, for every $u \leq v$ and every $\widehat{v} \leq \widehat{u}$.

Conditions (i)–(iii) are sometimes referred to as the Kamke conditions for mixed monotonicity[1] as developed in (Abate et al., 2021); see also (Coogan, 2020) for a equivalent infinitesimal characterization of mixed monotonicity. Every locally Lipschitz map F is mixed monotone with respect to 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 an optimal decomposition function is easily available in closed-form.

3. Implicit neural networks

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

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

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 $\Phi(x) = (\phi_1(x_1), \ldots, \phi_n(x_n))^\mathsf{T}$ is the diagonal matrix of activation functions, where, for every $i \in \{1, \ldots, n\}$, $\phi_i : \mathbb{R} \to \mathbb{R}$ 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 $U = [u_1, \ldots, u_m] \in \mathbb{R}^{r\times m}$ and its corresponding output data $Y = [y_1, \ldots, y_m] \in \mathbb{R}^{q\times m}$, the goal of the training optimization problem is to learn weights and biases which minimizes $\mathcal{L}(Y, CX + c)$ subject to the fixed-point equation $X = \Phi(AX + BU)$, where $\mathcal{L} : \mathbb{R}^{q\times m} \times \mathbb{R}^{q\times m} \to \mathbb{R}$ is a suitable cost function. Thus, the training optimization problem is given by

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

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 $\mu_{\infty,[\eta]^{-1}}(A) \leq \gamma$ are proposed, for some $\gamma < 1$ and some $\eta \in \mathbb{R}^n_{>0}$.

4. Robustness certificates via mixed monotone theory

One of the crucial features of neural networks in safety- and security-critical applications is their input-output robustness; the effect of input perturbations on the output. In this paper, we use theory of mixed monotone systems to study robustness of implicit neural networks.

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.

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

$$\begin{bmatrix}\underline{x}\\ \overline{x}\end{bmatrix} = \begin{bmatrix}\Phi(\lceil A\rceil^{\mathrm{Mzl}}\underline{x} + \lfloor A\rfloor^{\mathrm{Mzl}}\overline{x} + [B]^+\underline{u} + [B]^-\overline{u} + b)\\ \Phi(\lceil A\rceil^{\mathrm{Mzl}}\overline{x} + \lfloor A\rfloor^{\mathrm{Mzl}}\underline{x} + [B]^+\overline{u} + [B]^-\underline{u} + b)\end{bmatrix} := \begin{bmatrix}\mathsf{N}^{\mathrm{E}}(\underline{x},\overline{x},\underline{u},\overline{u})\\ \mathsf{N}^{\mathrm{E}}(\overline{x},\underline{x},\overline{u},\underline{u})\end{bmatrix},$$
$$\begin{bmatrix}\underline{y}\\ \overline{y}\end{bmatrix} = \begin{bmatrix}[C]^+ & [C]^-\\ [C]^- & [C]^+\end{bmatrix}\begin{bmatrix}\underline{x}\\ \overline{x}\end{bmatrix} + \begin{bmatrix}c\\ c\end{bmatrix}. \tag{3}$$

The embedded implicit neural network (3) can be considered as a neural network with the box input $[\underline{u},\overline{u}]$ and the box output $[\underline{y},\overline{y}]$ (see Figure 1). The following theorem studies well-posedness of the embedded implicit neural network (3) and its connection with robustness of the implicit neural network (1). We refer the reader to (Jafarpour et al., 2021a) for proof of this Theorem.

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

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

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

(ii) the α-average map $(\underline{x},\overline{x}) \to \begin{bmatrix}\mathsf{N}^{\mathrm{E}}_\alpha(\underline{x},\overline{x},\underline{u},\overline{u})\\ \mathsf{N}^{\mathrm{E}}_\alpha(\overline{x},\underline{x},\overline{u},\underline{u})\end{bmatrix}$ is a contraction mapping with minimum contraction factor $\mathrm{Lip}(\begin{bmatrix}\mathsf{N}^{\mathrm{E}}_{\alpha^*}\\ \mathsf{N}^{\mathrm{E}}_{\alpha^*}\end{bmatrix}) = 1 - \frac{1-\mu_{\infty,[\eta]^{-1}}(A)^+}{1-\min_{i\in\{1,\dots,n\}}(A_{ii})^-}$;

(iii) the α-average map N_α is a contraction mapping with minimum contraction factor $\mathrm{Lip}(\mathsf{N}_{\alpha^*}) = 1 - \frac{1-\mu_{\infty,[\eta]^{-1}}(A)^+}{1-\min_{i\in\{1,\dots,n\}}(A_{ii})^-}$;

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

$$\begin{bmatrix}\underline{x}^{k+1}\\ \overline{x}^{k+1}\end{bmatrix} = \begin{bmatrix}\mathsf{N}^{\mathrm{E}}_{\alpha^*}(\underline{x}^k,\overline{x}^k,\underline{u},\overline{u})\\ \mathsf{N}^{\mathrm{E}}_{\alpha^*}(\overline{x}^k,\underline{x}^k,\overline{u},\underline{u})\end{bmatrix}, \qquad \text{for every } k \in \mathbb{Z}_{\geq 0},\ \begin{bmatrix}\underline{x}^0\\ \overline{x}^0\end{bmatrix} \in \mathbb{R}^{2n}; \tag{4}$$

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

$$x^{k+1}_u = \mathsf{N}_{\alpha^*}(x^k_u, u), \qquad \text{for every } k \in \mathbb{Z}_{\geq 0},\ x^0 \in \mathbb{R}^n. \tag{5}$$

Remark 2 *The following remarks are in order.*

(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 $\frac{d}{dt}\begin{bmatrix}\underline{x}\\ \overline{x}\end{bmatrix} = -\begin{bmatrix}\underline{x}\\ \overline{x}\end{bmatrix} + \begin{bmatrix}\mathsf{N}^{\mathrm{E}}(\underline{x},\overline{x},\underline{u},\overline{u})\\ \mathsf{N}^{\mathrm{E}}(\overline{x},\underline{x},\overline{u},\underline{u})\end{bmatrix}$ (resp. $\frac{dx}{dt} = -x + \mathsf{N}(x,u)$).

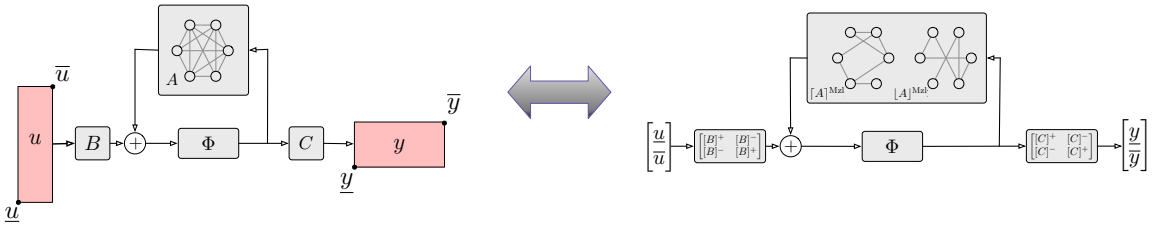

Figure 1: The original implicit neural network and its associated embedding network. The input-output behavior of the embedding system provides a box estimate for robustness of the original implicit neural network.

(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 $A \in \mathbb{R}^{kn \times kn}$ and a vector $\eta \in \mathbb{R}^{kn}_{>0}$ such that $\mu_{\infty,[\eta]^{-1}}(A) < 1$. In this case, the fixed-point of the embedded implicit network (3) is unique, can be computed explicitly, and corresponds exactly to the approach taken in (Gowal et al., 2018).

Robustness verification for classifiers. The embedded network output $[\underline{y}, \overline{y}]$ 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 to study classification problem by introducing a new classifier variable.

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 $i \in \{1, \dots, m\}$. Given bounds on the control input $u \in [\underline{u}, \overline{u}]$, we are interested to study the robustness of our classifier with respect to the input perturbations in $[\underline{u}, \overline{u}]$. To do this, we propose the *classifier variable* $z^u \in \mathbb{R}^{q-1}$ defined by

$$z^u := y^u_i \mathbb{1}_{q-1} - y^u_{-i}, \tag{6}$$

where $y^u_{-i} = (y^u_1, \dots, y^u_{i-1}, y^u_{i+1}, \dots, y^u_q) \in \mathbb{R}^{q-1}$. Note that $z^u \geq 0$ only when i is the correct label for the input u. Using (1), we can write (6) as follows

$$z^u = T^u y^u = T^u C x^* + T^u c, \tag{7}$$

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

$$\underline{z}^u = [T^u C]^+ \underline{x}^* + [T^u C]^- \overline{x}^* + T^u c, \tag{8}$$

where $\underline{x}^*, \overline{x}^*$ solve (3) and $\underline{u}, \overline{u}$ are the bounds on the input. It is easy to see that if $\underline{z}^u \geq 0$ then for all $u \in [\underline{u}, \overline{u}]$ the classification of u is i.

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. Given an implicit neural network (1) its certified adversarial robustness is its accuracy for detection of the correct label. To this end, we define the *deviation function* $\delta : \mathbb{R}_{\geq 0} \times \mathbb{R}^m \to \mathbb{R}$ by

$$\delta(\epsilon, u) = \max_{v \in \mathbb{R}^m} \{y_i^v - \max_{j \neq i} y_j^v \mid \|u - v\|_\infty \leq \epsilon, \;\; i = \operatorname{argmax}_j \, y_j^u\},$$

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 provide upper bounds on the deviation function and thus a 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 $\mathrm{Lip}_\infty^{u \to y} \in \mathbb{R}_{\geq 0}$, the output can be bounded as $\|y^u - y^v\|_\infty \leq \mathrm{Lip}_\infty^{u \to y} \|u - v\|_\infty$. We define $\delta^{\mathrm{Lip}}(\epsilon, u) := (y_i^u - \max_{j \neq i} y_j^u) - 2(\mathrm{Lip}_{u \to y}^\infty)\epsilon$. One can see that $\delta^{\mathrm{Lip}}(\epsilon, u) > 0$ 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 (Anonymous, 2022). Given an implicit neural network (1) with input perturbation $\|u - v\|_\infty \leq \epsilon$, the following fixed-point equation is proposed in (Anonymous, 2022):

$$\begin{bmatrix} \underline{x} \\ \overline{x} \end{bmatrix} = \begin{bmatrix} \Phi([A]^+ \underline{x} + [A]^- \overline{x} + [B]^+ \underline{u} + [B]^- \overline{u} + b) \\ \Phi([A]^+ \overline{x} + [A]^- \underline{x} + [B]^+ \overline{u} + [B]^- \underline{u} + b) \end{bmatrix}, \tag{9}$$

$$\begin{bmatrix} \underline{y} \\ \overline{y} \end{bmatrix} = \begin{bmatrix} [C]^+ & [C]^- \\ [C]^- & [C]^+ \end{bmatrix} \begin{bmatrix} \underline{x}^* \\ \overline{x}^* \end{bmatrix} + \begin{bmatrix} c \\ c \end{bmatrix}, \tag{10}$$

where $\underline{u} = u - \epsilon \mathbb{1}_m$, $\overline{u} = u + \epsilon \mathbb{1}_m$, and $(\underline{x}^*, \overline{x}^*)$ are the solutions of the fixed-point equation (9). 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 (9). The output of the neural network then can be bounded by the box $[\underline{y}, \overline{y}]$. We define $\delta^{\mathrm{IBP}}(\epsilon, u) = \underline{y}_i^* - \max_{j \neq i} \overline{y}_i^*$. One can see that $\delta^{\mathrm{IBP}}(\epsilon, u) > 0$ is a sufficient condition for certified adversarial robustness.

Method 3: Mixed monotone 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 map $\mathsf{N}_\alpha^{\mathrm{E}}$ with $\underline{u} = u - \epsilon \mathbb{1}_m$, $\overline{u} = u + \epsilon \mathbb{1}_m$ converges to $(\underline{x}^*, \overline{x}^*)$ and therefore, we have $y^v \in [\underline{y}^*, \overline{y}^*]$. Moreover, we can define $\delta^{\mathrm{MM}}(\epsilon, u) = \underline{y}_i^* - \max_{j \neq i} \overline{y}_i^*$. One can see that $\delta^{\mathrm{MM}}(\epsilon, u) > 0$ 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 $\delta^{\mathrm{MM-C}}(\epsilon, u) = \min_{i \in \{1,\dots,q-1\}} \underline{z}_i^u$, where $\underline{z}^u$ is as defined in equation (8). Then, one can obtain the tighter sufficient condition $\delta^{\mathrm{MM-C}}(\epsilon, u) > 0$ 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 $A = \begin{pmatrix} -\frac{1}{4} & -\frac{1}{4} \\ \frac{3}{4} & -\frac{1}{4} \end{pmatrix}$, $B = \begin{pmatrix} \frac{1}{2} & 1 \\ 1 & \frac{1}{2} \end{pmatrix}$, $C = I_2$, $b = c = \mathbb{0}_2$, and $\Phi(\cdot) = \mathrm{ReLU}(\cdot)$. Suppose that the nominal input is $u = \begin{pmatrix} \frac{1}{4} \\ \frac{3}{2} \end{pmatrix}$ and due to uncertainty, the input is in the box $v \in [\underline{u}, \overline{u}]$, where $\underline{u} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ and $\overline{u} = \begin{pmatrix} \frac{1}{3} \\ 2 \end{pmatrix}$. We compare the robustness bounds obtained using the Lipschitz bound approach, the interval bound propagation method, and our mixed monotone 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 (9), 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 approach, using the α-average iteration $\mathsf{N}_\alpha^{\mathrm{E}}$ 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.

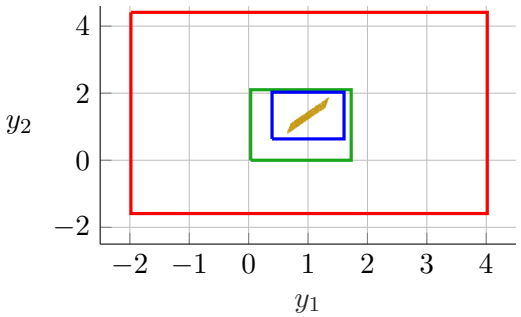

Figure 2: Problem Setting of Section 5.1: Comparing the application of Theorem 1 to existing verification methods for implicit neural networks. The yellow parallelogram shows different value of y^u for 1000 iid uniformly randomly selected $u = (u_1, u_2)^\mathsf{T}$ satisfying $0 \leq u_1 \leq \frac{1}{3}$ and $1 \leq u_2 \leq 2$. 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 dataset, a dataset of 70000 28×28 pixel images, 60000 of which are

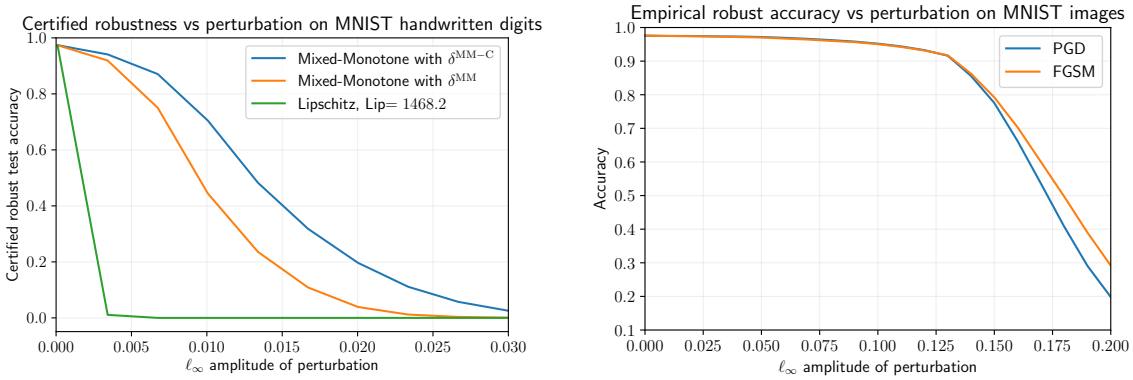

Figure 3: On the left is a plot of the certified adversarial robustness of the trained NEMON model using a Lipschitz method and two mixed monotonicity methods. For fixed ϵ, the fraction of test inputs which are certified robust are plotted. On the right is a plot of the empirical robustness of the same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

for training, and 10000 for testing. Pixel values are normalized in $[0, 1]$. 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 was trained for 15 epochs with a learning rate of 10^{-3}. After training, the model was validated 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 (PDG) 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 approach are tighter than the bounds provided by the Lipschitz constant. Third, we note the additional tightness in the bounds provided by computing the classification variable z^u. Finally, we note that although mixed monotonicity approaches to certified adversarial robustness 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 sizable ℓ_∞-perturbations.

6. Conclusions

⋆ From MA: remember to include

References

M. Abate and S. Coogan. Computing robustly forward invariant sets for mixed-monotone systems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 4553–4559, 2020. doi:10.1109/CDC42340.2020.9304461.

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.

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.

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.

Anonymous. Certified robustness for deep equilibrium models via interval bound propagation. In *Submitted to The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=y1PXylgrXZ. under review.

A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples, 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.

N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017. doi: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 *Thirty-Fifth Conference on 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 *36th International Conference 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.

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.

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.

S. Jafarpour, M. Abate, S. Davydov, F. Bullo, and S. Coogan. Robustness certificates for implicit neural networks: A mixed monotone contractive approach. 2021a. URL https://sites.engineering.ucsb.edu/~saber.jafarpour/Caltech_2021.pdf. 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*, November 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.

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.

Pierre-Jean Meyer, Alex Devonport, and Murat Arcak. Tira: Toolbox for interval reachability analysis. In *Proceedings of the 22nd ACM International Conference on Hybrid Systems: Computation and Control*, HSCC '19, page 224–229. Association for Computing Machinery, 2019. ISBN 9781450362825. doi:10.1145/3302504.3311808. URL https://doi.org/10.1145/3302504.3311808.

M. Mirman, T. Gehr, and M. Vechev. Differentiable abstract interpretation for provably robust neural networks. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3578–3586. PMLR, 10-15 Jul 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 (SP)*, pages 582–597, 2016. doi: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://proceedings.neurips.cc/paper/2018/file/d54e99a6c03704e95e6965532dec148b-Paper.pdf.

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 *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 5318–5322, 2019. doi:10.1109/CDC40024.2019.9030065.

L. Yang, O. Mickelin, and N. Ozay. On sufficient conditions for mixed monotonicity. *IEEE Transactions on Automatic Control*, 64(12):5080–5085, 2019. doi:10.1109/TAC.2019.2909815.

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. Gowal, 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.