Robust Implicit Networks via Non-Euclidean Contractions

Anonymous Author(s)
Affiliation
Address
email

Abstract

Implicit neural networks, a.k.a., deep equilibrium networks, are a class of implicit-depth learning models where function evaluation is performed by solving a fixed point equation. They generalize classic feedforward models and are equivalent to infinite-depth weight-tied feedforward networks. While implicit models show improved accuracy and significant reduction in memory consumption, they can suffer from ill-posedness and convergence instability.

This paper provides a new framework to design well-posed and robust implicit neural networks based upon contraction theory for the non-Euclidean norm ℓ_∞. Our framework includes (i) a novel condition for well-posedness based on one-sided Lipschitz constants, (ii) an average iteration for computing fixed-points, and (iii) explicit estimates on input-output Lipschitz constants. Additionally, we design a training problem with the well-posedness condition and the average iteration as constraints and, to achieve robust models, with the input-output Lipschitz constant as a regularizer. Our ℓ_∞ well-posedness condition leads to a larger polytopic training search space than existing conditions and our average iteration enjoys accelerated convergence. Finally, we perform several numerical experiments for function estimation and digit classification through the MNIST data set. Our numerical results demonstrate improved accuracy and robustness of the implicit models with smaller input-output Lipschitz bounds.

1 Introduction

Implicit neural networks are infinite-depth learning models with layers defined implicitly through a fixed-point equation. Examples of implicit neural networks include deep equilibrium models [Bai et al., 2019] and implicit deep learning models [El Ghaoui et al., 2019]. Implicit networks can be considered as generalizations of feedforward neural networks with input-injected weight tying, i.e., training parameters are transferable between layers. Indeed, in implicit networks, function evaluation is executed by solving a fixed-point equation and backpropagation is implemented by computing gradients using implicit differentiation. Due to these unique features, implicit models enjoy more flexibility and improved memory efficiency compared to traditional neural networks. At the same time, implicit networks can suffer from instability in their training due to the nonlinear nature of their fixed-point equations and can show brittle input-output behaviors due to their model flexibility.

It is known that implicit neural networks require careful tuning and initialization to avoid ill-posed training procedures. Indeed, without additional assumptions, their fixed-point equation may not have a unique solution and the numerical algorithms for finding their solutions might not converge. Several recent works in the literature have focused on studying well-posedness and convergence of the fixed-point equations of implicit networks using frameworks such as monotone operator theory [Winston and Kolter, 2020], contraction theory [El Ghaoui et al., 2019], and a mixture of both [Revay et al.,

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

2020]. Despite several insightful results, important questions about conditions for well-posedness of implicit networks and efficient algorithms that converge to their solutions are still open.

One of the key features of implicit neural networks is their flexibility, which might come at the cost of low input-output robustness. As first noted in [Szegedy et al., 2014], the input-output behavior of deep neural networks can be vulnerable to perturbations; close enough input data can lead to completely different outputs. This lack of robustness can lead to unreliable performance of neural networks in safety-critical applications. Among several notions of robustness, the Lipschitz constant of a neural network is a coarse but rigorous measure which can be used to estimate input-output sensitivity of the network [Szegedy et al., 2014]. For this reason, there has been a growing interest in estimating the input-output Lipschitz constant of deep neural networks with respect to the ℓ_2-norm [Fazlyab et al., 2019, Combettes and Pesquet, 2020]. However, it turns out that in some applications, the input-output Lipschitz constants with respect to non-Euclidean norms are more informative measures for studying robustness. One such application appears in the robustness analysis of neural networks with large-scale inputs under widely-distributed adversarial perturbations (examples of these adversarial perturbations can be found in [Szegedy et al., 2014]). For these examples, the input-output ℓ_2-Lipschitz constant does not provide complete information about robustness of the network; a neural network with small input-output ℓ_2-Lipschitz constant can be very sensitive to widespread entrywise-small perturbations of the input signal. On the other hand, the input-output ℓ_∞-Lipschitz constant provides a different metric which appears to be well-suited for the analysis of widespread distributed perturbations. Another application is the estimation of input signal confidence intervals from output deviations, where the input-output ℓ_∞-Lipschitz constant of the network provides more scalable bounds than its ℓ_2 counterpart.

Related works

Implicit learning models. Numerous works in learning theory have shown the power of deep learning models with implicit layers. In these learning models, the notion of layers are replaced by a composition rule, which can be either a fixed-point iteration or a solution to a differential equation. Well-known frameworks for deep learning using implicit infinite-depth layers include deep equilibrium networks [Bai et al., 2019], implicit deep learning [El Ghaoui et al., 2019], and Neural ODEs [Chen et al., 2018]. In [Travacca et al., 2020] a class of implicit learning models has been proposed that includes deep learning, nonlinear control, and mixed-integer programming as special cases. Implicit layers have also been used to study convex optimization problems [Agrawal et al., 2019] and to design control strategies [Amos et al., 2018]. Convergence to global minima of certain classes of implicit networks is studied in [Kawaguchi, 2021].

Well-posedness and numerical algorithms for fixed-point equations. There has been a recent interest in studying well-posedness and numerical stability of implicit-depth learning models. [El Ghaoui et al., 2019] proposes a sufficient spectral condition for well-posedness and 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. A recent influential survey on monotone operators is [Ryu and Boyd, 2016]. A recent survey on fixed point strategies in data science is given by [Combettes and Pesquet, 2021].

Robustness of learning models It is known that neural networks can be vulnerable to adversarial input perturbations [Szegedy et al., 2014]. A large body of literature is devoted to improve robustness of neural networks using various defense strategies against adversarial examples [Madry et al., 2018, Wong and Kolter, 2018]. While these strategies are effective in many scenarios, they do not provide formal guarantees for robustness. The input-output Lipschitz constant of a neural network is a rigorous metric for its worst-case sensitivity with respect to input perturbations. Several recent works have focused on estimating the Lipschitz constant and enforcing its boundedness. For example, [Fazlyab et al., 2019, 2020] propose a convex optimization framework using quadratic constraints and semidefinite programming to obtain upper bounds on Lipschitz constants of deep neural networks. In [Pauli et al., 2021], a training algorithm is designed to ensure boundedness of the Lipschitz constant of the neural network via a semidefinite program. Other methods for estimating the Lipschitz constant of deep neural networks include [Krishnan et al., 2020, Revay et al., 2020, 2021, Combettes and Pesquet, 2020].

Contributions

In this paper, we design novel implicit neural networks and study their well-posedness, stability, and robustness using tools and computational techniques from non-Euclidean contraction theory. First, we develop elements of a novel non-Euclidean monotone operator theory akin to the frameworks in [Bauschke and Combettes, 2017, Ryu and Boyd, 2016]. Using the concept of matrix measure, we introduce the essential notion of one-sided Lipschitz constant of a map. Based upon this notion, we prove a general fixed-point theorem with weaker requirements than classical results on Picard and Krasnosel'skii–Mann iterations. For maps with one-sided Lipschitz constant less than unity, we show that the average iteration (a.k.a. the Krasnosel'skii–Mann iteration) converges for sufficiently small step sizes and optimize its rate of convergence. For the special case of the weighted ℓ_∞-norm, we show that this average iteration can be accelerated by choosing a larger step size. Finally, we study perturbed fixed-point equations and establish a bound on the distance between perturbed and nominal equilibrium points as a function of one-sided Lipschitz condition. Second, for implicit neural networks, we use our new fixed-point theorem to (i) establish ℓ_∞-norm conditions for their well-posedness, (ii) design accelerated numerical algorithms for computing their solutions, and (iii) provide upper bounds on their input-output ℓ_∞-Lipschitz constants. Third, we use the average iteration and a polytopic characterization of the one-sided Lipschitz condition to design a training optimization problem. Additionally, by adding the input-to-output ℓ_∞-Lipschitz bound as a regularizer in the training problem, we study the tradeoff between the accuracy and the robustness of the trained neural network. Finally, we perform several numerical experiments illustrating improved performance for function estimation and digit classification on the MNIST data set compared to the state-of-the-art conditions in [El Ghaoui et al., 2019, Winston and Kolter, 2020]. Additionally, by adding the input-output Lipschitz constant as regularizer in the training problem, we observe improved robustness to some classes of adversarial perturbations. We include all relevant proofs in Appendix C

2 Review material

Matrix measures Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ and its induced norm on $\mathbb{R}^{n\times n}$. The matrix measure of $A \in \mathbb{R}^{n\times n}$ with respect to $\|\cdot\|$ is defined by $\mu(A) := \lim_{h\to 0^+} \frac{\|I_n + hA\| - 1}{h}$, that is, the one-sided directional derivative of the induced norm in direction of A, evaluated at I_n. Remarkably, the matrix measure is a tighter upper bound on the spectral abscissa of A than $\|A\|$ and $\mu(A) \leq 1$ is an unbounded subset of $\mathbb{R}^{n\times n}$ strictly containing the compact ball $\|A\| \leq 1$. We refer to [Desoer and Haneda, 1972] for a list of properties enjoyed by matrix measures.

We will be specifically interested in diagonally weighted ℓ_∞ norms defined by

$$\|x\|_{\infty,[\eta]^{-1}} = \max_i \frac{1}{\eta_i}|x_i|, \tag{1}$$

where, given a positive vector $\eta \in \mathbb{R}^n_{>0}$, we use $[\eta]$ to denote the diagonal matrix with diagonal entries η. The corresponding matrix norm and measure are

$$\|A\|_{\infty,[\eta]^{-1}} = \max_{i\in\{1,\dots,n\}} \sum_{j=1}^n \frac{\eta_j}{\eta_i}|a_{ij}|, \quad \mu_{\infty,[\eta]^{-1}}(A) = \max_{i\in\{1,\dots,n\}} \Big(a_{ii} + \sum_{j=1, j\neq i}^n |a_{ij}|\frac{\eta_j}{\eta_i}\Big). \tag{2}$$

Lipschitz maps Given a norm $\|\cdot\|$ with induced matrix measure $\mu(\cdot)$, a differentiable map $\mathsf{F} : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with constant $\mathrm{Lip}(\mathsf{F}) \in \mathbb{R}_{\geq 0}$ if

$$\|D\mathsf{F}(x)\| \leq \mathrm{Lip}(\mathsf{F}) \qquad \text{for all } x \in \mathbb{R}^n. \tag{3}$$

For example, for an affine $\mathsf{F}(x) = Ax + b$, the (smallest) Lipschitz constant is $\mathrm{Lip}(\mathsf{F}) = \|A\|$.

One-sided Lipschitz maps Given a norm $\|\cdot\|$, a differentiable map $\mathsf{F} : \mathbb{R}^n \to \mathbb{R}^n$ is one-sided Lipschitz continuous with constant $\mathrm{osL}(\mathsf{F}) \in \mathbb{R}$ if

$$\mu(D\mathsf{F}(x)) \leq \mathrm{osL}(\mathsf{F}) \qquad \text{for all } x \in \mathbb{R}^n. \tag{4}$$

For example, for an affine $\mathsf{F}(x) = Ax + b$, the (smallest) one-sided Lipschitz constant is $\mathrm{osL}(\mathsf{F}) = \mu(A)$. Note that (i) the one-sided Lipschitz constant is upper bounded by the Lipschitz constant, (ii) a Lipschitz continuous map is always one-sided Lipschitz continuous, and (iii) the one-sided

Lipschitz constant may be negative. For a more in-depth review we refer to Appendix A. The notion of one-sided Lipschitz continuity unifies several important concepts in dynamical systems and optimization theory. In operator theory, the map F is called a monotone operator if it is one-sided Lipschitz continuous with respect to the ℓ_2-norm with the constant $\mathrm{osL}(\mathsf{F}) > 0$ [Ryu and Boyd, 2016, Bauschke and Combettes, 2017]. In control theory, the vector field F is called strongly infinitesimally contracting if it is one-sided Lipschitz continuous with the constant $\mathrm{osL}(\mathsf{F}) < 0$ [Desoer and Haneda, 1972, Lohmiller and Slotine, 1998, Pavlov et al., 2004]. In what follows, we let $\mathrm{osL}_{\infty,[\eta]^{-1}}(\mathsf{F}) \in \mathbb{R}$ denote the one-sided Lipschitz constants with respect to the weighted ∞-norm.

3 Fixed-point equations and one-sided Lipschitz constants

In this section, we show that the notion of one-sided Lipschitz constant can be used to study solvability of fixed-point equation:

$$x = \mathsf{F}(x), \tag{5}$$

where $\mathsf{F} : \mathbb{R}^n \to \mathbb{R}^n$ is a differentiable map. Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$, then in view of the Banach fixed-point theorem, a simple sufficient condition for existence of a unique solution for the fixed-point equation (5) is $\mathrm{Lip}(\mathsf{F}) < 1$. We note that the sufficient condition $\mathrm{Lip}(\mathsf{F}) < 1$ depends on the specific form of the fixed-point equation (5) and can be relaxed by a suitable rewriting of this fixed-point equation. Given an averaging parameter $\alpha \in (0, 1]$ we define the <u>average map</u> $\mathsf{F}_\alpha : \mathbb{R}^n \to \mathbb{R}^n$ by $\mathsf{F}_\alpha := (1-\alpha)\mathsf{I} + \alpha\mathsf{F}$, where I is the identity map. Using this notion, an equivalent reformulation of the fixed-point equation (5) is:

$$x = (1-\alpha)x + \alpha\mathsf{F}(x) = \mathsf{F}_\alpha(x). \tag{6}$$

For $\alpha = 1$, we have $\mathsf{F}_\alpha(x) = \mathsf{F}(x)$ and equation (6) coincides with equation (5). For every $\alpha \in (0, 1)$, the map F_α is different from F but equations (5) and (6) are equivalent. Hence, if $\mathrm{Lip}(\mathsf{F}_\alpha) < 1$, then by the Banach fixed point theorem, the fixed point equation (6) (and therefore the fixed point equation (5)) has a unique solution x^* and the sequence $\{y_k\}_{k=1}^{\infty}$ defined by

$$y_{k+1} = (1-\alpha)y_k + \alpha\mathsf{F}(y_k), \qquad \text{for all } k \in \mathbb{Z}_{\geq 0} \tag{7}$$

converges geometrically to x^* with rate $\mathrm{Lip}(\mathsf{F}_\alpha)$. As a result of the parametrization (6), the condition $\mathrm{Lip}(\mathsf{F}) < 1$ for existence and uniqueness of the fixed-point can be relaxed to sufficient conditions

$$\mathrm{Lip}(\mathsf{F}_\alpha) < 1, \tag{8}$$

parametrized by $\alpha \in (0, 1]$. Additionally, if condition (8) is satisfied, then algorithm (7) computes the fixed point x^*. It can be shown that the condition (8) becomes less conservative as α decreases. The next theorem shows that in the limit as $\alpha \to 0^+$, condition (8) approaches the condition $\mathrm{osL}(\mathsf{F}) < 1$.

Theorem 1 (Fixed points via one-sided Lipschitz conditions)**.** *Let $\mathsf{F} : \mathbb{R}^n \to \mathbb{R}^n$ be differentiable and Lipschitz with respect to a norm $\|\cdot\|$. Define the average map $\mathsf{F}_\alpha = (1-\alpha)\mathsf{I} + \alpha\mathsf{F}$ and, for $\ell, c > 0$, the function $\gamma_{\ell,c} :]0, \frac{c}{(c+\ell+1)(\ell+1)}[\to \mathbb{R}$ by:*

$$\gamma_{\ell,c}(\alpha) := \left(1 + \alpha c - \frac{\alpha^2(\ell+1)^2}{1-\alpha(\ell+1)}\right)^{-1}.$$

Then the following statements are equivalent:

(i) $\mathrm{osL}(\mathsf{F}) < 1 - c$,

(ii) $\mathrm{Lip}(\mathsf{F}_\alpha) = \gamma_{\ell,c}(\alpha)$, *for* $\ell = \mathrm{Lip}(\mathsf{F})$ *and* $0 < \alpha < \frac{c}{(c+\ell+1)(\ell+1)}$.

Moreover, if the equivalent conditions (i) or (ii) hold, then, for condition number $\kappa = \frac{1+\mathrm{Lip}(\mathsf{F})}{1-\mathrm{osL}(\mathsf{F})}$,

(iii) F *has a unique fixed point* x^*;

(iv) for $0 < \alpha < \frac{1}{\kappa(\kappa+1)}$, F_α *is a contraction mapping with contraction factor* $\gamma_{\ell,c}(\alpha) < 1$;

(v) the <u>minimum</u> contraction factor $\gamma^*_{\ell,c} = 1 - \frac{1}{4\kappa^2} + \frac{1}{8\kappa^3} + \mathcal{O}\left(\frac{1}{\kappa^4}\right)$ *and the minimizing averaging parameter* α^* *of* F_α *is*

$$\alpha^* = \frac{\kappa}{1-\mathrm{osL}(\mathsf{F})}\left(1 - \frac{1}{\sqrt{1+1/\kappa}}\right) = \frac{1}{1-\mathrm{osL}(\mathsf{F})}\left(\frac{1}{2\kappa^2} - \frac{3}{8\kappa^3} + \mathcal{O}\left(\frac{1}{\kappa^4}\right)\right).$$

The average iteration (6) is often referred to as the Krasnosel'skii–Mann iteration or the damped iteration [Bauschke and Combettes, 2017]. Compared to [Bauschke and Combettes, 2017, Theorem 5.15], Theorem 1(iv) studies convergence of the Krasnosel'skii–Mann iteration for arbitrary norms, proposes a weaker convergence condition of the form $\mathrm{osL}(\mathsf{F}) < 1$ (hence, F need not be non-expansive). However, it ensures convergence for only sufficiently small $\alpha > 0$ and assumes that F is differentiable (as will be shown, however, the latter assumption can be relaxed).

3.1 Accelerated convergence for weighted ℓ_∞ norms

For diagonally weighted ℓ_∞ norms, one can strengthen Theorem 1(iv) to prove the convergence of the average iteration (6) on a larger domain of the parameter α.

Theorem 2 (Accelerated fixed point algorithm for ℓ_∞ norms)**.** *Let $\mathsf{F} : \mathbb{R}^n \to \mathbb{R}^n$ be differentiable and Lipschitz with respect to the weighted non-Euclidean norm $\|\cdot\|_{\infty,[\eta]^{-1}}$. Define the average map $\mathsf{F}_\alpha = (1-\alpha)\mathsf{I} + \alpha\mathsf{F}$ and pick $\mathrm{diagL}(\mathsf{F}) \in [-\mathrm{Lip}(\mathsf{F}), \mathrm{osL}(\mathsf{F})]$ to satisfy*

$$\mathrm{diagL}(\mathsf{F}) \leq \min_{i\in\{1,\dots,n\}} \inf_{x\in\mathbb{R}^n} D\mathsf{F}_{ii}(x). \tag{9}$$

If $\mathrm{osL}(\mathsf{F}) < 1$, then F has a unique fixed-point x^* and

(i) for $0 < \alpha \leq \dfrac{1}{1-\mathrm{diagL}(\mathsf{F})}$, F_α is a contraction mapping with the contraction factor $1-\alpha(1-\mathrm{osL}(\mathsf{F})) < 1$;

(ii) the minimum contraction factor and minimizing averaging parameter of F_α are, respectively,

$$\mathrm{Lip}(\mathsf{F}_{\alpha^*}) = 1 - \frac{1-\mathrm{osL}(\mathsf{F})}{1-\mathrm{diagL}(\mathsf{F})} = 1 - \frac{1}{\kappa_\infty}, \qquad \textit{for } \kappa_\infty = \frac{1-\mathrm{diagL}(\mathsf{F})}{1-\mathrm{osL}(\mathsf{F})} \leq \frac{1+\mathrm{Lip}(\mathsf{F})}{1-\mathrm{osL}(\mathsf{F})},$$
$$\alpha^* = \frac{1}{1-\mathrm{diagL}(\mathsf{F})}.$$

Note that $\mathrm{diagL}(\mathsf{F})$ is well-defined because ofthe Lipschitz continuity assumption. Specifically, one can show that $\mathrm{diagL}(\mathsf{F})$ is the minus the minimum over $i \in \{1,\dots,n\}$ of the one-sided Lipschitz constants of the maps $x_i \mapsto -\mathsf{F}(x_i, x_{-i})$ at $x_{-i} = (x_1,\dots,x_{i-1},x_{i+1},\dots,x_n)$ fixed.

It is instructive to compare the minimum contraction factor in the general Theorem 1 with the minimum contraction factor for ℓ_∞ norms in Theorem 2 and how they depend upon the corresponding condition numbers κ and κ_∞. We note that (i) the relevant condition number diminishes $\kappa \geq \kappa_\infty$, and (ii) the minimum contraction factor $\mathrm{Lip}(\mathsf{F}_{\alpha^*}) = 1 - \frac{1}{4\kappa^2} + \mathcal{O}(1/\kappa^4)$ improves to $\mathrm{Lip}(\mathsf{F}_{\alpha^*}) = 1 - \frac{1}{\kappa_\infty}$. This acceleration justifies the title of this section.

3.2 Perturbed fixed-point problems

In this subsection, we focus on solvability of the perturbed fixed-point equation:

$$x = \mathsf{F}(x, u), \tag{10}$$

where $\mathsf{F} : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}^n$ is differentiable in x. We define $\mathsf{F}_u(x) = \mathsf{F}(x,u)$ and $\mathsf{F}_x(u) = \mathsf{F}(x,u)$. Given a norm $\|\cdot\|_{\mathcal{X}}$ in $\mathbb{R}^n$ and $\|\cdot\|_{\mathcal{U}}$ in $\mathbb{R}^r$, F is Lipschitz in its first argument with constant $\mathrm{Lip}_x(\mathsf{F}) \in \mathbb{R}_{\geq 0}$ if

$$\|\mathsf{F}(x_1,u) - \mathsf{F}(x_2,u)\|_{\mathcal{X}} \leq \mathrm{Lip}_x(\mathsf{F})\|x_1 - x_2\|_{\mathcal{X}} \quad \text{for all } x_1, x_2 \in \mathbb{R}^n \text{ and } u \in \mathbb{R}^r,$$

and it is Lipschitz in its second argument with constant $\mathrm{Lip}_u(\mathsf{F}) \in \mathbb{R}_{\geq 0}$ if

$$\|\mathsf{F}(x,u_1) - \mathsf{F}(x,u_2)\|_{\mathcal{X}} \leq \mathrm{Lip}_u(\mathsf{F})\|u_1 - u_2\|_{\mathcal{U}} \quad \text{for all } x \in \mathbb{R}^n \text{ and } u_1, u_2 \in \mathbb{R}^r,$$

and it is one-sided Lipschitz in its first argument with constant $\mathrm{osL}_x(\mathsf{F}) \in \mathbb{R}$ if

$$\mu(D_x\mathsf{F}(x,u)) \leq \mathrm{osL}_x(\mathsf{F}) \quad \text{for all } x_1, x_2 \in \mathbb{R}^n \text{ and } u \in \mathbb{R}^r.$$

The following result, which is in the spirit of Lim's Lemma [Lim, 1985], provides an upper bound on the distance between fixed-points of the perturbed equation (10).

Theorem 3 (Perturbed fixed-points)**.** *Given a norm $\|\cdot\|_{\mathcal{X}}$ in $\mathbb{R}^n$ and a norm $\|\cdot\|_{\mathcal{U}}$ in $\mathbb{R}^r$, consider a map $\mathsf{F}:\mathbb{R}^n\times\mathbb{R}^r\to\mathbb{R}^n$ differentiable in the first argument and Lipschitz in both arguments. If F is one-sided Lipschitz with constant $\mathrm{osL}_x(\mathsf{F})<1$, then*

(i) for every $u\in\mathbb{R}^m$, the map F_u has a unique fixed point x_u^*;

(ii) for every $u,v\in\mathbb{R}^m$, $\|x_u^*-x_v^*\|_{\mathcal{X}}\leq\dfrac{\mathrm{Lip}_u(\mathsf{F})}{1-\mathrm{osL}_x(\mathsf{F})}\|u-v\|_{\mathcal{U}}$.

Finally, Theorems 1, 2, and 3 are not directly applicable to activation function that are not differentiable. In Appendix C.3, we show that for specific form of the fixed-point equation (5), where $\mathsf{F}=\Phi\circ\mathsf{H}$ and $\Phi:\mathbb{R}^n\to\mathbb{R}^n$ is a weakly increasing, non-expansive, diagonal activation function and $\mathsf{H}:\mathbb{R}^n\times\mathbb{R}^r\to\mathbb{R}^n$ is a differentiable function, all of the conclusions of Theorems 1, 2, and 3 by requiring equation (9) to hold almost everywhere.

4 Contraction analysis of implicit neural networks

Given $A\in\mathbb{R}^{n\times n}$, $B\in\mathbb{R}^{n\times r}$, $C\in\mathbb{R}^{q\times n}$, and $D\in\mathbb{R}^{q\times r}$, we consider the implicit neural network

$$x=\Phi(Ax+Bu):=\mathsf{N}(x,u),\qquad y=Cx+Du, \tag{11}$$

where $x\in\mathbb{R}^n$, $u\in\mathbb{R}^r$, $y\in\mathbb{R}^q$, and $\Phi:\mathbb{R}^n\to\mathbb{R}^n$ is defined by $\Phi(x)=(\phi_1(x_1),\dots,\phi_n(x_n))$. For every $i\in\{1,\dots,n\}$, we assume the activation function $\phi_i:\mathbb{R}\to\mathbb{R}$ is weakly increasing, i.e., $\phi_i(x_i)\geq\phi_i(z_i)$ for $x_i\geq z_i$, and non-expansive, i.e., $|\phi_i(x_i)-\phi_i(z_i)|\leq|x_i-z_i|$ for all x_i and z_i; if ϕ_i is differentiable, these conditions are equivalent to $0\leq\phi_i'(x_i)\leq 1$ for all $x_i\in\mathbb{R}$.

We are able to provide the following estimates on all relevant Lipschitz constants.

Theorem 4 (Lipschitz and one-sided Lipschitz constants for the implicit neural network)**.** *Consider the implicit neural network in equation (11) with weakly increasing and non-expansive activation functions Φ. With respect to $\|\cdot\|_{\infty,[\eta]^{-1}}$, $\eta\in\mathbb{R}^n_{>0}$, on $\mathbb{R}^n$ and $\|\cdot\|_{\mathcal{U}}$ on the input space $\mathbb{R}^r$, the map $\mathsf{N}:\mathbb{R}^n\times\mathbb{R}^r\to\mathbb{R}^n$ is one-sided Lipschitz continuous in the first variable and Lipschitz continuous in both variables with constants:*

$$\mathrm{osL}_x(\mathsf{N})=\mu_{\infty,[\eta]^{-1}}(A)_+\ ,\qquad \mathrm{Lip}_x(\mathsf{N})=\|A\|_{\infty,[\eta]^{-1}}\ , \tag{12}$$

$$\mathrm{Lip}_u(\mathsf{N})=\|B\|_{(\infty,[\eta]^{-1}),\mathcal{U}}\ ,\qquad \mathrm{diagL}(\mathsf{N})=\min_{i\in\{1,\dots,n\}}(A_{ii})_-\ , \tag{13}$$

where $(z)_+=z$ if $z\geq 0$ and $(z)_+=0$ if $z<0$; and $(z)_-=0$ if $z\geq 0$ and $(z)_-=z$ if $z<0$.

We now use these estimates to establish multiple properties of the implicit neural network.

Corollary 5 (Well posedness, input-state Lipschitz constant, and computation)**.** *Consider the model (11), with parameters (A,B,C,D) and with weakly increasing and non-expansive activation functions Φ. Define the average map $\mathsf{N}_\alpha:=(1-\alpha)\mathsf{I}+\alpha\mathsf{N}$ and consider the norms $\|\cdot\|_{\infty,[\eta]^{-1}}$, $\eta\in\mathbb{R}^n_{>0}$, on $\mathbb{R}^n$, $\|\cdot\|_{\mathcal{U}}$ on the input space $\mathbb{R}^r$ and $\|\cdot\|_{\mathcal{Y}}$ on the output space $\mathbb{R}^q$. Then*

(i) if $\mu_{\infty,[\eta]^{-1}}(A)<1$, then (11) is well posed, i.e., there exists a unique fixed point,

(ii) the map N_α is a contraction mapping for $0<\alpha\leq\alpha^*:=\big(1-\min_{i\in\{1,\dots,n\}}(A_{ii})_-\big)^{-1}$ with minimum contraction factor $\mathrm{Lip}(\mathsf{N}_{\alpha^*})=1-\frac{1-\mu_{\infty,[\eta]^{-1}}(A)_+}{1-\min_{i\in\{1,\dots,n\}}(A_{ii})_-}$.

(iii) the Lipschitz constants from input u to fixed point x_u^* and to the output $y=Cx_u^*+Du$ are

$$\mathrm{Lip}_{u\to x^*}:=\frac{\mathrm{Lip}_u(\mathsf{N})}{1-\mathrm{osL}_x(\mathsf{N})}=\frac{\|B\|_{(\infty,[\eta]^{-1}),\mathcal{U}}}{1-\mu_{\infty,[\eta]^{-1}}(A)_+}, \tag{14}$$

$$\mathrm{Lip}_{u\to y}:=\frac{\|B\|_{(\infty,[\eta]^{-1}),\mathcal{U}}\|C\|_{\mathcal{Y},(\infty,[\eta]^{-1})}}{1-\mu_{\infty,[\eta]^{-1}}(A)_+}+\|D\|_{\mathcal{Y},\mathcal{U}}. \tag{15}$$

In [El Ghaoui et al., 2019] a well-posedness condition of the form $\lambda_{\mathrm{pf}}(|A|)<1$ is proposed, where $|A|$ denotes the entrywise absolute value of the matrix A and λ_{pf} denotes the Perron-Frobenius eigenvalue. However, for the training procedure, this condition is relaxed to the convex condition $\|A\|_\infty<1$. It is easy to see that our well-posedness condition in Corollary 5(i) is less conservative than the condition $\lambda_{\mathrm{pf}}(|A|)<1$ and its convex relaxation of the form $\|A\|_\infty<1$ proposed in [El Ghaoui et al., 2019].

5 Training implicit neural networks

Problem setup Given an input data matrix $U = [u_1, \dots, u_m] \in \mathbb{R}^{r\times m}$ and a corresponding output data matrix $Y = [y_1, \dots, y_m] \in \mathbb{R}^{q\times m}$, we aim to learn matrices A, B, C, D so that the neural network (11) approximates the input-output relationship. We rewrite the model for matrix inputs as $\widehat{Y} = CX + DU$, where $X = \Phi(AX + BU)$. From Corollary 5(i), if each ϕ_i is weakly increasing and non-expansive, the fixed point problem is well-posed when $\mu_{\infty,[\eta]^{-1}}(A) < 1$ for some $\eta \in \mathbb{R}^n_{>0}$. We consider a training problem of the form

$$\begin{aligned} \min_{A,B,C,D,X} \quad & \mathcal{L}(Y, CX + DU) + \mathcal{P}(A, B, C, D) \\ & X = \Phi(AX + BU), \quad \mu_{\infty,[\eta]^{-1}}(A) \le \gamma, \end{aligned} \tag{16}$$

where $\mathcal{L}$ is a loss function assumed to be convex in its second argument, $\mathcal{P}$ is a convex penalty function, and $\gamma < 1$ is a hyperparameter ensuring the fixed point problem is well-posed. The penalty term serves to impose additional desired properties for the model, which could include sparsity, an upper triangular structure, or a bounded input-output Lipschitz constant.

Based upon equation (2) the set of matrices with $\mu_{\infty,[\eta]^{-1}}(A) \le \gamma$ is an unbounded convex polytope:

$$\begin{aligned} \mathcal{M}_{\gamma,\eta} &:= \{A \in \mathbb{R}^{n\times n} \mid \mu_{\infty,[\eta]^{-1}}(A) \le \gamma\} \\ &= \{A \in \mathbb{R}^{n\times n} \mid \exists T \in \mathbb{R}^{n\times n} \text{ s.t. } (A_{\mathrm{d}} + T)\eta \le \gamma\eta, -T \le (A - A_{\mathrm{d}}) \le T\}, \end{aligned} \tag{17}$$

where $A_{\mathrm{d}} \in \mathbb{R}^{n\times n}$ is the diagonal matrix with same diagonal entries as A.

Improving robustness via Lipschitz regularization We now focus on learning robust implicit neural networks with bounded Lipschitz constants via a regularization strategy. Setting both $\|\cdot\|_{\mathcal{U}}$ and $\|\cdot\|_{\mathcal{Y}}$ as $\|\cdot\|_\infty$ in the input-output Lipschitz bound (15), we get

$$\begin{aligned} \mathrm{Lip}_{u\to y} &= \frac{\|B\|_{(\infty,[\eta]^{-1}),(\infty)}\|C\|_{(\infty),(\infty,[\eta]^{-1})}}{1 - \mu_{\infty,[\eta]^{-1}}(A)_+} + \|D\|_{\infty,\infty} \\ &\le \frac{1}{2}\frac{\|B\|^2_{(\infty,[\eta]^{-1}),(\infty)} + \|C\|^2_{(\infty),(\infty,[\eta]^{-1})}}{1 - \mu_{\infty,[\eta]^{-1}}(A)_+} + \|D\|_{\infty,\infty}, \end{aligned}$$

where the inequality provides a convex upper bound for the input-output Lipschitz constant. Therefore, using the hyperparameter $\lambda > 0$, the regularized optimization problem is written as

$$\begin{aligned} \min_{A,B,C,D,X} \quad & \mathcal{L}(Y, CX + DU) + \lambda\Big(\frac{1}{2}\frac{\|B\|^2_{(\infty,[\eta]^{-1}),(\infty)} + \|C\|^2_{(\infty),(\infty,[\eta]^{-1})}}{1 - \mu_{\infty,[\eta]^{-1}}(A)_+} + \|D\|_{\infty,\infty}\Big) \\ & X = \Phi(AX + BU), \quad \mu_{\infty,[\eta]^{-1}}(A) \le \gamma. \end{aligned} \tag{18}$$

Backpropagation of gradients via average iteration From [El Ghaoui et al., 2019] we now show how the average iteration can be used to perform backpropagation via the implicit function theorem. For simplicity, we assume that each activation function ϕ_i is differentiable and consider mini-batches of size 1, i.e., we have $X = x \in \mathbb{R}^n$, $U = u \in \mathbb{R}^r$ and $\widehat{Y} = \widehat{y} \in \mathbb{R}^q$. Let x^* be the unique solution of the fixed-point equation (11). Then the chain rule implies

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial A} &= (\nabla_{x^*}\mathcal{L})x^\top, & \frac{\partial \mathcal{L}}{\partial B} &= (\nabla_{x^*}\mathcal{L})u^\top, \\ \frac{\partial \mathcal{L}}{\partial C} &= (\nabla_{\widehat{y}}\mathcal{L})x^\top, & \frac{\partial \mathcal{L}}{\partial D} &= (\nabla_{\widehat{y}}\mathcal{L})u^\top. \end{aligned}$$

Since $\mathcal{L}$ depends explicitly on $\widehat{y}$, computing $\nabla_{\widehat{y}}\mathcal{L}$ is straightforward. Computing $\nabla_{x^*}\mathcal{L}$ is more complicated since X^* is defined only implicitly. However, one can be shown that

$$\nabla_{x^*}\mathcal{L} = (C(I - D\Phi A)^{-1}D\Phi)^\top \nabla_{\widehat{y}}\mathcal{L}.$$

Since $\mu_{\infty,[\eta]^{-1}}(A) < 1$, by Lemma 8 we get that $\mu_{\infty,[\eta]^{-1}}(D\Phi A) < 1$. This implies that the matrix $G := (I_n - D\Phi A)^{-1}D\Phi \in \mathbb{R}^{n\times n}$ exists and is the solution to the following fixed-point equation [El Ghaoui et al., 2019, Lemma G.1]

$$G = D\Phi(AG + I_n). \tag{19}$$

Moreover, $\mu_{\infty,[\eta]^{-1}}(D\Phi A) < 1$ and Theorem 2 together imply that the fixed-point equation (19) has a unique solution G^* and, for every $0 < \alpha \leq \alpha^* := \big(1 - \min_i (A_{ii})_-\big)^{-1}$, the average iterations

$$G_{k+1} = (1-\alpha)I_n + \alpha D\Phi(AG_k + I_n), \qquad \text{for all } k \in \mathbb{Z}_{\geq 0}$$

is contracting with the minimum contraction factor $1 - \alpha^*\big(1 - \mu_{\infty,[\eta]^{-1}}(A)_+\big)$ at step size α^*.

6 Numerical Experiments

6.1 Learning a scalar nonlinear function

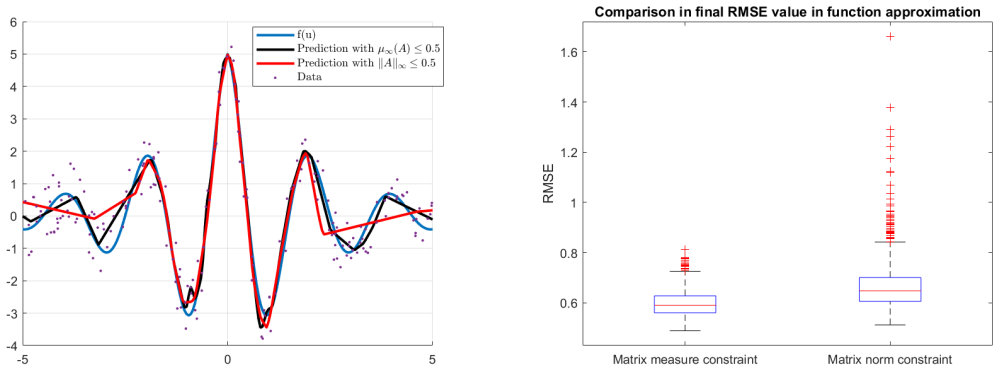

Figure 1: On the left is a plot of learned models for $\mu_\infty(A) \leq 0.5$ and the $\|A\|_\infty \leq 0.5$ compared to the data and the true function. On the right are final RMSE values over 1000 trials of function approximation.

We start[1] by comparing our model with the implicit model with an infinity norm constraint[2] as proposed in [El Ghaoui et al., 2019]. We aim to learn the scalar nonlinear function

$$f(u) = 5\cos(\pi u)\exp\Big(-\frac{1}{2}|u|\Big).$$

We select the input $u_i, i \in \{1, \ldots, m\}$ uniformly at random in $[-5, 5]$ with $m = 200$. We add noise at the output, $y(u) = f(u) + w$, where w is taken from a uniform distribution in $[-1, 1]$. We consider a model of order $n = 75$. For the comparison with the infinity norm constraint, in the training problem, $\mu_{\infty,[\eta]^{-1}}(A) \leq \gamma$ is replaced by $\|A\|_{\infty,[\eta]^{-1}} \leq \gamma$. For simulations, a value of $\gamma = 0.5$ was considered and a uniform vector $\eta = \mathbb{1}_n$ was used. Each model was trained using a LeakyReLU function $\phi_i(x) = \max\{x, 0.2x\}$ over 300 epochs. In Figure 1, sample approximations are plotted and final RMSE values are shown over 1000 trials. We see that the RMSE for our model with $\mu_\infty(A) \leq 0.5$ is about 10% lower than the RMSE corresponding to the model with $\|A\|_\infty \leq 0.5$.

6.2 MNIST experiments

In the digit classification data set MNIST [3] , input data are 28×28 pixel images of handwritten digits between 0-9. There are 60000 training images and 10000 test images. For training, images are reshaped into 784 dimensional column vectors and entries are scaled into the range $[0, 1]$. As a loss function, we use the cross-entropy. All models are of order $n = 100$, used the ReLU activation function $\phi_i(x) = (x)_+$, and are trained with a batch size of 300 over 10 epochs with a learning rate of 1.5×10^{-2}. For the model driven by the matrix measure constraint, $\mu_\infty(A) \leq 0.95$ is imposed, while for the model driven by the matrix norm constraint, $\|A\|_\infty \leq 0.95$ is imposed. Both models are additionally compared to a fully connected monotone operator network (MON)[4] [Winston and Kolter, 2020]. Curves for accuracy and loss versus epochs for the three models are shown in Figure 2. We observe that our model with $\mu_\infty(A) \leq 0.95$ performs better than the model with $\|A\|_\infty \leq 0.95$ and has a comparable perfomance to MON.

[1] All subsequent models are trained on a laptop with Intel Core i7-8565U CPU with 1.80 GHz processor.
[2] The software implementation is available at `https://github.com/beeperman/idl`.
[3] MNIST is licensed by CC BY-SA 3.0. and is available at `http://yann.lecun.com/exdb/mnist/`.
[4] The MON implementation is available at `https://github.com/locuslab/monotone_op_net`.

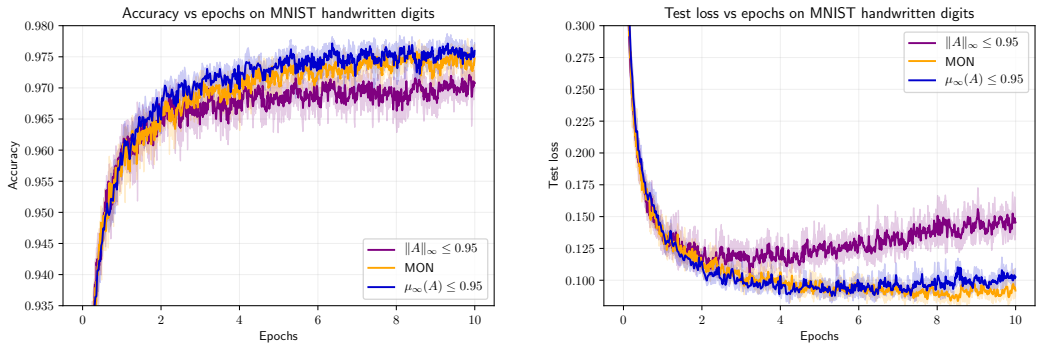

Figure 2: Performance comparison between $\mu_\infty(A) \leq 0.95$, $\|A\|_\infty \leq 0.95$, and monotone operator network (MON) on MNIST. The curves are generated by mean accuracy and loss over 5 different runs while light envelopes around the curves correspond to the standard deviation over the runs. Average best accuracy for $\mu_\infty(A) \leq 0.95$ is 0.9772, while it is 0.9721 for $\|A\|_\infty \leq 0.95$ and 0.9762 for MON. Note how the test loss begins to increase as epochs increase for $\|A\|_\infty \leq 0.95$, but not as substantially for $\mu_\infty(A) \leq 0.95$ or MON.

6.3 Robustness regularization experiments

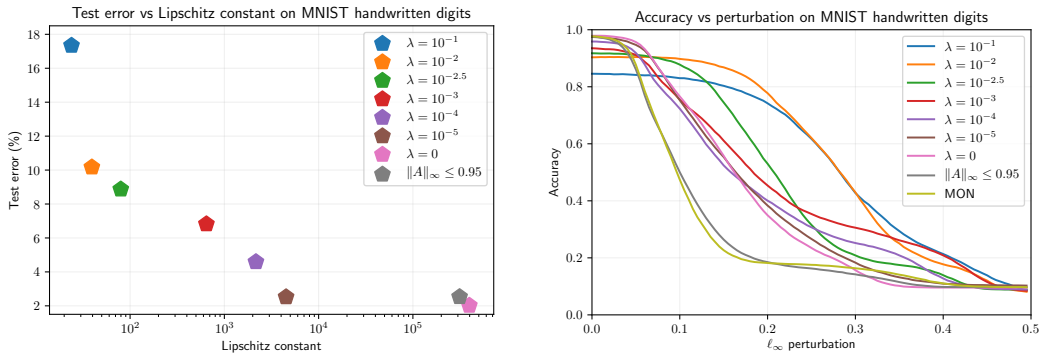

Figure 3: On the left is a plot of test error versus Lipschitz constant as parametrized by the regularization hyperparameter λ. We define the test error as 1 minus the accuracy. On the right is a plot of accuracy versus ℓ_∞ perturbation of a deterministic adversarial image inversion attack.

We study the robustness of our model compared to the model with $\|A\|_\infty \leq 0.95$ and MON on the MNIST data set. We train various models regularized by the input-output Lipschitz constant as in (18). Additionally, to verify robustness of the different models, we consider several adversarial attacks and plot the accuracy versus perturbation of such an attack. In Figure 3, we consider a continuous image inversion attack [Hosseini et al., 2017], where each pixel is perturbed in the direction of pixel value inversion with amplitude given by the ℓ_∞ perturbation. For more details on this and other types of adversarial perturbations, we refer to Appendix D. We observe that for $\lambda = 10^{-5}$, the regularized model achieves a two order of magnitude decrease in its input-output Lipschitz constant compared to the unregularized models. In addition, we see that the model with $\|A\|_\infty \leq 0.95$ and MON are more sensitive to the continuous image inversion attack than our proposed model. Moreover, as the regularization parameter λ increases, our model becomes increasingly robust to this attack.

7 Conclusion

Using non-Euclidean contraction theory, we propose a framework to study stability of fixed-point equations. We apply this framework to analyze well-posedness and convergence of implicit neural networks and to design an efficient training algorithm to incorporate robustness guarantees. For future research, we envision that our framework is applicable to study stability and robustness of implicit learning models with additional structure such as graph neural networks. Another interesting direction is establishing the tightness of our Lipschitz estimates.

References

A. Agrawal, B. Amos, S. Barratt, S. Boyd, S. Diamond, and J. Z. Kolter. Differentiable convex optimization layers. In Advances in Neural Information Processing Systems, volume 32, 2019.

B. Amos, I. Jimenez, J. Sacks, B. Boots, and J. Z. Kolter. Differentiable MPC for end-to-end planning and control. In Advances in Neural Information Processing Systems, volume 31, 2018.

S. Bai, J. Z. Kolter, and V. Koltun. Deep equilibrium models. In Advances in Neural Information Processing Systems, volume 32, 2019. URL https://arxiv.org/pdf/1909.01377.

H. H. Bauschke and P. L. Combettes. Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, 2 edition, 2017. ISBN 978-3-319-48310-8.

R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural ordinary differential equations. In Advances in Neural Information Processing Systems, page 6572–6583, 2018.

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.

P. L. Combettes and J.-C. Pesquet. Fixed point strategies in data science. IEEE Transactions on Signal Processing, 2021. doi:10.1109/TSP.2021.3069677.

A. Davydov, S. Jafarpour, and F. Bullo. Non-Euclidean contraction theory via semi-inner products. arXiv preprint arXiv:2103.12263, 2021.

C. A. Desoer and H. Haneda. The measure of a matrix as a tool to analyze computer algorithms for circuit analysis. IEEE Transactions on Circuit Theory, 19(5):480–486, 1972. doi:10.1109/TCT.1972.1083507.

L. El Ghaoui, F. Gu, B. Travacca, A. Askari, and A. Y. Tsai. Implicit deep learning. arXiv preprint arXiv:1908.06315, 2019.

Y. Fang and T. G. Kincaid. Stability analysis of dynamical neural networks. IEEE Transactions on Neural Networks, 7(4):996–1006, 1996. doi:10.1109/72.508941.

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, pages 11423–11434, 2019.

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.

W. He and J. Cao. Exponential synchronization of chaotic neural networks: a matrix measure approach. Nonlinear Dynamics, 55:55–65, 2009. doi:10.1007/s11071-008-9344-4.

H. Hosseini, B. Xiao, M. Jaiswal, and R. Poovendran. On the limitation of convolutional neural networks in recognizing negative images. In 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 352–358, 2017. doi:10.1109/ICMLA.2017.0-136.

K. Kawaguchi. On the theory of implicit deep learning: Global convergence with implicit layers. In International Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=p-NZIuwqhI4.

V. Krishnan, A. A. A. Makdah, and F. Pasqualetti. Lipschitz bounds and provably robust training by Laplacian smoothing. In Advances in Neural Information Processing Systems, 2020. URL https://arxiv.org/abs/2006.03712.

T. C. Lim. On fixed point stability for set-valued contractive mappings with applications to generalized differential equations. Journal of Mathematical Analysis and Applications, 110(2):436–441, 1985. doi:10.1016/0022-247X(85)90306-3.

W. Lohmiller and J.-J. E. Slotine. On contraction analysis for non-linear systems. Automatica, 34(6): 683–696, 1998. doi:10.1016/S0005-1098(98)00019-3.

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.

P. Pauli, A. Koch, J. Berberich, P. Kohler, and F. Allgower. Training robust neural networks using Lipschitz bounds. IEEE Control Systems Letters, 2021. doi:10.1109/LCSYS.2021.3050444.

A. Pavlov, A. Pogromsky, N. Van de Wouw, and H. Nijmeijer. Convergent dynamics, a tribute to Boris Pavlovich Demidovich. Systems & Control Letters, 52(3-4):257–261, 2004. doi:10.1016/j.sysconle.2004.02.003.

H. Qiao, J. Peng, and Z.-B. Xu. Nonlinear measures: A new approach to exponential stability analysis for Hopfield-type neural networks. IEEE Transactions on Neural Networks, 12(2):360–370, 2001. doi:10.1109/72.914530.

M. Revay, R. Wang, and I. R. Manchester. Lipschitz bounded equilibrium networks. arXiv preprint arXiv:2010.01732, 2020.

M. Revay, R. Wang, and I. R. Manchester. A convex parameterization of robust recurrent neural networks. IEEE Control Systems Letters, 5(4):1363–1368, 2021. doi:10.1109/LCSYS.2020.3038221.

E. K. Ryu and S. Boyd. Primer on monotone operator methods. Applied Computational Mathematics, 15(1):3–43, 2016.

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.

B. Travacca, L. El Ghaoui, and S. Moura. Implicit optimization: Models and methods. In IEEE Conf. on Decision and Control, pages 408–415, Dec 2020. doi:10.1109/CDC42340.2020.9304169.

E. Winston and J. Z. Kolter. Monotone operator equilibrium networks. In Advances in Neural Information Processing Systems, volume 33, pages 10718–10728, 2020. URL https://arxiv.org/pdf/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.

Checklist

(i) For all authors...

(a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

(b) Did you describe the limitations of your work? [Yes] We explicitly mention in Sections 3 and 4 that our proposed framework can be applied to a specific class of learning problems and all of the conditions obtained in this paper for well-posedness, convergence, and robustness of implicit networks are sufficient conditions (see Section 3).

(c) Did you discuss any potential negative societal impacts of your work? [No] Our work contributes to the fundamental study of supervised learning and neural networks and we do not see any direct path to any negative societal application.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

(ii) If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes] For every theorem, we state the full set of assumptions in the statement of the theorem. For details see Sections 3 and 4.

(b) Did you include complete proofs of all theoretical results? [Yes] We included all the proofs of the theoretical results in Appendix C.

(iii) If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We include the code in the supplementary material and we will make it available on github. We also used the standard MNIST dataset and two github software packages; we include links to all of these items.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We include all the training details in Sections 5 and 6.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] For Figure 2 we include error bars. Regarding Figure 3, the adversarial "image inversion" attack is deterministic so there is no need for error bars in this figure.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We have included the detail of the hardware used for training in Section 6.

(iv) If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] In the body of the paper, we cite references for every model used in our theoretical and experimental setting. Regarding our code, we adapt some existing software packages and we make references to these packages in the paper and in the code itself.

(b) Did you mention the license of the assets? [Yes] Yes, we did mention the license for the MNIST dataset and we also mention that the two github packages we use do not have an explicit license (hence most copyrights remain with the authors). We intend to license our code with a CC-BY-SA 4.0 license.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Our code for the numerical experiments in Section 6 is available in the supplemental material.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

(v) If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]