LIPSCHITZ-BOUNDED EQUILIBRIUM NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper introduces new parameterizations of equilibrium neural networks, i.e. networks defined by implicit equations. This model class includes standard multilayer and residual networks as special cases. The new parameterization admits a Lipschitz bound during training via unconstrained optimization: no projections or barrier functions are required. Lipschitz bounds are a common proxy for robustness and appear in many generalization bounds. Furthermore, compared to previous works we show well-posedness (existence of solutions) under less restrictive conditions on the network weights and more natural assumptions on the activation functions: that they are monotone and slope restricted. These results are proved by establishing novel connections with convex optimization, operator splitting on non-Euclidean spaces, and contracting neural ODEs. In image classification experiments we show that the Lipschitz bounds are very accurate and improve robustness to adversarial attacks.

1 INTRODUCTION

Deep neural network models have revolutionized the field of machine learning: their accuracy on practical tasks such as image classification and their scalability have led to an enormous volume of research on different model structures and their properties (LeCun et al., 2015). In particular, deep residual networks with skip connections He et al. (2016) have had a major impact, and neural ODEs have been proposed as an analog with "implicit depth" (Chen et al., 2018). Recently, a new structure has gained interest: *equilibrium networks* (Bai et al., 2019; Winston & Kolter, 2020), a.k.a. *implicit deep learning models* (El Ghaoui et al., 2019), in which model outputs are defined by implicit equations incorporating neural networks. This model class is very flexible: it is easy to show that includes many previous structures as special cases, including standard multi-layer networks, residual networks, and (in a certain sense) neural ODEs.

However model flexibility in machine learning is always in tension with model *regularity* or *robustness*. While deep learning models have exhibited impressive generalisation performance in many contexts it has also been observed that they can be very brittle, especially when targeted with adversarial attacks (Szegedy et al., 2014). In response to this, there has been a major research effort to understand and certify robustness properties of deep neural networks, e.g. Raghunathan et al. (2018); Tjeng et al. (2018); Liu et al. (2019); Cohen et al. (2019) and many others. Global Lipschitz bounds (a.k.a. incremental gain bounds) provide a somewhat crude but nevertheless highly useful proxy for robustness (Tsuzuku et al., 2018; Fazlyab et al., 2019), and appear in several analyses of generalization (e.g. (Bartlett et al., 2017; Zhou & Schoellig, 2019)).

Inspired by both of these lines of research, in this paper we propose new parameterizations of equilibrium networks with guaranteed Lipschitz bounds. We build directly on the monotone operator framework of Winston & Kolter (2020) and the work of Fazlyab et al. (2019) Lipschitz bounds.

The main contribution of our paper is the ability to enforce tight bounds on the Lipschitz constant of an equilibrium network during training with essentially *no extra computational effort*. In addition, we prove existence of solutions with less restrictive conditions on the weight matrix and more natural assumptions on the activation functions via novel connections to convex optimization and contracting dynamical systems. Finally, we show via small-scale image classification experiments that the proposed parameterizations can provide significant improvement in robustness to adversarial attacks with little degradation in nominal accuracy. Furthermore, we observe small gaps between certified Lipschitz upper bounds and observed lower bounds computed via adversarial attack.

2 RELATED WORK

Equilibrium networks, Implicit Deep Models, and Well-Posedness. As mentioned above, it has been recently shown that many existing network architectures can be incorporated into a flexible model set called an equilibrium network (Bai et al., 2019; Winston & Kolter, 2020) or implicit deep model (El Ghaoui et al., 2019). In this unified model set, the network predictions are made not by forward computation of sequential hidden layers, but by finding a solution to an implicit equation involving a single layer of all hidden units. One major question for this type of networks is its well-posedness, i.e. the existence and uniqueness of a solution to the implicit equation for all possible inputs. El Ghaoui et al. (2019) proposed a computationally verifiable but conservative condition on the spectral norm of hidden unit weight. In Winston & Kolter (2020), a less conservative condition was developed based on monotone operator theory. Similar monotonicity constraints were previously used to ensure well-posedness of a different class of implicit models in the context of nonlinear system identification (Tobenkin et al., 2017, Theorem 1). On the question of well-posedness, our contribution is a more flexible model set and more natural assumptions on the activation functions: that they are monotone and slope-restricted.

Neural Network Robustness and Lipschitz Bounds. The Lipschitz constant of a function measures the worst-case sensitivity of the function, i.e. the maximum "amplification" of difference in inputs to differences in outputs. The key features of a good Lipschitz bounded learning approach include a tight estimation for Lipschitz constant and a computationally tractable training method with bounds enforced. For deep networks, Tsuzuku et al. (2018) proposed a computationally efficient but conservative approach since its Lipschitz constant estimation method is based on composition of estimations for different layers. Similarly, El Ghaoui et al. (2019) proposed an estimation for equilibrium networks via input-to-state (ISS) stability analysis. Fazlyab et al. (2019) estimates for deep networks based on incremental quadratic constraints and semidefinite programming (SDP) were shown to give state-of-the-art results, however this results was limited to analysis of an already-trained network. The SDP test incorporated into training via the alternating direction method of multipliers (ADMM) in Pauli et al. (2020), however due to the complexity of the SDP the training times recorded were almost 50 times longer than for unconstrained networks. Our approach uses a similar condition to Fazlyab et al. (2019) applied to equilibrium networks, however we introduce a novel direct parameterization method that enables learning robust models via unconstrained optimization, removing the need for computationally-expensive projections or barrier terms.

3 PROBLEM FORMULATION AND PRELIMINARIES

3.1 PROBLEM STATEMENT

We consider the weight-tied network in which $x \in \mathbb{R}^d$ denotes the input, and $z \in \mathbb{R}^n$ denotes the hidden units, $y \in \mathbb{R}^p$ denotes the output, given by the following implicit equation

$$z = \sigma(Wz + Ux + b_z), \quad y = W_o z + b_y \tag{1}$$

where $W \in \mathbb{R}^{n\times n}$, $U \in \mathbb{R}^{n\times d}$, and $W_o \in \mathbb{R}^{p\times n}$ are the hidden unit, input, and output weights, respectively, $b_z \in \mathbb{R}^n$ and $b_y \in \mathbb{R}^p$ are bias terms. The implicit framework includes most current neural network architectures (e.g. deep and residual networks) as special cases. To streamline the presentation we assume that $\sigma : \mathbb{R} \to \mathbb{R}$ is a single nonlinearity applied elementwise, although our results also apply in the case that each channel has a different activation function, nonlinear or linear.

Equation (1) is termed as an equilibrium network since its solutions are equilibrium points of the difference equation $z^{k+1} = \sigma(Wz^k + Ux + b_z)$ or the ODE $\dot{z}(t) = -z(t) + \sigma(Wz(t) + Ux + b_z)$. Our goal is to learn equilibrium networks (1) possessing the following two properties:

- **Well-posedness:** For every input x and bias b_z, equation 1 admits a unique solution z.
- **γ-Lipschitz:** It has a finite Lipschitz bound of γ, i.e., for any input-output pairs (x_1, y_1), (x_2, y_2) we have $\|y_1 - y_2\|_2 \leq \gamma\|x_1 - x_2\|_2$.

3.2 PRELIMINARIES

Monotone operator theory. The theory of monotone operators on Euclidean space (see the survey Ryu & Boyd (2016)) has been extensively applied in the development of equilibrium network (Winston & Kolter, 2020). In this paper, we will use the monotone operator theory on non-Euclidean spaces (Bauschke et al., 2011), in particular, we are interested in a finite-dimensional Hilbert space $\mathcal{H}$, which we identify with $\mathbb{R}^n$ equipped with a weighted inner product $\langle x, y\rangle_Q := y^\top Qx$ where $Q \succ 0$. The main benefit is that we can construct a more expressive equilibrium network set. A brief summary or relevant theory can be found in Appendix B; here we give some definitions that are frequently used throughout the paper. An operator is a set-valued or single-valued function defined by a subset of the space $A \subseteq \mathcal{H} \times \mathcal{H}$. A function $f : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ is proper if $f(x) < \infty$ for at least one x. The proximal operators of a proper function f is defined as

$$\mathbf{prox}_f^\alpha(x) := \{z \in \mathcal{H} \mid z = \arg\min_u \frac{1}{2}\|u - x\|_Q^2 + \alpha f(u)\},$$

where $\|x\|_Q := \sqrt{\langle x, x\rangle_Q}$ is the induced norm. For $n = 1$, we only consider the case of $Q = 1$. An operator A is monotone if $\langle u - v, x - y\rangle_Q \geq 0$ and strongly monotone with parameter m if $\langle u - v, x - y\rangle_Q \geq m\|x - y\|_Q^2$ for all $(x, u), (y, v) \in A$.

Dynamical systems theory. In this paper, we will also treat the solutions of (1) as equilibrium points of certain dynamical systems $\dot{z}(t) = f(z(t))$. Then, the well-posedness and robustness properties of (1) can be guaranteed by corresponding properties of the dynamical system's solution set. A central focus in robust and nonlinear control theory for more than 50 years – and largely unified by the modern theory of integral quadratic constraints (Megretski & Rantzer, 1997) – has been on systems which are interconnections of linear mappings and "simple" nonlinearities, i.e. those easily bounded in some sense by quadratic functions. Fortuitously, this characteristic is shared with deep, recurrent, and equilibrium neural networks, a connection that we use heavily in this paper and has previously been exploited by Fazlyab et al. (2019); El Ghaoui et al. (2019); Revay et al. (2020) and others. A particular property we are interested in is called *contraction* (Lohmiller & Slotine, 1998), i.e., any pair of solutions $z_1(t)$ and $z_2(t)$ exponentially converge to each other:

$$\|z_1(t) - z_2(t)\| \leq \alpha\|z_1(0) - z_2(0)\|e^{-\beta t}$$

for all $t > 0$ and some $\alpha, \beta > 0$. Contraction can be established by finding a Riemannian metric with respect to which nearby trajectories converge, which is a differential analog of a Lyapunov function. A nice property of a contracting dynamical system is that if it is time-invariant, a unique equilibrium exists and possess certain level of robustness. Moreover, contraction can also be linked to monotone operators, i.e. a system is contracting w.r.t. to a constant (state-independent) metric Q if and only if the operator $-f$ is strongly monotone w.r.t. Q-weighted inner product. We collect some directly relevant results from systems theory in Appendix G.

4 MAIN RESULTS

This section contains the main theoretical results of the paper: conditions implying well-posedness and Lipschitz-boundedness of equilibrium networks, and direct (unconstrained) parameterizations such that these conditions are automatically satisfied.

Assumption 1. *The activation function σ is monotone and slope-restricted in $[0, 1]$, i.e.,*

$$0 \leq \frac{\sigma(x) - \sigma(y)}{x - y} \leq 1, \ \forall x, y \in \mathbb{R}, \ x \neq y. \tag{2}$$

Remark 1. *We will show below (Proposition 1 in Section 4.2) that Assumption 1 is equivalent to the assumption on σ in Winston & Kolter (2020), i.e. that $\sigma(\cdot) = \mathbf{prox}_f^1(\cdot)$ for some proper convex function f. However, the above assumption is arguably more natural, since it is easily verified for standard activation functions. Note also that if different channels have different activation functions, then we simply require that they all satisfy (2).*

The following conditions are central to our results on well-posedness and Lipschitz bounds:

Condition 1. *There exists a $\Lambda \in \mathbb{D}^+$ such that W satisfies*

$$2\Lambda - \Lambda W - W^T\Lambda \succ 0. \tag{3}$$

Condition 2. *Given a prescribed Lipschitz bound $\gamma > 0$, there exists $\Lambda \in \mathbb{D}^+$, with $\mathbb{D}^+$ denoting diagonal positive-definite matrices, such that W, W_o, U satisfy*

$$2\Lambda - \Lambda W - W^T\Lambda - \frac{1}{\gamma}W_o^T W_o - \frac{1}{\gamma}\Lambda U U^T \Lambda \succ 0. \tag{4}$$

Remark 2. *Note that Condition 2 implies Condition 1 since $1/\gamma(W_o^T W_o + \Lambda U U^T \Lambda) \succeq 0$. As a partial converse, if Condition 1 holds, then for any W_o, U there exist a sufficiently large γ such that Condition 2 is satisfied.*

The main theoretical results of this paper are the following:

Theorem 1. *If Assumption 1 and Condition 1 hold, then the equilibrium network (1) is well-posed, i.e. for all x and b_z, equation (1) admits a unique solution z. Moreover, it has a finite Lipschitz bound from x to y.*

Theorem 2. *If Assumption 1 and Condition 2 hold, then the equilibrium network (1) is well-posed and has a Lipschitz bound of γ.*

As a consequence, we call (1) a *Lipschitz bounded equilibrium network* (LBEN) if its weights satisfy either (3) or (4). We will prove these theorems below, but first we make some straightforward remarks. As depicted in Figure 1, we can represent (1) by the algebraic feedback interconnection:

$$v = Wz + Ux + b_z, \quad z = \sigma(v), \qquad y = W_o z + b_y. \tag{5}$$

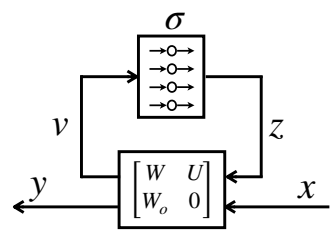


Figure 1: Equilibrium network as a feedback interconnection of a linear mapping and nonlinear activation functions.

Now, for each activation function, equation 2 can be rewritten as $(x - y)(\sigma(x) - \sigma(y)) \geq (\sigma(x) - \sigma(y))^2$. Clearly any conic (non-negative) combinations of this inequality applied to the individual activations is also true, i.e. σ satisfies the incremental sector condition

$$(v_a - v_b)^T\Lambda(z_a - z_b) \geq (z_a - z_b)^T\Lambda(z_a - z_b) \tag{6}$$

for any pair of solutions $z_a = \sigma(v_a), z_b = \sigma(v_b)$ and any $\Lambda \in \mathbb{D}^+$. Now, let $\Delta_v = v_a - v_b$ and $\Delta_z = z_a - z_b$, then the sector condition (6) can be rewritten as:

$$\langle \Delta_v - \Delta_z, \Delta_z \rangle_\Lambda \geq 0. \tag{7}$$

On the other size, the relation (3) states that pairs of solutions of (5) with the same input x satisfy

$$\langle \Delta_v - \Delta_z, \Delta_z \rangle_\Lambda \leq -\epsilon |\Delta_z|_\Lambda^2 \tag{8}$$

for some $\epsilon > 0$. From these it follows that if a solution exists to (5) then it is unique: (7) and (8) taken together imply that $\epsilon\|\Delta_z\|_\Lambda \leq 0$ where $\epsilon > 0$ and $\Lambda \in \mathbb{D}^+$, i.e. $\Delta_z = 0$. The existence of a solution will be proven via different perspectives in Sections 4.2 and 4.3.

We can also sketch a proof of Theorem 2. Since Condition 2 implies Condition 1, from Theorem 1 the solutions z exist for all input x. For any pair of inputs x_a and x_b, let (v_a, z_a, y_a) and (v_b, v_b, y_b) be the solutions to (5), respectively. Their differences satisfy $\Delta_v = W\Delta_z + U\Delta_x$ and $\Delta_y = W_o\Delta_z$. To obtain the Lipschitz bound, we first apply Schur complement to (4), then left-multiply by $\begin{bmatrix}\Delta_z^\top & \Delta_x^\top\end{bmatrix}$ and right-multiply by $\begin{bmatrix}\Delta_z^\top & \Delta_x^\top\end{bmatrix}^\top$, yielding the following:

$$\gamma\|\Delta_x\|_2^2 - \frac{1}{\gamma}\|\Delta_y\|_2^2 \geq 2\langle \Delta_v - \Delta_z, \Delta_z \rangle_\Lambda \geq 0,$$

where the inequality comes (7). It directly follows that $\|\Delta_y\|_2 \leq \gamma\|\Delta_x\|_2$ so the network has a Lipschitz bound of γ. See Appendix C for a detailed proof.

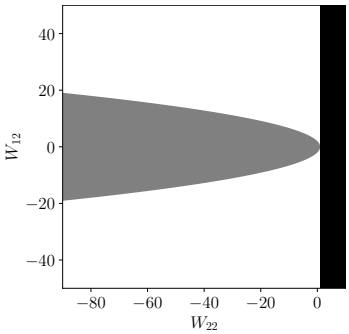


Figure 2: Valid coefficient ranges for Example 1.

Gray region: the condition from Winston & Kolter (2020) is feasible: $2I - W - W^T \succ 0$.

White region (including gray region): our well-posedness condition is feasible: $\exists \Lambda \in \mathbb{D}^+ : 2\Lambda - \Lambda W - W^T \Lambda \succ 0$.

Black region: neither condition feasible.

Remark 3. *In Fazlyab et al. (2019) it was claimed that (7) holds with a richer (more powerful) class of multipliers Λ previously introduced for robust stability analysis of systems with repeated nonlinearities, e.g. recurrent neural networks (Chu & Glover, 1999; D'Amato et al., 2001; Kulkarni & Safonov, 2002). However this is not true: a counterexample was given in Pauli et al. (2020), and here we provide a brief explanation: even if the nonlinearities $\sigma(v_i)$ are repeated when considered as functions of v_i, their increments $\Delta_{zi} = \sigma(v_i + \Delta_{vi}) - \sigma(v_i)$ are not repeated when considered as functions of Δ_{vi}, since they depend on the particular v_i which generally differs between units.*

Example 1. *We illustrate the extra flexibility of Condition 1 compared to the condition of Winston & Kolter (2020) by a toy example. Consider $W \in \mathbb{R}^{2\times 2}$ and take a slice near $W = 0$ of the form*

$$W = \begin{bmatrix} 0 & W_{12} \\ 0 & W_{22} \end{bmatrix}, \text{ for which we have: } \quad 2I - W - W^T = \begin{bmatrix} 2 & -W_{12} \\ -W_{12} & 2 - 2W_{22} \end{bmatrix}. \tag{9}$$

By Sylvester's criterion, this matrix is positive-definite if and only if $W_{22} < 1$ and $\det(2I - W - W^T) = 4(1 - W_{22}) - W_{12}^2 > 0$, which defines a parabolic region in the W_{12}, W_{22} plane.

Applying our condition (3), without loss of generality take $\Lambda = \mathrm{diag}(1, \alpha)$ with $\alpha > 0$ and we have

$$2\Lambda - \Lambda W - W^T \Lambda = \begin{bmatrix} 2 & -W_{12} \\ -W_{12} & 2\alpha - 2\alpha W_{22} \end{bmatrix}.$$

The positivity test is now $W_{22} < 1$ and $4\alpha(1 - W_{22}) - W_{12}^2 > 0$. For each W_{12} there is sufficiently large α such that the second condition is satisfied, since the first implies $1 - W_{22} > 0$. Hence the only constraint on W is that $W_{22} < 1$, which yields a much larger region in the W_{12}, W_{22} plane (see Figure 2). Interestingly, in this simple example with ReLU activation, the condition $W_{22} < 1$ is also a necessary condition for well-posedness (El Ghaoui et al., 2019, Theorem 2.8).

4.1 Direct Parameterization for Unconstrained Optimization

Training a network that satisfies Condition 1 or 2 can be formulated as a constrained optimization problem. In fact, Condition 1 is a linear matrix inequality (LMI) in the variables Λ and ΛW, from which W can be determined uniquely. Similarly, via Schur complement, Condition 2 is an LMI in the variables $\Lambda, \Lambda W, \Lambda U, W_o$, and γ, from which all network weights can be determined. In a certain theoretical sense LMI constraints are tractable – Nesterov & Nemirovskii (1994) proved they are polynomial-time solvable – however for even for moderate-scale networks (e.g. ≤ 100 activations) the associated barrier terms or projections become a major computational bottleneck, and for large-scale networks they quickly become prohibitive.

In this paper we propose direct parameterizations that allows learning via unconstrained optimization problem, i.e. all network parameters are transformations of free (unconstrained) matrix variables, in such a way that LMI constraints (3) or (4) are automatically satisfied.

For Condition (3), we parameterize via the following free variables: a matrix $V \in \mathbb{R}^{n\times n}$, a vector $d \in \mathbb{R}^n$, and skew-symmetric[1] matrix $S = -S^T \in \mathbb{R}^{n\times n}$, we can construct the hidden unit weight

$$W = I - \Psi(V^T V + \epsilon I + S), \tag{10}$$

[1]Note that S can be parameterized via its upper or lower triangular components, or via $S = N - N^T$ with N free, which can be more straightforward if W is defined implicitly via linear operators, e.g. convolutions.

where $\Psi = \mathrm{diag}\left(e^d\right)$ and $\epsilon > 0$ is some small constant to ensure strict positive-definiteness. Then it follows from straightforward manipulations that Condition 1 holds with $\Lambda = \Psi^{-1}$ if and only if W can be constructed as in (10). When $\Psi = I$, i.e. $d = 0$, this is exactly the parameterization used in Winston & Kolter (2020).

Similarly, for Condition 2, we add to the parameterization the free input and output weights U and W_o, and arbitrary $\gamma > 0$, we can construct

$$W = I - \Psi\left(\frac{1}{2\gamma}W_o^T W_o + \frac{1}{2\gamma}\Psi^{-1}UU^T\Psi^{-1} + V^T V + \epsilon I + S\right), \tag{11}$$

for which (4) is automatically satisfied. Again, it can easily be verified that this construction is necessary and sufficient, i.e. any W satisfying (4) can be constructed via (11).

4.2 Convex Optimization and Monotone Operator Perspective

In this section, we will show that the equilibrium network (1) is an optimality condition for a strongly convex optimization problem, and hence a solution exists. First, we need the following observation on the activation function σ.

Proposition 1. *Assumption 1 holds if and only if there exists a convex proper function $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ such that $\sigma(\cdot) = \mathbf{prox}_f^1(\cdot)$.*

The proof of Proposition 1 with a construction of f appears in Appendix D. Here we provide a list of f for popular σ in Table 3. It is well-known in monotone operator theory (Ryu & Boyd, 2016) that for any convex closed proper function f, the proximal operator $\mathbf{prox}_f^1(x)$ is monotone and non-expansive (i.e. slope-restricted in $[0, 1]$). Proposition 1 is a converse result for scalar functions.

Remark 4. *To our knowledge Proposition 1 is novel, however for several popular activation functions the corresponding functions f were computed in Li et al. (2019) (see also Table 3 in Appendix E). Compared with Li et al. (2019), our work gives a necessary and sufficient conditions.*

Now we connect the equilibrium network (1) to a convex optimization problem.

Proposition 2. *If Assumption 1 and Condition 1 hold, then finding a solution of the equilibrium network (1) is equivalent to solving the following strongly convex optimization problem*

$$\min_z \; J(z) = \left\langle \frac{1}{2}(I - W)z - Ux - b_z, z \right\rangle_\Lambda + \mathfrak{f}(z). \tag{12}$$

where $\mathfrak{f}(z) := \sum_{i=1}^n \lambda_i f(z_i)$ with λ_i as the ith diagonal element of Λ.

The proof appears in Appendix E and Theorem 1 follows directly since $J(z)$ is a strongly convex function on $\mathbb{R}^n$ and has a unique minimizer, which is also the solution of (1).

Computing an equilibrium. The convex optimization problem (12) has a structure that is amenable to various methods based on "splitting", since splits into a strongly-convex quadratic term and a summation of scalar convex functions with simple proximal operators. In particular, ADMM (Boyd et al., 2011) and FISTA (Beck & Teboulle, 2009), and Peaceman-Rachford splitting (Kellogg, 1969) can be directly applied. Winston & Kolter (2020) found that Peaceman-Rachford splitting converges very rapidly when properly tuned, and our experience agrees with this. As an alternative, FISTA often requires more iteration steps but does not require computing a matrix inverse or solving a linear system, which can be a significant advantage for large-scale networks.

Gradient backpropagation. As shown in (Winston & Kolter, 2020, Section 3.5), the gradients of the loss function $\ell(\cdot)$ can be represented by

$$\frac{\partial \ell}{\partial(\cdot)} = \frac{\partial \ell}{\partial z_\star}(I - JW)^{-1}J\frac{\partial(Wz_\star + Ux + b_z)}{\partial(\cdot)} \tag{13}$$

where $z_\star$ denotes the solution of (1), $(\cdot)$ denotes some learnable parameters in the parameterization (10) or (11), and $J \in \mathrm{D}\sigma(Wz_\star + Ux + b_z)$ with $\mathrm{D}\sigma$ as the Clarke generalized Jacobian of σ. Since σ is piecewise differentiable, then the set $\mathrm{D}\sigma(Wz_\star + Ux + b_z)$ is a singleton almost everywhere. The following proposition reveals that (13) is well-defined, see proof in Appendix F.

Proposition 3. *The matrix $I - JW$ is invertible for all $z_\star$, x and b_z.*

4.3 CONTRACTING NEURAL ODES

In this section, we will prove existence of a solution to (1) from a different perspective: by showing it is the equilibrium of a contracting dynamical system (a "neural ODE"). We first add a smooth state $v(t) \in \mathbb{R}^n$ to avoid the algebraic loop in (5). This idea has long been recognized as helpful for well-posedness questions (Zames, 1964). We define the dynamics of $v(t)$ by the following ODE:

$$\dot{v}(t) = -v(t) + Wz(t) + Ux + b_z, \quad z(t) = \sigma(v(t)). \tag{14}$$

The well-posedness of (1) is equivalent to the existence and uniqueness of an equilibrium of (14) for all x and b_z, which is established by the following proposition.

Proposition 4. *If Assumption 1 and Condition 1 hold, then the neural ODE (14) is contracting w.r.t. some constant metric $P \succ 0$.*

The proof is in Appendix H. Moreover, the metric P can be found via semidefinite programming. The above proposition also proves that the nonlinear operator $-f$ with $f(v) = -v + W\sigma(v) + Ux + b_z$, zeros of which define solutions of the equilibrium network (1), is actually monotone w.r.t. the P-weighted inner product, which gives a first-order cutting-plane oracle for the zero location $v_\star$ such that $f(v_\star) = 0$. I.e. given a test point $v_t \neq v^\star$, it proves that $v_\star$ is in the half-space defined by $\langle v_\star - v_t, f(v_t)\rangle_P > 0$. This may offer alternative ways to solve the equilibrium network (1), e.g. via Nemirovski (2004); Nesterov (2007).

Note also that the contraction property is independent of the input x and biases, and so extends directly to the case when these are time-varying. Roughly speaking: for any well-posed equilibrium network, there corresponds a contracting (strongly stable) neural ODE.

5 EXPERIMENTS

In this section we test our approach on the MNIST and SVHN image classification problems. Our numerical experiments focus on model robustness, the trade-off between model performance and the Lipschitz constant, and the tightness of the Lipschitz bound. We compare the the proposed LBEN to unconstrained equilibrium networks, monotone operator equilibrium network (MON) of Winston & Kolter (2020), and fully connected networks trained using Lipschitz margin training (LMT) (Tsuzuku et al., 2018). When studying model robustness to adversarial attacks, we use the L2 Fast Gradient Sign Method, implemented as part of the Foolbox toolbox (Rauber et al., 2020). All models are trained on a standard desktop computer with an NVIDIA GeForce RTX 2080 graphics card. Details of the models and training procedure can be found in Appendix J, all code will be made available online but links are omitted due to the double-blind review process.

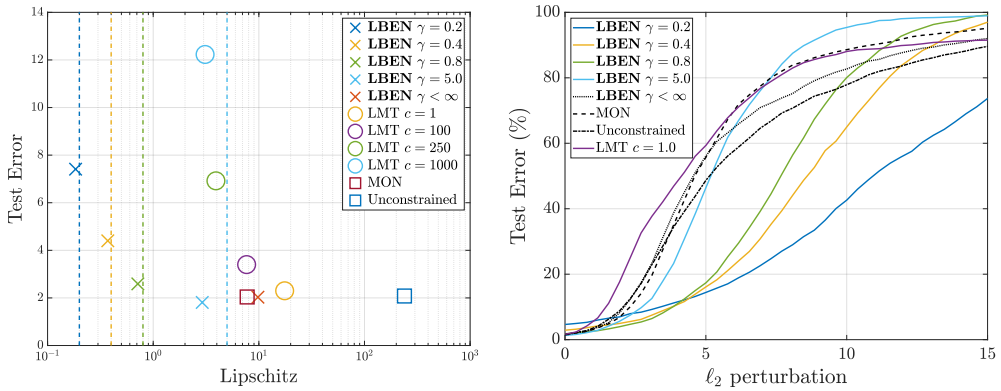
5.1 MNIST EXPERIMENTS WITH FULLY-CONNECTED NETWORKS

In Figure 3a the test error versus the observed Lipschitz constant, computed via adversarial attack for each of the models trained. We can see clearly that the parameter γ in LBEN offers a trade-off between test error and Lipschitz constant. Comparing the $\mathrm{LBEN}_{\gamma=5}$ with both MON and $\mathrm{LBEN}_{\gamma<\infty}$, we also note a slight regularizing effect in the lower test error.

By comparison, LMT (Tsuzuku et al., 2018) with c as a tunable regularization parameter displays a qualitatively similar trade-off, but underperforms LBEN in terms of both test error and robustness. If we examine the unconstrained equilibrium model, we observe a Lipschitz constant more than an order of magnitude higher, i.e. this model has regions of extremely high sensitivity, without gaining any accuracy in terms of test error.

For the LBEN models, the lower and upper bounds on the Lipschitz constant are very close: the markers are very close to their corresponding lines in Figure 3a, see also the table of numerical results in Appendix A in which the approximation accuracy is in many cases around 90%.

Next we tested robustness of classification accuracy to adversarial attacks of various sizes, the results are shown in Figure 3b and summarized in Table 1. We can clearly see that decreasing γ (i.e. stronger regularization) in the LBEN models results in a far more gradual degradation of performance as perturbation size increases, with only a mild impact on nominal (zero perturbation) test error.



(a) Nominal test error vs Lipschitz constant estimates: markers indicate observed lower bounds for all methods, vertical lines indicate certified upper bounds for LBEN

(b) Test error with adversarial perturbation versus size of adversarial perturbation. Lower is better.

Figure 3: Image classification results on MNIST character recognition data set.

Finally, we examined the impact of our parameterization on computational complexity compared to other equilibrium models. The test and training errors versus number of epochs are plotted in Figure 4, and we can see that all models converge similarly, and also take roughly the same amount of time per epoch. This is a clear contrast to the results of Pauli et al. (2020) in which imposing Lipschitz constraints resulted in fifty-fold increase in training time. Interestingly, we can also see in Figure 4 the effect of regularisation for LBEN with $\gamma = 5$: higher training error but lower test error.

It should also be noted that we have observed a number of cases where the unconstrained equilibrium model can become unstable during training as solutions are not guaranteed. LBEN never exhibits this problem.

5.2 SVHN Experiments with Convolutional Networks

The previous example looked at a simple fully connected model, however the approach can also be applied to alternate model structures. We have performed experiments with convolutional equilibrium networks on the SVHN dataset, comparing convolutional LBEN with a convolutional MON.

Table 2 in Appendix A shows that for a slight decrease in nominal test performance we can reduce the Lipschitz sensitivity to adversarial attack by more than a factor of 10, and significantly increase robustness of classification performance. Note that the Lipschitz bound for this model is not as tight as the one observed in the MNIST example, perhaps because we have used a slightly more restrictive set of multipliers (c.f. Section J.2 for details). Further exploration of larger and more richly-structured convolutional networks is a topic of our on-going research.

6 Conclusions

In this paper we have shown that the flexible framework of equilibrium networks can be made robust via a simple and direct parameterization which results in guaranteed Lipschitz bounds. Although we have not explored it in detail in this paper, our results can also be directly applied (as a special case) to standard multilayer and residual deep neural networks, and also provide a direct parameterization of nonlinear ODEs satisfying strong stability and robustness properties. Furthermore, although in this paper we have limited attention to standard scalar activation functions such as ReLU or sigmoids, our results easily extend to certain multivariable "activations" that satisfy appropriate monotonicity properties, or more generally integral quadratic constraints. This includes, for example, computing the $\arg\min$ of a quadratic program of the sort that appears in constrained model predictive control (Heath & Wills, 2007). Exploring these variations will be a topic of our future research.

REFERENCES

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, pp. 690–701, 2019.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pp. 6240–6249, 2017.

Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, pp. 6571–6583, 2018.

Yun-Chung Chu and Keith Glover. Bounds of the induced norm and model reduction errors for systems with repeated scalar nonlinearities. *IEEE Transactions on Automatic Control*, 44(3): 471–483, 1999.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320, 2019.

Fernando J D'Amato, Mario A Rotea, AV Megretski, and UT Jönsson. New results for analysis of systems with repeated nonlinearities. *Automatica*, 37(5):739–747, 2001.

Laurent El Ghaoui, Fangda Gu, Bertrand Travacca, Armin Askari, and Alicia Y. Tsai. Implicit deep learning. *arXiv:1908.06315*, 2019.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 11427–11438, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

WP Heath and AG Wills. Zames-falb multipliers for quadratic programming. *IEEE Transactions on Automatic Control*, 10(52):1948–1951, 2007.

Joao P Hespanha. *Linear Systems Theory*. Princeton university press, 2018.

R Bruce Kellogg. A nonlinear alternating direction method. *Mathematics of Computation*, 23(105): 23–27, 1969.

Hassan K Khalil. *Nonlinear systems*. Prentice-Hall, 2002.

Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic gradient descent. In *ICLR: International Conference on Learning Representations*, 2015.

Vishwesh V Kulkarni and Michael G Safonov. All multipliers for repeated monotone nonlinearities. *IEEE Transactions on Automatic Control*, 47(7):1209–1212, 2002.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Jia Li, Cong Fang, and Zhouchen Lin. Lifted proximal operator machines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4181–4188, 2019.

Changliu Liu, Tomer Arnon, Christopher Lazarus, Clark Barrett, and Mykel J Kochenderfer. Algorithms for verifying deep neural networks. *arXiv preprint arXiv:1903.06758*, 2019.

Winfried Lohmiller and Jean-Jacques E Slotine. On contraction analysis for non-linear systems. *Automatica*, 34(6):683–696, 1998.

Alexandre Megretski and Anders Rantzer. System analysis via integral quadratic constraints. *IEEE Transactions on Automatic Control*, 42(6):819–830, 1997.

Arkadi Nemirovski. Prox-Method with Rate of Convergence O(1/t) for Variational Inequalities with Lipschitz Continuous Monotone Operators and Smooth Convex-Concave Saddle Point Problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

Patricia Pauli, Anne Koch, Julian Berberich, and Frank Allgöwer. Training robust neural networks using Lipschitz bounds. *arXiv preprint arXiv:2005.02929*, 2020.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018.

Anders Rantzer. On the Kalman-Yakubovich-Popov lemma. *Systems & Control Letters*, 28(1):7–10, 1996.

Jonas Rauber, Roland Zimmermann, Matthias Bethge, and Wieland Brendel. Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and jax. *Journal of Open Source Software*, 5(53):2607, 2020.

Max Revay, Ruigang Wang, and Ian R Manchester. Convex sets of robust recurrent neural networks. *arXiv preprint arXiv:2004.05290*, 2020.

Ernest K Ryu and Stephen Boyd. Primer on monotone operator methods. *Appl. Comput. Math*, 15 (1):3–43, 2016.

Hanie Sedghi, Vineet Gupta, and Philip M Long. The singular values of convolutional layers. In *International Conference on Learning Representations*, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR: International Conference on Learning Representations*, 2014.

Vincent Tjeng, Kai Y Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2018.

Mark M Tobenkin, Ian R Manchester, and Alexandre Megretski. Convex parameterizations and fidelity bounds for nonlinear identification and reduced-order modelling. *IEEE Transactions on Automatic Control*, 62(7):3679–3686, 2017.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in neural information processing systems*, pp. 6541–6550, 2018.

Ezra Winston and J. Zico Kolter. Monotone operator equilibrium networks. *arXiv:2006.08591*, 2020.

G. Zames. Realizability Condition for Nonlinear Feedback Systems. *IEEE Transactions on Circuit Theory*, 11(2):186–194, 1964.

SiQi Zhou and Angela P Schoellig. An analysis of the expressiveness of deep neural network architectures based on their lipschitz constants. *arXiv preprint arXiv:1912.11511*, 2019.

A EXPERIMENTAL RESULTS ON MNIST CHARACTER RECOGNITION

This appendix contains tables of results on MNIST and SVHN data sets.

Legend:

- Err: Test error (%),
- $\|a\|_2$: ℓ^2 norm of adversarial attack.
- γ_{up}: certified upper bound on Lipschitz constant (for models that provide one).
- γ_{low}: observed lower bound on Lipschitz constant via adversarial attack.
- γ approx: approximation ratio of Lipschitz constant as percentage = $100 \times \left(\frac{\gamma_{low}}{\gamma_{up}}\right)$.

Models:

- LBEN: the proposed Lipschitz bounded equilibrium network..
- MON: the monotone operator equilibrium network of Winston & Kolter (2020).
- UNC: an unconstrained equilibrium network, i.e. W directly parameterized.
- LMT: Lipschitz Margin Training model as in Tsuzuku et al. (2018).
- Lip-NN: The Lipschitz Neural Network model of Pauli et al. (2020). Note these figures are as reported in (Pauli et al., 2020), all other figures are calculated by the authors of the present paper.

Model	Err: $\|a\|_2 = 0$	Err: $\|a\|_2 \leq 5$	Err: $\|a\|_2 \leq 10$	γ_{up}	γ_{low}	γ approx
$\text{LBEN}_{\gamma<\infty}$	2.03	56.0	82	-	9.8	-
$\text{LBEN}_{\gamma=5}$	**1.81**	46.4	95.4	5	2.912	58.2%
$\text{LBEN}_{\gamma=1}$	2.36	19.4	85.5	1	0.865	86.5%
$\text{LBEN}_{\gamma=0.8}$	2.59	17.4	80.1	0.8	0.715	89.4%
$\text{LBEN}_{\gamma=0.4}$	4.44	16.1	65.0	0.4	0.372	93%
$\text{LBEN}_{\gamma=0.2}$	7.41	**14.4**	**42.6**	0.2	0.184	92%
MON	2.04	55.8	88.6	-	7.75	-
UNC	2.08	48.75	77.9	-	239.0	-
$\text{LMT}_{c=1}$	2.3	59.4	88.1	-	17.5	-
$\text{LMT}_{c=100}$	3.4	65.4	92.0	-	7.66	-
$\text{LMT}_{c=250}$	6.92	61.8	98.4	-	6.92	-
$\text{LMT}_{c=1000}$	12.23	78.4	98.9	-	3.10	-
Lip-NN	3.55	-	-	8.74	-	-

Table 1: Results from MNIST experiments.

Model	Err: $\|a\|_2 = 0$	Err: $\|a\|_2 \leq 5$	Err: $\|a\|_2 \leq 10$	γ_{up}	γ_{low}	γ approx
MON	19.5	40	72	-	8.3	-
$\text{LBEN}_{\gamma<2}$	22.75	37	56	2	0.8	40%

Table 2: Performance of convolutional LBEN versus convolutional MON on SVHN dataset.

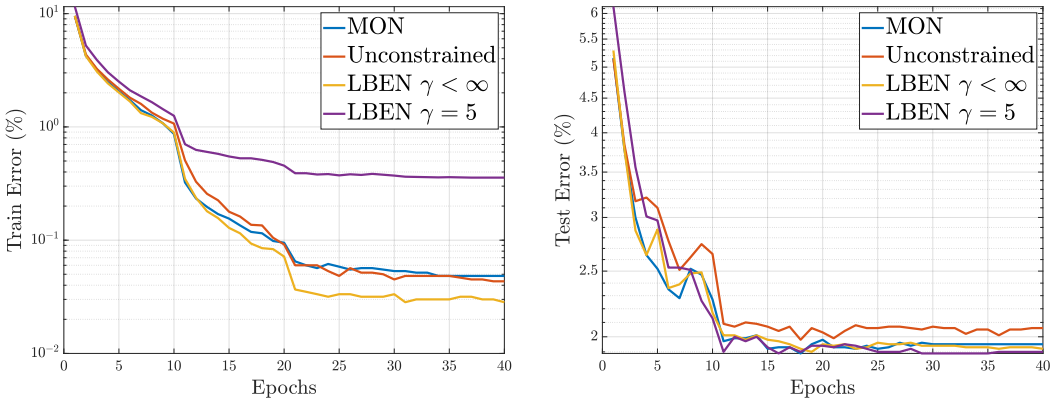


Figure 4: **Left:** Training set error versus epochs. **Right:** Test set error versus epochs. Note that the left and right plots are on different scales. The time per epoch for the MON, unconstrained, $\text{LBEN}_{\gamma<\infty}$ and $\text{LBEN}_{\gamma=5}$ networks are 14.4, 16.1, 14.9 and 14.8 seconds per epoch respectively.

B Monotone Operators with Non-Euclidean Inner Products

We present some basic properties of monotone operators on a finite-dimensional Hilbert space $\mathcal{H}$, which we identify with $\mathbb{R}^n$ equipped with a weighted inner product $\langle x, y\rangle_Q = y^\top Qx$ with $Q \succ 0$. For $n = 1$, we only consider the case of $Q = 1$. The induced norm $\|x\|_Q$ is defined as $\sqrt{\langle x, x\rangle_Q}$. A *relation* or *operator* is a set-valued or single-valued map defined by a subset of the space $A \subseteq \mathcal{H}\times\mathcal{H}$; we use the notation $A(x) = \{y \mid (x, y) \in A\}$. If $A(x)$ is a singleton, we called A a function. Some commonly used operators include: the linear operator $A(x) = \{(x, Ax) \mid x \in \mathcal{H}\}$; the operator sum $A + B = \{(x, y + z) \mid (x, y) \in A, (x, z) \in B\}$; the inverse operator $A^{-1} = \{(y, x) \mid (x, y) \in A\}$; and the subdifferential operator $\partial f = \{(x, \partial f(x))\}$ with $x = \mathbf{dom}\, f$ and $\partial f(x) = \{g \in \mathcal{H} \mid f(y) \geq f(x) + \langle y - x, g\rangle_Q, \forall y \in \mathcal{H}\}$. An operator A has Lipschitz constant L if for any $(x, u), (y, v) \in A$

$$\|u - v\|_Q \leq L\|x - y\|_Q. \tag{15}$$

An operator A is non-expansive if $L = 1$ and contractive if $L < 1$. An operator A is monotone if

$$\langle u - v, x - y\rangle_Q \geq 0, \ \forall (x, u), (y, v) \in A. \tag{16}$$

It is strongly monotone with parameter m if

$$\langle u - v, x - y\rangle_Q \geq m\|x - y\|_Q^2, \ \forall (x, u), (y, v) \in A. \tag{17}$$

A monotone operator A is maximal monotone if no other monotone operator strictly contains it, which is a property required for the convergence of most fixed point iterations. Specifically, an affine operator $A(x) = Wx + b$ is (maximal) monotone if and only if $QW + W^\top Q \succeq 0$ and strongly monotone if $QW + W^\top Q \succeq mI$. A subdifferential ∂f is maximal monotone if and only if f is a convex closed proper function.

The resolvent and Cayley operators for an operator A are denoted R_A and C_A and respectively defined as

$$R_A = (I + \alpha A)^{-1}, \quad C_A = 2R_A - I \tag{18}$$

for any $\alpha > 0$. When $A(x) = Wx + b$, then

$$R_A(x) = (I + \alpha W)^{-1}(x - \alpha b) \tag{19}$$

and when $A = \partial f$ for some CCP function f, then the resolvent is given by a proximal operator

$$R_A(x) = \mathbf{prox}_f^\alpha(x) := \arg\min_z \frac{1}{2}\|x - z\|_Q^2 + \alpha f(z). \tag{20}$$

The resolvent and Cayley operators are non-expansive for any maximal monotone A, and are contractive for strongly monotone A. Operator splitting methods consider finding a zero in a sum of operators (assumed here to be maximal monotone), i.e., find z such that $0 \in (A + B)(z)$. For example, the convex optimization problem in (12) can be formulated as an operator splitting problem

with $A(z) = (I - W)z - b$ and $B = \partial \mathfrak{f}$. Proposition 2 shows that A is strongly monotone and Lipschitz with some parameters of m and L. Here we give some popular operator splitting methods for this problem as follows.

- Forward-backward splitting: $z^{k+1} = R_B(z^k - \alpha A(z^k))$, i.e.,
$$\begin{aligned} u^k &= ((1-\alpha)I + \alpha W)z^k + \alpha b \\ z^{k+1} &= \mathbf{prox}_{\mathfrak{f}}^{\alpha}(u^k) \end{aligned} \tag{21}$$
- Peaceman-Rachford splitting: $u^{k+1} = C_A C_B(u^k)$, $z^k = R_B(u^k)$, i.e.,
$$\begin{aligned} u^{k+1/2} &= 2z^k - u^k, \\ z^{k+1/2} &= (I + \alpha(I-W))^{-1}(u^{k+1/2} + \alpha b), \\ u^{k+1} &= 2x^{k+1/2} - u^{k+1/2}, \\ z^{k+1} &= \mathbf{prox}_{\mathfrak{f}}^{\alpha}(u^{k+1}). \end{aligned} \tag{22}$$
- Douglas-Rachford splitting (or ADMM): $u^{k+1} = 1/2(I + C_A C_B)(u^k)$, $z^k = R_B(u^k)$, i.e.,
$$\begin{aligned} u^{k+1/2} &= 2z^k - u^k, \\ z^{k+1/2} &= (I + \alpha(I-W))^{-1}(u^{k+1/2} + \alpha b), \\ u^{k+1} &= 2x^{k+1/2} - u^{k+1/2}, \\ z^{k+1} &= \mathbf{prox}_{\mathfrak{f}}^{\alpha}(u^{k+1}). \end{aligned} \tag{23}$$
- Fast iterative shrinkage-thresholding algorithm (FISTA):
$$\begin{aligned} u^k &= \arg\min_u \mathfrak{f}(u) + \frac{L}{2}\left\| u - \frac{1}{L}\left[(L-1)z^k + Wz^k + b\right]\right\|_2^2 \\ t^{k+1} &= \frac{1+\sqrt{1+4(t^k)^2}}{2}, \\ z^{k+1} &= u^k + \left(\frac{t^k - 1}{t^{k+1}}\right)(u^k - u^{k-1}). \end{aligned} \tag{24}$$

A sufficient condition for forward-backward splitting to converge is $\alpha < 2m/L^2$. The Peacemance-Rachford and Douglas-Rachford methods converge for any $\alpha > 0$, although the convergence speed will often vary substantially based upon α. The FISTA method converges and it does not have any hyper-parameter. When the weighting W is updated, Peacemance-Rachford and Douglas-Rachford splitting need to compute a matrix inverse $(I + \alpha(I - W))^{-1}$ while FISTA requires to compute the maximum singular value of $I - W$.

C Proof of Theorem 2

Rearranging Eq. (4) yields
$$2\Lambda - \Lambda W - W^T\Lambda \succ \frac{1}{\gamma}(W_o^T W_o + \Lambda U U^T \Lambda) \succeq 0.$$
The well-posedness of the equilibrium network (1) follows by Theorem 1. To obtain the Lipschitz bound, we first apply Schur complement to (4):
$$\begin{bmatrix} 2\Lambda - \Lambda W - W^\top \Lambda - \frac{1}{\gamma} W_o^\top W_o & -\Lambda U \\ -U^\top \Lambda & \gamma I \end{bmatrix} \succ 0.$$
Left-multiplying $\begin{bmatrix}\Delta_z^\top & \Delta_x^\top\end{bmatrix}$ and right-multiplying $\begin{bmatrix}\Delta_z^\top & \Delta_x^\top\end{bmatrix}^\top$ gives
$$2\Delta_z^\top \Lambda \Delta_z - 2\Delta_z^\top \Lambda W \Delta_z - \frac{1}{\gamma}\Delta_z^\top W_o^\top W_o \Delta_z - 2\Delta_z^\top \Lambda U \Delta_x + \gamma \|\Delta_x\|_2^2 \geq 0.$$
Since (5) implies $\Delta_v = W\Delta_z + U\Delta_x$ and $\Delta_y = W_o \Delta_z$, the above inequality is equivalent to
$$\gamma\|\Delta_x\|_2^2 - \frac{1}{\gamma}\|\Delta_y\|_2^2 \geq 2\Delta_z^\top \Lambda \Delta_z - 2\Delta_z \Lambda \Delta_v = 2\langle \Delta_v - \Delta_z, \Delta_z\rangle_\Lambda.$$
Then, the Lipschitz bound of γ for the equilibrium network (1) follows by (7).

Activation	$\sigma(x)$	Convex $f(z)$	$\mathbf{dom}\, f$
ReLu	$\max(x, 0)$	0	$[0, \infty)$
LeakyReLu	$\max(x, 0.01x)$	$\frac{99}{2}\min(z, 0)^2$	$\mathbb{R}$
Tanh	$\tanh(x)$	$\frac{1}{2}\left[\ln(1-z^2) + z\ln\left(\frac{1+z}{1-z}\right) - z^2\right]$	$(-1, 1)$
Sigmoid	$1/(1+e^{-x})$	$z\ln z + (1-z)\ln(1-z) - \frac{z^2}{2}$	$(0, 1)$
Arctan	$\arctan(x)$	$-\ln(\lvert\cos z\rvert) - \frac{z^2}{2}$	$(-1, 1)$
Softplus	$\ln(1+e^x)$	$-\mathrm{Li}_2(e^z) - i\pi z - z^2/2$	$(0, \infty)$

Table 3: A list of common activation functions $\sigma(x)$ and associated convex proper $f(z)$ whose proximal operator is $\sigma(x)$. For $z \notin \mathbf{dom}\, f$, we have $f(z) = \infty$. In the case of Softplus activation, $\mathrm{Li}_s(z)$ is the polylogarithm function.

D Proof of Proposition 1

(*if*): It is well-known that if f is convex closed proper function, then $\mathbf{prox}_f^1$ is monotone and non-expansive, i.e., it is slope-restricted in $[0, 1]$. Here f is not necessary to be closed as $\mathbf{dom}\, f$ (i.e. the range of σ) could be open interval (z_l, z_r) or half-open interval $(z_l, z_r]$ or $[z_l, z_r)$. This can be resolved by defining $\hat{f}$ as the restriction of f on the closed interval $[\hat{z}_l, \hat{z}_r]$, and then make $\hat{z}_l \to z_l$ and $\hat{z}_r \to z_r$.

(*only if*): Assumption 1 implies that σ is a non-decreasing and piece-wise differentiable function on $\mathbb{R}$. Then, the range of σ is an interval, denoted by $\mathcal{Z}$. We will construct the derivative function f' on $\mathcal{Z}$ first and then integrate it to obtain f. Let $\{z_j \in \mathcal{Z}\}_{j\in\mathbb{Z}}$ be the sequence containing all points such that either $\sigma'(x_-) = 0$ or $\sigma'(x_+) = 0$ for all $x \in \sigma^{-1}(z_j)$. Note that $\sigma^{-1}(z)$ is a singleton for all $z \in (z_j, z_{j+1})$, whereas $\sigma^{-1}(z_j)$ is a closed interval of the forms $(-\infty, x_r]$, $[x_l, x_r]$ or $[x_l, \infty)$. Then, we define f' as follows

$$f'(z) = \begin{cases} \min[\sigma^{-1}(z)] - z, & \text{if } z = z_j \text{ and } \min\sigma^{-1}(z) > -\infty, \\ \max[\sigma^{-1}(z)] - z, & \text{if } z = z_j \text{ and } \min\sigma^{-1}(z) = -\infty, \\ \sigma^{-1}(z) - z, & \text{otherwise.} \end{cases}$$

Without loss of generality, we assume that $0 \in \mathcal{Z}$ and $\sigma^{-1}(0)$ is well-defined. We define the function f as follows

$$f(z) = \begin{cases} \int_0^z f'(\zeta)d\zeta + C & \text{if } z \in \mathcal{Z}, \\ \infty & \text{otherwise,} \end{cases}$$

where C is an arbitrary constant. Note that f is a convex function as f' is a piecewise differentiable function on $\mathcal{Z}$ and for those points where $x = \sigma^{-1}(z)$ is well-defined, f' is differentiable with $f''(z) = 1/\sigma'(x) - 1 \geq 0$ as $\sigma'(x) \in (0, 1]$. Finally, the definition of f' implies that $0 \in z - \sigma^{-1}(z) + \partial f(z)$, which implies that $z = \sigma(x)$ is the unique minimizer of $1/2(z-x)^2 + f(z)$. Furthermore, since σ is well-defined, we can conclude that f is bounded from below. We also provide a list of f for common activation functions in Table 3. A similar list can also be found in Li et al. (2019).

E Proof of Proposition 2

The problem (12) can be formulated as a operator splitting problem $0 \in (A + B)(z)$ where $A(z) = (I - W)(z) - (Ux + b_z)$ and $B = \partial\mathfrak{f}$. The cost function $J(z)$ in (12) is strongly convex as

A is strictly monotone by Condition 1 and $\mathfrak{f}$ is convex. Similar to Winston & Kolter (2020), we prove Proposition 2 by showing that the solution of (1), if it exists, is an fixed point of the forward-backward iteration (21) with $\alpha = 1$:

$$z^{k+1} = R_B(z^k - \alpha A z^k) = \mathbf{prox}_{\mathfrak{f}}^1(z^k - \alpha(I-W)z^k + \alpha(Ux + b_z)) = \sigma(Wz^k + Ux + b_z).$$

The last equality follows by

$$\sigma(x) = \begin{bmatrix} \arg\min_{z_1} \frac{1}{2}(z_1 - x_1)^2 + f(z_1) \\ \vdots \\ \arg\min_{z_n} \frac{1}{2}(z_n - x_n)^2 + f(z_n) \end{bmatrix} = \arg\min_z \frac{1}{2}\|z - x\|_\Lambda^2 + \sum_{i=1}^n \lambda_i f(z_i) = \mathbf{prox}_{\mathfrak{f}}^1(x).$$

Note that the necessary condition for $\sigma(\cdot)$ to be diagonal is that the weight matrix Λ is positive diagonal.

F Proof of Proposition 3

The matrix J is diagonal with elements in $[0, 1]$. Decompose $\Lambda = \Pi(J + \mu I)$ for some small $\mu > 0$, i.e. $\Pi = \Lambda(J + \mu I)^{-1}$, which is diagonal and positive-definite. By denoting $H = \Pi(I - W) + (I - W)^T \Pi$ we obtain the following inequality from (3):

$$\Pi J(I - W) + (I - W)^T J \Pi + \mu H \succeq \epsilon I,$$

which can be rearranged as

$$\Pi(I - JW) + (I - JW)^T \Pi \succeq \epsilon I + 2\Pi(I - J) - \mu H.$$

Since $2\Pi(I - J) \succeq 0$, we can choose a sufficiently small μ such that

$$\Pi(I - JW) + (I - JW)^T \Pi \succ 0,$$

which further implies that $I - JW$ is strongly monotone w.r.t. Π-weighted inner product, and is therefore invertible.

G Dynamical System Theory

In this section, we present some concepts and results of dynamical system theory that are used in this paper. We consider a nonlinear system of the form

$$\dot{z}(t) = f(z(t)) \tag{25}$$

where $z(t) \in \mathbb{R}^n$ is the state, and the function f is assumed to be Lipschitz continuous. By Picard's existence theorem we have a unique a solution for any initial condition. The above system is time-invariant since f is not explicitly depends on t. System (25) is called linear time-invariant (LTI) system if $f(z) = Az + b$ for some matrix $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$. The point $z_\star \in \mathbb{R}^n$ is call an equilibrium of (25) if $f(z_\star) = 0$.

The central concern in dynamical system theory is *stability*. While there are many different stability notions (Khalil, 2002), here we mainly focus on two of them: exponential stability and contraction w.r.t a constant metric $Q \succ 0$. System (25) is said to be locally exponentially stable at the equilibrium $z_\star$ w.r.t. to the metric Q if there exist some positive constants α, β, δ such that for any initial condition $z(0) \in \mathcal{B}_\delta(z_\star) := \{z \mid \|z - z_\star\|_Q < \delta\}$, the following condition holds:

$$\|z(t) - z_\star\| \leq \alpha \|z(0) - z_\star\|_Q e^{-\beta t}, \quad \forall t > 0. \tag{26}$$

And it is said to be globally exponentially stable if the above condition also holds for any $\delta > 0$. The exponentially stability can be verified via Lyapunov's second method, i.e., finding a Lyapunov function $V = \|z\|_P^2$ with $P \succ 0$ such that $\dot{V}(t) \leq -2\beta V(t)$ along the solutions, i.e.,

$$(z - z_\star)^\top P f(z) + f(z)^\top P(z - z_\star) + 2\beta(z - z_\star)^\top P(z - z_\star) \leq 0. \tag{27}$$

System (25) is said to be contracting w.r.t. the metric Q if there exist some positive constants α, β such that for any pair of solutions $z_1(t)$ and $z_2(t)$, we have

$$\|z_1(t) - z_2(t)\|_Q \leq \alpha \|z_1(0) - z_2(0)\|_Q e^{-\beta t}, \quad \forall t > 0. \tag{28}$$

Note that contraction is a much stronger notion than global exponential stability as Condition (26) can be implied by Condition (28) by setting $z_1 = z$ and $z_2 = z_\star$. However, unlike the Lyapunov analysis, contraction analysis can be done via simple local analysis which does not require any prior-knowledge about the equilibrium $z_\star$. Specifically, contraction can be established by the local exponential stability of the associated differential system defined by

$$\dot{\Delta}_z = \mathrm{D}f(z)\Delta_z$$

where $\Delta_z(t)$ is the infinitesimal variation between $z(t)$ and its neighborhood solutions, and $\mathrm{D}f$ is Clarke generalized Jacobian. The condition for (25) to be contracting can be represented as a state-dependent Linear Matrix Inequality (LMI) as follows

$$P\mathrm{D}f(z) + \mathrm{D}f(z)^\top P + 2\beta P \prec 0 \tag{29}$$

for some $P \succ 0$ and all $z \in \mathbb{R}^n$. For an LTI system, exponential stability and contraction are equivalent and the stability condition can be s if A is Hurwitz stable (i.e. all eigenvalues of A have strictly negative real part).

For most applications, the dynamic system usually involves an external input $x(t) \in \mathbb{R}^m$ and an output $y(t) \in \mathbb{R}^p$, whose state-space representation takes the form of

$$\dot{z}(t) = f(z(t), x(t)), \quad y(t) = h(z(t), x(t)). \tag{30}$$

Here we measure the robustness of the above system under input perturbation by incremental L_2-gain. That is, system (30) has an incremental L_2-gain bound of γ if for any pair of inputs $x_1(\cdot), x_2(\cdot)$ with $\int_0^T \|x_1(t) - x_2(t)\|_2^2 dt < \infty$ for all $T > 0$, and any initial conditions $z_1(0)$ and $z_2(0)$, the solutions of (30) exists and satisfy

$$\int_0^T \|y_1(t) - y_2(t)\|_2^2 \, dt \leq \gamma^2 \int_0^T \|x_1(t) - x_2(t)\|_2^2 \, dt + \kappa(z_1(0), z_(0)) \tag{31}$$

for some function $\kappa(z_1, z_2) \geq 0$ with $\kappa(z, z) = 0$. Note that γ can be viewed as a Lipschitz bound of all the mappings defined by (30) with some initial condition from the input signal $x(\cdot)$ to $y(\cdot)$. For any two constant inputs x_1, x_2, let z_1, z_2 and y_1, y_2 be the corresponding equilibrium and steady-state output, respectively. From (31) we have

$$\|y_1 - y_2\|_2^2 \leq \|x_1 - x_2\|_2^2 + \kappa(z_1, z_2)/T,$$

which implies a Lipschitz bound of γ as $T \to \infty$.

A particular class of nonlinear systems that have strong connections to various neural networks is the so-called Luré system, which takes the form of

$$\dot{z}(t) = Az(t) + B\phi(Cz(t)) \tag{32}$$

where A, B, C are constant matrices with proper size, and ϕ is a static nonlinearity with sector bounded of $[\alpha, \beta]$: for all solution (v, w) with $w = \phi(v)$

$$(w - \alpha v)^\top (\beta v - w) \geq 0 \tag{33}$$

or equivalently $\begin{bmatrix} v \\ w \end{bmatrix}^\top \Pi \begin{bmatrix} v \\ w \end{bmatrix} \geq 0$ with

$$\Pi = \begin{bmatrix} 2\alpha\beta I & (\alpha+\beta)I \\ (\alpha+\beta)I & -2I \end{bmatrix}. \tag{34}$$

This implies that the origin is an equilibrium since $\phi(0) = 0$. The above system can be viewed as a feedback interconnection of a linear system

$$G: \begin{cases} \dot{z}(t) = Az(t) + Bw(t) \\ v(t) = Cz(t) \end{cases} \tag{35}$$

and a nonlinear memoryless component $w(t) = \phi(v(t))$. The above linear system can also be described by a transfer function $G(s)$ with $s \in \mathbb{C}$. We refer to Hespanha (2018) for details about

frequency-domain concepts and results of linear systems. The frequency-domain representation for the sector bounded condition (33) can be written as

$$\begin{bmatrix} \hat{v}(j\omega) \\ \hat{w}(j\omega) \end{bmatrix}^* \Pi \begin{bmatrix} \hat{v}(j\omega) \\ \hat{w}(j\omega) \end{bmatrix} \geq 0 \quad \forall \omega \in \mathbb{R} \tag{36}$$

where $\hat{v}(j\omega)$ and $\hat{w}(j\omega)$ are Fourier transforms of v and w, respectively, $(\cdot)^*$ denotes the complex conjugate. Then, the closed-loop stability of the feedback interconnection can be verified by the Integral Quadratic Constraint (IQC) theorem (Megretski & Rantzer, 1997). Although the IQC framework allows for more general dynamic multipliers, here we only focus on the simple constant multiplier defined in (34).

Theorem 3. *Let G be stable and ϕ be a static nonlinearity with sector bound of $[\alpha, \beta]$. The feedback interconnection of G and ϕ is stable if here exists $\epsilon > 0$ such that*

$$\begin{bmatrix} G(j\omega) \\ I \end{bmatrix}^* \Pi \begin{bmatrix} G(j\omega) \\ I \end{bmatrix} \preceq -\epsilon I, \quad \forall \omega \in \mathbb{R}. \tag{37}$$

The Kalman-Yakubovich-Popov (KYP) lemma (Rantzer, 1996) can be applied to demonstrate the equivalence of Condition 3 in Theorem 3 to an LMI condition. The result is stated as follows.

Theorem 4. *There exists a $\epsilon > 0$ such that (37) holds if and only if there exists a matrix $P = P^\top$ such that*

$$\begin{bmatrix} A^\top P + PA & PB \\ B^\top P & 0 \end{bmatrix} + \begin{bmatrix} C^\top & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} \prec 0.$$

H Proof of Proposition 4

From (14) the dynamics of Δ_v and Δ_z can be formulated as a feedback interconnection of a linear system $\dot{\Delta}_v = -\Delta_v + W\Delta_z$ and a static nonlinearity $\Delta_z = \sigma(v_a) - \sigma(v_b)$. The linear system can be represented by a transfer function is $G(s) = 1/(s+1)W$. The nonlinear component can be rewritten as $\Delta_z = \Phi(v_a, v_b)\Delta_v$ where Φ as a diagonal matrix with each $\Phi_{ii} \in [0, 1]$. For the nonlinear component Φ, its input and output signals satisfies the quadratic constraint (7). For the linear system G, we have the following lemma.

Lemma 1. *If Condition 1 holds, then for all $\omega \in \{\mathbb{R} \cup \infty\}$*

$$\begin{bmatrix} G(j\omega) \\ I \end{bmatrix}^* \begin{bmatrix} 0 & \Lambda \\ \Lambda & -2\Lambda \end{bmatrix} \begin{bmatrix} G(j\omega) \\ I \end{bmatrix} \prec 0. \tag{38}$$

The KYP Lemma (Theorem 4) states that (38) is equivalent to the existence of a $P = P^\top$ such that

$$\begin{bmatrix} -2P & PW \\ W^T P & 0 \end{bmatrix} + \begin{bmatrix} 0 & \Lambda \\ \Lambda & -2\Lambda \end{bmatrix} \prec 0.$$

It is clear from the upper-left block that $P \succ 0$. The above inequality also implies

$$2\langle -\Delta_v + W\Delta_z, \Delta_v \rangle_P \leq \langle \Delta_z - \Delta_v, \Delta_z \rangle_\Lambda - \epsilon(\|\Delta_z\|_2^2 + \|\Delta_v\|_2^2) \leq -\epsilon(\|\Delta_z\|_2^2 + \|\Delta_v\|_2^2)$$

for some $\epsilon > 0$. The contraction property of the neural ODE (14 follows since

$$\frac{d}{dt}\|\Delta_v\|_P^2 = 2\langle -\Delta_v + W\Delta_z, \Delta_v \rangle_P \leq -\epsilon(\|\Delta_z\|_2^2 + \|\Delta_v\|_2^2) \leq -2\beta\|\Delta_v\|_P^2$$

for some sufficiently small $\beta > 0$. As a byproduct of the above inequality, we will show that the operator $-f$ with with $f(v) = -v + W\sigma(v) + Ux + b_z$ is strictly monotone w.r.t. the P-weighted inner product since

$$\langle -f(v_a) + f(v_b), v_a - v_b \rangle_P = \langle \Delta_v - W\Delta_z, \Delta_v \rangle_P \geq \beta\|\Delta_v\|_P^2.$$

I PROOF OF LEMMA 1

Note that (38) is equivalent to

$$2\Lambda - G_0(j\omega)\Lambda W - G_0(-j\omega)W^T\Lambda \succeq \mu I \tag{39}$$

where $G_0(j\omega) = \frac{1}{1+j\omega}$. For some $\omega \in (\mathbb{R} \cup \infty)$ let $g = \Re G_0(j\omega) = \Re G_0(-j\omega)$, where $\Re$ denotes real part. It is easy to verify that $g = 1/(\omega^2 + 1) \in [0, 1]$. From (3) we have

$$2g\Lambda - g\Lambda W - gW^T\Lambda \succeq g\epsilon I$$

for some $\epsilon > 0$. Rearranging the above inequality yields

$$2\Lambda - g\Lambda W - gW^T\Lambda \succeq g\epsilon I + (1 - g)2\Lambda$$

Now, since $g \in [0, 1]$ the right-hand-side is a convex combination of two positive definite matrices: ϵI and 2Λ, therefore (39) holds for some $\mu > 0$ and all $\omega \in (\mathbb{R} \cup \infty)$.

J TRAINING DETAILS

J.1 MNIST EXAMPLE

This section contains the model structures and the details of the training procedure used for the MNIST examples. All models are trained using the ADAM optimizer Kingma & Ba (2015) with an initial learning rate of 1×10^3. All models are trained for 40 Epochs, and the learning rate is reduced by a factor of 10 every 10 epochs.

The models in the MNIST example are all fully connected models with 80 hidden neurons and ReLU activations. For the equilibrium models, the forward and backward passes models are performed using the Peaceman-Rachford iteration scheme with $\epsilon = 1$ and a tolerance of 1×10^{-2}. When evaluating the models, we decrease the tolerance of the spitting method to 1×10^{-4}. We use the same α tuning procedure as Winston & Kolter (2020). All models were trained using the same initial point. Note that for LBEN, this requires initializing the metric $\Lambda = I$.

The feed-forward models trained using Lipschitz margin training were trained using the original author's code which can be found at `https://github.com/ytsmiling/lmt`.

J.2 SVHN EXAMPLE

This section contains the model structures and the details of the training procedure used for the SVHN examples. All models are trained using the ADAM optimizer Kingma & Ba (2015) with an initial learning rate of 1×10^3. The models were trained for 5 epochs and the learning rate was reduced by a factor of 10 every 10 epochs. Each model contains a single convolutional layer with 40 channels and a linear output layer. To encourage quick convergence of the equilbirum network solver, we set $\epsilon = 5$.

The MON was evaluated using the Peaceman-Rachford Iteration scheme.

CONVOLUTIONAL LBEN

Following the approach of Winston & Kolter (2020), we parametrize U and V in equation 11 via convolutions. The skew symmetric matrix is constructed by taking the skew symmetric part of a convolution $\bar{S}$, so that $S = \frac{1}{2}(\bar{S} - \bar{S}^\top)$.

For computational simplicity, we impose a block constant structure on the contraction metric $\Lambda = \Psi^{-1}$. In particular, if the hidden layer of the convolutional network has n channels and size $s \times s$ and $W \in \mathbb{R}^{ns^2 \times ns^2}$, then we parametrize the metric as $\Psi = \bar{\Psi} \otimes I_n$ with $\bar{\Psi} \in \mathbb{R}^{s^2 \times s^2}$.

In Winston & Kolter (2020) Peaceman-Rachford is used and the operator $I - W$ can be quickly inverted using the fast Fourier transform. This situation is more complicated in our case as the term $W_{\text{out}}^\top W_{\text{out}}$ cannot be represented as a convolution. Instead, we apply FISTA algorithm shown

in equation 24 which only requires the evaluation of the proximal operator. FISTA requires the calculation of the singular values of the $I - W$ which can be upper bounded via the following:

$$||I - W||_2 = \left|\left|\Psi\left(\frac{1}{2\gamma}W_o^T W_o + \frac{1}{2\gamma}\Psi^{-1}UU^T\Psi^{-1} + V^T V + \epsilon I + S\right)\right|\right|_2, \tag{40}$$

$$\leq \left|\left|\Psi\left(\frac{1}{2\gamma}\Psi^{-1}UU^T\Psi^{-1} + V^T V + \epsilon I + S\right)\right|\right|_2 + \frac{1}{2\gamma}||W_o^T W_o||_2. \tag{41}$$

The first term can be quickly calculated using the approach in Sedghi et al. (2018). The second term can be calculated using a low rank singular value decomposition.

It should also be noted that we observed a similar trend to the Winston & Kolter (2020), where the Lipschitz constant of $I - W$ increases during training. This results in the number of iterations required for FISTA to converge to increase over time.

The gradient in equation 13 is calculated using forward backward splitting with fixed $\alpha = 5 \times 10^{-2}$.