Robust Implicit Networks via Non-Euclidean Contractions

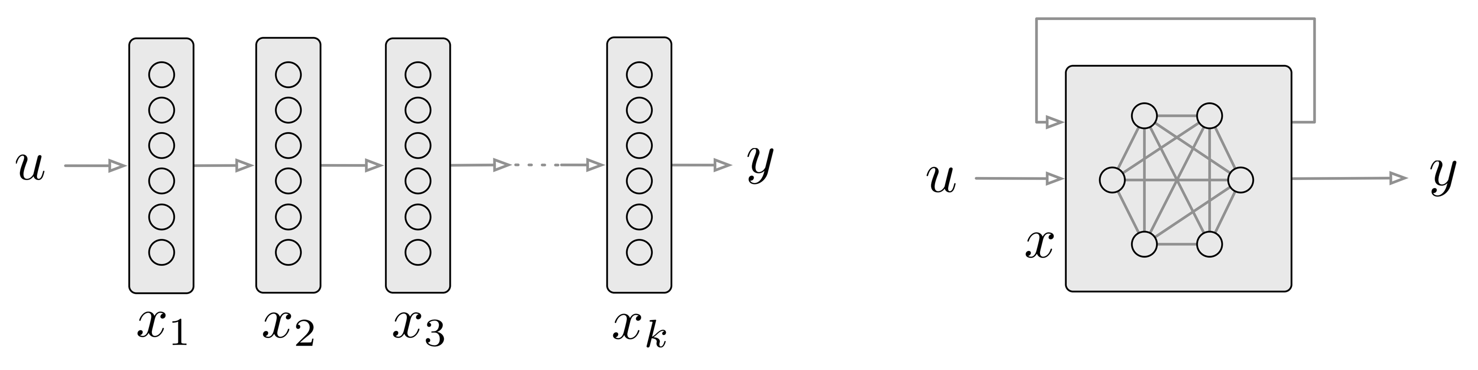
Saber Jafarpour*(1), Alexander Davydov*(1), Anton V. Proskurnikov (2), and Francesco Bullo (1)

(1) Center for Control, Dynamical Systems and Computation
University of California at Santa Barbara
{saber, davydov, bullo}@ucsb.edu

(2) Department of Electronics and Telecommunications
Politecnico di Torino,Turin, Italy
anton.p.1982@ieee.org

Implicit Neural Networks (INNs)

- INNs: Replacing the layers in NNs with implicit algebraic equations



Feedforward neural network — Implicit neural network

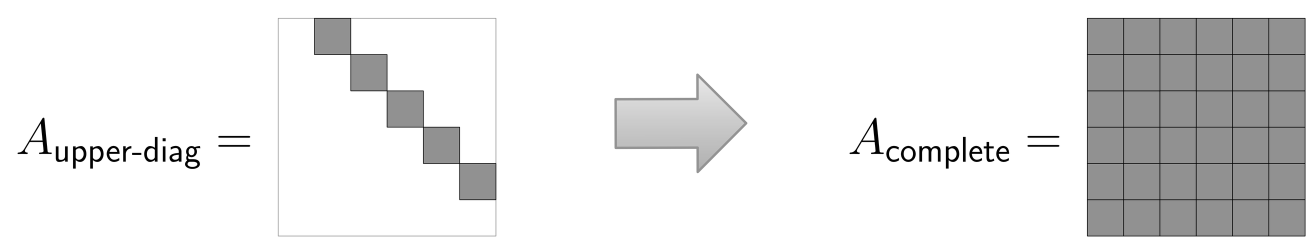
$$x^{i+1} = \Phi(A_i x^i + B_i u + b_i) \quad y = Cx^k + c$$

$$x = \Phi(Ax + Bu + b) \quad y = Cx + c$$

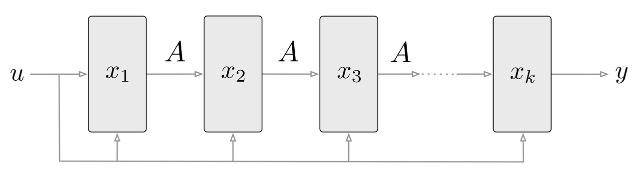
Motivations

- Inspired by neuronal circuits, implicit neural networks feature improved accuracy, improved input-output robustness, and reduced memory consumption [1,2]
- INNs generalize feedforward NNs to fully-connected synaptic matrices

$x^{i+1} = \Phi(A_i x^i + B_i u + b_i) \quad \Leftrightarrow \quad x = \Phi(Ax + Bu + b),\ A$ upper diag

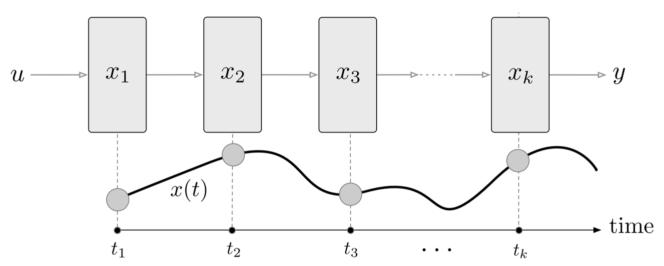


- INNs generalize weight-tied infinite-depth NNs



$x^{i+1} = \Phi(Ax^i + B_i u + b_i) \implies \lim_{i\to\infty} x^i = x^*$ solution to the INN

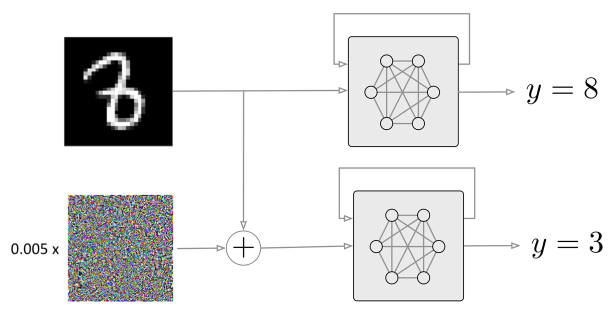
- INNs are a special case of Neural ODE models (infinite time)



$\dot{x} = -x + \Phi(Ax + Bu + b) \quad \Longrightarrow \quad \lim_{t\to\infty} x(t) = x^*$ solution to the INN

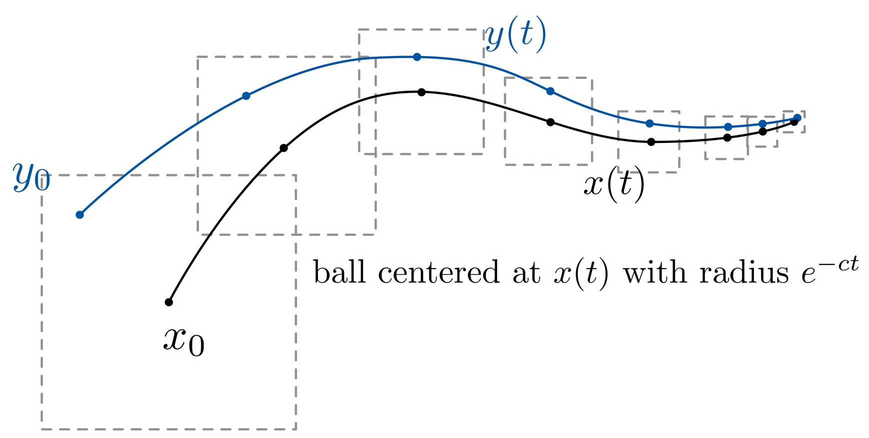
Challenges

- Existence and uniqueness of a fixed-point
- Efficient methods to compute the fixed-point
- Robustness to adversarial perturbations



Non-Euclidean Contraction Theory

A vector field is contracting if its flow is a contraction mapping for all times



ℓ_∞-matrix measure: $\mu_\infty(A) = \max_i \big(a_{ii} + \sum_{j\neq i} |a_{ij}|\big)$

A vector field $G : \mathbb{R}^n \to \mathbb{R}^n$ is contracting with respect to ℓ_∞-norm iff

$$\mu_\infty(D_x G(x)) \le -c, \qquad \text{for all } x$$

Well-Posedness of INNs

Key insight

Fixed-point of $x = \Phi(Ax + Bu + b)$ $\Longleftrightarrow$ Equilibrium point of $\dot{x} = -x + \Phi(Ax + Bu + b)$

Fixed-point approach: $\|A\|_\infty < 1$ then the Picard iteration

$$x^{k+1} = \Phi(Ax^k + Bu + b)$$

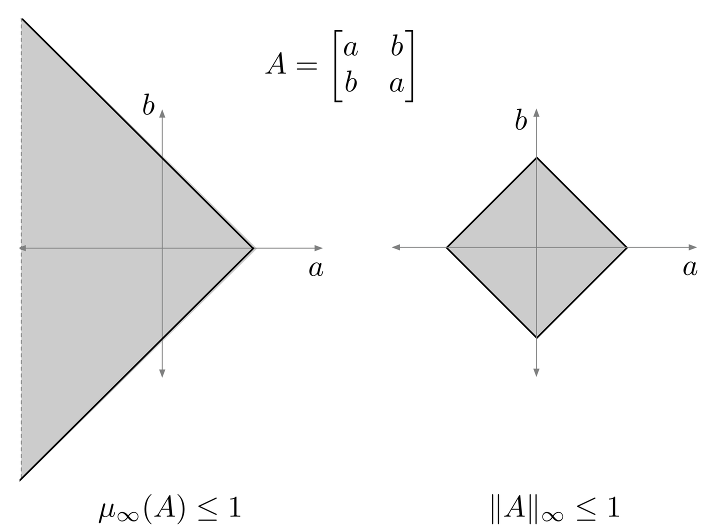
converges to a unique fixed-point.

Contraction theory approach: $\mu_\infty(A) < 1$ then the α-average iteration

$$x^{k+1} = (1-\alpha)x^k + \alpha\Phi(Ax^k + Bu + b)$$

converges to a unique fixed-point.

- **Accelerated convergence**: increased range of α compared to classical monotone operator methods
- **Neural ODE interpretation**: α-average iteration corresponds to forward Euler discretization of the ODE with step-size α



Input-Output Lipschitz Constant of INNs

$$\underbrace{u \mapsto x^*}_{\text{Lip}_{u\to x^*}} \underbrace{\mapsto y}_{\text{Lip}_{x^*\to y}} \qquad \text{Lip}_{u\to y} = \text{Lip}_{u\to x^*}\text{Lip}_{x^*\to y}$$

Input-output Lipschitz constant

if $\mu_\infty(A) < 1$ then

$$\text{Lip}_{u\to y} = \frac{\|B\|_\infty \|C\|_\infty}{1 - \mu_\infty(A)_+}.$$

Non-Euclidean Monotone Operator Network (NEMON)

- INN model:

$$x = \Phi(Ax + Bu + b) \quad y = Cx + c$$

- INN training: Training data $\{\widehat{u}_i, \widehat{y}_i\}_{i=1}^N$

$$\min_{A,B,C,b,c} \sum_{i=1}^N \mathcal{L}(\widehat{y}_i, Cx_i + c) + \underbrace{\lambda\ \text{Lip}_{u\to y}}_{\text{Promoting robustness}}$$

$$x_i = \Phi(Ax_i + B\widehat{u}_i + b)$$

$$\underbrace{\mu_\infty(A) \le \gamma}_{\text{Well-posedness}},$$

- $\gamma < 1$ is a hyperparameter
- $\lambda \ge 0$ is a regularization parameter
- α-average iterations for solving $x_i = \Phi(Ax_i + B\widehat{u}_i + b)$

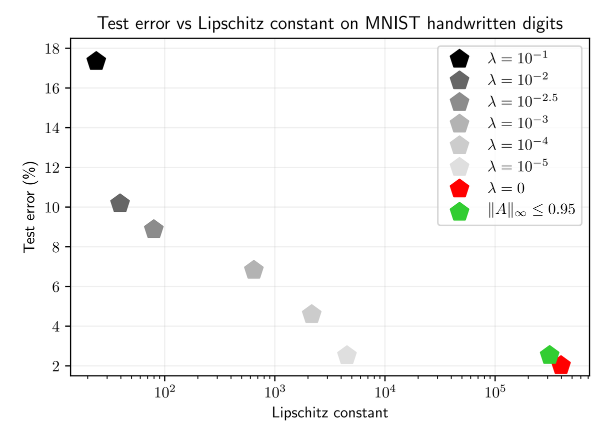
Parametrization of μ_∞-constraint

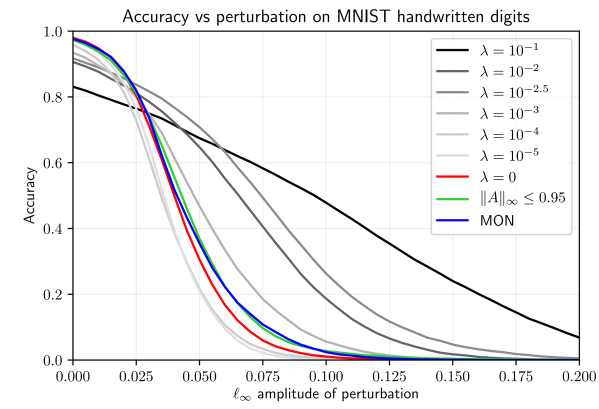
$$\mu_\infty(A) \le \gamma \quad \Longleftrightarrow \quad \exists T \ \text{ s.t. } \ A = T - \text{diag}(|T|\mathbb{1}_n) + \gamma I_n.$$

Numerical Experiments

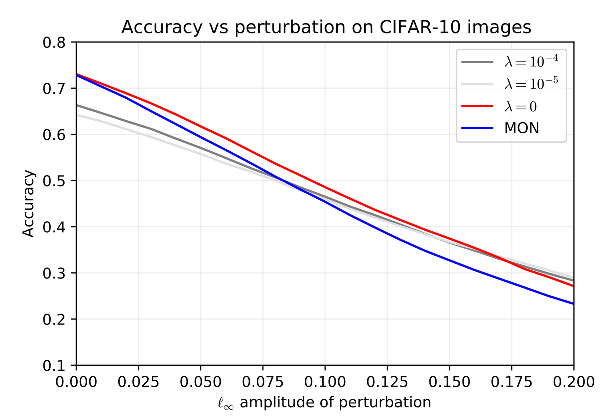
Setup: INNs models:
- MON from [1] with $m = 0.05$,
- IDL from [2] with $\|A\|_\infty \le 0.95$,
- NEMON with $\gamma = 0.95$

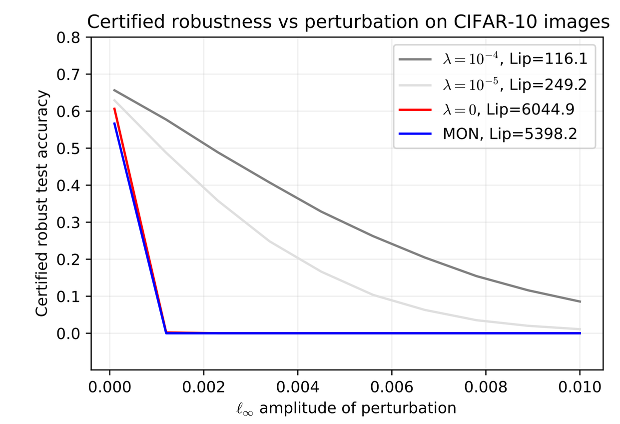
$\Phi = \text{ReLU}$
MNIST INNs size: $n = 100$
CIFAR-10 INNs size: 81 channel
Attack: Projected Gradient Descent (PGD)





- Pareto-optimal curve for Lipschitz constant vs. test error on MNIST
- Empirical robustness on MNIST and CIFAR-10 to PGD attacks: By losing few percentages in clean performance we observe improvements in robust accuracy





- Certified adversarial robustness: training with $\lambda > 0$ lead to a dramatic improvement in percentage of test examples which can be certified
- Code: https://github.com/davydovalexander/Non-Euclidean_Mon_Op_Net

References

(1) E. Winston and J. Z. Kolter. Monotone operator equilibrium networks. *In NeurIPS*, 2020.
(2) L. El Ghaoui, et al., Implicit deep learning. *SIMODS*, 3(3):930–958, 2021.