On Robust Classification using Contractive Hamiltonian Neural ODEs

Muhammad Zakwan, Liang Xu, Giancarlo Ferrari-Trecate
Institute of Mechanical Engineering, EPFL, Switzerland.

Neural ODEs are Fragile

- A neural ODE (NODE) is a continuous-depth model

$$\boldsymbol{\xi}(0) = \phi(\boldsymbol{x}, \omega)\ , \ \text{(input layer)} \tag{1}$$
$$\dot{\boldsymbol{\xi}}(t) = f_t(\boldsymbol{\xi}(t), \theta(t))\ , \ \text{(continuum of hidden layers)} \tag{2}$$
$$\boldsymbol{y}(T) = \psi(\boldsymbol{\xi}(T), \eta)\ , \ \text{(output layer)} \tag{3}$$

where $t \in [0,\ T]$, $\boldsymbol{x}$ is the input data (e.g. an image), $\boldsymbol{\xi}$ represents the state of the NODE, and ϕ, f, ψ are neural networks.

- Neural ODEs may not be robust to noise in features.

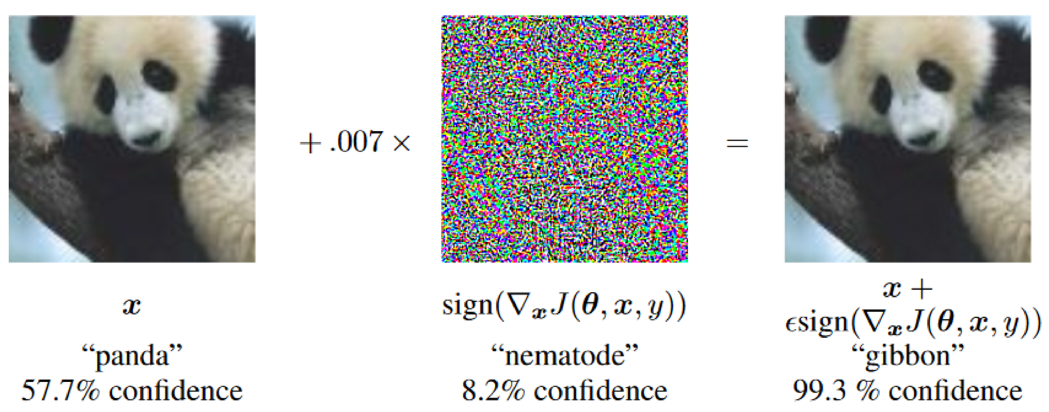

Figure 1: Fast gradient sign method attack (Goodfellow et al., 2014)

Contractivity Promotes Robustness

Definition 1. *Let $\boldsymbol{\xi}(t)$ and $\tilde{\boldsymbol{\xi}}(t)$ be two solutions of* (2) *starting from $\boldsymbol{\xi}(0)$ and $\tilde{\boldsymbol{\xi}}(0)$, respectively. Then* (2) *is contractive if $\exists C, \rho > 0$, such that $||\tilde{\boldsymbol{\xi}}(t) - \boldsymbol{\xi}(t)|| \le Ce^{-\rho t}||\tilde{\boldsymbol{\xi}}(0) - \boldsymbol{\xi}(0)||$ for all $t > 0$.*

If the ODE (2) is contractive, then perturbations in initial conditions vanish exponentially fast.

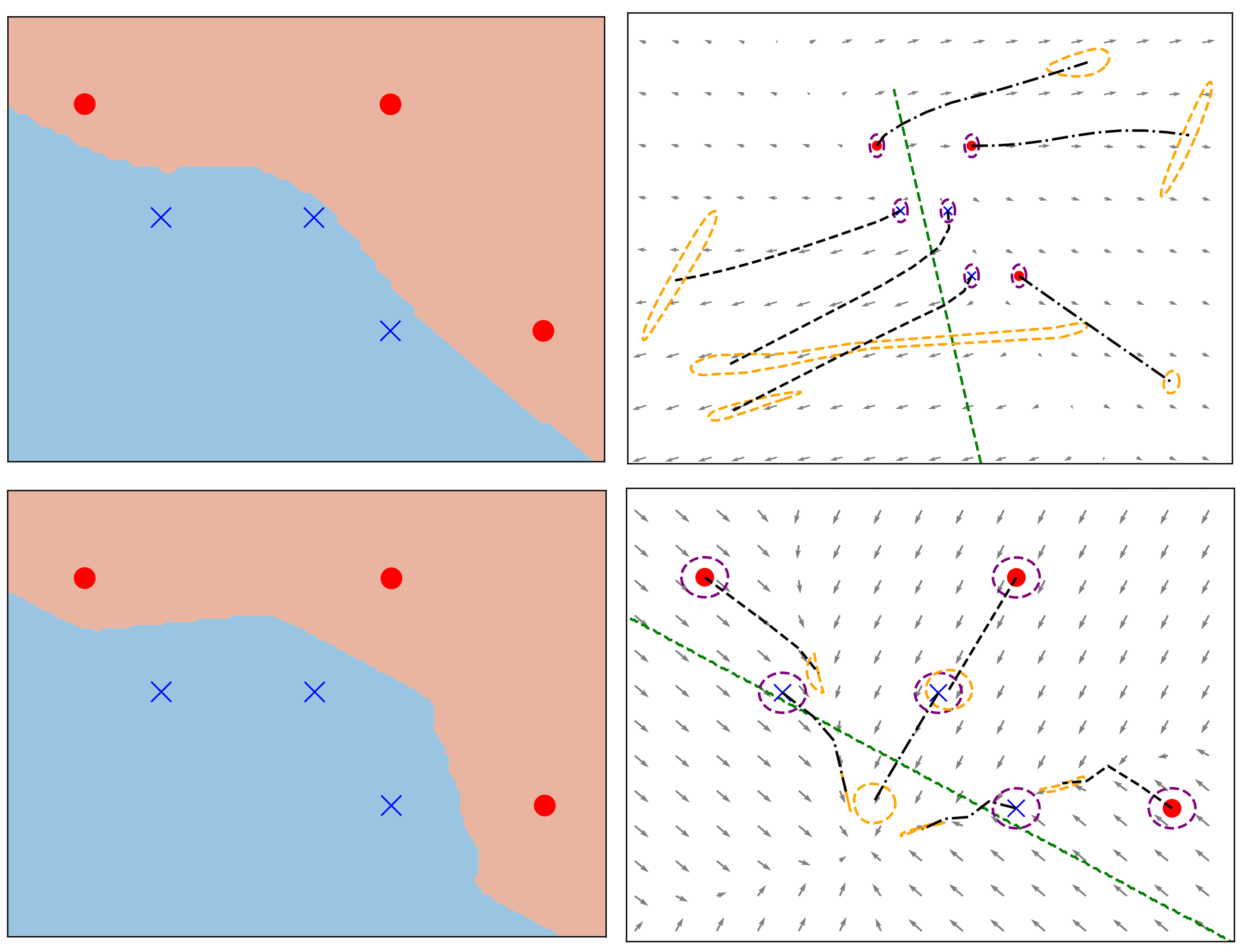

Figure 2: Comparison between a vanilla NODE (top) exhibiting sensitivity, and a contractive-NODE (bottom) showing robustness against input perturbations on a 2D binary classification task.

Neural ODEs with Contractivity by Design

By design: Allows almost free parametrization of weights, and decrease computationally complexity by lifting the need of regularizers.

Theorem 1. *For a given constant skew-symmetric matrix $\boldsymbol{J} = -\boldsymbol{J}^\top$, let*

$$\dot{\boldsymbol{\xi}} = (\boldsymbol{J} - \gamma \boldsymbol{I})\Big(\boldsymbol{K}^\top(t)\sigma(\boldsymbol{K}(t)\boldsymbol{\xi} + \boldsymbol{b}(t)) + (\boldsymbol{L}^\top(t)\boldsymbol{L}(t) + \kappa \boldsymbol{I})\boldsymbol{\xi}\Big), \tag{4}$$

where $\sigma(\cdot)$ is the activation function and has bounded derivative $0 \le \sigma'(\cdot) \le S$ for $S > 0$, $\kappa > 0$ is a constant, $\boldsymbol{K}$, $\boldsymbol{b}$, and $\boldsymbol{L}$ are trainable parameters, and define $c_1 = \inf_{s\in[0,T]} \underline{\lambda}(\boldsymbol{L}^\top(s)\boldsymbol{L}(s)) + \kappa, c_2 = \sup_{s\in[0,T]}(\bar{\lambda}(\boldsymbol{L}^\top(s)\boldsymbol{L}(s)) + S\bar{\lambda}(\boldsymbol{K}^\top(s)\boldsymbol{K}(s))) + \kappa, \alpha = \dfrac{c_2 - c_1}{c_2 + c_1}$. If $\epsilon > 0$ is such that $1 - \alpha^2 - \epsilon > 0$ and $\gamma \ge \sqrt{\dfrac{(\alpha^2+\epsilon)\bar{\lambda}(\boldsymbol{J}\boldsymbol{J}^\top)}{1 - \alpha^2 - \epsilon}}$, then, NODE (4) *is contractive.*

The ODE (4) is a Hamiltonian system without input-output ports. Therefore, is called *contractive Hamiltonian neural ODE (CH-NODE).*

Non-exploding Gradients

Backward Sensitivity Matrices (BSM) for NODE (4) is

$$\frac{\partial \boldsymbol{\xi}(T)}{\partial \boldsymbol{\xi}(T-t)}, \quad \forall t \in [0, T]. \tag{5}$$

- **Vanishing/Exploding Gradients :** convergence to zero or the divergence of BSM during training. **Causes numerical instability**.

Theorem 2. *The BSM* (5) *associated with the CH-NODE* (4) *satisfies*

$$\left|\left|\frac{\partial \boldsymbol{\xi}(T)}{\partial \boldsymbol{\xi}(T-t)}\right|\right| \le \exp\left(-\frac{\rho}{2}t\right), \quad \forall t \in [0, T]\ , \tag{6}$$

where $\rho = \frac{\epsilon\beta(\gamma^2 + \bar{\lambda}(\boldsymbol{J}\boldsymbol{J}^\top))}{\gamma}$, and $\beta = \frac{1}{2}(c_1 + c_2)$. Moreover, we have $\left|\left|\dfrac{\partial \boldsymbol{\xi}(T)}{\partial \boldsymbol{\xi}(0)}\right|\right| \le 1$, i.e., the input-output sensitivity is smaller than 1 (robustness guarantees).

Experiments

1. **MNIST**

		Nominal		$\mathcal{N}(0,\sigma)$		$s\&p(\sigma)$	
N	NN	Train	Test	$\sigma = 0.05$	$\sigma = 0.2$	$\sigma = 0.05$	$\sigma = 0.2$
4	ResNet	**98.91**	**97.01**	63.00	52.56	59.8	42.02
	H-DNN	94.68	94.60	31.12	26.65	30.52	23.83
	C-HNN	94.03	92.38	**81.30**	**77.69**	**79.86**	**63.84**
8	ResNet	**99.12**	**97.28**	32.99	30.56	30.27	28.11
	H-DNN	95.30	95.17	60.8	49.88	61.15	45.62
	C-HNN	89.55	89.01	**86.33**	**81.85**	**84.22**	**72.18**
12	ResNet	**99.11**	**96.86**	39.13	34.04	41.04	29.80
	H-DNN	95.36	95.23	26.79	23.53	27.48	22.75
	C-HNN	90.01	89.76	**85.68**	**80.97**	**84.88**	**72.82**

Table 1: Robustness comparison among ResNets (He et al., 2016), H-DNNs (Galimberti et al., 2021), and C-HNNs under the zero-mean Gaussian and the salt and pepper noise.

2. **Non-exploding gradients**

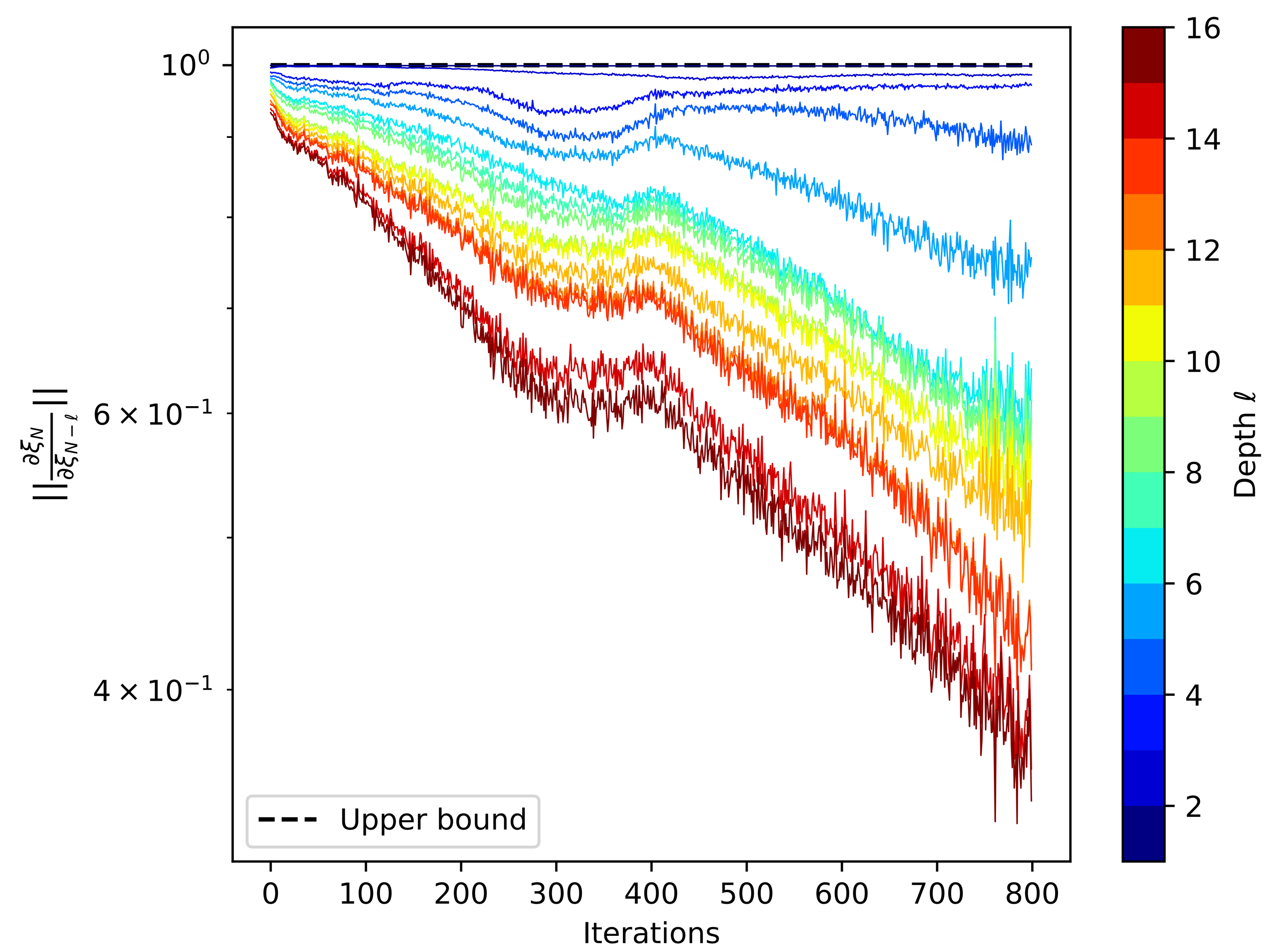

Figure 3: Evolution of 2-norm of the BSM during the training of a 16-layer C-HNN exhibiting non-exploding gradients.

Conclusion and Future Work

- NODEs based on Hamiltonian dynamics that are contractive by design, enjoys non-exploding gradients, and improved robustness guarantees.
- Analyze the robustness of CH-NODEs against adversarial attacks (e.g. FGSM, PGM).