Probabilistic Robustness Certificates against Adversarial Attacks

Sara Taheri[1], Majid Zamani[1,2]

[1] LMU Munich, Germany

[2] CU Boulder, USA

ICML 2026

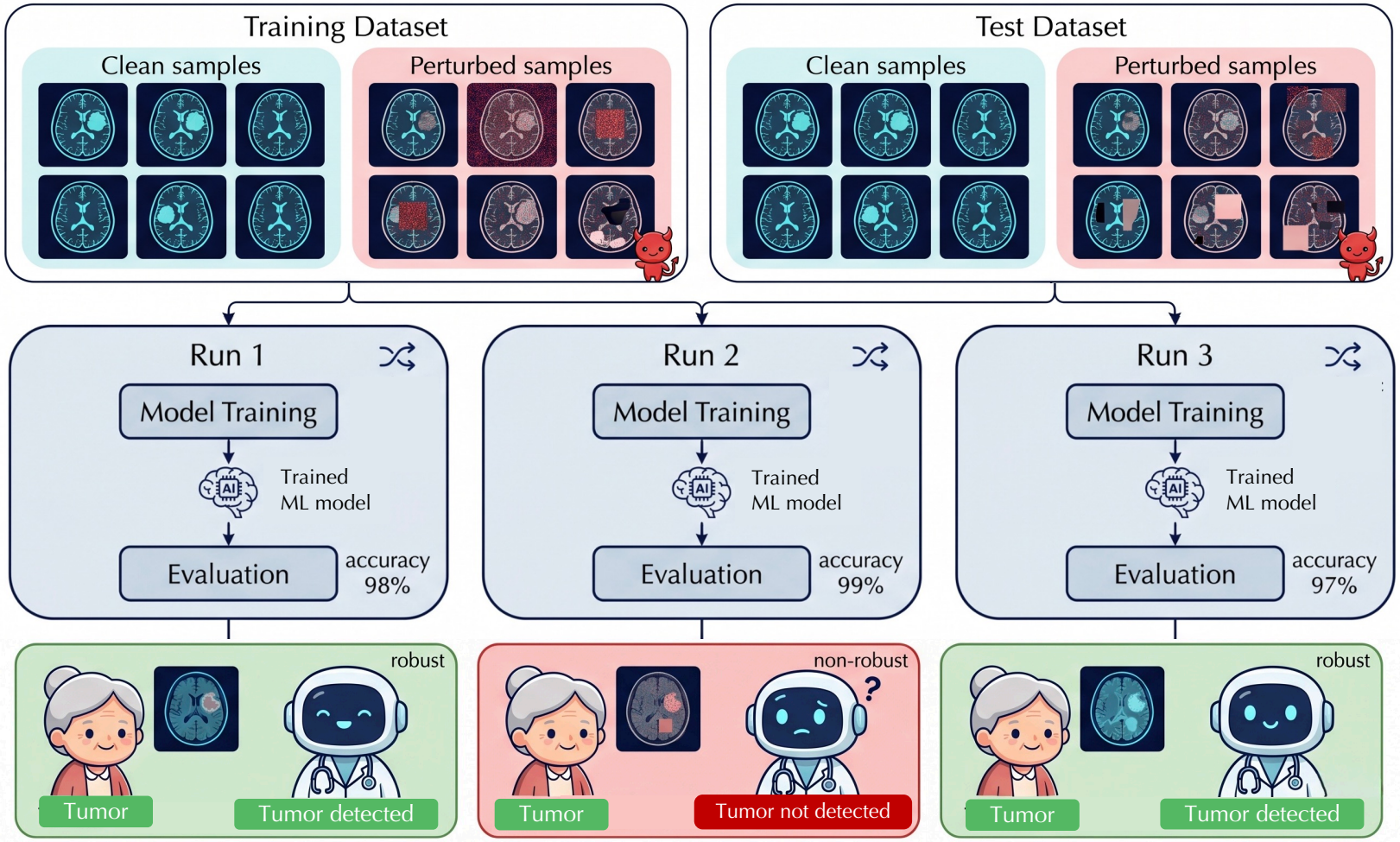
HyConSys Lab

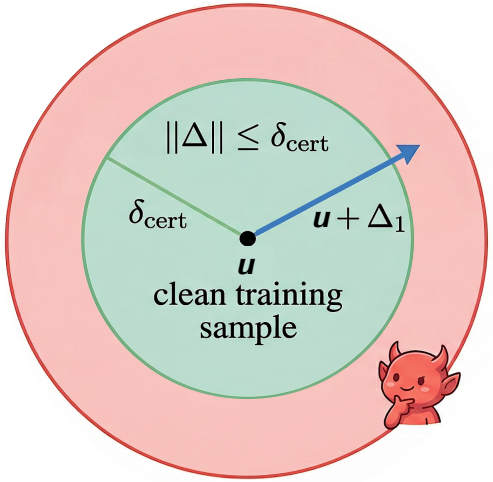
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

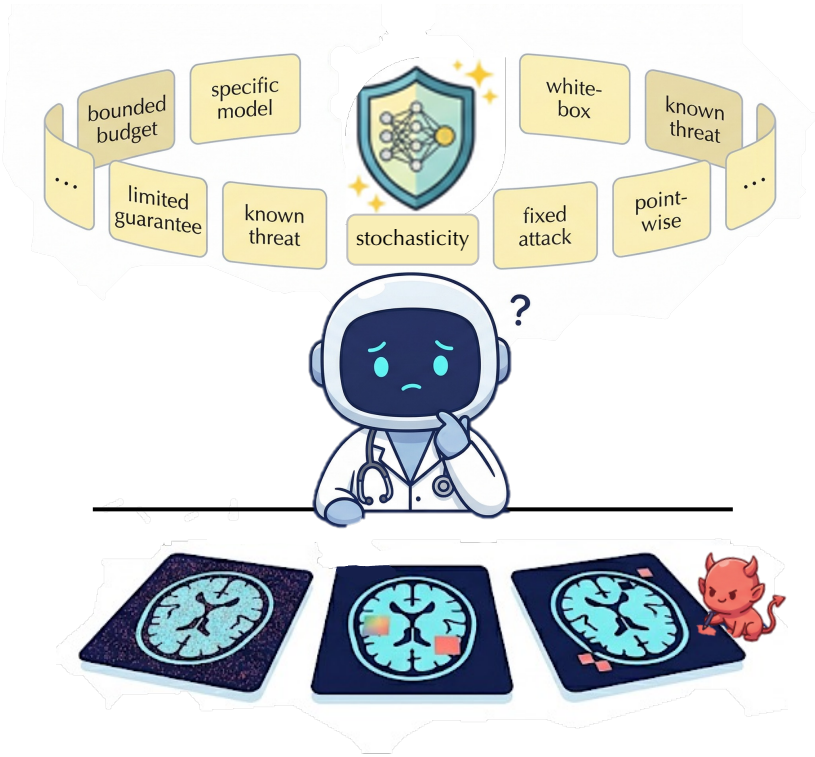
ICML International Conference On Machine Learning

University of Colorado Boulder

ConVeY


Training Dataset
Clean samples
Perturbed samples
Test Dataset
Clean samples
Perturbed samples
Run 1
Model Training
Trained ML model
Evaluation
accuracy 98%
Run 2
Model Training
Trained ML model
Evaluation
accuracy 99%
Run 3
Model Training
Trained ML model
Evaluation
accuracy 97%
robust
Tumor
Tumor detected
non-robust
Tumor
Tumor not detected
robust
Tumor
Tumor detected

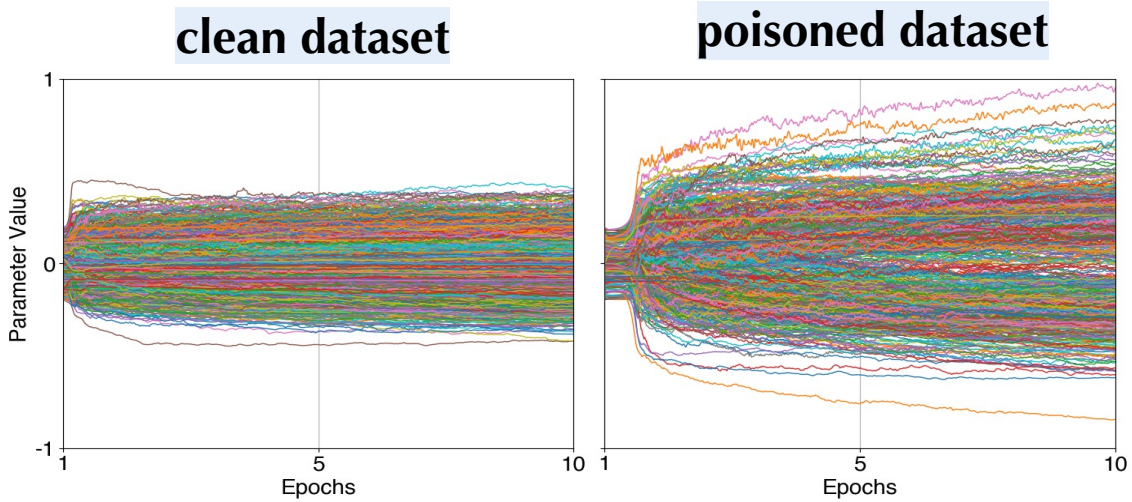


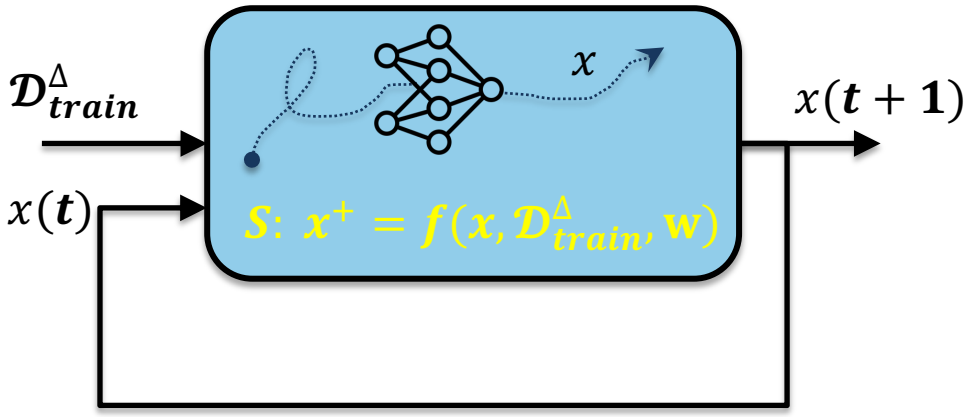


Robustness Problem

Find: The maximum certified poisoning radius ($\boldsymbol{\delta_{\text{cert}}}$)

Guarantee: For any attacker budget bounded by ($\|\boldsymbol{\Delta}\|_p \leq \boldsymbol{\delta_{\text{cert}}}$) the model's accuracy strictly stays above the safety threshold with a certified probability.

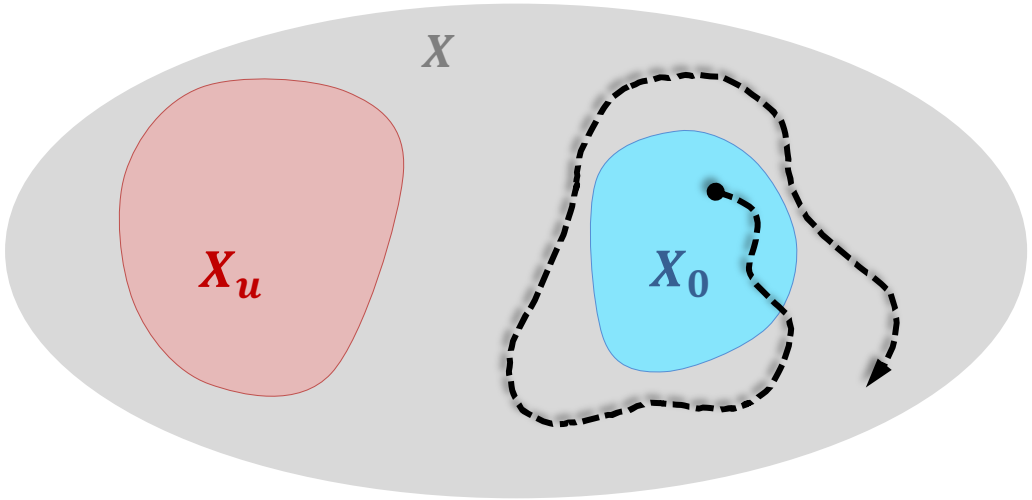




- **State**: model parameters $x \in X \subseteq \mathbb{R}^d$
- **Input**: poisoned training data $\mathcal{D}_{\text{train}}$
- **Dynamics**: Parameter update map f
- **Uncertainty:** Stochasticity in ML pipeline $w \in \mathcal{W}$

Safe set (X_s):

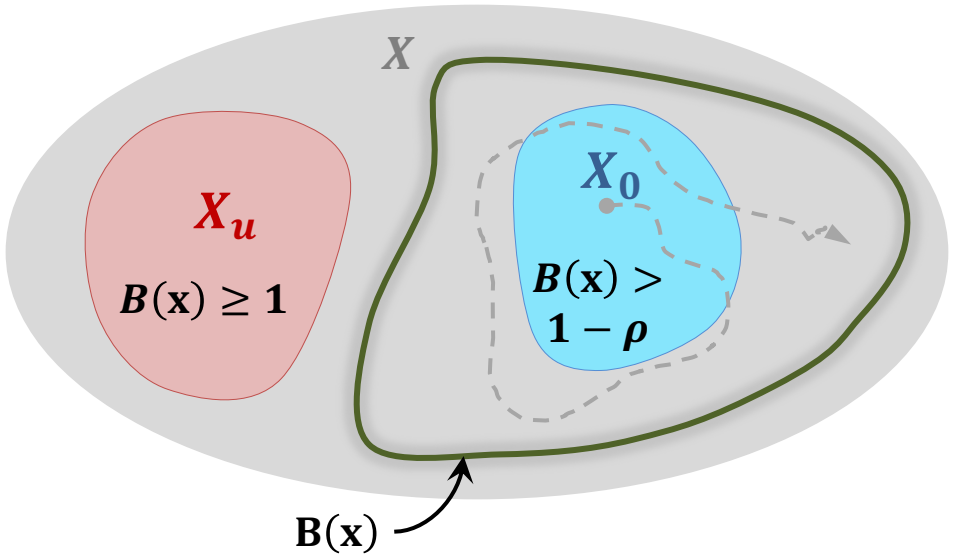
The region in the parameter space where the model's accuracy remains above the critical threshold.



$X_0 \subseteq X$: **Initial parameter set**
$X_u \subseteq X$: **Unsafe parameter set**
X: **Parameter set**

Can we certify that the final trained parameters remain in the safe set with some probability?

$X_0 \subseteq X$: Initial set
$X_u \subseteq X$: Unsafe set
X: State set

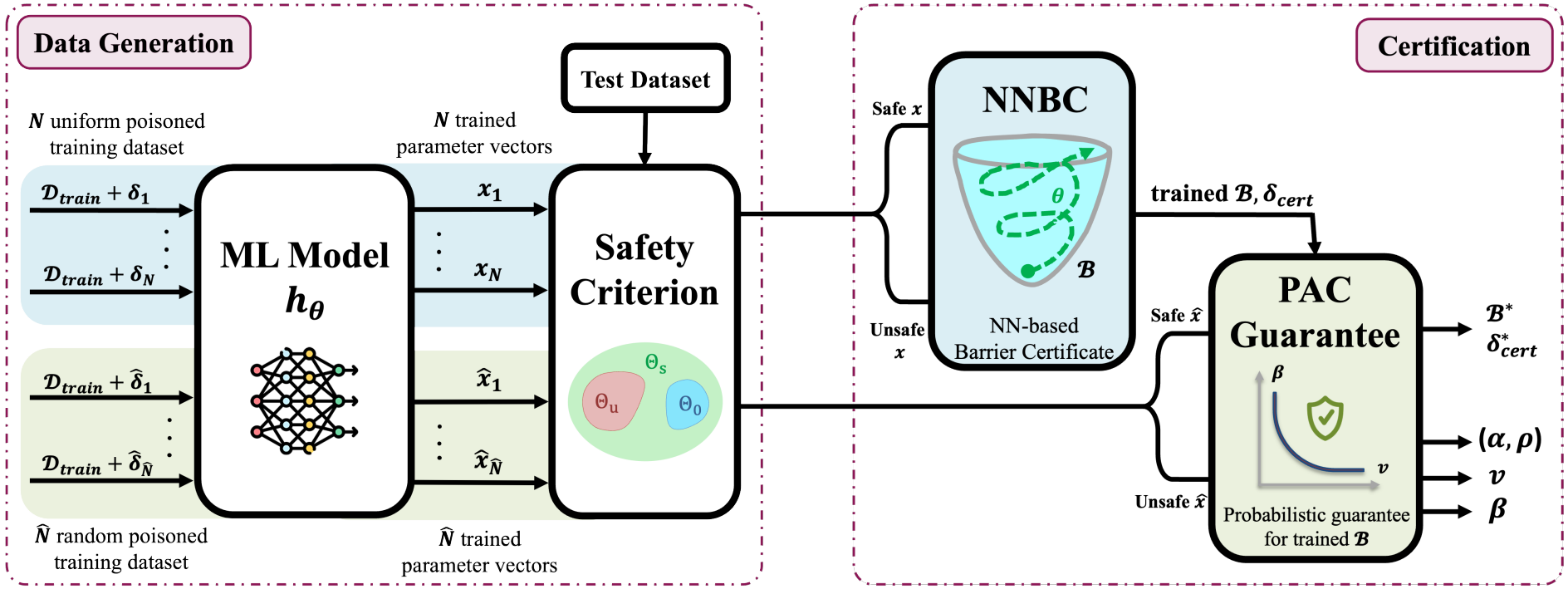


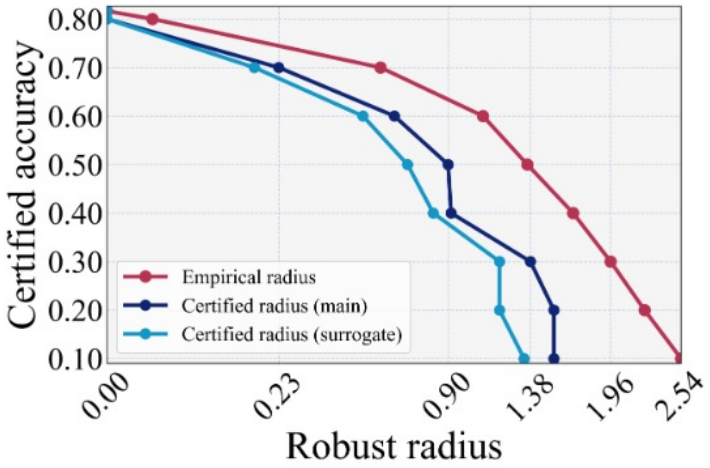
Theorem. BC for a dt-DS

Given a threshold $\rho \in [0,1]$, If there exists a function $\mathbf{B}$ and a certified robust radius δ such that:

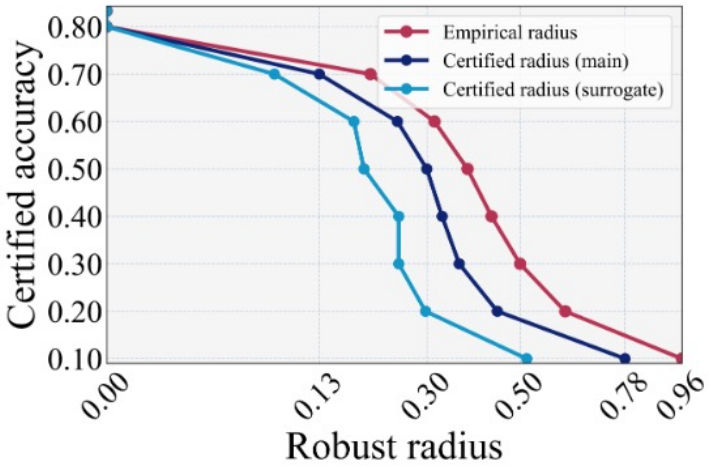
- $\mathbf{B}(x, 0) \leq 1 - \rho, \quad \forall x \in X_0,$
- $\mathbf{B}(x, \mathrm{T}) \geq 1, \quad \forall x \in X_u,$
- $\mathbb{E}[\,\mathbf{B}(x^+, \mathrm{t}) \mid x\,] \leq \mathbf{B}(x, \mathrm{t}-1), \quad \forall x \in X, \|\Delta\|_p \leq \delta.$

$\Longrightarrow$ $x(t) \notin X_u$ with probability at least ρ.

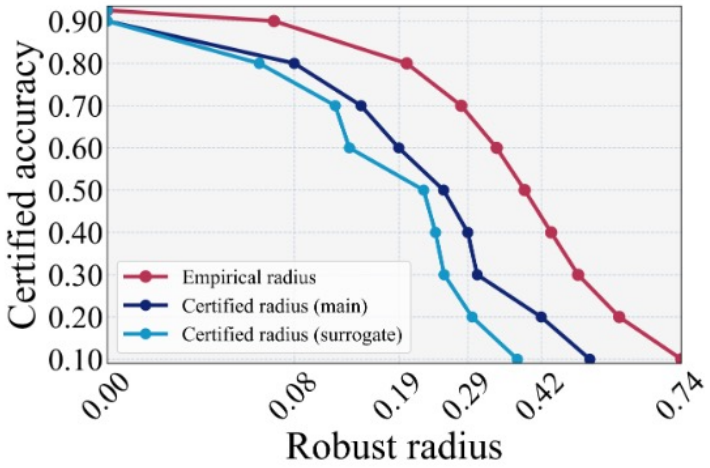

Data Generation
Test Dataset
N uniform poisoned training dataset
N trained parameter vectors
$\mathcal{D}_{train} + \delta_1$
$\mathcal{D}_{train} + \delta_N$
ML Model h_θ
x_1
x_N
Safety Criterion
Θ_s
Θ_u
Θ_0
$\mathcal{D}_{train} + \hat{\delta}_1$
$\mathcal{D}_{train} + \hat{\delta}_{\hat{N}}$
$\hat{x}_1$
$\hat{x}_{\hat{N}}$
$\hat{N}$ random poisoned training dataset
$\hat{N}$ trained parameter vectors
Certification
Safe x
Unsafe x
NNBC
θ
$\mathcal{B}$
NN-based Barrier Certificate
trained $\mathcal{B}, \delta_{cert}$
Safe $\hat{x}$
Unsafe $\hat{x}$
PAC Guarantee
β
v
Probabilistic guarantee for trained $\mathcal{B}$
$\mathcal{B}^*$
δ^*_{cert}
(α, ρ)
v
β



(a) SVHN, CNN, SGD, BPA ($\ell_2, \tau = 0.2$)



(b) CIFAR10, ResNet, Adam, PGD ($\ell_\infty, \tau = 0.8$)



(c) MNIST, CNN, SGD, BPA ($\ell_2, \tau = 0.3$)

Our framework provides model- and attack-agnostic probabilistic certificates for robustness under stochastic training, using neural barrier certificates and PAC-style verification.

Thanks for your attention!

HyConSys Lab

ICML
International Conference
On Machine Learning

University of Colorado
Boulder