Asymmetric Certified Robustness via Feature-Convex Neural Networks

Samuel Pfrommer [1,*] Brendon G. Anderson [1,*] Julien Piet [1] Somayeh Sojoudi [1]
[1]University of California, Berkeley [*]Co-First Authors

Introduction

- We consider **certified robustness** of neural networks (NNs) against adversarial attacks.
- Real-world adversaries are **asymmetric**; change spam into non-spam, red light to green, etc.
- Past methods do not take asymmetric structure into account, causing loose or expensive certificates.
- We propose architecture tailored to this setting.

Contributions

Propose **feature-convex neural networks** to achieve:

1. **Closed-form certified radii** for arbitrary ℓ_p-norms.
2. Characterize decision region geometry, extend universal approximation theorem to convex classification, and prove **high expressivity** in high dimensions.
3. **Faster certificates that meet/exceed state-of-the-art** on MNIST, Malimg, Fashion-MNIST, and CIFAR-10.

Proposed Architecture

- Binary **feature-convex** classifier $f\colon \mathbb{R}^d \to \{0,1\}$:

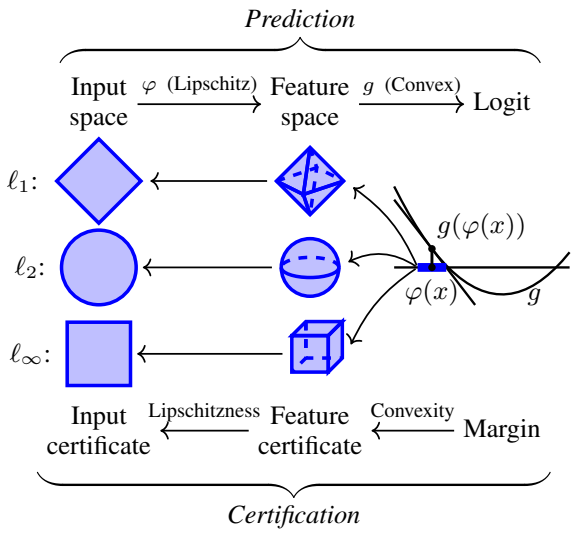
$$f(x) = \mathbf{1}_{>0}(g(\varphi(x))),$$

g convex, φ Lipschitz "feature map."

- Our numerical implementations:

g ReLU NN with nonnegative weights,
$\varphi(x) = (x, |x|)$, so $\mathrm{Lip}_p(\varphi) \le 2$.

- **Key insight:** Simple, low-Lipschitz φ yields highly expressive nonconvex $g \circ \varphi$ with large certified radii.



Theoretical Analysis

- **Closed-form** radius for certified robustness:

Theorem 1 *If $f(x) = 1$, then $f(x+\delta) = 1$ for all δ such that*

$$\|\delta\|_p < r(x) := \frac{g(\varphi(x))}{\mathrm{Lip}_p(\varphi)\|\nabla g(\varphi(x))\|_{p,*}}.$$

- Universal approximation for convex classifiers:

Theorem 2 *For any convex classifier h, any convex compact $X \subseteq \mathbb{R}^d$, and any $\epsilon > 0$, there exists feature-convex ReLU NN f such that*

$$\mathrm{Vol}(\{x \in X : f(x) \neq h(x)\}) < \epsilon.$$

- Can **perfectly fit** high-dimensional datasets:

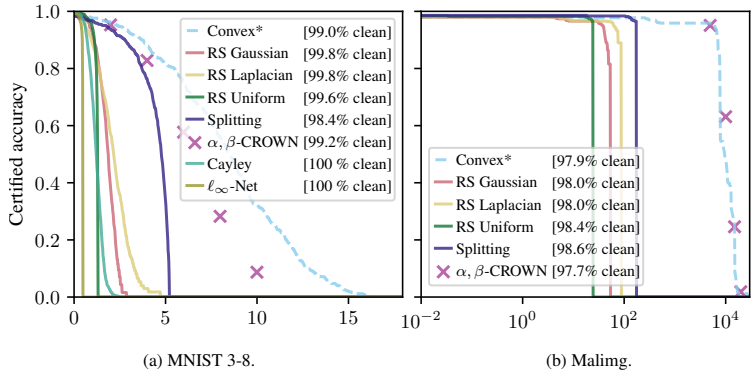
Finding 1 *There exists feature-convex f achieving* $\mathbf{100\%}$ ***training accuracy on CIFAR-10 cats-versus-dogs.***

Theorem 3 *Consider i.i.d.* $\mathrm{Unif}([-1,1]^d)$ *datasets* $X_1 = \{x^{(1)}, \dots, x^{(M)}\}$, $X_0 = \{y^{(1)}, \dots, y^{(N)}\}$. *Then*

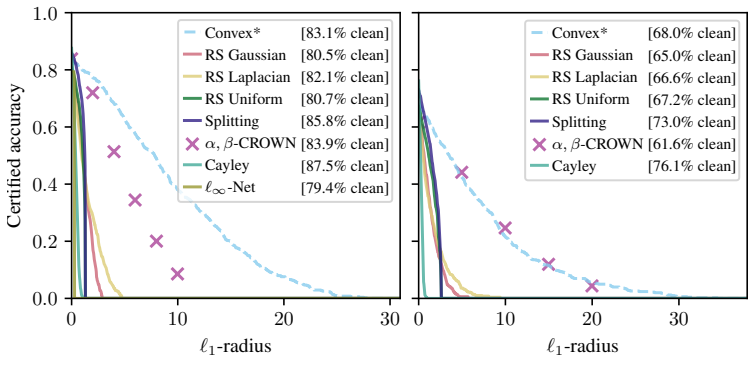
$$\mathbb{P}((X_1, X_0) \text{ is } f\text{-separable}) \ge \begin{cases} 1 - \left(1 - \frac{M!N!}{(M+N)!}\right)^d & \text{for all } d \in \mathbb{N}, \\ 1 & \text{if } d \ge M+N. \end{cases}$$

Experiments

- Our method $\mathbf{1000}\times$ **faster** than randomized smoothing, $\mathbf{3000}\times$ **faster** than α, β-CROWN.
- Our certified accuracy meets/exceeds state-of-the-art.



(a) MNIST 3-8. (b) Malimg.



(c) Fashion-MNIST shirts. (d) CIFAR-10 cats-dogs.