EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATHEMATISCH-
NATURWISSENSCHAFTLICHE
FAKULTÄT

Provably Adversarially Robust Nearest Prototype Classifiers

Václav Voráček Matthias Hein

July 15, 2022

Adversarial Robustness

$$f : [0,1]^d \to \{1, 2, \ldots, Y\}$$

Definition

The robust radius of f at point x is the maximal r such that

$$\forall x' : d(x, x') < r \implies f(x') = y,$$

where y is the correct class of x.

Nearest Prototype Classifier (NPC)

NPC

- (semi)-metric $d(\cdot, \cdot)$
- set of prototypes $W_y = \{w_i \in \mathbb{R}^d\}_{i=1}^N$ for every class y
- $f(x) = \underset{y=1,\dots,K}{\arg\min} \min_{w \in W_y} d(x, w)$

Advantages:

- Fast
- Interpretable
- Exact certification

Robust Radius Computation - ϵ

Exact Radius Computation (of class y)

$$r(x)_w = \min_{x' \in [0,1]^d} \quad d_2(x, x')$$

$$\text{sbj. to: } d_1(x', w_y) \leq d_1(x', w), \forall\, w_y \in W_y$$

$$\epsilon(x) = \min_{w \in W_{yc}} r(x)_w$$

Lower-bound Computation (of class y)

$$\rho(x)_{w,w_y} = \min_{x' \in [0,1]^d} \quad d_2(x, x')$$

$$\text{sbj. to: } d_1(x', w_y) \leq d_1(x', w),$$

$$\epsilon(x) \geq \text{lower-bound}(x) = \min_{w \in W_{yc}} \max_{w_y \in W_y} \rho(x)_{w,w_y}$$

Computational Complexities

ℓ_p-distance	ℓ_q-threat model		
	ℓ_1	ℓ_2	ℓ_∞
ℓ_1	NP-hard	NP-hard	Poly
ℓ_2	Poly	Poly	Poly
ℓ_∞	NP-hard	NP-hard	NP-hard

Table: Computational Complexity of $r(x)$ and $\epsilon(x)$.

ℓ_p-distance	ℓ_q-threat model		
	ℓ_1	ℓ_2	ℓ_∞
ℓ_1	NP-hard	NP-hard	$O(d \log(d))$
ℓ_2	$\Theta(d)$	$\Theta(d)$	$\Theta(d)$
ℓ_∞	$\Theta(d)$	$O(d \log(d))$	$\Theta(d)$

Table: Computational complexity of $\rho(x)_{w,w_y}$.

Certified Robust Accuracy (CRA) - ℓ_2 - MNIST

	std. acc.	$\epsilon_2 = 1.5$ CRA	$\epsilon_2 = 1.58$ CRA	$\epsilon_2 = 2$ CRA
NPC	97.3	**75.5**	**73.0**	**56.1**
1-NN	96.9	52.1	47.3	23.7
GloRob	97.0	-	62.8	-
OrthConv	**98.1**	-	61.0	-
LocLip	96.3	-	55.8	-
BCP	92.4	-	47.9	-
CAP	88.1	-	44.5	-
SmoothLip$_{\sigma=0.5}$	**98.7**	**81.8***	-	0*
SmoothLip$_{\sigma=1}$	93.7	62.7*	-	44.9*

- Cifar10 - :(

LPIPS distance

- For a deep network g, we use $\hat{g}^{(l)}(x)$ to denote normalized (over channels) activations in the l-th layer.

Definition

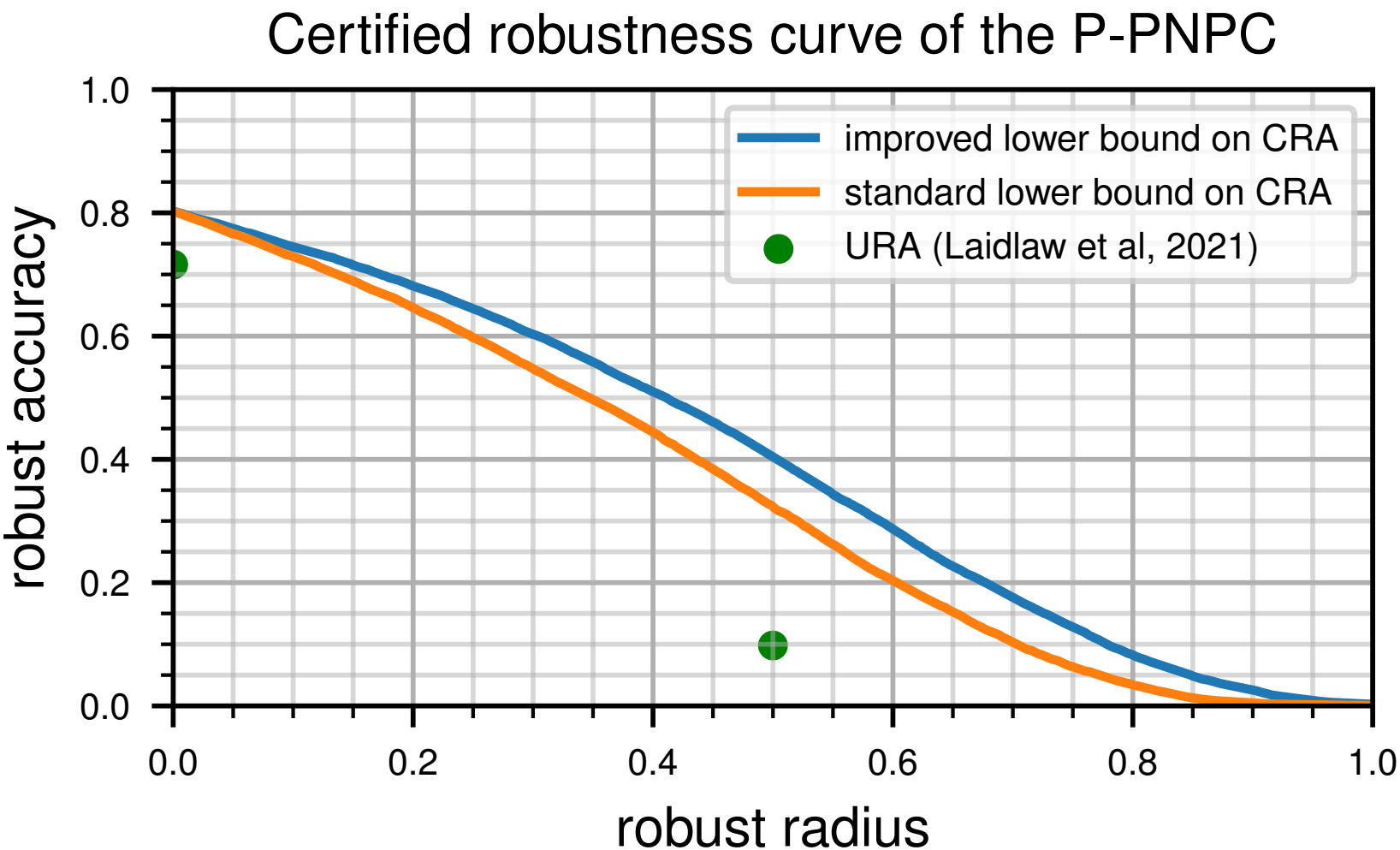
LPIPS-distance induced by model g is

$$d(x,y) = \sqrt{\sum_{l \in I_L} \frac{1}{H_l W_l} \sum_{h,w} \left\| \hat{g}^{(l)}_{h,w}(x) - \hat{g}^{(l)}_{h,w}(y) \right\|_2^2}$$

Equivalent definition

After a feature mapping $\phi(x) = \left(\frac{\hat{g}^{(1)}}{\sqrt{H_1 W_1}}, \dots, \frac{\hat{g}^{(L)}}{\sqrt{H_L W_L}} \right)$,
LPIPS-distance induced by model g is $d(x,y) = \|\phi(x) - \phi(y)\|_2$

LPIPS experiment



Box Constraints