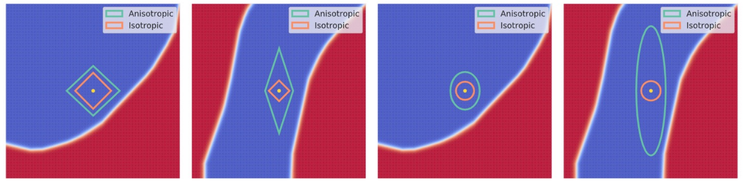
ANCER: Anisotropic Certification via Sample-wise Volume Maximization

Francisco Eiras[1, 2], Motasem Alfarra[3], M. Pawan Kumar[1], Philip H.S. Torr[1], Puneet K. Dokania[1, 2], Bernard Ghanem[3], Adel Bibi[1]

[1]University of Oxford [2]Five AI Limited [3]KAUST

Motivating Anisotropic Certificates



- **Isotropic ℓ_p-norm** ($\|\delta\|_p \leq r$) certificates are:
  - Worst-case: limited on all directions by the closest adversary
  - Uniformative w.r.t. decision boundary: do not indicate regions around the input which are safe
- **Our proposed anisotropic $\ell_p^{\mathbf{A}}$** ($\|\delta\|_{\mathbf{A},p} = \|\mathbf{A}^{-1/p}\delta\|_p \leq r$ for $p \in \{1, 2\}$) certificates:
  - Maintain worst-case in the directions of the closest adversary
  - Include larger safe regions in other directions

CIFAR-10 perturbed images (bottom) inside an optimized anisotropic certificate, but not in an optimized isotropic certificate

Anisotropic Certification

- **Theoretical foundation**: a simplified Lipschitz analysis

Proposition 1. *Consider a differentiable function $g : \mathbb{R}^n \to \mathbb{R}$. If $\sup_x \|\nabla g(x)\|_* \leq L$ where $\|\cdot\|_*$ has a dual norm $\|z\| = \max_x z^\top x$ s.t. $\|x\|_* \leq 1$, then g is L-Lipschitz under norm $\|\cdot\|_*$, that is $|g(x) - g(y)| \leq L\|x - y\|$.*

Given the previous proposition, we formalize $\|\cdot\|$ certification as follows:

Theorem 1. *Let $g : \mathbb{R}^n \to \mathbb{R}^K$, g^i be L-Lipschitz continuous under norm $\|\cdot\|_*$ $\forall i \in \{1, \ldots, K\}$, and $c_A = \arg\max_i g^i(x)$. Then, we have $\arg\max_i g^i(x + \delta) = c_A$ for all δ satisfying:*

$$\|\delta\| \leq \frac{1}{2L}\left(g^{c_A}(x) - \max_c g^{c \neq c_A}(x)\right).$$

- **Certifying ellipsoids**: $g_\Sigma(x) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\Sigma)}[f(x + \epsilon)]$

Let $c_A = \arg\max_i g_\Sigma^i(x)$, then $\arg\max_i g_\Sigma^i(x + \delta) = c_A$ for all δ satisfying:

$$\|\delta\|_{\Sigma,2} \leq \frac{1}{2}\left(\Phi^{-1}\left(g_\Sigma^{c_A}(x)\right) - \Phi^{-1}\left(\max_c g_\Sigma^{c \neq c_A}(x)\right)\right).$$

- **Certifying generalized cross-polytopes**: $g_\Lambda(x) = \mathbb{E}_{\epsilon \sim \mathcal{U}[-1,1]^n}[f(x + \Lambda\epsilon)]$

Let $c_A = \arg\max_i g_\Lambda^i(x)$, then $\arg\max_i g_\Lambda^i(x + \delta) = c_A$ for all δ satisfying:

$$\|\delta\|_{\Lambda,1} = \|\Lambda^{-1}\delta\|_1 \leq \left(g_\Lambda^{c_A}(x) - \max_c g_\Lambda^{c \neq c_A}(x)\right).$$

Evaluating Anisotropic Certificates

Given two general (isotropic or anisotropic) certificates, how can we compare them?

- **Superset argument**:

Definition 1. *For a given input point x, consider the two certification regions $\mathcal{R}_1$ and $\mathcal{R}_2$ obtained for two classifiers f_1 and f_2, i.e. $\mathcal{A}_1 = \{\delta : \arg\max_c f_1^c(x) = \arg\max_c f_1^c(x + \delta), \forall \delta \in \mathcal{R}_1\}$ and $\mathcal{A}_2 = \{\delta : \arg\max_c f_2^c(x) = \arg\max_c f_2^c(x + \delta), \forall \delta \in \mathcal{R}_2\}$ where $\arg\max_c f_1^c(x) = \arg\max_c f_2^c(x)$. We say $\mathcal{A}_1$ is a "superior certificate" to $\mathcal{A}_2$ (i.e. $\mathcal{A}_1 \succ \mathcal{A}_2$), if and only if, $\mathcal{A}_1 \supset \mathcal{A}_2$.*

- **Comparing ℓ_2-balls to ℓ_2^Σ-ellipsoids**: compare the isotropic certificate to the maximal enclosed ℓ_2-ball in the anisotropic one.
- **Quantifying ℓ_2^Σ certificates**: compare the nth root volume of the ellipsoid – or *proxy radius* for short - with the isotropic ℓ_2-ball.
- **Comparing ℓ_1-balls to ℓ_1^Λ-generalized cross-polytopes**: compare the isotropic certificate to the maximal enclosed ℓ_1-ball in the anisotropic one.
- **Quantifying ℓ_1^Λ certificates**: compare the nth root volume of the generalized cross-polytope – or *proxy radius* for short - with the isotropic ℓ_1-ball.

ANCER

How can we choose the anisotropic Σ and Λ?

- Memory-based sample-wise certification based on [1]: maximize the certified region's *proxy radius*, while keeping it a superset of the isotropic one.

$$\arg\max_{\Theta^x} \underbrace{r^p(x, \Theta^x)\sqrt[n]{\prod_i \theta_i^x}}_{\textit{Volume per sample (Proxy radius)}} \quad \text{s.t.} \quad \underbrace{\min_i \theta_i^x r^p(x, \Theta^x) \geq r_{\text{iso}}^*}_{\textit{Superset of isotropic}}$$

- Memory-based certification as in [1] for soundness purposes.

Pre-print: Code:

Results

- **Comparison metrics**:
  - certified accuracy @ radius R (% of dataset correctly classified and w/radius at least R) and average certified radius (ACR)
  - certified accuracy @ proxy radius R (% of dataset correctly classified and w/proxy radius at least R) and average certified proxy radius ($AC\tilde{R}$)
- **Ellipsoid certification results**: comparison with Cohen [2], SmoothAdv [3] and MACER [4] on CIFAR-10 and ImageNet.

CIFAR-10	Certification	Accuracy @ ℓ_2 radius (%) 0.0	0.25	0.5	1.0	1.5	2.0	2.5	ℓ_2 ACR	ℓ_2^Σ $AC\tilde{R}$
COHEN Cohen et al. (2019)	Fixed σ	86	71	51	27	14	6	2	0.722	0.722
	Isotropic DD	82	76	62	39	24	14	8	1.117	1.117
	ANCER	86	85	77	53	31	17	10	**1.449**	**1.772**
SMOOTHADV Salman et al. (2019a)	Fixed σ	82	72	55	32	19	9	5	0.834	0.834
	Isotropic DD	82	75	63	40	25	15	7	1.011	1.011
	ANCER	83	81	73	48	30	17	8	**1.224**	**1.573**
MACER Zhai et al. (2019)	Fixed σ	87	76	59	37	24	14	9	0.970	0.970
	Isotropic DD	88	80	66	40	17	9	6	1.007	1.007
	ANCER	84	80	67	34	15	11	9	**1.136**	**1.481**

ImageNet	Certification	Accuracy @ ℓ_2 radius (%) 0.0	0.5	1.0	1.5	2.0	2.5	3.0	ℓ_2 ACR	ℓ_2^Σ $AC\tilde{R}$
COHEN Cohen et al. (2019)	Fixed σ	70	56	41	31	19	14	12	1.098	1.098
	Isotropic DD	71	59	46	36	24	19	15	1.234	1.234
	ANCER	70	70	62	61	42	36	29	**1.810**	**1.981**
SMOOTHADV Salman et al. (2019a)	Fixed σ	65	59	44	38	26	20	18	1.287	1.287
	Isotropic DD	66	62	53	41	32	24	20	1.428	1.428
	ANCER	66	66	62	58	44	37	32	**1.807**	**1.965**

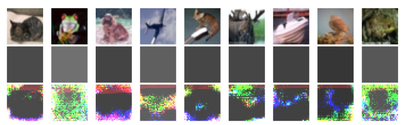
- **Generalized cross-polytope certification results**: comparison with RS4A [5] on CIFAR-10 and ImageNet.

CIFAR-10	Certification	Accuracy @ ℓ_1 radius (%) 0.0	0.25	0.5	0.75	1.0	1.5	2.0	ℓ_1 ACR	ℓ_1^Λ $AC\tilde{R}$
RS4A Yang et al. (2020)	Fixed σ	92	83	75	71	46	0	0	0.775	0.775
	Isotropic DD	92	89	82	76	58	6	2	0.946	0.946
	ANCER	92	90	84	80	63	6	2	**0.980**	**1.104**
ImageNet										
RS4A Yang et al. (2020)	Fixed σ	78	73	67	63	0	0	0	0.683	0.683
	Isotropic DD	79	76	70	65	46	0	0	0.729	0.729
	ANCER	78	76	70	66	48	0	0	**0.730**	**1.513**

- **(a) certification runtime comparison and (b) certificate visualization**

	Fixed σ	Isotropic DD	ANCER
(a)	1.6s	1.8s	2.7s
(b)	7.4s	9.5s	11.5s
(c)	109.5s	136.0s	147.0s

(a)



(b)

References
[1] Motasem Alfarra, *et al.* Data dependent randomized smoothing. *arXiv preprint*, 2020
[2] Jeremy M Cohen, *et al.* Certified adversarial robustness via randomized smoothing. In *ICML*, 2019
[3] Hadi Salman, *et al.* Provably robust deep learning via adversarially trained smoothed classifiers. In *NeurIPS*, 2019
[4] Runtian Zhai, *et al.* Macer: Attack-free and scalable robust training via maximizing certified radius. In *ICLR*, 2019
[5] Greg Yang, *et al.* Randomized smoothing of all shapes and sizes. In *ICML*, 2020