# An information geometry approach to Randomized smoothing

Pol Labarbarie (IRT SystemX)

Marc Arnaudon (IMB)

Hatem Hajri (IRT SystemX)

# Neural nets are not robust

- Invisible perturbations which break network behavior!

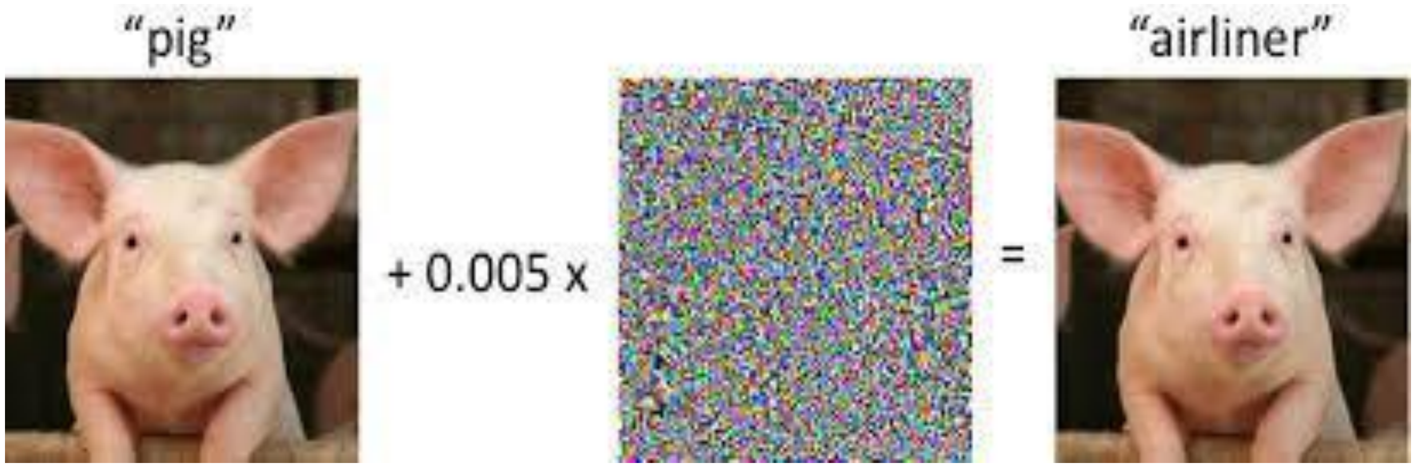



Illustration of an adversarial example (Szegedy et al. 2014).

# Making neural nets robust

- Empirical defenses
  - Only tell you if a specific attack is successful, no provable guarantees
  - Adversarial training

- Certified defenses
  - Aim to provide a guarantee in a specific neighborhood
  - MILP, formal methods, …

  ! Assume specific network architecture!

  - Randomized smoothing: no assumption on the network architecture

# Randomized Smoothing (Cohen et al. 2019, Salman et al. 2019a)

- Given a soft classifier F : $\mathbb{R}^d \to P(\mathcal{Y})$, the associated smoothed classifier is given by

$$G_F(x) = \left(F * \mathcal{N}(0, \sigma^2 I_d)\right)(x) = \mathbb{E}_{\varepsilon \sim \mathcal{N}(0,\sigma^2 I_d)}[F(x+\varepsilon)]$$

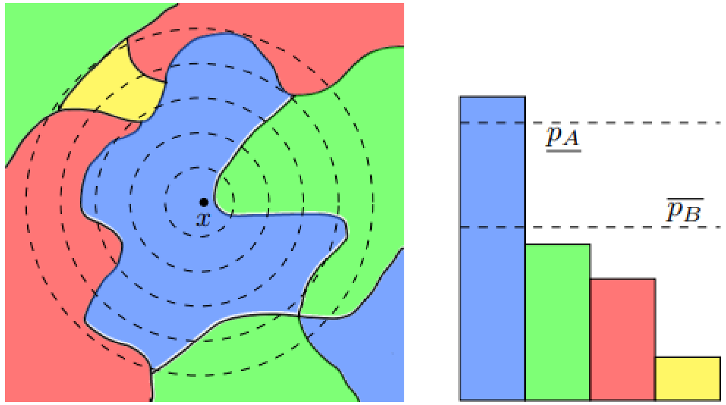



Illustration of the randomized smoothing (Cohen et al. 2019).

# R.S yields certifiable $\ell_2$ robustness

**Theorem (Cohen et al. 2019):** Let F be a soft classifier, g its smoothing with $\mathcal{N}(0, \sigma^2 I_d)$ and $x \in \mathbb{R}^d$. Let

$$a = \operatorname{argmax}_{c \in \mathcal{Y}} G_F(x)_c, \qquad p_a = G_F(x)_a$$
$$b = \operatorname{argmax}_{c \in \mathcal{Y}, c \neq a} G_F(x)_c, \qquad p_b = G_F(x)_b$$

Then $G_F$ is robust at $x$ for any $\ell_2$ perturbation of size

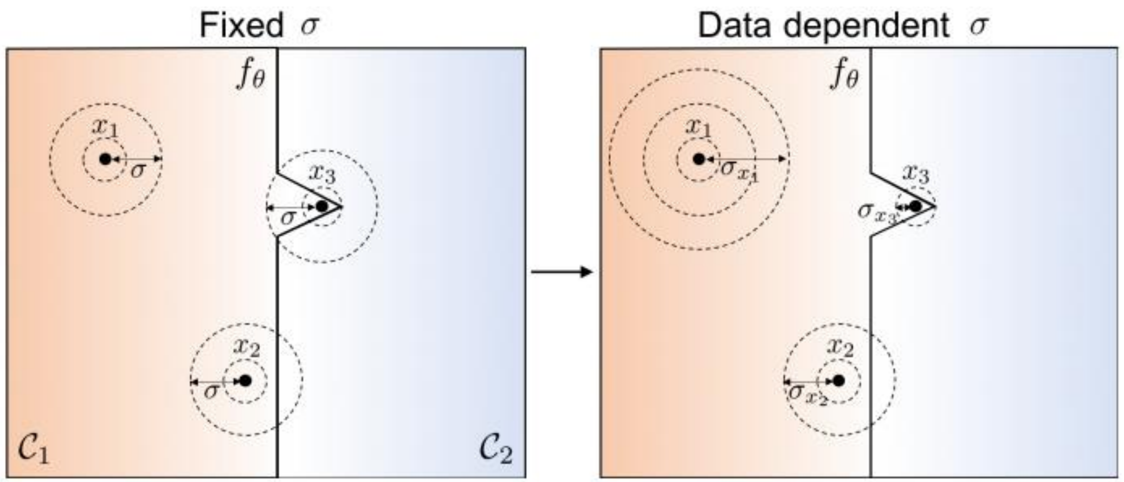
$$R = \frac{\sigma^2}{2}\left(\phi^{-1}(p_a) - \phi^{-1}(p_b)\right)$$

# Data-dependent R.S (Alfarra et al. 2020)

## Idea:

- RS fixes $\sigma$ and $x$, then finds $R = R(x, \sigma)$
- Data dependent RS only fixes $x$, and then optimizes to find

$$\sigma_x^* = \text{argmax}_{\sigma_x} R(x, \sigma_x)$$

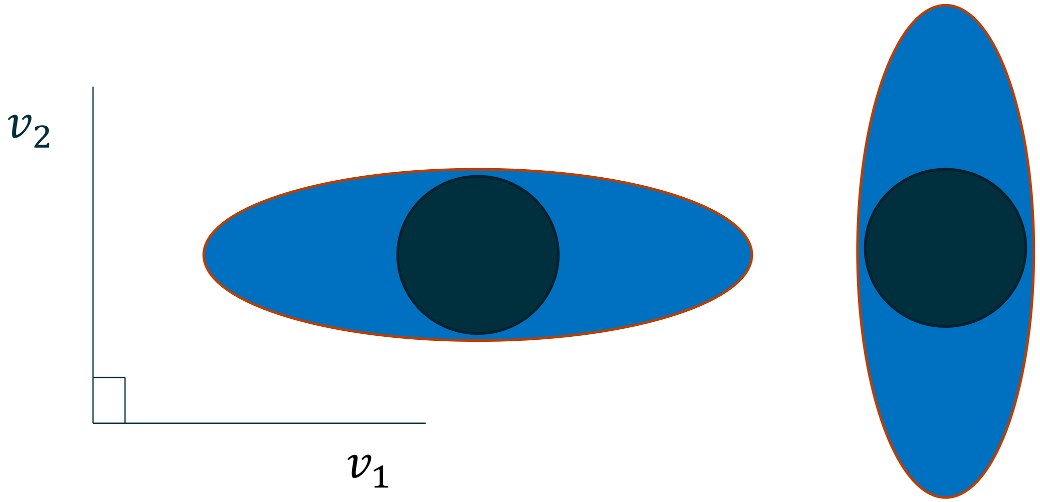



From fixed to data-dependent RS (Alfarra et al. 2020).

# ANCER : ANISOTROPIC CERTIFICATION VIA SAMPLE-WISE VOLUME MAXIMIZATION (Eiras et al. 2020)

## Idea:

- Smooth with an anisotropic gaussian noise
- Solved $\Sigma_x^* = \text{argmax}_{\Sigma_x} R(x, \Sigma_x)$ where $\Sigma_x$ is diagonal

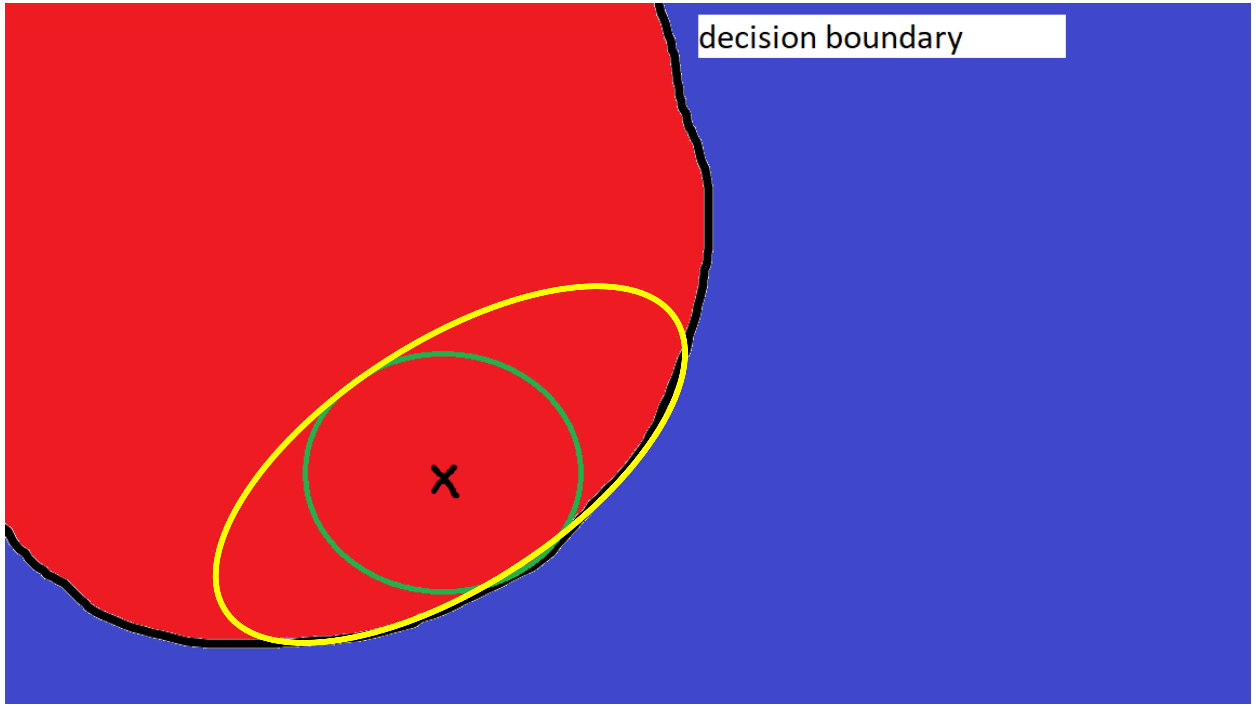



Comparison between ANCER and Data-dependent R.S.

**The certified regions by ANCER are ellipsoids containing the balls obtained by RS: all axis are parallel to the canonical ones.**

# An information geometry approach (joint work with Marc Arnaudon and Hatem Hajri)

Idea: Smoothing by a diagonal noise is not optimal and not invariant under rotation



# An information geometry approach

- $M = \{\Sigma, \Sigma \; is \; SPD\}$ is a Riemannian manifold
- Solve $\Sigma_x^* = \operatorname{argmax}_{\Sigma_x} R(x, \Sigma_x)$ on this manifold.

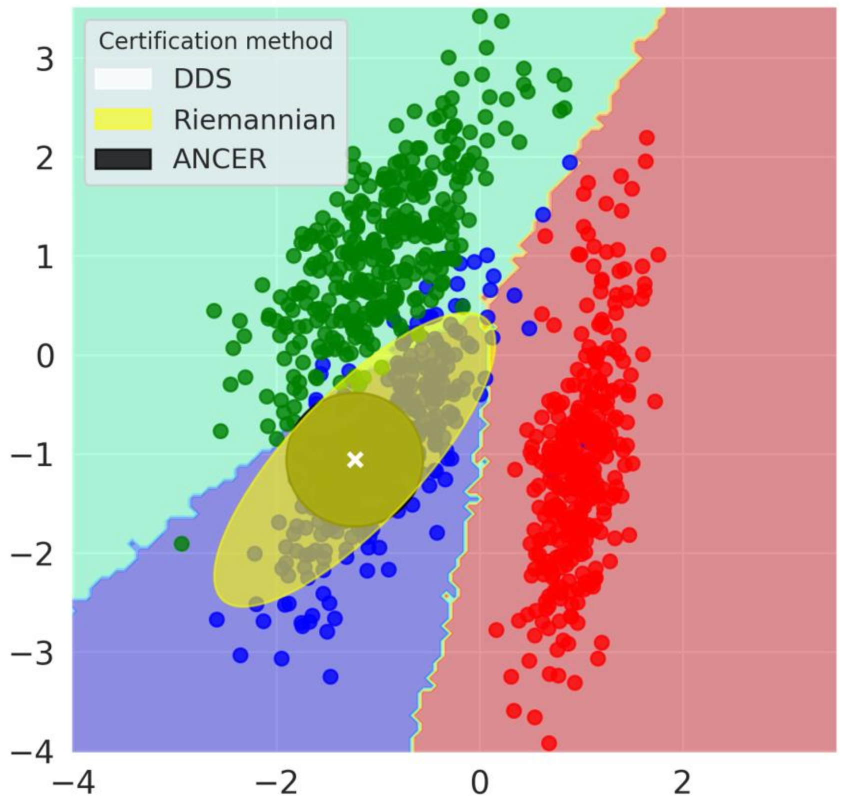
Resolution:

- Information-geometry metric: Rao-Fisher

$$d^2(Y, Z) = \operatorname{tr}\,[\log(Y^{-1/2} Z Y^{-1/2})]^2.$$

- Automatic differentiation of matrices: Geomstats Library (JLMR 2020)

# An information geometry approach




Comparison between methods for a 2D classification problem.

Parameters:

- Failure probability $\alpha = 0{,}001$

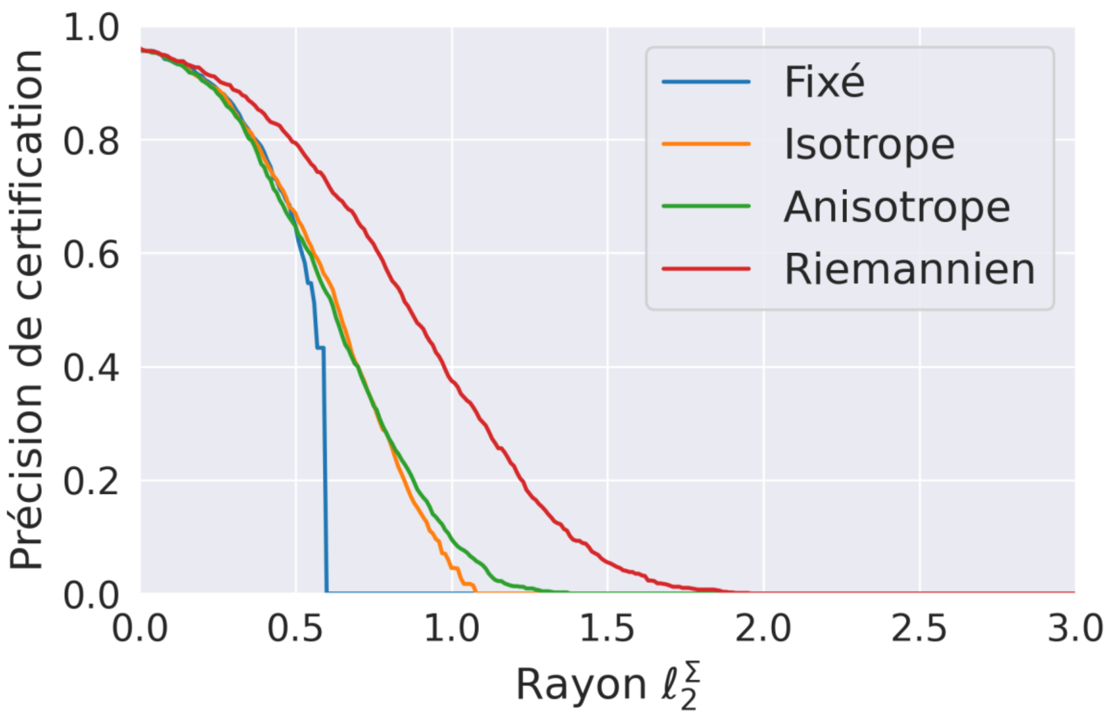
- $N = 100000$ samples

- 100 iterations for ANCER's method and our method

# An information geometry approach

## Results:

- For a fixed R>o, we evaluate the percentage of inputs $x$ such that $B_2(x, R)$ is included in the certified domain.



Confiance.ai

www.confiance.ai
contact@irt-systemx.fr