Canonical foliations of neural networks:
application to robustness

Eliot Tron (ENAC OPTIM), Nicolas Couellan (ENAC, IMT), Stéphane Puechmorel (ENAC, IMT).

Modeling

- A Neural network $\mathcal{N} : \mathcal{X} \to \mathcal{Y}$,
- associated with a probability distribution $p_\theta(z \mid x) = \mathcal{N}_\theta(x)_i$
- and a (semi-definite) metric: the Fisher Information Metric

$$g_{ij} = \mathbb{E}_{z|x}\left[\partial_{x_i} \log p_\theta(z \mid x)\, \partial_{x_j} \log p_\theta(z \mid x)\right].$$

Robustness

We want to reduce vulnerability to *adversarial attacks* defined for a budget $\varepsilon > 0$ as solutions to the following optimisation problem:

$$(AAP) \begin{cases} \max_{x_a \in \mathcal{X}} d_{\text{geo}}(x, x_a) \text{ s.t.} \\ d_{\text{human}}(x, x_a) < \varepsilon. \end{cases}$$

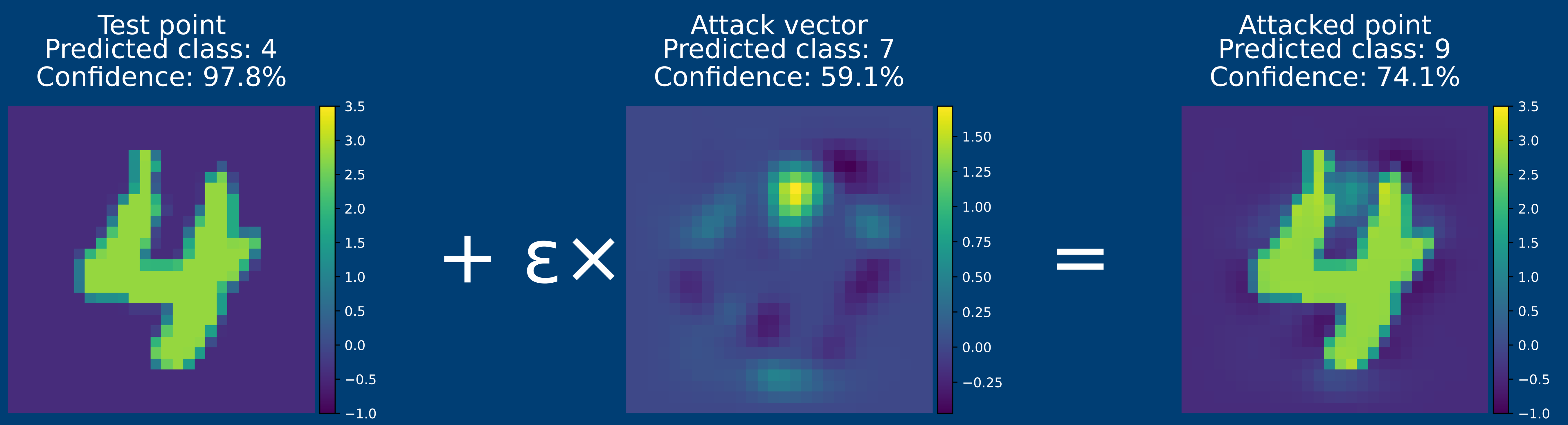
Two Step Spectral Attack

We approximate (AAP) with the sum of v, highest eigenvector of G_x, and w, solution of the following:

$$\max_w \|w\|^2_{\mathcal{X}} \text{ s.t.} \begin{cases} \|v\|_2 + \|w\|_2 \leq \varepsilon \\ \|v\|_2 = \mu < \varepsilon \\ v \text{ eigenvector of } G_x \end{cases}$$

The second step w helps by taking into account the curvature of $(\mathcal{X}, g)$.

What does the AI see?



Preprint on ArXiv: `https://arxiv.org/abs/2203.00922`

Curvature counts

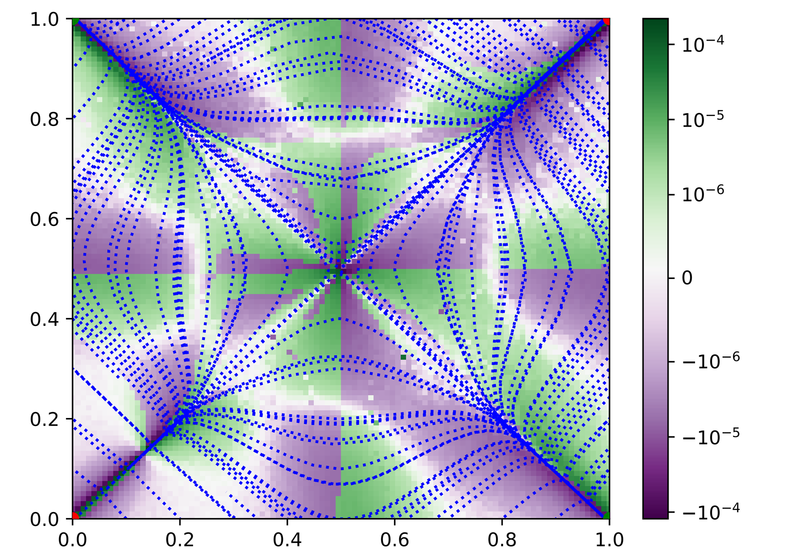


Figure 1: Xor kernel foliation.

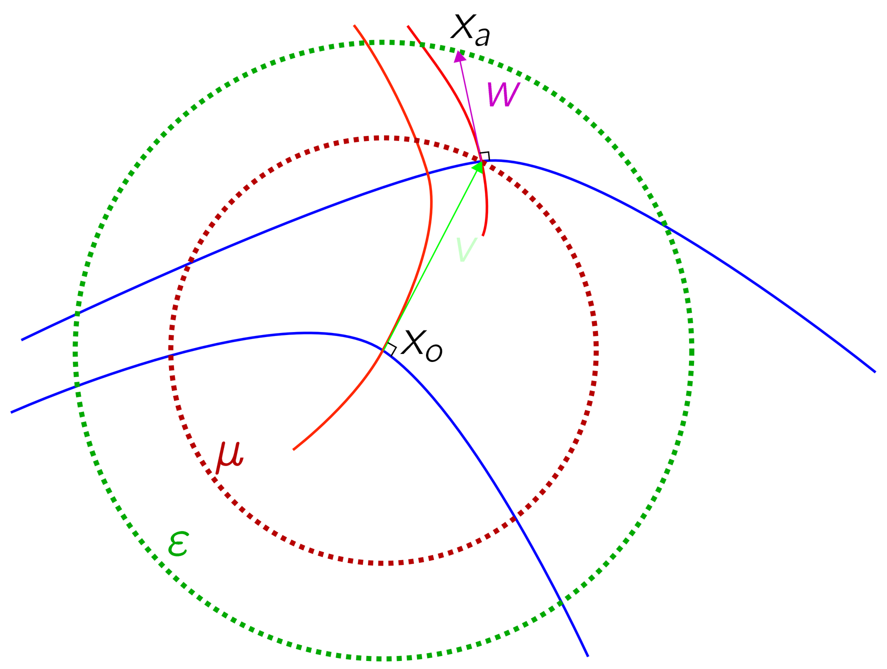


Figure 2: Two-step attack.

Experiments on MNIST

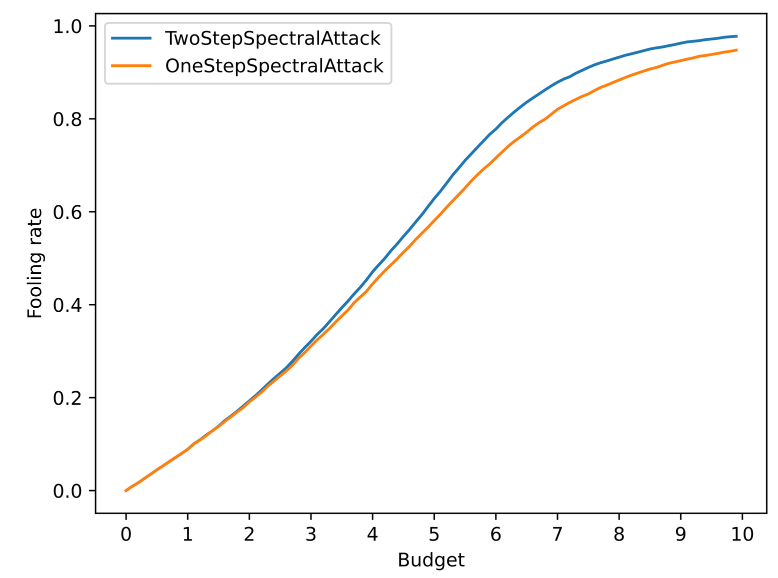


Figure 3: MNIST neural network.

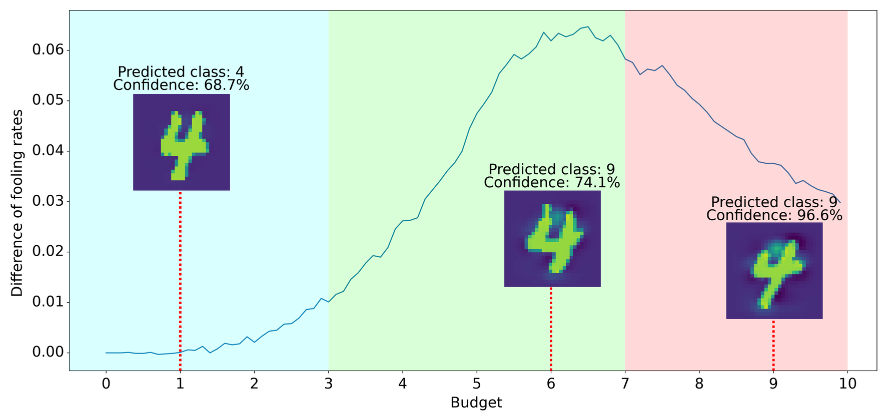


Figure 4: Difference TSSA - OSSA.