Certified defences hurt generalisation

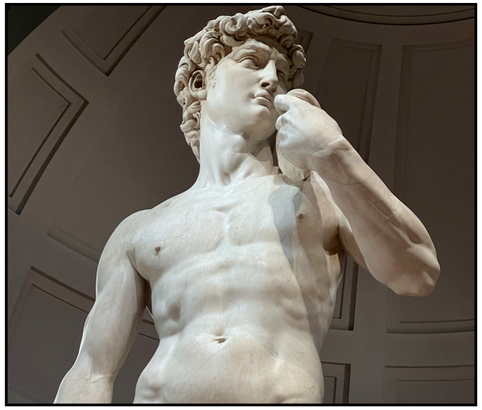
Piersilvio De Bartolomeis, joint work with J. Clarysse, F. Yang and A. Sanyal

ICBINB Workshop @ NeurIPS 2022

ETH zürich

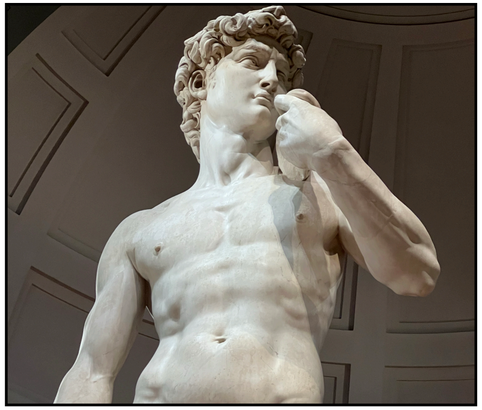
What are adversarial examples?

What are adversarial examples?

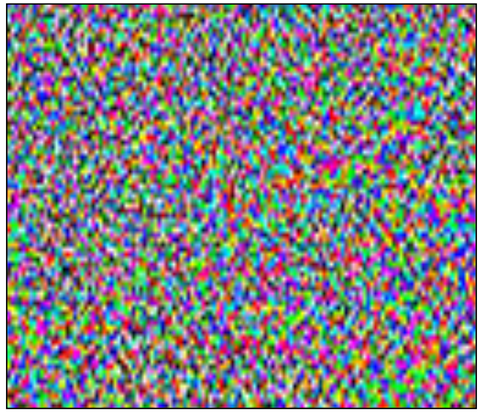


class: "Sculpture"

What are adversarial examples?

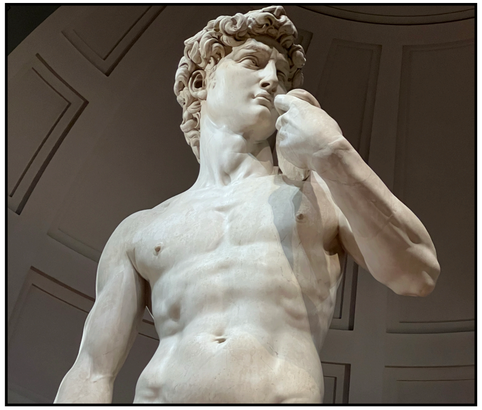


$+ \delta$

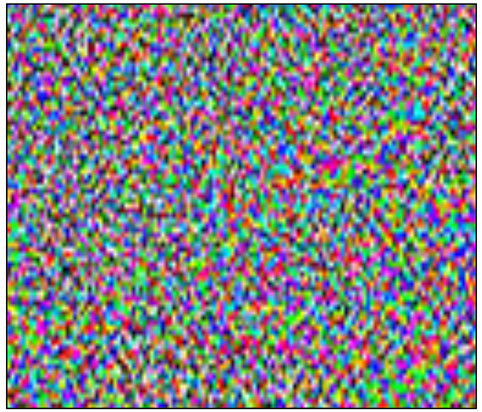


class: "Sculpture"

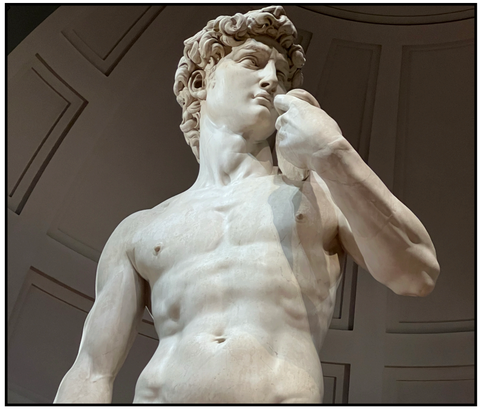
What are adversarial examples?



$+ \delta$



$=$



class: "Sculpture"

class: "Painting"

A formal definition

Given:

A formal definition

Given:

- A classifier $f_\theta : \mathbb{R}^d \to \mathcal{Y}$

A formal definition

Given:

- A classifier $f_\theta : \mathbb{R}^d \to \mathcal{Y}$
- A threat model $\mathcal{B}_\epsilon := \{\delta : \|\delta\|_p \leq \epsilon\}$

A formal definition

Given:

- A classifier $f_\theta : \mathbb{R}^d \to \mathcal{Y}$
- A threat model $\mathcal{B}_\epsilon := \{\delta : \|\delta\|_p \leq \epsilon\}$

An adversarial example for $x \in \mathbb{R}^d$ is a point $x + \delta$ for $\delta \in \mathcal{B}_\epsilon$ such that:

A formal definition

Given:

- A classifier $f_\theta : \mathbb{R}^d \to \mathcal{Y}$
- A threat model $\mathcal{B}_\epsilon := \{\delta : \|\delta\|_p \leq \epsilon\}$

An adversarial example for $x \in \mathbb{R}^d$ is a point $x + \delta$ for $\delta \in \mathcal{B}_\epsilon$ such that:

$$f_\theta(x) \neq f_\theta(x + \delta)$$

Robust risk minimisation

Given:

- A classifier $f_\theta : \mathbb{R}^d \to \mathcal{Y}$
- A threat model $\mathcal{B}_\epsilon := \{\delta : \|\delta\|_p \leq \epsilon\}$

Robust risk minimisation

Given:

- A classifier $f_\theta : \mathbb{R}^d \to \mathcal{Y}$
- A threat model $\mathcal{B}_\epsilon := \{\delta : \|\delta\|_p \leq \epsilon\}$
- A dataset $D = \{(x_i, y_i)\}_{i=1}^n$ where $(x_i, y_i) \in \mathbb{R}^d \times \mathcal{Y}$

Robust risk minimisation

Given:

- A classifier $f_\theta : \mathbb{R}^d \to \mathcal{Y}$
- A threat model $\mathcal{B}_\epsilon := \{\delta : \|\delta\|_p \leq \epsilon\}$
- A dataset $D = \{(x_i, y_i)\}_{i=1}^n$ where $(x_i, y_i) \in \mathbb{R}^d \times \mathcal{Y}$
- A loss function L

Robust risk minimisation

Given:

- A classifier $f_\theta : \mathbb{R}^d \to \mathcal{Y}$
- A threat model $\mathcal{B}_\epsilon := \{\delta : \|\delta\|_p \leq \epsilon\}$
- A dataset $D = \{(x_i, y_i)\}_{i=1}^n$ where $(x_i, y_i) \in \mathbb{R}^d \times \mathcal{Y}$
- A loss function L

$$\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{B}_\epsilon} L(f_\theta(x+\delta), y) \right]$$

Robust risk minimisation

Given:

- A classifier $f_\theta : \mathbb{R}^d \to \mathcal{Y}$
- A threat model $\mathcal{B}_\epsilon := \{\delta : \|\delta\|_p \leq \epsilon\}$
- A dataset $D = \{(x_i, y_i)\}_{i=1}^n$ where $(x_i, y_i) \in \mathbb{R}^d \times \mathcal{Y}$
- A loss function L

$$\min_\theta \sum_{(x,y)\in D} \max_{\delta\in\mathcal{B}_\epsilon} L(f_\theta(x+\delta), y)$$

Solving the inner-maximisation is NP-Hard

$$\min_{\theta} \sum_{(x,y)\in D} \max_{\delta\in\mathcal{B}_\epsilon} L(f_\theta(x+\delta), y)$$

Solving the inner-maximisation is NP-Hard

$$\min_{\theta} \sum_{(x,y)\in D} \max_{\delta\in\mathcal{B}_\epsilon} L(f_\theta(x+\delta), y)$$

Solving the inner-maximisation is NP-Hard

$$\min_{\theta} \sum_{(x,y)\in D} \max_{\delta\in\mathcal{B}_\epsilon} L(f_\theta(x+\delta), y)$$

NP-Hard

Solving the inner-maximisation is NP-Hard

$$\min_{\theta} \sum_{(x,y)\in D} \max_{\delta\in\mathcal{B}_\epsilon} L(f_\theta(x+\delta), y)$$

NP-Hard

- Lower bound: empirical defences

Solving the inner-maximisation is NP-Hard

$$\min_{\theta} \sum_{(x,y)\in D} \boxed{\max_{\delta\in\mathcal{B}_\epsilon} L(f_\theta(x+\delta), y)}$$

NP-Hard

- Lower bound: empirical defences $\Longrightarrow$ works in practice but no guarantees

Solving the inner-maximisation is NP-Hard

$$\min_{\theta} \sum_{(x,y)\in D} \max_{\delta\in\mathcal{B}_\epsilon} L(f_\theta(x+\delta), y)$$

NP-Hard

- Lower bound: empirical defences $\Longrightarrow$ works in practice but no guarantees
- Upper bound: certified defences

Solving the inner-maximisation is NP-Hard

$$\min_{\theta} \sum_{(x,y)\in D} \max_{\delta\in\mathcal{B}_\epsilon} L(f_\theta(x+\delta), y)$$

NP-Hard

- Lower bound: empirical defences $\Longrightarrow$ works in practice but no guarantees
- Upper bound: certified defences $\Longrightarrow$ provably robust on the training data

Solving the inner-maximisation is NP-Hard

$$\min_{\theta} \sum_{(x,y)\in D} \max_{\delta\in\mathcal{B}_\epsilon} L(f_\theta(x+\delta), y)$$

NP-Hard

- Lower bound: empirical defences $\Longrightarrow$ works in practice but no guarantees
- Upper bound: certified defences $\Longrightarrow$ provably robust on the training data

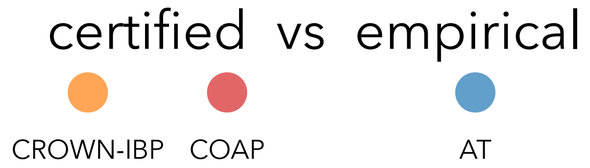
This talk: How effective are certified defences in practice?

Certified defences hurt generalisation

ℓ_2-ball perturbations

Certified defences hurt generalisation

ℓ_2-ball perturbations



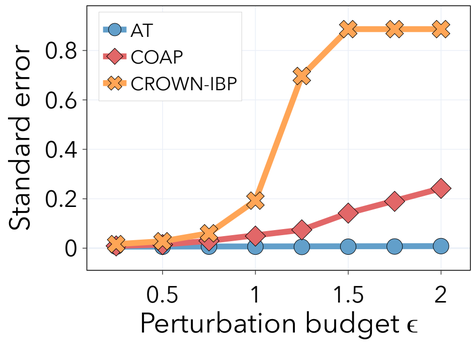
Certified defences hurt generalisation

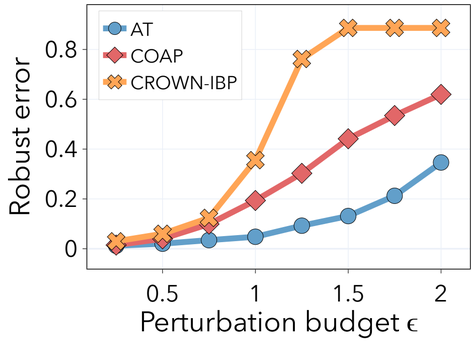
ℓ_2-ball perturbations

certified vs empirical

CROWN-IBP COAP AT

MNIST





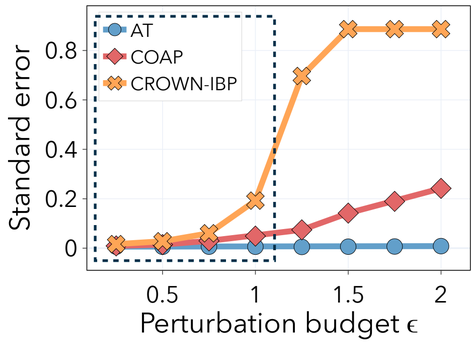
Certified defences hurt generalisation

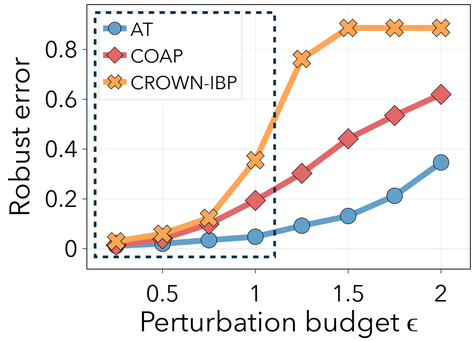
ℓ_2-ball perturbations

certified vs empirical

CROWN-IBP COAP AT

MNIST





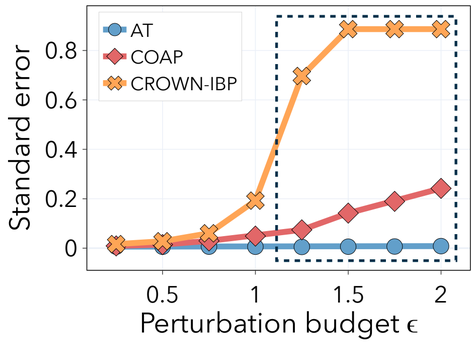
Certified defences hurt generalisation

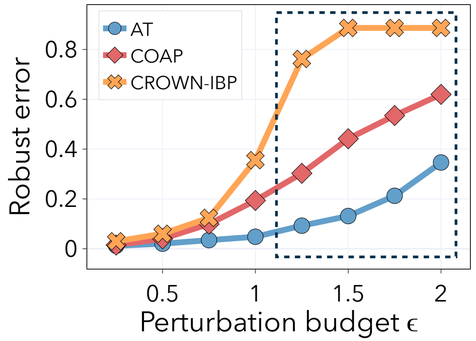
ℓ_2-ball perturbations

certified vs empirical

CROWN-IBP COAP AT

MNIST





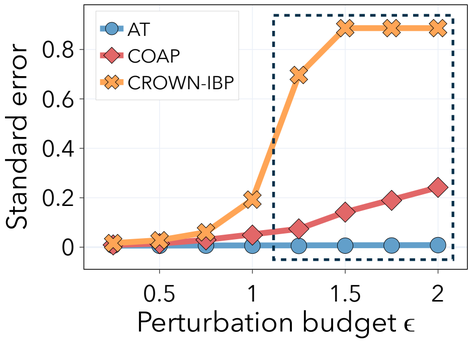
Certified defences hurt generalisation

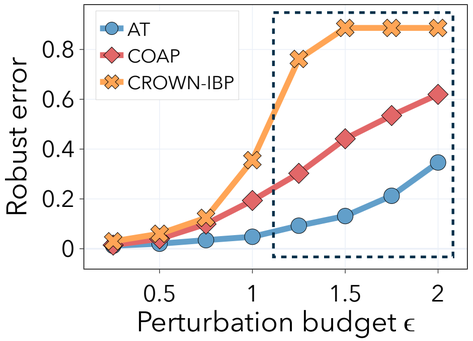
ℓ_2-ball perturbations

certified vs empirical

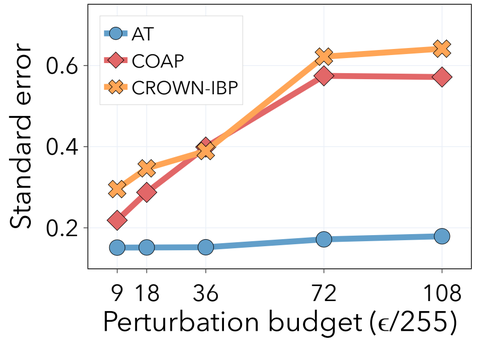
CROWN-IBP COAP AT

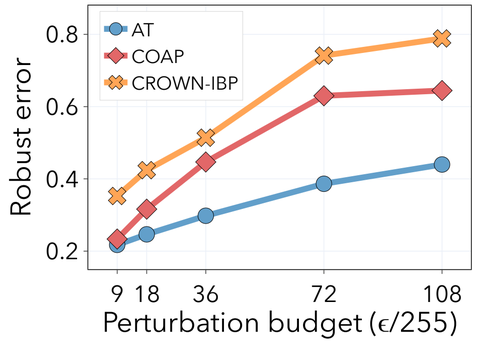
MNIST





CIFAR-10





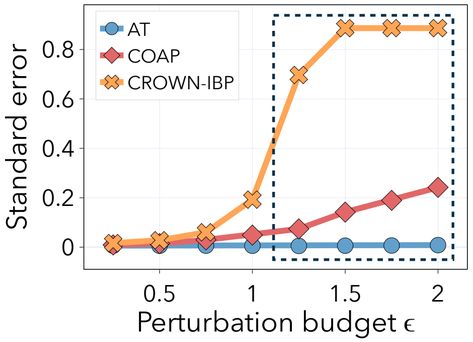
Certified defences hurt generalisation

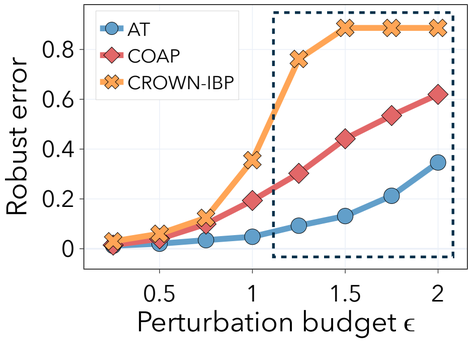
ℓ_2-ball perturbations

certified vs empirical

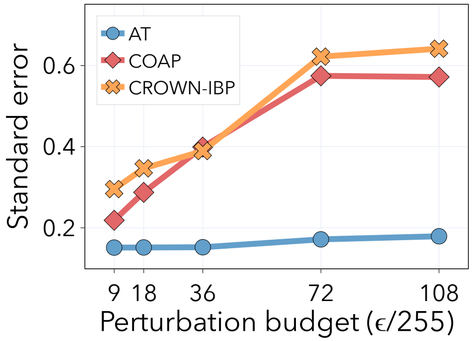
CROWN-IBP COAP AT

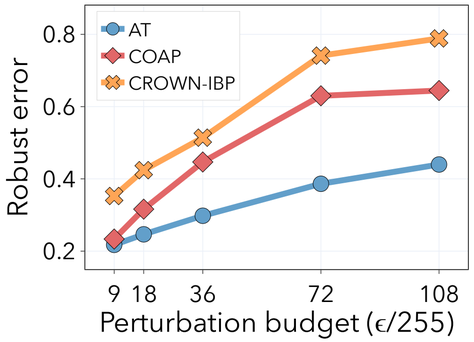
MNIST



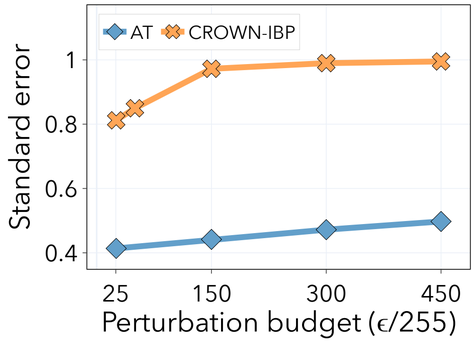


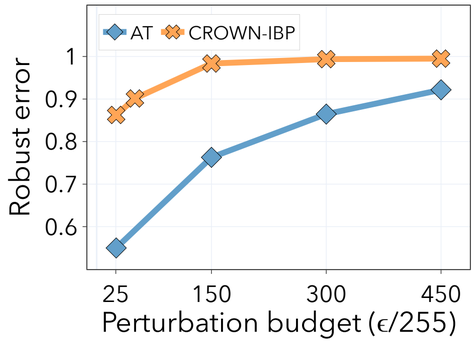
CIFAR-10



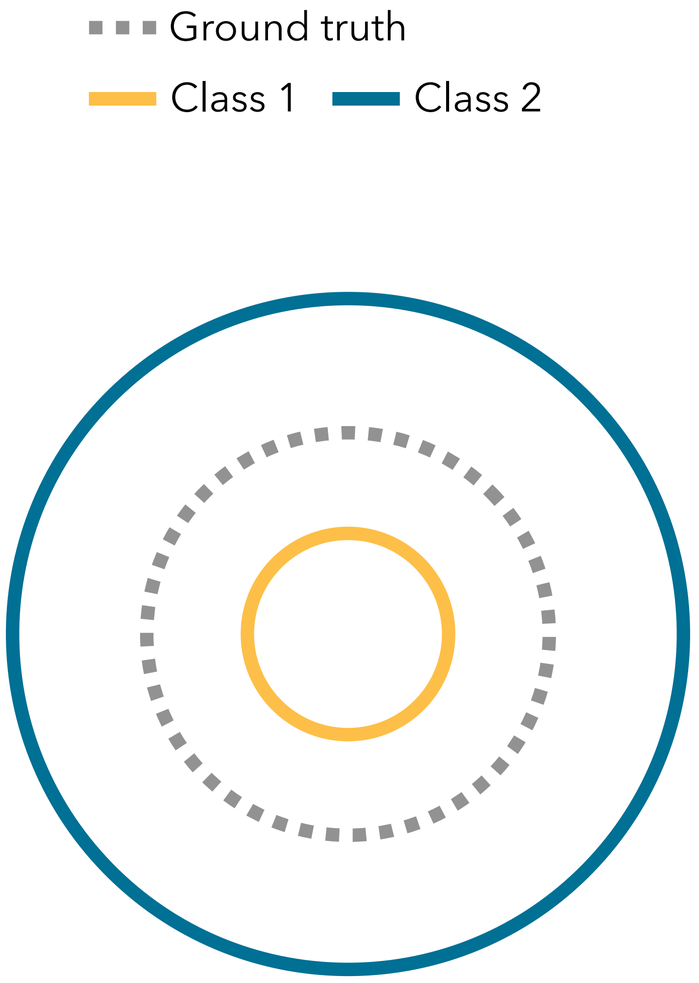


Tiny ImageNet

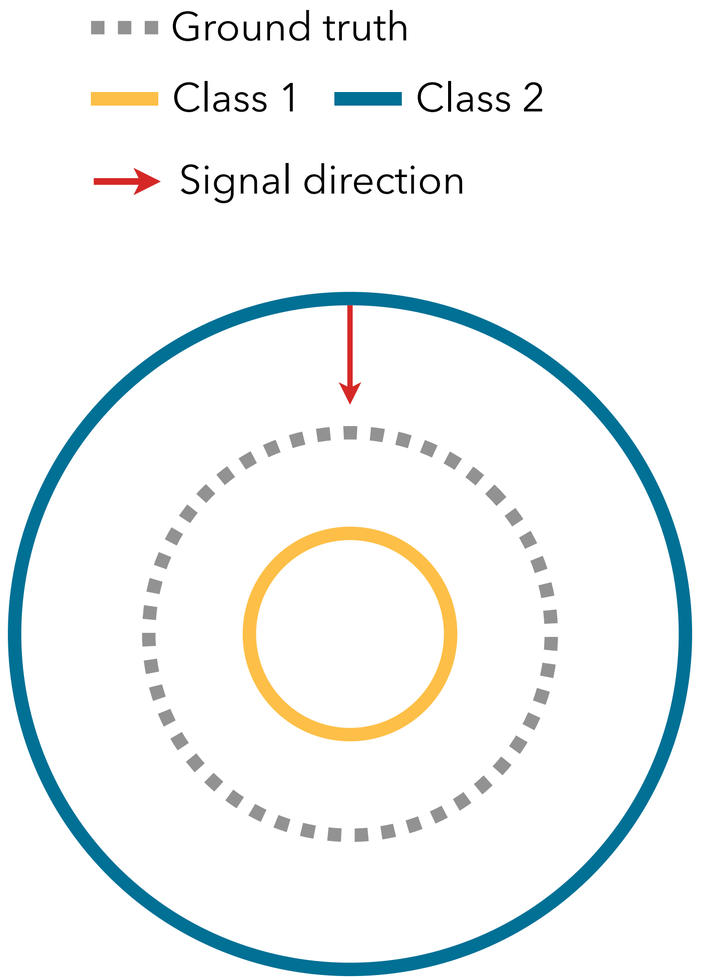




Why do certified defences hurt generalisation?

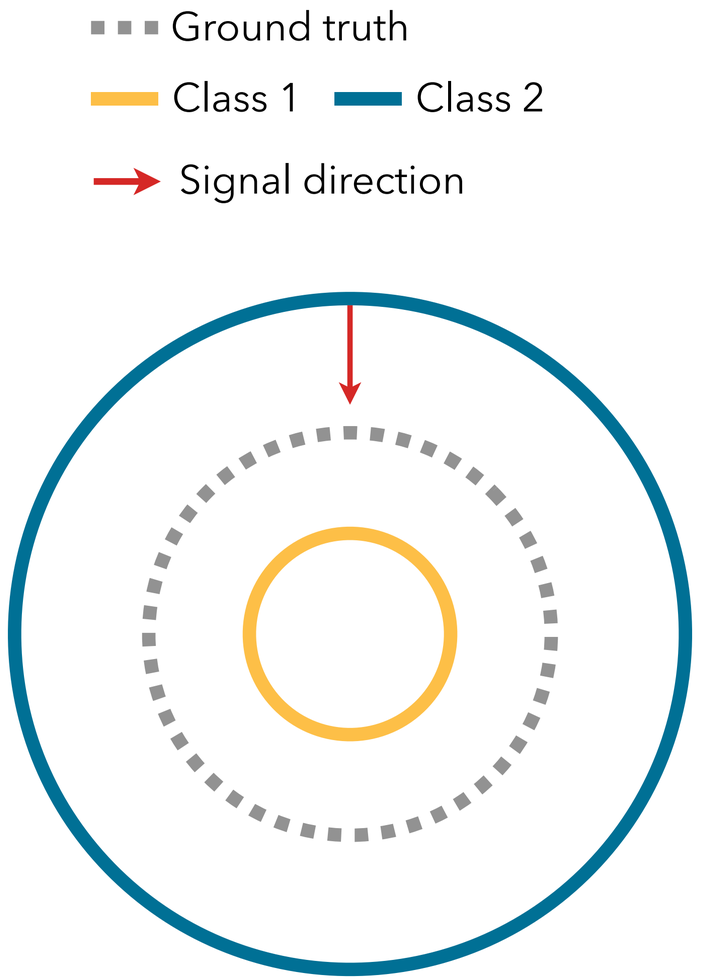


Why do certified defences hurt generalisation?



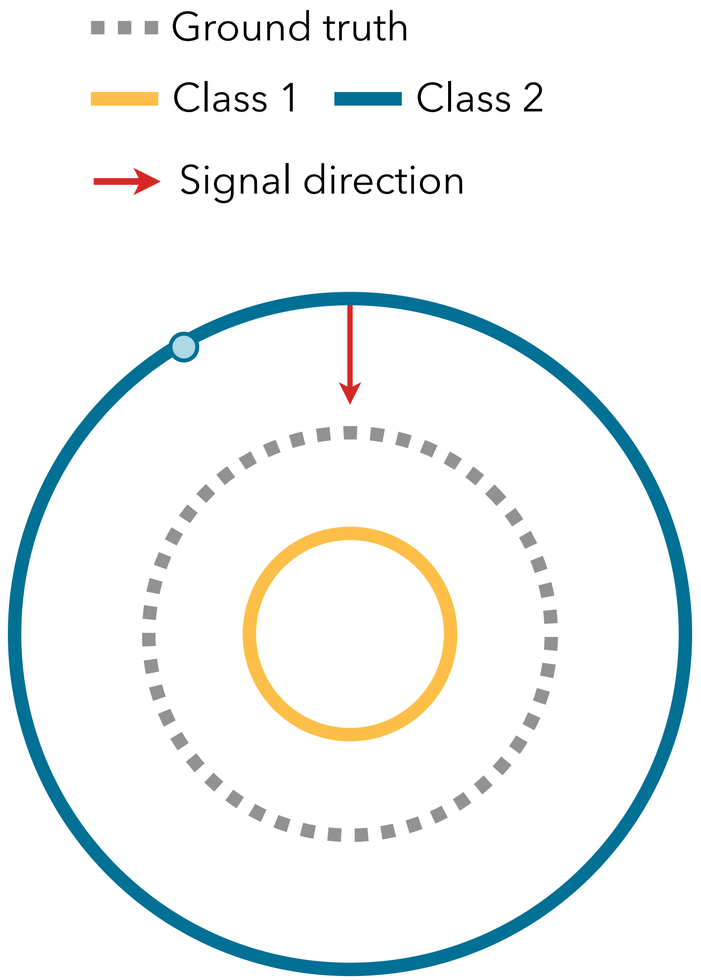
Why do certified defences hurt generalisation?

- (i) Magnitude of the perturbation



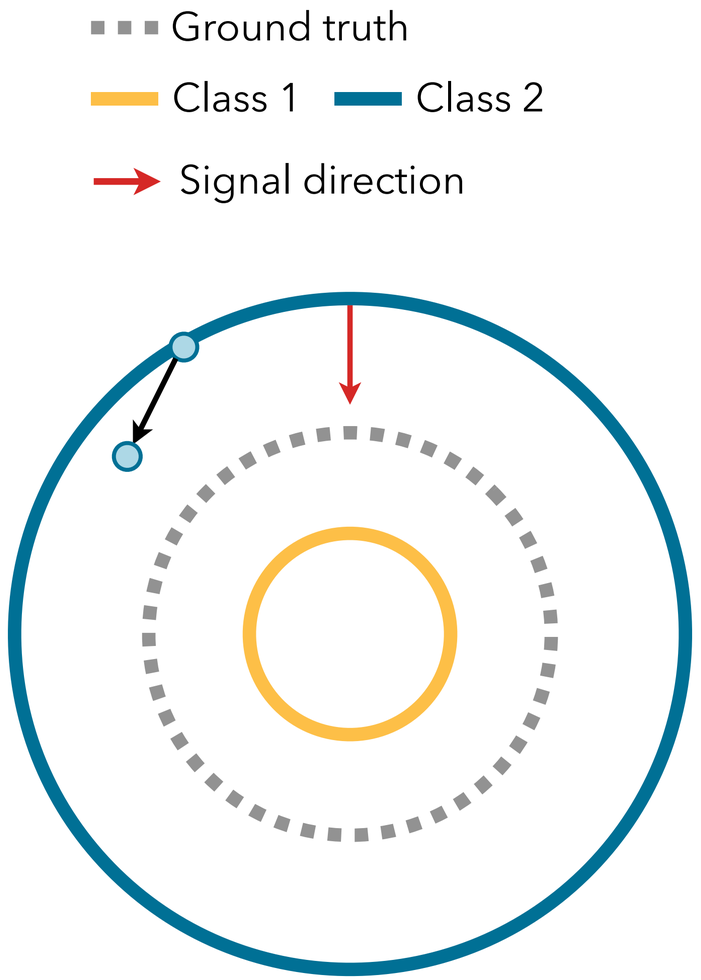
Why do certified defences hurt generalisation?

- (i) Magnitude of the perturbation



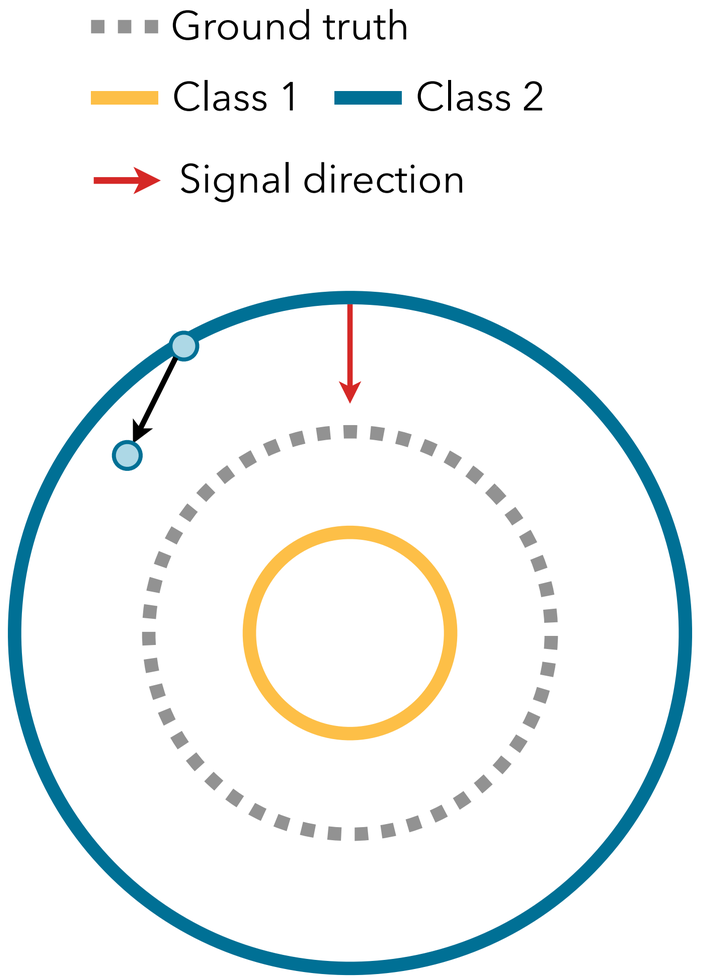
Why do certified defences hurt generalisation?

- (i) Magnitude of the perturbation



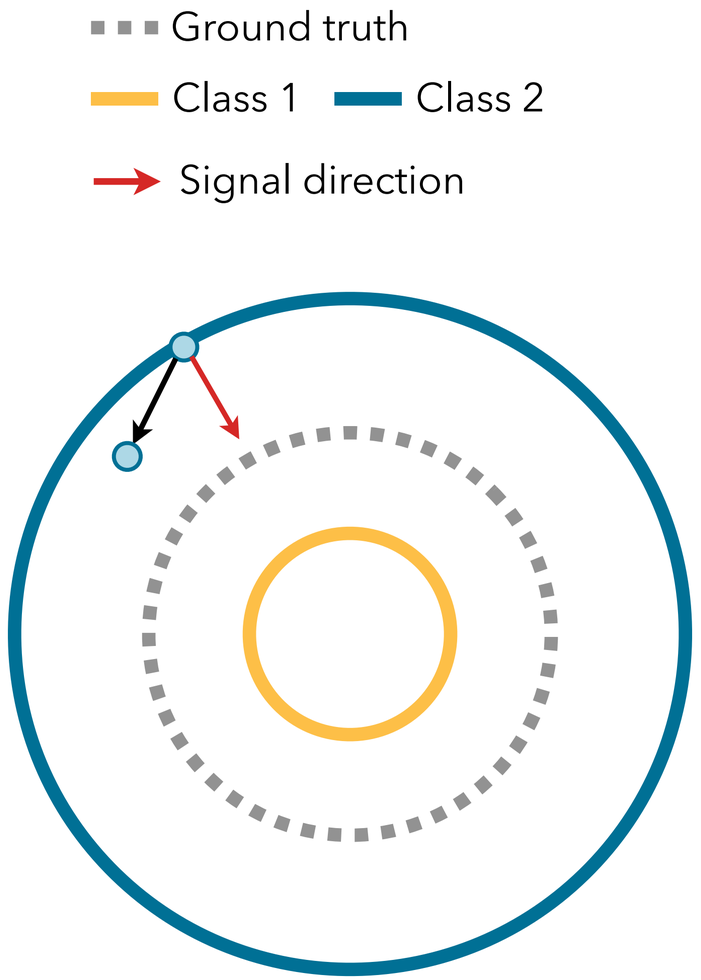
Why do certified defences hurt generalisation?

- (i) Magnitude of the perturbation
- (ii) Alignment with the signal direction



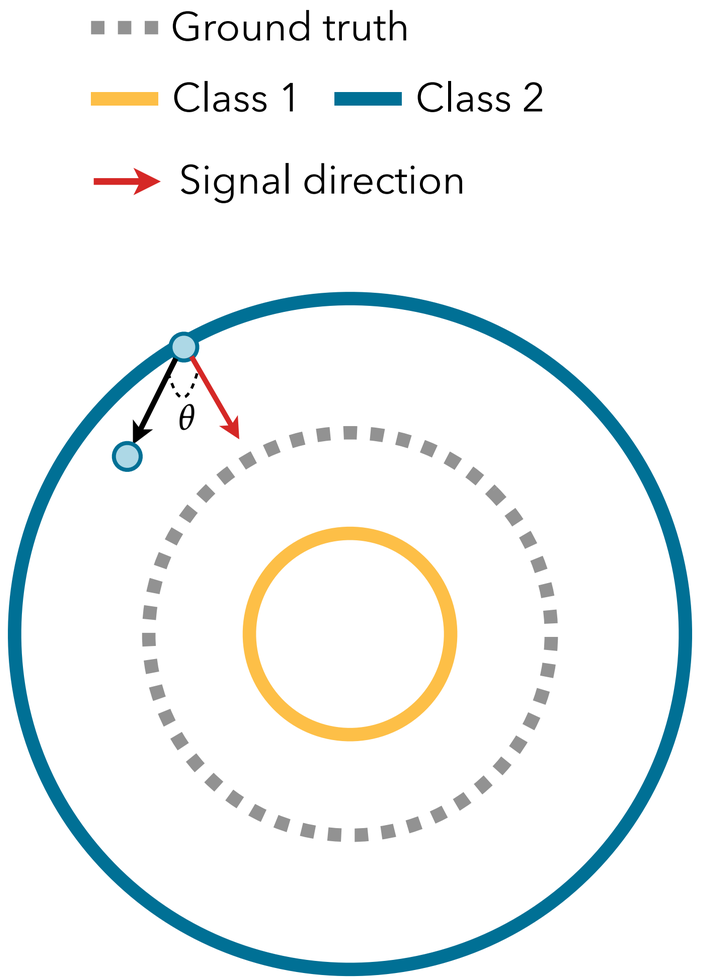
Why do certified defences hurt generalisation?

- (i) Magnitude of the perturbation
- (ii) Alignment with the signal direction

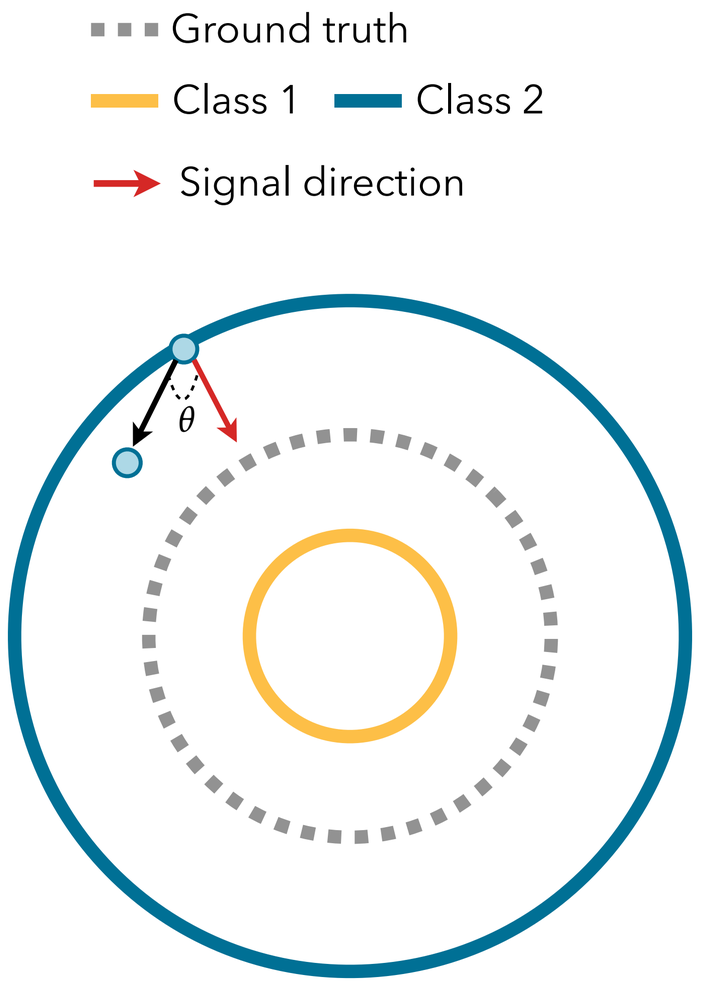


Why do certified defences hurt generalisation?

- (i) Magnitude of the perturbation
- (ii) Alignment with the signal direction



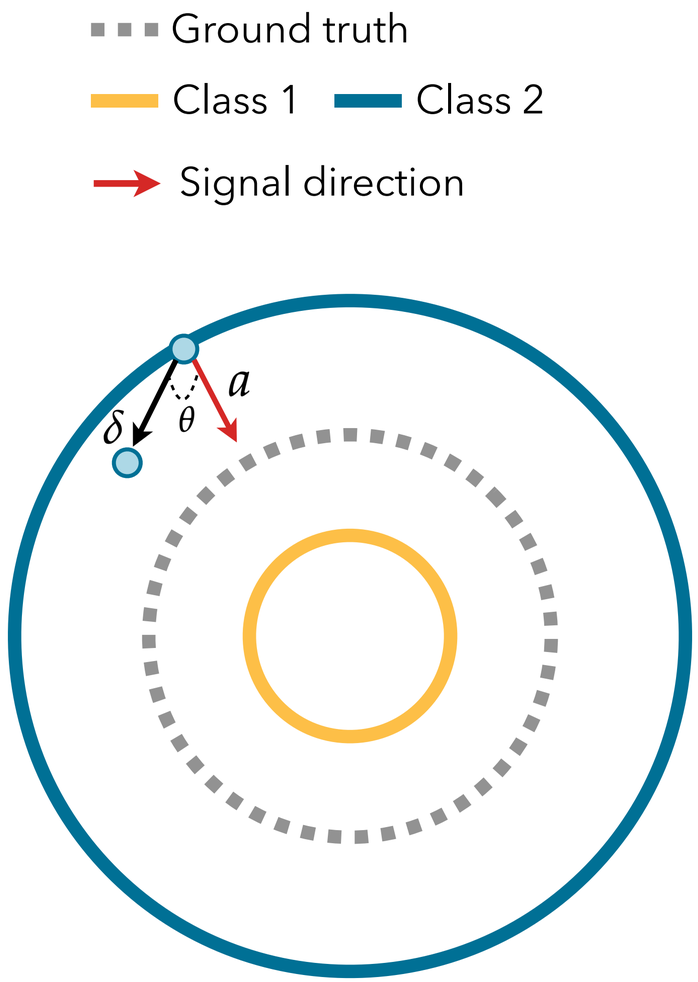
Why do certified defences hurt generalisation?



Why do certified defences hurt generalisation?

$$\frac{a \cdot \delta}{\|a\| \, \|\delta\|}$$

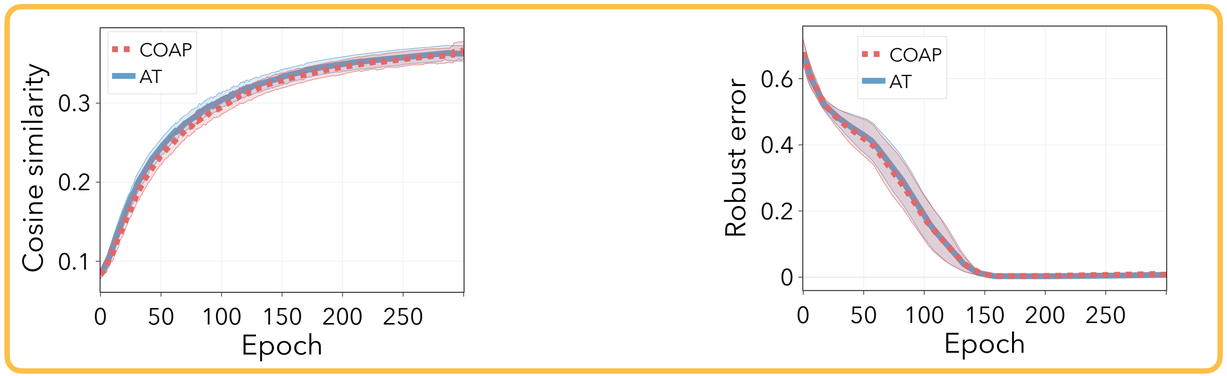
Cosine similarity



Why do certified defences hurt generalisation?

Small ϵ

$$\frac{a \cdot \delta}{\|a\| \, \|\delta\|}$$

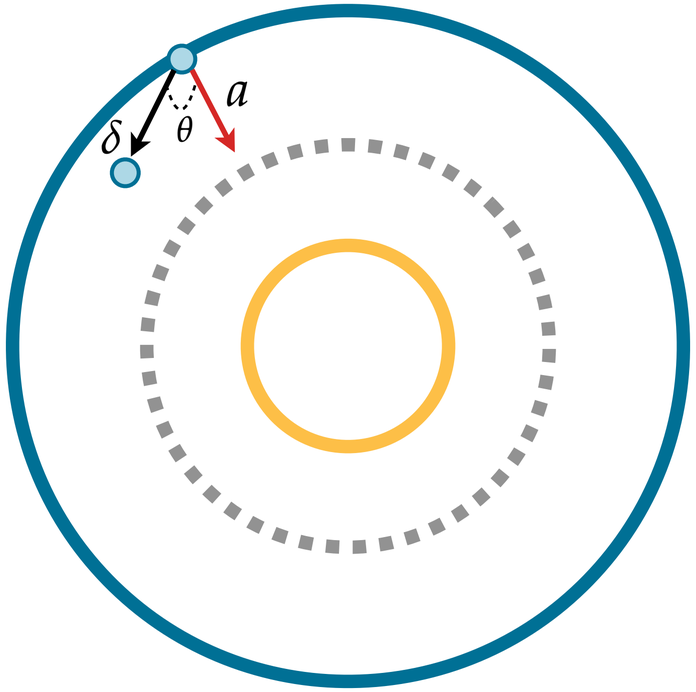


certified vs empirical

Ground truth

Class 1 Class 2

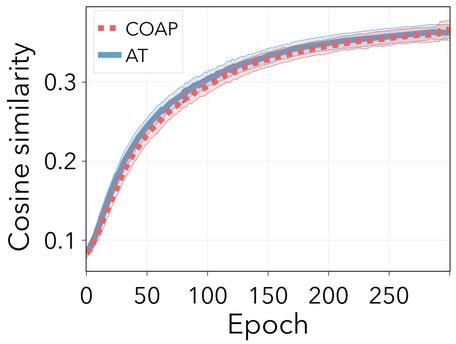
Signal direction

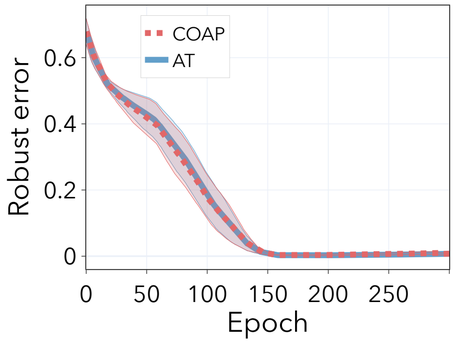


Why do certified defences hurt generalisation?

$$\frac{a \cdot \delta}{\|a\| \, \|\delta\|}$$

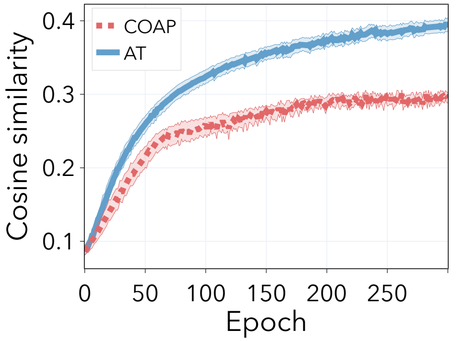
Small ϵ

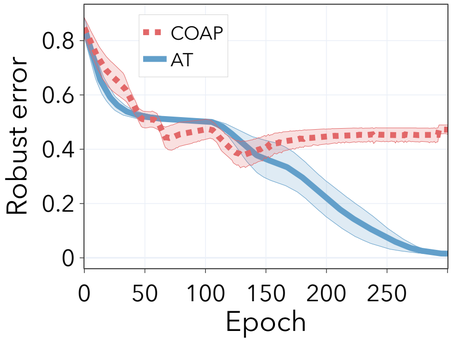




Large ϵ

certified vs empirical

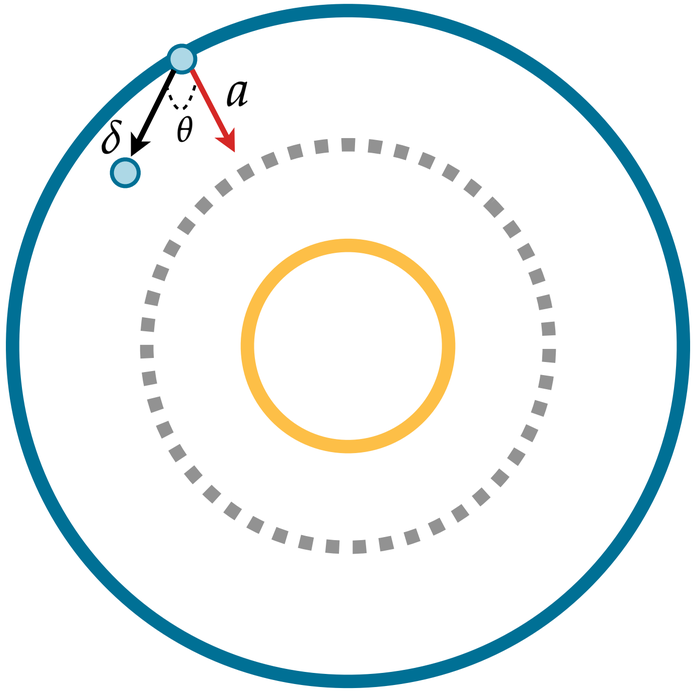




Ground truth

Class 1 Class 2

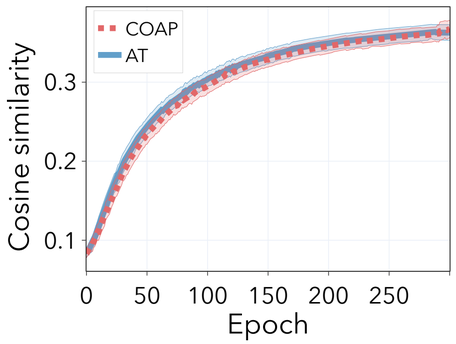
Signal direction

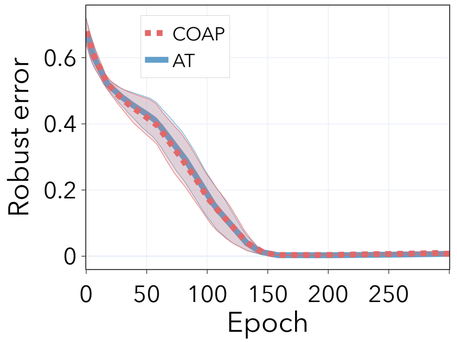


Why do certified defences hurt generalisation?

$$\frac{a \cdot \delta}{\|a\| \, \|\delta\|}$$

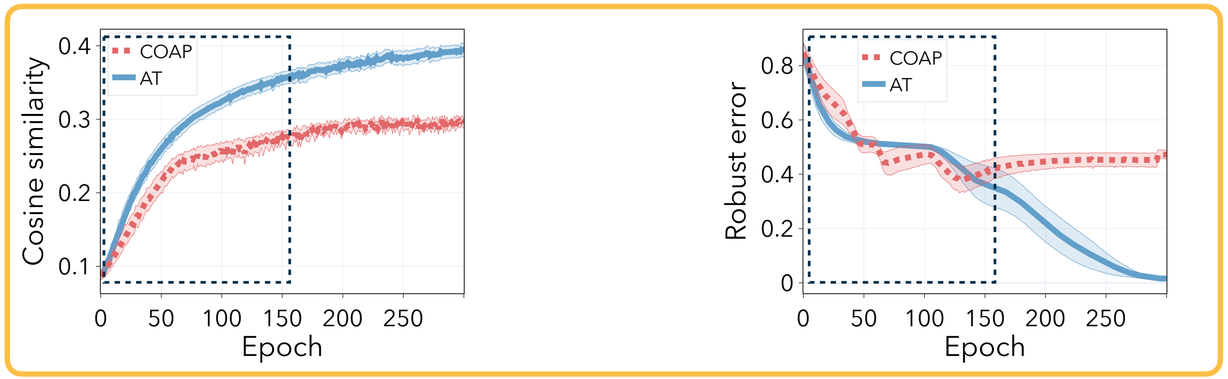
Small ϵ

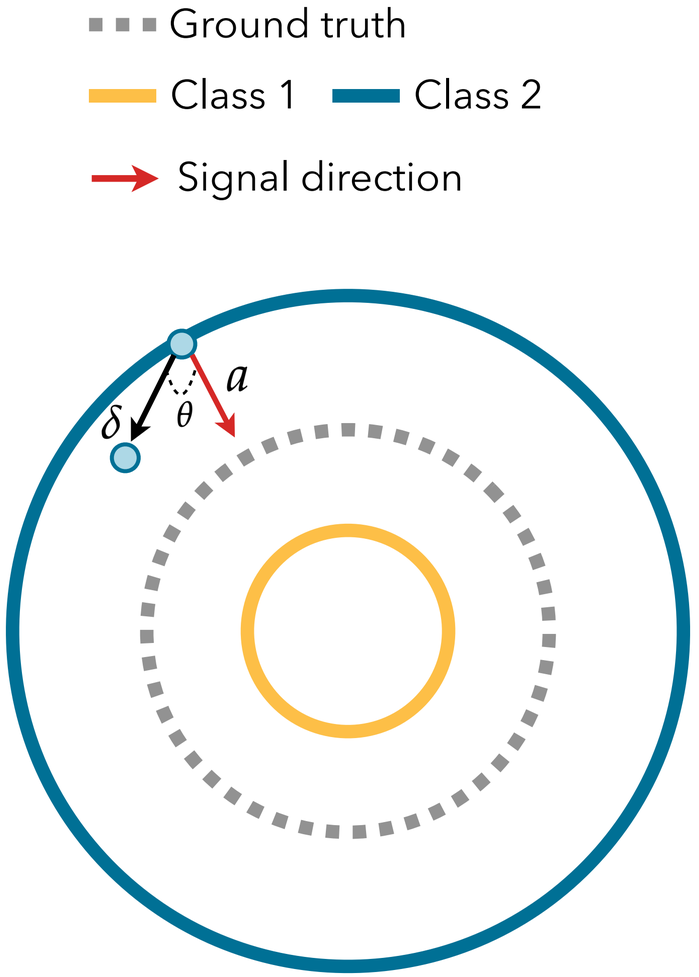




Large ϵ

certified vs empirical

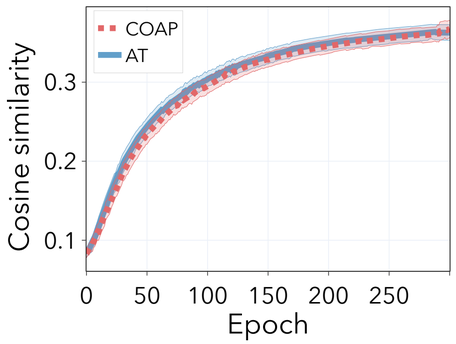


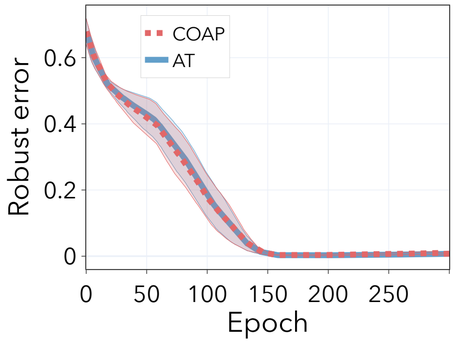


Why do certified defences hurt generalisation?

$$\frac{a \cdot \delta}{\|a\| \, \|\delta\|}$$

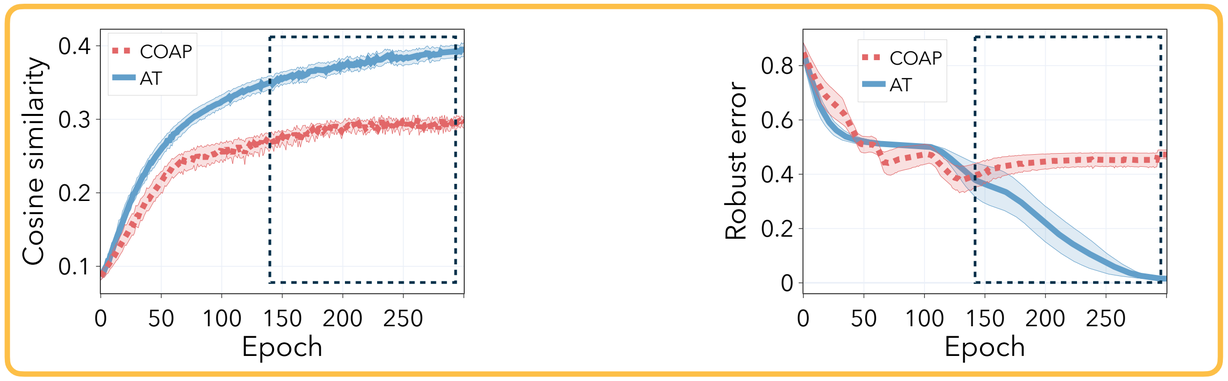
Small ϵ

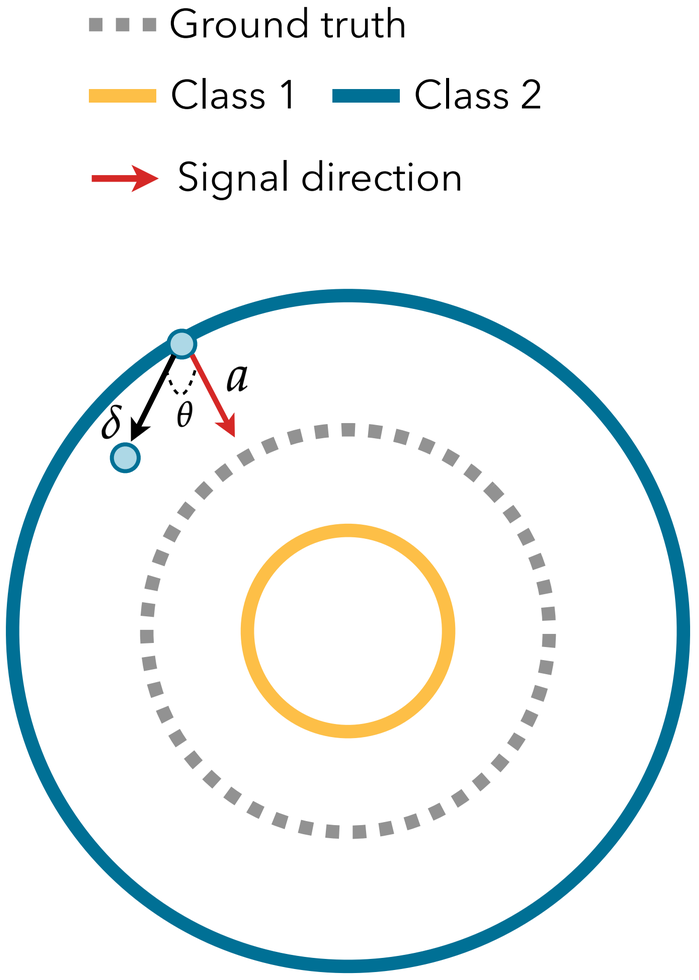




Large ϵ

certified vs empirical





Thanks for your attention! Questions?

Piersilvio

Jacob

Fanny

Amartya

Statistical Machine Learning group

ETH zürich