Certified defences hurt generalisation

Piersilvio De Bartolomeis[1], Jacob Clarysse[1], Fanny Yang[1], Amartya Sanyal[2]

[1]Department of Computer Science, ETH Zürich
[2]ETH AI Center

ETH AI CENTER

ETH zürich

Adversarially robust classification

Goal: Low robust test error for a class of perturbations $\mathcal{B}_\epsilon(x)$

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

- Empirical defences
  - Solve the inner maximisation with first-order optimisation methods
  - Find a lower bound solution $x^\star$

$$L(f_\theta(x^\star), y) \leq \max_{x'\in\mathcal{B}_\epsilon(x)} L(f_\theta(x'), y)$$

- Certified defences
  - Solve a convex relaxation of the inner-maximisation
  - Find an upper bound solution $x^\star$

$$\max_{x'\in\mathcal{B}_\epsilon(x)} L(f_\theta(x'), y) \leq L(f_\theta(x^\star), y)$$

Fairness

We measure the degree of unfairness as follows:

$$\frac{\max_k \mathbf{R}^k(\theta) - \mathbf{R}(\theta)}{1 - \mathbf{R}(\theta)}$$

where

- $\mathbf{R}(\theta)$ is the error of the classifier
- $\mathbf{R}^k(\theta)$ is the error conditioned on the label k

References

[1] Aleksander Madry et al, Towards Deep Learning Models Resistant to Adversarial Attacks. ICLR, 2018.
[2] Eric Wong and J. Zico Kolter, Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope. ICML, 2018.
[3] Huan Zhang et al., Towards Stable and Efficient Training of Verifiably Robust Neural Networks. ICLR, 2020.

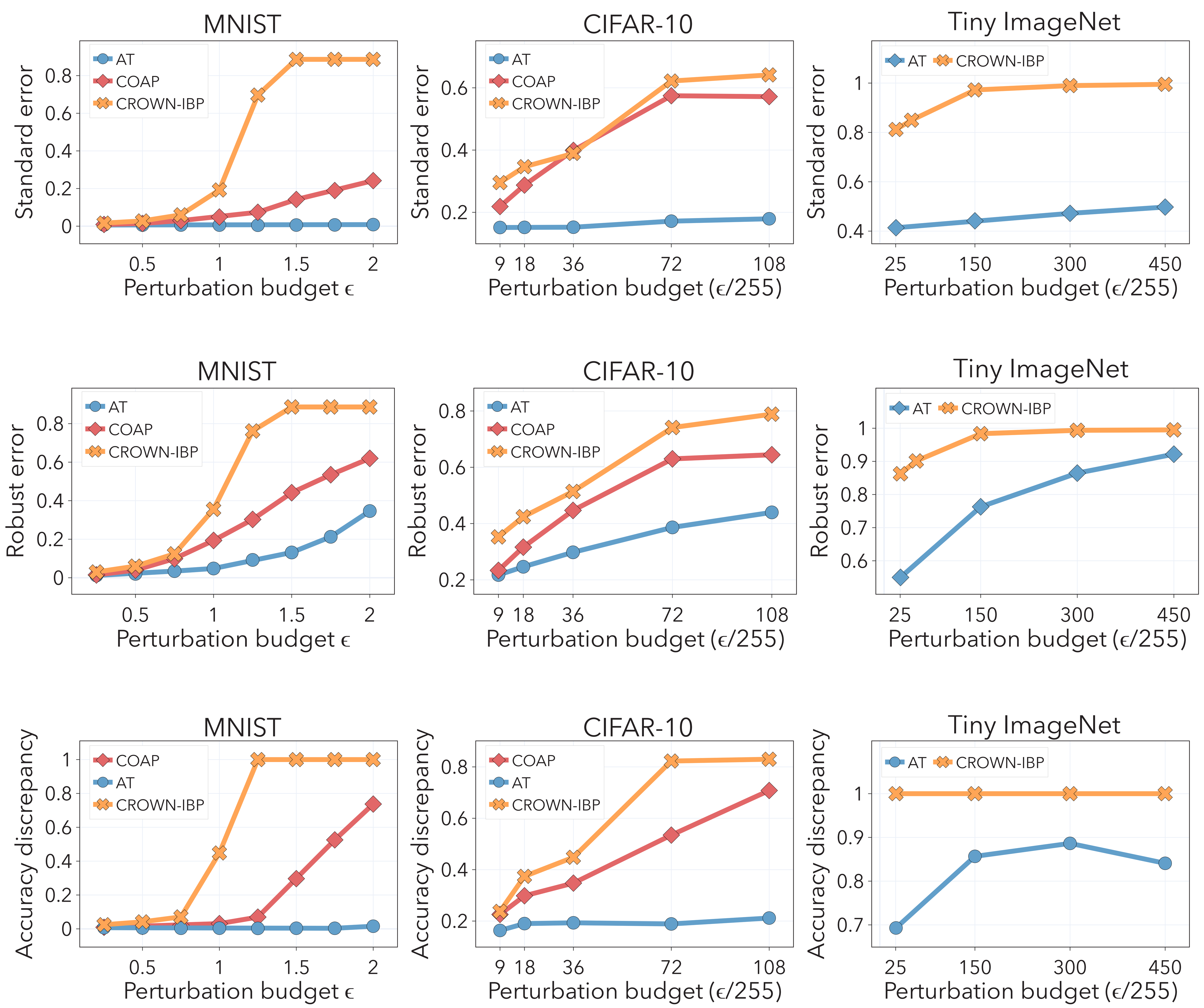
How effective are certified defences in practice?

Certified defences hurt generalisation and fairness

Threat model: ℓ_2-ball perturbations of radius ϵ

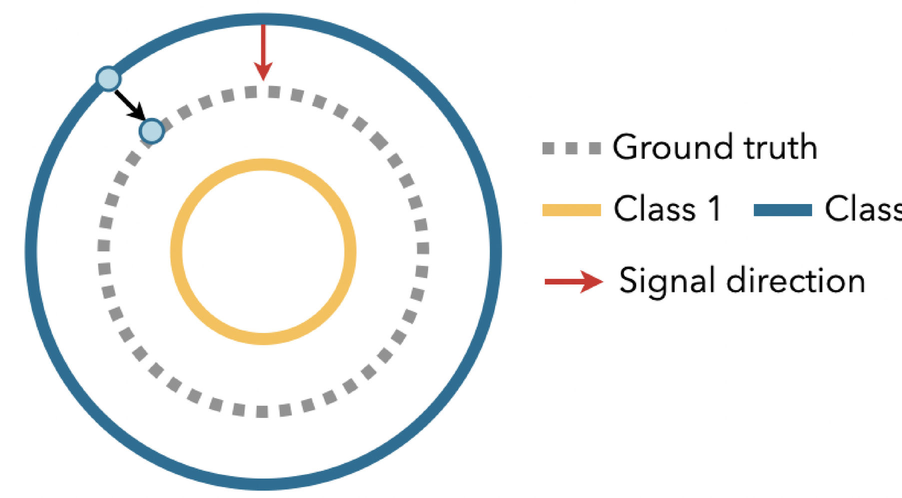
Empirical defences: adversarial training (AT) [1]

Certified defences: convex outer adversarial polytope (COAP) [2] and CROWN-IBP [3]



Why? Intuition on synthetic data

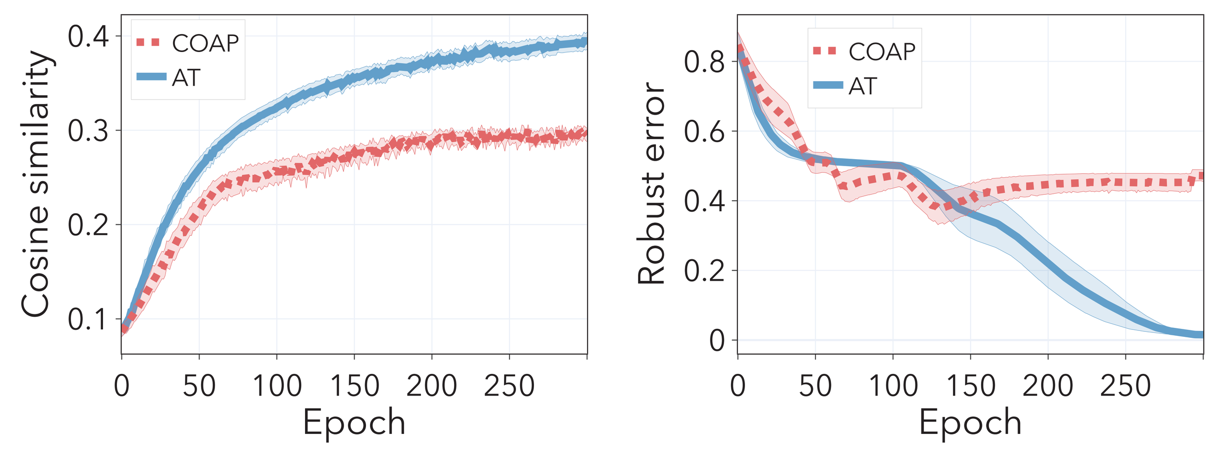
A more controlled setting: concentric spheres



Certified defences hurt generalisation when

- perturbation magnitude is large and
- perturbation aligns with the signal direction

Large ϵ



Theoretical result

Setting

- $y \sim \{-1, 1\}$, $x_1 = \gamma\,\mathrm{sgn}(y)$, $x_{2:d} \sim \mathcal{N}(0, \sigma^2 I_{d-1})$
- $\mathcal{B}_\epsilon(x) = \{z_1 = x + e_1\beta \mid |\beta| \leq \epsilon\}$
- $f_\theta(x) = a\,\mathrm{ReLU}(\theta^\top x) + b$
- θ_1 is the only trainable parameter

Theorem (informal)

For $\frac{2}{3}\gamma < \epsilon < \gamma$ and d large enough, after one step of gradient descent with respect to the COAP and AT objectives, we have: $\mathbf{R}_\epsilon(\theta^{\mathrm{COAP}}) > \mathbf{R}_\epsilon(\theta^{\mathrm{AT}})$