Why adversarial training can hurt robust accuracy

Jacob Clarysse, Julia Hörrmann, Fanny Yang

Department of Computer Science, ETH Zurich

Adversarial robust classification

Goal: Low robust test error for a class of perturbations $T(x, \epsilon_{te})$

- Most common approach: **adversarial training**. One minimizes

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathbb{P}} \left[\max_{x' \in T(x,\epsilon)} L(f_\theta(x'), y) \right]$$

- Better than **standard training** in the high sample regime

Can adversarial training lead to a lower robust accuracy than standard training?

Failure of AT in the low sample regime

- Waterbirds dataset, CIFAR10, ResNets
- Motion blur, adversarial illumination and mask attack

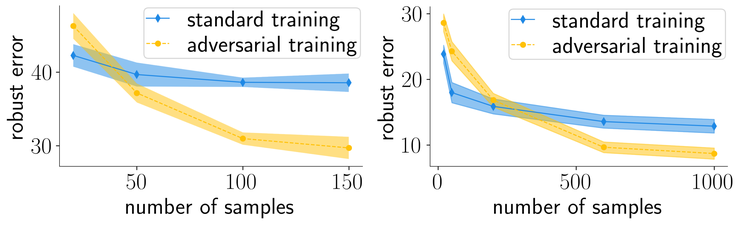


Figure 1: Motion blur

Figure 2: Adversarial illumination

Conclusion: AT can hurt robust accuracy in the low sample size regime.

Directed attacks

Attacks that reduce the information about the object in the image.

Original | Motion blur | Adversarial illumination

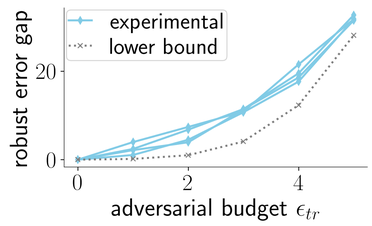
Theoretical result

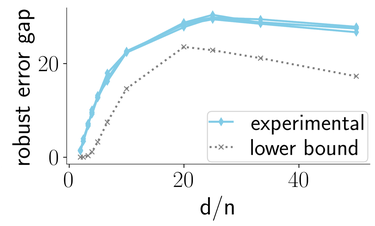
Setting: Binary classification; Gaussian mixture, linear separable;

- Logistic regression with $n < d$
- Directed attacks along signal θ^*

Theorem [informal] W.h.p. over draws of data, we prove

1. that the robust error is monotonically increasing with ϵ
2. a lower bound on the robust error gap $\rightarrow$ the robust error gap increases for increasing d/n until AT classifier $\approx$ trivial

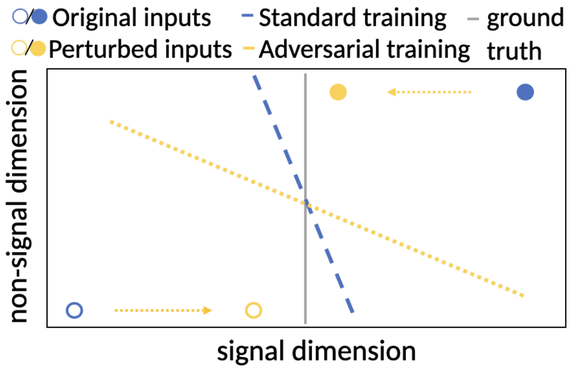




Robust error gap = robust error (AT) − robust error (ST)

Proof intuition I

1. Training with points closer to $\theta^\star$ hurts robust generalization when $n < d$
2. Directed attacks bring points closer to $\theta^\star$



Take-away: Points close to the optimal decision boundary can hurt robust generalization.

Proof intuition II

The robust error can be decomposed in

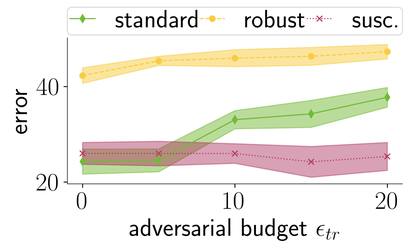
- Standard error, i.e. classification accuracy
- Attack susceptibility, i.e. robustness against attacks

Commonly AT with increasing ϵ

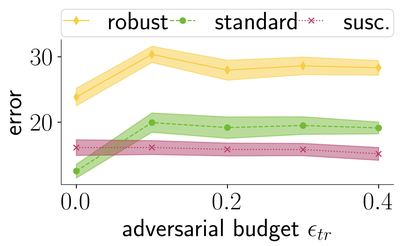
$\uparrow$ standard error $\ll \Delta$ susceptibility $\rightarrow\uparrow$ robust error

AT with directed attacks for increasing ϵ

$\uparrow$ standard error $\gg \downarrow$ susceptibility $\rightarrow\downarrow$ robust error



Motion blur



Adversarial illumination

Related work

For the low dimensional regime and ℓ_p-ball attacks, the literature shows that

- AT can hurt standard accuracy, but improves robust accuracy
- Setting $\epsilon \approx \epsilon_{te}$ is the optimal setting

We show that in the low sample size regime for AT for directed attacks

- can hurt robust accuracy compared to ST
- robust accuracy decreases with increasing ϵ

Conclusion

For directed attacks in the low sample size regime AT

- can hurt robust accuracy compared to ST.
- decreases attack-susceptibility but largely increases the standard error.