Adversarial Training Can Hurt Generalization

Aditi Raghunathan* **Sang Michael Xie*** **Fanny Yang**
John C. Duchi **Percy Liang**

{aditir, xie, pliang}@cs.stanford.edu, {fannyang, jduchi}@stanford.edu

Introduction

Observation on common image datasets: Adversarial training yields higher robust test accuracy but sacrifices standard test accuracy

	Standard training	Adversarial training
Robust test	3.5%	45.8%
Robust train	-	100%
Standard test	95.2%	87.3%
Standard train	100%	100%

Results on CIFAR-10 [2].

Prior work: Tradeoff even with infinite data

- Fundamental conflict: No predictor can achieve both optimal standard accuracy and high robust accuracy [4, 5]
- Lack of expressivity: Hypothesis class doesn't contain an optimally robust and accurate predictor [3]

However for real world datasets following conditions hold:

- Perturbations do not change the true label (e.g., small ℓ_∞ in vision)
- Model class expressive enough (100% standard & robust train accuracy)

If there's no tradeoff with infinite data, why does adversarial training on finite samples lead to lower standard accuracy?

Result in a nutshell

Possible intuition: Additional structural information provided via adversarial training should *help* generalization (like Lasso)

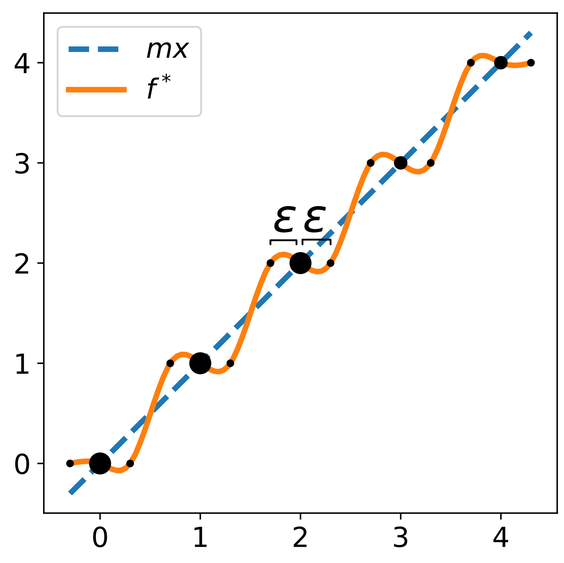
$\Longrightarrow$ Tradeoff due to training dynamics of neural networks?

*Result: No! There is a **purely statistical reason** for the tradeoff.*

We construct a convex learning problem where

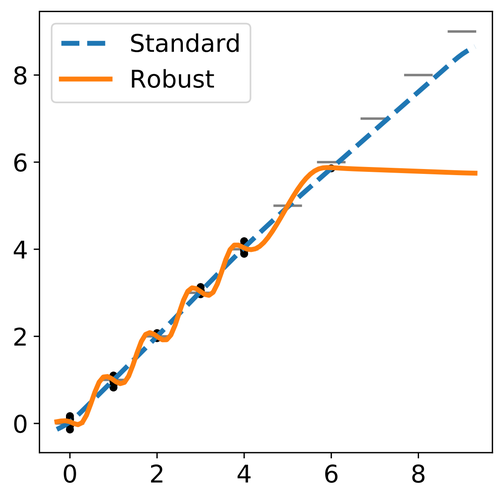
- adversarial training obtains lower standard accuracy for small sample sizes (worse *generalization*)
- robust self-training [1] mostly eliminates the tradeoff by leveraging unlabeled data to increase sample size
- trends match observations on real datasets like CIFAR-10

Convex learning problem: the spline staircase



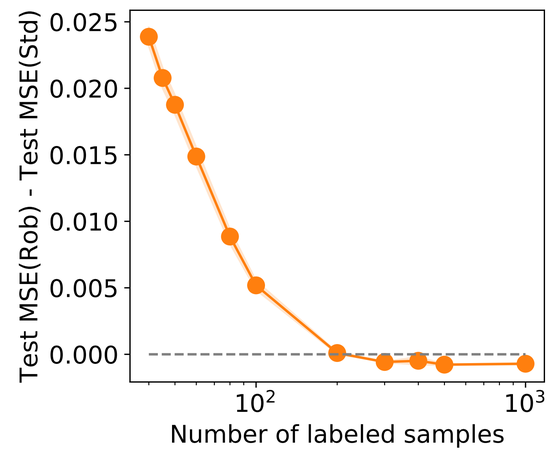
- A simple linear predictor fits most of the distribution
- Perturbations of most points have low probability
- Perturbations require complex staircase fit

Fitting the staircase with small samples



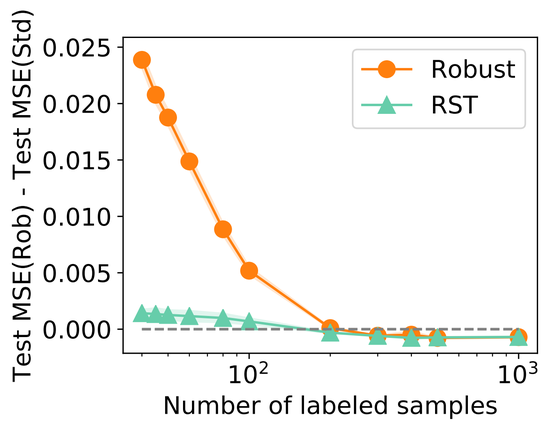
- Standard training only fits most probable points
  $\Longrightarrow$ simple linear predictor (generalizes well)
- Adversarial training fits low probability perturbations
  $\Longrightarrow$ complex staircase predictor (generalizes worse)

Staircase: effect of sample size



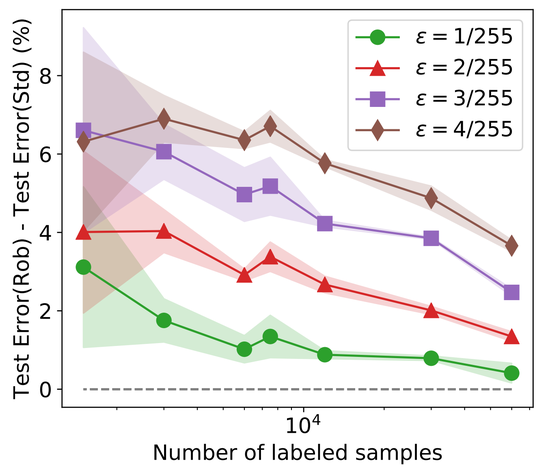
- Gap between standard and adversarial training decreases as sample size increases
- There is no tradeoff with large samples when adversarial training generalizes well

Robust self-training mostly eliminates the tradeoff



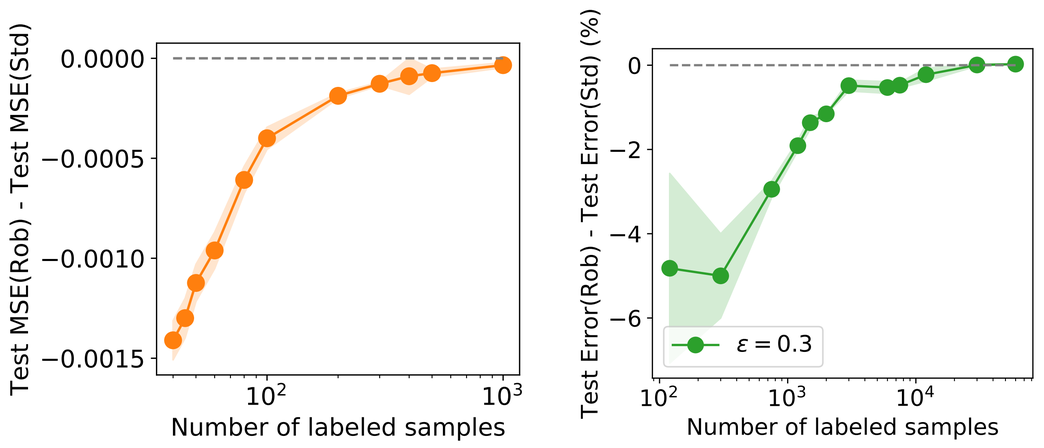
- **Robust self-training (RST) uses unlabeled data** as follows:
  - Learn standard estimator from labeled data (generalizes well)
  - Generate pseudo-labels on unlabeled data and augment dataset
  - Learn robust estimator on augmented dataset (increased sample size)

Parallels in CIFAR-10



- Explore effect of sample size by subsampling CIFAR-10
- The gap between test errors (%) of standard and adversarially trained models decreases with more samples just like the staircase problem

Robustness can also help



(a) Flat staircase ($m = 0$) (b) MNIST

- When the optimal robust predictor is simple, adversarial training can help
- For our staircase problem with slope $m = 0$, optimal predictor is linear
- Adversarial training is less sensitive to target noise
- This observation is mirrored in MNIST

References

[1] Y. Carmon, A. Raghunathan, L. Schmidt, P. Liang, and J. C. Duchi. Unlabeled data improves adversarial robustness, 2019.
[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks, 2018.
[3] P. Nakkiran. Adversarial robustness may be at odds with simplicity, 2019.
[4] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy, 2019.
[5] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy, 2019.