Understanding and Mitigating the Tradeoff Between Robustness and Accuracy

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, Percy Liang

Presented by:

Wissam Kontar, Abhirav Gholba

THE UNIVERSITY of WISCONSIN MADISON

Intriguing properties of Neural Networks

- Deep Neural Networks are highly expressive; reason they succeed but also why they produce uninterpretable solutions with counter-intuitive properties.

- Any linear combination of activations of a layer stores feature information invariantly. It is the space rather than individual units of neural networks that contains the semantic information.

- Input-output mapping in NN is not perfect. Imperceptible perturbations can cause a model to misclassify.

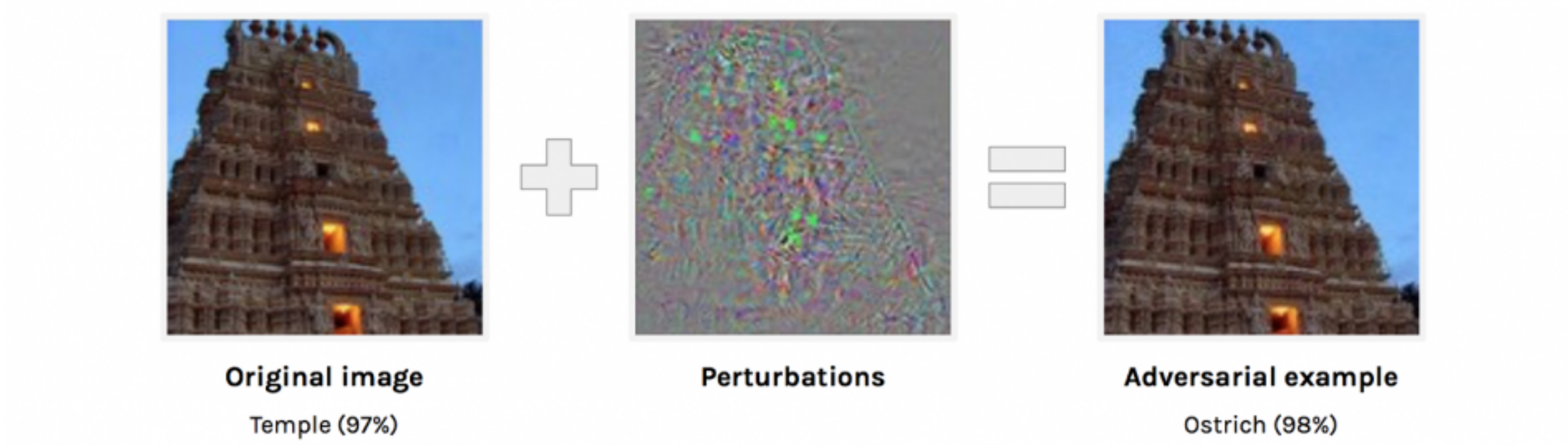
Szegedy et al, 2014

Bad news: machine learning is not robust

Common adversarial attacks

Two broad types:

1) Black box
2) White box (our focus)



THE UNIVERSITY of WISCONSIN MADISON

Common adversarial attacks

The Fast Gradient Sign Method (FGSM) attack

$$x + \varepsilon\, \mathrm{sgn}(\nabla_x L(\theta, x, y)).$$

	Error rate	Confidence	ε
MNIST (softmax)	99.9%	79.3%	0.25
MNIST (maxout)	89.4%	97.6%	0.25
CIFAR-10 (maxout)	87.15%	96.6%	0.1

Goodfellow, 2015

Common adversarial attacks

The Projected Gradient Descent (PGD) attack

$$x^{t+1} = \Pi_{x+\mathcal{S}} \left(x^t + \alpha \, \mathrm{sgn}(\nabla_x L(\theta, x, y)) \right).$$

- Very strong first order attack.
- Iterative.
- Finds perturbations in l_2 and l_∞ balls.

Tsiparas et al, 2019

Dataset	Norm	ε	Standard Accuracy: Standard	Standard Accuracy: Half-half	Standard Accuracy: Robust	Robust Accuracy: Standard	Robust Accuracy: Half-half	Robust Accuracy: Robust
MNIST	ℓ_∞	0	99.31%	-	-	-	-	-
		0.1	99.31%	99.43%	99.36%	29.45%	95.29%	95.05%
		0.2	99.31%	99.22%	98.99%	0.05%	90.79%	92.86%
		0.3	99.31%	99.17%	97.37%	0.00%	89.51%	89.92%
	ℓ_2	0	99.31%	-	-	-	-	-
		0.5	99.31%	99.35%	99.41%	94.67%	97.60%	97.70%
		1.5	99.31%	99.29%	99.24%	56.42%	87.71%	88.59%
		2.5	99.31%	99.12%	97.79%	46.36%	60.27%	63.73%
CIFAR10	ℓ_∞	0	92.20%	-	-	-	-	-
		$2/255$	92.20%	90.13%	89.64%	0.99%	69.10%	69.92%
		$4/255$	92.20%	88.27%	86.54%	0.08%	55.60%	57.79%
		$8/255$	92.20%	84.72%	79.57%	0.00%	37.56%	41.93%
	ℓ_2	0	92.20%	-	-	-	-	-
		$20/255$	92.20%	92.04%	91.77%	45.60%	83.94%	84.70%
		$80/255$	92.20%	88.95%	88.38%	8.80%	67.29%	68.69%
		$320/255$	92.20%	81.74%	75.75%	3.30%	34.45%	39.76%

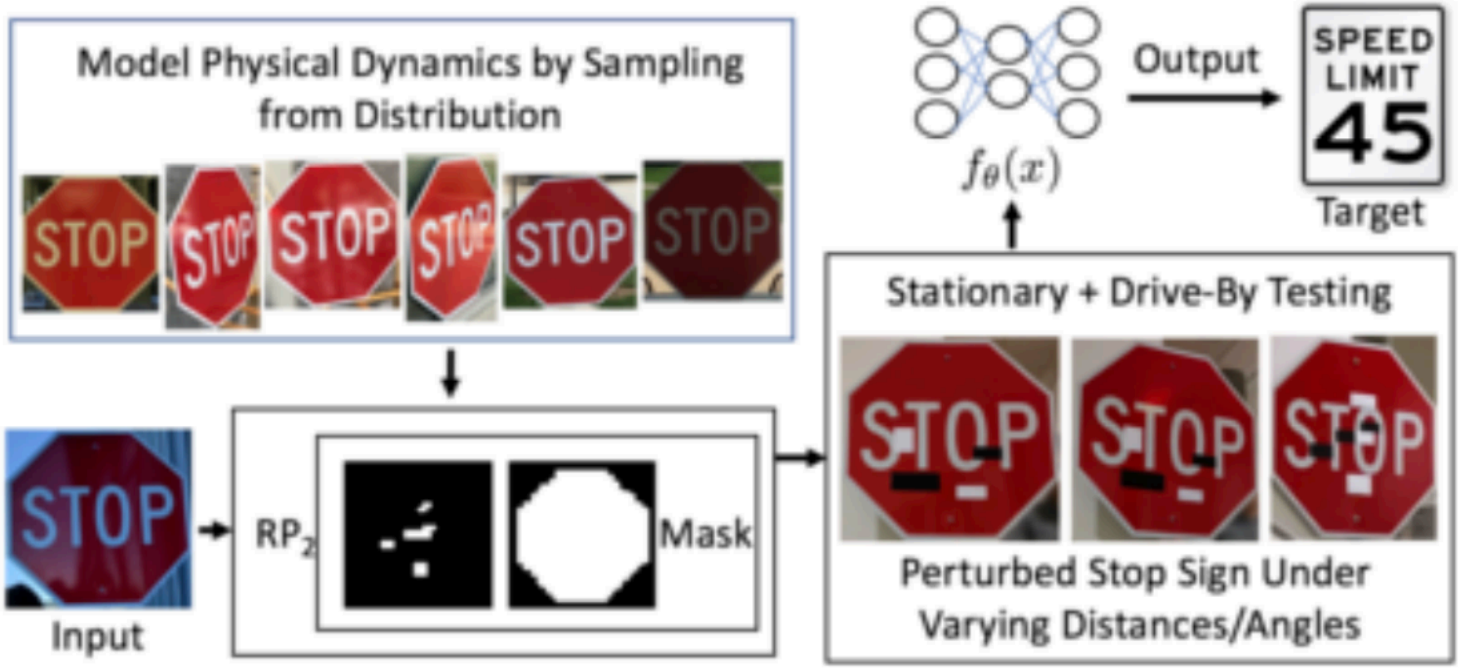
Tsiparas et al, 2019

Robust Physical-World Attacks

- Robust Physical Perturbation (RP_2)
- Targeted misclassification on real-world example of traffic stop sign.
- Generates robust perturbations that achieve high misclassification rates under various environmental conditions, including viewpoints.

Eykholt et al, 2019

Robust Physical-World Attacks

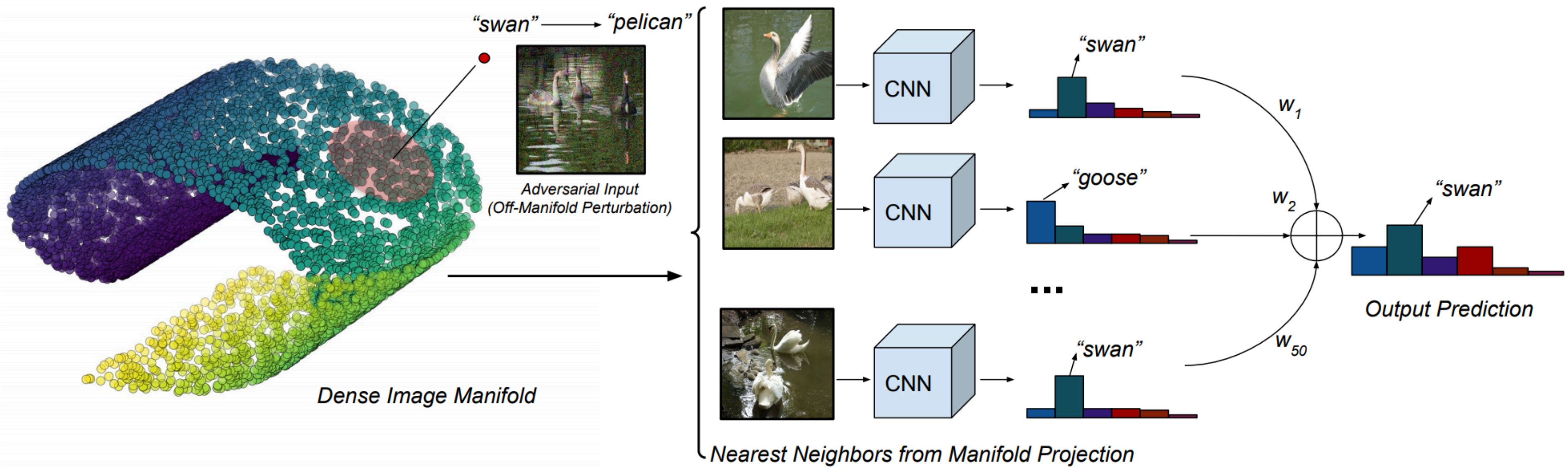


$$\min \quad H(x+\delta, x), \quad \text{s.t.} \quad f_\theta(x+\delta) = y^*$$

$$\underset{\delta}{\operatorname{argmin}}\ \lambda ||\delta||_p + J(f_\theta(x+\delta), y^*) \qquad (1)$$

Eykholt et al, 2019

Robust Defense



Defense Against Adversarial Images using Web-Scale Nearest-Neighbor Search

Dubey et al, 2019

Robust Defense

Method

- “Off-manifold” adversarial images.
- Approximate the projection of an adversarial example onto the image manifold by the finding nearest neighbors in the image database.
- Classify the “projection” of the adversarial example.

Robust Defense

Defense	Clean	Gray box	Black box
No defense	0.761	0.038	0.046
Crop ensemble [10]	0.652	0.456	0.512
TV Minimization [10]	0.635	0.338	0.597
Image quilting [10]	0.414	0.379	**0.618**
Ensemble training [35]	–	–	0.051
ALP [16]	0.557	0.279	0.348
RA-CNN [39]*	0.609	0.259	–
Our Results			
IG-50B-All (conv_5_1-RMAC)	0.676	0.427	0.491
IG-1B-Targeted (conv_5_1)	**0.681**	**0.462**	0.587
YFCC-100M (conv_5_1)	0.613	0.309	0.395
IN-1.3M (conv_5_1)	0.471	0.286	0.312

Table 2. ImageNet classification accuracies of ResNet-50 models using state-of-the-art defense strategies against the PGD attack, using a normalized ℓ_2 distance of 0.06. * RA-CNN [39] experiments were performed using a ResNet-18 model.

THE UNIVERSITY of WISCONSIN MADISON

General Robustness Problem

Training distribution $\neq$ Test distribution

- Robust Statistics: Hard train & Normal test
- Robust Optimization: Normal train & Hard test

THE UNIVERSITY of WISCONSIN MADISON

Dilemma

Our goal: robust (test) accuracy (test; adversarial examples)
Direct instinct: optimize robust (training) accuracy (training; adversarial training)
Problem: standard accuracy is affected

Results on CIFAR 10

Training	Standard Accuracy	Robust Accuracy
Standard Training	95.2%	0%
Adversarial Training (Modry et al. 2018)	87.3%	45.8%
TRADES (Zhang et al. 2019)	84.8%	55.4%

There is a tradeoff between standard accuracy and robust accuracy

THE UNIVERSITY of WISCONSIN MADISON

Key Questions

1. Why is there a tradeoff ?

2. When does it happen?

3. How to mitigate it?

THE UNIVERSITY of WISCONSIN MADISON

Setup Formulation

Standard accuracy: average over training distribution

$$\mathrm{P}[f(x) \neq y]$$

Standard training: find f to optimize standard error on the training data

Robust error: average over worst-case perturbations

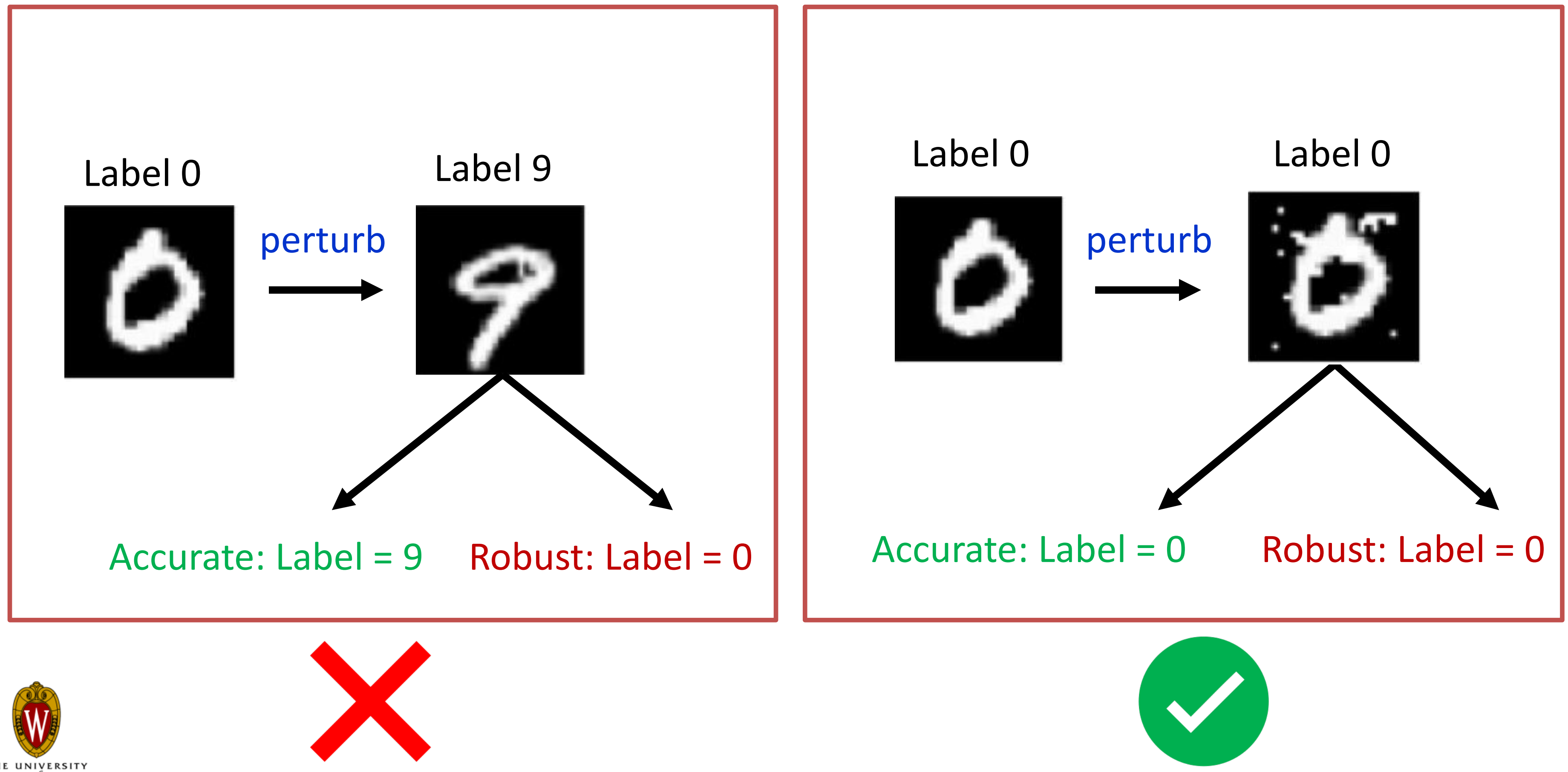
$$P[\exists \tilde{x} \in B(x) \; such \; that \; f(\tilde{x}) \neq y]$$

$$B(x) = \{\tilde{x} \mid \|\tilde{x} - x\|_\infty \leq \varepsilon\}$$

Robust training: find f to optimize robust error on training data

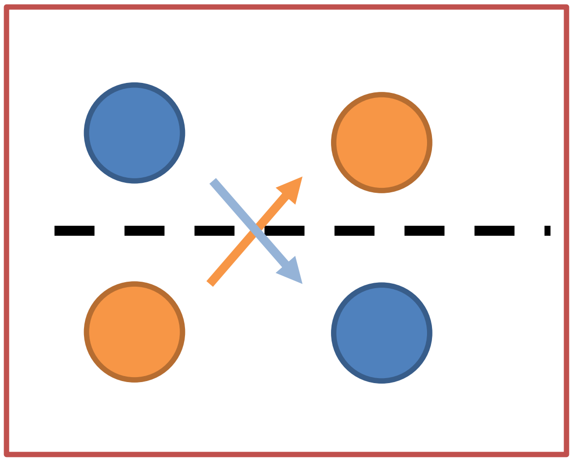
Why can robust training affect standard accuracy?

1) The optimal accurate predictor is not robust: *(Tsipras et al. 2019, Zhang et al. 2019, Fawzi et al. 2018)*



Why can robust training affect standard accuracy?

2) Model class is not expressive enough: *(Nakkiran et al. 2019)*



Well specified problem

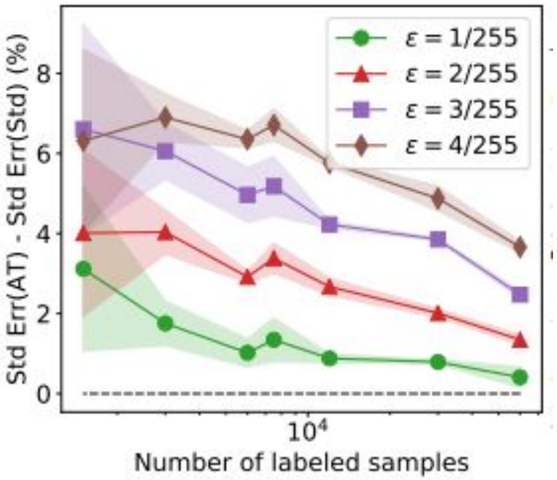
Over parametrized network can fit data perfectly

THE UNIVERSITY of WISCONSIN MADISON

Why can robust training affect standard accuracy?

When things are consistent $f^*(x) = f^*(\tilde{x})$, and we have a well specified setting , why is there still a tradeoff ?

- Suggests that tradeoff exists even with infinite data



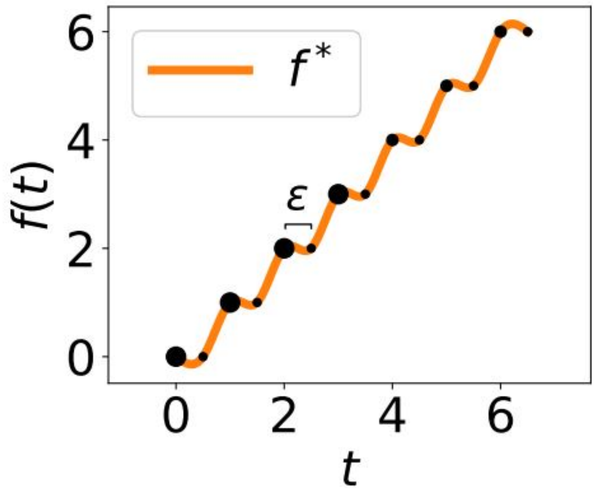
Gap between standard and robust error decreases with more data

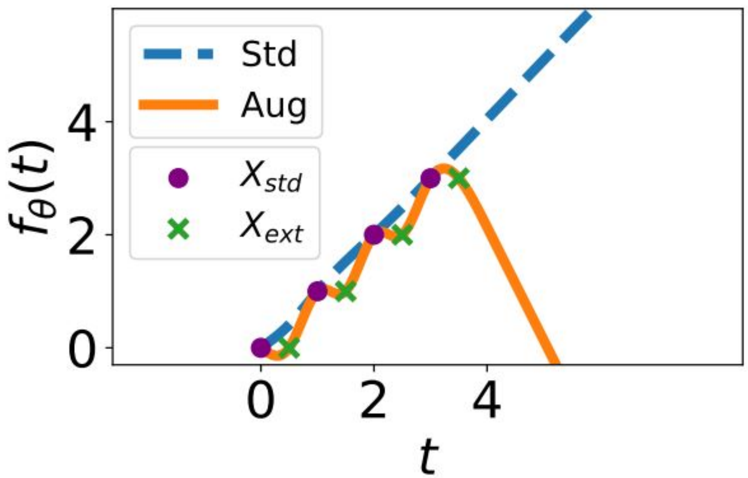
THE UNIVERSITY of WISCONSIN MADISON

Spline Setting

Spline setting: consider a well-specified model (no approximation issues), and convex (no optimization issues)

Surprisingly: tradeoff still exists





Extra data commanded local fit at the expense of global fit

THE UNIVERSITY of WISCONSIN MADISON

General Linear Setting

Simple linear model: $y = x^T\theta^*$

Standard data: $X_{std}, y_{std} = X_{std}\theta^*$

Extra data (adversarial data): $X_{ext}, y_{ext} = X_{ext}\theta^*$

Analysis of the estimators:

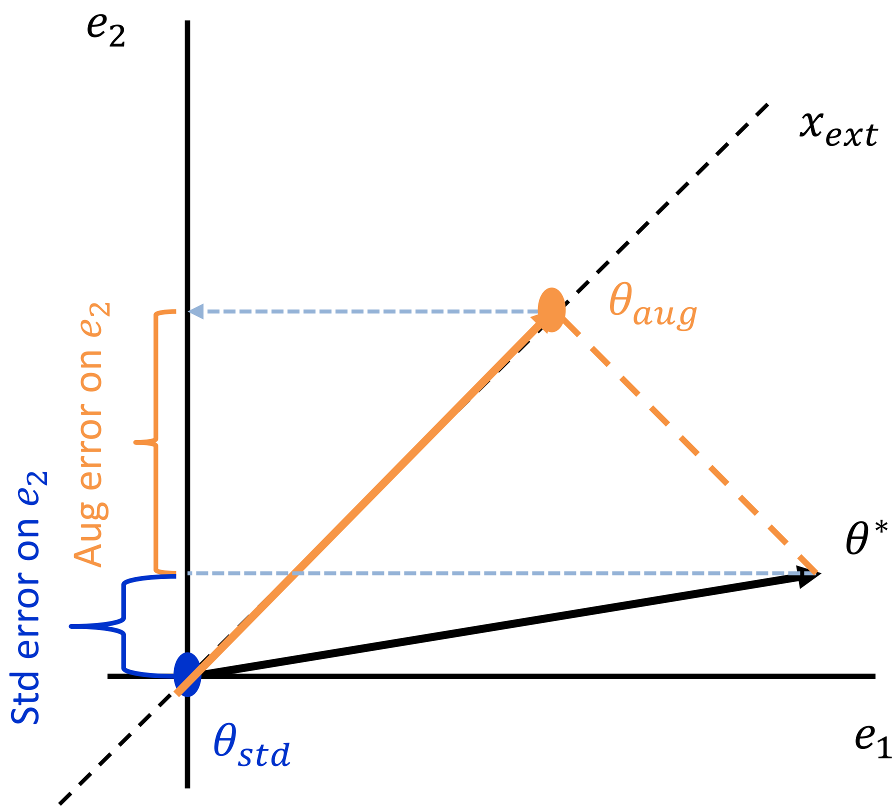
- $\theta_{std} = argmin_\theta\{\|\theta\|_2 : X_{std}\theta = y_{std}\}$
- $\theta_{aug} = argmin_\theta\{\|\theta\|_2 : X_{std}\theta = y_{std}, X_{est}\theta = y_{ext}\}$

How are these two estimators related, and why adding extra points will make error worse.

THE UNIVERSITY of WISCONSIN MADISON

Extra data increasing error

- $\theta_{std} = argmin_{\theta}\{\|\theta\|_2 : X_{std}\theta = y_{std}\}$
- $\theta_{aug} = argmin_{\theta}\{\|\theta\|_2 : X_{std}\theta = y_{std}, X_{est}\theta = y_{ext}\}$



Standard test error:

$$L_{std}(\theta) = (\theta - \theta^*)^T \Sigma (\theta - \theta^*)$$

Σ is population covariance; governs which space is more expensive

If Σ has large weight on direction of e_2
Then errors on e_2 are expensive

Augmented estimator θ_{aug} has much higher standard error

THE UNIVERSITY of WISCONSIN MADISON

Extra data increasing error

$$L_{std}(\theta_{std}) - L_{std}(\theta_{aug}) = \boxed{v^T \Sigma v} + \boxed{2w^T \Sigma v}$$

$$v = \Pi_{std}^{\perp} \Pi_{aug} \theta^* \; and \; w = \Pi_{aug}^{\perp} \theta^*$$

Always Positive term (PSD): decrease in standard error of θ_{aug} by fitting extra data in some direction ***BENEFIT***

Can be negative: measures the cost of a possible increase in parameter error along a certain direction (like e_2 previously) ***COST***

Cost > Benefit : standard error of θ_{aug} is higher

No tradeoff scenario:

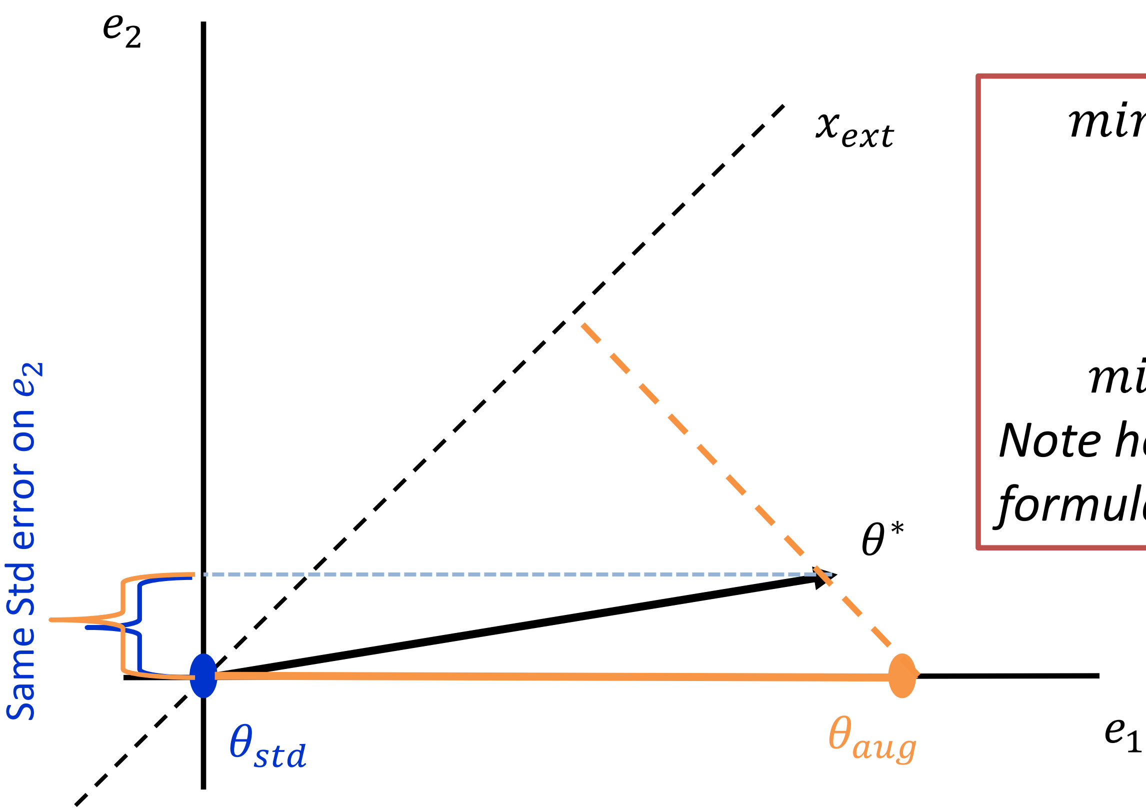
- $w = \Pi_{aug}^{\perp} \theta^* = 0$
  - Perfect Augmentation on entire space
- $\Sigma = \mathrm{I}$
  - No direction is more costly than the other (augmentation is always beneficial)

THE UNIVERSITY of WISCONSIN MADISON

How Can We Mitigate the Tradeoff ?

Our Intentions:

- Keep θ_{std} the same
- Find a robust estimator for X_{ext}



Robust Self Training

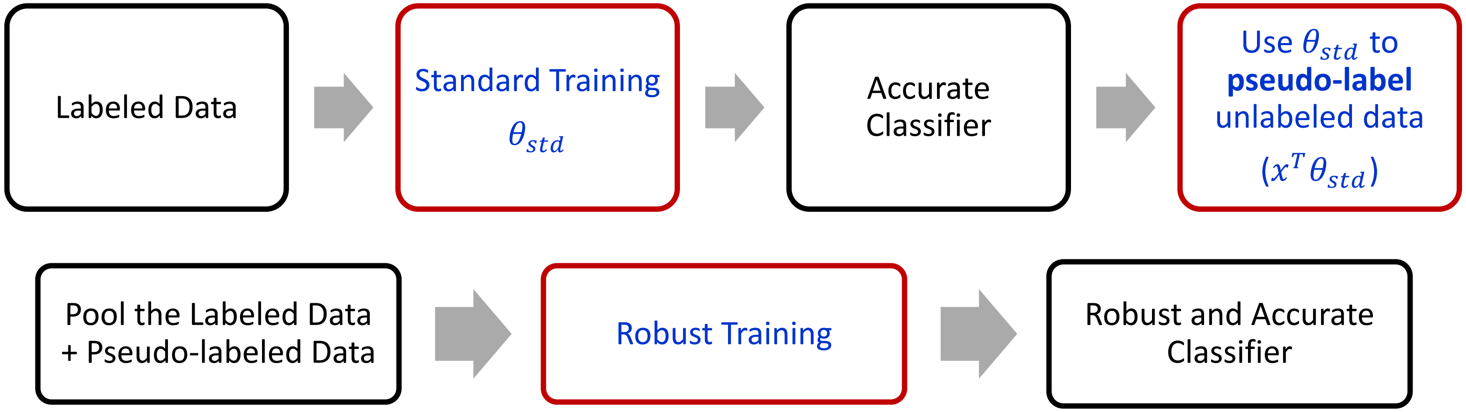
$$\min \mathbb{E}_x[(x^T\theta - x^T\theta_{std})^2] \; s.t.$$
$$X_{std}\theta = y_{std}$$
$$X_{ext}\theta = y_{ext}$$

$$\min (\theta - \theta_{std})^T \, \Sigma (\theta - \theta_{std})$$

Note how Σ is included in the formulation

- You train your estimator $x^T\theta$ to have predictions close to the pseudo-labels $x^T\theta_{std}$
- And still interpolating the available data

Robust Self Training



	Standard	Robust (X_{ext} or X_{adv})
Labeled Data	Input x and label y	Input x_{ext} and label y
Unlabeled Data	Input $\tilde{x}$ and pseudo-label $\tilde{y}$	Input $\tilde{x}_{ext}$ and pseudo-label $\tilde{y}_{ext}$

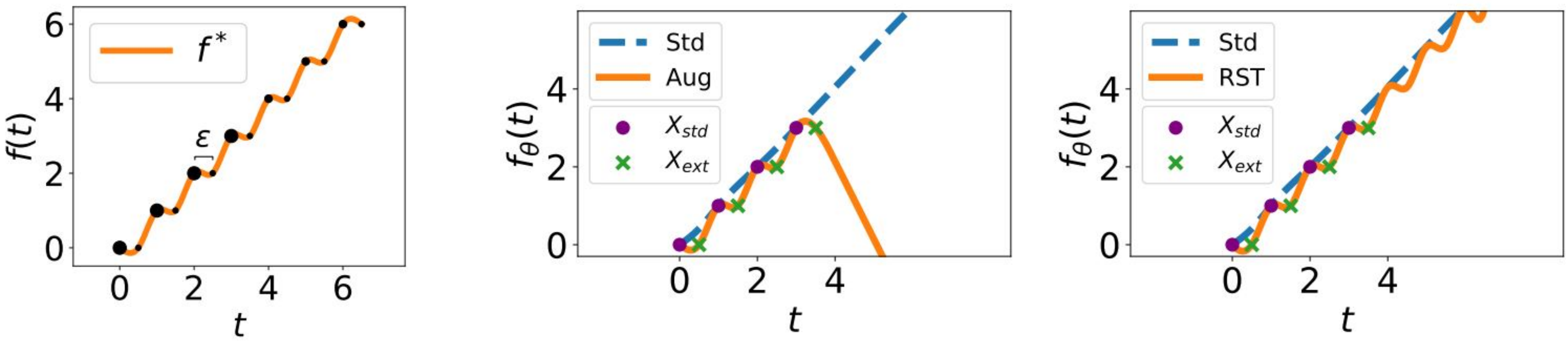
$$min\mathbb{E}_x[(x^T\theta - x^T\theta_{std})^2]\ s.t.$$
$$X_{std}\theta = y_{std}$$
$$X_{ext}\theta = y_{ext}$$

$$L_{std}(\theta_{rst}) \leq L_{std}(\theta_{std})$$
$$L_{rob}(\theta_{rst}) \leq L_{rob}(\theta\_aug)$$

[Carmon et al. 2019]

How Does All This Help ?

Revisit our Spline example



We achieved a global structure and a local structure

THE UNIVERSITY of WISCONSIN MADISON

Take Out Points

1. Sometimes adding true data to the model can hurt (spline example)
2. Unlabeled data when added can in fact help in robustness (Robust Self Training)
3. We might think that NN can be very expressive and fit anything, but the key problem remains in inductive bias and generalization; if done wrong will hurt the model a lot

Question & Discussion

Quiz Questions

1) What happens to the gap between standard error of adversarial training and standard training when training data increases ?
   a) Stays the same
   b) Increases
   c) Decreases

2) What is the approach the authors take in explaining the tradeoff between standard and robust error? Tradeoff occurs due to:
   a) Hypothesis class is not expressive enough
   b) Generalization from finite data
   c) Standard and robust error being fundamentally at odds
   d) Robust accuracy being hard to optimize

3) Which of the following statements is true?
   a) When the population covariance Σ is equal to the identity matrix I, the standard error does not increase when fitting augmented data
   b) The parameter error does not change with data augmentation
   c) Robust Self Training (RST) improves the robust error and hurts (increases) the standard error
   d) Augmenting the training data set with perturbed examples will decrease the standard error.

THE UNIVERSITY of WISCONSIN MADISON