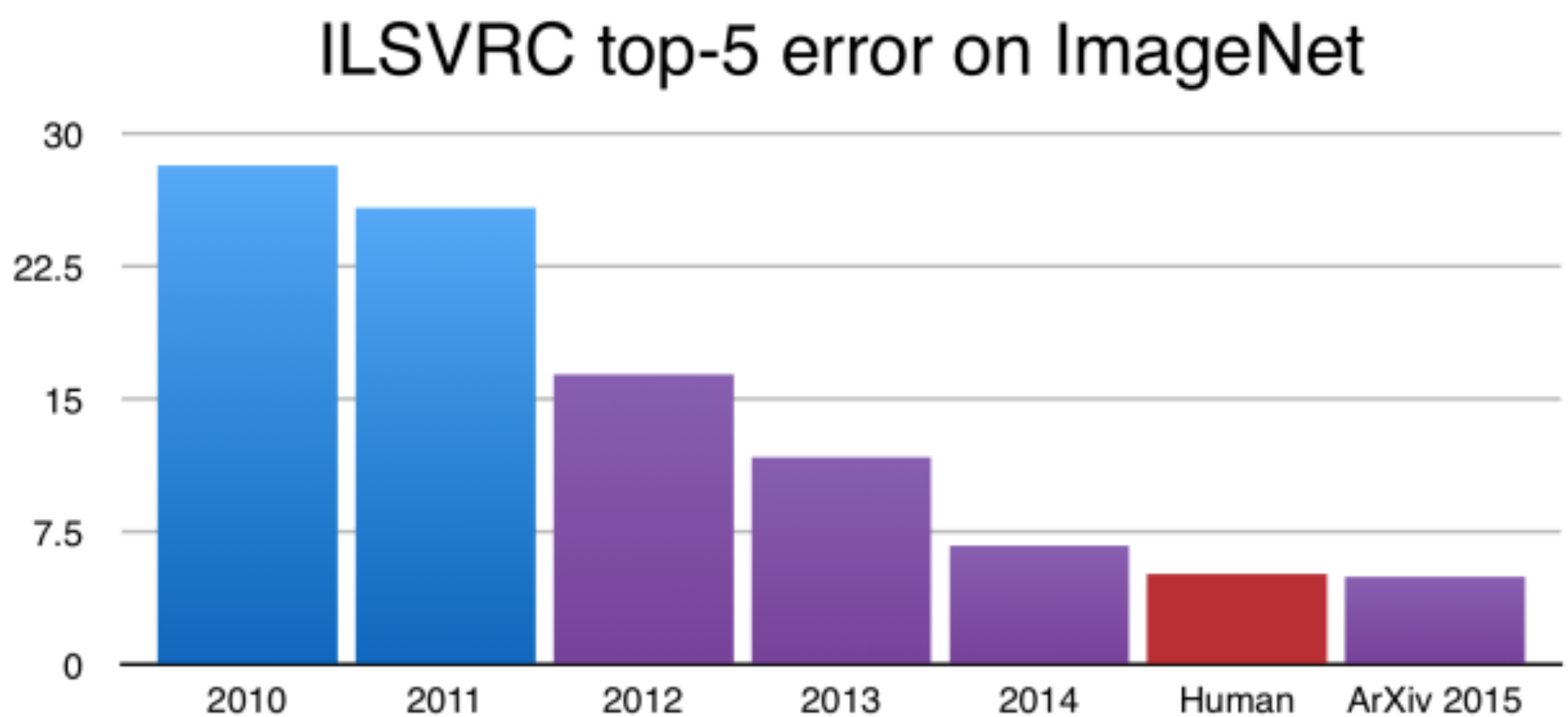
Adversarially Robust Optimization and Generalization

Ludwig Schmidt

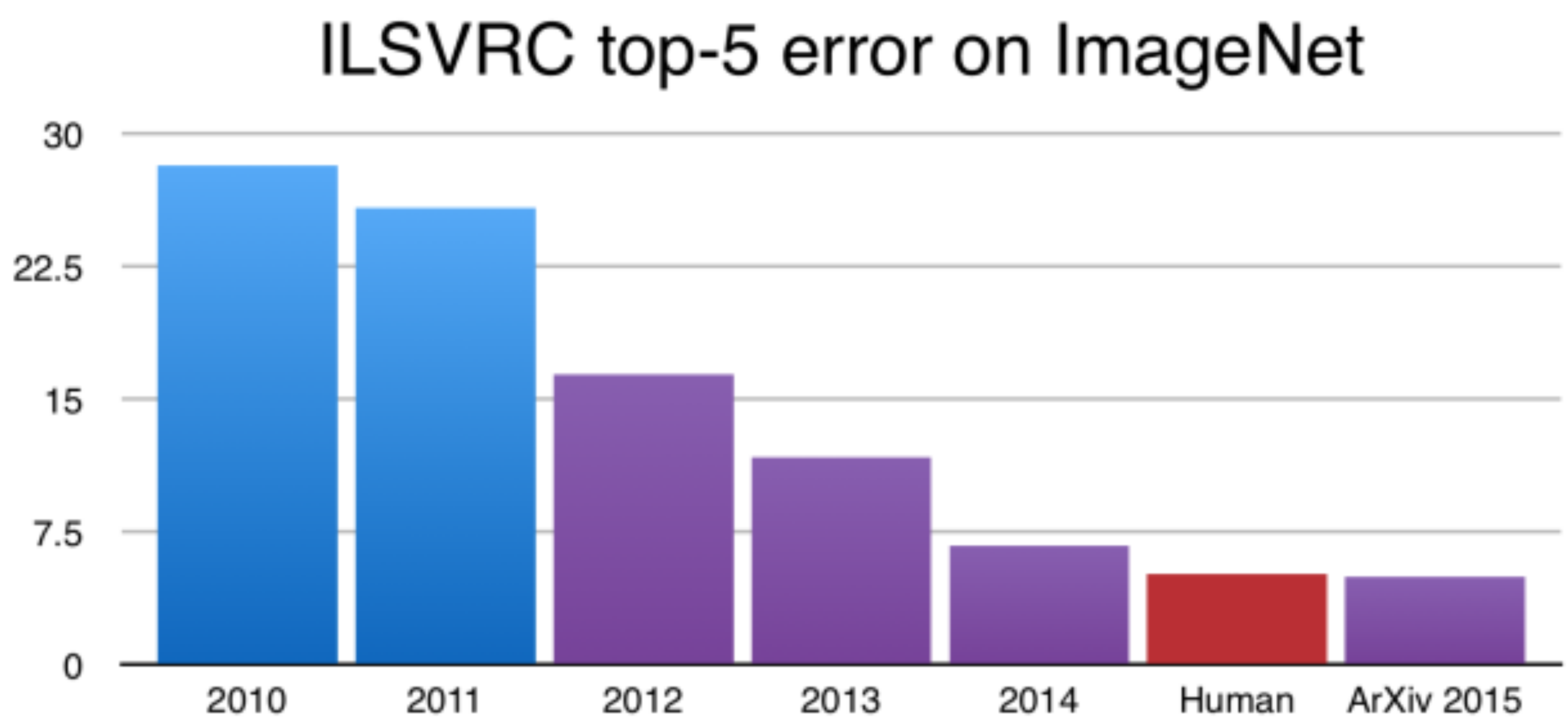
MIT → UC Berkeley

Based on joint works with Logan Engstrom (MIT), Aleksander Madry (MIT), Aleksandar Makelov (MIT), Dimitris Tsipras (MIT), Kunal Talwar (Google), and Adrian Vladu (Boston University).

Recent Progress in ML



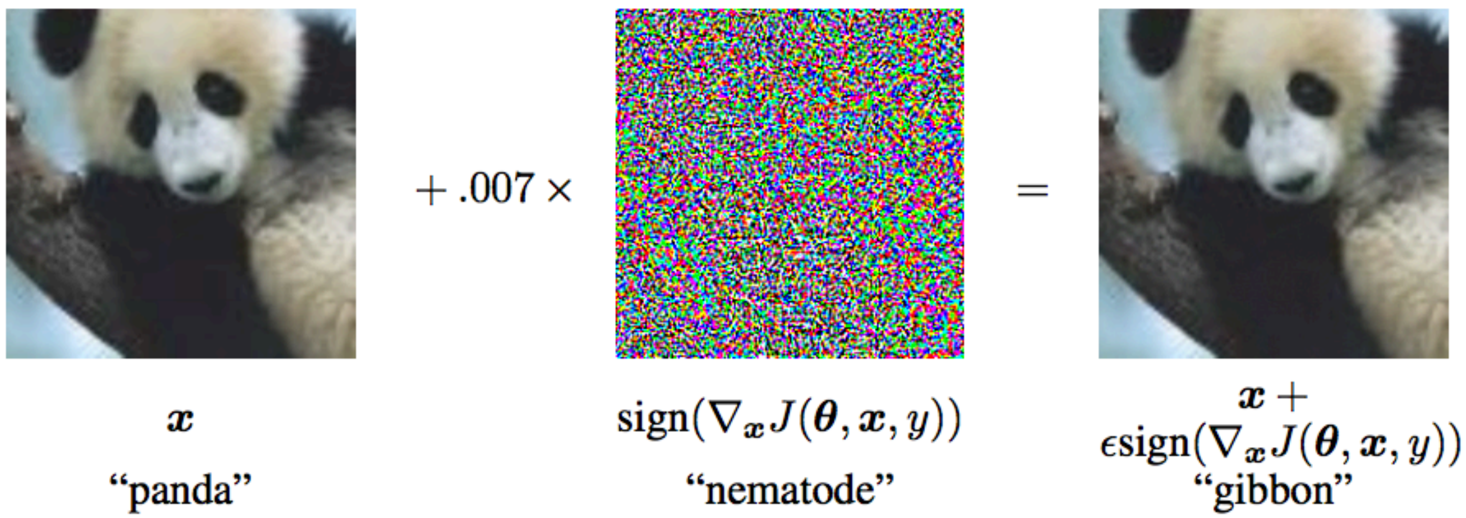
Recent Progress in ML



Have we *really* achieved human-level performance?

Lack of Robustness

Adversarial Examples

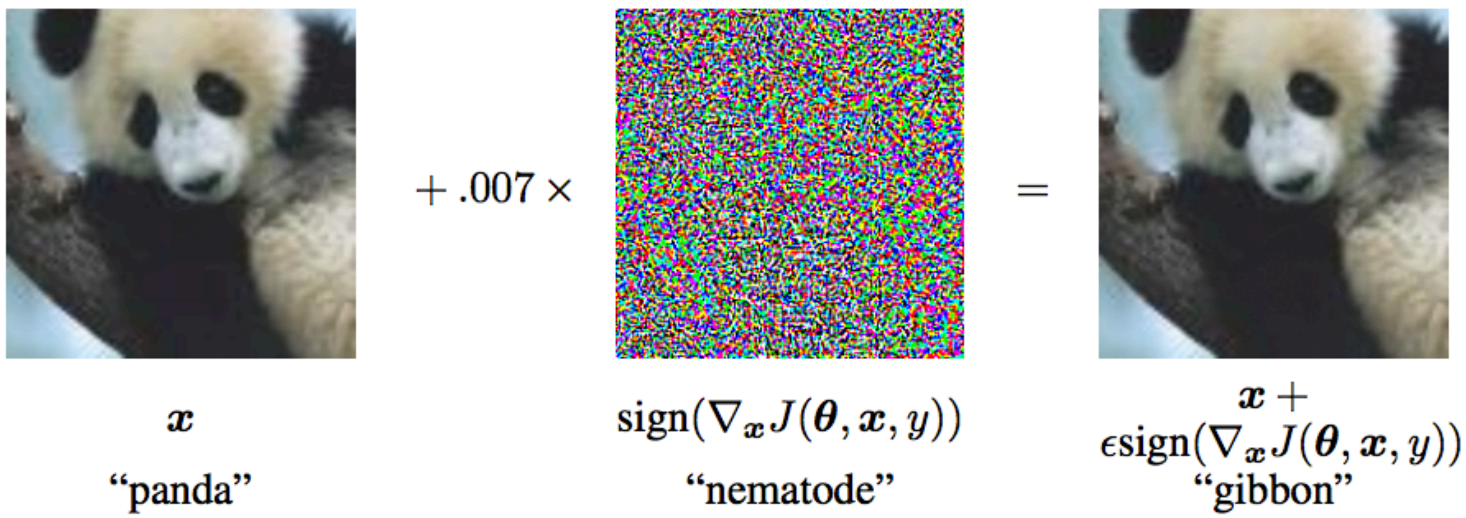


[Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus, 2014]

[Athalye, Engstrom, Ilyas, Kwok, 2017]

Lack of Robustness

Adversarial Examples



[Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus, 2014]

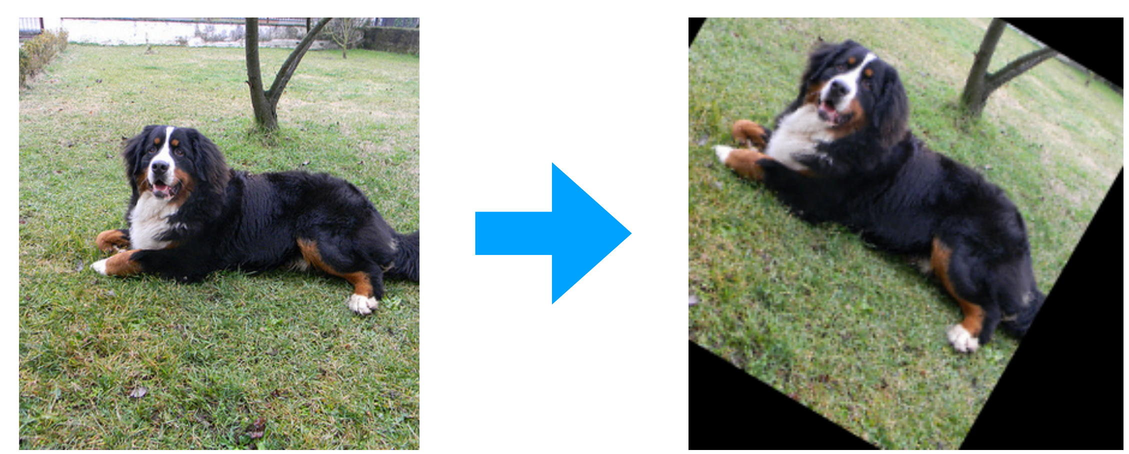
[Athalye, Engstrom, Ilyas, Kwok, 2017]

Translations + rotations

(shifts by <10% pixels, <30° rotations)

CIFAR10: 93% → **8%** accuracy

ImageNet: 76% → **31%** accuracy



[Engstrom, Tsipras, Schmidt, Madry, 2017]

Adversarially Robust Generalization

“Standard” Generalization

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\, \text{loss}(x, \theta) \, \right]$$

Adversarially Robust Generalization

"Standard" Generalization

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\,\text{loss}(x,\theta)\,\right]$$

Adversarially Robust Generalization

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\max_{x'\in P(x)} \text{loss}(x',\theta)\right]$$

Perturbation set: rotations, translations,
small ℓ_∞ perturbations, …

Adversarially Robust Generalization

"Standard" Generalization

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\,\text{loss}(x,\theta)\,\right]$$

Adversarially Robust Generalization

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\max_{x'\in P(x)}\text{loss}(x',\theta)\right]$$

Perturbation set: rotations, translations,
small ℓ_∞ perturbations, …

What is the right set of perturbations?

This talk: assume the set P is given.

Long History

ANNALS OF MATHEMATICS
Vol. 46, No. 2, April, 1945

STATISTICAL DECISION FUNCTIONS WHICH MINIMIZE THE MAXIMUM RISK

BY ABRAHAM WALD

(Received November 7, 1944)

1. Introduction

In some previous publications (see [1] and the last chapter in [2]) the author outlined a theory of statistical inference which deals with the following general problem: Let $X = (X_1, \cdots, X_n)$ be a set of random variables and suppose that the joint cumulative distribution function $F(t_1, \cdots, t_n)$ of the random variables $X_1, \cdots, X_n$ is not known. However it is known that $F(t_1, \cdots, t_n)$ is an element of a given class Ω of distribution functions. Consider a system S of subsets of Ω and for each element ω of S let H_ω denote the hypothesis that the joint distribution function of $X_1, \cdots, X_n$ is an element of ω. Furthermore, denote by H_S the system of all hypotheses H_ω corresponding to all elements ω of S. Let $E = (x_1, \cdots, x_n)$ denote an observation on X, i.e., x_i denotes an observed value of X_i $(i = 1, 2, \cdots, n)$. The totality of all possible observations E on X is the n-dimensional Cartesian space and is called the sample space. Any point of the sample space is called a sample point. The problem

Why This Guarantee?

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\max_{x' \in P(x)} \text{loss}(x', \theta) \right]$$

Why This Guarantee?

$$\mathbb{E}_{x \sim \mathcal{D}} \left[\max_{x' \in P(x)} \text{loss}(x', \theta) \right]$$

If a classifier satisfies this property, we avoid **arms races**.

JSMA → Defensive Distillation → Tuned JSMA

[Papernot et al. '15], [Papernot et al. '16], [Carlini et al. '17]

FGSM → Feature Squeezing, Ensembles → Tuned Lagrange

[Goodfellow et al. '15], [Abbasi et al. '17], [Xu et al. '17]; [He et al. '17]

How Can We Get There?

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\max_{x'\in P(x)} \text{loss}(x',\theta)\right]$$

How Can We Get There?

$$\mathbb{E}_{x \sim \mathcal{D}} \left[\max_{x' \in P(x)} \text{loss}(x', \theta) \right]$$

Standard image classifiers do not satisfy this property.

How Can We Get There?

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\max_{x'\in P(x)} \mathrm{loss}(x', \theta)\right]$$

Standard image classifiers do not satisfy this property.

How does robustness affect **optimization** and **sample complexity**?

Robust Optimization

Main problem:

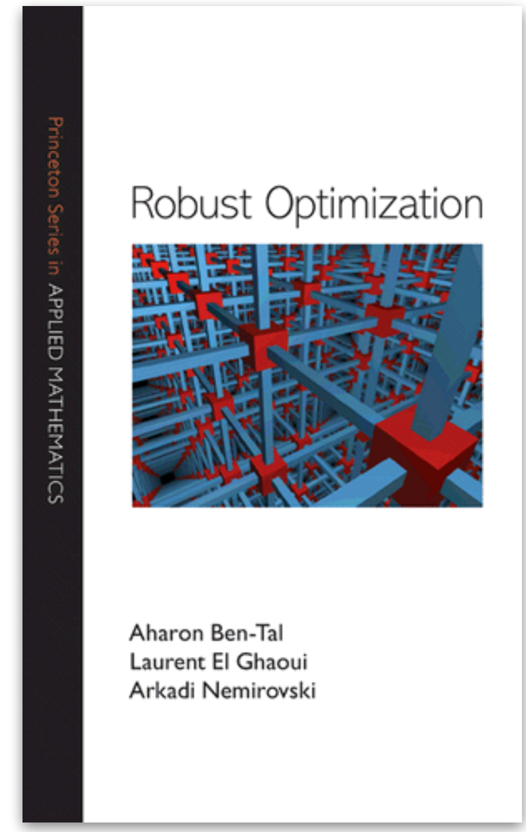
$$\min_{\theta} \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\max_{x' \in P(x)} \text{loss}(x', \theta) \right]$$

[Madry, Makelov, Schmidt, Tsipras, Vladu, 2017]

Robust Optimization

Main problem:

$$\min_{\theta} \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \left[\max_{x' \in P(x)} \text{loss}(x', \theta) \right]$$



[Madry, Makelov, Schmidt, Tsipras, Vladu, 2017]

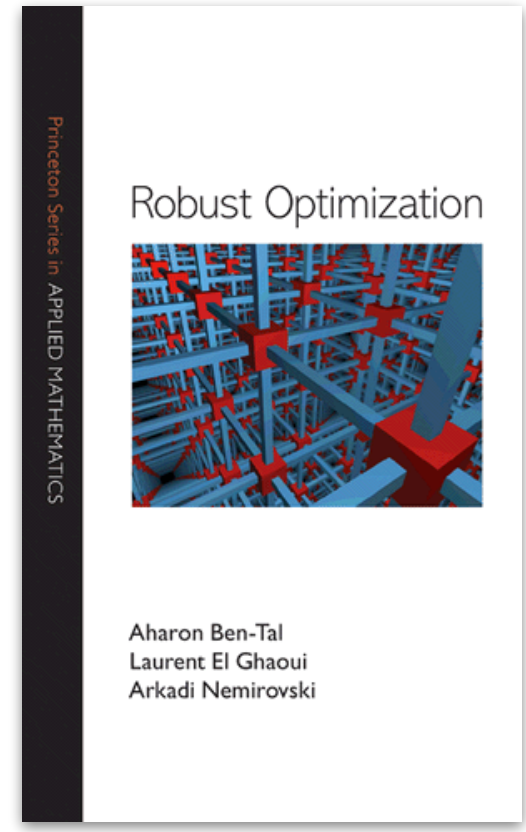
Robust Optimization

Main problem:

$$\min_{\theta} \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\max_{x' \in P(x)} \text{loss}(x', \theta) \right]$$

Convert to empirical risk:

$$\min_{\theta} \sum_{i=1}^{n} \max_{x' \in P(x_i)} \text{loss}(x', \theta)$$

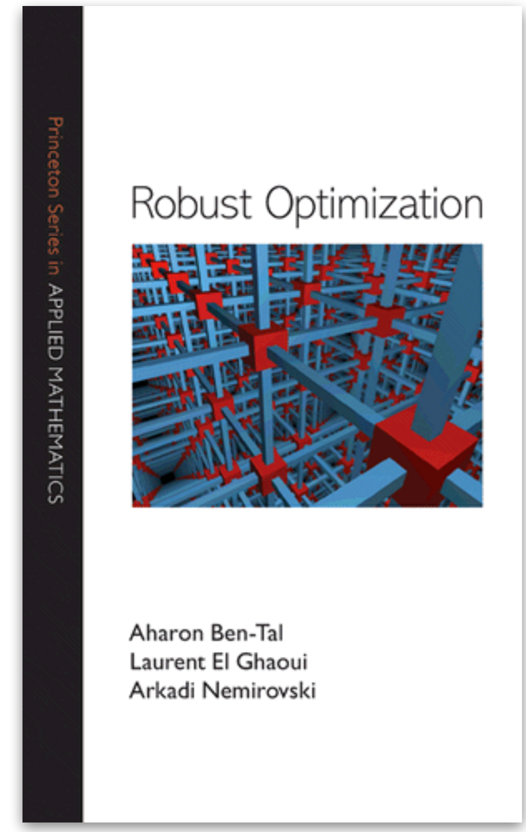


[Madry, Makelov, Schmidt, Tsipras, Vladu, 2017]

Robust Optimization

Main problem:

$$\min_{\theta} \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\max_{x' \in P(x)} \text{loss}(x', \theta) \right]$$



Convert to empirical risk:

$$\min_{\theta} \sum_{i=1}^{n} \max_{x' \in P(x_i)} \text{loss}(x', \theta)$$

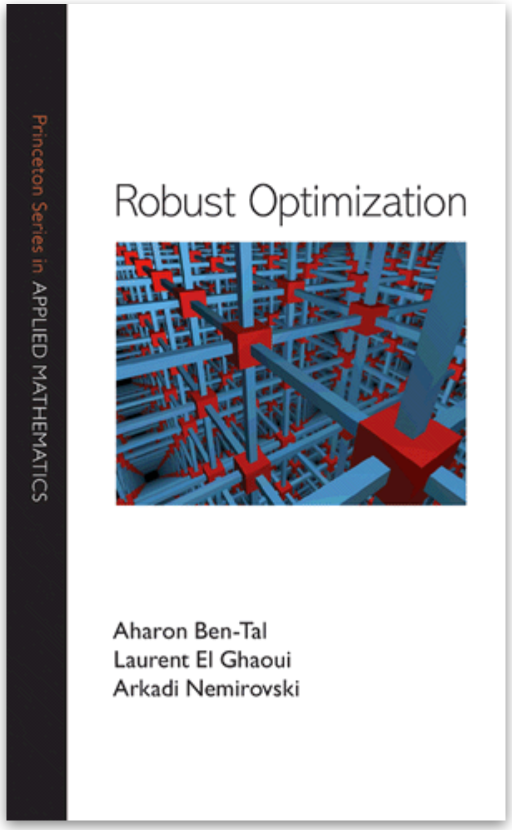
min part:
run SGD

[Madry, Makelov, Schmidt, Tsipras, Vladu, 2017]

Robust Optimization

Main problem:

$$\min_{\theta} \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\max_{x'\in P(x)} \text{loss}(x',\theta)\right]$$



Convert to empirical risk:

$$\min_{\theta} \sum_{i=1}^{n} \max_{x'\in P(x_i)} \text{loss}(x',\theta)$$

min part:
run SGD

How do we get gradients for the inner max?

[Madry, Makelov, Schmidt, Tsipras, Vladu, 2017]

Good Gradients = Good Attacks

Danskin's Theorem

Simplified, but holds for **non-convex** losses: Let

$$\phi_x(\theta) = \max_{x' \in P(x)} \text{loss}(x', \theta)$$

and let x^*_θ be a constrained maximizer of $\text{loss}(\cdot, \theta)$. Then

$$\nabla \phi_x(\theta) \ = \ \nabla_\theta \, \text{loss}(x^*_\theta, \theta)$$

Good Gradients = Good Attacks

Danskin's Theorem

Simplified, but holds for **non-convex** losses: Let

$$\phi_x(\theta) = \max_{x' \in P(x)} \text{loss}(x', \theta)$$

and let x^*_θ be a constrained maximizer of $\text{loss}(\cdot, \theta)$. Then

$$\nabla \phi_x(\theta) \; = \; \nabla_\theta \, \text{loss}(x^*_\theta, \theta)$$

Overall algorithm: **adversarial training.**

→ Principled approach for $\min_\theta \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\max_{x' \in P(x)} \text{loss}(x', \theta) \right]$

Good Gradients = Good Attacks

Danskin's Theorem

Simplified, but holds for **non-convex** losses: Let

$$\phi_x(\theta) = \max_{x' \in P(x)} \text{loss}(x', \theta)$$

and let x^*_θ be a constrained maximizer of $\text{loss}(\cdot, \theta)$. Then

$$\nabla \phi_x(\theta) \;=\; \nabla_\theta \, \text{loss}(x^*_\theta, \theta)$$

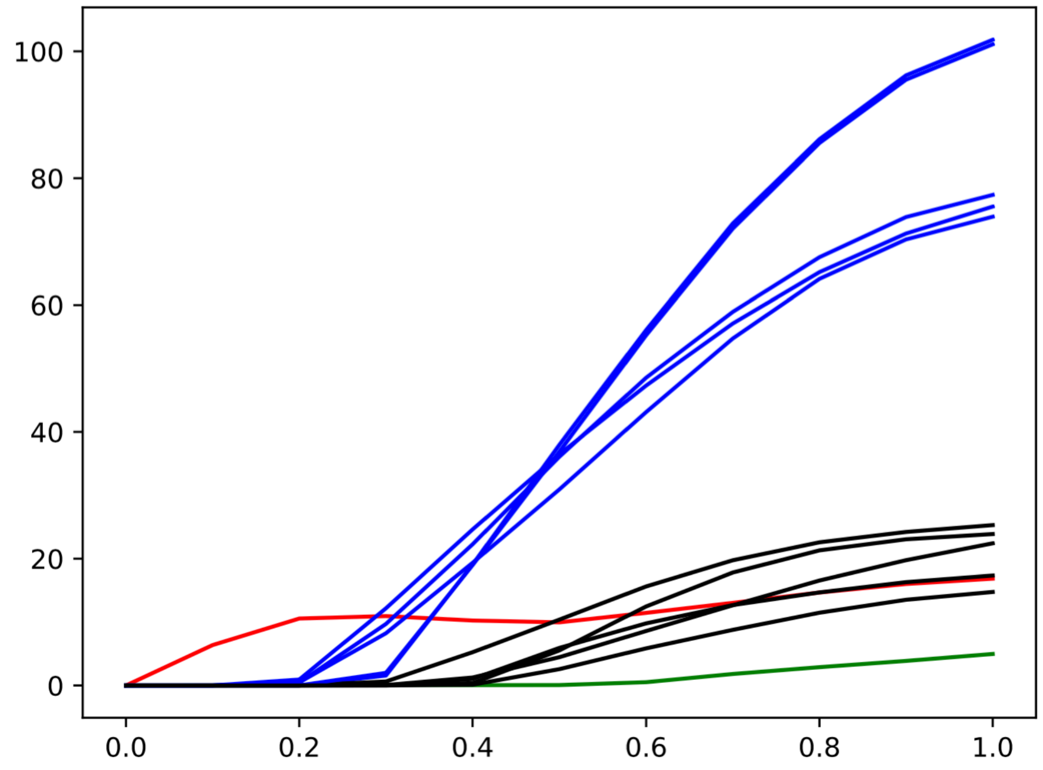
Overall algorithm: **adversarial training.**

→ Principled approach for $\min_\theta \mathbb{E}_{x \sim \mathcal{D}} \left[\color{red}{\max_{x' \in P(x)}} \text{loss}(\color{red}{x'}, \theta) \right]$

Crucial point: need to find the best possible attack.

Is There Any Hope?

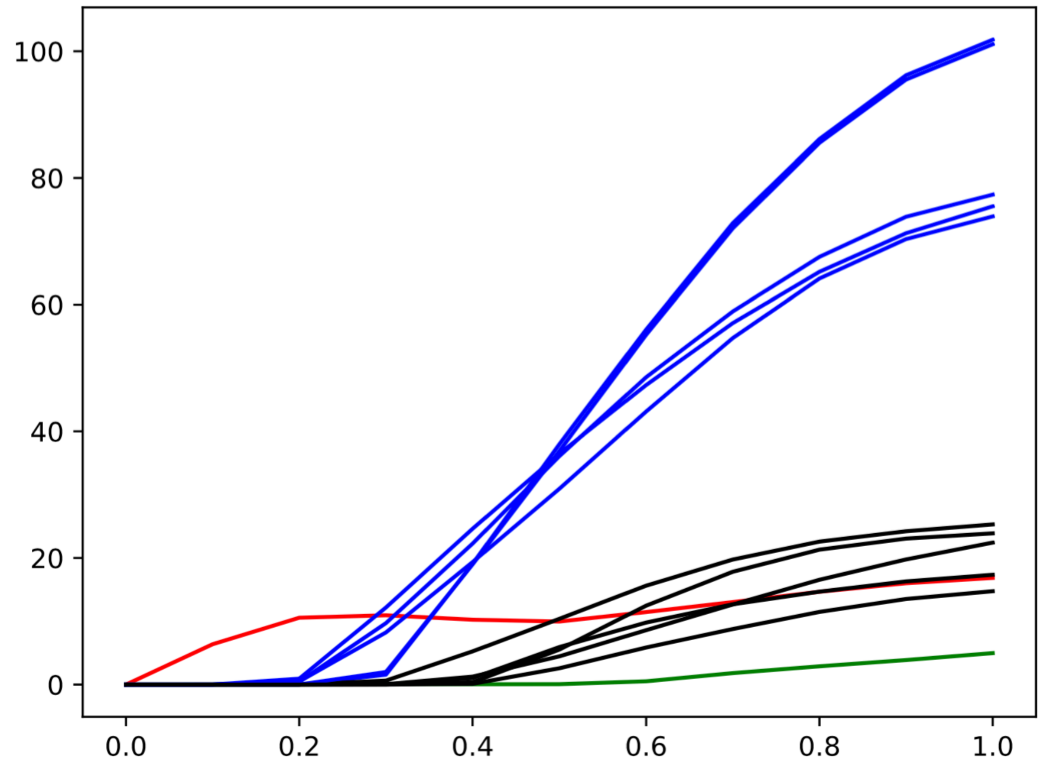
Non-concave maximization problem.



FGSM (single gradient)
PGD (100 steps with η=0.3)
Transfer FGSM
Transfer PGD

Is There Any Hope?

Non-concave maximization problem.

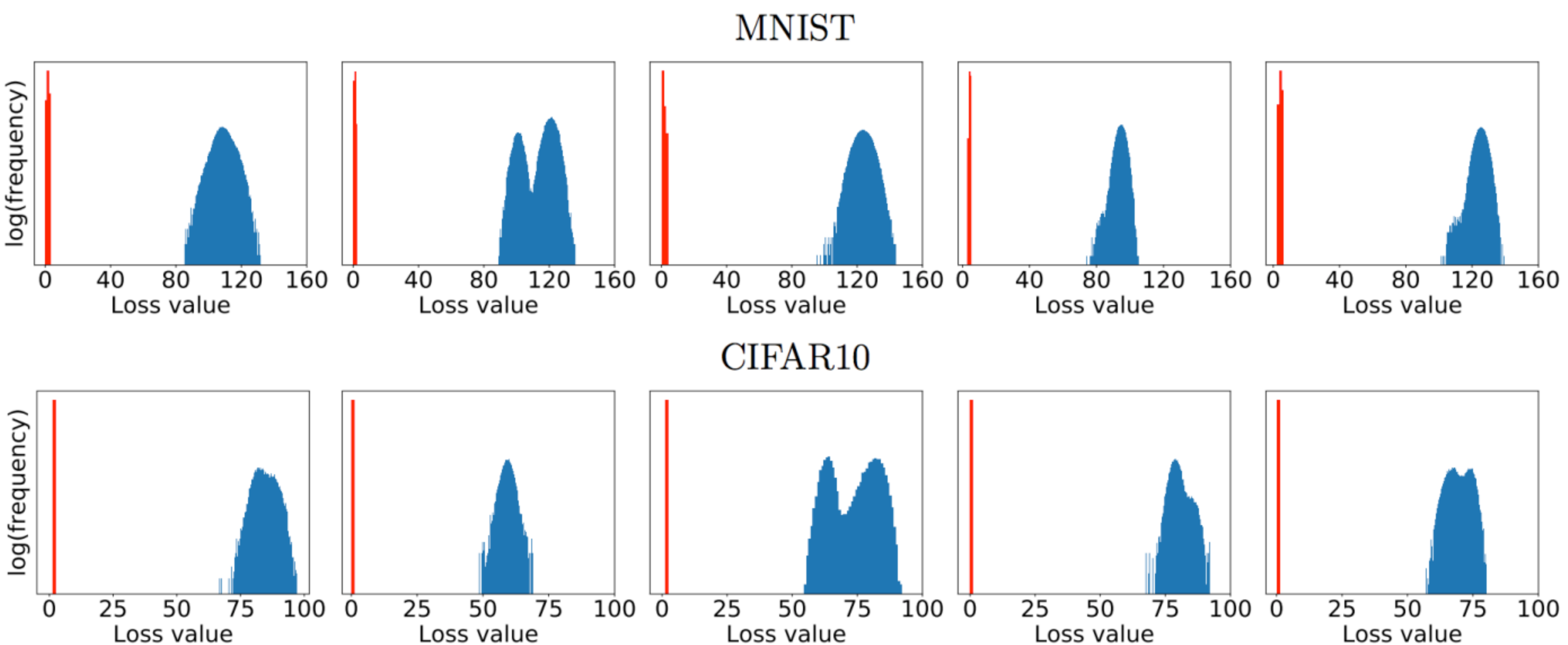


FGSM (single gradient)
PGD (100 steps with η=0.3)
Transfer FGSM
Transfer PGD

Explains failure of FGSM

Loss Landscape

Explore loss surface with randomly restarted PGD (100k trials):



Many local maxima, but loss values **concentrate.**

Results: Robust Classifiers?

Results

MNIST (eps = 0.3): 90% accuracy vs white-box
93% accuracy vs black-box

CIFAR10 (eps = 8): 46% accuracy vs white-box
63% accuracy vs black-box

Results: Robust Classifiers?

Results

MNIST (eps = 0.3): 90% accuracy vs white-box
93% accuracy vs black-box

CIFAR10 (eps = 8): 46% accuracy vs white-box
63% accuracy vs black-box

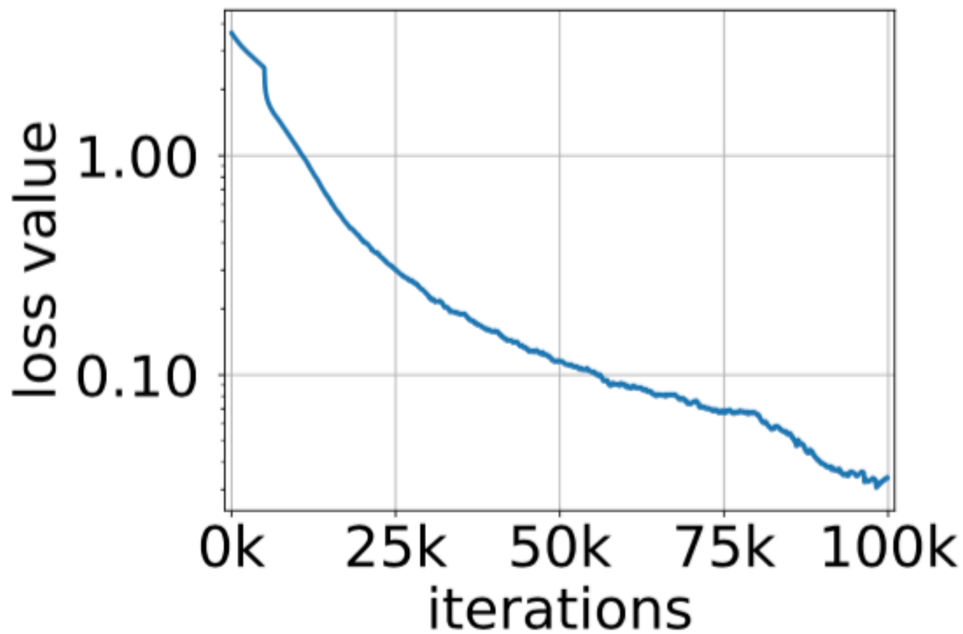
Public challenges since June (see github).

Top black-box attacks

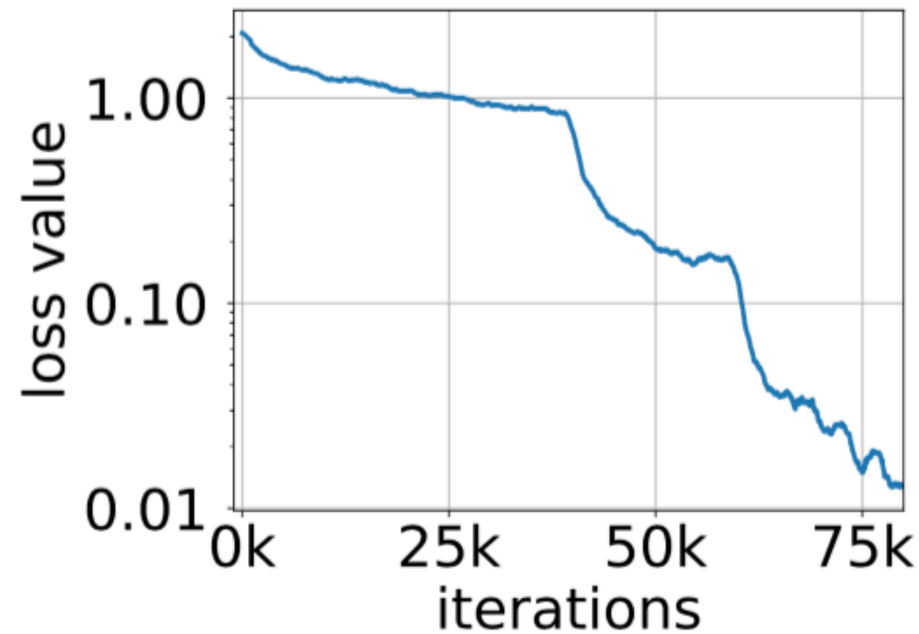
92.8% “Generating Adversarial Examples with Adversarial Networks”

93.5% PGD against three copies of the network (Florian Tramer)

What About CIFAR10?

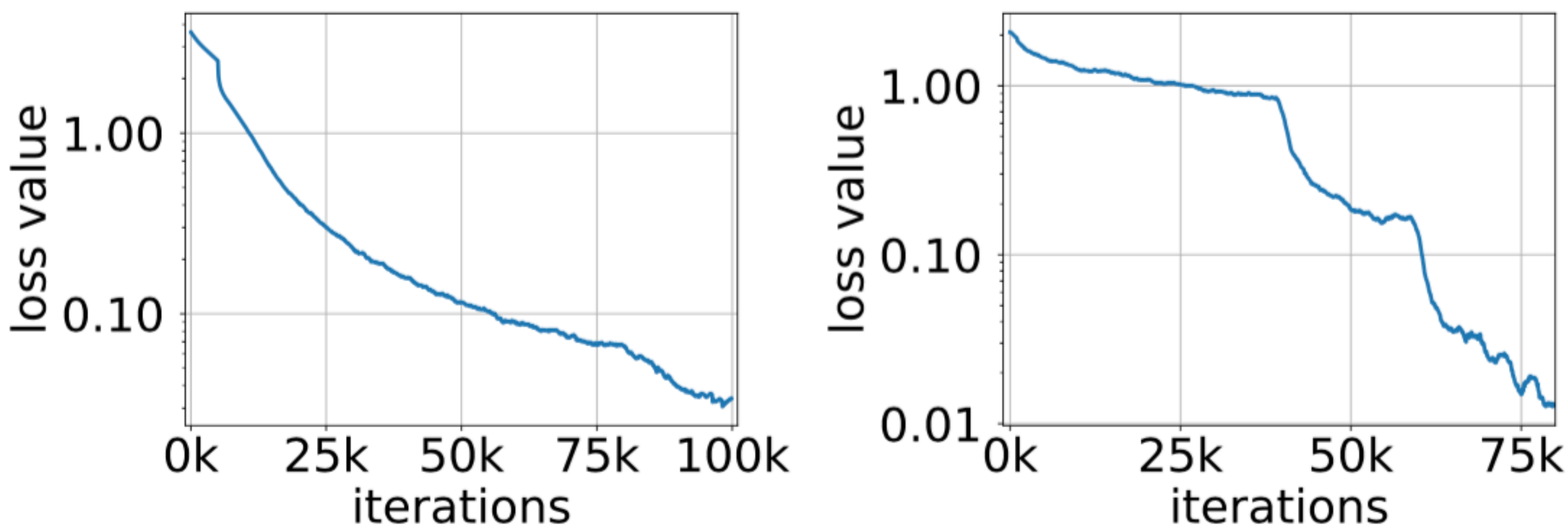


(a) MNIST



(b) CIFAR10

What About CIFAR10?



(a) MNIST (b) CIFAR10

Optimization succeeds, but the model **overfits** on CIFAR10:

100% train **adv.** accuracy, but only 48% on test.

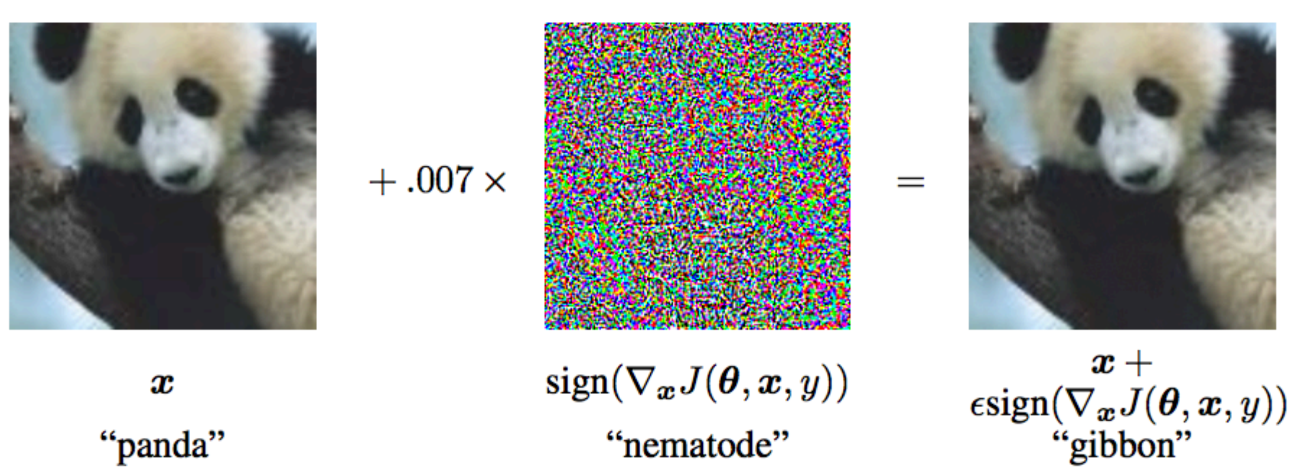
Robust Generalization

Does robustness require more data?

Theorem (informal): There is a distribution over points in R^d with the following property: Learning a ℓ_∞ robust linear classifier for this distribution requires $\sqrt{d}$ more samples than learning a non-robust classifier.

Conclusions

- **Robust generalization** is a prerequisite for secure ML.
- Adversarial training (a.k.a. robust optimization) **with strong enough attacks** is a principled defense.
- Optimization is only half of the picture: We need to take care of adversarially robust generalization too



Questions

- What robustness guarantees should ML-based systems provide?
- Are there trade-offs between robust and standard generalization?
- What compromises in mathematical rigor are acceptable?
- How can we verify ML-based systems?