Adversarial Robustness via Optimization Lens

Aleksander Mądry

Joint work with

madry-lab.ml

Aleksandar Makelov

Ludwig Schmidt

Dimitris Tsipras

Adrian Vladu

Why am I (are we?) here?

IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

WHY DEEP LEARNING IS SUDDENLY CHANGING YOUR LIFE

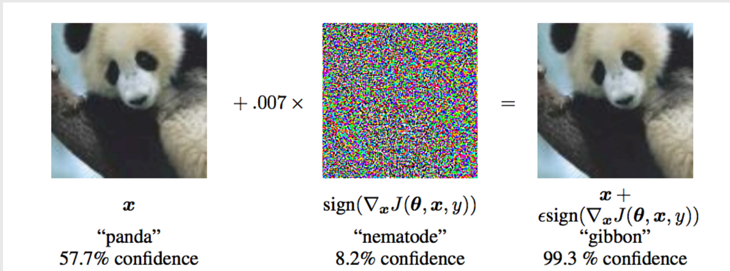
2016: The Year That Deep Learning Took Over the Internet

Crucial question:

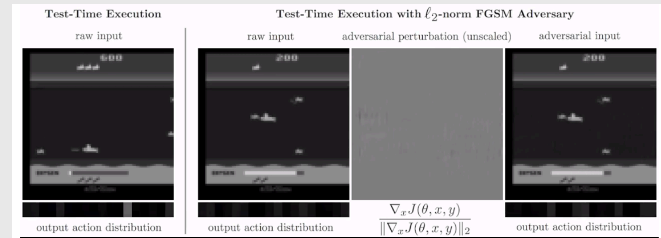
Can you **really** trust your deep learning model?

Goal: Make deep learning safe and reliable

Focus today: Adversarial Examples **[Szegedy et al. ‘14]**

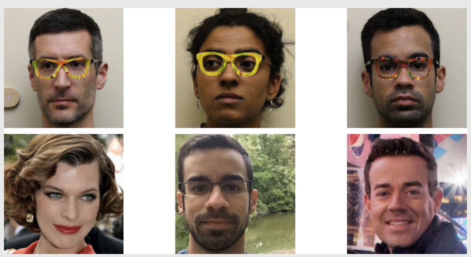


[Goodfellow et al. ’15]

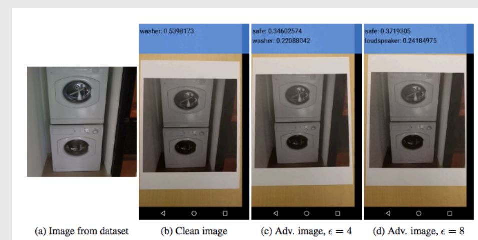


This is not only about pandas…

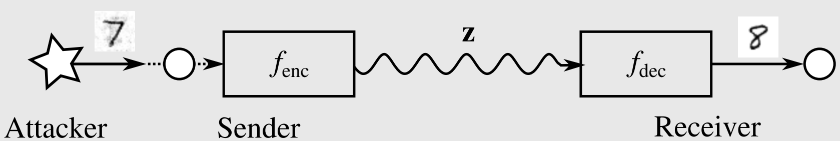
[Huang et al. ’17] [Behzadan-Munir ‘17]



[Sharif et al. ’16]



[Kurakin et al. ’17]



[Kos et al. ’17]

…or only about security

Our models do **not** generalize as reliably as we thought

Focus so far:

→ Exploration of the structure of adversarial examples

→ Mostly interest in their construction, i.e., attacks

→ Proposed defense mechanism tend to be bypassed by new, more sophisticated attacks

"Arms race" between attacks and defenses

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]

→ No good understanding yet of the extent to which one can or cannot be resistant to adversarial examples

Our work: Attempt a principled (re)look at adv. robustness

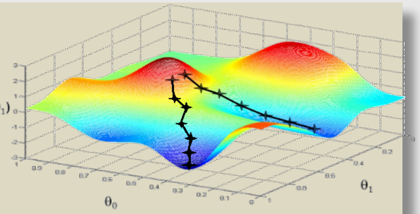
Three principles underlying our approach:

→ Be precise about your threat model, i.e., what you want to be secure against (and what is ok to be vulnerable to)

→ Use (robust) optimization as a lens on adv. robustness

→ Let the intended security guarantees be the driver of the design of the corresponding defense mechanism

Resulting framework:



→ Enables us to train **reliably*** robust models

→ Provides a perspective on adversarial robustness (that also unifies and explains much of previous findings)

Optimization-based View on Adversarial Robustness

$$\min_{\theta} \; E_D\,[\mathrm{loss}(\theta, x, y)]$$

Optimization-based View on Adversarial Robustness

$$\min_\theta E_D\,[\max_{\delta\in\Delta} \text{loss}(\theta, x+\delta, y)]$$

(Also see **[Huang et al. '15]** and **[Shaham et al. '15]**)

Δ = set of "allowed" adversarial perturbations (attack model)

Here: Focus on images & **Δ =** each pixel changed by **≤ ε**

Equivalently: $$\min_\theta E_D\,[\varphi(\theta,x,y)]$$

$$\varphi(\theta,x,y) = \max_{\delta\in\Delta} \text{loss}(\theta, x+\delta, y)$$ ("adversarial" loss)

Note: If we find **θ** that makes the objective small

⇒ security against <u>**any**</u> attack in **Δ**

So, now it is "just" about optimization

Evaluation of Adversarial Loss

$$\min_\theta E_D[\boldsymbol{\varphi}(\theta,x,y)]$$

$\boldsymbol{\varphi}(\theta,x,y) = \max_{\delta\in\Delta} \text{loss}(\theta, x+\delta, y)$ ("adversarial" loss)

Observe: Evaluation of adversarial loss
$\Leftrightarrow$ finding best attack

→ Quality of evaluation = reliability of the attacks
→ Most prior attacks thus correspond to evaluation of this adversarial loss (often in a quite ad-hoc manner)

What is the "best" way to evaluate adv. loss/attack?

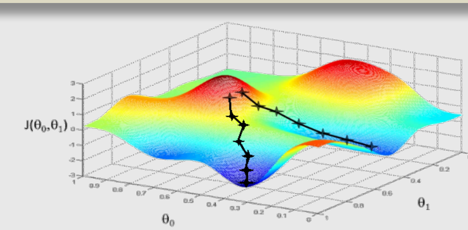
Evaluation of Adversarial Loss

$$\min_\theta E_D\,[\boldsymbol{\varphi}(\theta,x,y)]$$

$\boldsymbol{\varphi}(\theta,x,y) = \max_{\delta\in\Delta} \text{loss}(\theta, x+\delta, y)$ (“adversarial” loss)

A priori: Evaluating $\boldsymbol{\varphi}(\theta,x,y)$ corresponds to maximizing a **non-concave** function (loss)

What is the best we can do here?
(If loss has no special structure)



Natural (only?) approach: (Multi-step) projected gradient descent/ascent (PGD) with random restarts

Indeed: PGD leads to strong “first order” attacks

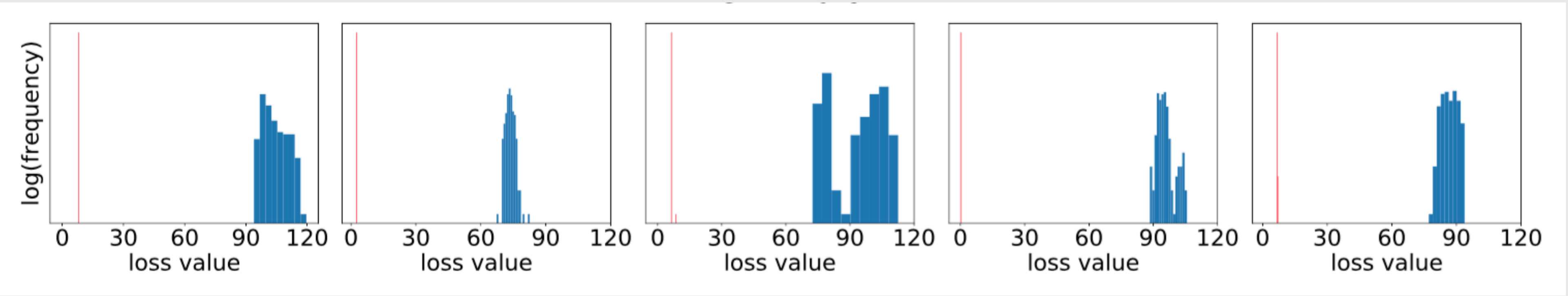
But why?

Evaluation of Adversarial Loss

$$\min_\theta E_D[\boldsymbol{\varphi}(\theta,x,y)]$$

$\boldsymbol{\varphi}(\theta,x,y) = \max_{\delta\in\Delta} \text{loss}(\theta, x+\delta, y)$ (“adversarial” loss)

Observation: Even though there is a lot of distinct local maxima of $\boldsymbol{\varphi}(\theta,x,y)$, their **values** are fairly concentrated



This suggests: Maxima we identify close to global ones ⇒ they give good descent directions (cf Danskin’s theorem)

Solving our saddle point problem

Recall: Evaluation of $\boldsymbol{\varphi}(\theta,x,y)$ ⟺ Finding best attack

Consequently: Solving our saddle point problem
⟺ Performing adversarial training

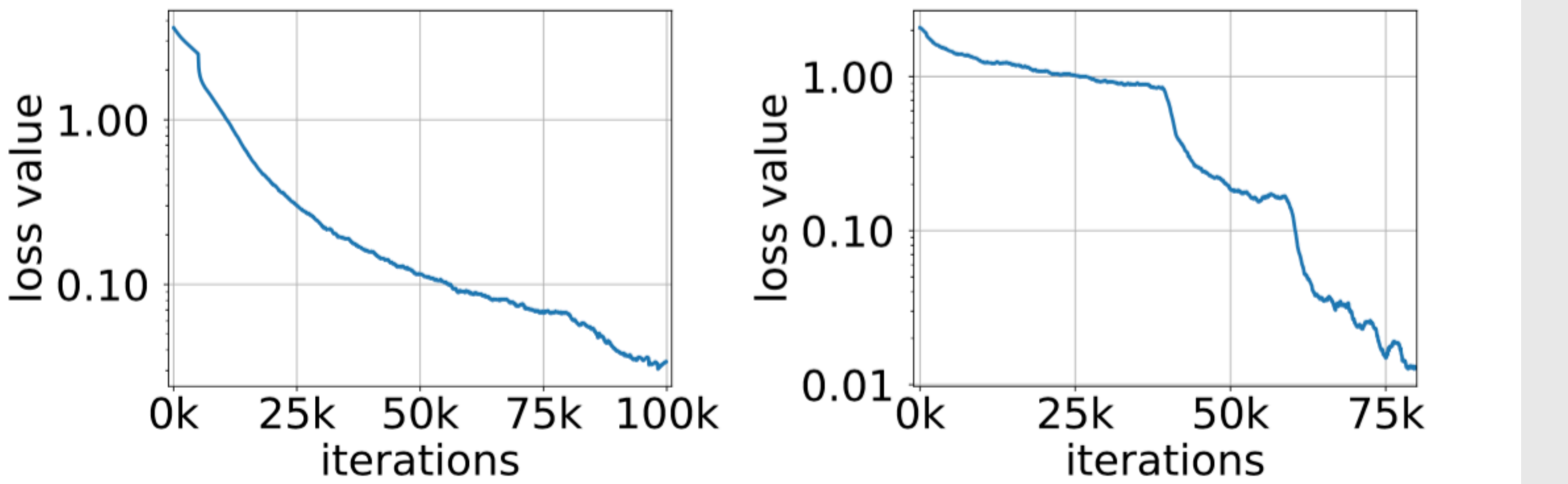
Our method = Best* adversarial training?

Key caveat: ”Reliability” of our attacks was verified
only from the “first order” perspective

⇒ Could have much better attacks/local maxima
we can’t easily access with first order methods

“First order” security model?

Solving our saddle point problem: Results



(a) MNIST (b) CIFAR10

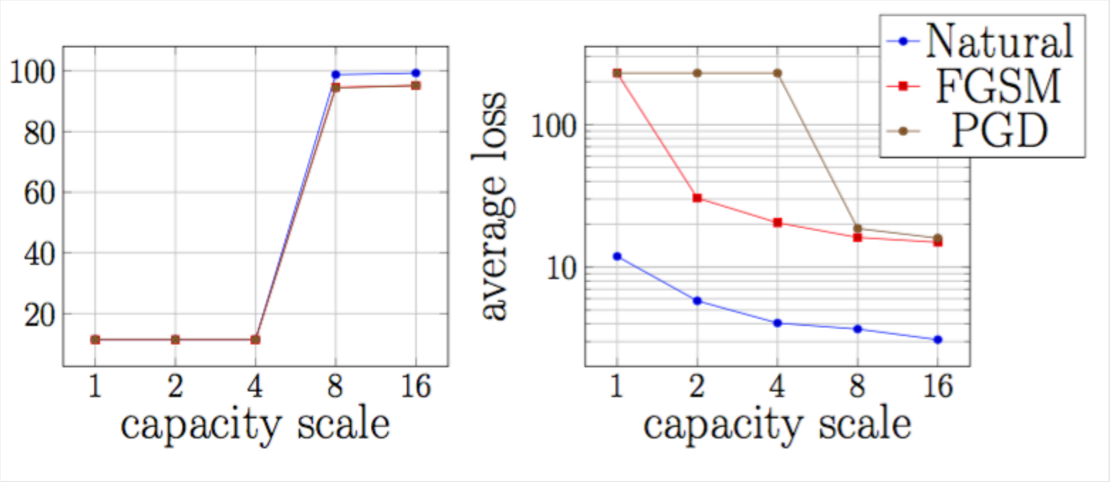
Our best models:

→ **MNIST (ε=0.3):** Accuracy 89% against the "best" white box attack and 95% against black box/transfer attacks

→ **CIFAR10 (ε=8):** Accuracy 46% (white box attack) and 64% (black box/transfer attack)

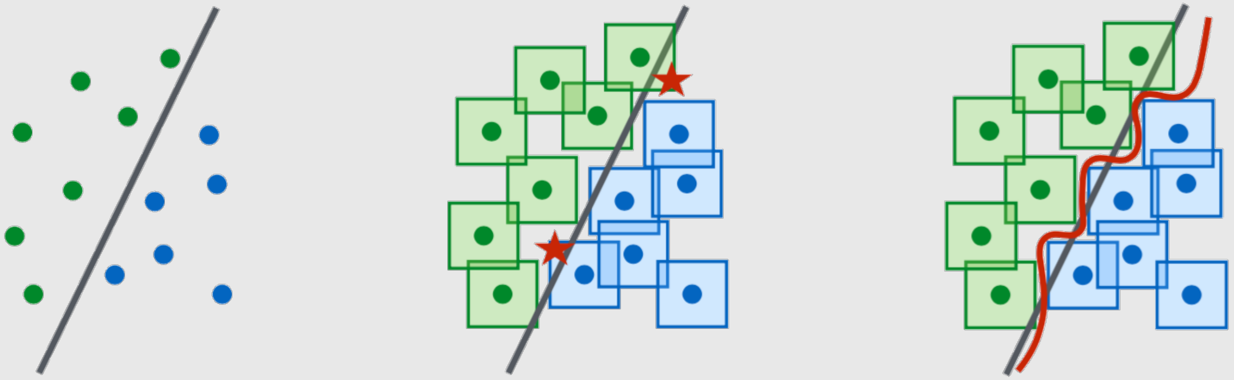
Important: Capacity of our model matters

Accuracy and loss vs. model capacity (PGD training on MNIST):



Why?

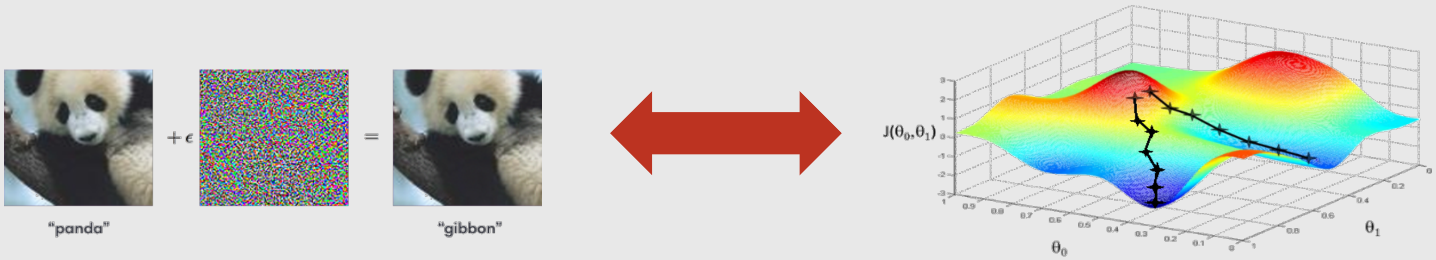
Need enough capacity to have the **final** value of our saddle point problem be small enough



Some Take Home Messages

→ Opt.-based perspective enables us to reason about adversarial robustness **guarantees** in a precise and principled manner

Key duality: If you can <u>reliably</u> attack it, you can also <u>reliably</u> defend



Attacks ⟺ Evaluation of adv. loss

Adv. training ⟺ Solving saddle point problem

→ Reliable optimization **and** enough capacity is crucial

(Most of quirks observed in past work seem to be tied to lack of one of these)

Truly adversarially robust ML might be possible after all!

Moving forward

→ Validate further the predictions of our framework
→ MNIST results pretty satisfying
but CIFAR10, although promising, still needs more work
→ Different data sets? Different/better attack models?
Non-differentiable attacks?
→ Faster training time/smaller models?

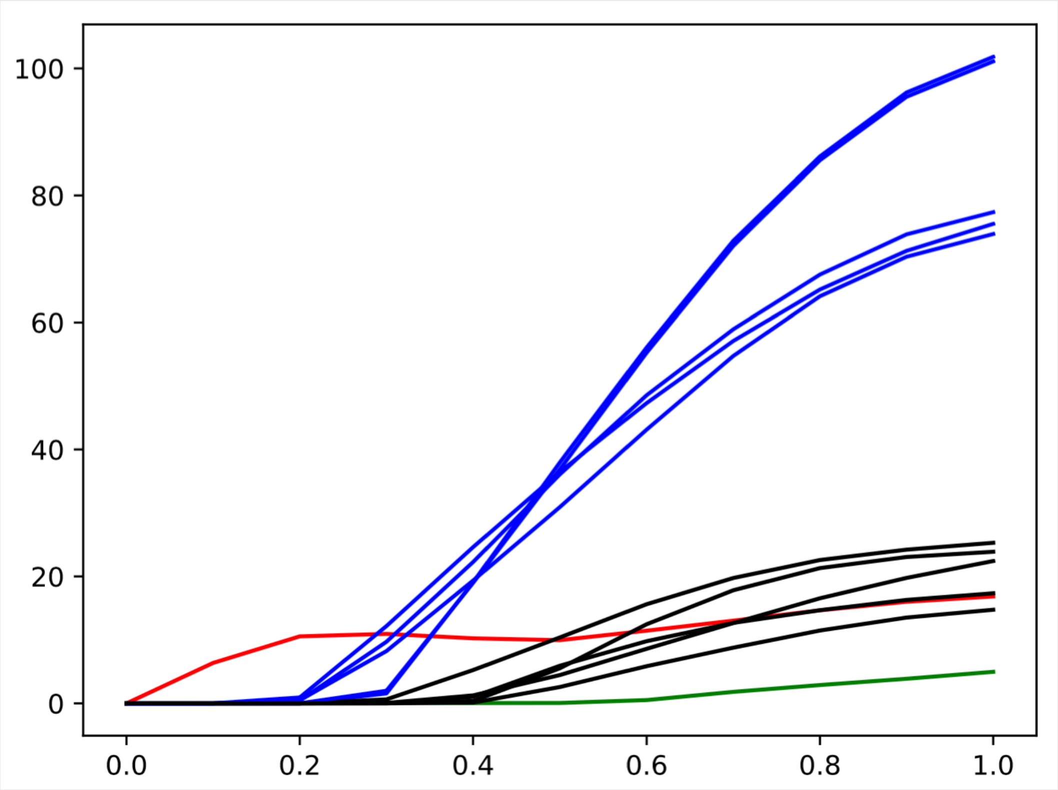
Also: MNIST/CIFAR10 black box/transfer security challenge

→ Break our model, because we couldn't
→ Details:
https://github.com/MadryProj/mnist_challenge
→ Aim to host more such challenges soon
(crucial to get truly reliable ML security)

Thank you

PGD = a universal "first order" adversary?

Change of loss in the direction identified by different attacks:



FGSM (single gradient)

PGD (100 steps with η=0.3)

Transfer FGSM

Transfer PGD

Why am I (are we?) here?

WHY DEEP LEARNING IS SUDDENLY CHANGING YOUR LIFE

IS "DEEP LEARNING" A REVOLUTION IN ARTIFICIAL INTELLIGENCE?

2016: The Year That Deep Learning Took Over the Internet

"Obvious" tantalizing question:
Why deep learning works (even though it "should" not)?

But: Would you **really** trust your deep learning model?

Can we make deep learning safe and reliable?