A THEORETICAL LOOK AT ADVERSARIAL EXAMPLES

Tom Goldstein

...and also...

Ali Shafahi, Ronny Huang,
Mahyar Najibi, Octavian Suciu,
Christoph Studer, Soheil Feizi, Tudor Dumitras

UNIVERSITY OF MARYLAND
OVERVIEW

What are adversarial examples, and what are their risks?

Poison attacks!

Are they an escapable problem?
ADVERSARIAL EXAMPLES

Add "tweaks" to image

Compute "tweak" to each pixel

One's hot label

ball  car  ↑  hat

grumpy cat
ADVERSARIAL ATTACKS

“Egyptian Cat” 28%

“Traffic Light” 97%
ADVERSARIAL ATTACKS

“Ox” 85%

“Traffic Light” 96%
SECURITY RISKS

Eykholt et al, 2018  Sharif et al, 2016
THREAT MODEL: POISON

Train-time attacks: adversary controls training data

Does this *actually* happen?

- Scraping images from the web
- Harvesting system inputs (spam detector)
- Bad actors/inside agents
HOW POISONING WORKS

Training data

Base

Testing example

Plane

Frog
HOW POISONING WORKS

Training data

Base

+ = Poison!

Testing example

Plane

Frog
HOW POISONING WORKS

Training data

Base

Poison!

Testing example

Plane

Frog
CLEAN-LABEL + TARGETED

Base + Poison!

Clean label: poisons are labeled “correctly”
Targeted: Performance only changes on selected target

Attacks can be executed by outsider
Poison data can be placed on the web
TWO CONTEXTS

**Transfer learning**
- Standard, pre-trained net is used
- “Feature extraction” layers frozen
- Classification layers re-trained
- Common practice in industry

“One-shot kill” possible

**End-to-end re-training**
- Pre-trained net is used
- All-layers are re-trained

Multiple poisons required
C O L L I S I O N  A T T A C K

\[ p = \arg \min_{x} \| f(x) - f(t) \|^2 + \beta \| x - b \|^2 \] (1)

Decision boundary

Base

Target
COLLISION ATTACK

\[ p = \arg\min_{x} \| f(x) - f(t) \|^2 + \beta \| x - b \|^2 \] (1)
COLLISION ATTACK

\[ p = \arg\min_{\forall x} \| f(x) - f(t) \|^2 + \beta \| x - b \|^2 \] (1)
Target instances from Fish class

Clean Base

Target instances from Dog class

Shafahi et al. “Poison frogs! Targeted poisoning attacks on neural nets”
Target instances from Fish class

Clean Base

Poison instances made for fish class from dog base

Poisons made for dog class from fish bases

Shafahi et al. “Poison frogs! Targeted poisoning attacks on neural nets”
Shafahi et al. “Poison frogs! Targeted poisoning attacks on neural nets”
Target instances from Fish class

Shafahi et al. “Poison frogs! Targeted poisoning attacks on neural nets”
Target instances from Fish class

Target instances from Dog class

Clean Base

Poisoned instances made for fish class from dog base instances

Poison fish

Shafahi et al. “Poison frogs! Targeted poisoning attacks on neural nets”
BLACK BOX ATTACK

\[ p = \text{argmin}_{\forall x} \| f(x) - f(t) \|^{2} + \beta \| x - b \|^{2} \] (1)
BLACK BOX ATTACK

\[
p = \arg\min \forall x \quad \| f(x) - f(t) \|^2 + \beta \| x - b \|^2
\]  

(1)
BLACK BOX ATTACK

\[ p = \arg \min_{x} \| f(x) - f(t) \|^2 + \beta \| x - b \|^2 \]
POISON POLYTOPE

Decision boundary

Base

Target

Zhu et al. "Transferable clean-label poisoning attacks"
POISON POLYTOPE

Decision boundary

Base

Target

Zhu et al. "Transferable clean-label poisoning attacks"
POISON POLYTOPE

Zhu et al. "Transferable clean-label poisoning attacks"
POISON POLYTOPE

Target (fish)
THEORY OF ADVERSARIAL EXAMPLES
ATTACK & DEFENSES

**Adversarial attacks**
- Szegedy et al, 2013
- Biggio et al, 2013

**Multi-stage attacks**
- Kurakin et al, 2016
- Tramer et al, 2017

**Optimization attacks**
- Carlini & Wagner ‘17

**Approximation attacks**
- Athalye et al, 2018

**Adversarial training**
- Goodfellow et al 2015

**Distillation**
- Papernot ‘16
- Zantedeschia ‘16

**Bounded relu**
- Meng & Chen ‘17

**MagNet**

**Thermometer Detection**
- Buckman ‘18
- Ma et al, ‘18

**Compression**
- Guo, ‘18

**GANs**
- Samangouei, ‘18

…and LOTS more
ARE ADVERSARIAL EXAMPLES INEVITABLE?
RELATED WORK

**K-nearest neighbors classifier**
“Analyzing the Robustness of Nearest Neighbors to Adversarial Examples”
Wang, Jha, Chaudhuri, 2017

**Datasets produced by GAN-type generator**
“Adversarial vulnerability for any classifier”
Fawzi, Fawzi, Fawzi, 2018

**Classes lie on concentric spheres**
“Adversarial spheres”
Gilmer, Metz, Faghri, Schoenholz, Raghu, Wattenberg, Goodfellow, 2018

**Most similar to ours...**
“The Curse of Concentration in Robust Learning”
Mahloujifar, Diochnos, Mahmood, 2018
ARE ADVERSARIAL EXAMPLES INEVITABLE?

**spoiler alert**

...and the answer is...

YES!

...if the adversary is strong enough.
ARE ADVERSARIAL EXAMPLES INEVITABLE?

...but computer scientists think...

NO!

Common assumptions...

Human perception is not exploitable

High dimensional spaces aren’t that weird
THE SETUP

Adversarial example

\[ \| x - \hat{x} \|_p < \epsilon. \]
TOY PROBLEM

Dimension

3
TOY PROBLEM

Dimension 3

Surface area 50%

Adversarial examples?
TOY PROBLEM

Dimension
3

Surface area
55%

$\epsilon = 0.1$
TOY PROBLEM

Dimension
100

Surface area
84%

$\epsilon = 0.1$
TOY PROBLEM

Dimension
1000

Surface area
99.8%

random sampling $\rightarrow$ adversarial susceptibility

$\epsilon = 0.1$

$A$ $B$
Theorem (Levy & Pellegrino, 1951)
The $\epsilon$-expansion of any set that occupies half the sphere is at least as big as the $\epsilon$-expansion of a semi-sphere.
WHAT ABOUT REALISTIC MODELS?
Images
Points in a unit cube
THE SETUP

**Images**
Points in a unit cube

**Class**
Probability density function on cube
(bounded by $U_c$)
THE SETUP

**Images**
Points in a unit cube

**Class**
Probability density function on cube (bounded by $U_c$)

**Classifier**
Partitions cube into disjoint sets (measurable)
“MOST”’THINGS ARE ADVERSARIAL

**Theorem**

Choose a class $c$ that occupies less than half the cube according to the classifier. Define...

$U_c :$ supremum of the density function for class $c$
"MOST" THINGS ARE ADVERSARIAL

**Theorem**
Choose a class $c$ that occupies less than half the cube according to the classifier. Define...

$U_c$ : supremum of the density function for class $c$

Sample a random point $x$ from the class distribution.
"MOST" THINGS ARE ADVERSARIAL

**Theorem**

Choose a class $c$ that occupies less than half the cube according to the classifier. Define...

$U_c :$ supremum of the density function for class $c$

Sample a random point $x$ from the class distribution. With probability at least

$$1 - U_c \exp(-\pi \epsilon^2)$$

One of the following conditions holds:

- $x$ is misclassified by the classifier
- $x$ has an adversarial example $\hat{x}$ with $\|x - \hat{x}\|_2 < \epsilon$. 

"Are adversarial examples inevitable?"  arXiv ‘18
"MOST" THINGS ARE ADVERSARIAL

\[ 1 - U_c \exp(-\pi \epsilon^2) \]

\[ \epsilon = 10 \]
WHAT HAPPENS IN THE ZERO NORM?

\[ \|x - \hat{x}\|_p < \epsilon. \]

\[ p = 0 \]

\[ \|x - \hat{x}\|_0 = \text{card}\{i | x_i \neq \hat{x}_i\} \]

**Sparse adversarial example**
SPARSE ATTACKS

“Ox”

3% pixels changed

“Traffic Light”
Theorem

Choose a class $c$ that occupies less than half the cube according to the classifier. Define...

$U_c : \text{supremum of the density function for class } c$

Sample a random point $x$ from the class distribution. With probability at least

$$1 - 2U_c \exp(-k^2/n)$$

One of the following conditions holds:

- $x$ is misclassified by the classifier
- The label of $x$ can be changed by modifying at most $k$ pixels.
WHAT ABOUT HIGH DIMENSIONS?
WHAT ABOUT HIGH Dimensions?

Clean

Adversarial

“dog” 9%

“traffic light” 97%
WHAT ABOUT HIGH DIMENSIONS?

Clean

Adversarial

“dog” 9%

“traffic light” 97%

90+% Robust

Shafahi et al. “Adversarial training for free!”
BOUND IN HIGH
DIMENSIONS

$1 - U_c \exp(-\pi \epsilon^2)\,$

Does this stay the same for large $n$?

NOPE!
BIG MNIST

28

56

112
Theorem

For all classifiers, a random image has an \( \varepsilon \)-adversarial example with probability \( p \).

For all classifiers, a random image has an \( 2\varepsilon \)-adversarial example with probability \( p \).
There is no fundamental relation between dimensionality and robustness!
MNIST hardened using PGD (30 steps)

High accuracy
ADVERSARIAL TRAINING

MNIST hardened using PGD (30 steps)

Low accuracy
ADVERSARIAL TRAINING

MNIST hardened using PGD (30 steps)
ADVERSARIAL TRAINING

MNIST hardened using PGD (30 steps)
WHAT AFFECTS ROBUSTNESS?

MNIST < susceptibility < CIFAR
WHAT AFFECTS ROBUSTNESS?

$$1 - U_c \exp(-\pi \epsilon^2)$$

concentration

pixels correlated
low-dimensional

low pixel correlations
high-dimensional
WHAT AFFECTS THE BOUND?

56x56 MNIST
- 3136 features
- 10 classes

CIFAR-10
- 3072 features
- 10 classes
ADVERSARIAL TRAINING

![Graph showing accuracy on adversarial examples for various datasets and perturbation levels.](image)

- **56x56 MNIST**
- **hardened big MNIST**
- **Original Big MNIST**

Accuracy on adversarial examples (%)
ADVERSARIAL TRAINING

Accuracy on adversarial examples (%) vs $\epsilon$

- 56x56 MNIST
- CIFAR-10
IMAGE COMPLEXITY LOWERS ROBUSTNESS

1 − \( U_c \exp(-\pi \epsilon^2) \)

“Complex” image classes have low density
lower pixel correlations
higher-dimensional manifolds

MNIST  CIFAR  ImageNet

complexity
TAKEAWAYS

Robustness has *fundamental* limits

Not specific to neural nets

Can’t escape by being clever

Robustness limit for neural nets might be far worse than intuition tells us!
Poison frogs! Targeted poisoning attacks on neural nets
Ali Shafahi, Ronny Huang, Mahyar Najibi, Octavian Suciu, C Studer, T Dimitras, T Goldstein

Transferable clean-label poisoning attacks
Chen Zhu, Ronny Huang, Ali Shafahi, Hengduo Li, Gavin Taylor, Chris Studer, Tom Goldstein

Adversarial training for free!
Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, Dickerson, Studer, Davis, Taylor, Goldstein

Are adversarial examples inevitable?
Ali Shafahi, Ronny Huang, Soheil Feize, Christoph Studer, Tom Goldstein