Introduction to Adversarial ML
What can Adversaries do to ML?
Manipulate Classification

“panda”
57.7% confidence

$+ \epsilon$

“gibbon”
99.3% confidence

https://openai.com/blog/adversarial-example-research/
Manipulate Classification

without the dataset the article is useless

okay google, browse to evil.com

https://nicholas.carlini.com/code/audio_adversarial_examples/
Manipulate Regression

\[ 31.51 + \text{[Random Image]} = 18.76 \]

BMI [Levin et al 2019]
Physical Attacks

Physical Attacks

Physical Attacks

Physical Attacks

Those were Test-Time Attacks

test $X$

$\downarrow$

$\theta$

$\downarrow$

$y$
Training-Time Attacks

\[ \min_{\theta} \mathbb{E}_{D^\dagger} \mathcal{L}(x, y, \theta) \]

Backdoor

\[ \theta^\dagger(x) = \theta(x) \]

\[ \theta^\dagger(x^\dagger) = \theta(x^\dagger) + 1 \]

Unfairness

\[ \min_{D^\dagger} \frac{P(\theta^\dagger(x) = 1 \mid \text{woman})}{P(\theta^\dagger(x) = 1 \mid \text{man})} \]

\[ \theta^\dagger = \arg \min_{\theta} \mathbb{E}_{D^\dagger} \ell(x, y, \theta) \]

Manipulate Model Interpretation
(linear regression)

Manipulate Model Interpretation
(latent Dirichlet allocation)

Manipulate Model Interpretation
(deep network attribution)

Model Stealing

Privacy Identification

\[ \in D? \]
Basic AdvML Math
Test-time Attack

$$\max_{\delta \in \Delta} \ell(x + \delta, y, \theta)$$
Feasible Set $\Delta$

$p$-norm ball

not match perception

Nonconvexity

projected gradient descent

\[ \delta \leftarrow \Pi_{\Delta} [ \delta + \eta \nabla \ell (x + \delta, y, \theta) ] \]

w.r.t. \( \delta \)

\[ \Pi_{\Delta} \]

\[ \eta \nabla \ell (x + \delta, y, \theta) \]

\[ \delta = 0 \]
Nonconvexity
projected gradient descent

\[ \delta \leftarrow \Pi_{\Delta} [\delta + \eta \nabla \ell(x + \delta, y, \theta)] \]

\[ \ell_{\infty} \text{ ball, } \eta \rightarrow \infty \]

\[ \delta = \epsilon \cdot \text{sign}(\nabla \ell(x + \delta, y, \theta)) \]

Nonconvexity

Projected gradient descent

\[ \delta \leftarrow \Pi_\Delta \left[ \delta + \eta \nabla \ell(x + \delta, y, \theta) \right] \]

Small \( \eta \)

PGD  Madry et al 2019  [link]
Nonconvexity

Projected gradient descent
\[ \delta \leftarrow \Pi_\Delta[\delta + \eta \nabla \ell(x + \delta, y, \theta)] \]

White-box attack: knows

Black-box attack: derivative-free optimization

With convex relaxation one may certify there is no adv. examples

(One) Defense against Test-time Attack

Adversarial Training

$$\min_{\theta} \mathbb{E}_D \max_{\delta \in \Delta} \ell(x + \delta, y, \theta)$$

Training-Set Poisoning

Bi-level optimization

\[
\min_{D^*, \theta} \|\theta - \theta^*\|
\]

\[
d(D^*, D) \leq \epsilon
\]

\[
\theta = \arg \min_{\nu} \mathbb{E}_{D^*} \ell(x, y, \nu)
\]

The Cat and Mouse Game

Attack
- Approximation attacks
  (Athalye et al., 2018)
- Optimization attacks
  (Carlini, 2017)
- Multi-stage attacks
  (Kurakin, 2016)
- Single Step attacks
  (Goodfellow, 2014)

Defense
- GANs
  (Samangouei et al., 2018)
- Detection
  (Ma et al., 2018)
- Distillation
  (Papernot et al., 2016)
- Adversarial training
  (Goodfellow et al., 2015)

Siegelmann 2019
What Next?

“Adversarial Machine Learning”
AdvML “at Large”

- Observation, Reward
- Action
- System
- Environment / Plant
- Adversarial attacks
Autonomous Vehicle

Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day

By James Vincent | @jjvincent | Mar 24, 2016, 6:43am EDT
Human-Bot Team Trust

AdvML at Large are Dynamic Systems

Attack = Optimal Control

\[
\min_{a_{0:T-1}} \sum_{t=0}^{T-1} \|s_{t+1} - s_{_{t+1}}^{\dagger}\| + \|a_t\|
\]

\[s_{t+1} = f(s_t, a_t)\]

... and Rational Agents
Game theory

Multi-class logistic regression is incentive incompatible

\[
\theta(\square) = \text{green}
\]

\[
\theta^\top(\square) = \text{red}
\]

Wu, Tzamos, Z. In preparation
Selected References

• Siegelmann. DARPA GARD Program. 2019

• Kolter and Madry. NeurIPS tutorial on Adversarial Robustness. 2018


• Zhu. An Optimal Control View of Adversarial Machine Learning. 2018