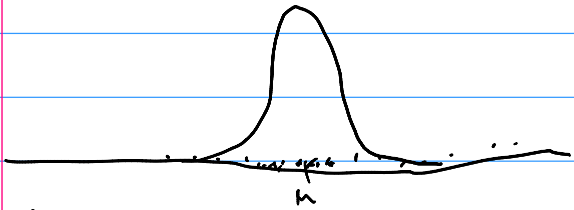
Robustness

- Model Misspecification
- Measurement error
- Dirty data
- Adversary

Simple example:

$X_1,\dots,X_n \sim N(\mu,1)$ Goal: Est μ from $X_1,\dots,X_n$



$\hat{\mu} = \frac{1}{n}\sum X_i$. If n is large enough, $\hat{\mu} \approx \mu$

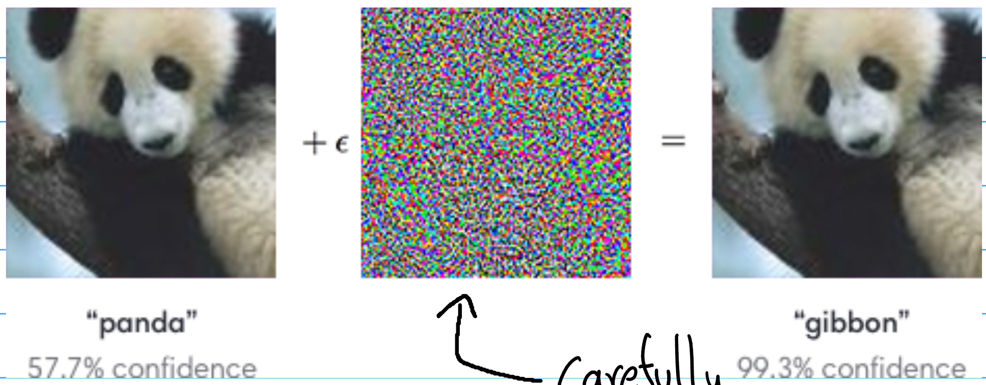
Adversary: Can add pts to dataset (few)

If • is very large ($>100n+\mu$) then $|\hat{\mu}-\mu|$ will be large (>100)

Defenses:

- Prune dataset
- Median, instead of mean

Today: Robustness to adversarial Examples



Carefully crafted

Setting: Model trained on some training data.
At test time, adversary can modify each point "a bit"
Adversary goal: Reduce test accuracy as much as possible

x' is An adversarial example for x on model f_θ if
1. (Informal) $x \approx x' \Leftrightarrow d(x, x')$ is small $\Leftrightarrow$ x and x' have same label according to human
2. $f_\theta(x) \neq f_\theta(x')$

Proxy for "human perception" in 1:
- Use dist instead

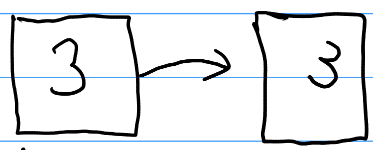
Common: ℓ_p dist $\left(\|x - x'\|_p = \left(\sum (x_i - x_i')^p \right)^{1/p} \right)$

Adversary may output x' in $\{ y : \|x - y\|_p \le \varepsilon \}$ ← small

Common: $p = 0, 2, \infty$ ← Today: Can change each pixel by $\le \varepsilon$.

$p = 0$: Can change ε pts arbitrarily

Other dists
- Wasserstein
- Translations, Rotations, Resizing



NN: 3 NN: 5?

Attacker:

Given trained model f_θ, test example x.
Construct x':

1. $\|x - x'\|_\infty \le \varepsilon$
2. $f_\theta(x) \ne f_\theta(x')$

Notes:
a) White-box
b) Untargeted vs Targeted attacks ← ($f_\theta(x') = c \ne f_\theta(x)$, for target c)

Optimize NN:

$$\min_\theta \frac{1}{n} \sum \ell(x_i, y_i, \theta)$$

Data fixed, optimize θ

$$\theta \leftarrow \theta - \eta \frac{1}{n} \sum \nabla_\theta \ell(x_i, y_i, \theta)$$

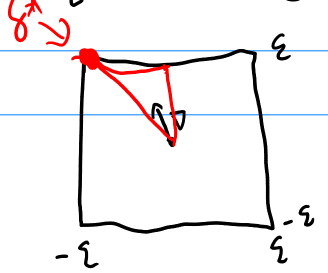
Generate Adversarial example

$$\delta' = \arg\max_{\delta \in \mathbb{R}^d} \ell(x+\delta, y, \theta) \quad \text{s.t. } \|\delta\|_\infty \le \varepsilon. \implies x' = x + \delta'$$

Gradient-Based opt

- Simple: Fast Gradient Sign Method (FGSM)

$\nabla_\delta \ell(x+\delta, y, \theta) \Rightarrow$ Take biggest step in ∇ direction, s.t. constraint

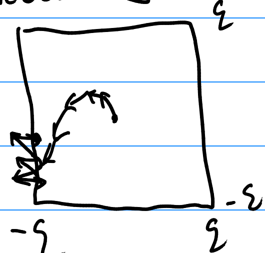


$$\delta^* = \varepsilon \cdot \text{sign}(\nabla_\delta \ell(x+\delta, y, \theta)) \in \{\pm\varepsilon\}^d$$

Better: Projected Gradient Descent (PGD)

$$\delta^{t+1} = \mathrm{Proj}\Big(\underbrace{\delta^t + \eta\, \mathrm{Sign}\big(\nabla_\delta \ell(x+\delta^t, y, \theta)\big)}_{\text{Same as before, but } \eta \text{ is hyper param}}\Big)$$

Projects into $[-\varepsilon, \varepsilon]^d$ if necessary



$$\mathrm{Proj}([3\varepsilon, -2\varepsilon, 0.5\varepsilon]) = [\varepsilon, -\varepsilon, 0.5\varepsilon]$$

Untargeted: $\max_\delta \ell(x+\delta, y, \theta)$ s.t. $\|\delta\|_\infty \le \varepsilon$

Targeted to c: $\max_\delta \ell(x+\delta, y, \theta) - \ell(x+\delta, c, \theta)$ s.t. $\|\delta\|_\infty \le \varepsilon$

Defenses

Adversarial Training

- Usual goal: $\min_\theta \mathbb{E}_{(x,y)\sim p}[\ell(x,y,\theta)]$
- Robust setting: $\min_\theta \mathbb{E}_{(x,y)\sim p}\Big[\max_{\delta: \|\delta\|_\infty \le \varepsilon} \ell(x+\delta, y, \theta)\Big]$

$\hookrightarrow \min_\theta \frac{1}{n} \sum \max_{\delta_i: \|\delta_i\| \le \varepsilon} \ell(x_i + \delta_i, y_i, \theta)$

How to solve

1. Draw minibatch B
2. For each (x_i, y_i) in B, compute $\delta_i^* = \operatorname{argmax}_{\delta_i: \|\delta_i\| \le \varepsilon} \ell(x_i + \delta_i, y_i, \theta)$ (use methods above)
3. $\theta \leftarrow \theta - \eta \frac{1}{|B|} \sum_{i \in B} \nabla_\theta \ell(x_i + \delta_i^*, y_i, \theta)$
4. Repeat

- Athalye, Carlini, Wagner '18
- ICLR '18 accepted 9 papers on adv defs.

Defense	Dataset	Distance	Accuracy
Buckman et al. (2018)	CIFAR	0.031 (ℓ_∞)	0%*
Ma et al. (2018)	CIFAR	0.031 (ℓ_∞)	5%
Guo et al. (2018)	ImageNet	0.005 (ℓ_2)	0%*
Dhillon et al. (2018)	CIFAR	0.031 (ℓ_∞)	0%
Xie et al. (2018)	ImageNet	0.031 (ℓ_∞)	0%*
Song et al. (2018)	CIFAR	0.031 (ℓ_∞)	9%*
Samangouei et al. (2018)	MNIST	0.005 (ℓ_2)	55%**
Madry et al. (2018)	CIFAR	0.031 (ℓ_∞)	47%
Na et al. (2018)	CIFAR	0.015 (ℓ_∞)	15%

Backdoor attacks