Regularization properties of adversarially-trained linear regression

Antônio H. Ribeiro[1,*], Dave Zachariah[1], Francis Bach[2], Thomas B. Schön[1]

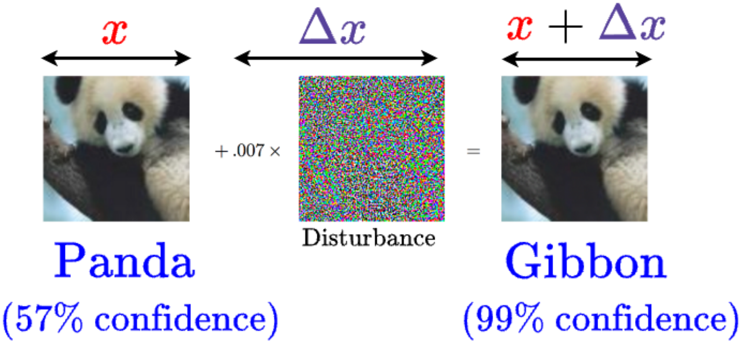
[1]Uppsala University, Sweden
[2]INRIA / PSL research university, France

[*]**Presenting**

NeurIPS

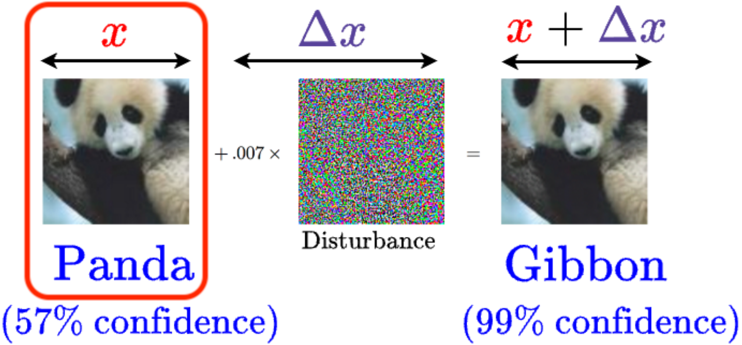
2023

Adversarial attacks



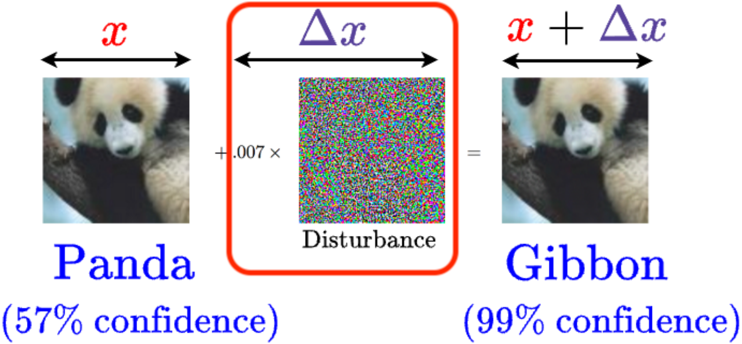
Explaining and Harnessing Adversarial Examples
I. J. Goodfellow, J. Shlens, C. Szegedy
ICLR (2015)

Adversarial attacks



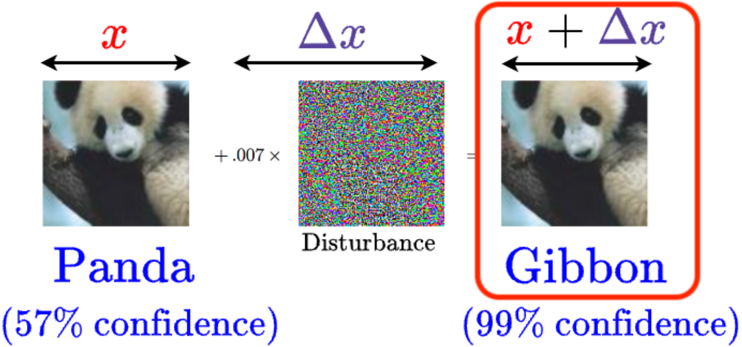
Explaining and Harnessing Adversarial Examples
I. J. Goodfellow, J. Shlens, C. Szegedy
ICLR (2015)

Adversarial attacks



Explaining and Harnessing Adversarial Examples
I. J. Goodfellow, J. Shlens, C. Szegedy
ICLR (2015)

Adversarial attacks



Explaining and Harnessing Adversarial Examples
I. J. Goodfellow, J. Shlens, C. Szegedy
ICLR (2015)

Adversarial training: *Each training sample is modified by an adversary.*

Adversarially-trained linear regression

- **Linear regression:**

$$\min_{\beta} \sum_{i=1}^{\#train} (y_i - \beta^\top x_i)^2$$

Adversarially-trained linear regression

- **Linear regression:**

$$\min_{\beta} \sum_{i=1}^{\#train} \Big(\underbrace{\boxed{y_i}}_{\text{observed}} - \underbrace{\boxed{\beta^\top x_i}}_{\text{linear prediction}} \Big)^2$$

Adversarially-trained linear regression

- **Linear regression:**

$$\min_{\beta} \sum_{i=1}^{\#train} (y_i - \beta^\top x_i)^2$$

- **Adversarial training** in linear regression:

$$(y_i - \beta^\top (x_i + \boxed{\Delta x_i}))^2$$

Adversarially-trained linear regression

- **Linear regression:**

$$\min_{\beta} \sum_{i=1}^{\#train} (y_i - \beta^\top x_i)^2$$

- **Adversarial training** in linear regression:

$$\max_{\|\Delta x_i\| \leq \delta} (y_i - \beta^\top (x_i + \Delta x_i))^2$$

Adversarially-trained linear regression

- **Linear regression:**

$$\min_{\beta} \sum_{i=1}^{\#train} (y_i - \beta^\top x_i)^2$$

- **Adversarial training** in linear regression:

$$\min_{\beta} \sum_{i=1}^{\#train} \max_{\|\Delta x_i\| \leq \delta} (y_i - \beta^\top (x_i + \Delta x_i))^2$$

Adversarially-trained linear regression

$$\sum_{i=1}^{\#train} \max_{\|\Delta x_i\| \leq \delta} (y_i - (x_i + \Delta x_i)^\mathsf{T} \beta)^2$$

Adversarially-trained linear regression

$$\sum_{i=1}^{\#train} \max_{\|\Delta x_i\| \leq \delta} (y_i - (x_i + \Delta x_i)^\mathsf{T} \beta)^2$$

It can be ***rewritten*** *as:*

$$\sum_{i=1}^{\#train} \left(|y_i - x_i^\mathsf{T} \beta| + \delta \|\beta\|_* \right)^2$$

where $\| \cdot \|_*$ is the ***dual norm****.*

Adversarially-trained linear regression

$$\sum_{i=1}^{\#train} \max_{\boxed{\|\Delta x_i\|_\infty \leq \delta}} (y_i - (x_i + \Delta x_i)^\mathsf{T} \beta)^2$$

It can be **rewritten** *as:*

$$\sum_{i=1}^{\#train} \left(|y_i - x_i^\mathsf{T} \beta| + \delta \boxed{\|\beta\|_1}\right)^2$$

where $\boxed{\|\cdot\|_1}$ *is the* **dual norm**.

Similarities with Lasso

- **ℓ_∞-adversarial attacks:**

$$\sum_{i=1}^{\#train} \left(|y_i - x_i^{\mathsf{T}} \beta| + \delta \|\beta\|_1 \right)^2$$

- **Lasso:**

$$\sum_{i=1}^{\#train} \left(|y_i - x_i^{\mathsf{T}} \beta| \right)^2 + \lambda \|\beta\|_1 .$$

Similarities with Lasso

- **ℓ_∞-adversarial attacks:**

$$\sum_{i=1}^{\#train} \left(|y_i - x_i^\mathsf{T}\beta| + \boxed{\delta\|\beta\|_1}\right)^2$$

- **Lasso:**

$$\sum_{i=1}^{\#train} \left(|y_i - x_i^\mathsf{T}\beta|\right)^2 + \boxed{\lambda\|\beta\|_1}.$$

Similarities with Lasso

- **ℓ_∞-adversarial attacks:**

$$\sum_{i=1}^{\#train} \left(|y_i - x_i^\mathsf{T} \beta| + \boxed{\delta \|\beta\|_1} \right)^2$$

- **Lasso:**

$$\sum_{i=1}^{\#train} \left(|y_i - x_i^\mathsf{T} \beta| \right)^2 + \boxed{\lambda \|\beta\|_1}.$$

Main results:

#1. **Map** $\lambda \leftrightarrow \delta$ for which they yield the **same result.**

Similarities with Lasso

- **ℓ_∞-adversarial attacks:**

$$\sum_{i=1}^{\#train} \left(|y_i - x_i^\mathsf{T} \beta| + \boxed{\delta \|\beta\|_1} \right)^2$$

- **Lasso:**

$$\sum_{i=1}^{\#train} \left(|y_i - x_i^\mathsf{T} \beta| \right)^2 + \boxed{\lambda \|\beta\|_1}.$$

Main results:

#1. **Map** $\lambda \leftrightarrow \delta$ for which they yield the **same result.**

#2. **More parameters than data**: abrupt transition into interpolation.

Similarities with Lasso

- **ℓ_∞-adversarial attacks:**

$$\sum_{i=1}^{\#train} \left(|y_i - x_i^\mathsf{T} \beta| + \boxed{\delta \|\beta\|_1} \right)^2$$

- **Lasso:**

$$\sum_{i=1}^{\#train} \left(|y_i - x_i^\mathsf{T} \beta| \right)^2 + \boxed{\lambda \|\beta\|_1}.$$

Main results:

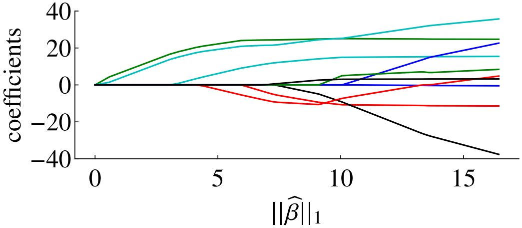
#1. **Map** $\lambda \leftrightarrow \delta$ for which they yield the **same result.**

#2. **More parameters than data**: abrupt transition into interpolation.

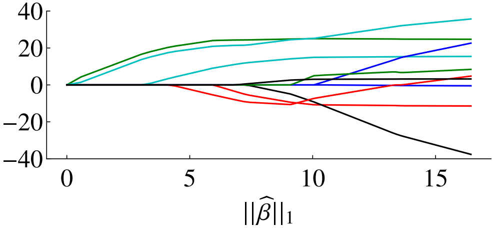
#3. **Optimal choice** of δ independent on noise level.

1. Equivalence with Lasso

Map $\lambda \leftrightarrow \delta$ for which they yield the **same result.**



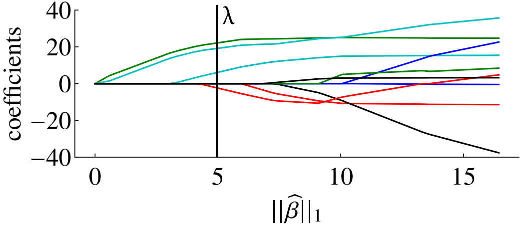
Lasso.



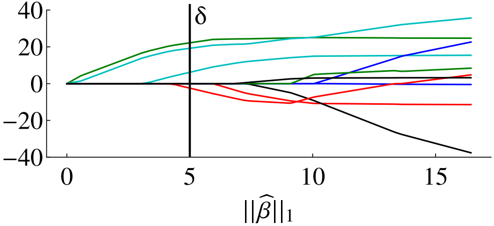
ℓ_∞-adv. training.

1. Equivalence with Lasso

Map $\lambda \leftrightarrow \delta$ for which they yield the **same result.**



Lasso.



ℓ_∞-adv. training.

The that yield the **same result** are **not** necessarily the same, i.e.: $\delta \neq \lambda$

2. More parameters than data

Lasso: transition **only in the limit**

$$\lambda \to 0^+ \Rightarrow \text{Mean square error} \to 0$$

2. More parameters than data

Lasso: transition **only in the limit**

$$\lambda \to 0^+ \Rightarrow \text{Mean square error} \to 0$$

Adversarial training:

$$\delta \in (0, \text{threshold}] \Rightarrow \text{Mean square error} = 0$$

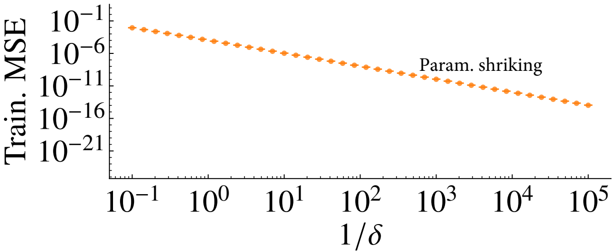
\# 2. More parameters than data

Lasso: transition **only in the limit**

$$\lambda \to 0^{+} \Rightarrow \text{Mean square error} \to 0$$

Adversarial training:

$$\delta \in (0, \text{threshold}] \Rightarrow \text{Mean square error} = 0$$



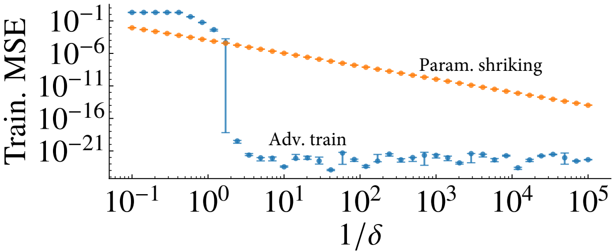
2. More parameters than data

Lasso: transition **only in the limit**

$$\lambda \to 0^+ \Rightarrow \text{Mean square error} \to 0$$

Adversarial training:

$$\delta \in (0, \text{threshold}] \Rightarrow \text{Mean square error} = 0$$



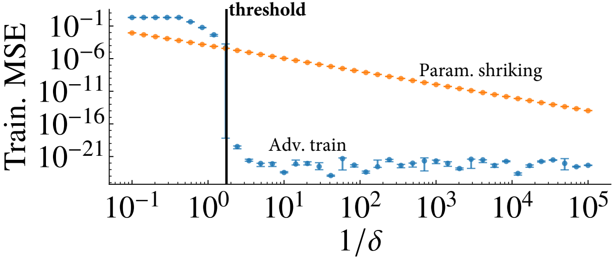
2. More parameters than data

Lasso: transition **only in the limit**

$$\lambda \to 0^+ \Rightarrow \text{Mean square error} \to 0$$

Adversarial training:

$$\delta \in (0, \text{threshold}] \Rightarrow \text{Mean square error} = 0$$



2. Equivalence with minimum norm interpolator

For $\delta \in (0, \mathrm{threshold}]$, the minimum-norm interpolator is the solution to adversarial training.

2. Equivalence with minimum norm interpolator

For $\delta \in (0, \mathrm{threshold}]$, the minimum-norm interpolator is the solution to adversarial training.

Relevance

Connect **adversarial training** with **double descent** and **benign overfitting**

3. Invariance to noise levels

To obtain near-oracle performance.

- *Lasso:*

$$\lambda \propto \sigma \sqrt{\log(\#\mathit{params}) / \#\mathit{train}}$$

- ℓ_∞*-adversarial attack*:

$$\delta \propto \sqrt{\log(\#\mathit{params}) / \#\mathit{train}}$$

3. Invariance to noise levels

To obtain near-oracle performance.

- *Lasso:*

$$\lambda \propto \underbrace{\sigma}_{\text{unknown}} \sqrt{\log(\#\textit{params})/\#\textit{train}}$$

- ℓ_∞-*adversarial attack*:

$$\delta \propto \sqrt{\log(\#\textit{params})/\#\textit{train}}$$

Data model

$$y = \underbrace{x^\top \beta^*}_{\text{signal}} + \underbrace{\sigma}_{\text{noise std.}} \varepsilon.$$

arXiv:2310.10807

- ℓ_2-**adv.** attacks and **ridge regression.**

arXiv:2310.10807

- ℓ_2**-adv.** attacks and **ridge regression.**
- Generalization to **other loss** functions

arXiv:2310.10807

- ℓ_2**-adv.** attacks and **ridge regression.**
- Generalization to **other loss** functions
- Connection to **robust regression** and $\sqrt{\text{Lasso}}$.