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]ÍNRIA/PSL Research University (France)
[*]*contact: antonio.horta.ribeiro@it.uu.se*

Adversarial Examples

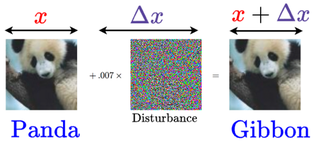


Fig.: Adv. example on image classification.

[1] I. J. Goodfellow, J. Shlens, C. Szegedy, *"Explaining and Harnessing Adversarial Examples"*, ICLR 2015.

Adversarial training in linear models

Adversarially-trained linear regression minimizes:

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

Cost function is **convex**. *It can be* **rewritten** *as:* [2]

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

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

[2] **A H Ribeiro** and T B Schön, *"Overparameterized Linear Regression under Adversarial Attacks"*, IEEE Transactions on Signal Processing, 2023

Pairs of dual norms

ℓ_2-adversarial attacks	
Adv. constraint: $\|\Delta x\|_2 \leq \delta$	Cost function
	$\sum_{i=1}^{n} \left(\|y_i - x_i^\mathsf{T}\beta\| + \delta\|\beta\|_2\right)^2$
ℓ_∞-adversarial attacks	
Adv. constraint: $\|\Delta x\|_\infty \leq \delta$	Cost function
	$\sum_{i=1}^{n} \left(\|y_i - x_i^\mathsf{T}\beta\| + \delta\|\beta\|_1\right)^2$

Motivation

How does adversarial training compare with other regularization methods?

Regularization methods:

1. Parameter shrinking methods.

2. Minimum-norm interpolators.

3. Robust regression and $\sqrt{\text{Lasso}}$.

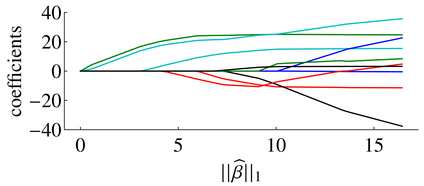
#1: Equivalence with parameter shrinking

Backgroud:

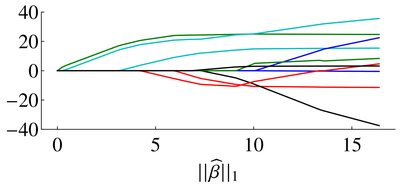
- Ridge: $\min_\beta \sum_{i=1}^{n} \left(|y_i - x_i^\mathsf{T}\beta|\right)^2 + \lambda\|\beta\|_2^2$.
- Lasso: $\min_\beta \sum_{i=1}^{n} \left(|y_i - x_i^\mathsf{T}\beta|\right)^2 + \lambda\|\beta\|_1$.

Our result: **map** $\lambda \leftrightarrow \delta$ *for which the results are asymptotically equivalent if* $\mathbb{E}[x] = 0$ *and* $x \sim -x$.

Numerical experiments:



(a) Lasso.



(b) ℓ_∞-adv. training.

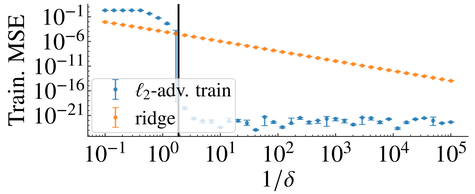
More parameters than data

Lasso and ridge regression: transition **only in the limit**

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

Adversarial training: we observe an **abrupt transition**:

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



#2: Minimum norm interpolators

Definition: Minimum $\|\cdot\|_*$-norm interpolator:

$$\min_\beta \|\beta\|_* \quad \text{subject to} \quad x_i\beta = y_i, \forall i.$$

Backgroud: (a) $\beta^{\text{ridge}}(\lambda) \to \beta^{\min-\ell_2}$; (b) $\beta^{\text{lasso}}(\lambda) \to \beta^{\min-\ell_1}$.

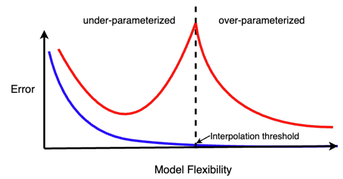
Our result: *When* $\#train < \#params$ *and* $\delta \in (0, \text{threshold}]$

(a) *Minimum* ℓ_2*-norm interpolator* **is solution to** ℓ_2*-adv. training.*

(b) *Minimum* ℓ_1*-norm interpolator* **is solution to** ℓ_∞*-adv. training.*

Why minimum-norm interpolators?

- **Gradient descent** on linear regression ($\beta^0 = 0$) converges to $\beta^{\min-\ell_2}$.
- Used in the study of **double-descent** [3] and **benign-overfitting**.



[3] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," PNAS 2019.

Invariance to noise magnitude

Background: If data is generated as

$$y = \underbrace{x^\top\beta^*}_{\text{signal}} + \sigma \underbrace{\varepsilon}_{\mathcal{N}(0,1)}$$

For Lasso, **near-oracle** performance is attained with:

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

Our result: *For* ℓ_∞*-adv. attack,* **near-oracle** *performance is attained with*

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

That is, the adv. radius δ can be set **without knowledge of the noise magnitude**.

#3: Relation to robust regression and $\sqrt{\text{Lasso}}$

$\sqrt{\text{Lasso}}$: Has the **same property**. It is the estimator: [4]

$$\min_\beta \sqrt{\textstyle\sum_{i=1}^{n} |y_i - x_i^\top\beta|^2} + \lambda\|\beta\|_1.$$

and it attains **near-oracle** performance if:

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

Robust regression: [5]

$$\min_\beta \max_{\Delta\in\mathcal{S}} \|y - (X + \Delta)\beta\|_2.$$

It is **equivalent** to $\sqrt{\text{Lasso}}$ if **columns** are constrained: $\|\Delta_{i,:}\|_2 \leq \delta, \forall i$.

Our result: *Robust regression is* **equivalent** *to adversarial training if* **rows** *are constrained:* $\|\Delta_{:,j}\|_2 \leq \delta, \forall j$.

[4] A. Belloni, V. Chernozhukov, and L. Wang, *"Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming,"* Biometrika 2011.
[5] H. Xu, C. Caramanis, and S. Mannor, *"Robust regression and lasso,"* NIPS 2008

Future work

- **Error-in-variables**: adv. training making it more robust to input disturbance Δx that are stochastic.
- **Tailored solver:** enableuse in high-dimensional applications (genetics).
- **Nonlinear models:** most results still hold for inputs in infinite spaces.