Adversarial training in linear regression

Antônio Horta Ribeiro
Uppsala University

Seminar on Advances in Probabilistic Machine Learning
Aalto University and ELLIS unit Helsinki
Online, Nov 2022

Outline

Motivation

Adversarial training

Robustness in high-dimensions

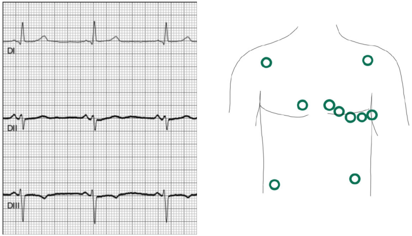
Electrocadiogram exam

Goal: *Build data-driven ECG analysis tools.*

- The ECG is the major diagnostic tool.
- Cardiovascular diseases: 32% of all deaths (GBD 2019).
- Example. CODE dataset: annotated historical data
  $n =$1.6M patients

- Model for automatic diagnosis:

  A. H. Ribeiro , M.H. Ribeiro, Paixão, G.M.M., et al. "Automatic diagnosis of the 12-lead ECG using a deep neural network," Nature Communications, 2020



Left: ECG signal **Right:** Electrode placement.

Adversarial examples

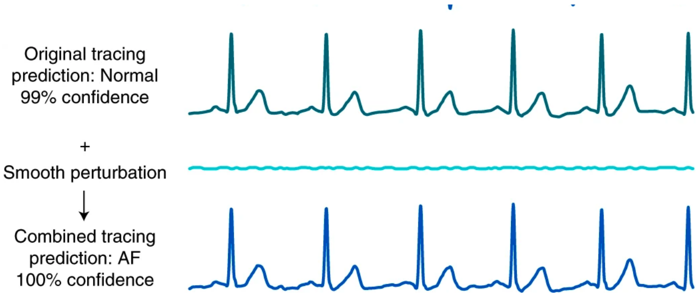


Figure: Effect of adversarial examples on ECG Classification.

Source: Han, X., Hu, Y., Foschini, L. et al. Deep learning models for electrocardiograms are susceptible to adversarial attack. Nature Medicine 26, 360–363 (2020).

Adversarial Training: *Model is trained training on samples that have been modified by an adversary.*

Adversarial training

"Is it fundamentally different than other regularization methods?"'

Surprises in adversarially-trained linear regression (2022). **Antônio H. Ribeiro**, Dave Zachariah, Thomas B. Schön. arXiv:2205.12695.

Adversarial robustness

"what is the role of high-dimensionality in model robustness?"

Overparameterized Linear Regression under Adversarial Attacks (2022). **Antônio H. Ribeiro**, Thomas B. Schön. arXiv:2204.06274

Outline

Motivation

Adversarial training

Robustness in high-dimensions

Framework: Linear regression

Simplest case where adversarial vulnerability has been observed.

I. J. Goodfellow, J. Shlens, C. Szegedy , *"Explaining and Harnessing Adversarial Examples"*, ICLR 2015
D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Ma, *"Robustness May Be At Odds with Accuracy,"* ICLR, p. 23, 2019.

- Training dataset:
$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n) \Rightarrow \widehat{\beta}$$
- Model prediction
$$\widehat{y} = \widehat{\beta}^\mathsf{T} x$$
- $\text{Error}(\widehat{\beta}) = |y - x^\mathsf{T}\widehat{\beta}|$
- $\text{Adv-error}(\widehat{\beta}) = \max_{\|\Delta x\| \le \delta} \left| y - (x + \Delta x)^\mathsf{T}\widehat{\beta} \right|$

Adversarial training

Empirical risk minimization:

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^\mathsf{T} \beta)^2$$

Adversarial training:

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

Adversarial error in linear regression

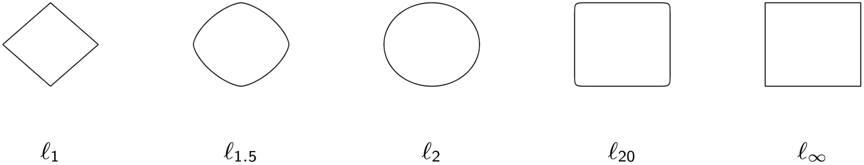
- $\text{Error}(\widehat{\beta}) = |y - x^{\mathsf{T}}\widehat{\beta}|$
- $\text{Adv-error}(\widehat{\beta}) = \max_{\|\Delta x\| \leq \delta} \left| y - (x + \Delta x)^{\mathsf{T}}\widehat{\beta} \right|$
- *Dual formula for the adversarial error*

$$\left(\text{Adv-error}(\widehat{\beta})\right)^2 = \left(|\text{Error}(\widehat{\beta})| + \delta\|\widehat{\beta}\|_*\right)^2$$

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

ℓ_p-adversarial attacks

- ℓ_∞-adversarial attack: $\{\|\Delta x\|_\infty \leq \delta\} \Rightarrow$ dual norm: $\|\Delta x\|_1$
- ℓ_2-adversarial attack: $\{\|\Delta x\|_2 \leq \delta\} \Rightarrow$ dual norm: $\|\Delta x\|_2$
- ℓ_p-adversarial attack: $\{\|\Delta x\|_p \leq \delta\} \Rightarrow$ dual norm: $\|\Delta x\|_q$ for $1/p + 1/q = 1$



Consequences to adversarial training

- Adversarial training,

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

can be reformulated as

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

The above expression is **convex**

Lasso and ℓ_∞-adversarial training

- ℓ_∞-adversarial training:

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

- Lasso:

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

Ridge regression and ℓ_2-adversarial training

- ℓ_2-adversarial training:

$$\frac{1}{n}\sum_{i=1}^{n}\left(|y_i - x_i^\mathsf{T}\beta| + \delta\|\beta\|_2\right)^2$$

- Ridge:

$$\frac{1}{n}\sum_{i=1}^{n}\left(|y_i - x_i^\mathsf{T}\beta|\right)^2 + \delta\|\beta\|_2^2$$

Diabetes example

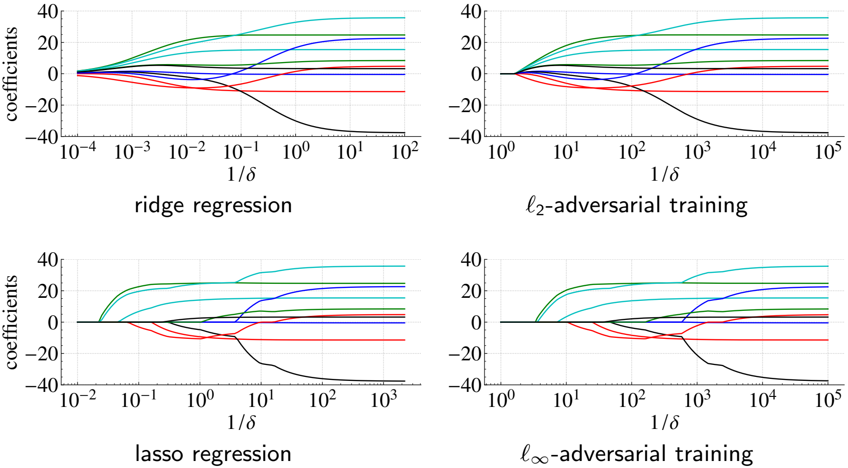


Figure: Regularization paths.

Overparametrized models and interpolators

Can a model perfectly fit the training data and still generalize well?

- **Benign overfitting**
  P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," Proceedings of the National Academy of Sciences, vol. 117, no. 48, pp. 30063–30070, Apr. 2020.
- **Double descent**
  M. Belkin, D. Hsu, S. Ma, and S. Mandal , "Reconciling modern machine-learning practice and the classical bias–variance trade-off," PNAS (2019)
- **Example:**

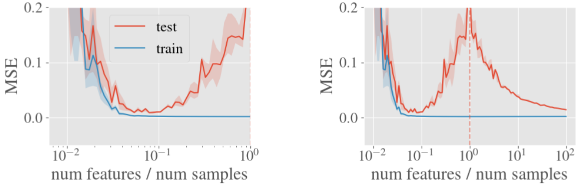


Figure: nonlinear ARX mean squared error (MSE).

A. H. Ribeiro, J. N. Hendriks, A. G. Wills, T. B. Schön. "Beyond Occam's Razor in System Identification: Double-Descent when Modeling Dynamics". IFAC SYSID 2021. *Honorable mention:* **Young author award**

Minimum-norm solution

Minimum ℓ_2-norm solution

$$\min_{\beta} \|\beta\|_2 \quad \text{subject to} \quad X\beta = y$$

- Gradient descent in linear regression converges to $\widehat{\beta}^{\text{min}-\ell_2}$.
- Ridge $\widehat{\beta}^{\text{ridge}}(\delta) \to \widehat{\beta}^{\text{min}-\ell_2}$ as $\delta \to 0^+$.

Minimum ℓ_1-norm solution

$$\min_{\beta} \|\beta\|_1 \quad \text{subject to} \quad X\beta = y$$

- Basis pursuit: i.e. allow you to recover sparse signals.
- Ridge $\widehat{\beta}^{\text{lasso}}(\delta) \to \widehat{\beta}^{\text{min}-\ell_1}$ as $\delta \to 0^+$ (LARS algorithm)..

Interpolation for finite δ

Theorem

For $0 < \delta < \bar{\delta}$, adversarial training is minimized at some $\widehat{\beta}$ that satisfies:

$$X\widehat{\beta} = y$$

Corollary

$\widehat{\beta}^{\text{min-}\ell_2}$ is the solution to ℓ_2-adversarial training for all $0 < \delta < \bar{\delta}$.

Corollary

$\widehat{\beta}^{\text{min-}\ell_1}$ is the solution to ℓ_∞-adversarial training for all $0 < \delta < \bar{\delta}$.

Overparametrized model

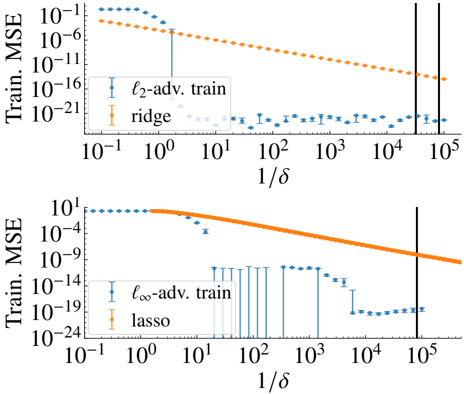


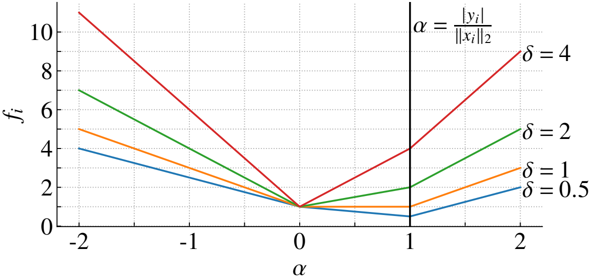
Figure: Training MSE *vs* regularization parameter.

Discussion

- New interpretation for minimum-norm solution.
- Distinct behavior from other parameter shrinking methods (overparametrized).
- Explanation for abrupt transitions. Let:

$$f_i(\beta) = |y_i - x_i^{\mathsf{T}} \beta| + \delta \|\beta\|_2.$$

and assume $|y_i| = \|x_i\|_2 = 1$



Outline

Motivation

Adversarial training

Robustness in high-dimensions

Can a model perfectly fit the data and still be robust?

Analysing adversarial robustness

From:

$$\mathbb{E}\left[\left(\text{Adv-error}(\widehat{\beta})\right)^2\right] = \mathbb{E}\left[\left(|\text{Error}(\widehat{\beta})| + \delta\|\widehat{\beta}\|_*\right)^2\right]$$

It follows:

$$\mathbb{E}[\text{Error}(\widehat{\beta})^2] + \delta^2\|\widehat{\beta}\|_*^2 \leq \mathbb{E}[(\text{Adv. error})^2] \leq \left(\sqrt{\mathbb{E}[\text{Error}(\widehat{\beta})^2]} + \delta\|\widehat{\beta}\|_*\right)^2.$$

Application: plug-and-play from other analysis

Analysing minimum norm inteporlation:

$$(x_i, \epsilon_i) \sim P_x \times P_\epsilon, \qquad y_i = x_i^\mathsf{T} \beta + \epsilon_i,$$

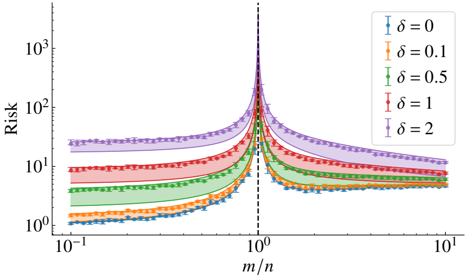


Figure: Adversarial risk *vs* number of features m.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in High-Dimensional Ridgeless Least Squares Interpolation," Annals of Statistics. 50(2): 949-986 (2022).

Is robustness at odds with accuracy?

Can a good model to be arbitrarely vulnerable to adversarial attacks as you add more features?

Proposition

If $\mathbb{E}[\text{Error}(\widehat{\beta})^2] < M$ as $\#\textit{features} \to \infty$:

$$\mathbb{E}[(\text{Adv. error}(\widehat{\beta}))^2] \to \infty$$

if and **only if**

$$\delta \|\widehat{\beta}\|_* \to \infty.$$

Example

Minimum ℓ_2-norm interpolator and Gaussian features:

$$\|\widehat{\beta}\|_1 = \mathcal{O}(1) \quad \|\widehat{\beta}\|_2 = \mathcal{O}(1/\sqrt{m})$$

Now, if we scale

$$\delta \propto \mathbb{E}\|x\|_2 = \mathcal{O}(\sqrt{m}).$$

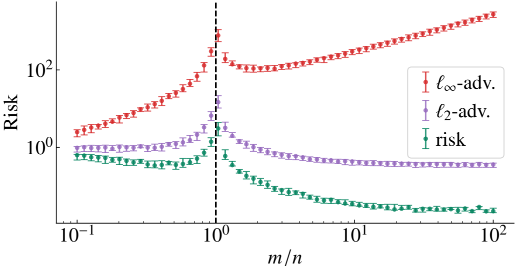


Figure: Adv. risk.

I. J. Goodfellow, J. Shlens, C. Szegedy , *"Explaining and Harnessing Adversarial Examples"*, ICLR 2015
D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Ma, *"Robustness May Be At Odds with Accuracy,"* ICLR, p. 23, 2019.

The effect of regularization

- Ridge, ℓ_2-adversarial training and min-ℓ_2-norm solution

$$\|\widehat{\beta}\|_1 = \mathcal{O}(1)$$

- Lasso, ℓ_∞-adversarial training and min ℓ_2-norm solution

$$\|\widehat{\beta}\|_1 = \mathcal{O}(1/\sqrt{m})$$

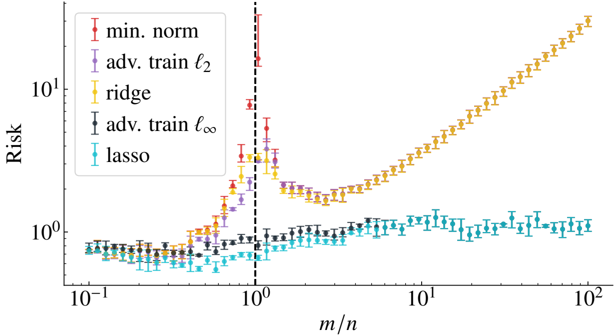


Figure: Adversarial ℓ_∞ risk and $\delta \propto \mathbb{E}\|x\|_2$.

Summary

- Dual formula for the adversarial error:

$$\left(\text{Adv-error}(\widehat{\beta})\right)^2 = \left(|\text{Error}(\widehat{\beta})| + \delta\|\widehat{\beta}\|_*\right)^2$$

- Consequences to adversarial training:
  - Convex formula / Similarities with parameter shrink methods
  - ℓ_∞-adversarial training $\Rightarrow$ sparse solutions
  - Can interpolate for disturbance bounded by $\delta > 0$.
- Consequences to adversarial robustness
  - Simplify analysis of adversarial robustness.
  - Sufficient and necessary conditions for good models to be vulnerable to adversary .

Thank you!

✉ antonio.horta.ribeiro@it.uu.se

🌐 antonior92.github.io