Adversarially-trained linear regression

Antônio Horta Ribeiro

Systems and control division
Uppsala, Nov 2022

$\mu-$seminar

Adversarial examples

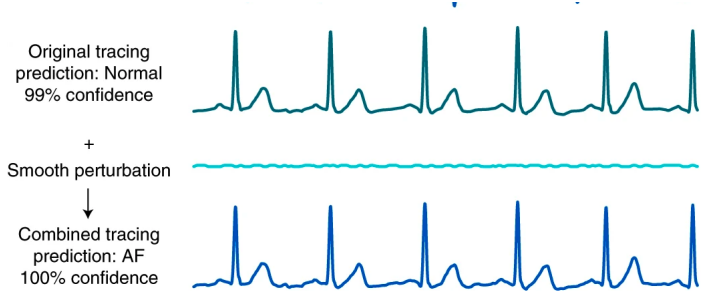


Figure: Effect of adversarial examples on ECG Classification.

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.*

"Is adversarial training fundamentally different than other regularization methods?"

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

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

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

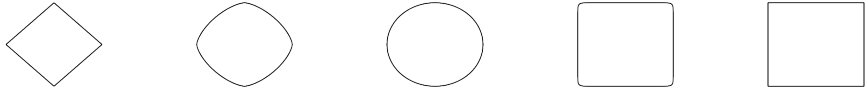
- $\text{Error}(\widehat{\beta}) = y - x^{\mathsf{T}}\widehat{\beta}$
- $\text{Adv-error}(\widehat{\beta}) = \max_{\|\Delta x\| \leq \delta} \left(y - (x + \Delta x)^{\mathsf{T}}\widehat{\beta} \right)$
- *Dual formula for the adversarial error*

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

- 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 $1/p + 1/q = 1$



ℓ_1 $\ell_{1.5}$ ℓ_2 ℓ_{20} ℓ_∞

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:

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

- Lasso:

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

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

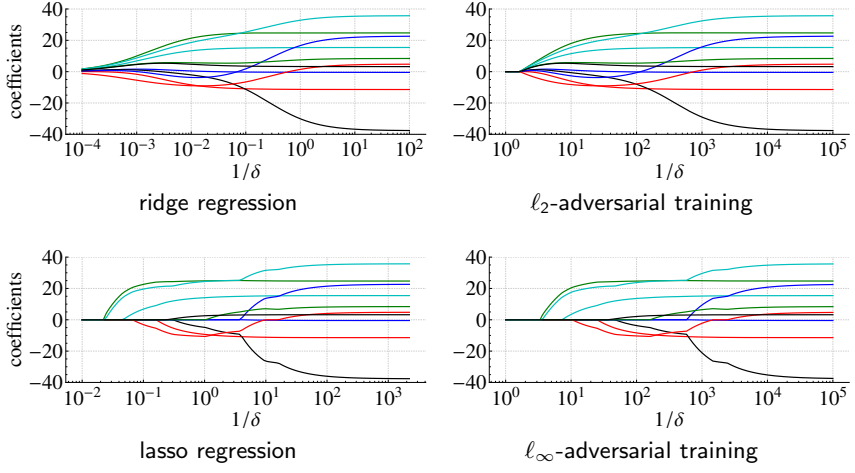


Figure: Regularization paths.

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:

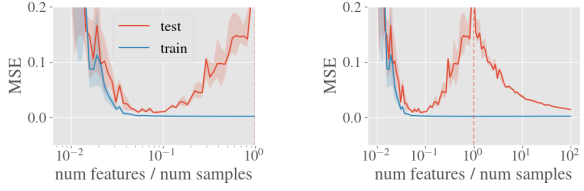


Figure: nonlinear ARX mean squared error (MSE).

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

If X has full row-rank, for $0 < \delta < \overline{\delta}$, adversarial training is minimized at

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

Corollary

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

Corollary

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

Overparametrized model

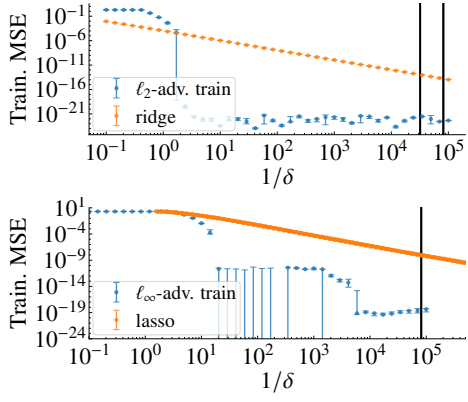


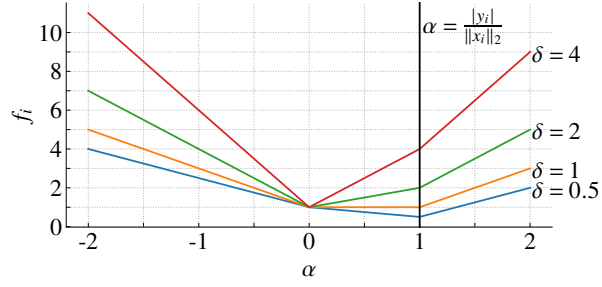
Figure: Training MSE *vs* regularization parameter.

Discussion

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

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

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



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 nonzero disturbance $\delta > 0$.

Thank you!

✉ antonio.horta.ribeiro@it.uu.se

🌐 antonior92.github.io