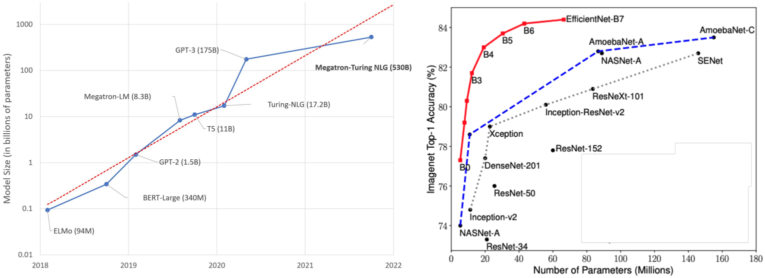
Overparameterized Linear Regression under Adversarial Attacks

Antônio H. Ribeiro
Uppsala University

Online Presentation @ University of British Columbia
June 3rd, 2022

Model size in neural networks



(a) Language models (b) Image classification

Figure: Models number of parameters

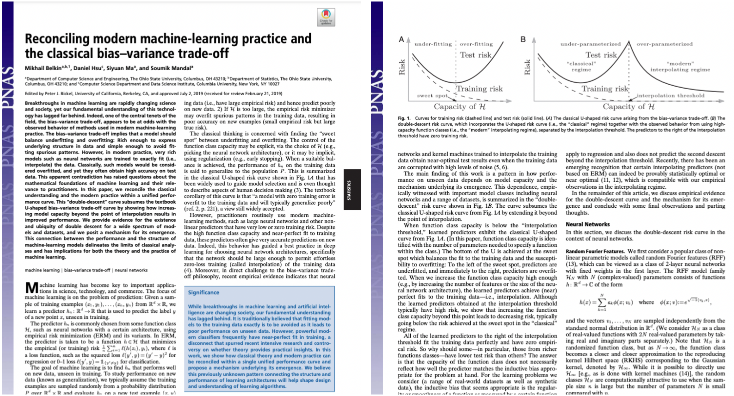
Sources: J. Simon (2021) "Large Language Models: A New Moore's Law?". Online (acessed: 2021-11-09). URL: huggingface.co/blog/large-language-models .
M. Tan and Q. V. Le (2019) "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," ICML

Overparametrized models

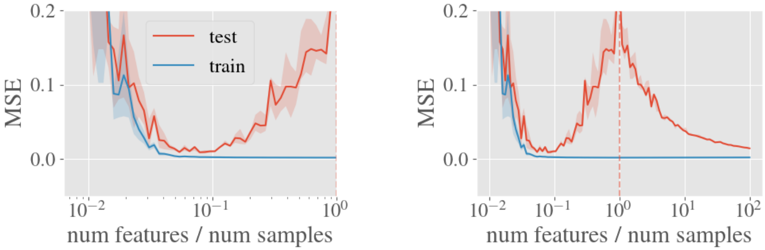
Seminars: overparameterized machine learning models (2021, Fall) — PhD level course - together with Dave Zachariah and Per Mattsson.

All material available in: https://github.com/uu-sml/seminars-overparam-ml

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



Double-descent



(a) U-Shape (b) Double-descent

Figure: Nonlinear ARX performance in Couple Eletric Drives benchmark.

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**

Double-descent in linear models

Estimated parameter: using train dataset (x_i, y_i), $i = 1, \cdots, n$:

- Underparametrized:

$$\hat{\beta} = \arg\min_{\beta} \sum_i (y_i - x_i{}^\mathsf{T}\beta)^2$$

- Overparametrized:

$$\begin{aligned}\hat{\beta} =& \arg\min_{\beta} \|\beta\|_2^2 \\ & \text{subject to } y_i = x_i{}^\mathsf{T}\beta \\ & \text{for every } i\end{aligned}$$

Random features: Belking et.al. (2019) generates the features through the nonlinear mapping: $\phi : u_i \mapsto x_i$ obtained from Random Fourier Features.

Overparametrized models can generalize effectively when train and test come from the **same** *distribution...*

are they robust?

Adversarial Attacks

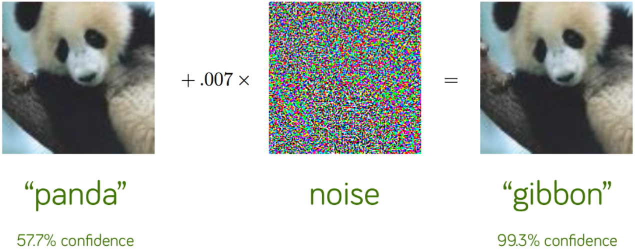


Figure: Illustration of adversarial attack.

Source: I. J. Goodfellow, J. Shlens, C. Szegedy , *"Explaining and Harnessing Adversarial Examples"*, ICLR 2015.

The role of high-dimensionality

- High-dimensionality as a source of vulnerability:
  - I. J. Goodfellow, J. Shlens, C. Szegedy , "*Explaining and Harnessing Adversarial Examples*", ICLR 2015
  - J. Gilmer et al., *"Adversarial Spheres,"* arXiv:1801.02774, Sep. 2018.
  - D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Ma, "Robustness May Be At Odds with Accuracy," ICLR, p. 23, 2019.
- High-dimensionality as a source of robustness:
  - S. Bubeck and M. Sellke, "A Universal Law of Robustness via Isoperimetry," Advances in Neural Information Processing Systems, 2021

Outline

Paper I **A. H. Ribeiro** and T. B. Schön, "Overparametrized Linear Regression under Adversarial Attacks," arXiv:2204.06274, April 2022.

Paper II **A. H. Ribeiro**, D. Zachariah, and T. B. Schön, "Surprises in adversarially-trained linear regression," arXiv:2205.12695, May 2022.

Linear regression under adversarial attacks

Given a data point not seen during training (x, y).

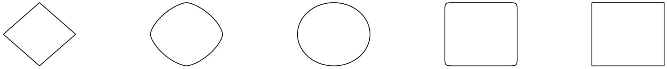
Standard risk:

$$R = E\left\{(y - x^{\mathsf{T}}\hat{\beta})^2\right\}$$

Adversarial risk:

$$R_p^{\text{adv}} = E\left\{\max_{\|\Delta x\|_p \leqslant \delta} (y - (x + \Delta x)^{\mathsf{T}}\hat{\beta})^2\right\}$$

$\Delta x \rightsquigarrow$ Adversarially generated disturbance



(a) ℓ_1 (b) $\ell_{1.5}$ (c) ℓ_2 (d) ℓ_{20} (e) ℓ_∞

Linear regression is a special case

The original formula

$$R_p^{\text{adv}} = E\left\{ \max_{\|\Delta x\|_p \leqslant \delta} (y - (x + \Delta x)^\mathsf{T} \hat{\beta})^2 \right\}$$

Can be reformulated. Let q, such that $\frac{1}{p} + \frac{1}{q} = 1$

$$R_p^{\text{adv}} = E\left(|y - x^\mathsf{T} \hat{\beta}| + \delta \|\hat{\beta}\|_q\right)^2.$$

Bounds on the adversarial risk

$$R + \delta^2 \|\widehat{\beta}\|_q^2 \leqslant R^{\text{adv}} \leqslant \left(\sqrt{R} + \delta \|\widehat{\beta}\|_q\right)^2$$

- $R^{\text{adv}} \rightsquigarrow$ Adversarial risk
- $R \rightsquigarrow$ Risk
- $\delta \rightsquigarrow$ Adv. disturbance magnitude

Note: *in the Gaussian case*

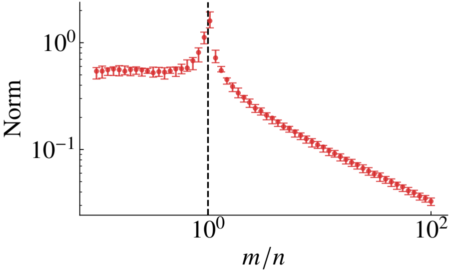
$$R^{\text{adv}}(\beta) = \left(1 - \sqrt{\frac{2}{\pi}}\right)(\text{Upper bound}) + \sqrt{\frac{2}{\pi}}(\text{Lower bound}).$$

Decay rate of the ℓ_2-norm

- Data model:

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

- ℓ_2-norm of the estimated parameter: decays with $\frac{1}{\sqrt{\#\text{ features}}}$

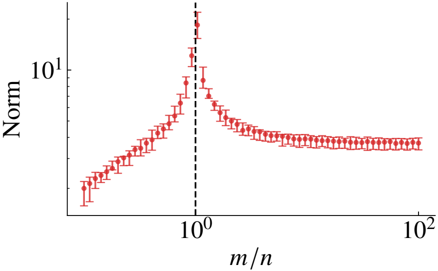


Decay rate of the ℓ_1-norm

- Relation between p-norm

$$\|\hat{\beta}\|_2 \leqslant \|\hat{\beta}\|_1 \leqslant \sqrt{m}\|\hat{\beta}\|_2.$$

- ℓ_1-norm of the estimated parameter: approaches a constant



- Hence:

$$\|\hat{\beta}\|_1 \to c\sqrt{m}\|\hat{\beta}\|_2.$$

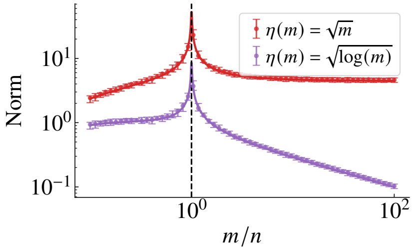
Scaling

- Model prediction: $\hat{\beta}^\mathsf{T} x$.
- Equivalent model prediction: $\tilde{\beta}^\mathsf{T} \tilde{x}$.

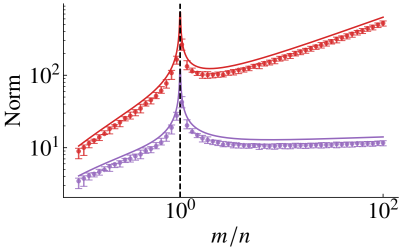
$$\tilde{x} = \frac{1}{\eta} x$$
$$\tilde{\beta} = \eta \hat{\beta}$$

- x be an isotropic $\rightarrow \mathbb{E}\left[\|x\|_2^2\right] = m$.
  $\rightarrow \eta(m) = \sqrt{m}$
- x is a sub-Gaussian $\rightarrow \mathbb{E}\left[\|x\|_\infty\right] = \Theta(\sqrt{\log(m)})$
  $\rightarrow \eta(m) = \sqrt{\log m}$

Norm



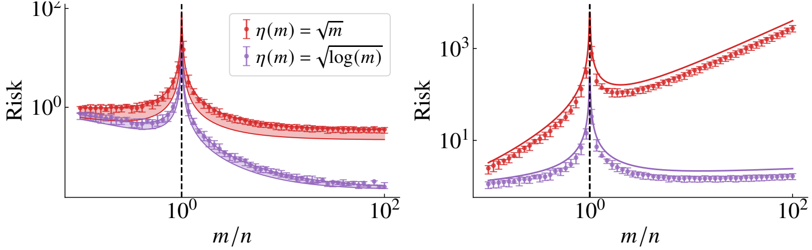
(a) $\|\hat{\beta}\|_2$



(b) $\|\hat{\beta}\|_1$

Adversarial Risk

$$R + \delta^2 \|\widehat{\beta}\|_q^2 \leqslant R^{\mathrm{adv}} \leqslant \left(\sqrt{R} + \delta \|\widehat{\beta}\|_q\right)^2$$



(a) ℓ_2 adv. risk (b) ℓ_∞ adv. risk

Discussion

- Different metrics in the input space → different assessments of the robustness.

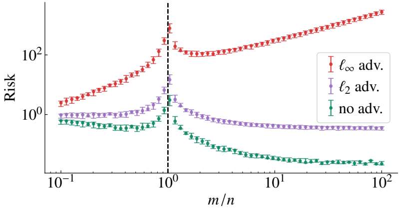


Figure: Adv. risk.

Discussion

- Can be seen as one aspect of the curse of dimensionality.
- Most pathological results for mismatched setup: $\mathbb{E}_x\left[\|x\|_2^2\right]$ const. attack while ℓ_∞ attack

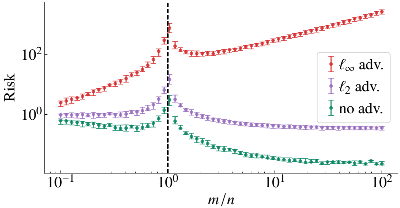


Figure: Adv. risk. **Fixed** $\eta(m) = \sqrt{n}$

- Brittleness to adversarial examples is reproducible in linear models is highly influential. The mismatch usually appears hidden in the examples.

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 and adversarial training

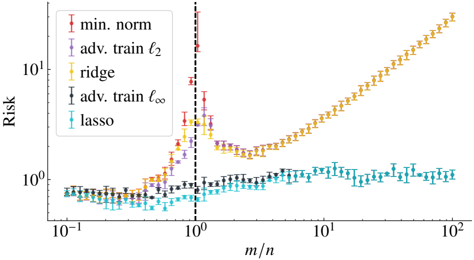


Figure: Adversarial ℓ_∞ risk.

Concentration of the norm

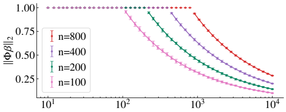
- The parameter estimated is

$$\begin{aligned}\hat{\beta} &= (X^{\mathsf{T}}X)^{\dagger}X^{\mathsf{T}}y,\\ &= (X^{\mathsf{T}}X)^{\dagger}X^{\mathsf{T}}(X\beta + \epsilon),\\ &= \underbrace{(X^{\mathsf{T}}X)^{\dagger}X^{\mathsf{T}}X}_{\Phi}\beta + (X^{\mathsf{T}}X)^{\dagger}X^{\mathsf{T}}\epsilon\end{aligned}$$

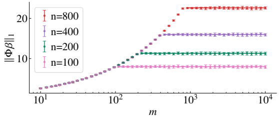
- $\Phi \in \mathbb{R}^{m\times m}$ is an orthogonal projector into a subspace of dimension n.
- If the entries of X are Gaussian, then Φ projects onto a random subspace uniformly sampled from Grassmannian $G(m, n)$.
- It is well know (Vershynin 2018, High-Dimensional Probability, Lemma 5.3.2) probability greater then $1 - 2\exp(-ct^2 n)$:

$$(1-t)\sqrt{\frac{n}{m}}\|\beta\|_2 \leqslant \|\Phi\beta\|_2 \leqslant (1+t)\sqrt{\frac{n}{m}}\|\beta\|_2 \qquad (1)$$

Concentration of the norm



(a) ℓ_2 norm of projection



(b) ℓ_1 norm of projection

Figure: **Random projection and norms**.

$\|\Phi\beta\|_1$ concentrate with high-probability around $c\sqrt{m}\|\beta\|_2$.

Outline

Paper I **A. H. Ribeiro** and T. B. Schön, "Overparametrized Linear Regression under Adversarial Attacks," arXiv:2204.06274, April 2022.

Paper II **A. H. Ribeiro**, D. Zachariah, and T. B. Schön, "Surprises in adversarially-trained linear regression," arXiv:2205.12695, May 2022.

Adversarial Training

- Empirical risk minimization (ERM). Minimizes:

$$\widehat{R}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i^\mathsf{T}\beta\right)^2,$$

- Adversarial training, minimizes *empirical adversarial risk*:

$$\widehat{R}_p^{\text{adv}}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\max_{\|\Delta x_i\|_p \leqslant \delta}(y_i - (x_i + \Delta x_i)^\mathsf{T}\beta)^2$$

Adversarial Training in linear regression

- The same simplification applies:

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

- The above expression is **convex**

Lasso and ℓ_∞-adversarial training

- ℓ_∞-adversarial training:

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

- Lasso:

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

Ridge regression and ℓ_2-adversarial training

- ℓ_2-adversarial training:

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

- Ridge:

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

Diabetes example

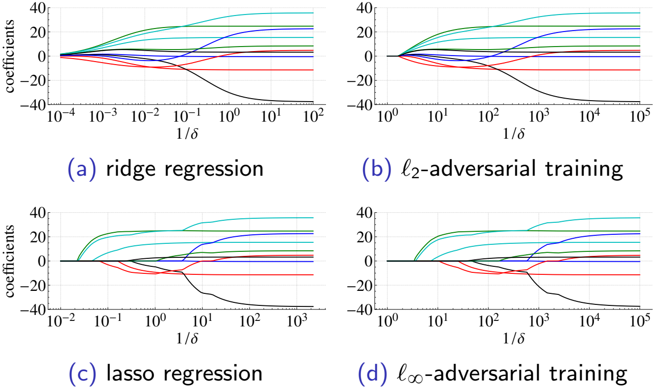
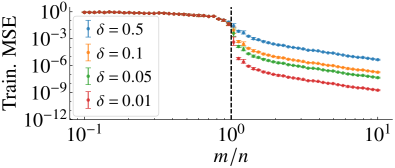
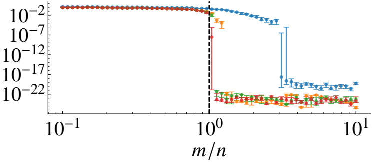


Figure: Regularization paths.

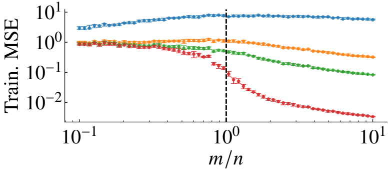
Diferences in the overparametrized region



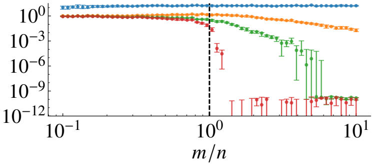
(a) ridge regression



(b) ℓ_2-adversarial training



(c) lasso regression

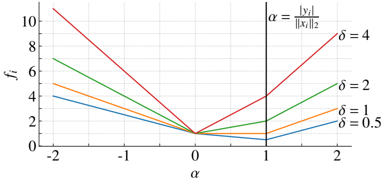


(d) ℓ_∞-adversarial training

Figure: Mean square error in training data.

Discussion

- Adversarial training can go through abrupt transitions in behavior.
- Looking at one point can be instructive:
  $f_i(\beta) = |y_i - x_i^\mathsf{T}\beta| + \delta\|\beta\|_2$



- Related work

H. Xu, C. Caramanis, and S. Mannor, "Robust regression and lasso," Advances in neural information processing systems, vol. 21, 2008

  - Robust regression → feature-wise perturbation
  - Adversarial training → sample-wise perturbation

Thank you!

Contact info:

antonio.horta.ribeiro@it.uu.se

@ahortaribeiro

antonior92.github.io

github.com/antonior92