Overparametrized Linear Regression under Adversarial Attacks

Antônio Horta Ribeiro

Uppsala University (Sweden)

SIERRA group meeting
INRIA Paris, March 2023

Outline

Motivation

Robustness in high-dimensions

Adversarial training

The state of Minas Gerais and telehealth

- Minas Gerais

The state of Minas Gerais and telehealth

- Minas Gerais
  - approximately the same area as France.

The state of Minas Gerais and telehealth

- Minas Gerais
  - approximately the same area as France.
  - 853 municipalities

The state of Minas Gerais and telehealth

- Minas Gerais
  - approximately the same area as France.
  - 853 municipalities
- Telehealth center of Minas Gerais

The state of Minas Gerais and telehealth

- Minas Gerais
  - approximately the same area as France.
  - 853 municipalities
- Telehealth center of Minas Gerais
- More than 4000 ECGs per day

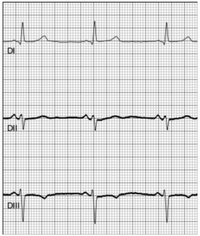
Electrocadiogram exam and machine learning

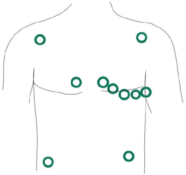
Goal: *Build data-driven ECG analysis tools.*

Electrocadiogram exam and machine learning

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

- Cardiovascular diseases: 32% of all deaths (GBD 2019).



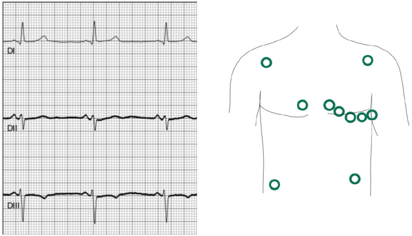


Left: ECG signal **Right:** Electrode placement.

Electrocadiogram exam and machine learning

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

- Cardiovascular diseases: 32% of all deaths (GBD 2019).
- The ECG is the major diagnostic tool.

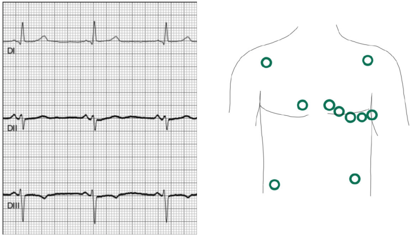


Left: ECG signal **Right:** Electrode placement.

Electrocadiogram exam and machine learning

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

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

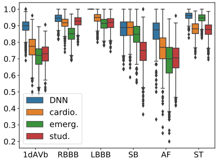


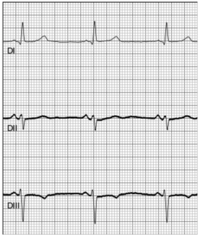
Left: ECG signal **Right:** Electrode placement.

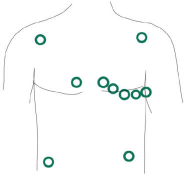
Electrocadiogram exam and machine learning

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

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





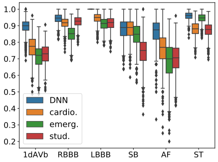


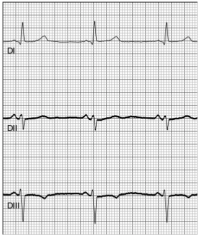
Left: ECG signal **Right:** Electrode placement.

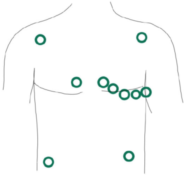
Electrocadiogram exam and machine learning

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

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





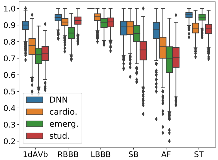


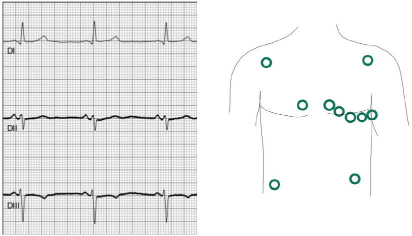
Left: ECG signal **Right:** Electrode placement.

Electrocadiogram exam and machine learning

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

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





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

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

Adversarial examples

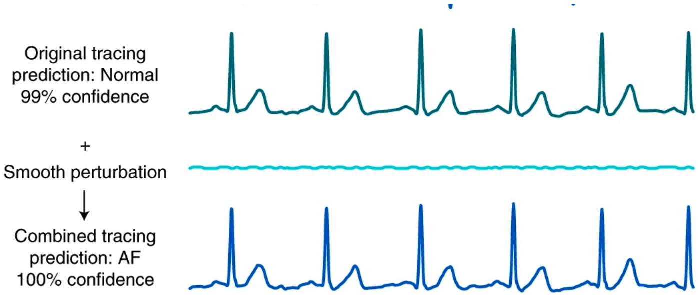


Figure: Effect of adversarial examples on ECG Classification.

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 examples

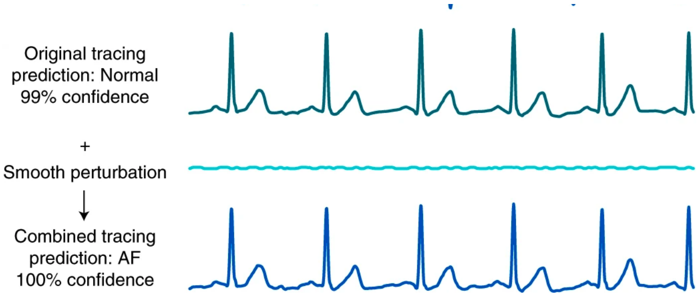


Figure: Effect of adversarial examples on ECG Classification.

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 robustness

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

Overparameterized Linear Regression under Adversarial Attacks (2023). **Antônio H. Ribeiro**, Thomas B. Schön. *IEEE Transactions on Signal Processing (preprint: arxiv.org/abs/2204.06274).*

Adversarial robustness

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

Overparameterized Linear Regression under Adversarial Attacks (2023). **Antônio H. Ribeiro**, Thomas B. Schön. *IEEE Transactions on Signal Processing (preprint: arxiv.org/abs/2204.06274).*

Adversarial training

"How is it connected to other regularization methods?"'

Surprises in adversarially-trained linear regression (2022). **Antônio H. Ribeiro**, Dave Zachariah, Francis Bach, Thomas B. Schön. *Work in progress.*

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:

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n) \Rightarrow \widehat{\beta}$$

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

$$\widehat{y} = \widehat{\beta}^{\mathsf{T}} x$$

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

$$\widehat{y} = \widehat{\beta}^{\mathsf{T}} x$$

- $\mathrm{Error}(\widehat{\beta}) = |y - x^{\mathsf{T}} \widehat{\beta}|$

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

$$\widehat{y} = \widehat{\beta}^{\mathsf{T}} x$$

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

Adversarial 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

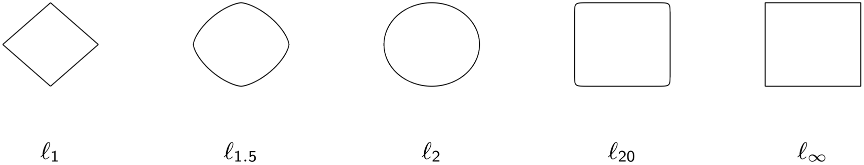
- ℓ_∞-adversarial attack: $\{\|\Delta x\|_\infty \leq \delta\} \Rightarrow$ dual norm: $\|\Delta x\|_1$

ℓ_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 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 $1/p + 1/q = 1$



Outline

What is the role of high-dimensionality in model robustness?

- High-dimensionality as a source of vulnerability:

  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, 2019.
  J. Gilmer et al., *"Adversarial Spheres,"* arXiv:1801.02774, Sep. 2018.

What is the role of high-dimensionality in model robustness?

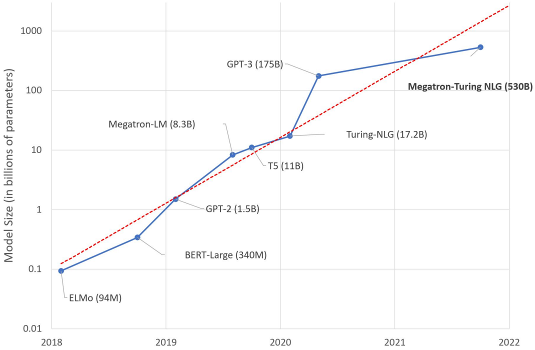
- High-dimensionality as a source of vulnerability:

  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, 2019.
  J. Gilmer et al., *"Adversarial Spheres,"* arXiv:1801.02774, Sep. 2018.

- 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

How neural networks can perform well?



Language models

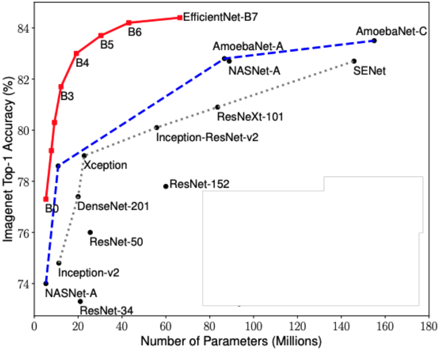


Image classification

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

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.

Overparametrized models

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)

Overparametrized models

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:

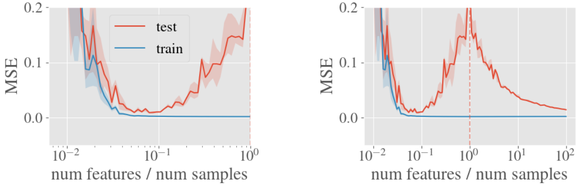


Figure: nonlinear ARX mean squared error (MSE).

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**

Setup: minimum-norm interpolator

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

- Gradient descent in linear regression converges to $\widehat{\beta}^{\text{min}-\ell_2}$.

Setup: minimum-norm interpolator

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

- Gradient descent in linear regression converges to $\widehat{\beta}^{\text{min}-\ell_2}$.
- Used to study *benign overfitting* in:

  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.

Setup: minimum-norm interpolator

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

- Gradient descent in linear regression converges to $\widehat{\beta}^{\mathsf{min}-\ell_2}$.
- Used to study *benign overfitting* in:

  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.
- Use to study *double descent* in:

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

Analysing adversarial robustness

From:

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

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 that:

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

Double-descent in the adversarial loss

$$\mathbb{E}[(\ell_2\text{-adv. error}(\widehat{\beta}))^2] \propto \mathbb{E}[\text{Error}(\widehat{\beta})^2] + \delta^2 \|\widehat{\beta}\|_2^2.$$

Double-descent in the adversarial loss

$$\mathbb{E}[(\ell_2\text{-adv. error}(\widehat{\beta}))^2] \propto \mathbb{E}[\text{Error}(\widehat{\beta})^2] + \delta^2\|\widehat{\beta}\|_2^2.$$

$\|\widehat{\beta}\|_2$ also present a double descent behavior: *As we increase the problem dimension, it becomes possible to find solutions with smaller norm.*

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

Asymptotic results

Analysing minimum-norm inteporlation:

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

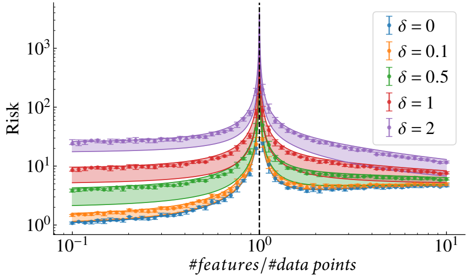


Figure: Adversarial risk *vs* number of features m.

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

High-dimensionality as a source of brittleness

- Model:

$$y \sim \mathcal{N}(0, 1)$$
$$x_i \sim \mathcal{N}(y, 1)$$

High-dimensionality as a source of brittleness

- Model:

$$y \sim \mathcal{N}(0,1)$$
$$x_i \sim \mathcal{N}(y,1)$$

- Optimal predictor: $\widehat{\beta} = \left[\frac{1}{\# features}, \cdots, \frac{1}{\# features}\right]$

High-dimensionality as a source of brittleness

- Model:

$$y \sim \mathcal{N}(0,1)$$
$$x_i \sim \mathcal{N}\left(y, 1\right)$$

- Optimal predictor: $\widehat{\beta} = \left[\frac{1}{\#features}, \cdots, \frac{1}{\#features}\right]$
- $\delta = \mathbb{E}\|x\|_2 \propto \sqrt{\#features}$

High-dimensionality as a source of brittleness

- Model:

$$y \sim \mathcal{N}(0,1)$$
$$x_i \sim \mathcal{N}\left(y,1\right)$$

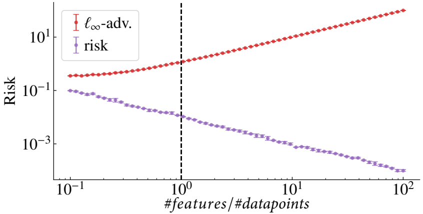
- Optimal predictor: $\widehat{\beta} = \left[\frac{1}{\#features}, \cdots, \frac{1}{\#features}\right]$
- $\delta = \mathbb{E}\|x\|_2 \propto \sqrt{\#features}$

High-dimensionality as a source of brittleness

- Model:

$$y \sim \mathcal{N}(0,1)$$
$$x_i \sim \mathcal{N}\left(y,1\right)$$

- Optimal predictor: $\widehat{\beta} = \left[\frac{1}{\#features}, \cdots, \frac{1}{\#features}\right]$
- $\delta = \mathbb{E}\|x\|_2 \propto \sqrt{\#features}$



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.

Is robustness at odds with accuracy?

When good model become more vulnerable to adversarial attacks as we add features?

Proposition

If $\mathbb{E}[\text{Error}(\widehat{\beta})^2] < \epsilon$:

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

if and **only if**

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

Example: optimal predictor vulnerable to adversarial attacks

- Optimal predictor: $\widehat{\beta} = \left[\frac{1}{\#features}, \cdots, \frac{1}{\#features}\right]$
- $\Delta x = \mathbb{E}\|x\|_2 \propto \sqrt{\#features}$
- For our example,
$$\delta\|\widehat{\beta}\|_1 = \sqrt{\#features}$$
hence
$$\mathbb{E}[(\ell_\infty\text{-adv. error}(\widehat{\beta}))^2] = \mathcal{O}(\#features)$$

Example: optimal predictor vulnerable to adversarial attacks

- Optimal predictor: $\widehat{\beta} = \left[\frac{1}{\#features}, \cdots, \frac{1}{\#features}\right]$
- $\Delta x = \mathbb{E}\|x\|_2 \propto \sqrt{\#features}$
- For our example,

$$\delta\|\widehat{\beta}\|_1 = \sqrt{\#features}$$

hence

$$\mathbb{E}[(\ell_\infty\text{-adv. error}(\widehat{\beta}))^2] = \mathcal{O}(\#features)$$

Mismatched example

- ℓ_∞-adv. attack $\|\Delta x\|_\infty \leq \delta$.
- $\Delta x \propto \mathbb{E}\|x\|_2$.

:

Example: minimum ℓ_2-norm interpolator

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

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

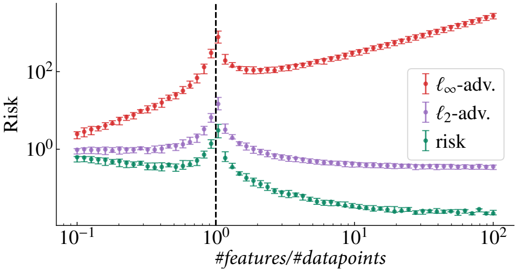


Figure: Adv. risk.

The effect of adversarial training and regularization

Empirical risk minimization:

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

Adversarial training:

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

The effect of regularization

- Ridge and ℓ_2-adversarial training

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

- Lasso, ℓ_∞-adversarial training

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

The effect of regularization

- Ridge and ℓ_2-adversarial training

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

- Lasso, ℓ_∞-adversarial training

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

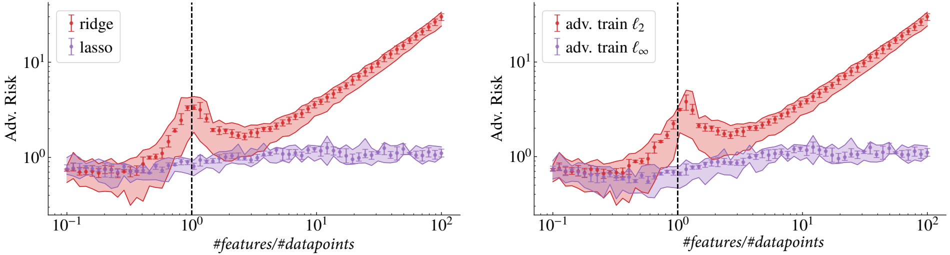


Figure: Adversarial ℓ_∞ risk and $\delta \propto \mathbb{E}\|x\|_2$.

Outline

Motivation

Robustness in high-dimensions

Adversarial training

Adversarial training in linear models

- Adversarial training,

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

Adversarial training in linear models

- Adversarial training,

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

can be reformulated as

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

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

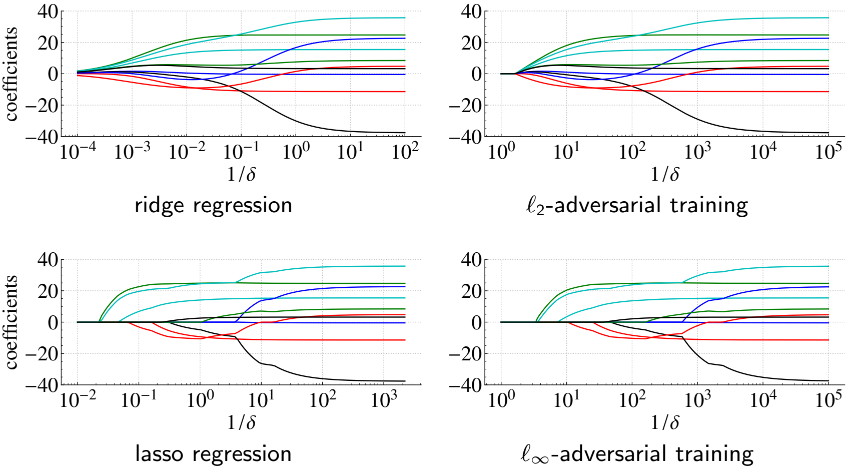


Figure: Regularization paths.

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 $\widehat{\beta}^{\text{min}-\ell_2}$.
- Used to study *benign overfitting* in:

  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.
- Use to study *double descent* in:

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

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 $\widehat{\beta}^{\mathsf{min}-\ell_2}$.
- Used to study *benign overfitting* in:

  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.
- Use to study *double descent* in:

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

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.
- Can also be used to study *benign overfitting* and *double descent*, i.e.,

  F. Koehler, L. Zhou, D. J. Sutherland, and N. Srebro, "Uniform Convergence of Interpolators: Gaussian Width, Norm Bounds and Benign Overfitting," presented at the Advances in Neural Information Processing Systems, 2021

Minimum-norm interpolator and adversarial training

Theorem

Adversarial training is minimized at the minimum norm interpolator

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

iff $0 < \delta < \bar{\delta}$.

Surprises in adversarially-trained linear regression (2022). **Antônio H. Ribeiro**, Dave Zachariah, Francis Bach, Thomas B. Schön. *Work in progress.*

ℓ_2-adversarial training *vs* ridge regression

Relation to min-norm solution

- **(corollary)** $\widehat{\beta}^{\text{min-}\ell_2}$ is the solution to ℓ_2-adversarial training iff $0 < \delta < \bar{\delta}$.
- Ridge $\widehat{\beta}^{\text{ridge}}(\delta) \to \widehat{\beta}^{\text{min}-\ell_2}$ as $\delta \to 0^+$.

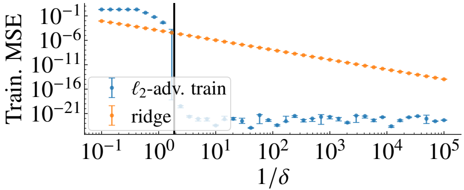


Figure: Training MSE *vs* regularization parameter.

ℓ_∞-adversarial training *vs* Lasso

Relation to min-norm solution

- **(corollary)** $\widehat{\beta}^{\text{min-}\ell_1}$ is the solution to ℓ_∞-adversarial training iff $0 < \delta < \bar{\delta}$.
- Lasso $\widehat{\beta}^{\text{lasso}}(\delta) \to \widehat{\beta}^{\text{min}-\ell_1}$ as $\delta \to 0^+$ (LARS algorithm)..

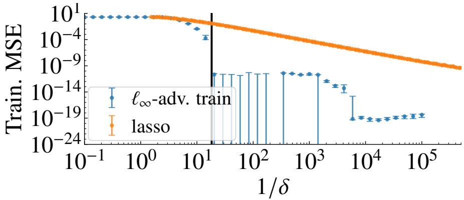


Figure: Training MSE *vs* regularization parameter.

Discussion

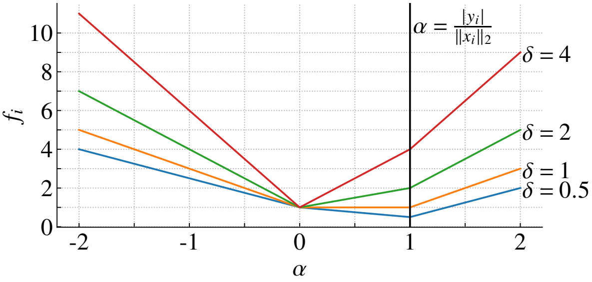
- Distinct behavior from other parameter shrinking methods (overparametrized).

Discussion

- 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 $|y_i| = \|x_i\|_2 = 1$



New interpretation for minimum-norm interpolator

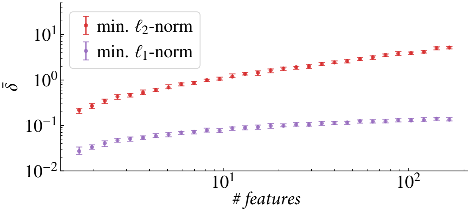


Figure: Threshold $\bar{\delta}$ *vs* number of features m.

New interpretation for minimum-norm interpolator

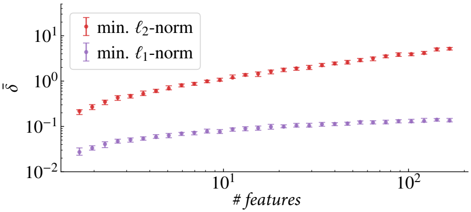


Figure: Threshold $\bar{\delta}$ *vs* number of features m.

- Setup: $y_i = x_i^\top \beta + \epsilon_i$, $x_i \sim \mathcal{N}(0, r^2 I_m)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- Increasing the number of features for minimum-norm interpolators increases the maximum disturbance $\bar{\delta}$ in the corresponding adversarial training problems.

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 robustness
  - Simplify analysis of adversarial robustness:

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

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

Thank you!

INRIA de Paris - Room C407 (from now to mid-June)
antonio.horta.ribeiro@it.uu.se
antonior92.github.io