Ataques adversariais em modelos lineares

Antônio Horta Ribeiro
Uppsala University
Sweden

Laboratório Nacional de Computação Científica

Brasil, 30 de Outubro 2023 (Online)

Supervised learning

- Train dataset:
$$(x_i, y_i),\ \ i = 1, \cdots, \#train.$$

Supervised learning

- Train dataset:

$$(x_i, y_i),\ \ i = 1, \cdots, \#train.$$

- Model:

$$f_\beta : x \mapsto \widehat{y}$$

Supervised learning

- Train dataset:

$$(x_i, y_i),\ \ i = 1, \cdots, \#train.$$

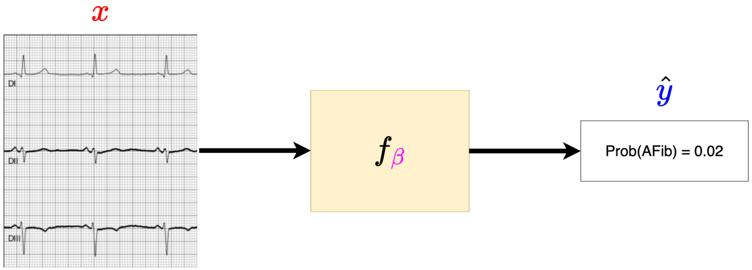
- Model:

$$f_\beta : x \mapsto \widehat{y}$$

- Parameter estimation method:

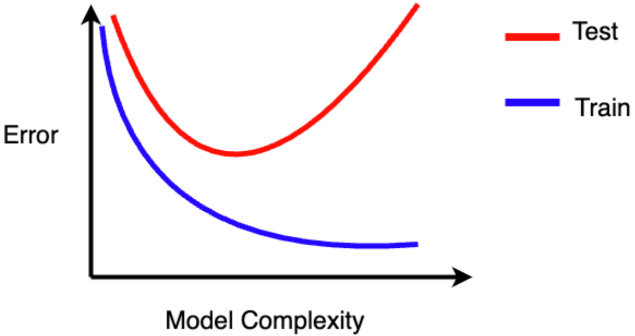
$$\min_\beta \sum_{i=1}^{\#train} \ell(y_i, f_\beta(x_i))$$

Example: automatic diagnosis of the ECG

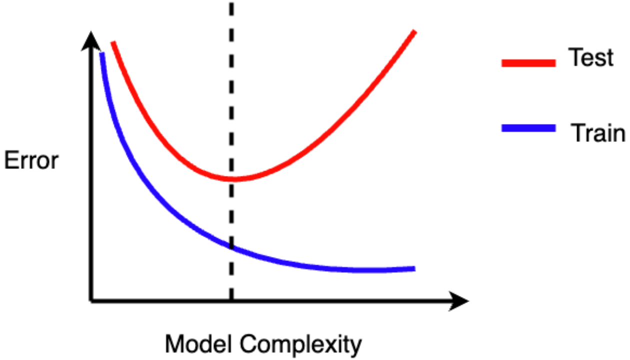


Automatic diagnosis of the 12-lead ECG using a deep neural network
A. H. Ribeiro , M.H. Ribeiro, Paixão, G.M.M. Paixão et al
Nature Communications (2020)

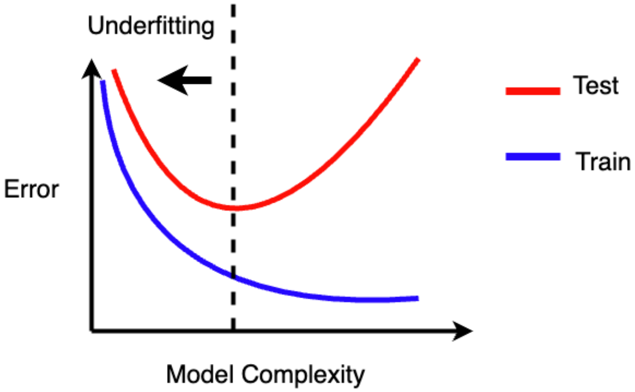
Generalization to new test points



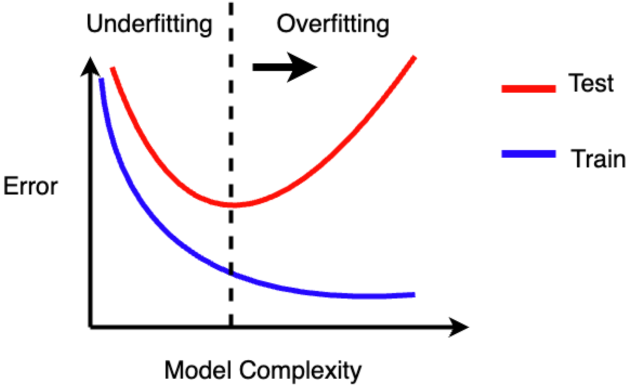
Generalization to new test points



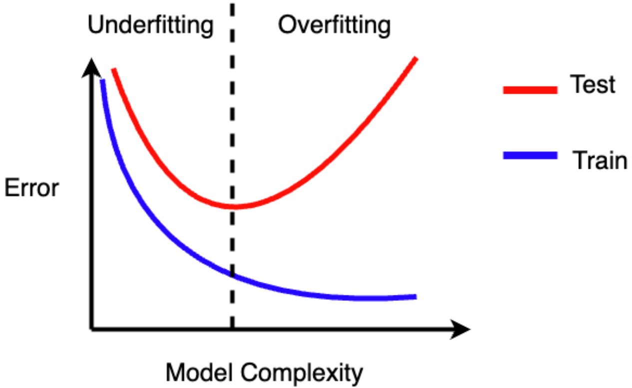
Generalization to new test points



Generalization to new test points



Generalization to new test points



Regularization

- "*Mechanism to explicitly or implicitly* **prioritize lower complexity** *when choosing a predictive model*"

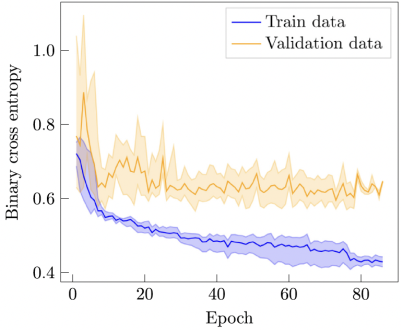
Regularization

- *"Mechanism to explicitly or implicitly* **prioritize lower complexity** *when choosing a predictive model"*
- Example: **Parameter shrinking**

$$\min_{\beta} \underbrace{\sum_{i=1}^{\#train} \ell(y_i, f_\beta(x_i))}_{\text{error in training}} + \underbrace{\|\beta\|^2}_{\text{complexity penalty term}}$$

Robustness and external validation

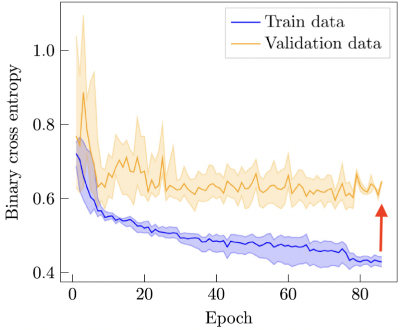
- **Generalization** gap



Screening for Chagas disease from the electrocardiogram using a deep neural network
Carl Jidling, Daniel Gedon, Thomas B. Schön, Claudia Di Lorenzo Oliveira, Clareci Silva Cardoso, Ariela Mota Ferreira, Luana Giatti, Sandhi Maria Barreto, Ester C. Sabino, Antônio L. P. Ribeiro, **Antônio H. Ribeiro**
Plos Neglected Tropical Diseases (2023)

Robustness and external validation

- **Generalization** gap



Screening for Chagas disease from the electrocardiogram using a deep neural network
Carl Jidling, Daniel Gedon, Thomas B. Schön, Claudia Di Lorenzo Oliveira, Clareci Silva Cardoso, Ariela Mota Ferreira, Luana Giatti, Sandhi Maria Barreto, Ester C. Sabino, Antônio L. P. Ribeiro, **Antônio H. Ribeiro**
Plos Neglected Tropical Diseases (2023)

Robustness and external validation

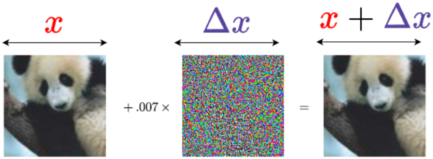
- **Generalization** gap
- **Robustness** gap

Split	Cohort	ROC-AUC
Test	CODE + SaMi-Trop	0.80
External validation 1	REDS-II	0.68
External validation 2	ELSA-Brasil	0.59

Screening for Chagas disease from the electrocardiogram using a deep neural network
Carl Jidling, Daniel Gedon, Thomas B. Schön, Claudia Di Lorenzo Oliveira, Clareci Silva Cardoso, Ariela Mota Ferreira, Luana Giatti, Sandhi Maria Barreto, Ester C. Sabino, Antônio L. P. Ribeiro, **Antônio H. Ribeiro**
Plos Neglected Tropical Diseases (2023)

Adversarial attacks

- $x \rightarrow \widehat{y}$:
  Panda (Probability $= 0.57$)

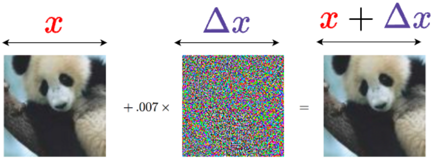


Explaining and Harnessing Adversarial Examples
I. J. Goodfellow, J. Shlens, C. Szegedy
ICLR (2015)

x

Adversarial attacks

- $x \to \widehat{y}$:
  Panda (Probability $= 0.57$)
- $\|\Delta x\|_\infty < 0.007$

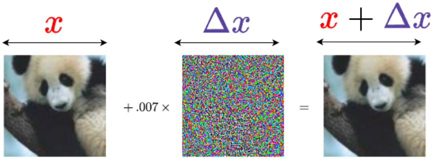


Explaining and Harnessing Adversarial Examples
I. J. Goodfellow, J. Shlens, C. Szegedy
ICLR (2015)

x

Adversarial attacks

- $x \to \widehat{y}$:
  Panda (Probability $= 0.57$)
- $\|\Delta x\|_\infty < 0.007$
- $x + \Delta x \to \widetilde{y}$:
  Gibbon (Probability $= 0.99$)

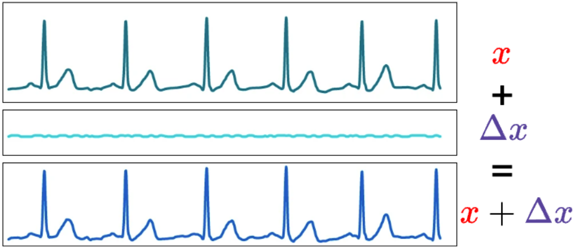


Explaining and Harnessing Adversarial Examples
I. J. Goodfellow, J. Shlens, C. Szegedy
ICLR (2015)

x

Adversarial attacks

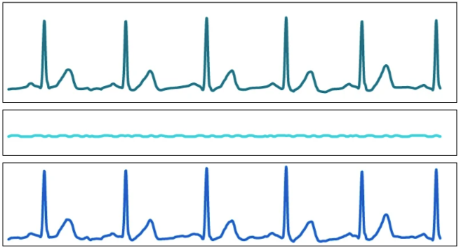
- $x \to \widehat{y}$:
  Normal (Probability $= 0.99$)



Deep learning models for electrocardiograms are susceptible to adversarial attacks
Han, X., Hu, Y., Foschini, L. et al.
Nature Medicine (2020)

Adversarial attacks

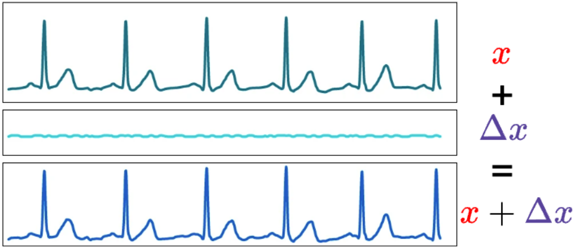
- $x \rightarrow \widehat{y}$:
  Normal (Probability $= 0.99$)
- $\|\Delta x\| < \delta$



Deep learning models for electrocardiograms are susceptible to adversarial attacks
Han, X., Hu, Y., Foschini, L. et al.
Nature Medicine (2020)

Adversarial attacks

- $x \to \hat{y}$:
  Normal (Probability $= 0.99$)
- $\|\Delta x\| < \delta$
- $x + \Delta x \to \tilde{y}$:
  AFib (Probability $= 1.00$)



Deep learning models for electrocardiograms are susceptible to adversarial attacks
Han, X., Hu, Y., Foschini, L. et al.
Nature Medicine (2020)

Adversarial attacks

Disturbance chosen to maximize the error

$$\max_{\|\Delta x'\| \leq \delta} \ell(y_i, f_\beta(x_i + \Delta x'))$$

Adversarial attacks

Disturbance chosen to maximize the error

$$\max_{\|\Delta x'\| \leq \delta} \ell(y_i, f_\beta(x_i + \Delta x'))$$

We will analyze a **simplified** case:

- **Linear** model: $f_\beta(x) = \beta^\top x$

Regularization properties of adversarially-trained linear regression
Antônio H. Ribeiro, Dave Zachariah, Francis Bach, Thomas B. Schön.
NeurIPS (Spotlight) (2023)

Adversarial attacks

Disturbance chosen to maximize the error

$$\max_{\|\Delta x'\| \leq \delta} \ell(y_i, f_\beta(x_i + \Delta x'))$$

We will analyze a **simplified** case:

- **Linear** model: $f_\beta(x) = \beta^\top x$
- **Squared-error** loss: $\ell(y, \beta^\top x) = (y - \beta^\top x)^2$

Regularization properties of adversarially-trained linear regression
Antônio H. Ribeiro, Dave Zachariah, Francis Bach, Thomas B. Schön.
NeurIPS (Spotlight) (2023)

Adversarial attacks

Disturbance chosen to maximize the error

$$\max_{\|\Delta x'\| \leq \delta} \ell(y_i, f_\beta(x_i + \Delta x'))$$

We will analyze a **simplified** case:

- **Linear** model: $f_\beta(x) = \beta^\top x$
- **Squared-error** loss: $\ell(y, \beta^\top x) = (y - \beta^\top x)^2$
- Resulting problem:

$$\max_{\|\Delta x\| \leq \delta} (y_i - \beta^\top (x_i + \Delta x))^2$$

Regularization properties of adversarially-trained linear regression
Antônio H. Ribeiro, Dave Zachariah, Francis Bach, Thomas B. Schön.
NeurIPS (Spotlight) (2023)

Why linear models?

- **Simplest model class** where adversarial vulnerability has been observed.

  Explaining and Harnessing Adversarial Examples
  I. J. Goodfellow, J. Shlens, C. Szegedy
  ICLR (2015)

Why linear models?

- **Simplest model class** where adversarial vulnerability has been observed.

  Explaining and Harnessing Adversarial Examples
  I. J. Goodfellow, J. Shlens, C. Szegedy
  ICLR (2015)

- Ameanable to **mathematical analysis**

Why linear models?

- **Simplest model class** where adversarial vulnerability has been observed.

  Explaining and Harnessing Adversarial Examples
  I. J. Goodfellow, J. Shlens, C. Szegedy
  ICLR (2015)

- Ameanable to **mathematical analysis**
- Using **infinite dimensional** spaces we can analyze nonlinear extensions.

Adversarial training

- **Linear regression:**

$$\min_{\beta} \sum_{i=1}^{\#train} (y_i - \beta^\top x_i)^2$$

Adversarial training

- **Linear regression:**

$$\min_{\beta} \sum_{i=1}^{\#train} (y_i - \beta^\top x_i)^2$$

- **Adversarial training** in linear regression:

$$\min_{\beta} \sum_{i=1}^{\#train} \max_{\|\Delta x_i\| \leq \delta} (y_i - \beta^\top (x_i + \Delta x_i))^2$$

Adversarially-trained linear regression

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

- **Convex** *optimization problem;*

Overparameterized Linear Regression under Adversarial Attack.
Antônio H. Ribeiro, Thomas B. Schön.
IEEE Transactions on Signal Processing (2023)

Adversarially-trained linear regression

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

- **Convex** *optimization problem;*
- *It can be* **rewritten** *as:*

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

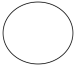
where $\|\cdot\|_*$ *is the* **dual norm**.

Overparameterized Linear Regression under Adversarial Attack.
Antônio H. Ribeiro, Thomas B. Schön.
IEEE Transactions on Signal Processing (2023)

Pairs of dual norms

(a) ℓ_2-adversarial attacks: $(\|\cdot\|_2 \leftrightarrow \|\cdot\|_2)$

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



$\|\Delta x\|_2 \leq \delta$

(b) ℓ_∞-adversarial attacks: $(\|\cdot\|_\infty \leftrightarrow \|\cdot\|_1)$

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

$\|\Delta x\|_\infty \leq \delta$

Driving question

How does adversarial training **compare with** *other regularization methods?*

Regularization methods:

1. Parameter shrinking methods.
  - Lasso.
  - Ridge regression.

2. $\sqrt{\text{Lasso}}$.

3. Minimum-norm interpolators.

Driving question

How does adversarial training **compare with** *other regularization methods?*

Regularization methods:

1. Parameter shrinking methods.

- Lasso.
- Ridge regression.

2. $\sqrt{\text{Lasso}}$.

3. Minimum-norm interpolators.

#1. Equivalence with Lasso

- **Lasso:**

$$\sum_{i=1}^{\#train} \left(|y_i - x_i^\mathsf{T}\beta|\right)^2 + \lambda\|\beta\|_1.$$

- **ℓ_∞-adversarial attacks:**

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

#1. Equivalence with Lasso

- **Lasso:**

$$\sum_{i=1}^{\#train} \left(|y_i - x_i^T \beta| \right)^2 + \lambda \|\beta\|_1 .$$

- **ℓ_∞-adversarial attacks:**

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

#1. Equivalence with Lasso

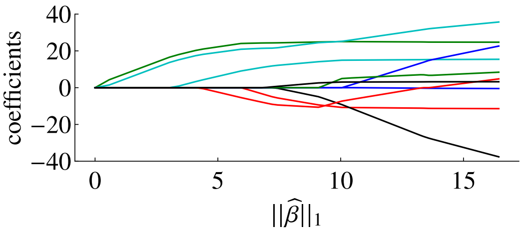
Result

if $\mathcal{E}[x] = 0$ and $x \sim -x$ there is a **map** $\lambda \leftrightarrow \delta$ for which the results are asymptotically **equivalent**.

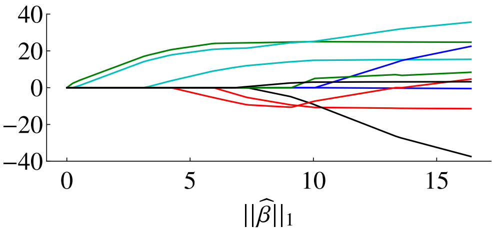
#1. Equivalence with Lasso

Result

if $\mathcal{E}[x] = 0$ and $x \sim -x$ there is a **map** $\lambda \leftrightarrow \delta$ for which the results are asymptotically **equivalent**.



Lasso.



ℓ_∞-adv. training.

"Is there an **advantage** *in using adversarial training?"*

Driving question

How does adversarial training **compare with** *other regularization methods?*

Regularization methods:

\# 1. Parameter shrinking methods.

- Lasso.
- Ridge regression.

\# 2. $\sqrt{\text{Lasso}}$.

\# 3. Minimum-norm interpolators.

#2: Similarities with $\sqrt{\text{Lasso}}$

$\sqrt{\text{Lasso}}$ minimizes:

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

A. Belloni, V. Chernozhukov, and L. Wang, *"Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming,"* Biometrika 2011.

#2: Similarities with $\sqrt{\text{Lasso}}$

$\sqrt{\text{Lasso}}$ minimizes:

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

A. Belloni, V. Chernozhukov, and L. Wang, *"Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming,"* Biometrika 2011.

- It allows the regularization parameter λ to be set **without the knowledge of the noise variance** [3].

#2: Similarities with $\sqrt{\text{Lasso}}$

$\sqrt{\text{Lasso}}$ minimizes:

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

A. Belloni, V. Chernozhukov, and L. Wang, *"Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming,"* Biometrika 2011.

- It allows the regularization parameter λ to be set **without the knowledge of the noise variance** [3].
- ℓ_∞-adversarial attacks have the **same property**.

#2: Similarities with $\sqrt{\text{Lasso}}$

Data model

We assume the data generated as:

$$y = x^\top \beta^* + \varepsilon$$

where the **noise** has variance $\mathrm{Var}(\varepsilon) = \sigma^2$.

#2: Similarities with $\sqrt{\text{Lasso}}$

Data model

We assume the data generated as:

$$y = x^\top \beta^* + \varepsilon$$

where the **noise** has variance $\mathrm{Var}(\varepsilon) = \sigma^2$.

To obtain near-oracle performance.

- *Lasso:*

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

- $\sqrt{\textit{Lasso}}$:

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

#2: Similarities with $\sqrt{\text{Lasso}}$

Data model

We assume the data generated as:

$$y = x^\top \beta^* + \varepsilon$$

where the **noise** has variance $\mathrm{Var}(\varepsilon) = \sigma^2$.

To obtain near-oracle performance.

- *Lasso:*

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

- $\sqrt{Lasso}$:

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

#2: Similarities with $\sqrt{\text{Lasso}}$

Data model

We assume the data generated as:

$$y = x^\top \beta^* + \varepsilon$$

where the **noise** has variance $\mathrm{Var}(\varepsilon) = \sigma^2$.

To obtain near-oracle performance.

- *Lasso:*

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

- $\sqrt{\textit{Lasso}}$:

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

- **ℓ_∞-adversarial attack:**

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

Driving question

How does adversarial training **compare with** *other regularization methods?*

Regularization methods:

1. Parameter shrinking methods.
- Lasso.
- Ridge regression.

2. $\sqrt{\text{Lasso}}$.

3. Minimum-norm interpolators.

#3: Equivalence min ℓ_1-norm interp.

Minimum ℓ_1-norm interpolator

Let $\#\text{train} < \#\text{params}$, the minimum ℓ_1-norm interpolator is

$$\min_{\beta} \|\beta\|_1 \quad \text{subject to} \quad x_i\beta = y_i, \forall i.$$

Result:

If $0 < \delta \leq \bar{\delta}$, the minimum ℓ_1-norm interpolator is **a solution of ℓ_∞-adv. training**.

#3: Equivalence min ℓ_2-norm interp.

Minimum ℓ_2-norm interpolator

Let $\#\text{train} < \#\text{params}$, the minimum ℓ_2-norm interpolator is

$$\min_{\beta} \|\beta\|_2 \quad \text{subject to} \quad x_i\beta = y_i, \forall i.$$

Result:

If $0 < \delta \leq \delta$, the minimum ℓ_2-norm interpolator is **a solution of ℓ_2-adv. training**.

#3: Equivalence min-norm interp.

Parameter shrinking: transition **only in the limit**

- $\beta^{\text{lasso}}(\lambda) \to \beta^{\text{min}-\ell_1}$ as $\lambda \to 0^+$ (for LARS algorithm).

#3: Equivalence min-norm interp.

Parameter shrinking: transition **only in the limit**

- $\beta^{\text{lasso}}(\lambda) \to \beta^{\text{min}-\ell_1}$ as $\lambda \to 0^+$ (for LARS algorithm).
- $\beta^{\text{ridge}}(\lambda) \to \beta^{\text{min}-\ell_2}$ as $\lambda \to 0^+$

#3: Equivalence min-norm interp.

Parameter shrinking: transition **only in the limit**

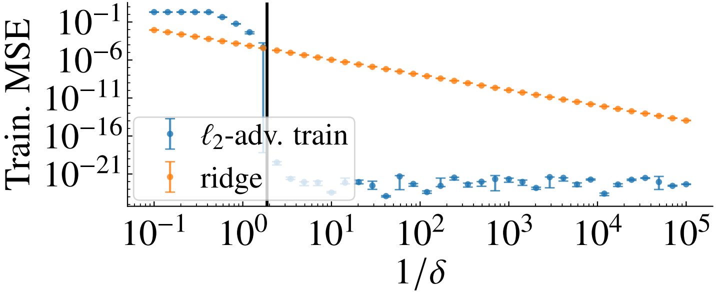
- $\beta^{\text{lasso}}(\lambda) \to \beta^{\text{min}-\ell_1}$ as $\lambda \to 0^+$ (for LARS algorithm).
- $\beta^{\text{ridge}}(\lambda) \to \beta^{\text{min}-\ell_2}$as $\lambda \to 0^+$

#3: Equivalence min-norm interp.

Parameter shrinking: transition **only in the limit**

- $\beta^{\text{lasso}}(\lambda) \to \beta^{\text{min}-\ell_1}$ as $\lambda \to 0^+$ (for LARS algorithm).
- $\beta^{\text{ridge}}(\lambda) \to \beta^{\text{min}-\ell_2}$as $\lambda \to 0^+$

Adversarial training: **abrupt transition**



Driving question

How does adversarial training **compare with** *other regularization methods?*

Regularization methods:

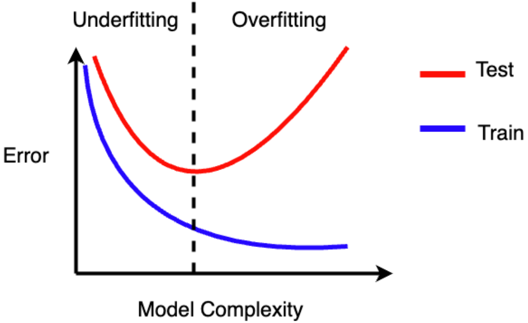
\# 1. Parameter shrinking methods.

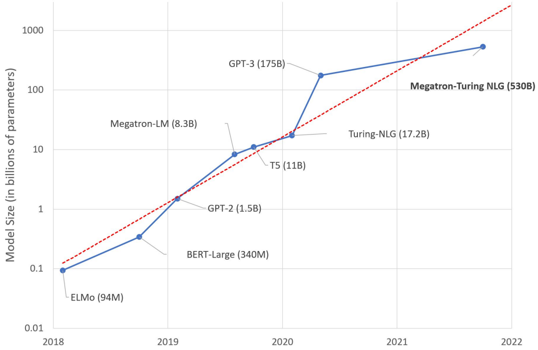
- Lasso.
- Ridge regression.

\# 2. $\sqrt{\text{Lasso}}$

\# 3. Minimum-norm interpolators.

Generalization of deep neural networks





C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. ICLR, 2017

Robustness in large-scale models

"Everything should be made as simple as possible, but not simpler"

Robustness in large-scale models

"Everything should be made as simple as possible, but not simpler"

Questions:

1. **Generalization**.
2. **Robustness**.

The importance of implicit regularization

Solutions of a linear system

$$X\beta = y$$

- **no** solution if $\#\textit{features} < \#\textit{train}$

The importance of implicit regularization

Solutions of a linear system

$$X\beta = y$$

- **no** solution if $\#features < \#train$
- one **unique** solution if $\#features = \#train$

The importance of implicit regularization

Solutions of a linear system

$$X\beta = y$$

- **no** solution if $\#features < \#train$
- one **unique** solution if $\#features = \#train$
- **multiple** solutions if $\#features > \#train$

The importance of implicit regularization

Solutions of a linear system

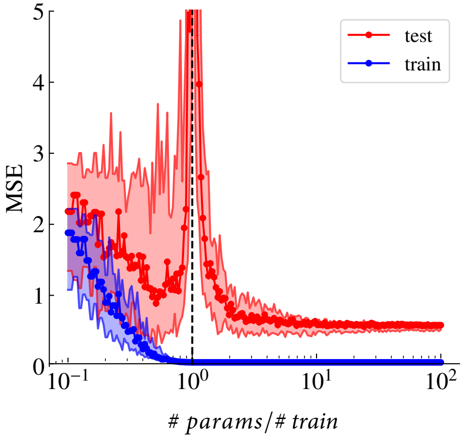
$$X\beta = y$$

- **no** solution if $\#features < \#train$
- one **unique** solution if $\#features = \#train$
- **multiple** solutions if $\#features > \#train$

Gradient descent converges to the minimum-norm solution:

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

Double-descent and benign overfitting



Beyond Occam's Razor in System Identification: Double-Descent when Modeling Dynamics
Antônio H. Ribeiro, Johannes N. Hendriks, Adrian G. Wills, Thomas B. Schön.
IFAC Symposium on System Identification (SYSID), 2021. Honorable mention: Young author award

Simple model of study

- Nonlinear **map** $\phi(x)$, input to feature space

$$\phi : \mathbb{R}^{\#inputs} \mapsto \mathbb{R}^{\#features}.$$

Simple model of study

- Nonlinear **map** $\phi(x)$, input to feature space

$$\phi : \mathbb{R}^{\#inputs} \mapsto \mathbb{R}^{\#features}.$$

- **Linear** model:

$$\hat{y} = \widehat{\beta}^\top \phi(x)$$

Simple model of study

- Nonlinear **map** $\phi(x)$, input to feature space

$$\phi : \mathbb{R}^{\#inputs} \mapsto \mathbb{R}^{\#features}.$$

- **Linear** model:

$$\hat{y} = \widehat{\beta}^\top \phi(x)$$

- **Estimation** procedure:

$$\min_{\beta} \sum_{i=1}^{\#train} (y_i - \widehat{\beta}^\top \phi(x_i))^2$$

Simple model of study

- Nonlinear **map** $\phi(x)$, input to feature space

$$\phi : \mathbb{R}^{\#inputs} \mapsto \mathbb{R}^{\#features}.$$

- **Linear** model:

$$\hat{y} = \widehat{\beta}^\top \phi(x)$$

- **Estimation** procedure:

$$\min_{\beta} \sum_{i=1}^{\#train} (y_i - \widehat{\beta}^\top \phi(x_i))^2$$

- **Optimization** procedure: *Gradient descent starting from zero.*

$$\beta^{i+1} = \beta^i - \gamma \nabla V(\beta^i)$$

Can double descent be observed in adversarial settings?

Can double descent be observed in adversarial settings?

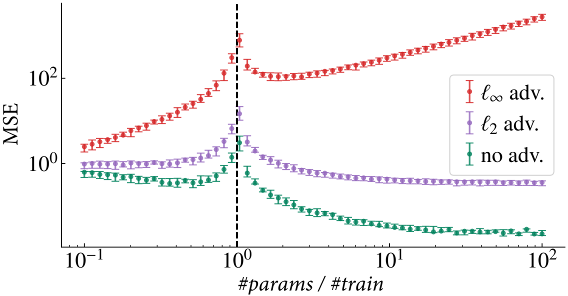


Figure: Adv. risk. minimum ℓ_2-norm interpolator

Overparameterized Linear Regression under Adversarial Attack.
Antônio H. Ribeiro, Thomas B. Schön.
IEEE Transactions on Signal Processing (2023)

Future work

- **Error-in-variables**: adv. train considers worst-case **input disturbances** Δx.
- **Tailored solver:** use in **high-dimensional applications** (genetics).
- **Nonlinear models:** most results still hold for inputs in **infinite spaces.**

Thank you!

✉ antonio.horta.ribeiro@it.uu.se
🌐 antonior92.github.io