Linear adversarial training, robustness in machine learning and applications to cardiology

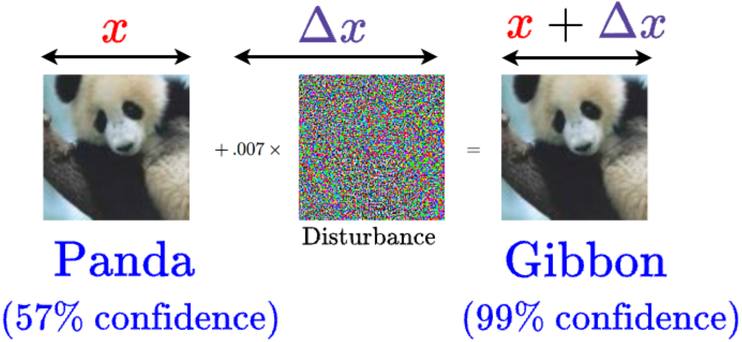
Antônio H. Ribeiro

Uppsala University, Sweden

KTH, Royal Institute of Technology
Stockholm, Sweden

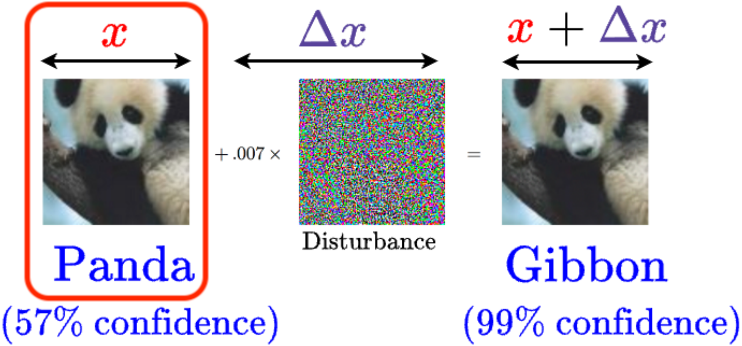
2023

Adversarial attacks



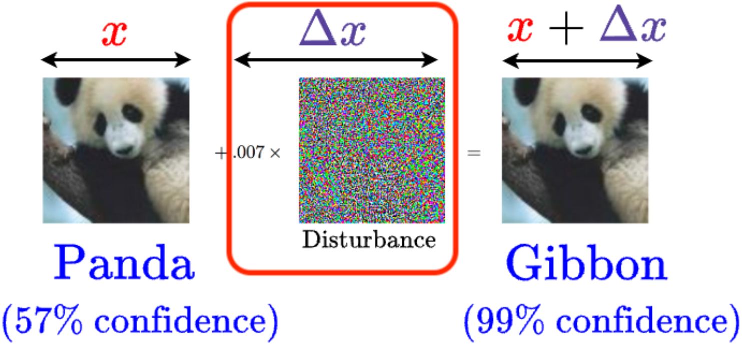
I. J. Goodfellow, J. Shlens, C. Szegedy. Explaining and Harnessing Adversarial Examples, ICLR (2015)

Adversarial attacks



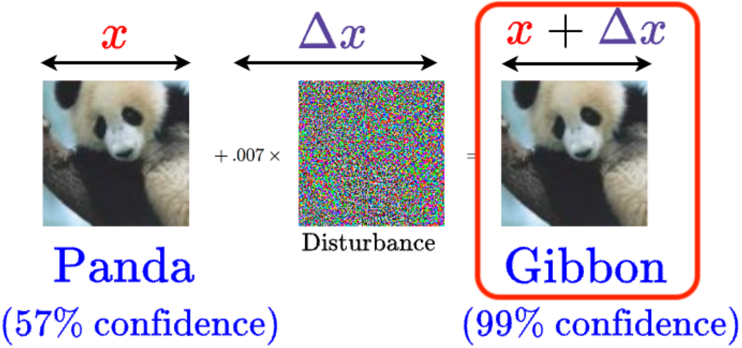
I. J. Goodfellow, J. Shlens, C. Szegedy. Explaining and Harnessing Adversarial Examples, ICLR (2015)

Adversarial attacks



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

Adversarial attacks



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

Adversarial training: *Each training sample is modified by an adversary.*

Part I. Linear adversarial training

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

Part II. Robustness of overparameterized models

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

Part III. Automatic ECG analysis

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

Part I. Linear adversarial training

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

Part II. Robustness of overparameterized models

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

Part III. Automatic ECG analysis

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

Adversarially-trained linear regression

- **Linear regression:**

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

Adversarially-trained linear regression

- **Linear regression:**

$$\min_{\beta} \sum_{i=1}^{\#train} (\underbrace{y_i}_{\text{observed}} - \underbrace{\beta^\top x_i}_{\text{linear prediction}})^2$$

Adversarially-trained linear regression

- **Linear regression:**

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

- **Adversarial training** in linear regression:

$$(y_i - \beta^\top (x_i + \Delta x_i))^2$$

Adversarially-trained linear regression

- **Linear regression:**

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

- **Adversarial training** in linear regression:

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

Adversarially-trained linear regression

- **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

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

Adversarially-trained linear regression

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

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***.*

Adversarially-trained linear regression

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

It can be **rewritten** *as:*

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

where $\|\cdot\|_1$ *is the* **dual norm***.*

Similarities with Lasso

- **ℓ_∞-adversarial attacks:**

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

- **Lasso:**

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

Similarities with Lasso

- **ℓ_∞-adversarial attacks:**

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

- **Lasso:**

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

Similarities with Lasso

- **ℓ_∞-adversarial attacks:**

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

- **Lasso:**

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

Main results:

$\#1$. **Map** $\lambda \leftrightarrow \delta$ for which they yield the **same result.**

Similarities with Lasso

- **ℓ_∞-adversarial attacks:**

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

- **Lasso:**

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

Main results:

#1. **Map** $\lambda \leftrightarrow \delta$ for which they yield the **same result.**

#2. **More parameters than data**: abrupt transition into interpolation.

Similarities with Lasso

- **ℓ_∞-adversarial attacks:**

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

- **Lasso:**

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

Main results:

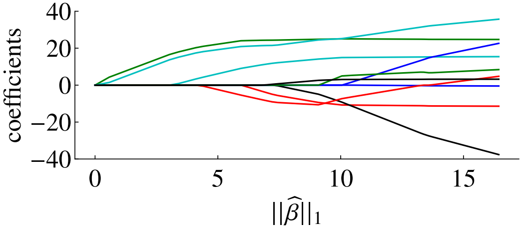
#1. **Map** $\lambda \leftrightarrow \delta$ for which they yield the **same result.**

#2. **More parameters than data**: abrupt transition into interpolation.

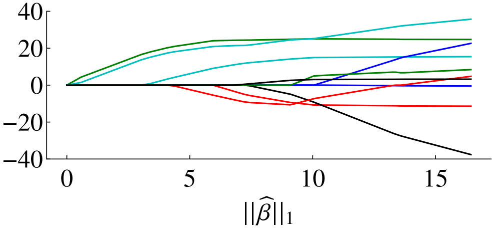
#3. **Optimal choice** of δ independent on noise level.

1. Equivalence with Lasso

Map $\lambda \leftrightarrow \delta$ for which they yield the **same result.**



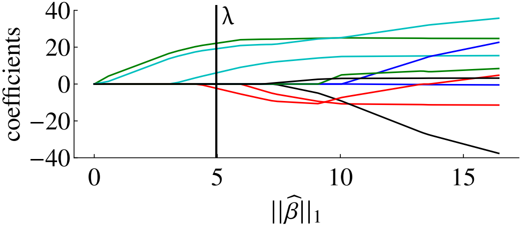
Lasso.



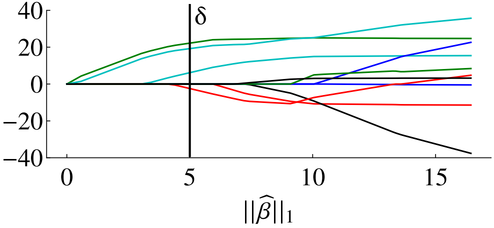
ℓ_∞-adv. training.

1. Equivalence with Lasso

Map $\lambda \leftrightarrow \delta$ for which they yield the **same result.**



Lasso.



ℓ_∞-adv. training.

The that yield the **same result** are **not** necessarily the same, i.e.: $\delta \neq \lambda$

2. More parameters than data

Lasso: transition **only in the limit**

$$\lambda \to 0^+ \Rightarrow \sum_{i=1}^{\#train} \left(y_i - x_i^\top \beta \right)^2 \to 0$$

2. More parameters than data

Lasso: transition **only in the limit**

$$\lambda \to 0^+ \Rightarrow \sum_{i=1}^{\#train} \left(y_i - x_i^\mathsf{T} \beta\right)^2 \to 0$$

Adversarial training:

$$\delta \in (0, \mathrm{threshold}] \Rightarrow \sum_{i=1}^{\#train} \left(y_i - x_i^\mathsf{T} \beta\right)^2 = 0$$

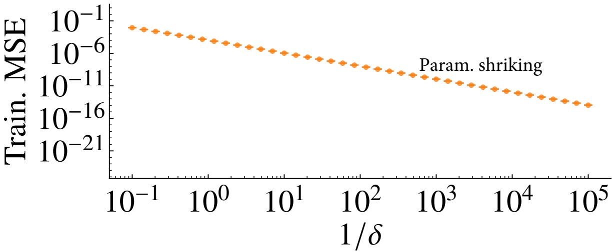
2. More parameters than data

Lasso: transition **only in the limit**

$$\lambda \to 0^+ \Rightarrow \sum_{i=1}^{\#train} \left(y_i - x_i^\mathsf{T}\beta\right)^2 \to 0$$

Adversarial training:

$$\delta \in (0, \text{threshold}] \Rightarrow \sum_{i=1}^{\#train} \left(y_i - x_i^\mathsf{T}\beta\right)^2 = 0$$



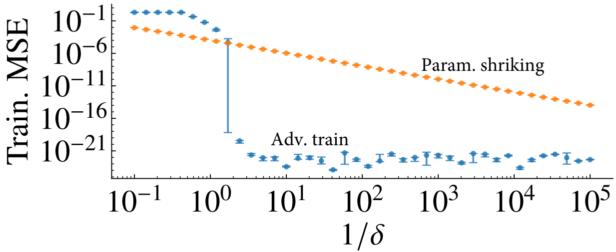
2. More parameters than data

Lasso: transition **only in the limit**

$$\lambda \to 0^+ \Rightarrow \sum_{i=1}^{\#train} \left(y_i - x_i^\mathsf{T} \beta\right)^2 \to 0$$

Adversarial training:

$$\delta \in (0, \text{threshold}] \Rightarrow \sum_{i=1}^{\#train} \left(y_i - x_i^\mathsf{T} \beta\right)^2 = 0$$



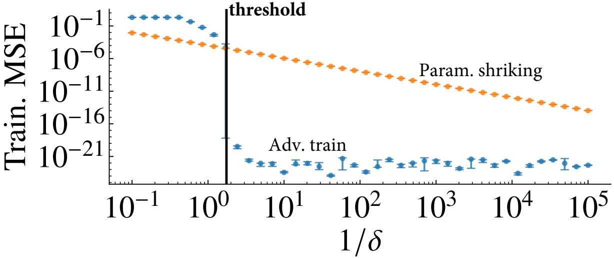
2. More parameters than data

Lasso: transition **only in the limit**

$$\lambda \to 0^+ \Rightarrow \sum_{i=1}^{\#train} \left(y_i - x_i^\mathsf{T}\beta\right)^2 \to 0$$

Adversarial training:

$$\delta \in (0, \text{threshold}] \Rightarrow \sum_{i=1}^{\#train} \left(y_i - x_i^\mathsf{T}\beta\right)^2 = 0$$



2. Equivalence with minimum norm interpolator

For $\delta \in (0, \mathrm{threshold}]$, the minimum-norm interpolator is the solution to adversarial training.

2. Equivalence with minimum norm interpolator

For $\delta \in (0, \mathrm{threshold}]$, the minimum-norm interpolator is the solution to adversarial training.

Relevance

Connect **adversarial training** with **double descent** and **benign overfitting**

3. Invariance to noise levels

To obtain near-oracle performance.

- *Lasso:*

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

- ℓ_∞*-adversarial attack*:

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

3. Invariance to noise levels

To obtain near-oracle performance.

- *Lasso:*

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

- ℓ_∞*-adversarial attack*:

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

Data model

$$y = \underbrace{x^\top \beta^*}_{\text{signal}} + \underbrace{\sigma}_{\text{noise std.}} \varepsilon.$$

Regularization properties of adversarially-trained linear regression

Additional results:

- ℓ_2**-adv.** attacks and **ridge regression.**

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

Regularization properties of adversarially-trained linear regression

Additional results:

- ℓ_2**-adv.** attacks and **ridge regression.**
- Generalization to **other loss** functions

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

Regularization properties of adversarially-trained linear regression

Additional results:

- ℓ_2**-adv.** attacks and **ridge regression.**
- Generalization to **other loss** functions
- Connection to **robust regression** and $\sqrt{\text{Lasso}}$.

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

Part I. Linear adversarial training

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

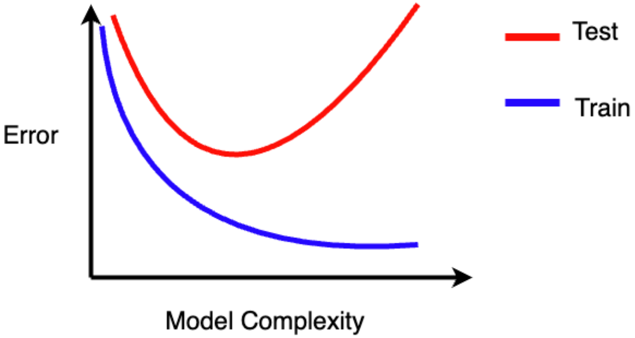
Part II. Robustness of overparameterized models

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

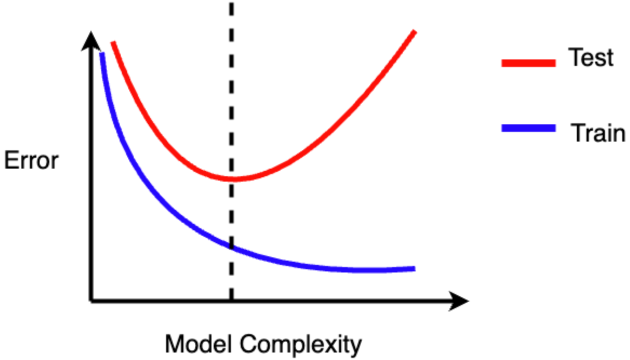
Part III. Automatic ECG analysis

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

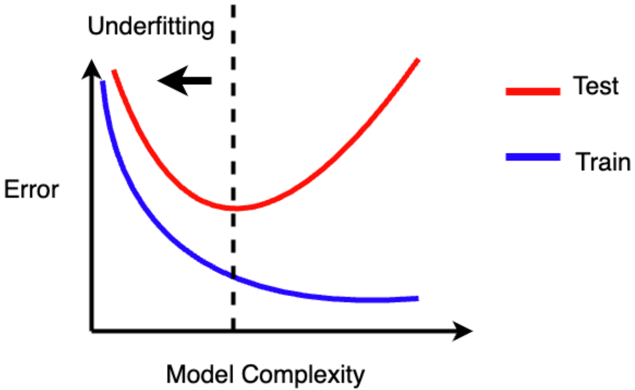
Generalization to new test points



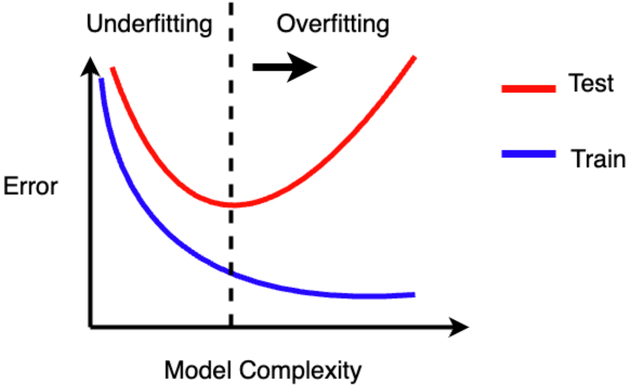
Generalization to new test points



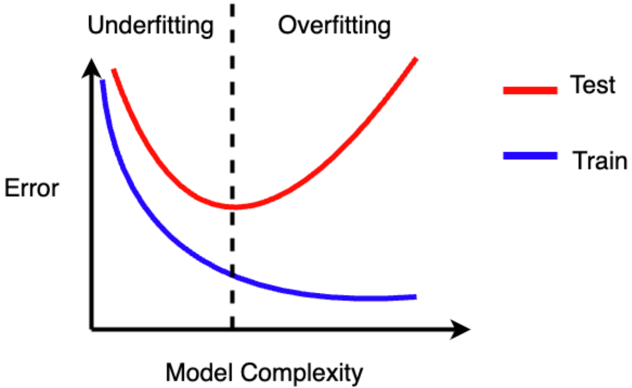
Generalization to new test points



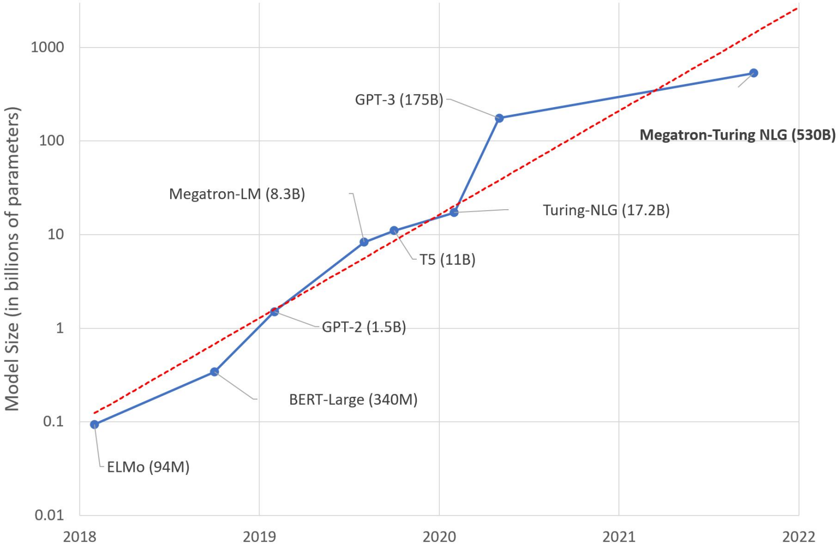
Generalization to new test points



Generalization to new test points

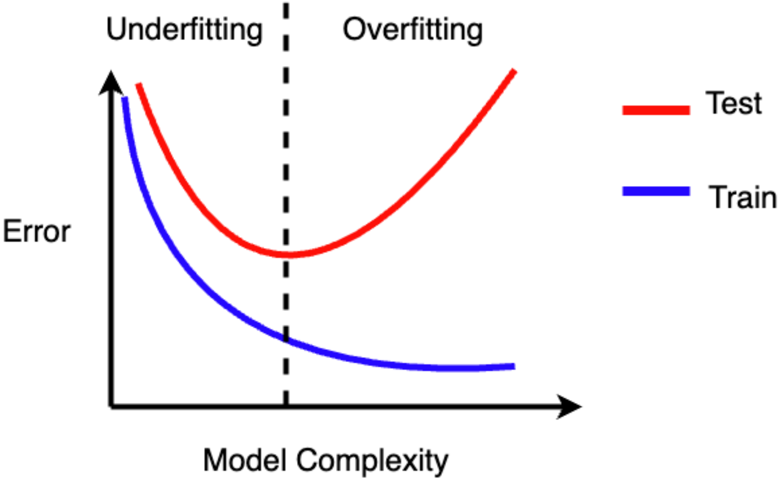


Generalization of deep neural networks



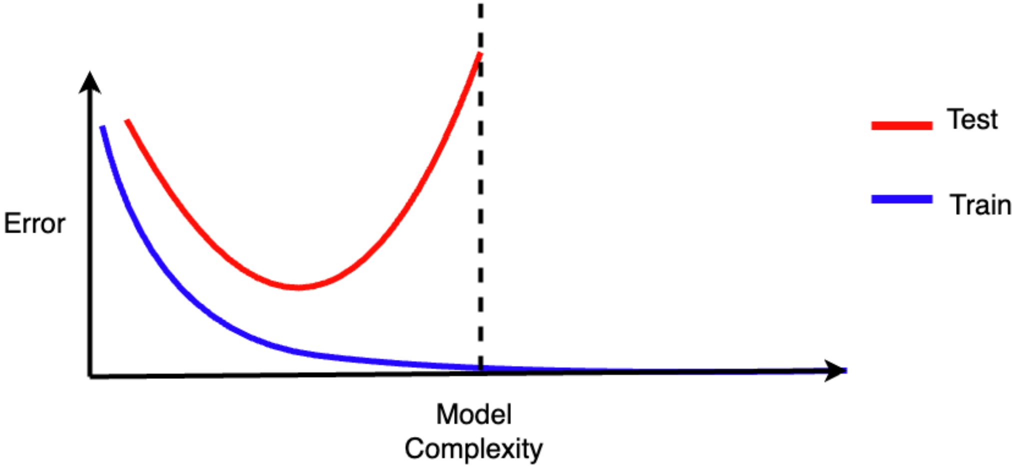
C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. ICLR, 2017

Double-descent curves



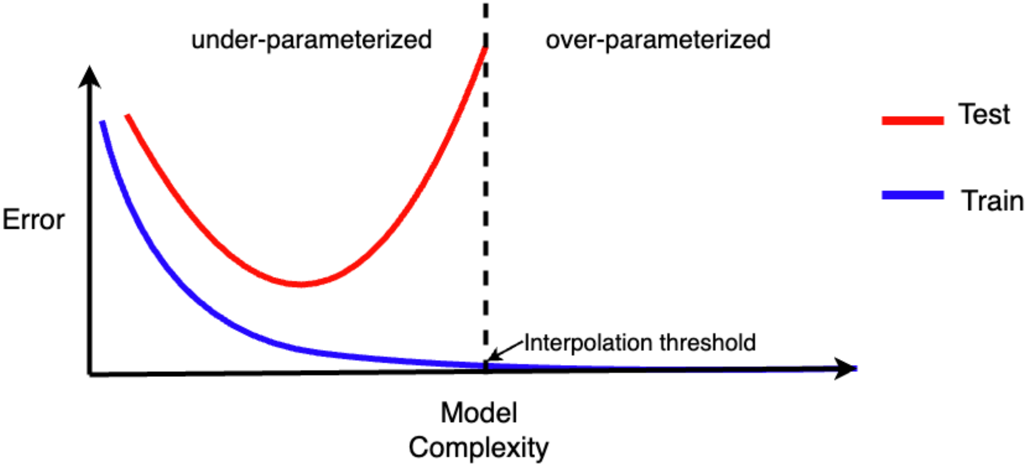
M. Belkin, D. Hsu, S. Ma, S. Mandal. Reconciling modern machine learning practice and the bias-variance trade-off. PNAS, 2020.

Double-descent curves



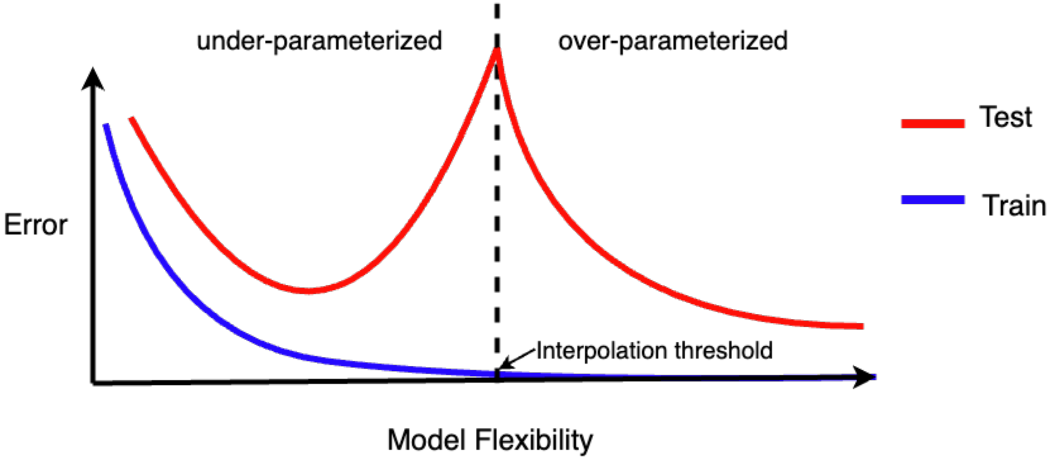
M. Belkin, D. Hsu, S. Ma, S. Mandal. Reconciling modern machine learning practice and the bias-variance trade-off. PNAS, 2020.

Double-descent curves



M. Belkin, D. Hsu, S. Ma, S. Mandal. Reconciling modern machine learning practice and the bias-variance trade-off. PNAS, 2020.

Double-descent curves



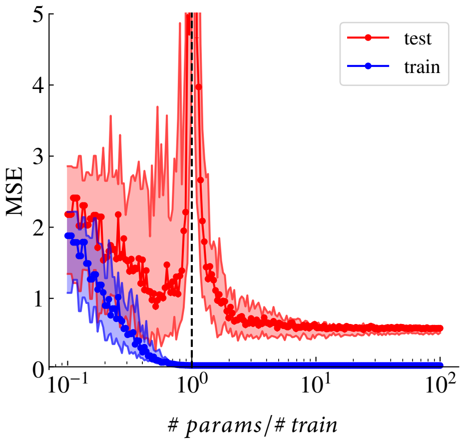
M. Belkin, D. Hsu, S. Ma, S. Mandal. Reconciling modern machine learning practice and the bias-variance trade-off. PNAS, 2020.

Double-descent

- Ph.D. seminar **course**:
  The unreasonable effectiveness of overparameterized machine learning models (3 hp), **2021**

Double-descent

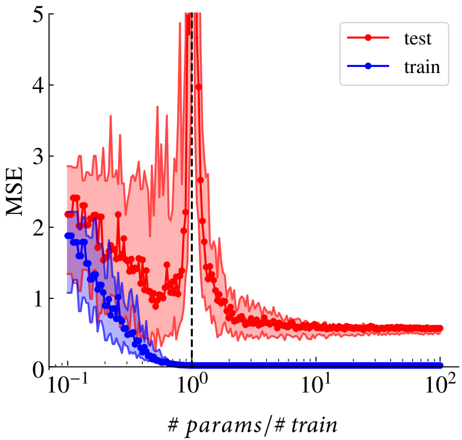
- Ph.D. seminar **course**:
  The unreasonable effectiveness of overparameterized machine learning models (3 hp), **2021**
- Double descent in **dynamical systems**.



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**

Double-descent

- Ph.D. seminar **course**:
  The unreasonable effectiveness of overparameterized machine learning models (3 hp), **2021**
- Double descent in **dynamical systems**.



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**

Deep networks for system identification: a Survey
Gianluigi Pillonetto, Aleksandr Aravkin, Daniel Gedon, Lennart Ljung, Antonio H. Ribeiro, Thomas Bo Schön.
Automatica (Provisionally accepted), 2023.

Can **double descent** be observed **in adversarial settings**?

Can double descent be observed in adversarial settings?

Given a **test point** (x_0, y_0), the error is:

- no adversary

$$(y_0 - \beta^\top x_0)^2$$

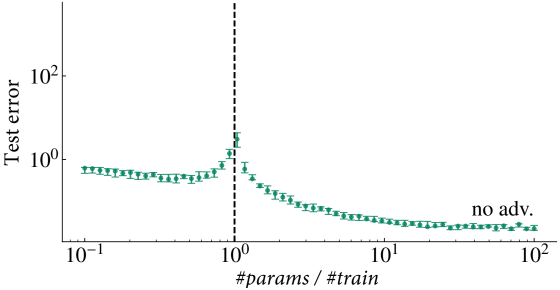


Figure: Adv. risk. min. ℓ_2-norm interpolator

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

Can double descent be observed in adversarial settings?

Given a **test point** (x_0, y_0), the error is:

- no adversary

$$(y_0 - \beta^\top x_0)^2$$

- ℓ_2-adversary

$$\max_{\|\Delta x_0\|_2 \leq \delta} (y_0 - \beta^\top (x_0 + \Delta x_0))^2$$

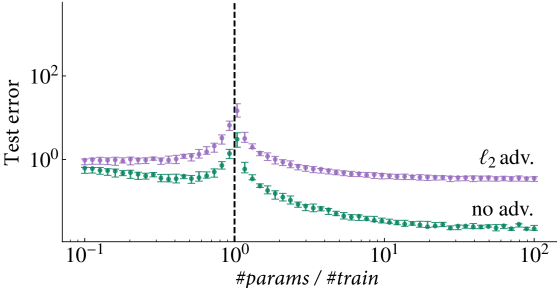


Figure: Adv. risk. min. ℓ_2-norm interpolator

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

Can double descent be observed in adversarial settings?

Given a **test point** (x_0, y_0), the error is:

- no adversary

$$(y_0 - \beta^\top x_0)^2$$

- ℓ_2-adversary

$$\max_{\|\Delta x_0\|_2 \leq \delta} (y_0 - \beta^\top (x_0 + \Delta x_0))^2$$

- ℓ_∞-adversary

$$\max_{\|\Delta x_0\|_\infty \leq \delta} (y_0 - \beta^\top (x_0 + \Delta x_0))^2$$

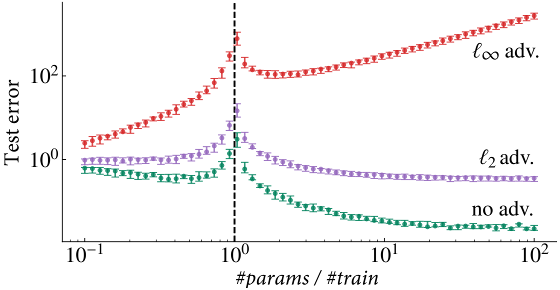


Figure: Adv. risk. min. ℓ_2-norm interpolator

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

Overparameterized Linear Regression under Adversarial Attack

Interpretation

Minimum ℓ_2-norm interpolation $\Leftrightarrow$ ℓ_2**-adversarial training**. (Result #2, Part I)

Overparameterized Linear Regression under Adversarial Attack

Interpretation

Minimum ℓ_2-norm interpolation $\Leftrightarrow$ **ℓ_2-adversarial training**. (Result #2, Part I)

Analysis:

- **Assimptotic results** showing the phenomena
- **Non-asymptotic results**: concentration inequalities

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

Part I. Linear adversarial training

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

Part II. Robustness of overparameterized models

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

Part III. Automatic ECG analysis

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

The electrocardiogram (ECG) exam

Cardiovascular diseases:

- $\approx$18 million **deaths** in 2019 (**32%**).

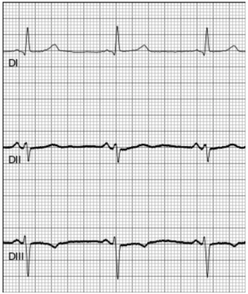
The electrocardiogram (ECG) exam

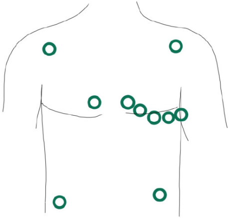
Cardiovascular diseases:

- $\approx$18 million **deaths** in 2019 (**32%**).

The ECG is the **major diagnostic tool**.

- Low-cost, safe and non-invasive
- Can detect arrhythmias, myocardial infarction, cardiomyopathy...





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

Computational electrocardiography

Figure Automated ECG interpretation Glasgow (1971).

Macfarlane, P.W.; Kennedy, J. "Automated ECG Interpretation—A Brief History from High Expectations to Deepest Networks." Hearts 2021.

The transition into end-to-end learning

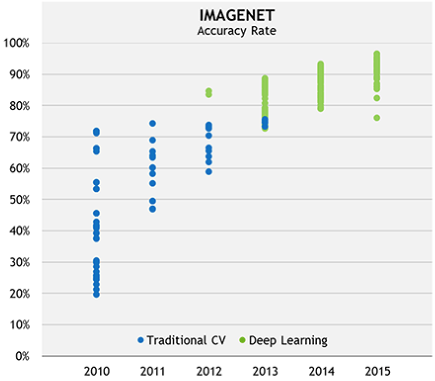


Figure: **Accuracy on Imagenet** as models **transitioned** from feature extraction to end-to-end.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," CVPR (2009)

Telehealth and automatic diagnosis

Figure: State of Minas Gerais

Telehealth and automatic diagnosis

Telehealth Center of Minas Gerais

- 1100 municipalities
- $> 3\,500$ ECGs per day

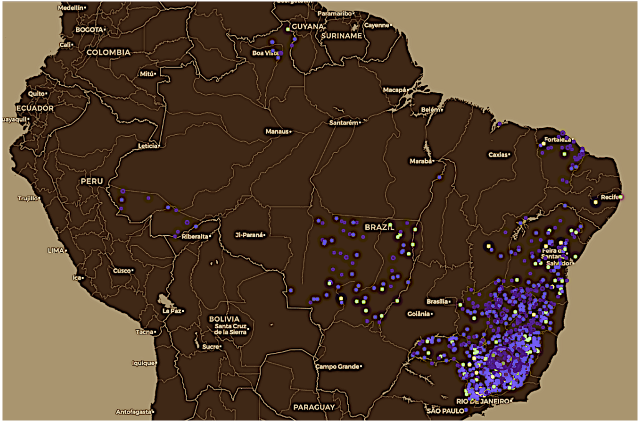


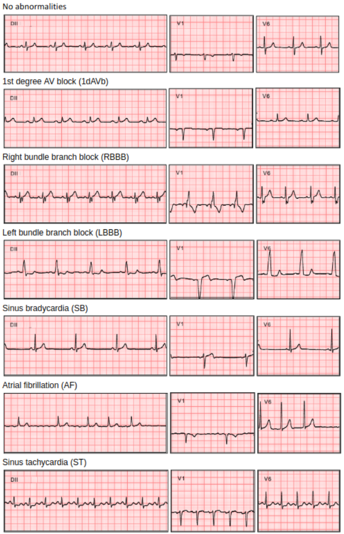
Figure: Municipalities assisted by the telehealth center

Automatic diagnosis of the ECG

- CODE dataset: historical data 2010 to 2017.
  - n =1.6M patients

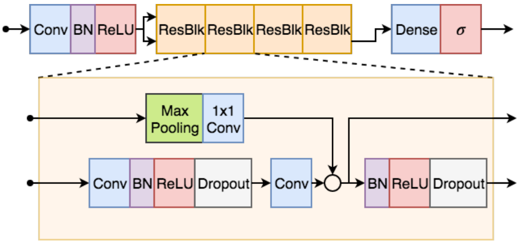
Automatic diagnosis of the ECG

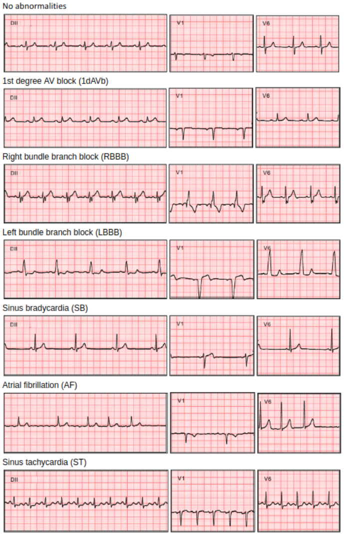
- CODE dataset: historical data 2010 to 2017.
  - n =1.6M patients



Automatic diagnosis of the ECG

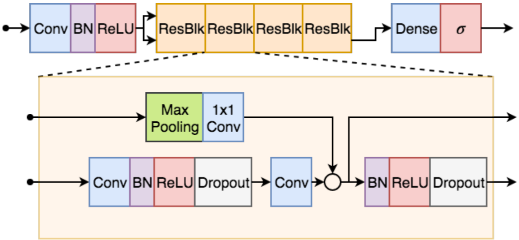
- CODE dataset: historical data 2010 to 2017.
  - n =1.6M patients
- Develop and evaluate deep neural network



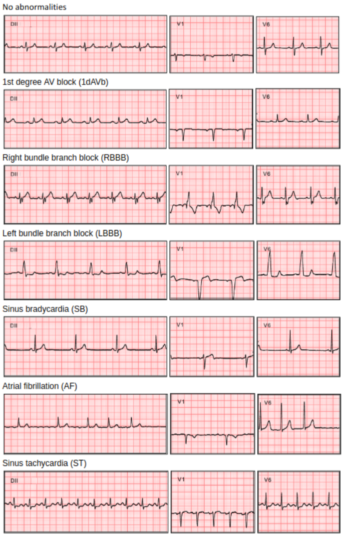


Automatic diagnosis of the ECG

- CODE dataset: historical data 2010 to 2017.
  - n =1.6M patients
- Develop and evaluate deep neural network



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



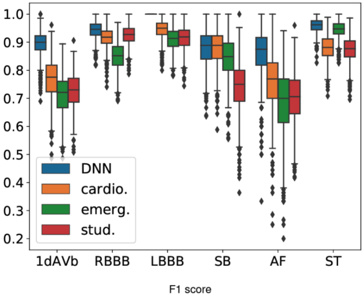
Automatic diagnosis of the ECG (cont.)

- **Result**: Deep neural network (DNN) performs at least as well as experts

cardio. → 4th year cardiology residents

emerg. → 3rd year emergency residents

stud. → 5th year Medical students



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

Automatic diagnosis of the ECG (cont.)

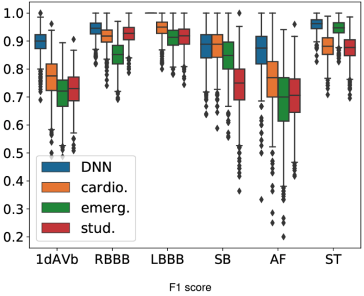
- **Result**: Deep neural network (DNN) performs at least as well as experts

  cardio. → 4th year cardiology residents

  emerg. → 3rd year emergency residents

  stud. → 5th year Medical students

- **Goal:** Improve the **accuracy**



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

Automatic diagnosis of the ECG (cont.)

- **Result**: Deep neural network (DNN) performs at least as well as experts

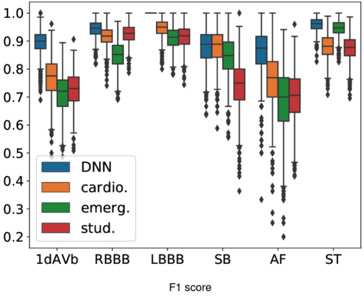
  cardio. → 4th year cardiology residents

  emerg. → 3rd year emergency residents

  stud. → 5th year Medical students

- **Goal:** Improve the **accuracy**

  assist **more patients**



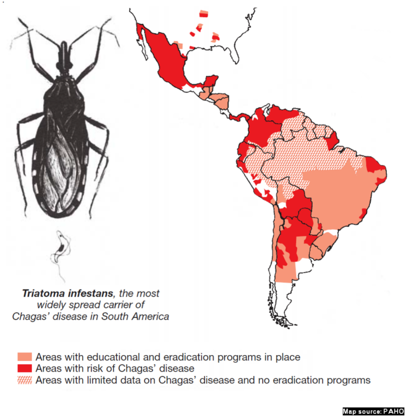
Automatic diagnosis of the 12-lead ECG using a deep neural network
A. H. Ribeiro , M.H. Ribeiro, Paixão, G.M.M. et al
Nature Communications (2020)

Three directions

1. Automatic diagnosis;
2. Screening;
3. Prognosis.

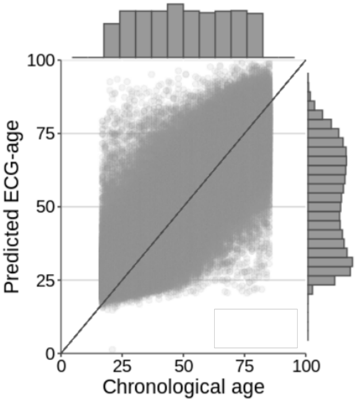
Screening for Chagas disease from the ECG using deep neural networks

- **6 million** people infected.
- Diagnosed with **blood test**.
- Early diagnosis and treatment **halt progression**.
- **Low detection rates**



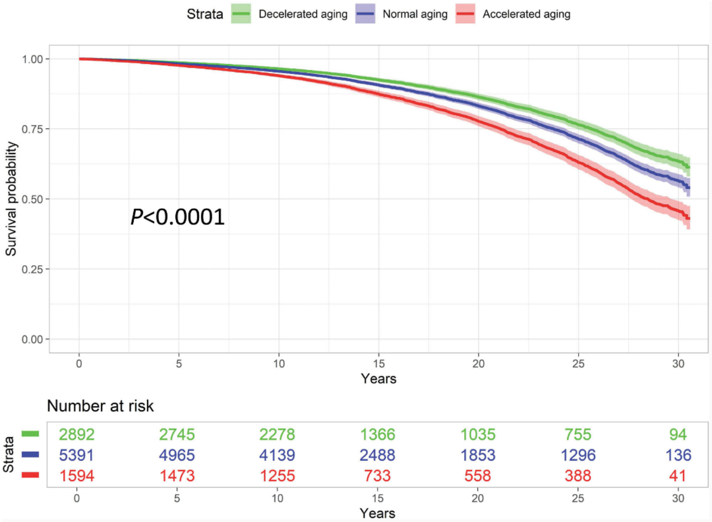
Screening for Chagas disease from the electrocardiogram using a deep neural network
Carl Jidling, Daniel Gedon, Thomas B. Schön, Claudia Di Lorenzo Oliveira, Clareci Silva Cardoso, Ariela Mota Ferreira, Luana Giatti, Sandhi Maria Barreto, Ester C. Sabino, Antônio L. P. Ribeiro, **Antônio H. Ribeiro**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

ECG predicted-age



Deep neural network estimated electrocardiographic-age as a mortality predictor
Emilly M. Lima*, **Antônio H. Ribeiro***, Gabriela MM Paixão*, et. al. *Equal contribution*
Nature Communications (2021)

Risk predictor of cardiovascular events



Electrocardiographic Age Predicts Cardiovascular Events in Community: The Framingham Heart Study
Luisa C C Brant, **Antônio H Ribeiro**, Marcelo M Pinto-Filho, et. al.
Circulation: Cardiovascular Quality and Outcomes (2023)

Challenges

- **Interpretability** Attempt to draw real electrocardiographic **knowledge**.

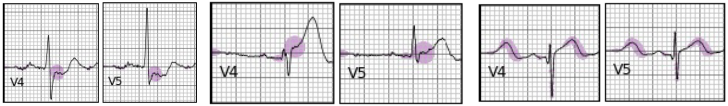


Figure: Grad-CAM plots. **(Left)** STEMI. **(Middle)** STEMI. **(Right)** NSTEMI.

Development and validation of deep learning ECG-based prediction of myocardial infarction in emergency department patients.
S. Gustafsson, D. Gedon, E. Lampa, **Antônio H. Ribeiro**, M. J. Holzmann, T. B. Schön, J. Sundström.
Scientific Reports (2022)

Challenges

- **Interpretability** Attempt to draw real electrocardiographic **knowledge**.

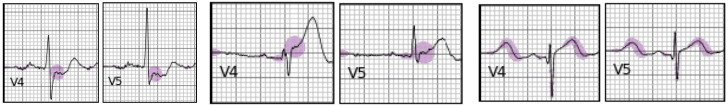


Figure: Grad-CAM plots. **(Left)** STEMI. **(Middle)** STEMI. **(Right)** NSTEMI.

Development and validation of deep learning ECG-based prediction of myocardial infarction in emergency department patients.
S. Gustafsson, D. Gedon, E. Lampa, **Antônio H. Ribeiro**, M. J. Holzmann, T. B. Schön, J. Sundström.
Scientific Reports (2022)

- **Robustness**. Ability to work in **real situations**.

Challenges

- **Interpretability** Attempt to draw real electrocardiographic **knowledge**.

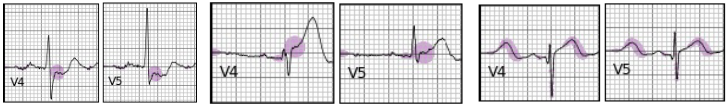


Figure: Grad-CAM plots. **(Left)** STEMI. **(Middle)** STEMI. **(Right)** NSTEMI.

Development and validation of deep learning ECG-based prediction of myocardial infarction in emergency department patients.
S. Gustafsson, D. Gedon, E. Lampa, **Antônio H. Ribeiro**, M. J. Holzmann, T. B. Schön, J. Sundström.
Scientific Reports (2022)

- **Robustness**. Ability to work in **real situations**.

Challenges

- **Interpretability** Attempt to draw real electrocardiographic **knowledge**.

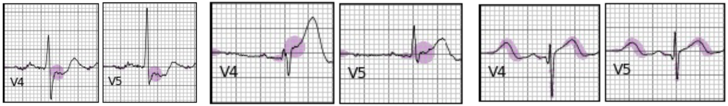


Figure: Grad-CAM plots. **(Left)** STEMI. **(Middle)** STEMI. **(Right)** NSTEMI.

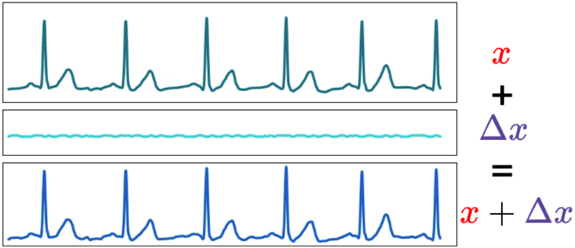
Development and validation of deep learning ECG-based prediction of myocardial infarction in emergency department patients.
S. Gustafsson, D. Gedon, E. Lampa, **Antônio H. Ribeiro**, M. J. Holzmann, T. B. Schön, J. Sundström.
Scientific Reports (2022)

- **Robustness**. Ability to work in **real situations**.

ML algorithms don't need to be really interpreable to be useful in clinical practice.
But they need to be robust!

Adversarial attacks in ECGs

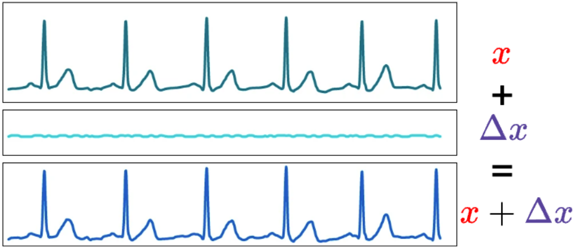
- $x \rightarrow \widehat{y}$:
  Normal (Probability $= 0.99$)



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

Adversarial attacks in ECGs

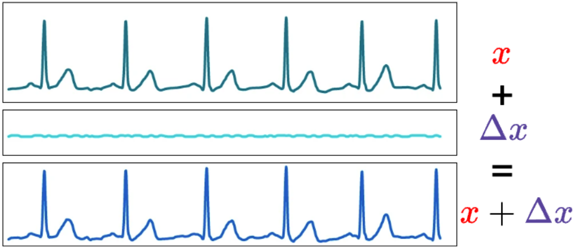
- $x \rightarrow \widehat{y}$: **Normal** (Probability = 0.99)
- $\|\Delta x\| < \delta$



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

Adversarial attacks in ECGs

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



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

Conclusion

- **Large-scale models** have great potential for medicine (and critical applications).
- **Robustness** is a major challenge.
- **Adversarial attacks** framework allows for analysis of **worst-case scenarios**.
- **Linear models** for insight and analysis.
- **Adversarially-trained linear regression** is a competitive regression method.

Thank you!

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