Deciphering Lasso-based Classification Through a Large Dimensional Analysis of the Iterative Soft-Thresholding Algorithm

Malik Tiomoko, Ekkehard Schnoor, Mohamed El Amine Seddik,
Igor Colin, Aladin Virmaux

HUAWEI Noah's Ark Lab, France.
Chair for Mathematics of Information Processing, RWTH Aachen University, Germany.

July 15, 2022

Motivation

- **Lasso:** Amongst the most well-known tools in statistics and signal processing.
- Employ ℓ_1-regularization to impose sparsity on the solution sought by selecting limited number of features.
- Interests recently in the field of classification but lack of interpretability (choice of hyperparameter, statistical understanding)
- Need for a deep theoretical understanding of Lasso scheme for classification
- **State of the art:** Statistical physics-based analysis of Lasso and analysis using CGMT in the regression context
- **In this talk:** Large dimensional of Lasso in a classification context using Random Matrix Theory.
- Application to hyperparameter selection

Context

Observations:

- Samples/data points from two classes $\mathbf{x}_i^{(1)} \in \mathcal{C}_1$ and $\mathbf{x}_i^{(2)} \in \mathcal{C}_2$.

Context

Observations:

- Samples/data points from two classes $\mathbf{x}_i^{(1)} \in \mathcal{C}_1$ and $\mathbf{x}_i^{(2)} \in \mathcal{C}_2$.
- Data matrix $\mathbf{X} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}]$ with $\mathbf{X}^{(\ell)} = [\mathbf{x}_1^{(\ell)}, \ldots, \mathbf{x}_{n_\ell}^{(\ell)}]$, $\mathbf{x}_i^{(\ell)} \in \mathbb{R}^p$.
- Associated labels $y_i^{(\ell)}$ in $\mathbf{y} = [y_1^{(1)}, \ldots, y_{n_1}^{(1)}, y_1^{(2)}, \ldots, y_{n_2}^{(2)}]^\mathsf{T} \in \{-1, 1\}^n$.

Context

Observations:

- Samples/data points from two classes $\mathbf{x}_i^{(1)} \in \mathcal{C}_1$ and $\mathbf{x}_i^{(2)} \in \mathcal{C}_2$.
- Data matrix $\mathbf{X} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}]$ with $\mathbf{X}^{(\ell)} = [\mathbf{x}_1^{(\ell)}, \ldots, \mathbf{x}_{n_\ell}^{(\ell)}]$, $\mathbf{x}_i^{(\ell)} \in \mathbb{R}^p$.
- Associated labels $y_i^{(\ell)}$ in $\mathbf{y} = [y_1^{(1)}, \ldots, y_{n_1}^{(1)}, y_1^{(2)}, \ldots, y_{n_2}^{(2)}]^\mathsf{T} \in \{-1, 1\}^n$.

Objective:

- Given a new test datum $\mathbf{x}$, our goal is to predict its associated label $\mathbf{y}$ using a linear classifier obtained through Lasso.

Context

Observations:

- Samples/data points from two classes $\mathbf{x}_i^{(1)} \in \mathcal{C}_1$ and $\mathbf{x}_i^{(2)} \in \mathcal{C}_2$.
- Data matrix $\mathbf{X} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}]$ with $\mathbf{X}^{(\ell)} = [\mathbf{x}_1^{(\ell)}, \ldots, \mathbf{x}_{n_\ell}^{(\ell)}]$, $\mathbf{x}_i^{(\ell)} \in \mathbb{R}^p$.
- Associated labels $y_i^{(\ell)}$ in $\mathbf{y} = [y_1^{(1)}, \ldots, y_{n_1}^{(1)}, y_1^{(2)}, \ldots, y_{n_2}^{(2)}]^\mathsf{T} \in \{-1, 1\}^n$.

Objective:

- Given a new test datum $\mathbf{x}$, our goal is to predict its associated label $\mathbf{y}$ using a linear classifier obtained through Lasso.

Prediction steps:

- Sep. hyperplane: solution $\boldsymbol{\omega}^\star$ of the (convex, but non-smooth!) min. problem

$$\operatorname{argmin}_{\boldsymbol{\omega} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}^\mathsf{T} \boldsymbol{\omega}\|_2^2 + \lambda \|\boldsymbol{\omega}\|_1. \qquad \text{(Lasso)}$$

Context

Observations:

- Samples/data points from two classes $\mathbf{x}_i^{(1)} \in \mathcal{C}_1$ and $\mathbf{x}_i^{(2)} \in \mathcal{C}_2$.
- Data matrix $\mathbf{X} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}]$ with $\mathbf{X}^{(\ell)} = [\mathbf{x}_1^{(\ell)}, \ldots, \mathbf{x}_{n_\ell}^{(\ell)}]$, $\mathbf{x}_i^{(\ell)} \in \mathbb{R}^p$.
- Associated labels $y_i^{(\ell)}$ in $\mathbf{y} = [y_1^{(1)}, \ldots, y_{n_1}^{(1)}, y_1^{(2)}, \ldots, y_{n_2}^{(2)}]^\mathsf{T} \in \{-1, 1\}^n$.

Objective:

- Given a new test datum $\mathbf{x}$, our goal is to predict its associated label $\mathbf{y}$ using a linear classifier obtained through Lasso.

Prediction steps:

- Sep. hyperplane: solution $\boldsymbol{\omega}^\star$ of the (convex, but non-smooth!) min. problem

$$\operatorname{argmin}_{\boldsymbol{\omega} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}^\mathsf{T} \boldsymbol{\omega}\|_2^2 + \lambda \|\boldsymbol{\omega}\|_1. \qquad \text{(Lasso)}$$

- Given the optimal separating hyperplane $\boldsymbol{\omega}^\star$, classification performed by sign of

$$g(\mathbf{x}) = \boldsymbol{\omega}^{\star\mathsf{T}} \mathbf{x}.$$

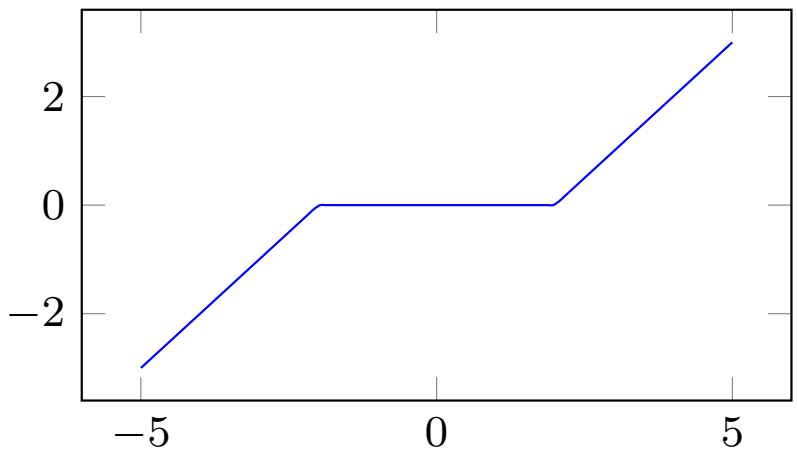
- Solve (Lasso) via the **iterative soft-thresholding algorithm (ISTA)**.

Iterative soft-thresholding algorithm

- For a sparse minimization of the differentiable function $h(\boldsymbol{\omega}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}^\mathsf{T}\boldsymbol{\omega}\|_2^2$, do

$$\textbf{Gradient step: } \mathbf{z}^j = \boldsymbol{\omega}^{j-1} - \tau \nabla h\left(\boldsymbol{\omega}^{j-1}\right),$$
$$\textbf{Sparsity step: } \boldsymbol{\omega}^j = S_{\tau\lambda}\left(\mathbf{z}^j\right),$$

  with τ the step size and $S_{\tau\lambda}$ the soft threshold function defined below.



$$S_\lambda(x) = \text{sign}(x) \cdot \max(0, |x| - \lambda)$$

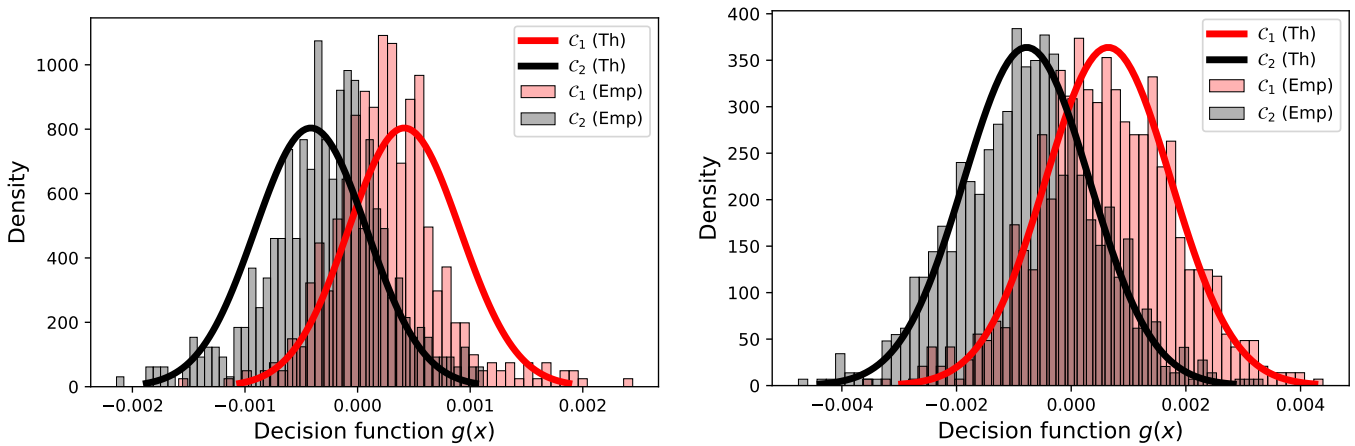
- Applied to Lasso-based classification $\boldsymbol{\omega}^\star$ via ISTA (initialization $\boldsymbol{\omega}^0 = \mathbf{0} \in \mathbb{R}^p$):

$$\boldsymbol{\omega}^{j+1} = S_{\tau\lambda}\left(\boldsymbol{\omega}^j + \tau\mathbf{X}\left(\mathbf{y} - \mathbf{X}^\mathsf{T}\boldsymbol{\omega}^j\right)\right)$$

- **Goal:** Predict (asymptotically precise) classification accuracy under this framework.

Experiments

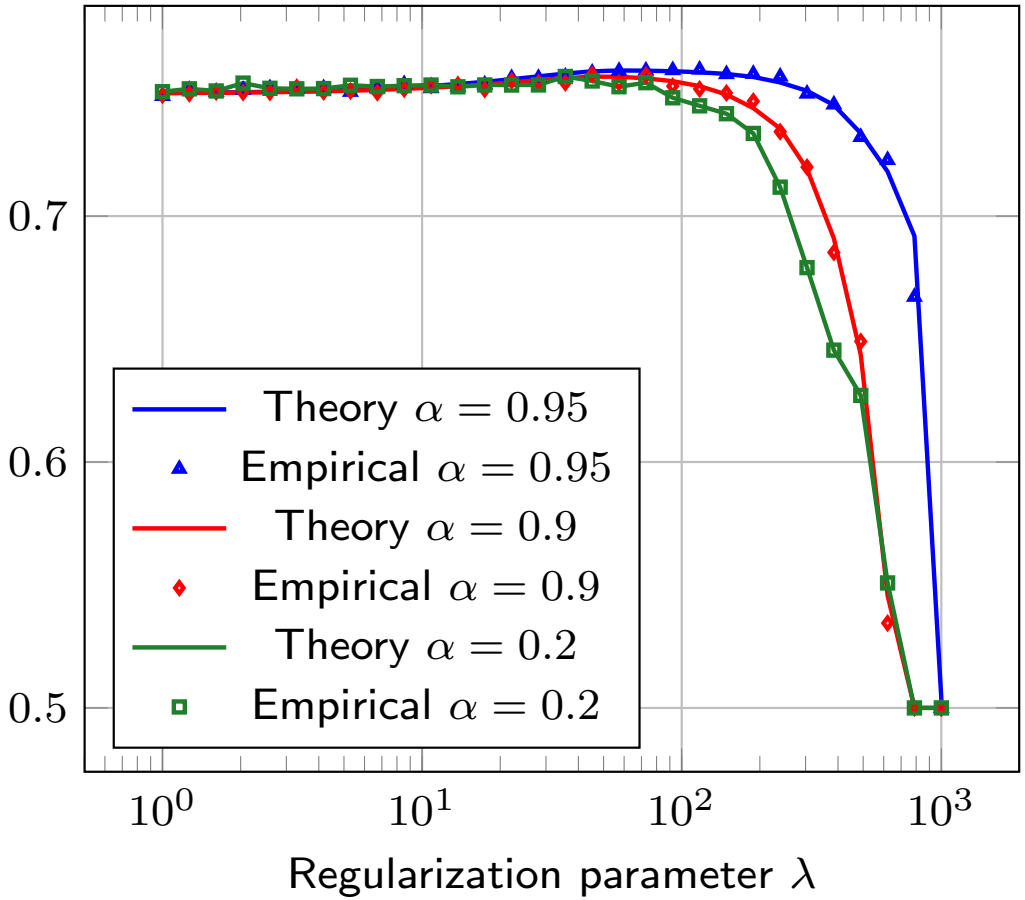
Goal: Predict classification accuracy from only statistical properties (mean, covariance) of the training set!



(Left) Amazon review dataset ("review to score - positiv vs. negative") for two score classes with dim. $p = 400$ and $n_1 = n_2 = 100$. (right) MNIST dataset ("4" vs. "9"). Histogram of the values of the classification score $g(\mathbf{x}) = \boldsymbol{\omega}^{\star\mathsf{T}}\mathbf{x}$ generated from 400 test samples.

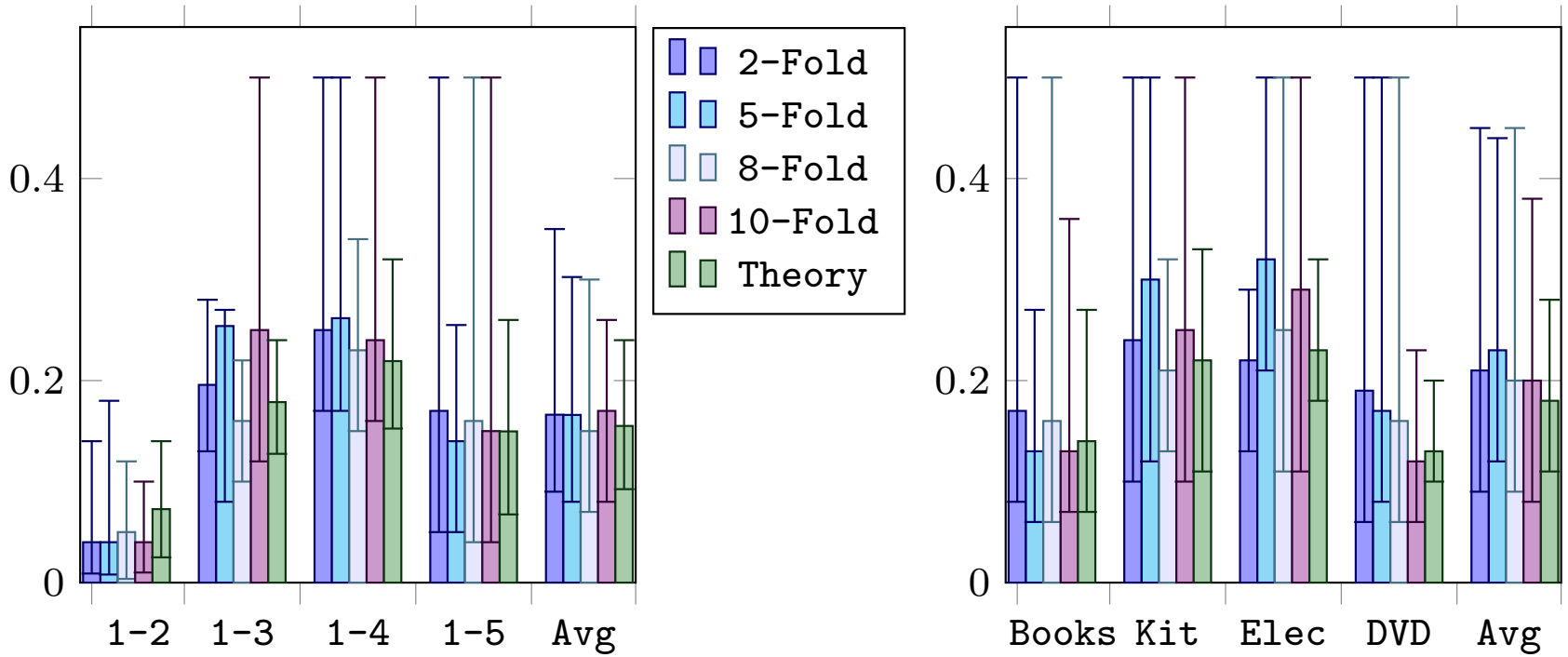
- Close fit between the theoretical decision score and the empirical even on real data.
- Possibility to predict in advance the classification error and best hyperparameters.

Regularization parameter analysis



Close fit between the theoretical and empirical (averaged over $1\,000$ test samples) classification accuracy (as a function of λ), for three different values of α (sparsity level). Gaussian mixture model with class sizes $n_1, n_2 = 500$ and $\mathbf{x}_i^{(\ell)} \sim \mathcal{N}(\boldsymbol{\mu}_\ell, \mathbf{I}_p)$, for $\ell = 1, 2$, with mean $\boldsymbol{\mu}_\ell = (-1)^\ell \mathbf{b} \odot \mathbf{m}$, where $\mathbf{m} \sim \mathcal{N}(\mathbf{0}_p, \frac{1}{p}\mathbf{I}_p)$, and where $\mathbf{b}$ is a Bernoulli random vector that puts each single entry to zero with probability α/p, with the feature size $p = 100$.

Application to hyperparameter selection



Empirical classification error for different tasks; **(Left)** MNIST: $p = 100$-PCA preprocessing, $n_1 = n_2 = 20$, 500 test samples. **(Right)** Amazon Review dataset: Positive vs. negative review for different classes (Books, Kitchen, Electronics, DVD) with $n_1 = n_2 = 20$, 2 000 test samples.

Concluding remarks

- Theoretical analysis of a Lasso-based classification through the analysis of an iterative algorithm (ISTA).
- Interesting insights into its applicability in a classification context, but also offers a reliable alternative to cross-validation.
- Theoretical perspectives on the analysis of iterative processes that induce very strong dependencies between data (Stochastic Gradient Descent and tensor-based classification algorithms).
- Efficient use of the Lasso in real applications by appropriately choosing the regularization parameter.