HUAWEI

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

Malik Tiomoko [1] Ekkehard Schnoor [2] Mohamed El Amine Seddik [1] Igor Colin [1] Aladin Virmaux [1]

[1]Huawei Noah's Ark Lab, Paris, France [2]Chair for Mathematics of Information Processing, RWTH Aachen University, Germany

RWTH AACHEN UNIVERSITY

Abstract

- **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 work:** Large dimensional of Lasso in a classification context using Random Matrix Theory (RMT).
- Application to hyperparameter selection.

Context

Observations:

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

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

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

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

- Solve equation 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

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

- 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.

Growth Rate

As $n \to \infty$, $p \to \infty$, we assume $p/n \to c_0 > 0$ and $n_\ell/n \to c_\ell \in (0,1)$, $\ell = 1, 2$.

Distribution of $\mathbf{X}$ and $\mathbf{x}$

There exist two constants $C, c > 0$ (independent of n, p) such that, for any 1-Lipschitz function $f : \mathbb{R}^{p\times n} \to \mathbb{R}$,

$$\mathbb{P}(|f(\mathbf{X}) - m_{f(\mathbf{X})}| \geq t) \leq Ce^{-(t/c)^2} \qquad \forall t > 0,$$

where m_Z is a median of the random variable Z. We require that the columns of $\mathbf{X}$ are independent and that for $\ell \in \{1, 2\}$, $\mathbf{x}_1^{(\ell)}, \ldots, \mathbf{x}_{n_\ell}^{(\ell)}$ are i.i.d. such that $\mathrm{Cov}(\mathbf{x}_i^{(\ell)}) = \mathbf{I}_p$.

Main ingredients of the theory

- **Goal:** Track how the randomness of the data $\mathbf{X}$ induces randomness onto $\boldsymbol{\omega}^j$ (and, in the limit: $\boldsymbol{\omega}^\star$), which is calculated through ISTA.
- Main focus on estimating mean $\mathbb{E}[\boldsymbol{\omega}^\star]$ and covariance $\mathrm{Cov}(\boldsymbol{\omega}^\star)$ of $\boldsymbol{\omega}^\star$.
- Construct an iterative scheme (with $\mathbf{z}^j = \boldsymbol{\omega}^j - \tau\mathbf{X}\mathbf{X}^\mathsf{T}\boldsymbol{\omega}^j + \tau\mathbf{X}\mathbf{y}$)

$$\mathbb{E}\left[\boldsymbol{\omega}^{j+1}\right] = \mathbb{E}\left[S_{\tau\lambda}(\mathbf{z}^j)\right].$$

- Prove that $\mathbf{z}^j$ is gaussian random vector which allows to write

$$\mathbb{E}\left[\boldsymbol{\omega}^{j+1}\right] = \varphi\left(\tau\lambda, \bar{\mathbf{z}}^j, \boldsymbol{\sigma}_{\mathbf{z}^j}\right),$$

where for random vector $\mathbf{v}$ we denote $\bar{\mathbf{v}} = \mathbb{E}[\mathbf{v}]$ and $\boldsymbol{\sigma}_\mathbf{v}$ the diagonal of $\mathrm{Cov}(\mathbf{v})$ with

$$\varphi : \mathbb{R}_{>0} \times \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^p,$$
$$(\lambda, \bar{\mathbf{v}}, \boldsymbol{\sigma}_\mathbf{v}) \mapsto \mathbb{E}_{\mathbf{v}\sim\mathcal{N}(\bar{\mathbf{v}}, \boldsymbol{\Sigma}_\mathbf{v})}[S_\lambda(\mathbf{v})].$$

- Final step: Estimate the quantities $\bar{\mathbf{z}}^j$ and $\boldsymbol{\sigma}_{\mathbf{z}^j}$.

Proof idea (continued)

- For illustration, focus just on $\bar{\mathbf{z}}^j$. By linearity of expectation

$$\bar{\mathbf{z}}^j = \mathbb{E}\left[\boldsymbol{\omega}^j - \tau\mathbf{X}\mathbf{X}^\mathsf{T}\boldsymbol{\omega}^j + \tau\mathbf{X}\mathbf{y}\right]$$
$$= \bar{\boldsymbol{\omega}}^j - \tau\sum_{i=1}^n \mathbb{E}\left[(\boldsymbol{\omega}^{j\mathsf{T}}\mathbf{x}_i)\mathbf{x}_i\right] + \tau\mathbb{E}[\mathbf{X}\mathbf{y}],$$

- Disentangle strong dependency between $\boldsymbol{\omega}^j$ at iteration j and the (columns of the) data matrix $\mathbf{X}$ → **leave-one out** approach.
- Approximate $\mathbb{E}[\boldsymbol{\omega}^{j\mathsf{T}}\mathbf{x}_i]$ for both classes using the functions

$$\zeta_{\mathcal{C}_{\pi(i)}}\left(\mathbb{E}\left[\mathbf{x}_i^\mathsf{T}\boldsymbol{\omega}_{-i}^j\right]\right), \qquad \pi(i) \in \{1, 2\},$$

- Functions $\zeta_{\mathcal{C}_{\pi(i)}}$ are established through determining the difference between $\boldsymbol{\omega}_{-i}^j$ and $\boldsymbol{\omega}^j$ using the interpolating function

$$\boldsymbol{\omega}_{-i}^j(t) = S_{\tau\lambda}\left(\boldsymbol{\omega}_{-i}^j(t) + \tau\mathbf{X}_{-i}\left(\mathbf{y}_{-i} - \mathbf{X}_{-i}^\mathsf{T}\boldsymbol{\omega}_{-i}^j(t)\right)\right.$$
$$\left. + \tau t\mathbf{x}_i(\mathbf{y}_i - \boldsymbol{\omega}_{-i}^j(t)^\mathsf{T}\mathbf{x}_i)\right), \quad t \in [0, 1],$$

Theory versus simulations

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

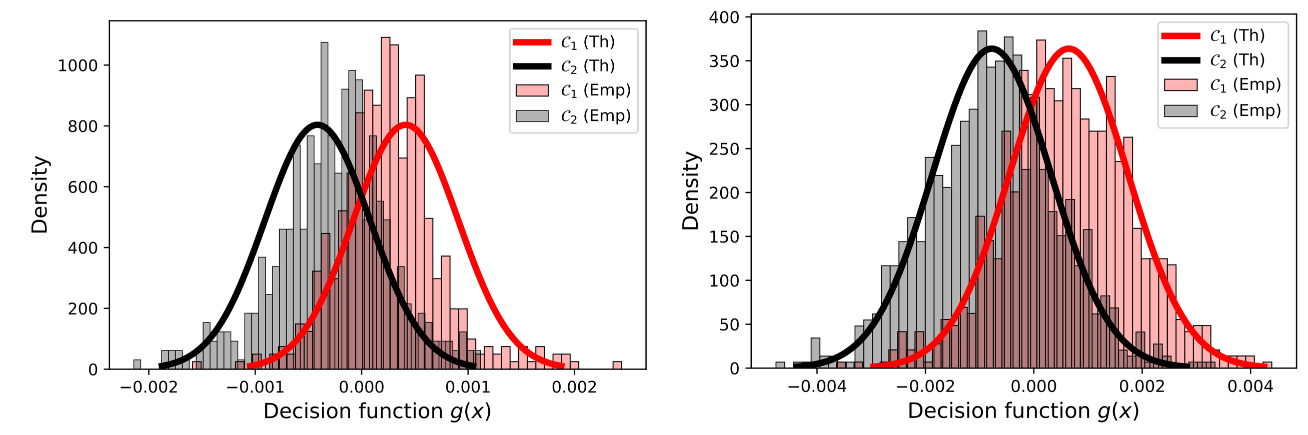

Figure 1. (Left) Amazon review dataset ("review to score - positiv vs. negative") for two score classes with dim. $p = 400$ and $n_1 = n_2 = 100$. (right) MNIST dataset ("4" vs. "9"). Histogram of the values of the classification score $g(\mathbf{x}) = \boldsymbol{\omega}^{\star\mathsf{T}}\mathbf{x}$ 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.

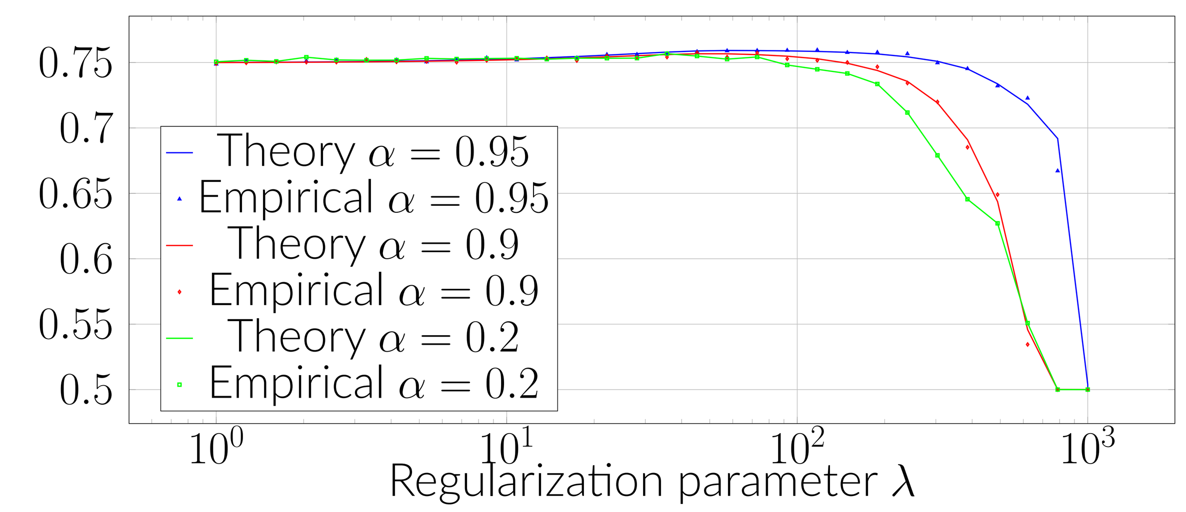

Figure 2. Close fit between the theoretical and empirical (averaged over 1 000 test samples) classification accuracy (as a function of λ), for three different values of α (sparsity level). Gaussian mixture model with class sizes $n_1, n_2 = 500$ and $\mathbf{x}_i^{(\ell)} \sim \mathcal{N}(\boldsymbol{\mu}_\ell, \mathbf{I}_p)$, 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 $p = 100$.

Conclusion

- 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

International Conference on Machine Learning - July 2022
Link to article: `https://melaseddik.github.io/files/icml22_lasso.pdf`
Get in touch: `https://tiomokomalik.wixsite.com/mysite`