SCHOOL OF DATA SCIENCE
數據科學學院

# Internship proposal:

## Empirical risk minimization in High Dimension.

| | |
|---|---|
| **Place:** | Chinese University of Hongkong, Shenzhen, China mainland |
| | School of Data science. |
| **Supervisor:** | Cosme Louart – cosmelouart@cuhk.edu.cn |
| | Assistant professor in the Department of Statistics |
| **Working period:** | For a period of 4 to 6 months, starting from 2025-01-01 |
| **Stipend:** | 10000 RMB/month (around 1330 EUR) - Airfare covered |
| **Underlying knowledge:** | Random matrix theory – Concentration of measure theory – |
| | Optimization – Monte Carlo techniques – Kernel regression |

### Problem

The projects concerns the study of a broad range of regularized empirical risk minimization problems that write:

$$\text{Minimize} \quad \frac{1}{n}\sum_{i=1}^{n} L_i(x_i^\top \beta) + \rho(\beta)^2, \qquad \beta \in \mathbb{R}^p, \tag{1}$$

where $x_1, \dots, x_n \in \mathbb{R}^p$ are the data, $L_1, \dots, L_n : \mathbb{R} \to \mathbb{R}$ are convex loss functions that could possibly contain label information (in supervised setting), and $\rho : \mathbb{R}^p \to \mathbb{R}_+$ is a regularization convex mapping. A large number of applications fit into this framework (all kinds of regression, support vector machines...), but we are interested here first of all in the theoretical insight.

### Literature review

Extensive theoretical work has been carried out for several years, mainly inspired by statistical physics techniques such as approximate message passing in [SC19]; [DM16], convex Gaussian minmax theorem in [TOS20]; [DKT22], mean field theory in [Mig+20a], or replica method in [Mig+20b]; [SZ22]. In the case of a $\ell_2$ regularization, random matrix theory can give an exact prediction of the error depending on the mean and covariance of the data [Kar13]. It was then applied to softmax classifiers in [ML19]; [Sed+21] and later extended to lasso regression [Tio+22], showing that exact error prediction may be achievable for general convex regularizations.

### Our approach

These seemingly perfect fit predictions may look impressive, but they have not yet yet led to significant improvements or new insights on available algorithms (choice of loss and regularization function, optimization of the weighting between them...). This is mainly due to the complexity of the formulation of the first two moments (mean and covariance) of the parameter vector $\beta \in \mathbb{R}^p$ solution to (1). Under general assumptions (e.g.,$\rho = \frac{1}{2}\|\cdot\|^2$ and concentration of measure and indendence assumption on $x_1, \dots x_n$) the mean and the covariance of $\beta$ are each close to $\mu_\beta = Q\bar{\mu}, \quad C_\beta = \frac{1}{n}Q\bar{C}Q$, for $Q \in \mathbb{R}^{p\times p}$, $\bar{\mu} \in \mathbb{R}^p$ and $\bar{C} \in \mathbb{R}^{p\times p}$ solutions of the fixed point equation:

- $z \sim \mathcal{N}(\bar{z}, \sigma)$
- $\xi : \mathbb{R}^n \to \mathbb{R}$ satisfying for any $z \in \mathbb{R}^n$, $\xi(z) = L'(z + \Delta\xi(z))$
- $K = \mathbb{E}[\xi'(z)] \otimes C, \quad Q = (I_p - K)^{-1}, \quad \Delta = \frac{1}{n}\operatorname{tr}(CQ)$

- $\bar{\mu} = \mathbb{E}[\xi(z)] \otimes \mu, \quad \bar{C} = \mathbb{E}[\xi(z)]\mathbb{E}[\xi(z)]^\top \otimes C$
- $\bar{z} = \mu^\top Q \bar{\mu}, \quad \sigma = \frac{1}{n} \operatorname{tr}\left(CQ\bar{C}Q\right) + \bar{\mu}^\top Q \bar{C} Q \bar{\mu}$

where $\mu = \mathbb{E}[x]$ and $C \equiv \mathbb{E}[xx^\top] - \mu\mu^\top$.

### Expected outcome

We give below the program of the intership (it is arguably too long to be fullfilled)

1. Help with the writing of an article in progress on low density transfer learning relying on this equation. This will help familiarize the intern with the mathematical tools and simulation techniques.
2. Optimize the choice of the regularizing weight for simple cases ($C = I_n$, $L$ with trivial second derivative). Provide more keys of interpretation and progressively increase the complexity of problem.
3. Extend the validity of these formulas to general regularization function $\rho$ (the theoretical approach relies on a fixed point equation which is naturally inferred from the identity $\nabla\|\beta\|^2 = 2\beta$, but there exists an equivalent minimization formulation which should adapt smoothly to general regularization functions).
4. Design an optimized algorithm that takes advantage of these theoretical conclusions.
5. Explore the extension of this approach to Kernel regression.

The proof of the uniqueness and existence of the solution to the fixed point problem is still lacking some justification. Our intuition is that we lack an important insight on the problem to be able to end the proof. That would be of course a great achievement of the internship if it is identified.

## References


[Kar13] Nourreddine El Karoui. "Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators : rigorous results". In: *arXiv preprint arXiv:1311.2445* (2013).

[DM16] David Donoho and Andrea Montanari. "High dimensional robust m-estimation: Asymptotic variance via approximate message passing". In: *Probability Theory and Related Fields* 166.3 (2016), pp. 935–969.

[ML19] Xiaoyi Mai and Zhenyu Liao. "High dimensional classification via regularized and unregularized empirical risk minimization: Precise error and optimal loss". In: *arXiv preprint arXiv:1905.13742* (2019).

[SC19] Pragya Sur and Emmanuel J Candès. "A modern maximum-likelihood theory for high-dimensional logistic regression". In: *Proceedings of the National Academy of Sciences* 116.29 (2019), pp. 14516–14525.

[Mig+20a] Francesca Mignacco et al. "Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9540–9550.

[Mig+20b] Francesca Mignacco et al. "The role of regularization in classification of high-dimensional noisy Gaussian mixture". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6874–6883.

[TOS20] Christos Thrampoulidis, Samet Oymak, and Mahdi Soltanolkotabi. "Theoretical insights into multiclass classification: A high-dimensional asymptotic view". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 8907–8920.

[Sed+21] Mohamed El Amine Seddik et al. "The Unexpected Deterministic and Universal Behavior of Large Softmax Classifiers". In: *AISTATS* (2021).

[DKT22] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. "A model of double descent for high-dimensional binary linear classification". In: *Information and Inference: A Journal of the IMA* 11.2 (2022), pp. 435–495.

[SZ22] Luca Saglietti and Lenka Zdeborová. "Solvable model for inheriting the regularization through knowledge distillation". In: *Mathematical and Scientific Machine Learning*. PMLR. 2022, pp. 809–846.

[Tio+22] Malik Tiomoko et al. "Deciphering Lasso-based Classification Through a Large Dimensional Analysis of the Iterative Soft-Thresholding Algorithm". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 21449–21477.