University Grenoble-Alpes **GIPSA-lab, GSTATS Data Science Chair**
INTERNSHIP PROJECT **Romain COUILLET & Nicolas KERIVEN**

2021 **Office D1178**

Towards "Digital Sobriety": A Random Matrix Analysis of On-Line Learning

Project context and application

With the recent compelling showcases of deep learning in performing advanced classfication tasks, deep neural networks are given more and more attention these days. However, deep learning is costly, its performances badly understood and its stability guarantees almost inexistant. Random matrix theory has recently managed to tip the scales in demonstrating that many machine learning algorithms (some of them quite standard) in fact suffer from their being ill-used when dealing with large dimensional realistic data. Correction measures have already been envisioned which sometimes largely improve over existing techniques, in passing providing strong theoretical guarantees [1-4].

However, for large dimensional and numerous datasets, also these methods become practically costly, especially when they involve matrix inverses or SVD. A recent line of investigated has been launched which surprisingly demonstrates that "cheap and efficient" learning is in fact accessible if one discards a vast majority of the calculus involved in simple learning methods, without significantly impacting the performance (sometimes even with virtually no loss in performance) [5].

The subject of this internship is to go beyond [5] by merging the recent appealing concept of "data sketching" [6] to (supervised or unsupervised) learning: that is, assuming data are too expensive to store or handle, as data arrive in the learning pipeline, they are processed at a low computational cost before being discarded altogether. Using random matrix theory to track and fine-tune the performance of such cheap learning algorithms, the objective of the iternship is then to propose a new family of performance-guaranteed machine learning algorithms based on data sketching.

The internship may lead to a PhD position opening as of October 2021.

Main steps

- Review of the literature on sketching, spectral clustering, random matrix theory.
- Implementation of the main documented solutions.
- Theoretical analysis and development of improved solutions and guarantees using random matrix theory.

Associated domains: Random matrix theory, machine learning, time series, signal processing, neural networks.

Requirements: Good coding skill in Matlab or Python, knowledge of the basics of random matrix theory, good understanding of general machine learning concepts.

Location: The internship will take place at GIPSA-lab, University of Grenoble-Alpes, in the Grenoble area.

References

[1] M. Tiomoko, H. Tiomoko, R. Couillet, **"Deciphering and Optimizing Multi-Task Learning: a Random Matrix Approach"**, (submitted to) International Conference on Learning Representation (ICLR'21), 2021.
[2] R. Couillet, F. Benaych-Georges, **"Kernel Spectral Clustering of Large Dimensional Data"**, Electronic Journal of Statistics, vol. 10, no. 1, pp. 1393-1454, 2016.
[3] C. Louart, Z. Liao, R. Couillet, **"A Random Matrix Approach to Neural Networks"**, The Annals of Applied Probability, vol. 28, no. 2, pp. 1190-1248, 2018.
[4] Z. Liao, R. Couillet, **"A Large Dimensional Analysis of Least Squares Support Vector Machines"**, IEEE Transactions on Signal Processing, vol. 67, no.4, pp. 1065-1074, 2018.
[5] T. Zarrouk, R. Couillet, F. Chatelain, N. Le Bihan, **"Performance-Complexity Trade-Off in Large Dimensional Statistics"**, International Workshop on Machine Learning for Signal Processing (MLSP'20), Espoo, Finland, 2020.
[6] Keriven, N., Bourrier, A., Gribonval, R., & Pérez, P. (2018). **Sketching for large-scale learning of mixture models**. *Information and Inference: A Journal of the IMA*, *7*(3), 447-508.