University of Grenoble-Alpes
GIPSA-lab, MIAI LargeDATA chair

INTERNSHIP
2020 - 2021

Romain COUILLET, Steeve ZOZOR, Lorenzo DALL'AMICO
romain.couillet@gipsa-lab.grenoble-inp.fr

Performance Optima of Large Dimensional Machine Learning: A Random Matrix, Statistical Physics and Information Theory Analysis

Project context and application

Using a large dimensional analysis (as the number n and size p of the data increase at the same rate), random matrix theory has recently opened the possibility to analyze the performance of so far intractable machine learning algorithms, starting with kernel methods, spectral clustering, semi-supervised learning, neural networks [1], etc. The universality results, classical in random matrix theory, further demonstrate that the performances of many machine learning methods applied to realistic data models (such as intricate generative models) coincide with the performance achieved on mere Gaussian mixture models [2]. As a consequence, even simplistic Gaussian-based methods can be proved optimal.

The question of the optimality of machine learning algorithms (unsupervised clustering, semi-supervised learning, transfer learning, etc.) thus becomes extremely meaningful and, since sufficient under Gaussian model assumptions, amenable to theoretical analysis. Such a study is performed in [3] in the context of semi-supervised learning, exploiting recent advances in random matrix theory, statistical physics and information theory.

The objective of the internship is to exploit the standard tools and methods of the aforementioned three disciplines to analyze the performance of key methods and algorithms in machine learning, such as transfer learning, under sufficient data settings (that is, mostly Gaussian mixture models), and compare with tools recently developed in the context of large dimensional statistics (and random matrix theory) for machine learning.

The internship will be part of an important project on the Grenoble area (the MIAI artificial intelligence institute), will take place within the GIPSA-lab on the Grenoble university campus under the joint supervision of Romain Couillet (Professor, head of the MIAI LargeDATA chair, expert in random matrices and machine learning), Steeve Zozor (DR CNRS, expert in information theory) and Lorenzo Dall'Amico (PhD student, expert in statistical physics), and may conduct to a PhD thesis position as of October 2021.

Related domains

Large dimensional statistics, random matrix theory, information theory, statistical physics, concentration of measure theory, data processing, machine learning, graph theory.

Main steps

- Getting to grasp with random matrices for machine learning and the information theory and statistical physics tools notably exploited in [3].
- Developping exact asymptotic optima (or bounds) on the performance of some targeted basic ML problems (such as transfer learning).
- Simulating and confronting the performances achieved on synthetic data and real data, versus theoretical optima.

Requirements

Strong knowledge in probability/statistics and machine learning, ideally with additional basic knowledge in random matrix theory, statistical physics and/or information theory; good coding skills in either Python or Matlab.

Contact

Romain COUILLET ***Steeve ZOZOR***
http://romaincouillet.hebfree.org/ http://www.gipsa-lab.grenoble-inp.fr/~steeve.zozor/
romain.couillet@gipsa-lab.grenoble-inp.fr steeve.zozor@gipsa-lab.grenoble-inp.fr

References

[1] Couillet, Romain, and Florent Benaych-Georges. "Kernel spectral clustering of large dimensional data." *Electronic Journal of Statistics* 10.1 (2016): 1393-1454.
[2] Louart, Cosme, and Romain Couillet. "Concentration of Measure and Large Random Matrices with an application to Sample Covariance Matrices." *arXiv preprint arXiv:1805.08295* (2018).
[3] Marc Lelarge, and Léo Miolane, (2019). Asymptotic Bayes risk for Gaussian mixture in a semi-supervised setting. *arXiv preprint arXiv:1907.03792*.