Randomized linear algebra for large dimensional data

Advisors: Nicolas Tremblay[†], Simon Barthelmé[†], Pierre-Olivier Amblard[†](HdR).

[†]: GIPSA-Lab, Grenoble, France.

Email : `firstname.lastname@gipsa-lab.fr`

Motivation. Some of the core operations in linear algebra, including inversion, the SVD or the eigendecomposition, scale cubically in the dimension of the input. For large dimensional datasets, this is too slow, and much effort has gone into developing approximate faster methods. One recent direction has been to use Monte Carlo methods that sample the input matrix, and perform all expensive operations on the sampled matrix rather than the original input. This class of techniques is called "randomized Linear Algebra" (RLA, [1]). For example, when computing the SVD of a "wide" matrix $\mathbf{A}$, an RLA algorithm will sample only a few, well-chosen columns of $\mathbf{A}$, and compute the SVD on these columns only. Under certain conditions this leads to a large increase in performance with only a small estimation error on the singular vectors.

We have recently begun to explore another way of performing RLA, one that does not involve sampling the columns or rows of a matrix. Instead, we use links between linear algebra and graph theory. Specifically, we focus on matrices that can be viewed as graph Laplacians. In that case some linear-algebraic properties of the matrix reflect graph-theoretic properties on the corresponding graph, and vice-versa (that's the basis of spectral methods for finding communities in a graph, for example). Uniform Spanning Trees, and variants thereof, are objects of particular interest in this context. A spanning tree is a tree that connects all nodes in a graph. There can be many spanning trees in a graph, and Wilson's algorithm is a (fast) way of sampling one of these trees at random. Due to the matrix-tree theorem, there are many deep links between linear algebraic quantities and Uniform Spanning Trees. We have been able to exploit these links to construct a fast estimator of regularised inverse traces [2], i.e. quantities of the form $\mathrm{Tr}(q\mathbf{I} + \mathbf{A})^{-\mathbf{1}}$.

PhD description. The goal of the PhD is to further this work by estimating other linear-algebraic quantities. We have ways of constructing estimators for the eigenspectrum of a matrix, and ways of solving certain linear systems. The student will investigate the theoretical properties of these estimators, and seek to find efficient implementations. The student will apply the resulting methods to problems in graph signal processing and/or semi-supervised machine learning, where the properties of the underlying graph Laplacian are of great importance.

Organization. The position is held within the GIPSA-lab at the University Grenoble-Alpes, as part of the MIAI chair on "Large Dimensional Statistics for AI" (LargeDATA). The LargeDATA chair develops expertise in large dimensional statistics for AI, notably focusing on random matrix theory, statistical physics and graphs. The PhD position is located at Gipsa-lab. Besides the two co-supervisors, the other people involved in the project are: Pierre-Olivier Amblard (also at Gipsa), Luca Avena (Univ. of Leiden), and Alexandre Gaudillière (CNRS, Institut de Mathématiques, Marseilles). P-O Amblard has worked on a large variety of estimation problems and has extensive expertise in multivariate and higher-order statistics. L. Avena and A. Gaudillière are probability specialists, with a growing body of work on the statistical properties and applications of random spanning trees and forests [3].

Profile. The candidate should hold a MSc in statistical signal processing/applied mathematics.

References

[1] Mahoney. "Randomized algorithms for matrices and data", Found. and Trends in Mach. Lear., 2011.
[2] Barthelmé et al. "Estimating the inverse trace using random forests on graphs", arXiv 1905.02086, 2019.
[3] Gaudillière, and Avena. "Two Applications of Random Spanning Forests", J. of Theor. Proba., 2017.