

GDR MIA THEMATIC DAYS ON HIGH-DIMENSIONAL DATA ANALYSIS

October 21-22, 2021

FRUMAM St Charles, Marseille, FRANCE



Summary

The analysis of high-dimensional data sets is now ubiquitous in many applications and blends tools from probability theory, machine learning, geometry, graph theory, statistics and optimization. The objective of this two-day event is to gather researchers from different backgrounds to explore recent advances in this field and to stimulate discussions. Special attention will be given to young researchers who are encouraged to participate with an introductory lecture on the first day and dedicated funding for their travel and accommodation.

Format

The event is organized into two parts:

Thursday, October 21 Introductory lecture on “Measure concentration and statistics in high-dimension: an introduction” by Aurélien Garivier (ENS Lyon).

Friday, October 22 Research talks

Participants will be able to attend both physically and virtually.

Venue

The event will be held at:

Frumam, St Charles
Aix-Marseille Université
3, place Victor Hugo - MARSEILLE Cedex 03

(5min walk from the railway station St Charles)

Virtual participation will also be possible via Zoom (the link will be sent by email to participants).

Registration

Registration is free but *mandatory*. To register, please fill the form:

<https://framaforms.org/gdr-mia-thematic-day-on-high-dimensional-data-analysis-1626874509>.

Day 1 - Doctoral course

“Measure concentration and statistics in high-dimension: an introduction” by Aurélien Garivier (ENS Lyon).

This course is dedicated to students and any person willing to acquire a background in measure concentration.

From 9h to 17h30.

Day 2 - Research talks

Speakers

- Elsa Cazelles (CNRS IRIT, Toulouse)
- Lénaïc Chizat (CNRS, Orsay)
- Agnès Desolneux (CNRS Centre Borelli, Paris)
- Nicolas Keriven (CNRS GIPSA, Grenoble)
- Olga Mula (CEREMADE, Paris)
- Geneviève Robin (LaMME, Évry)
- François-Xavier Vialard (LIGM, Marne-la-Vallée)

Program

09h30 - 10h00	<i>Coffee and introduction</i>	
10h00 - 10h45	Agnès Desolneux	<i>A Wasserstein-type distance in the space of Gaussian mixture models</i>
10h45 - 11h15	FX Vialard	<i>TBD</i>
11h15 - 11h45	<i>Break</i>	
11h45 - 12h15	Elsa Cazelles	<i>A novel notion of barycenter for probability distributions based on optimal weak mass transport</i>
12h15 - 14h00	<i>Lunch</i>	
14h00 - 14h45	Lénaïc Chizat	<i>Analysis of gradient descent on wide two-layer ReLU neural networks</i>
14h45 - 15h15	Geneviève Robin	<i>Outliers detection in networks with missing links</i>
15h15 - 15h45	<i>Break</i>	
15h45 - 16h15	Nicolas Keriven	<i>Graph neural networks on large random graphs: convergence, stability, universality</i>
16h15 - 16h45	Olga Mula	<i>Model order reduction by interpolation in Wasserstein spaces</i>
16h45 - 17h00	<i>Conclusions</i>	

Abstracts

- **Elsa Cazelles** A novel notion of barycenter for probability distributions based on optimal weak mass transport
We introduce weak barycenters of a family of probability distributions, based on the recently developed notion of optimal weak transport of mass. We provide a theoretical analysis of this object and discuss its interpretation in the light of convex ordering between probability measures. In particular, we show that, rather than averaging in a geometric way the input distributions, as the Wasserstein barycenter based on classic optimal transport does, weak barycenters extract common geometric information shared by all the input distributions, encoded as a latent random variable that underlies all of them. We also provide an iterative algorithm to compute a weak barycenter for a finite family of input distributions, and a stochastic algorithm that computes them for arbitrary populations of laws. The latter approach is particularly well suited for the streaming setting, i.e., when distributions are observed sequentially.
- **Lénaïc Chizat** Analysis of gradient descent on wide two-layer ReLU neural networks
In this talk, we propose an analysis of gradient descent on wide two-layer ReLU neural networks that leads to sharp characterizations of the learned predictor. The main idea is to study the training dynamics when the width of the hidden layer goes to infinity, which is a Wasserstein gradient flow. While this dynamics evolves on a non-convex landscape, we show that for appropriate initializations, its limit, when it exists, is a global minimizer. We also study the implicit regularization of this algorithm when the objective is the unregularized logistic loss, which leads to a max-margin classifier in a certain functional space. We finally discuss what these results tell us about the generalization performance, and in particular how these models compare to kernel methods.
- **Agnès Desolneux** A Wasserstein-type distance in the space of Gaussian mixture models
In this talk, we introduce a Wasserstein-type distance on the set of Gaussian mixture models. This distance is defined by restricting the set of possible coupling measures in the optimal transport problem to Gaussian mixture models. We derive a very simple discrete formulation for this distance, which makes it suitable for high dimensional problems. We also study the corresponding multi-marginal and barycenter formulations. We show some properties of this Wasserstein-type distance, and we illustrate its practical use with some examples in image processing. This is a joint work with Julie Delon.
- **Nicolas Keriven** Graph neural networks on large random graphs: convergence, stability, universality
In this talk, we will discuss some theoretical properties of Graph Neural Networks (GNNs) on large graphs. Indeed, most existing analyses of GNNs are purely combinatorial and may fail to reflect their behavior regarding large-scale structures in graphs, such as communities of well-connected nodes or manifold structures. To address this, we assume that the graphs of interest are generated with classical models of random graphs. We first give non-asymptotic convergence bounds of GNNs toward “continuous” equivalents as the number of nodes grows. We then study their stability to small deformations of the underlying random graph model, a crucial property in traditional CNNs. Finally, we study their universality and approximation power, and show how some recent GNNs are more powerful than others.
- **Olga Mula** Model order reduction by interpolation in Wasserstein spaces
TBD
- **Geneviève Robin** Outliers detection in networks with missing links
Networks are powerful to analyse complex systems: agents are represented as nodes, and pair-

wise interactions between agents are recorded as edges between these nodes. An important field of application is sociology, where the recent development of online social networks offers unprecedented possibilities. In this example, estimating the underlying probabilities of interactions between agents is a key problem. However, in many examples existing models fail to describe networks containing a small number of outlier nodes with abnormal behaviour. For example, fake accounts created to spread fake news. In addition, real-life networks are likely to be incompletely observed, with missing links due to individual non-response or machine failures. Identifying outliers in the presence of missing links is therefore a crucial problem. In this work, we introduce a new algorithm to detect outliers in a network that simultaneously predicts the missing links. Under fairly general assumptions, our algorithm exactly detects the outliers, and achieves the best known error for the prediction of missing links with polynomial computation cost. We also illustrate the method numerically and present to applications in epidemiology and social network analysis.

- **François-Xavier Vialard** TBD
TBD