Data compression for constrained problems and clustering

College of Computing, University Mohammed 6 Polytechnic

May 13, 2025

1 Introduction

A large class of problems requires the ability to express an initial dataset in smaller dimensions to overcome the curse of dimensionality. For instance, in the biotechnology field, patient genomics or proteomics data have tens of thousands of features (genes or proteins), which make it extremely challenging to work with such a large number of features. Hence, dimensionality reduction/feature selection is crucial to efficiently work on such problems. In some settings, this data compression/reduction is subject to specific conditions or constraints that make it unfeasible to use standard methods such as forward or backward feature selection, PCA, POD, etc. In the genomics example, a dimensionality reduction would be more meaningful and interpretable if it can preserve the non-negativity of gene expression values. The aim of this PhD thesis is to explore alternative strategies that leverage a combination of some of the following fields : optimization, statistics, machine learning, and algebra to design algorithms that obey the data compression/reduction constraints and set efficacy/accuracy metrics and error bounds.

Such an endeavor can be beneficial for various fields of applications. It is initially motivated by a biotechnology application (genomics/proteomics data compression) and a contact mechanics application (Lagrange multiplier compression). However, other applications might be addressed as they arise.

Time-permitting, this thesis would also address the non-uniqueness of clustering solutions (e.g., spectral clustering as a case study) which can be considered as another form of data reduction.

2 Milestones

1. Get familiarized with the required technical fields and perform a literature review;
2. Work on the algorithm design;
3. Assess the accuracy levels of the algorithm(s).

3 Qualifications

Ideal candidates should have the following qualifications :

- Currently completing a masters' or engineering degree in applied mathematics, data science, operations research or a related field.
- A solid understanding of some of the following fields (preferably many): optimization, linear algebra, machine learning, dimensionality reduction, partial differential equations, scientific computing.
- Good knowledge and previous use of Python for academic projects or internships.
- A solid understanding of optimization is highly preferred.

4 How to apply

In your application, include a CV, your college transcripts, and a couple of potential references.