PhD project proposal:
Geometric Approximations of Massive Data, with Applications to Optimization.

Supervisor: Nabil H. Mustafa (mustafa@lipn.univ-paris13.fr)

The aim of this PhD project proposal is to study the construction of *sparse approximations* of geometric data, and their application to solving *geometric optimization* problems more efficiently. In particular, the goal is to

1. develop improved methods for **combinatorial partitions** of geometric data;
2. using these, construct **small-sized sketches**; and finally, to
3. use the sketches to design **effective algorithms** to optimization problems.

Adequation with themes at LIPN: This PhD subject falls within the scope of the combinatorics team (CALIN) at LIPN and two areas of the scientific program of MathSTIC (Axe 3: combinatorics and axe 1: optimisation). More precisely, the topic is the *sparse representation* of data using methods and tools from combinatorics and geometry. The study of sparsity has gained momentum recently in the field of combinatorics and especially in the team CALIN [2, 4, 3, 1].

The following text briefly outlines the background and some representative questions that will be studied.

⋆ ⋆ ⋆

I. Context: Geometric Data. Massive geometric data is increasingly common thanks to the proliferation of ubiquitous data-collecting devices: portable 3D scanners, remote sensing, mobile devices, GPS, medical imaging, and wireless sensor networks to name a few. Such data presents particularly vexing challenges for algorithmic processing, not only due to the fact that even algorithms that use super-linear time and space simply become unfeasible, but also because such data may not fit on individual machines and must be stored in a distributed system; or one has access to it only through a data stream.

II. Sketches of Data. One promising approach in dealing with this amount of data is the following. Given input data P and an approximation parameter ϵ, first construct a small-sized sketch S of P, then solve the problem on S, and finally extend this solution to a $(1+\epsilon)$-approximation to the original problem. This approach is possible only if there exists a set S such that

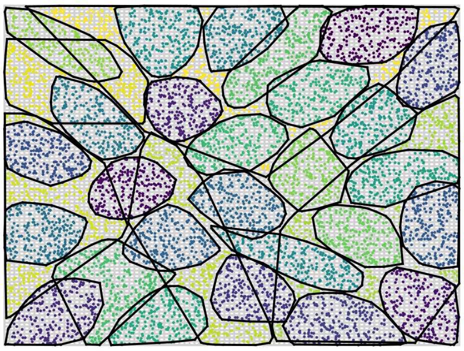
(i) the size of S is considerably less than that of P, and

(ii) an exact solution to S can be used to construct a good approximate solution to the problem on P.

The goal is to construct sketches whose size is independent of the size of the input data, while minimizing dependence on the approximation parameter ϵ. See the recent book by the supervisor of this project [4] on this topic.

A key component of the PhD work will be to understand the time and space complexity of constructing accurate sketches of data in high dimensions. Examples of some specific open questions include the use of shallow-cell complexity to improve bounds on relative (ρ, δ)-approximations, as well as the so-called sensitive ϵ-approximations. Further, the bounds in all known constructions of ϵ-approximations degrade substantially in the streaming and low-space settings.

III. Sparse Representations. A classical approach towards constructing small sketches is via random sampling. It has had great success due to its algorithmic simplicity, as well as its broad applicability (e.g., to general spaces with bounded VC dimension). However, such purely probabilistic approaches are unable to take full advantage of combinatorial and geometric properties present in data.



Indeed, it turns out that one can construct even better sketches by melding probabilistic methods with *structural partitioning* of the input data. Say, for example, that given a set P of n points in $\mathbb{R}^d$ and a parameter $\epsilon > 0$, we can partition the entire space $\mathbb{R}^d$ into $O\left(\frac{1}{\epsilon^d}\right)$ cells, such that each cell of this partition contains roughly the same number of points of P, *and*, any hyperplane intersects 'few' cells of this partition. Then it has been shown that this partition can be used to construct sketches of data which are provably better than purely random samples. See the right figure for a recent algorithm (Louvet *et al.*) that constructs such a partition.

Many basic questions here are open; for example, geometric partitions for objects other than point sets. See [2] for some recent progress in this area. Here is one potential question:

Problem: Design a $\tilde{O}_d\left(n\right)$ algorithm to compute simplicial partitions of points in $\mathbb{R}^d$ with respect to balls.

IV. Combinatorial Properties. Constructing sparse representations, in turn, requires one to understand the combinatorics of geometric configurations. For example, classical quantitative bounds on the sizes of random samples rely on a combinatorial property called VC dimension. Hence, bounds on the VC dimension have been studied in several communities, sometimes unrelated to geometry.

Despite all this work, many fundamental questions remain open. The PhD subject includes work on bounds for more general set-theoretic functions of geometric configurations. For example, even the basic question of the VC dimension of unions of hyperplanes in $\mathbb{R}^d$ remains unresolved.

V. Applications in Optimization. Finally the PhD work will turn to applications to various basic questions in optimization. Here is a partial list of the possible applications.

Coresets. Current constructions of coresets, besides being non-optimal, also required computationally intensive methods. Many key problems remain open and will be studied in this project. Examples include the existence and construction of coresets for the projective clustering problem (current bounds on *exponential* in the dimension d).

Low-space optimization. Very recently, sampling together with multiplicative weights update technique have been used to design algorithms that use low-space. This paradigm has already been successful for certain learning problems [5]. It is a natural question to extend this study to design low-space algorithms for problems in combinatorial optimization.

Deterministic linear programming. The current-best deterministic algorithms for linear programming (Chan, 2018) use ϵ-nets, together with the multiplicative weights technique. This PhD topic can thus lead to improved algorithms and bounds for spatial partitioning towards algorithms for deterministic linear programming.

Integer optimization. It is known that integer programs can be reduced to shortest paths in an appropriately defined graph on the integer lattice. Eisenbrand and Weismantel, 2018 proved that in fact this graph only needs to be constructed on the integer points within a distance of $O(d)$ to the origin. Unfortunately, the number of points within this ball still depends exponentially on the dimension d. It remains to explore the extension of the LP machinery developed (via ϵ-nets) for the linear case to the integer case.

References

[1] Jean-Philippe Chancelier, Michel de Lara, Antoine Deza, and Lionel Pournin. The Geometry of Sparse Optimization. *In preparation*, 2024.

[2] Monika Csikos and Nabil H. Mustafa. An Optimal Sparsification Lemma for Low-Crossing Matchings and its Applications to Discrepancy and Approximations. *ICALP*, 2024.

[3] Antoine Deza, Jean-Baptiste Hiriart-Urruty, and Lionel Pournin. Polytopal Balls Arising in Optimization. *Contributions to Discrete Mathematics* **16**(3), 125–138, 2021.

[4] Nabil H. Mustafa. Sampling in Combinatorial and Geometric Set Systems. *American Mathematical Society (AMS)*, 2022.

[5] Binghui Peng and Aviad Rubinstein. Near Optimal Memory-Regret Tradeoff for Online Learning. *FOCS*, 2023.