Probabilistic tools for high-dimensional geometric inference, topological data analysis and large-scale networks

PI: dr. Kunal Dutta

In the present age of "big data", the collection and storage of vast amounts of data, in nearly every field of science as well as industry, ranging from the internet and social media, to bio-genetics, astrophysics, and the social sciences, as well as VLSI design and many other areas, has become an everyday reality. The collected data, besides often coming in quantities that were unimaginable only a few decades ago, is also very often high-dimensional, that is, the number of data parameters can be of the order of hundreds, thousands or even millions. With this influx of data, comes a growing need to develop automated techniques and algorithms for the analysis, visualization and interpretation of data. Such analysis had so far often relied on human expertise, together with some statistical tools and algorithms. However, the sheer volume of the data, together with high dimensionality, forces us to develop new tools and techniques that are better geared to face large volumes and work as well, or even better, in high dimensions than they do in low dimensions. Such tools could aid human experts in their analysis and interpretation, or even be used as standalone data analysis packages.

In the area of *geometric inference*, the data points are represented in a geometric setting, say as a point cloud in a geometric space, and the aim is to utilize the geometry of the space, together with the *intrinsic* geometry of the data, to make inferences. For instance, one would like to be able to take in the point cloud in 3-D shown in Figure 1 as input, and conclude that it represents a double-torus (∞). In the closely allied area of *topological data analysis* (TDA), one aims to construct nested topological structures called *filtrations of simplicial complexes* from the data points, and then compute their *persistent homology*, from which topological information about the data can then be extracted. Both these areas utilize a variety of advanced mathematical techniques, and have enjoyed a fair bit of success over the past decade in several different fields of application. A third area is the analysis of large-scale networks, which has also seen increased interest due to the arrival of networks such as the world-wide web, social media networks, etc. which have billions of nodes and trillions of connections.

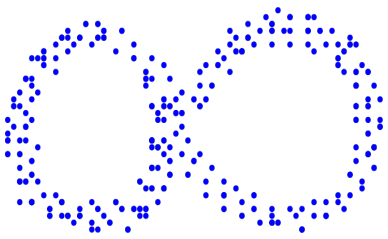

Figure 1: Point Cloud in 3 dimensions

The aims of this project are twofold. First, to bring together ideas and sophisticated mathematical tools from areas like probabilistic analysis, combinatorics, discrete and differential geometry and even geometric functional analysis, to develop better techniques and more efficient as well as practical algorithms, with mathematically proven guarantees, for geometric inference and toplogical data analysis. For example, many current algorithms used in TDA have an exponential or worse dependence of their running time, on the ambient dimension - a phenomenon often referred to as the *curse of dimensionality*. These algorithms are rendered almost impossible to use when the data has dimensionality in the order of hundreds or thousands, as can be the case in many potential applications of TDA. One of our challenges is to show the applicability of dimension reduction techniques such as random projections, to the computation of the persistent homology. A second problem is to analyze existing algorithms, or design new ones, and reduce their dependence on the ambient dimension by using randomization. Another problem relates to developing and extending the existing VC (Vapnik-Chervonenkis) dimension theory - which seeks to capture certain aspects of machine learning and statistical inference - and adapting it to the TDA setting.

Second, we plan to study the geometric and topological properties of the data itself, *on average*, by mathematically studying random models of the data, again using probabilistic tools. Random models of data, such as random networks, simplicial complexes, or points on random polytopes or manifolds, can provide important insights into the nature of the data, and therefore, the feasibility and effectiveness of techniques proposed for TDA. One problem relates to studying the recently proposed *strong collapse* procedure for reducing a dataset, on random simplicial complexes. Another problem is to study the spread of percolation processes, which seek to probabilistically model the spread of information, diseases, or rumors along a random or deterministic simplicial complex.

We thus expect to develop new algorithmic techniques for, as well as improve our understanding of, high-dimensional point cloud datasets.