Algorithmic Foundations of Geometry Understanding in Higher Dimensions (GUDHI)

Jean-Daniel Boissonnat

January 23, 2012

Abstract

High-dimensional spaces and shapes are ubiquitous in science. Let us mention the 4-dimensional *space-time*, the 6-dimensional *phase space* in particle physics, *configuration spaces* of mechanical systems, *conformational spaces* of macromolecules. Although their needs may be very different, numerical simulations, motion planning, molecular docking are among celebrated applications where approximating and processing high-dimensional objects is a central issue.

A maybe less intuitive place where one needs to understand high-dimensional geometry is in *data analysis.* Data analysis can be turned into a geometric problem by encoding heteregeneous data as clouds of points in high dimensional spaces equiped with some appropriate metric. Usually those points are not distributed in the whole embedding space but, due to the very nature of the system that produced those data, lie close to some subset of much smaller intrinsic dimension. Hence, the data conveys some geometric structure whose extraction and analysis are key to understanding the underlying system. Since data are produced at an unprecedented rate in all sciences, understanding the geometry of high-dimensional point clouds has become a core task in science and engineering.

The last decade has seen tremendous progress in processing high-dimensional shapes and spaces in various fields like robotics, machine learning and computational geometry. Computational topology emerged as a new discipline to provide solid foundations for the effective understanding of complex shapes. Although of a fundamental nature, these advances attracted interest in several fields like data analysis, computer vision or sensor networks. However, the current theory has not demonstrated its scalability to real problems and, up to now, it has only been applied to rather simple cases and in low dimensions. This is due to the lack of satisfactory algorithmic solutions in high-dimensional spaces. Breaking the computational bottleneck is now the main issue. Settling the *algorithmic foundations* of geometry understanding in higher dimensions is a grand challenge of great theoretical and practical significance.

The tenet of this proposal is that, to take up the challenge, we need a global approach involving tight and long-standing interactions between mathematical developments, algorithmic design and advanced programming. We believe that this is key to obtaining methods with built in robustness, scalability and guarantees, and therefore impact in the long run. By following this paradigm, we want to give to geometry understanding in higher dimensional an effective theory and a reference software platform.

TODO.

1. Expand the abstract
2. Check the references : the best ones ? others ?
3. Positionning wrt Machine Learning ?
4. List precise problems and open questions
5. List of WP.
6. An application?

A cultural note. A gudhi is found hanging out of a window or otherwise prominently displayed in traditional Maharashtrian households. Gudhi is a bright green or yellow cloth adorned with brocade (zari) tied to the tip of a long bamboo over which gathi (sugar crystals), neem leaves[citation needed], a twig of mango leaves and a garland of red flowers is tied. A silver or copper pot is placed in the inverted position over it. This gudhi is then hoisted outside the house, in a window, terrace or a high place so that everybody can see it.

Some of the significances attributed to raising a Gudhi are as follows:

– Gudhi symbolizes the Brahmadhvaj (translation: Brahmas flag) mentioned in the Brahma Purana, because Lord Brahma created the universe on this day. It may also represent Indradhvaj (translation: the flag of Indra).

– Mythologically, the Gudhi symbolizes Lord Ramas victory and happiness on returning to Ayodhya after slaying Ravan. Since a symbol of victory is always held high, so is the gudhi (flag). It is believed that this festival is celebrated to commemorate the coronation of Rama post his return to Ayodhya after completing 14 years of exile.

– Maharashtrians also see the Gudhi as a symbol of victory associated with the conquests of the Maratha forces led by Chhatrapati Shivaji. It also symbolizes the victory of King Shalivahana over Sakas and was hoisted by his people when he returned to Paithan. Gudhi is believed to ward off evil, invite prosperity and good luck into the house.

1 Extended synopsis of the project

The need to understand higher-dimensional spaces and shapes is ubiquitous in science. Physicists are used to combine space and time into a single 4-dimensional *space-time* continuum. In particle physics, the *phase space* consists of all possible values of position and momentum variables and is 6-dimensional. *Configuration spaces* of mechanical systems, *conformational spaces* of macromolecules are other examples of common high-dimensional geometric objects. Although their needs may be very different, numerical simulations, motion planning, molecular docking are among celebrated applications where approximating and processing high-dimensional objects is a central issue.

A maybe less intuitive place where one needs to understand high-dimensional geometry is in *data analysis*. Natural and artificial systems like biological or sensor networks are often described by a large number of real parameters, whereas a collection of text documents can be represented as a set of term frequency vectors in Euclidean space; similar interpretations can be given for image and video data [16]. Data analysis can then be turned into a geometric problem by encoding those heteregeneous data as clouds of points in high dimensional spaces equiped with some appropriate metric. Usually those points are not distributed in the whole embedding space but, due to the very nature of the system that produced those data, lie close to some subset of much smaller intrinsic dimension. Hence, the data conveys some geometric structure whose extraction and analysis are key to understanding the underlying system. Since data are produced at an unprecedented rate in all sciences, understanding the geometry of high-dimensional point clouds has become a core task in science and engineering.

The curses of higher-dimensional geometry. High-dimensional geometries are much more difficult to process and analyse than 3D shapes. The dimensionality severely restricts our intuition and ability to visualize the data. Hence understanding higher dimensional shapes must rely on automated tools able to extract useful qualitative and quantitative information from the input high-dimensional data.

Moreover, the complexity of the data structures and of the algorithms used to approximate shapes rapidly grows as the dimensionality increases, which makes them intractable in high dimensions. This curse of *dimensionality* is exemplyfied by the size of one of the simplest representations of a point set, namely its convex hull, whose complexity depends exponentially on the dimension.

In addition, high-dimensional data often suffer from significant *defects*, including sparsity, noise, and outliers which make them much more difficult to process than 3D data routinely provided by scanning devices. This is particularly so in the case of biological data, such as high throughput data from microarray or other sources. Moreover, the structure and occurrence of geometric features in the data may depend on the *scale* at which it is considered, thus requiring the analysis process to be multiscale.

The emergence of computational topology. The last decade has seen tremendous progress in processing high-dimensional shapes and spaces. In robotics [13], randomized techniques have been proposed to capture the topology of configuration spaces and to search paths. In machine learning, a variety of techniques have been proposed to reduce the dimension of data, learn nonlinear manifolds and cluster data [15]. In computational geometry, new approaches have been proposed to solve basic problems like searching nearest neighbours, computing smallest enclosing ellipsoids or approximating convex sets. Most of these methods have limited guarantees and impose strong constraints on the dimension or the topology of the shapes they can successfully handle. Computational topology emerged as a new discipline to provide solid foundations for the effective understanding of complex shapes [12]. The concepts of ε-samples, restricted Delaunay triangulation, anisotropic meshes emerged together with the first efficient and provably correct algorithms for emblematic problems like mesh generation and shape reconstruction in 3-dimensions [4].

The attempt to extend these results to higher dimensions led to the development of beautiful pieces of theory with deep roots in various areas of mathematics like Riemannian geometry, geometric measure theory, differential and algebraic topology. Let us mention the emergence of a sampling theory of geometric objects and of geometric inference [7], and the groundbreaking invention and rapid growth of persistent homology [11]. Although of a fundamental nature, these advances attracted interest in several fields like data analysis, computer vision or sensor networks.

The computational bottleneck. However, the current theory has not demonstrated its scalability to real problems and, up to now, it has only been applied to rather simple cases and in low dimensions. This is due to the lack of satisfactory algorithmic solutions in high-dimensional spaces. Breaking the computational bottleneck is now the main issue. Settling the *algorithmic foundations* of geometry understanding in higher dimensions is a grand challenge of great theoretical and practical significance.

The tenet of this proposal is that, to take up the challenge, we need a global approach involving tight and long-standing interactions between mathematical developments, algorithmic design and advanced programming. We believe that this is key to obtaining methods with built in robustness, scalability and guarantees, and therefore impact in the long run. Such an approach has been successfully carried out in low dimensions with the development of the Computational Geometric Algorithms Library CGAL [17].

By following this paradigm, we want to give to geometry understanding in higher dimensional an effective theory and a reference software platform. We strongly believe that this ambitious objective is realistic and can be reached. To pave the way towards this goal, we have identified four main scientific challenges that will also correspond to the main worktasks of the project.

Scientific challenge 1 : Going beyond affine models and Euclidean spaces. In the last decades, a set of new geometric methods, known as manifold learning, have been developed with the intent of parametrizing nonlinear shapes embedded in high-dimensional spaces. Although widely used, those methods assume very restrictive hypotheses on the geometry of the manifolds sampled by the datapoints to ensure correctness. A different route, inspired by what has been done in 3-dimensions, consists in approximating highly nonlinear shapes by *simplicial complexes*, the analogue of triangulations in higher dimensions. Recent research has exhibited sampling conditions under which topological or geometric properties, or even a full approximation of the sampled shape can be recovered [14, 7, 1]. Still the sampling conditions are quite stringent, the simplicial complexes are huge objects and their construction may be problematic.

A fundamental issue is the choice of a metric which determines the type and quality of an approximation. The simple Euclidean distance in the ambient space, while easy to deal with, is often not the right choice. As already mentionned, most of the time, when working in high dimensional spaces, we are in fact interested by objets of much smaller intrinsic dimension. The intrinsic geometry on the objects provides the right framework. Computational intrinsic geometry has not been seriously tackled yet and a basic question like the existence of Delaunay triangulations on Riemannian manifolds is still open. This question is of utmost importance for anisotropic mesh generation and optimal approximations. Another important situation is when the data are not given as a point cloud in some Euclidean space, but rather as a matrix of pairwise distances (i.e. a discrete metric space). Although such data may not be sampled from geometric subsets of Riemannian manifolds, it may still carry some interesting topological structures that need to be understood.

In information theory, signal and image processing, other pseudo-distances such as Kullback-Leibler, Itakura-Saito or Bregman divergences are prefered. These divergences are usually not true distances (they may not be symmetric nor satisfy the triangular inequality) and it is necessary to revisit geometric data structures and algorithms in this context [3].

Scientific challenge 2 : Bypassing the curse of dimensionality. The complexities of many geometric algorithms and data structures grow exponentially with increasing dimension. This rules out most, even if not all, geometric algorithms developped in low dimensions. Hence, extending Computational Geometry in high dimensions cannot be done in a straightforward manner and one has to take advantage of additional structure of the problem or to restrict attention to approximation. This motivated a number of new algorithmic paradigms such as locality-sensitive hashing, smoothed analysis, intrinsic algorithms. In this project, we will address the curse of dimensionality by focusing on the inherent structure in the data which in some sense needs to be sparse or of low intrinsic dimension as well as by putting the emphasis on output-sensitive algorithms and average-case analysis [16]. First investigations led to very promissing results, such as the design of new simplicial complexes with good complexity and approximation algorithms that scale well with the dimension [6, 1].

An important feature is that, even if we go to approximate solutions, we do not want to sacrifice guarantees. This is mandatory since the behaviour of algorithms in high dimensions is much less intuitive and easy to predict than in small dimensions.

Scientific challenge 3 : Searching for stable models. When dealing with approximation and samples, one needs stability results to ensure that the quantities that are computed are good approximations of the real ones. This is especially true in higher-dimensions where data are usually corrupted by various types of noise. A number of groundbreaking new approaches appeared recently. *Topological persistence* was recently introduced as a powerful tool for the study of the topological invariants of sampled spaces [11].

To deal with *non-local noise* and outliers, another new paradigm for point cloud data analysis has emerged recently. Point clouds are no longer treated as mere compact sets but rather as empirical measures. A notion of distance to such measures has been defined and shown to be stable with respect to perturbations of the measure [8]. A big challenge is to find efficient algorithms in arbitrary dimensions to compute or approximate the topological structure of the sublevel-sets of the distance to a measure. Such algorithms would naturally find applications in topological inference in the presence of significant noise and outliers, but also in other less obvious contexts such as stable clustering.

Multiscale reconstruction is another novel approach [2]. Taking advantage of the ideas of persistence, the approach consists in building a one-parameter family of simplicial complexes approximating the input at various scales. Determining the topology and shape of the original object reduces to finding the stable sequences in the one-parameter family of complexes. Despite its nice features, multiscale reconstruction, in its current form, can only be applied to low-dimensional data sets in practice.

Scientific challenge 4 : Building up the reference platform for high-dimensional geometric modeling. Software development is a central issue in this project and we will devote a substantial part of the effort to build up a software platform that will provide easy access to efficient and reliable high-dimensional geometric algorithms in the form of a C++ library. We will follow the example of the Computational Geometry Algorithms Library (CGAL), one of the major success stories of computational geometry, by now the gold standard for low dimensional geometric computing [17]. Our project aims at extending this success story, combining efficiency with correctness guarantees to high-dimensional geometric computation.

The development of a reference platform for high-dimensional geometric modeling is dictated by three main motivations. *First*, the sofware platform will allow to experiment at a large scale, which is mandatory to design the right models and data structures. It will boost theoretical research in geometric modeling, computational geometry in high dimensions, and computational topology, leading towards a virtuous circle between theory and experimental research. This has proven to be of utmost importance when developing the CGAL library and will be even truer in high dimensional geometry.

Second, maintaining such a platform will help further effort and consolidation in the long run. Having a library with interoperable modules will allow to incrementally add more and more sophisticated tools based on solid foundations. This is consistent with our long-term vision and our conviction that it is only through such a long standing effort that true impact, both theoretical and applied, can be gained.

Third, the platform will serve as a unique tool to communicate with the computational geometry community and with researchers from other fields. In return, we will get feedback from practionners that will help shaping the platform.

Risks and feasibility of the project. Simultaneously pursuing basic research at the best international level and developing industrial strength software is not without risks. My personnal record as well as the record of the members of my research team Geometrica who will be involved in this project are strong indications of our ability to take up the challenge with good chances of success. Geometrica has been at the cutting edge of research in geometric data structures and algorithms [5, 4, 3], mesh generation, shape approximation [], geometric inference and computational topology [10, 9, 7, 2, 8]. Geometrica is also one of the leader teams of the CGAL project [17]. We were at the source of successfull developments in CGAL like interval arithmetics, triangulations (now integrated in the heart of Matlab) and meshing packages. We also took a prominent part in the animation of the CGAL project and community, and in the creation of the spinoff GeometryFactory. It can be argued that my research group Geometrica is the best team worldwide to take up this dual challenge and to make this project a success.

I will devote 70% of my time to this project, and I will dedicate all my expertise and efforts to conduct and supervise the research work. To this end, I will receive the precious help of 2 permanent researchers of the Geometrica team : Frédéric Chazal who is a leading researcher in geometric inference and computational topology and Mariette Yvinec who is an expert in geometric computing and a member of the CGAL Editorial Board. They will devote 20% of their time to this project to co-supervise with me the research and implementation work of the students, postdocs and engineers to be engaged in this project. Other members of Geometrica, not financially supported by this project, will also collaborate to the project.

We will benefit from our strong collaborations with the best groups in Europe and in the US. Research in computational geometry and topology is very active in Europe. The academic European community has been supported through several research projects by the European Commission leading to the successful development of CGAL over the years. The ICT Fet-Open project Computational Geometric Learning (CG-Learning) is closely related to this project. The focus, the timetable and the management though are different. The proposed project wants to take over the results of CG-Learning and to go beyong prototype developments.

We will also benefit from our long-standing collaborations in the USA with Stanford university (Pr. Guibas) and Ohio State university (Pr. Dey) on topics that are related to this project.

New horizons and opportunities. If successful, the project will put higher-dimensional geometric modeling on new theoretical and algorithmic ground. It will help strengthening a research team with a unique spectrum of expertise covering mathematics, algorithm design and software development. It will also provide an open platform with no equivalent in USA or Asia. We forsee the platform to become a catalyst for research in high-dimensional geometry inside and outside of the project. By implementing the most effective techniques in a reliable and scalable way, the platform will open the way to groundbreaking technological advances for applications as varied as numerical simulation, visualization, machine learning, computer vision, data analysis, robotics or molecular biology. We will keep close contacts with research groups working in those domains and leap on opportunities arising from our new results and tools. In return, we will get feedback from practitionners which will help shaping the theoretical models and the software platform. Based on our experience with CGAL, we will undertake a vigorous action towards code diffusion in applied domains.

References

[1] J.-D. Boissonnat and A. Ghosh. Manifold reconstruction using tangential Delaunay complexes. In *Proc. 26th Annual Symposium on Computational Geometry*, 2010.

[2] J.-D. Boissonnat, L. J. Guibas, and S. Oudot. Manifold reconstruction in arbitrary dimensions using witness complexes. *Discrete and Computational Geometry*, 42(1):37–70, 2009.

[3] J.-D. Boissonnat, F. Nielsen, and R. Nock. Bregman Voronoi diagrams. *Discrete and Computational Geometry*, 44(2), 2010.

[4] J.-D. Boissonnat and M. Teillaud, editors. *Effective Computational Geometry for Curves and Surfaces.* Springer-Verlag, 2006.

[5] J.-D. Boissonnat and M. Yvinec. *Algorithmic geometry.* Cambridge University Press, 1998.

[6] G. Carlsson and V. de Silva. Topological estimation using witness complexes. In *Symposium on Point-Based Graphics*, 2004.

[7] F. Chazal, D. Cohen-Steiner, and A. Lieutier. A sampling theory for compact sets in euclidean space. *Discrete Comput. Geom.*, 41(3):461–479, 2009.

[8] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for probability measures. *Journal on Foundations of Computational Mathematics*, 11(6):733–751, 2011.

[9] F. Chazal and S. Y. Oudot. Towards Persistence-Based Reconstruction in Euclidean Spaces. In *Proc. ACM Symp. on Computational Geometry*, pages 232–241, 2008.

[10] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete Comput. Geom.*, 37(1):103–120, 2007.

[11] H. Edelsbrunner and J. Harer. Persistent homology — a survey. In J. P. J. E. Goodman and R. Pollack, editors, *Surveys on Discrete and Computational Geometry. Twenty Years Later*, Contemporary Mathematics 453, pages 257–282. Amer. Math. Soc., 2008.

[12] H. Edelsbrunner and J. Harer. *Computational topology.* American Mathematical Society, 2010.

[13] S. M. LaValle. *Planning Algorithms.* Cambridge University Press, 2006.

[14] P. Niyogi, S. Smale, and S. Weinberger. Finding the Homology of Submanifolds with High Confidence from Random Samples. *Discrete and Computational Geometry*, 39(1):419–441, 2008.

[15] H. Samet. *Foundations of Multidimensional and Metric Data Structures.* Morgan Kaufmann, 2006.

[16] H. S. Seung and D. D. Lee. The manifold ways of perception. *Science*, 290:2268–2269, 2000.

[17] CGAL. Computational Geometry Algorithms Library. http://www.cgal.org.