RESEARCH CENTRES

Saclay - Île-de-France

Sophia Antipolis - Méditerranée

IN PARTNERSHIP WITH:

Université Paris-Saclay, CNRS

2021
ACTIVITY REPORT

Project-Team

DATASHAPE

Understanding the shape of data

IN COLLABORATION WITH: Laboratoire de mathématiques d'Orsay de l'Université de Paris-Sud (LMO)

DOMAIN

Algorithmics, Programming, Software and Architecture

THEME

Algorithmics, Computer Algebra and Cryptology

Contents

Project-Team DATASHAPE

Creation of the Project-Team: 2017 July 01

Keywords

Computer sciences and digital sciences

A3. – Data and knowledge

A3.4. – Machine learning and statistics

A7.1. – Algorithms

A8. – Mathematics of computing

A8.1. – Discrete mathematics, combinatorics

A8.3. – Geometry, Topology

A9. – Artificial intelligence

Other research topics and application domains

B1. – Life sciences

B2. – Health

B5. – Industry of the future

B9. – Society and Knowledge

B9.5. – Sciences

1 Team members, visitors, external collaborators

Research Scientists

- Frédéric Chazal [Team leader, Inria, Senior Researcher, HDR]
- David Cohen-Steiner [Team leader, Inria, Researcher]
- Jean-Daniel Boissonnat [Inria, Emeritus, HDR]
- Mathieu Carrière [Inria, Researcher]
- Florent Dewez [Inria, Starting Research Position, from May 2021 until Aug 2021]
- Marc Glisse [Inria, Researcher]
- Jisu Kim [Inria, Starting Research Position]
- Clément Maria [Inria, Researcher]
- Steve Oudot [Inria, Senior Researcher, HDR]

Faculty Members

- Gilles Blanchard [Univ Paris-Saclay, Professor]
- Blanche Buet [Univ Paris-Saclay, Associate Professor]
- Pierre Pansu [Univ Paris-Saclay, Professor]

Post-Doctoral Fellows

- Charles Arnal [Inria, from Nov 2021]
- Felix Hensel [Inria, from Oct 2021]
- Kristof Huszar [Inria]
- Theo Lacombe [Inria, until Aug 2021]
- Siddharth Pritam [Inria, until May 2021]

PhD Students

- Charly Boricaud [Univ Paris Saclay, from Oct 2021]
- Jeremie Capitao-Miniconi [Univ Paris-Saclay]
- Antoine Commaret [École Normale Supérieure de Paris, from Sep 2021]
- Alex Delalande [Inria]
- Vincent Divol [Univ Paris-Saclay, until Aug 2021]
- Bastien Dussap [Univ Paris-Saclay, from Oct 2021]
- Georg Grützner [Studienstiftung des Deutschen Volkes]
- Alexandre Jean Daniel Guerin [Sysnav, from Sep 2021]
- Olympio Hacquard [Univ Paris-Saclay]
- Etienne Lasalle [Univ Paris-Saclay]

- Vadim Lebovici [École Normale Supérieure de Paris]
- David Loiseaux [Inria, from Nov 2021]
- Daniel Perez [Paris Sciences et Lettres]
- Louis Pujol [Univ Paris-Saclay]
- Wojciech Reise [Inria]
- Owen Rouille [Inria]
- Christophe Vuong [Telecom ParisTech]

Technical Staff

- Aziz Ben Ammar [Inria, Engineer, from May 2021]
- Rudresh Mishra [Inria, Engineer, until Mar 2021]
- Hind Montassif [Inria, Engineer, from Apr 2021]
- Vincent Rouvreau [Inria, Engineer]

Interns and Apprentices

- Charly Boricaud [Inria, from Apr 2021 until Jul 2021]
- Antoine Commaret [École Normale Supérieure de Paris, until Mar 2021]
- Thibault De Surrel [Inria, from May 2021 until Aug 2021]
- Rayhan Ladjouze [Inria, from Apr 2021 until Aug 2021]
- David Loiseaux [Inria, from May 2021 until Oct 2021]
- Isaac Ren [Inria, from Mar 2021 until Aug 2021]

Administrative Assistants

- Laurence Fontana [Inria]
- Sophie Honnorat [Inria]

Visiting Scientist

- Eduardo Fonseca Mendes [Fundação Getulio Vargas, from Nov 2021]

External Collaborator

- Bertrand Michel [École centrale de Nantes]

2 Overall objectives

DataShape is a research project in Topological Data Analysis (TDA), a recent field whose aim is to uncover, understand and exploit the topological and geometric structure underlying complex and possibly high dimensional data. The overall objective of the DataShape project is to settle the mathematical, statistical and algorithmic foundations of TDA and to disseminate and promote our results in the data science community.

The approach of DataShape relies on the conviction that it is necessary to combine statistical, topological/geometric and computational approaches in a common framework, in order to face the challenges of TDA. Another conviction of DataShape is that TDA needs to be combined with other data science approaches and tools to lead to successful real applications. It is necessary for TDA challenges to be simultaneously addressed from the fundamental and applied sides.

The team members have actively contributed to the emergence of TDA during the last few years. The variety of expertise, going from fundamental mathematics to software development, and the strong interactions within our team as well as numerous well established international collaborations make our group one of the best to achieve these goals.

The expected output of DataShape is two-fold. First, we intend to set up and develop the mathematical, statistical and algorithmic foundations of Topological and Geometric Data Analysis. Second, we intend to pursue the development of the GUDHI platform, initiated by the team members and which is becoming a standard tool in TDA, in order to provide an efficient state-of-the-art toolbox for the understanding of the topology and geometry of data. The ultimate goal of DataShape is to develop and promote TDA as a new family of well-founded methods to uncover and exploit the geometry of data. This also includes the clarification of the position and complementarity of TDA with respect to other approaches and tools in data science. Our objective is also to provide practically efficient and flexible tools that could be used independently, complementarily or in combination with other classical data analysis and machine learning approaches.

3 Research program

3.1 Algorithmic aspects and new mathematical directions for topological and geometric data analysis

TDA requires to construct and manipulate appropriate representations of complex and high dimensional shapes. A major difficulty comes from the fact that the complexity of data structures and algorithms used to approximate shapes rapidly grows as the dimensionality increases, which makes them intractable in high dimensions. We focus our research on simplicial complexes which offer a convenient representation of general shapes and generalize graphs and triangulations. Our work includes the study of simplicial complexes with good approximation properties and the design of compact data structures to represent them.

In low dimensions, effective shape reconstruction techniques exist that can provide precise geometric approximations very efficiently and under reasonable sampling conditions. Extending those techniques to higher dimensions as is required in the context of TDA is problematic since almost all methods in low dimensions rely on the computation of a subdivision of the ambient space. A direct extension of those methods would immediately lead to algorithms whose complexities depend exponentially on the ambient dimension, which is prohibitive in most applications. A first direction to by-pass the curse of dimensionality is to develop algorithms whose complexities depend on the intrinsic dimension of the data (which most of the time is small although unknown) rather than on the dimension of the ambient space. Another direction is to resort to cruder approximations that only captures the homotopy type or the homology of the sampled shape. The recent theory of persistent homology provides a powerful and robust tool to study the homology of sampled spaces in a stable way.

3.2 Statistical aspects of topological and geometric data analysis

The wide variety of larger and larger available data - often corrupted by noise and outliers - requires to consider the statistical properties of their topological and geometric features and to propose new relevant statistical models for their study.

There exist various statistical and machine learning methods intending to uncover the geometric structure of data. Beyond manifold learning and dimensionality reduction approaches that generally do not allow to assert the relevance of the inferred topological and geometric features and are not well-suited for the analysis of complex topological structures, set estimation methods intend to estimate, from random samples, a set around which the data is concentrated. In these methods, that include support and manifold estimation, principal curves/manifolds and their various generalizations to name a few, the estimation problems are usually considered under losses, such as Hausdorff distance or symmetric difference, that are not sensitive to the topology of the estimated sets, preventing these tools to directly infer topological or geometric information.

Regarding purely topological features, the statistical estimation of homology or homotopy type of compact subsets of Euclidean spaces, has only been considered recently, most of the time under the quite restrictive assumption that the data are randomly sampled from smooth manifolds.

In a more general setting, with the emergence of new geometric inference tools based on the study of distance functions and algebraic topology tools such as persistent homology, computational topology has recently seen an important development offering a new set of methods to infer relevant topological and geometric features of data sampled in general metric spaces. The use of these tools remains widely heuristic and until recently there were only a few preliminary results establishing connections between geometric inference, persistent homology and statistics. However, this direction has attracted a lot of attention over the last three years. In particular, stability properties and new representations of persistent homology information have led to very promising results to which the DATASHAPE members have significantly contributed. These preliminary results open many perspectives and research directions that need to be explored.

Our goal is to build on our first statistical results in TDA to develop the mathematical foundations of Statistical Topological and Geometric Data Analysis. Combined with the other objectives, our ultimate goal is to provide a well-founded and effective statistical toolbox for the understanding of topology and geometry of data.

3.3 Topological and geometric approaches for machine learning

This objective is driven by the problems raised by the use of topological and geometric approaches in machine learning. The goal is both to use our techniques to better understand the role of topological and geometric structures in machine learning problems and to apply our TDA tools to develop specialized topological approaches to be used in combination with other machine learning methods.

3.4 Experimental research and software development

We develop a high quality open source software platform called GUDHI which is becoming a reference in geometric and topological data analysis in high dimensions. The goal is not to provide code tailored to the numerous potential applications but rather to provide the central data structures and algorithms that underlie applications in geometric and topological data analysis.

The development of the GUDHI platform also serves to benchmark and optimize new algorithmic solutions resulting from our theoretical work. Such development necessitates a whole line of research on software architecture and interface design, heuristics and fine-tuning optimization, robustness and arithmetic issues, and visualization. We aim at providing a full programming environment following the same recipes that made up the success story of the CGAL library, the reference library in computational geometry.

Some of the algorithms implemented on the platform will also be interfaced to other software platforms, such as the R software for statistical computing, and languages such as Python in order to make them usable in combination with other data analysis and machine learning tools. A first attempt in this direction has been done with the creation of an R package called TDA in collaboration with the group of Larry Wasserman at Carnegie Mellon University (Inria Associated team CATS) that already includes some functionalities of the GUDHI library and implements some joint results between our team and the CMU team. A similar interface with the Python language is also considered a priority. To go even further towards helping users, we will provide utilities that perform the most common tasks without requiring any programming at all.

4 Application domains

Our work is mostly of a fundamental mathematical and algorithmic nature but finds a variety of applications in data analysis, e.g., in material science, biology, sensor networks, 3D shape analysis and processing, to name a few.

More specifically, DATASHAPE is working on the analysis of trajectories obtained from inertial sensors (PhD theses of Wojtek Riese Alexandre Guérin with Sysnav, participation to the DGA/ANR challenge MALIN with Sysnav) and, more generally on the development of new TDA methods for Machine Learning and Artificial Intelligence for (multivariate) time-dependent data from various kinds of sensors in collaboration with Fujitsu, or high dimensional point cloud data with MetaFora.

DATASHAPE is also working in collaboration with the University of Columbia in New-York, especially with the Rabadan lab, in order to improve bioinformatics methods and analyses for single cell genomic data. For instance, there is a lot of work whose aim is to use TDA tools such as persistent homology and the Mapper algorithm to characterize, quantify and study statistical significance of biological phenomena that occur in large scale single cell data sets. Such biological phenomena include, among others: the cell cycle, functional differentiation of stem cells, and immune system responses (such as the spatial response on the tissue location, and the genomic response with protein expression) to breast cancer.

5 Social and environmental responsibility

5.1 Footprint of research activities

The weekly research seminar of DATASHAPE is now taking place online, and travels for the team members have decreased a lot this year, mainly because of the COVID-19 pandemic.

6 New results

6.1 Algorithmic aspects and new mathematical directions for topological and geometric data analysis

6.1.1 Efficient open surface reconstruction from lexicographic optimal chains and critical bases

Participant: David Cohen-Steiner.

In collaboration with André Lieutier (Dassault Systèmes) and Julien Vuillamy (Dassault Systèmes and TITANE)

Previous works on lexicographic optimal chains have shown that they provide meaningful geometric homology representatives while being easier to compute than their l_1-norm optimal counterparts. This work [36] presents a novel algorithm to efficiently compute lexicographic optimal chains with a given boundary in a triangulation of 3-space, by leveraging Lefschetz duality and an augmented version of the disjoint-set data structure. Furthermore, by observing that lexicographic minimization is a linear operation, we define a canonical basis of lexicographic optimal chains, called critical basis, and show how to compute it. In applications, the presented algorithms offer new promising ways of efficiently reconstructing open surfaces in difficult acquisition scenarios.

6.1.2 Lower bound on the Voronoi diagram of lines in $\mathbb{R}^d$

Participant: Marc Glisse.

We give [41] a lower bound of $\Omega(n^{\lceil 2d/3 \rceil})$ on the maximal complexity of the Euclidean Voronoi diagram of n non-intersecting lines in $\mathbb{R}^d$ for $d > 2$.

6.1.3 ISDE : Independence Structure Density Estimation

Participant: Louis Pujol.

Density estimation appears as a subroutine in many learning procedures, so it is of crucial interest to have efficient methods to perform it in practical situations. Multidimensional density estimation faces the curse of dimensionality. To tackle this issue, a solution is to add a structural hypothesis through an undirected graphical model on the underlying distribution. We propose [52] ISDE (Independence Structure Density Estimation) an algorithm designed to estimate a density and an undirected graphical model from a special family of graphs corresponding to an independence structure, where features can be separated into independent groups. It is designed for moderately high dimensional data (up to 15 features) and it can be used in parametric as well as nonparametric situations. Existing methods on nonparametric graph estimation focus on multidimensional dependencies only through pairwise ones. ISDE does not suffer from this restriction and can addresses structures not covered yet by available algorithms. In this paper, we present existing theory about independence structure, explain the construction of our algorithm and prove its effectiveness on simulated data both quantitatively, through measures of density estimation performance under Kullback-Leibler loss and qualitatively, in terms of recovering of independence structures. We also provide information about running time.

6.1.4 Hybrid transforms of constructible functions.

Participant: Vadim Lebovici.

In [47] we introduce a general definition of hybrid transforms for constructible functions. These are integral transforms combining Lebesgue integration and Euler calculus. Lebesgue integration gives access to well-studied kernels and to regularity results, while Euler calculus conveys topological information and allows for compatibility with operations on constructible functions. We conduct a systematic study of such transforms and introduce two new ones: the Euler-Fourier and Euler-Laplace transforms. We show that the first has a left inverse and that the second provides a satisfactory generalization of Govc and Hepworth's persistent magnitude to constructible sheaves, in particular to multi-parameter persistent modules. Finally, we prove index-theoretic formulae expressing a wide class of hybrid transforms as generalized Euler integral transforms. This yields expectation formulae for transforms of constructible functions associated to (sub)level-sets persistence of random Gaussian filtrations.

6.1.5 Computing persistent Stiefel-Whitney classes of line bundles.

Participant: Raphael Tinarrage.

In this work [24] we propose a definition of persistent Stiefel-Whitney classes of vector bundle filtrations. It relies on seeing vector bundles as subsets of some Euclidean spaces. The usual Čech filtration of such a subset can be endowed with a vector bundle structure, that we call a Čech bundle filtration. We show that this construction is stable and consistent. When the dataset is a finite sample of a line bundle, we implement an effective algorithm to compute its persistent Stiefel-Whitney classes. In order to use simplicial approximation techniques in practice, we develop a notion of weak simplicial approximation. As a theoretical example, we give an in-depth study of the normal bundle of the circle, which reduces to understanding the persistent cohomology of the torus knot $(1,2)$.

6.1.6 Quantitative Stability of Optimal Transport Maps under Variations of the Target Measure.

Participant: Alex Delalande.

In collaboration with Quentin Mérigot (Laboratoire de Mathématiques d'Orsay, Univ. Paris-Saclay).

This work [38] studies the quantitative stability of the quadratic optimal transport map between a fixed probability density ρ and a probability measure μ on $\mathbb{R}^d$, which we denote T_μ. Assuming that the source density ρ is bounded from above and below on a compact convex set, we prove that the map $\mu \to T_\mu$ is bi-Hölder continuous on large families of probability measures, such as the set of probability measures whose moment of order $p > d$ is bounded by some constant. These stability estimates show that the linearized optimal transport metric $W_{2,\rho}(\mu,\nu) = \|T_\mu - T_\nu\|_{L_2(\rho,\mathbb{R}^d)}$ is bi-Hölder equivalent to the 2-Wasserstein distance on such sets, justifying its use in applications.

6.1.7 Nearly Tight Convergence Bounds for Semi-discrete Entropic Optimal Transport.

Participant: Alex Delalande.

We derive nearly tight and non-asymptotic convergence bounds for solutions of entropic semi-discrete optimal transport [37]. These bounds quantify the stability of the dual solutions of the regularized problem (sometimes called Sinkhorn potentials) w.r.t. the regularization parameter, for which we ensure a better than Lipschitz dependence. Such facts may be a first step towards a mathematical justification of annealing or ε-scaling heuristics for the numerical resolution of regularized semi-discrete optimal transport. Our results also entail a non-asymptotic and tight expansion of the difference between the entropic and the unregularized costs.

6.1.8 Zeta functions and the topology of super level-sets of stochasic processes

Participant: Daniel Perez.

In this preprint [49], we study bar-codes of random functions on the real line using the lens of zeta functions of continuous processes, that encompasses the $Pers_p$ functional already in use in topological data analysis. We give a closed formula for the zeta function of Brownian motion. For α-stable Lévy processes, the zeta function has a meromorphic extension with a unique simple pole at α. This implies a sharp asymptotic expansion for the expected number of small bars of the corresponding random bar-codes. We use this theoretical information to propose a new statistical test for the parameter α. A local version of the zeta function, counting bars beginning near a given value x, allows to analyse zeta functions of semi-martingales, and to give closed formulae in certain cases, like the Ornstein-Uhlenbeck process.

6.1.9 A uniformization theorem for closed convex polyhedra in Euclidean 3-space

Participant: Georg Grützner.

This work [42] deals with various aspects of Möbius geometry, including a discrete version. A new discrete uniformization theorem for 3-dimensional convex Euclidean polytopes is proved. Here, two polytopes are discretely conformal if certain cusped hyperbolic surfaces canonically associated to them are isometric. We then show that every convex polytope is discretely conformal to a polytope inscribed in a sphere, which is unique up to Möbius transformations between spheres.

6.2 Statistical aspects of topological and geometric data analysis

6.2.1 Topologically penalized regression on manifolds

Participants: Olympio Hacquard, Gilles Blanchard.

In collaboration with C. Levrard (LPSM, Sorbonne Université), K. Balasubramanian (UC. Davis), W. Polonik (UC. Davis).

In this preprint [43] we study a regression problem on a compact manifold $\mathscr{M}$. In order to take advantage of the underlying geometry and topology of the data, the regression task is performed on the basis of the first several eigenfunctions of the Laplace-Beltrami operator of the manifold, that are regularized with topological penalties. The proposed penalties are based on the topology of the sub-level sets of either the eigenfunctions or the estimated function. The overall approach is shown to yield promising and competitive performance on various applications to both synthetic and real data sets. We also provide theoretical guarantees on the regression function estimates, on both its prediction error and its smoothness (in a topological sense). Taken together, these results support the relevance of our approach in the case where the targeted function is "topologically smooth".

6.2.2 Minimax adaptive estimation in manifold inference.

Participant: Vincent Divol.

In this work [18], we focus on the problem of manifold estimation: given a set of observations sampled close to some unknown submanifold M , one wants to recover information about the geometry of M. Minimax estimators which have been proposed so far all depend crucially on the a priori knowledge of parameters quantifying the underlying distribution generating the sample (such as bounds on its density), whereas those quantities will be unknown in practice. Our contribution to the matter is twofold. First, we introduce a one-parameter family of manifold estimators (M_t), $t \geq 0$ based on a localized version of convex hulls, and show that for some choice of t, the corresponding estimator is minimax on the class of models of C^2 manifolds introduced in [Genovese et al., Manifold estimation and singular deconvolution under Hausdorff loss]. Second, we propose a completely data-driven selection procedure for the parameter t, leading to a minimax adaptive manifold estimator on this class of models. This selection procedure actually allows us to recover the Hausdorff distance between the set of observations and M , and can therefore be used as a scale parameter in other settings, such as tangent space estimation.

6.2.3 Estimation and Quantization of Expected Persistence Diagrams.

Participants: Vincent Divol, Théo Lacombe.

Persistence diagrams (PDs) are the most common descriptors used to encode the topology of structured data appearing in challenging learning tasks; think e.g. of graphs, time series or point clouds sampled close to a manifold. Given random objects and the corresponding distribution of PDs, one may want to build a statistical summary-such as a mean-of these random PDs, which is however not a trivial task as the natural geometry of the space of PDs is not linear. In this article [27], we study two such summaries, the Expected Persistence Diagram (EPD), and its quantization. The EPD is a measure supported on $\mathbb{R}^2$, which may be approximated by its empirical counterpart. We prove that this estimator is optimal from a minimax standpoint on a large class of models with a parametric rate of convergence. The empirical EPD is simple and efficient to compute, but possibly has a very large support, hindering its use in practice. To overcome this issue, we propose an algorithm to compute a quantization of the empirical EPD, a measure with small support which is shown to approximate with near-optimal rates a quantization of the theoretical EPD.

6.2.4 A short proof on the rate of convergence of the empirical measure for the Wasserstein distance.

Participant: Vincent Divol.

We provide a short proof that the Wasserstein distance between the empirical measure of a n-sample and the estimated measure is of order $n^{\frac{-1}{d}}$, if the measure has a lower and upper bounded density on the d-dimensional flat torus [39].

6.2.5 Reconstructing measures on manifolds: an optimal transport approach.

Participant: Vincent Divol.

Assume that we observe i.i.d. points lying close to some unknown d-dimensional C^k submanifold M in a possibly high-dimensional space. We study the problem of reconstructing the probability distribution generating the sample. After remarking that this problem is degenerate for a large class of standard losses (L_p , Hellinger, total variation, etc.), we focus on the Wasserstein loss, for which we build an estimator, based on kernel density estimation, whose rate of convergence depends on d and the regularity $s \leq k-1$ of the underlying density, but not on the ambient dimension [40]. In particular, we show that the estimator is minimax and matches previous rates in the literature in the case where the manifold M is a d-dimensional cube. The related problem of the estimation of the volume measure of M for the Wasserstein loss is also considered, for which a minimax estimator is exhibited.

6.2.6 Heat diffusion distance processes: a statistically founded method to analyze graph data sets.

Participant: Etienne Lasalle.

In [46], we propose two multiscale comparisons of graphs using heat diffusion, allowing to compare graphs without node correspondence or even with different sizes. These multiscale comparisons lead to the definition of Lipschitz-continuous empirical processes indexed by a real parameter. The statistical properties of empirical means of such processes are studied in the general case. Under mild assumptions, we prove a functional Central Limit Theorem, as well as a Gaussian approximation with a rate depending only on the sample size. Once applied to our processes, these results allow to analyze data sets of pairs of graphs. We design consistent confidence bands around empirical means and consistent two-sample tests, using bootstrap methods. Their performances are evaluated by simulations on synthetic data sets.

6.3 Topological and geometric approaches for machine learning

6.3.1 Optimal quantization of the mean measure and applications to statistical learning

Participants: Frédéric Chazal, Martin Royer.

In collaboration with C. Levrard (LPSM, Sorbonne Université.

This work [16] addresses the case where data come as point sets, or more generally as discrete measures. Our motivation is twofold: first we intend to approximate with a compactly supported measure the mean of the measure generating process, that coincides with the intensity measure in the point process framework, or with the expected persistence diagram in the framework of persistence-based topological data analysis. To this aim we provide two algorithms that we prove almost minimax optimal. Second we build from the estimator of the mean measure a vectorization map, that sends every measure into a finite-dimensional Euclidean space, and investigate its properties through a clustering-oriented lens. In a nutshell, we show that in a mixture of measure generating processes, our technique yields a representation in $\mathbb{R}^k$, for $k \in \mathbb{N}^*$ that guarantees a good clustering of the data points with high probability. Interestingly, our results apply in the framework of persistence-based shape classification via the ATOL procedure described in [31].

6.3.2 Optimizing persistent homology based functions.

Participants: Mathieu Carrière, Frédéric Chazal, Marc Glisse.

In collaboration with Yuichi Ike (Fujitsu) and Hariprasad Kannan.

Solving optimization tasks based on functions and losses with a topological flavor is a very active, growing field of research in data science and Topological Data Analysis, with applications in non-convex optimization, statistics and machine learning. However, the approaches proposed in the literature are usually anchored to a specific application and/or topological construction, and do not come with theoretical guarantees. To address this issue, we study the differentiability of a general map associated with the most common topological construction, that is, the persistence map [26]. Building on real analytic geometry arguments, we propose a general framework that allows us to define and compute gradients for persistence-based functions in a very simple way. We also provide a simple, explicit and sufficient condition for convergence of stochastic subgradient methods for such functions. This result encompasses all the constructions and applications of topological optimization in the literature. Finally, we provide associated code, that is easy to handle and to mix with other non-topological methods and constraints, as well as some experiments showcasing the versatility of our approach.

6.3.3 Barcode Embeddings for Metric Graphs.

Participant: Steve Oudot.

In collaboration with Elchanan Solomon (Department of Mathematics, Duke University).

Stable topological invariants are a cornerstone of persistence theory and applied topology, but their discriminative properties are often poorly-understood. In [23] we study a rich homology-based invariant first defined by Dey, Shi, and Wang, which we think of as embedding a metric graph in the barcode space. We prove that this invariant is locally injective on the space of metric graphs and globally injective on a GH-dense subset. Moreover, we show that is globally injective on a full measure subset of metric graphs, in the appropriate sense.

6.3.4 A Framework for Differential Calculus on Persistence Barcodes.

Participant: Steve Oudot.

In collaboration with Jacob Leygonie and Ulrike Tillmann (Mathematical Institute of Oxford).

In [20] we define notions of differentiability for maps from and to the space of persistence barcodes. Inspired by the theory of diffeological spaces, the proposed framework uses lifts to the space of ordered barcodes, from which derivatives can be computed. The two derived notions of differentiability (respectively from and to the space of barcodes) combine together naturally to produce a chain rule that enables the use of gradient descent for objective functions factoring through the space of barcodes. We illustrate the versatility of this framework by showing how it can be used to analyze the smoothness of various parametrized families of filtrations arising in topological data analysis.

6.3.5 A Gradient Sampling Algorithm for Stratified Maps with Applications to Topological Data Analysis.

Participants: Mathieu Carrière, Théo Lacombe, Steve Oudot.

In collaboration with Jacob Leygonie (Mathematical Institute of Oxford).

We introduce a novel gradient descent algorithm extending the well-known Gradient Sampling methodology to the class of stratifiably smooth objective functions, which are defined as locally Lipschitz functions that are smooth on some regular pieces-called strata-of the ambient Euclidean space [20]. For this class of functions, our algorithm achieves a sub-linear convergence rate. We then apply our method to objective functions based on the (extended) persistent homology map computed over lower-star filters, which is a central tool of Topological Data Analysis. For this, we propose an efficient exploration of the corresponding stratification by using the Cayley graph of the permutation group. Finally, we provide benchmark and novel topological optimization problems, in order to demonstrate the utility and applicability of our framework.

6.3.6 Topological Uncertainty: Monitoring trained neural networks through persistence of activation graphs.

Participants: Mathieu Carrière, Frédéric Chazal, Marc Glisse, Théo Lacombe.

In collaboration with Yuhei Umeda (Fujitsu) and Yuichi Ike (Fujitsu).

Although neural networks are capable of reaching astonishing performances on a wide variety of contexts, properly training networks on complicated tasks requires expertise and can be expensive from a computational perspective. In industrial applications, data coming from an open-world setting might widely differ from the benchmark datasets on which a network was trained. Being able to monitor the presence of such variations without retraining the network is of crucial importance. In this article [28], we develop a method to monitor trained neural networks based on the topological properties of their activation graphs. To each new observation, we assign a Topological Uncertainty, a score that aims to assess the reliability of the predictions by investigating the whole network instead of its final layer only, as typically done by practitioners. Our approach entirely works at a post-training level and does not require any assumption on the network architecture, optimization scheme, nor the use of data augmentation or auxiliary datasets; and can be faithfully applied on a large range of network architectures and data types. We showcase experimentally the potential of Topological Uncertainty in the context of trained network selection, Out-Of-Distribution detection, and shift-detection, both on synthetic and real datasets of images and graphs.

6.3.7 ATOL: Measure Vectorization for Automatic Topologically-Oriented Learning

Participants: Frédéric Chazal, Martin Royer.

In collaboration with with Yuhei Umeda (Fujitsu), Yuichi Ike (Fujitsu) and Clément Levrard (Univ. Paris Diderot).

Robust topological information commonly comes in the form of a set of persistence diagrams, finite measures that are in nature uneasy to affix to generic machine learning frameworks. We introduce [31] a fast, learnt, unsupervised vectorization method for measures in Euclidean spaces and use it for reflecting underlying changes in topological behaviour in machine learning contexts. The algorithm is simple and efficiently discriminates important space regions where meaningful differences to the mean measure arise. It is proven to be able to separate clusters of persistence diagrams. We showcase the strength and robustness of our approach on a number of applications, from emulous and modern graph collections where the method reaches state-of-the-art performance to a geometric synthetic dynamical orbits problem. The proposed methodology comes with a single high level tuning parameter: the total measure encoding budget. We provide a completely open access software.

6.4 Miscellaneous

6.4.1 Fast rates for prediction with limited expert advice

Participant: Gilles Blanchard.

In collaboration with El Mehdi Saad (LMO, University Paris-Saclay)

Motivated by the question of frugal machine learning, we investigate in this paper [32] the problem of minimizing the excess generalization error with respect to the best expert prediction in a finite family in the stochastic setting, under limited access to information: we assume that the learner only has access to a limited number of expert advices per training round, as well as for prediction. Under suitable assumptions on the loss (strong convexity and Lipschitz) We design novel algorithms achieving fast rates in this setting, and show that it is necessary to query at least two experts per training round to attain such fast rates.

6.4.2 Error rate control for classification rules in multiclass mixture models

Participant: Gilles Blanchard.

Collaboration with T.Mary-Huard (AgroParisTech), V. Perduca (MAP5, Univ. Paris), M.-L. Martin-Magniette (AgroParisTech)

In the context of finite mixture models, in this paper [22] we consider the problem of classifying as many observations as possible in the classes of interest while controlling the classification error rate in these same classes using an appropriate adaptation of the FDR criterion. It is shown that finding an optimal classification rule boils down to searching an optimal region in the observation space where to apply the classical Maximum A Posteriori (MAP) rule, and we propose a heuristic to compute that optimal region and the classification rule in practice. It is shown on both simulated and real datasets that the FDR-like optimal rule may be significantly less conservative than the naive thresholded MAP rule.

6.4.3 High-Dimensional Multi-Task Averaging and Application to Kernel Mean Embedding

Participant: Gilles Blanchard.

Collaboration with: H. Marienwald (U. Potsdam), J.-B. Fermanian (LMO, U. Paris-Saclay)

In the multi-task averaging problem, the goal is the joint estimation of the means of multiple distributions using separate, independent data sets. This is also known as the "many-means estimation problem" which has a long history in statistics. In this paper [30] we are particularly interested in the setting where the ambient space is of large dimension and propose a method exploiting similarities between tasks, without any related information being known in advance. The principle is that each empirical mean is shrunk towards the local average of its neighbors, which are found by multiple testing. We prove theoretically that this approach provides a reduction in mean squared error. This improvement can be significant when the dimension of the input space is large, demonstrating a "blessing of dimensionality" phenomenon. An application of this approach is the estimation of multiple kernel mean embeddings.

6.4.4 On agnostic post hoc approaches to false positive control

Participant: Gilles Blanchard.

Collaboration with: P. Neuvial (CNRS, U. Toulouse), E. Roquain (LPSM, U Sorbonne université)

This article [33] is a chapter in the *Handbook of Multiple Comparisons*. It gives a partial survey on post hoc approaches to false positive control for multiple test procedures.

6.4.5 Nonasymptotic one-and two-sample tests in high dimension with unknown covariance structure

Participant: Gilles Blanchard.

Collaboration with: J.-B. Fermanian (LMO, U. Paris-Saclay)

Article to appear in the *Festschrift in the honor of V. Spokoiny*

In this paper [34] we are interested in the test of so-called *relevant differences* for closeness of two mean vectors in a two-sample testing setting, i.e. the null hypothesis allows for a certain distance (tolerance) between the mean vectors, making it a composite null. We give matching upper and lower bounds giving in particular the precise dependence of the minimax test separation distance as a function of the effective dimensionality of each sample, and of the tolerance parameter allowed in the null hypothesis. We construct a test achieving optimal performance also when the covariance matrix of both samples (which determines their effective dimensionality) is unknown.

6.4.6 Identifying homogeneous subgroups of patients and important features: a topological machine learning approach.

Participants: Frédéric Chazal, Mathieu Carrière.

In collaboration with Ewan Carr and Raquel Iniesta (King's College London) and Bertrand Michel (Ecole Centrale de Nantes).

This work [15] exploits recent developments in topological data analysis to present a pipeline for clustering based on Mapper, an algorithm that reduces complex data into a one-dimensional graph. We present a pipeline to identify and summarise clusters based on statistically significant topological features from a point cloud using Mapper. Key strengths of this pipeline include the integration of prior knowledge to inform the clustering process and the selection of optimal clusters; the use of the bootstrap to restrict the search to robust topological features; the use of machine learning to inspect clusters; and the ability to incorporate mixed data types.

6.4.7 Topology identifies emerging adaptive mutations in SARS-CoV-2.

Participant: Mathieu Carrière.

In collaboration with Michael Bleher (Heidelberg University), Lukas Hahn (Heidelberg University), Juan Patiño-Galindo (Columbia University), Ulrich Bauer (TUM), Raúl Rabadán (Columbia University) and Andreas Ott (Heidelberg University).

The COVID-19 pandemic has lead to a worldwide effort to characterize its evolution through the mapping of mutations in the genome of the coronavirus SARS-CoV-2. As the virus spreads and evolves it acquires new mutations that could have important public health consequences, including higher transmissibility, morbidity, mortality, and immune evasion, among others. Ideally, we would like to quickly identify new mutations that could confer adaptive advantages to the evolving virus by leveraging the large number of SARS-CoV-2 genomes. One way of identifying adaptive mutations is by looking at convergent mutations, mutations in the same genomic position that occur independently. The large number of currently available genomes, more than a million at this moment, however precludes the efficient use of phylogeny-based techniques. In this work [35], we establish a fast and scalable Topological Data Analysis approach for the early warning and surveillance of emerging adaptive mutations of the coronavirus SARS-CoV-2 in the ongoing COVID-19 pandemic. Our method relies on a novel topological tool for the analysis of viral genome datasets based on persistent homology. It systematically identifies convergent events in viral evolution merely by their topological footprint and thus overcomes limitations of current phylogenetic inference techniques. This allows for an unbiased and rapid analysis of large viral datasets. We introduce a new topological measure for convergent evolution and apply it to the complete GISAID dataset as of February 2021, comprising 303,651 high-quality SARS-CoV-2 isolates taken from patients all over the world since the beginning of the pandemic. A complete list of mutations showing topological signals of convergence is compiled. We find that topologically salient mutations on the receptor-binding domain appear in several variants of concern and are linked with an increase in infectivity and immune escape. Moreover, for many adaptive mutations the topological signal precedes an increase in prevalence. We demonstrate the capability of our method to effectively identify emerging adaptive mutations at an early stage. By localizing topological signals in the dataset, we are able to extract geo-temporal information about the early occurrence of emerging adaptive mutations. The identification of these mutations can help to develop an alert system to monitor mutations of concern and guide experimentalists to focus the study of specific circulating variants.

6.4.8 An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists.

Participant: Frédéric Chazal.

In collaboration with Bertrand Michel (Ecole Centrale de Nantes).

Topological Data Analysis (TDA)is a recent and fast growing field providing a set of new topological and geometric tools to infer relevant features for possibly complex data. This paper [17] is a brief introduction, through a few selected topics, to basic fundamental and practical aspects of TDA for non experts.

7 Bilateral contracts and grants with industry

7.1 Bilateral contracts with industry

- **Participants:** Alexandre Guerin, Frédéric Chazal.

  Collaboration with Sysnav, a French SME with world leading expertise in navigation and geopositioning in extreme environments, on TDA, geometric approaches and machine learning for the analysis of movements of pedestrians and patients equipped with inetial sensors (CIFRE PhD of Alexandre Guérin).

- **Participants:** Felix Hensel, Theo Lacombe, Marc Glisse, Mathieu Carrière, Frédéric Chazal.

Research collaboration with Fujitsu on the development of new TDA methods and tools for Machine learning and Artificial Intelligence (started in Dec 2017).

- **Participants:** Louis Pujol, Bastien Dussap, Marc Glisse, Gilles Blanchard.

Research collaboration with MetaFora on the development of new TDA-based and statistical methods for the analysis of cytometric data (started in Nov. 2019).

7.2 Bilateral grants with industry

- DataShape and Sysnav have been selected for the ANR/DGA Challenge MALIN (funding: 700 kEuros) on pedestrian motion reconstruction in severe environments (without GPS access).

8 Partnerships and cooperations

8.1 International research visitors

8.1.1 Visits of international scientists

Other international visits to the team

Eduardo Fonseca Mendes

Status: Researcher

Institution of origin: Fundação Getulio Vargas, EMAp - School of Applied Mathematics

Country: Brasil

Dates: Nov. 10 - Dec. 12

Context of the visit: Initiation of collaboration.

Mobility program/type of mobility: Research stay

8.2 National initiatives

8.2.1 ANR

ANR ASPAG

Participants: Marc Glisse.

- Acronym : ASPAG.
- Type : ANR blanc.
- Title : Analysis and Probabilistic Simulations of Geometric Algorithms.
- Coordinator : Olivier Devillers (équipe Inria Gamble).
- Duration : 4 years from January 2018 to December 2021.

- Others Partners: Inria Gamble, LPSM, LABRI, Université de Rouen, IECL, Université du Littoral Côte d'Opale, Telecom ParisTech, Université Paris X (Modal'X), LAMA, Université de Poitiers, Université de Bourgogne.

- Abstract:

The analysis and processing of geometric data has become routine in a variety of human activities ranging from computer-aided design in manufacturing to the tracking of animal trajectories in ecology or geographic information systems in GPS navigation devices. Geometric algorithms and probabilistic geometric models are crucial to the treatment of all this geometric data, yet the current available knowledge is in various ways much too limited: many models are far from matching real data, and the analyses are not always relevant in practical contexts. One of the reasons for this state of affairs is that the breadth of expertise required is spread among different scientific communities (computational geometry, analysis of algorithms and stochastic geometry) that historically had very little interaction. The Aspag project brings together experts of these communities to address the problem of geometric data. We will more specifically work on the following three interdependent directions.

(1) Dependent point sets: One of the main issues of most models is the core assumption that the data points are independent and follow the same underlying distribution. Although this may be relevant in some contexts, the independence assumption is too strong for many applications.

(2) Simulation of geometric structures: The phenomena studied in (1) involve intricate random geometric structures subject to new models or constraints. A natural first step would be to build up our understanding and identify plausible conjectures through simulation. Perhaps surprisingly, the tools for an effective simulation of such complex geometric systems still need to be developed.

(3) Understanding geometric algorithms: the analysis of algorithm is an essential step in assessing the strengths and weaknesses of algorithmic principles, and is crucial to guide the choices made when designing a complex data processing pipeline. Any analysis must strike a balance between realism and tractability; the current analyses of many geometric algorithms are notoriously unrealistic. Aside from the purely scientific objectives, one of the main goals of Aspag is to bring the communities closer in the long term. As a consequence, the funding of the project is crucial to ensure that the members of the consortium will be able to interact on a very regular basis, a necessary condition for significant progress on the above challenges.

- See also: https://members.loria.fr/Olivier.Devillers/aspag/

ANR Chair in AI

Participants: Frédéric Chazal, Marc Glisse, Louis Pujol, Wojciech Riese.

- Acronym : TopAI
- Type : ANR Chair in AI.
- Title : Topological Data Analysis for Machine Learning and AI
- Coordinator : Frédéric Chazal
- Duration : 4 years from September 2020 to August 2024.
- Others Partners: Two industrial partners, the French SME Sysnav and the French start-up MetaFora.
- Abstract:

The TopAI project aims at developing a world-leading research activity on topological and geometric approaches in Machine Learning (ML) and AI with a double academic and industrial/societal objective. First, building on the strong expertise of the candidate and his team in TDA, TopAI aims at designing new mathematically well-founded topological and geometric methods and tools for Data Analysis and ML and to make them available to the data science and AI community through state-of-the-art software tools. Second, thanks to already established close collaborations and the strong involvement of French industrial partners, TopAI aims at exploiting its expertise and tools to address a set of challenging problems with high societal and economic impact in personalized medicine and AI-assisted medical diagnosis.

ANR ALGOKNOT

Participants: Clément Maria.

- Acronym : ALGOKNOT.
- Type : ANR Jeune Chercheuse Jeune Chercheur.
- Title : Algorithmic and Combinatorial Aspects of Knot Theory.
- Coordinator : Clément Maria.
- Duration : 2020 – 2023 (3 years).
- Abstract: The project AlgoKnot aims at strengthening our understanding of the computational and combinatorial complexity of the diverse facets of knot theory, as well as designing efficient algorithms and software to study their interconnections.
- See also: https://www-sop.inria.fr/members/Clement.Maria/

ANR GeMfaceT

Participants: Blanche Buet.

- Acronym: GeMfaceT.
- Type: ANR JCJC -CES 40 – Mathématiques
- Title: A bridge between Geometric Measure and Discrete Surface Theories
- Coordinator: Blanche Buet.
- Duration: 48 months, starting October 2021.
- Abstract: This project positions at the interface between geometric measure and discrete surface theories. There has recently been a growing interest in non-smooth structures, both from theoretical point of view, where singularities occur in famous optimization problems such as Plateau problem or geometric flows such as mean curvature flow, and applied point of view where complex high dimensional data are no longer assumed to lie on a smooth manifold but are more singular and allow crossings, tree-structures and dimension variations. We propose in this project to strengthen and expand the use of geometric measure concepts in discrete surface study and complex data modelling and also, to use those possible singular disrcete surfaces to compute numerical solutions to the aforementioned problems.

8.2.2 Collaboration with other national research institutes

SHOM

Participants: Steve Oudot.

Research collaboration between DataShape and the Service Hydrographique et Océanographique de la Marine (SHOM) on bathymetric data analysis using a combination of TDA and deep learning techniques. This collaboration is funded by the AMI IA *Améliorer la cartographie du littoral*.

IFPEN

Participants: Frédéric Chazal, Marc Glisse, Jisu Kim.

Research collaboration between DataShape and IFPEN on TDA applied to various problems issued from energy transition and sustainable mobility.

8.3 Regional initiatives

Metafora

Participants: Gilles Blanchard, Bastien Dussap, Marc Glisse, Louis Pujol.

- Type : Paris Region PhD² - PhD 2021.
- Title : Analyse de données cytométriques.

The Île-de-France region funds two PhD theses in collaboration with Metafora biosystems, a company specialized in the analysis of cells through their metabolism. The first is supervised by Pascal Massart (Inria team Celeste) and Marc Glisse, and its goal is to improve clustering for this particular type of data. The second one is supervised by Gilles Blanchard and Marc Glisse and aims to compare samples instead of analyzing just one sample.

9 Dissemination

Member of the organizing committees

- Frédéric Chazal is a member of the organizing committee of the Trimester “Geometry and Statistics in Data Sciences” at Institut Henri Poincaré.

Member of the conference program committees

- Gilles Blanchard was an area chair for the conference ICML 2021.
- David Cohen-Steiner was a program committee member for SGP 2021 and SMI 2021.
- Steve Oudot was a program committee member for the SIAM Conference on Applied Algebraic Geometry.

Member of the editorial boards

- Gilles Blanchard was member of the following journal editorial boards: Electronic Journal of Statistics, Bernoulli.
- Frédéric Chazal is a member of the following journal editorial boards: Discrete and Computational Geometry (Springer), Graphical Models (Elsevier).
- Frédéric Chazal is the Editor-in-Chief of the Journal of Applied and Computational Topology (Springer).
- Steve Oudot is a member of the editorial board of the Journal of Computational Geometry.

9.0.1 Invited talks

- Gilles Blanchard: invited talk at MFO workshop "Mathematical Foundations of Machine Learning" (march 2021)
- Gilles Blanchard: invited talk at the "German probability and statistics days" (september 2021)
- Blanche Buet: invited talk at Mathematics and Image Analysis MIA’21 conference (online, January 2021)
- Frédéric Chazal: invited lecture at the Waseda Cherry Blossom Workshop on Topological Data Science, Tokyo (March 2021).
- Frédéric Chazal: invited lecture at Rencontres Maths-Méca organised by SMAI and AFM at the Institut Henri Poincaré (Nov. 2021).

- Steve Oudot: invited talk at the workshop "Metrics in Multiparameter Persistence" hosted by the Lorenz Center (July 2021).
- Steve Oudot: invited talk at the "Applied Topology Seminar" held by the Applied Algebraic Topology Research Network (Sept. 2021).
- Steve Oudot: invited talk at the conference "TOUTELIA 2021 : Geometry, Topology and AI" organized by the Institut de Mathématiques de Toulouse (Sept. 2021)

9.0.2 Leadership within the scientific community

- Frédéric Chazal is the Director of the DATAIA Institute at Université Paris-Saclay (https://www.dataia.eu/).
- Steve Oudot is co-head (with Luca Castelli Aleardi) of the GT GeoAlgo within the GdR IM (https://www.gdr-im.fr/organisation/).

9.0.3 Research administration

- Marc Glisse is president of the CDT at Inria Saclay.
- Steve Oudot is president of the Commission Scientifique at Inria Saclay.

9.1 Teaching - Supervision - Juries

9.1.1 Teaching

- Master: Frédéric Chazal, Analyse Topologique des Données, 30h eq-TD, Université Paris-Sud, France.
- Master: Marc Glisse and Clément Maria, Computational Geometry Learning, 36h eq-TD, M2, MPRI, France.
- Master: Frédéric Cazals and Mathieu Carrière, Foundations of Geometric Methods in Data Analysis, 24h eq-TD, M2, École Centrale Paris, France.
- Master: Frédéric Chazal and Julien Tierny, Topological Data Analysis, 38h eq-TD, M2, Mathématiques, Vision, Apprentissage (MVA), ENS Paris-Saclay, France.
- Master: Steve Oudot, Topological data analysis, 45h eq-TD, M1, École polytechnique, France.
- Master: Steve Oudot, Data Analysis: geometry and topology in arbitrary dimensions, 24h eq-TD, M2, graduate program in Artificial Intelligence & Advanced Visual Computing, École polytechnique, France.
- Master: Gilles Blanchard, Mathematics for Artificial Intelligence 1, 70h eq-TD, IMO, Université Paris-Saclay, France.
- Master: Blanche Buet, TD-Distributions et analyse de Fourier, 60h eq-TD, M1, Université Paris-Saclay, France.
- Undergrad-Master: Steve Oudot, Algorithms for data analysis in C++, 22.5h eq-TD, L3/M1, École polytechnique, France.
- Undergrad: Marc Glisse, Mécanismes de la programmation orientée-objet, 40h eq-TD, L3, École Polytechnique, France.

9.1.2 Supervision

- PhD in progress: Vadim Lebovici, Laplace transform for constructible functions. Started September 2020. Steve Oudot and François Petit.
- PhD in progress: Christophe Vuong, Random hypergraphs. Started November 2020. Laurent Decreusefond and Marc Glisse.
- PhD in progress: Louis Pujol, Partitionnement de données cytométriques, started November 1st, 2019, Pascal Massart and Marc Glisse.
- PhD in progress: Bastien Dussap, Comparaison de données cytométriques, started October 1st, 2021, Gilles Blanchard and Marc Glisse.
- PhD in progress: Olympio Hacquard, Apprentissage statistique par méthodes topologiques et géométriques, 2020, Gilles Blanchard and Clément Levrard.
- PhD in progress: El Mehdi Saad, Efficient online methods for variable and model selection, started September 2019. Gilles Blanchard and Sylvain Arlot.
- PhD in progress: Hannah Marienwald, Transfer learning in high dimension. Started September 2019. Gilles Blanchard and Klaus-Robert Müller.
- PhD in progress: Jean-Baptiste Fermanian, Estimation de Kernel Mean Embedding et tests multiples en grande dimension. Started September 2021. Gilles Blanchard and Magalie Fromont-Renoir.
- PhD in progress: Antoine Commaret, Persistent Geometry. Started September 2021. David Cohen-Steiner and Indira Chatterji.
- PhD in progress: David Loiseaux, Multivariate topological data analysis for statistical machine learning. Started November 2021. Mathieu Carrière and Frédéric Cazals.
- PhD in progress: Wojciech Rieser, TDA for curve data. Started October 2020. Frédéric Chazal and Bertrand Michel.
- PhD in progress: Alexandre Guérin, Movement analysis from inertial sensors. Started on October 2021. Frédéric Chazal and Bertrand Michel.
- PhD in progress: Etienne Lasalle, TDA and statistics on graphs. Started on October 2019. Frédéric Chazal and Pascal Massart.
- PhD in progress: Alex Delande, optimal transport. Started on December 2019. Frédéric Chazal and Quentin Mérigot.
- PhD in progress: Jérémie Capitao-Miniconi, deconvolution for singular measures with geometric support. Started on October 2020. Frédéric Chazal and Elisabeth Gassiat.
- PhD in progress: Charly Boricaud, Geometric inference for Data analysis: a Geometric Measure Theory perspective. Started on October 2021. Blanche Buet, Gian Paolo Leonardi et Simon Masnou.

9.1.3 Juries

- Gilles Blanchard was a member for the following Ph.D. defenses: Batiste Le Bars (ENS Paris-Saclay), Fabrice Grela (Univ. Rennes), Timothée Mathieu (U. Paris-Saclay), Amandine Dubois (ENSAI), Tien-Dat Nguyen (U. Paris-Saclay).

10 Scientific production

10.1 Major publications

[1] D. Attali, U. Bauer, O. Devillers, M. Glisse and A. Lieutier. 'Homological Reconstruction and Simplification in R3'. In: *Computational Geometry* (2014). DOI: 10.1016/j.comgeo.2014.08.010. URL: https://hal.archives-ouvertes.fr/hal-01132440.

[2] J.-D. Boissonnat, R. Dyer and A. Ghosh. 'Delaunay Triangulation of Manifolds'. In: *Foundations of Computational Mathematics* 45 (2017), p. 38. DOI: 10.1007/s10208-017-9344-1. URL: https://hal.inria.fr/hal-01509888.

[3] J.-D. Boissonnat, R. Dyer, A. Ghosh and S. Y. Oudot. 'Only distances are required to reconstruct submanifolds'. In: *Computational Geometry* 66 (2017), pp. 32–67. DOI: 10.1016/j.comgeo.2017.08.001. URL: https://hal.inria.fr/hal-01583086.

[4] J.-D. Boissonnat, K. C. Srikanta and S. Tavenas. 'Building Efficient and Compact Data Structures for Simplicial Complexe'. In: *Algorithmica* (Sept. 2016). DOI: 10.1007/s00453-016-0207-y. URL: https://hal.inria.fr/hal-01364648.

[5] B. Buet, G. P. Leonardi and S. Masnou. 'A Varifold Approach to Surface Approximation'. In: *Archive for Rational Mechanics and Analysis* 226.2 (Nov. 2017), pp. 639–694. DOI: 10.1007/s00205-017-1141-0. URL: https://hal.archives-ouvertes.fr/hal-02141325.

[6] F. Chazal, D. Cohen-Steiner and A. Lieutier. 'A Sampling Theory for Compact Sets in Euclidean Space'. In: *Discrete Comput. Geom.* 41.3 (2009), pp. 461–479. URL: http://dx.doi.org/10.1007/s00454-009-9144-8.

[7] F. Chazal, D. Cohen-Steiner and Q. Mérigot. 'Geometric Inference for Measures based on Distance Functions'. Anglais. In: *Foundations of Computational Mathematics* 11.6 (2011). RR-6930, pp. 733–751. DOI: 10.1007/s10208-011-9098-0. URL: http://hal.inria.fr/inria-00383685.

[8] F. Chazal, S. Y. Oudot, M. Glisse and V. De Silva. *The Structure and Stability of Persistence Modules.* SpringerBriefs in Mathematics. Springer Verlag, 2016, pp. VII, 116. URL: https://hal.inria.fr/hal-01330678.

[9] L. J. Guibas, S. Y. Oudot, P. Skraba and F. Chazal. 'Persistence-Based Clustering in Riemannian Manifolds'. Anglais. In: *Journal of the ACM* 60.6 (Nov. 2013), p. 38. URL: http://hal.archives-ouvertes.fr/hal-00923563.

[10] M. Mandad, D. Cohen-Steiner, L. Kobbelt, P. Alliez and M. Desbrun. 'Variance-Minimizing Transport Plans for Inter-surface Mapping'. In: *ACM Transactions on Graphics* 36 (2017), p. 14. DOI: 10.1145/3072959.3073671. URL: https://hal.inria.fr/hal-01519006.

[11] S. Y. Oudot. *Persistence Theory: From Quiver Representations to Data Analysis.* Mathematical Surveys and Monographs 209. American Mathematical Society, 2015, p. 218. URL: https://hal.inria.fr/hal-01247501.

10.2 Publications of the year

International journals

[12] S. Arya, J.-D. Boissonnat, K. Dutta and M. Lotz. 'Dimensionality reduction for k-distance applied to persistent homology'. In: *Journal of Applied and Computational Topology* 5 (19th Oct. 2021), pp. 671–691. DOI: 10.1007/s41468-021-00079-x. URL: https://hal-amu.archives-ouvertes.fr/hal-03412594.

[13] G. Blanchard, A. A. Deshmukh, U. Dogan, G. Lee and C. Scott. 'Domain Generalization by Marginal Transfer Learning'. In: *Journal of Machine Learning Research* 22.2 (2021), pp. 1–55. URL: https://hal.archives-ouvertes.fr/hal-02974216.

[14] J.-D. Boissonnat, R. Dyer, A. Ghosh, A. Lieutier and M. Wintraecken. 'Local Conditions for Triangulating Submanifolds of Euclidean Space'. In: *Discrete and Computational Geometry* 66.2 (Sept. 2021), pp. 666–686. DOI: 10.1007/s00454-020-00233-9. URL: https://hal-amu.archives-ouvertes.fr/hal-03372073.

[15] E. Carr, M. Carriere, B. Michel, F. Chazal and R. Iniesta. 'Identifying homogeneous subgroups of patients and important features: a topological machine learning approach'. In: *BMC Bioinformatics* (2021). URL: https://hal.inria.fr/hal-03368489.

[16] F. Chazal, C. Levrard and M. Royer. 'Optimal quantization of the mean measure and applications to statistical learning'. In: *Electronic Journal of Statistics* 15.1 (Apr. 2021), pp. 2060–2104. URL: https://hal.archives-ouvertes.fr/hal-02465446.

[17] F. Chazal and B. Michel. 'An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists'. In: *Frontiers in Artificial Intelligence*. Front. Artif. Intell. (29th Sept. 2021). URL: https://hal.inria.fr/hal-01614384.

[18] V. Divol. 'Minimax adaptive estimation in manifold inference'. In: *Electronic Journal of Statistics* (27th Dec. 2021). URL: https://hal.inria.fr/hal-02440881.

[19] R. Gribonval, G. Blanchard, N. Keriven and Y. Traonmilin. 'Compressive Statistical Learning with Random Feature Moments'. In: *Mathematical Statistics and Learning* 3.2 (21st Aug. 2021), pp. 113–164. DOI: 10.4171/msl/20. URL: https://hal.inria.fr/hal-01544609.

[20] J. Leygonie, S. Y. Oudot and U. Tillmann. 'A Framework for Differential Calculus on Persistence Barcodes'. In: *Foundations of Computational Mathematics* (2021). URL: https://hal.archives-ouvertes.fr/hal-02304300.

[21] H. Luo, A. Patania, J. Kim and M. Vejdemo-Johansson. 'Generalized penalty for circular coordinate representation'. In: *Foundations of Data Science* 3.4 (2021), pp. 729–767. DOI: 10.3934/fods.2021024. URL: https://hal.inria.fr/hal-03501929.

[22] T. Mary-Huard, V. Perduca, M. L. Martin-Magniette and G. Blanchard. 'Error rate control for classification rules in multiclass mixture models'. In: *The International Journal of Biostatistics* (2021). URL: https://hal-universite-paris-saclay.archives-ouvertes.fr/hal-03357461.

[23] S. Oudot and E. Solomon. 'Barcode Embeddings for Metric Graphs'. In: *Algebraic and Geometric Topology* (2021). URL: https://hal.inria.fr/hal-01708780.

[24] R. Tinarrage. 'Computing persistent Stiefel-Whitney classes of line bundles'. In: *Journal of Applied and Computational Topology* (2nd Oct. 2021). URL: https://hal.archives-ouvertes.fr/hal-02619607.

International peer-reviewed conferences

[25] J.-D. Boissonnat, S. Kachanovich and M. Wintraecken. 'Tracing isomanifolds in R^d in time polynomial in d using Coxeter-Freudenthal-Kuhn triangulations'. In: SoCG 2021 - 37th Symposium on Computational Geometry. 37th International Symposium on Computational Geometry (SoCG 2021). Buffalo, United States, 7th June 2021. DOI: 10.4230/LIPICS.SOCG.2021.17. URL: https://hal.inria.fr/hal-03006663.

[26] M. Carriere, F. Chazal, M. Glisse, Y. Ike and H. Kannan. 'Optimizing persistent homology based functions'. In: 38th International Conference on Machine Learning (ICML) 2021. Vol. PMLR 139. Proceedings of the 38th International Conference on Machine Learning, ICML 2021. Virtual conference, France, 18th July 2021, pp. 1294–1303. URL: https://hal.inria.fr/hal-02969305.

[27] V. Divol and T. Lacombe. 'Estimation and Quantization of Expected Persistence Diagrams'. In: ICML - 38th International Conference on Machine Learning. Virtual Conference, United States, 18th July 2021. URL: https://hal.archives-ouvertes.fr/hal-03222657.

[28] T. Lacombe, Y. Ike, M. Carriere, F. Chazal, M. Glisse and Y. Umeda. 'Topological Uncertainty: Monitoring trained neural networks through persistence of activation graphs'. In: IJCAI 2021 - International Joint Conference on Artificial Intelligence. Montréal, Canada, 19th Aug. 2021. URL: https://hal.archives-ouvertes.fr/hal-03213188.

[29] C. Maria and O. Rouillé. 'Computation of Large Asymptotics of 3-Manifold Quantum Invariants'. In: ALENEX 21 - SIAM Symposium on Algorithm Engineering and Experiments. Alexandria / Virtual, United States, 10th Jan. 2021. URL: https://hal.archives-ouvertes.fr/hal-03133238.

[30] H. Marienwald, J.-B. Fermanian and G. Blanchard. 'High-Dimensional Multi-Task Averaging and Application to Kernel Mean Embedding'. In: *Proceedings of Machine Learning Research*. AISTATS 2021 - International Conference on Artificial Intelligence and Statistics. Vol. 130. Proceedings of Machine Learning Research. Virtual, United States, 2021, pp. 1963–1971. URL: https://hal.archives-ouvertes.fr/hal-03002342.

[31] M. Royer, F. Chazal, C. Levrard, Y. Umeda and Y. Ike. 'ATOL: Measure Vectorization for Automatic Topologically-Oriented Learning'. In: The 24th International Conference on Artificial Intelligence and Statistics (AISTATS 2021). The 24th International Conference on Artificial Intelligence and Statistics. Virtual conference, France, Apr. 2021. URL: https://hal.archives-ouvertes.fr/hal-02296513.

[32] E. M. Saad and G. Blanchard. 'Fast rates for prediction with limited expert advice'. In: NeurIPS. Advances in Neural Information Processing Systems 34. Online conference, United States, 2021. URL: https://hal.archives-ouvertes.fr/hal-03405899.

Scientific book chapters

[33] G. Blanchard, P. Neuvial and E. Roquain. 'On agnostic post hoc approaches to false positive control'. In: *Handbook of Multiple Comparisons*. Handbooks of Modern Statistical Methods. Chapman & Hall/CRC, 16th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-02320543.

Reports & preprints

[34] G. Blanchard and J.-B. Fermanian. *Nonasymptotic one-and two-sample tests in high dimension with unknown covariance structure*. 7th Oct. 2021. URL: https://hal-universite-paris-saclay.archives-ouvertes.fr/hal-03329848.

[35] M. Bleher, L. Hahn, J. Patiño-Galindo, M. Carriere, U. Bauer, R. Rabadán and A. Ott. *Topology identifies emerging adaptive mutations in SARS-CoV-2*. 6th Oct. 2021. URL: https://hal.inria.fr/hal-03368477.

[36] D. Cohen-Steiner, A. Lieutier and J. Vuillamy. *Efficient open surface reconstruction from lexicographic optimal chains and critical bases*. 3rd Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03456390.

[37] A. Delalande. *Nearly Tight Convergence Bounds for Semi-discrete Entropic Optimal Transport*. 29th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03396206.

[38] A. Delalande and Q. Merigot. *Quantitative Stability of Optimal Transport Maps under Variations of the Target Measure*. 9th Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-03164147.

[39] V. Divol. *A short proof on the rate of convergence of the empirical measure for the Wasserstein distance*. 21st Jan. 2021. URL: https://hal.inria.fr/hal-03117283.

[40] V. Divol. *Reconstructing measures on manifolds: an optimal transport approach*. 15th Feb. 2021. URL: https://hal.inria.fr/hal-03141977.

[41] M. Glisse. *Lower bound on the Voronoi diagram of lines in* $\mathbb{R}^d$. 17th Dec. 2021. URL: https://hal.inria.fr/hal-03491732.

[42] G. A. Gruetzner. *A uniformization theorem for closed convex polyhedra in Euclidean 3-space*. 31st Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03479663.

[43] O. Hacquard, K. Balasubramanian, G. Blanchard, W. Polonik and C. Levrard. *Topologically penalized regression on manifolds*. 25th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03402076.

[44] O. Hacquard, E. Lasalle and V. Lebovici. *Challenge mathématiques et entreprises : Eurecam*. Université Paris-Saclay, 15th Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03345714.

[45] K. Huszár. *On the pathwidth of hyperbolic 3-manifolds*. 11th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03373577.

[46] E. Lasalle. *Heat diffusion distance processes: a statistically founded method to analyze graph data sets*. 5th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03366848.

[47] V. Lebovici. *Hybrid transforms of constructible functions*. 10th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03474277.

[48] J. Leygonie, M. Carrière, T. Lacombe and S. Oudot. *A Gradient Sampling Algorithm for Stratified Maps with Applications to Topological Data Analysis*. 3rd Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03330940.

[49] D. Perez. *ζ-functions and the topology of superlevel sets of stochastic processes*. 19th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03372822.

[50] D. Perez. *On Sovacool's et al. study on the differences in carbon emissions reduction between countries pursuing renewable electricity versus nuclear power*. 16th Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-03170325.

[51] M. Perrot-Dockès, G. Blanchard, P. Neuvial and E. Roquain. *Post hoc false discovery proportion inference under a Hidden Markov Model*. 1st May 2021. URL: https://hal.archives-ouvertes.fr/hal-03214472.

[52] L. Pujol. *ISDE : Independence Structure Density Estimation*. 12th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03401530.

[53] E. M. Saad, G. Blanchard and S. Arlot. *Online Orthogonal Matching Pursuit*. 14th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03141061.