Analysis of multi-parameter Reeb spaces and Mappers

Primary advisor: Mathieu Carrière - Centre Inria d'Université Côte d'Azur, DataShape team - 50%
Secondary advisor: Steve Oudot - Inria Saclay & École polytechnique (LIX), GeomeriX team - 50%

Main location: Centre Inria d'Université Côte d'Azur

For further information: steve.oudot@inria.fr, mathieu.carriere@inria.fr

1 Context

An important challenge in data science is the so-called *curse of dimensionality*: when data points live in high dimensional spaces—in other words, when data points have a lot of associated measurements or variables—standard analysis methods are known to perform poorly, due to measure concentration phenomena. This is why many data mining and machine learning pipelines include a data embedding step that exploits the *intrinsic* structure of the data. The underlying hypothesis is that, while the data live in high dimensional space, they are actually distributed near some lower-dimensional substructure. Among existing embedding techniques, the ones coming from Topological Data Analysis (TDA) have gained popularity thanks to their ability to extract and encode the hidden topological structure of data, which by nature is invariant under large classes of transformations.

One of these techniques is the *Mapper* [SMC07], a discrete approximation to a well-known mathematical construction called the *Reeb space* [Ree46]. The Mapper provides ways of representing the data in a readily readable form using simplicial complexes, which are combinatorial models of topological spaces that can be easily stored and processed in computers. These representations are known to capture the local and global topological structure of the data, including connected components, loops, branches, and so on. Such topological features in the data can be subsequently used in exploratory contexts, for instance to identify relevant variables or subpopulations among the data [LSL+13]. They can also be exploited in deep learning contexts, notably for the analysis of the behavior of neural networks [BCE23, ZHL24].

The Reeb space and Mapper complex are defined for a continuous function $f : X \to \mathbb{R}^d$, called a *lens* or *filter*, where X is either a discrete space corresponding to the input data (for the Mapper complex) or the underlying continuous topological space (for the Reeb space). The Mapper complex is constructed by: (1) covering the image of f with a family $\mathcal{I}$ of hypercubes, so that $\mathrm{im}(f) \subseteq \bigcup_{I\in\mathcal{I}} I$; (2) pulling back this cover to X by computing the preimages of the hypercubes under f; (3) refining the pulled-back cover by separating the various connected components of its elements via some clustering technique, which gives a *connected cover* $\mathcal{V} = \{\mathrm{cc}(f^{-1}(I))\}_{I\in\mathcal{I}}$; (4) computing the nerve of the connected cover:

$$\mathrm{M}_f(X, \mathcal{I}) := \mathcal{N}(\mathcal{V}),$$

whose faces are in one-to-one correspondence with the non-empty k-fold intersections of elements of $\mathcal{V}$ (for all values of k). See Figure 1. The Reeb space is the limit version of this construction, using covers with singleton hypercubes. It is more directly defined as a quotient space, where points with same filter value that belong to the same level set component are glued together:

$$\mathrm{R}_f(X) := X/\sim_f,$$

where $x \sim_f x'$ if and only if $f(x) = f(x')$ and x, x' belong to the same connected component of $f^{-1}(\{f(x)\}) = f^{-1}(\{f(x')\})$.

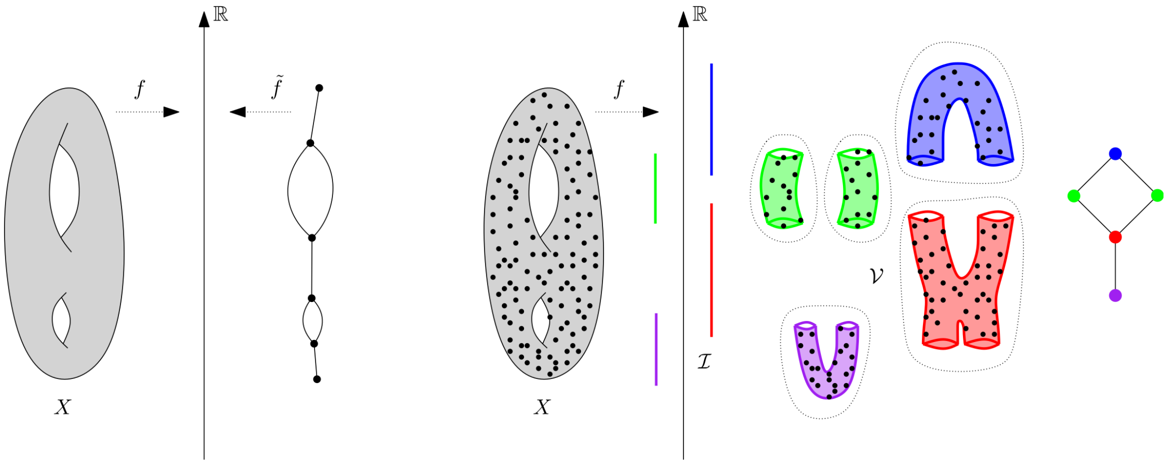

Figure 1: Example of Reeb graph **(left)** and Mapper graph **(right)** computed on a double torus X using the height function f as filter and a cover $\mathcal{I}$ with four intervals. This filter f induces a filter $\tilde{f} : \mathrm{R}_f(X) \to \mathbb{R}$ on the Reeb graph.

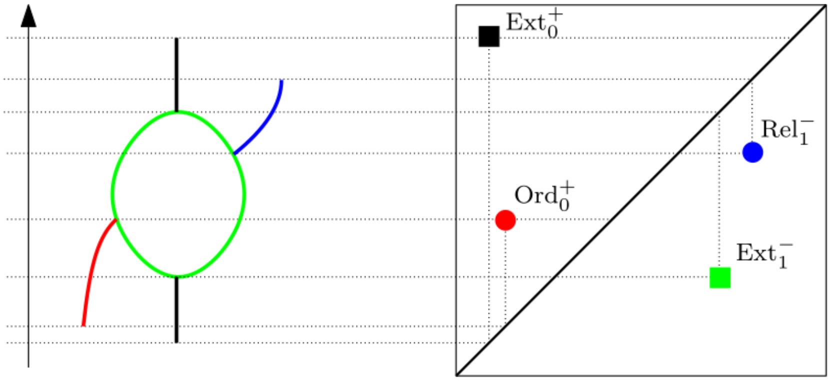

Figure 2: Example of Reeb graph and its corresponding extended persistence diagram. Each point in this diagram corresponds to a topological feature of the graph.

The theory of Reeb spaces and Mapper complexes (including the type of topological features they capture, as well as their statistical robustness) is now well established for filters taking values in $\mathbb{R}$ [BGW14, CMO18, CO17b]. Specifically, Reeb and Mapper graphs can be characterized by their *extended persistence diagrams*, which are standard topological descriptors from TDA. Diagrams are sets of points in the Euclidean plane, each point encoding the presence of a connected component, branch or a loop in the corresponding Mapper or Reeb graph, with its coordinates encoding the size of the feature. See Figure 2. These descriptors, in addition to having good visual interpretability and to being stable, can be compared efficiently with the bottleneck distance, allowing for subsequent statistical inference.

By contrast, much less is known on the *multi-parameter* case $f : X \to \mathbb{R}^d$ with $d > 1$, where filters capture the joint interactions of several variables at the same time (such as, e.g., scale and density or marker genes). In this case, a few distances have been defined [DMW17, MW16], with corresponding statistical results [BBMW21, CM22], but they remain both very hard to interpret (in the sense that the connection to the topological features that are actually encoded in Reeb spaces and Mapper complexes is unclear) and very difficult to compute from an algorithmic standpoint. Indeed, in the multi-parameter setting there is no easily computable and interpretable complete descriptor such as the extended persistence diagram.

2 Objectives and work program

The general aim of this Ph.D. will be to develop a rigorous mathematical framework for the analysis of the structure and stability of multi-parameter Reeb spaces and Mappers, relying on recent TDA contributions to the definition of stable and computable invariants for multi-parameter filters, in particular invariants coming from relative homological algebra. The impact of our framework on applications of Mapper in deep learning will also be investigated. The four axes of the work program are described below.

2.1 Expressiveness of signed barcodes and their variants

In full generality, the presence of topological features in a Reeb space or Mapper complex can be captured by the topology of the sublevel and superlevel sets of the filter $f\colon X \to \mathbb{R}^d$, via an abstract algebraic construction called the *extended persistence module* of f. This object is defined formally as the functor: $\mathbb{R}^d \to \mathrm{Vec}$ associating to every level $\alpha \in \mathbb{R}^d$ the homology group of the α-sublevel or α-superlevel set of f. Such modules do not admit a simple description in general, however many invariants have been proposed to encode part of their structure, some of which can be seen as generalizations of the extended persistence diagrams. Among these invariants, we will be interested in *signed barcodes* [BOO24] and their variants [BBH24], which are defined from projective resolutions of the modules relative to certain classes of intervals and admit interpretations in terms of decompositions in the corresponding relative Grothendieck groups.

The first objective of the Ph.D. will be to understand how much of the structure of the Reeb space or Mapper complex invariants of this type are able to capture. More precisely, we will seek to derive a dictionary of the features of the Reeb space or Mapper complex under consideration, from the interpretation of the structure of its corresponding persistence module's invariant as in the 1-parameter case [CO17b]. This task will likely be harder than in the 1-parameter case for two reasons: (1) the invariants considered for multi-parameter persistence modules are not complete, and (2) the connection between the persistent homology of the input filter f and that of the induced filter $\tilde{f}$ on the quotient Reeb space is less straightforward [BCP22].

2.2 Local metric equivalences for Reeb spaces and Mapper complexes

In the 1-parameter setting, there is a well-known local equivalence between the Gromov-Hausdorff distance between Reeb and Mapper graphs and the bottleneck distance between extended persistence diagrams [CO17a]. This equivalence, together with the aforementioned dictionary, made possible the study of the variations of Mapper and Reeb graphs locally, i.e., under small perturbations of the data. Ultimately, it enabled the statistical analysis of the behavior of the Mapper graph under certain noise models, and the automatic setup of its parameters via resampling techniques [CMO18].

The second objective of the Ph.D. will be to extend this local equivalence, and the ensuing statistical analysis, to the multi-parameter setting. Our strategy for the proof will to adapt the constructive approach developed by Vipond [Vip20] for locally comparing metrics on finitely presented persistence modules. For this adaptation we will leverage the interpretation of the signed barcodes developed in the first axis of the program, as well as existing stability results for signed barcodes [BOOS24, OS24].

2.3 Study of the geodesic space of Reeb spaces and Mapper complexes

Generally speaking, locally equivalent metrics induce globally equivalent path metrics. Based on this fact, the local equivalence worked out in the previous axis of the program will enable the third objective of the Ph.D., which will be to study the space of Reeb spaces and Mapper complexes as a path metric space. Questions such as its geodesic completeness and metric curvature will be considered, as they have an important impact on downstream applications such as optimization.

Optimization itself will be an important topic in this axis. In particular, we will focus on gradient descent and/or automatic differentiation processes, where the geodesic estimation is done in two steps, by first initializing interpolated Reeb spaces and Mapper complexes along the geodesic randomly, and then by optimizing them iteratively so that a global geodesic loss, computed by summing up local losses based on the multi-parameter topological invariants, is minimized. The convergence and quality associated to the local minima of such optimization processes, as well as the corresponding geodesic interpolations, will be investigated, building on recent topology-based optimization results [CCG+21, LCLO23, LOT21].

Statistical aspects will also be considered. In particular, classic subsampling and bootstrap techniques for estimating confidence regions in the single-parameter case [CMO18] will be generalized to the multi-parameter setting equipped with the path metric. The local equivalence will allow us to study and quantify the robustness of the subpopulations detected in the Reeb spaces and Mapper complexes, as well as their trajectories and interpolations along geodesics, opening the way to the use of Reeb spaces and Mapper complexes in trajectory inference and longitudinal studies.

2.4 Application to the monitoring of deep neural networks

The fourth objective of the Ph.D. will be to revisit recent contributions to the monitoring of deep neural networks using Mapper graphs or complexes. Of particular interest to us will be the use of Mapper complexes to analyze the weights and activation patterns in trained neural networks in order to assess their generalization capabilities. We believe the progress made in the previous axes of the program will provide theoretical grounding and further insight into existing TDA approaches for this question [CG20, Gab20, RCPW21].

3 Requested background

Considering the prominent role played by algebraic invariants of persistence modules in the first objective, which is a core part of the program, our primary request is a strong background in algebra, particularly in representation theory and in algebraic topology.

As a complement, for the other objectives we are requesting some background in statistics, in Riemannian or metric geometry, and in machine and deep learning—possibly with some hands-on experience with existing ML/DL libraries such as Scikit-learn and PyTorch.

4 Scientific environment

The Ph.D. student will be integrated in the DataShape team at Centre Inria d'Université Côte d'Azur, under the supervision of **Mathieu Carrière**. He will also interact with **Steve Oudot** on a weekly basis via videoconferencing, and through monthly visits to the GeomeriX team at LIX & Inria Saclay. Implementations and software developed within the Ph.D. project will potentially be included in the TDA open-source library `Gudhi`[1].

References

[BBH24] Benjamin Blanchette, Thomas Brüstle, and Eric J Hanson. Homological approximations in persistence theory. *Canadian Journal of Mathematics*, 76(1):66–103, 2024.

[BBMW21] Adam Brown, Omer Bobrowski, Elizabeth Munch, and Bei Wang. Probabilistic convergence and stability of random Mapper graphs. *Journal of Applied and Computational Topology*, 5:99–140, 2021.

[BCE23] Rubén Ballester, Carles Casacuberta, and Sergio Escalera. Topological data analysis for neural network analysis: A comprehensive survey. *arXiv preprint arXiv:2312.05840*, 2023.

[BCP22] Saugata Basu, Nathanael Cox, and Sarah Percival. On the reeb spaces of definable maps. *Discrete & Computational Geometry*, 68(2):372–405, 2022.

[BGW14] Ulrich Bauer, Xiaoyin Ge, and Yusu Wang. Measuring distance between Reeb graphs. In *30th Annual Symposium on Computational Geometry (SoCG 2014)*, pages 464–473. Association for Computing Machinery, 2014.

[1] https://gudhi.inria.fr/

[BOO24] Magnus Bakke Botnan, Steffen Oppermann, and Steve Oudot. Signed barcodes for multi-parameter persistence via rank decompositions and rank-exact resolutions. *Foundations of Computational Mathematics*, pages 1–60, 2024.

[BOOS24] Magnus Bakke Botnan, Steffen Oppermann, Steve Oudot, and Luis Scoccola. On the bottleneck stability of rank decompositions of multi-parameter persistence modules. *Advances in Mathematics*, 451:109780, 2024.

[CCG$^+$21] Mathieu Carriere, Frédéric Chazal, Marc Glisse, Yuichi Ike, Hariprasad Kannan, and Yuhei Umeda. Optimizing persistent homology based functions. In *International conference on machine learning*, pages 1294–1303. PMLR, 2021.

[CG20] Gunnar Carlsson and Rickard Brüel Gabrielsson. Topological approaches to deep learning. In *Topological Data Analysis: The Abel Symposium 2018*, pages 119–146. Springer, 2020.

[CM22] Mathieu Carrière and Bertrand Michel. Statistical analysis of mapper for stochastic and multivariate filters. *Journal of Applied and Computational Topology*, pages 2367–1734, 2022.

[CMO18] Mathieu Carrière, Bertrand Michel, and Steve Oudot. Statistical analysis and parameter selection for Mapper. *Journal of Machine Learning Research*, 19(12):1–39, 2018.

[CO17a] Mathieu Carrière and Steve Oudot. Local equivalence and intrinsic metrics between Reeb graphs. In *33rd International Symposium on Computational Geometry (SoCG 2017)*, volume 77, pages 25:1–25:15. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017.

[CO17b] Mathieu Carrière and Steve Oudot. Structure and stability of the one-dimensional Mapper. *Foundations of Computational Mathematics*, 18(6):1333–1396, 2017.

[DMW17] Tamal Dey, Facundo Mémoli, and Yusu Wang. Topological analysis of nerves, Reeb spaces, Mappers, and Multiscale Mappers. In *33rd International Symposium on Computational Geometry (SoCG 2017)*, volume 77, pages 36:1–36:16. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017.

[Gab20] Maxime Gabella. Topology of learning in feedforward neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8):3588–3592, 2020.

[LCLO23] Jacob Leygonie, Mathieu Carrière, Théo Lacombe, and Steve Oudot. A gradient sampling algorithm for stratified maps with applications to topological data analysis. *Mathematical Programming*, 202(1):199–239, 2023.

[LOT21] Jacob Leygonie, Steve Oudot, and Ulrike Tillmann. A framework for differential calculus on persistence barcodes. *Foundations of Computational Mathematics*, 2021.

[LSL$^+$13] Pek Y Lum, Gurjeet Singh, Alan Lehman, Tigran Ishkanov, Mikael Vejdemo-Johansson, Muthu Alagappan, John Carlsson, and Gunnar Carlsson. Extracting insights from the shape of complex data using topology. *Scientific reports*, 3(1):1236, 2013.

[MW16] Elizabeth Munch and Bei Wang. Convergence between categorical representations of Reeb space and Mapper. In *32nd International Symposium on Computational Geometry (SoCG 2016)*, volume 51, pages 53:1–53:16. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2016.

[OS24] Steve Oudot and Luis Scoccola. On the stability of multigraded betti numbers and hilbert functions. *SIAM Journal on Applied Algebra and Geometry*, 8(1):54–88, 2024.

[RCPW21] Archit Rathore, Nithin Chalapathi, Sourabh Palande, and Bei Wang. Topoact: Visually exploring the shape of activations in deep learning. In *Computer Graphics Forum*, volume 40, pages 382–397. Wiley Online Library, 2021.

[Ree46] Georges Reeb. Sur les points singuliers d'une forme de Pfaff complètement intégrable ou d'une fonction numérique. *Comptes Rendus de l'Académie des Sciences de Paris*, 222:847–849, 1946.

[SMC07] Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson. Topological methods for the analysis of high dimensional data sets and 3D object recognition. In *4th Eurographics Symposium on Point-Based Graphics (SPBG 2007)*, pages 91–100. The Eurographics Association, 2007.

[Vip20] Oliver Vipond. Local equivalence of metrics for multiparameter persistence modules. In *CoRR*. arXiv:2004.11926, 2020.

[ZHL24] Ben Zhang, Zitong He, and Hongwei Lin. A comprehensive review of deep neural network interpretation using topological data analysis. *Neurocomputing*, 609:128513, 2024.