6.5 Optimizing persistent homology based criteria for a parameterized family of filter functions

Elements of o-minimal geometry theory

We will use o-minimal geometry to study the differentiability of functions of persistence in the Mapper setting. We give here some bases of the theory, see [Cos02] for more details.

Definition 37. *o-minimal structure*
An o-minimal structure on the field on real numbers $\mathbb{R}$ is a collection $(S_n)_{n\in\mathbb{N}}$ where each S_n is a set of subsets of $\mathbb{R}^n$ that verify :

1. *S_1 is exactly the family of finite unions of points and intervals,*
2. *All algebraic subsets of $\mathbb{R}^n$ are in S_n,*
3. *S_n is a Boolean subalgebra of $\mathbb{R}^n$ (i.e. stable by finite union, finite intersection and complementary),*
4. *if $A \in S_n$ and $B \in S_m$ then $A \times B \in S_{n+m}$,*
5. *if $\pi\colon \mathbb{R}^{n+1} \to \mathbb{R}^n$ is the linear projection onto the first n coordinates and $A \in S_{n+1}$ then $\pi(A) \in S_n$.*

The elementary example of an o-minimal structure is the collection of semi-algebraic sets.
An element $A \in S_n$ for some $n \in \mathbb{N}$ is called a definable set. Furthermore, a map $f\colon A \to \mathbb{R}^m$ is called a definable map if its graph (i.e. $\{(x, f(x)) : x \in A\}$) is in S_{n+m}.
Definable maps are stable under addition, product and composition. Moreover, the *max* and *min* of real valued definable maps is also definable.
An important property of definable maps is that they admit a finite Whitney stratification. This means that if $f\colon A \to \mathbb{R}^m$ is definable with $A \in S_n$, then A can be decomposed into a finite union of smooth manifolds such that the restriction of f to each of these manifolds is a differentiable function.

Optimization problem

We put ourselves in the unsupervised setting where the choice of a filter function has to be made automatically through tuning a set of parameters. We express this choice in the form of an optimization problem that we decompose in the following manner :
let $\{x_i\}_{1\leq i\leq n}$ be a fixed set of points, N_I a fixed number of intervals in the cover of the filter image space, g a fixed overlap parameter and N_θ a fixed number of parameters that determine the filter function. We first define the F function that associates a vector of filter values evaluated on the point cloud to a set of parameters for the filter function :

$$\begin{aligned} F\colon \mathbb{R}^{N_\theta} &\to \mathbb{R}^n \\ \theta &\mapsto (f_\theta(x_i))_{1\leq i\leq n}. \end{aligned}$$

This map is differentiable depending on the map $\theta \mapsto f_\theta$. In particular, f_θ can be obtained through a deep learning architecture in which θ denotes the model weights.
Furthermore, such a filter function would guarantee that F is definable in any o-minimal structure as it would be the result of composition and Cartesian product of definable functions.
Let M_f be the Mapper built on $\{x_i\}_{1\leq i\leq n}$ using parameters g, N_I, a clustering algorithm for which we fix a number of clusters N_C and a filter function f. Then we can define a function C that associates filtration values on the vertices of M_f to the filter values associated to the point cloud :

$$\begin{aligned} C\colon \mathbb{R}^n &\to \mathbb{R}^{N_I \times N_C} \\ (f_i)_{1\leq i\leq n} &\mapsto (\tilde{f}(c_j))_{1\leq j\leq N_I \times N_C}. \end{aligned}$$

Several choices for $\tilde{f}$ can be made to satisfy that C is definable, one such choice would be :

$$\tilde{f}(c_j) = \frac{1}{\mid c_j \mid} \sum_{i:x_i \in c_j} f_i.$$

Indeed, this would mean that C is algebraic and therefore definable in any o-minimal structure.
Note that the number of vertices in M is $N_I \times N_C$ independently of the choice of a filter function.
Furthermore, we can associate filtration values to the simplices of the Mapper as follows :

$$\Phi \colon \mathbb{R}^{N_I \times N_C} \to \mathbb{R}^{|M_f|}$$
$$(\tilde{f}_j)_{1 \leq j \leq N_I \times N_C} \mapsto (\hat{f}(\sigma_k))_{1 \leq k \leq |M_f|},$$

where

$$\hat{f}(\sigma_k) = \max_{j:c_j \in \sigma_k} \tilde{f}_j.$$

This corresponds to a sublevel set filtration on the Mapper simplicial complex. As stated in [Car+21] it belongs to a definable family of filtrations.
To resolve the issue of Φ having a codomain that depends on the Mapper and therefore on the filter function, we can equivalently consider that it associates values to the simplices of K, the 1-skeleton of the simplicial complex that contains all the faces of the $(n-1)$-simplex containing the data points.

$$\Phi \colon \mathbb{R}^{N_I \times N_C} \to \mathbb{R}^{|K|}$$
$$(\tilde{f}_j)_{1 \leq j \leq N_I \times N_C} \mapsto (\hat{f}(\sigma_k))_{1 \leq k \leq |K|},$$

where

$$\hat{f}(\sigma_k) = (\max_{j:c_j \in \sigma_k} \tilde{f}_j).\mathbb{1}_{\sigma_k \in M_f} + A.\mathbb{1}_{\sigma_k \notin M_f},$$

and A is an arbitrarily large real value, undissociated from infinity in the persistence diagram. The terms we added in the expression are all constants with respect to the filtration values assigned to the clusters. Φ is therefore still definable for the same reasons as before.
Finally, in the same way as in [Car+21] we compute the persistence diagram using the map $Pers \colon \mathbb{R}^{|M_f|} \to (\mathbb{R}^2)^p \times \mathbb{R}^q$ and we use a function of persistence $E \colon (\mathbb{R}^2)^p \times \mathbb{R}^q \to \mathbb{R}$ that expresses our persistent homology based criteria. p and q denote here the number of paired and unpaired points in the persistence diagram respectively.
Taking into account all of the above, the function to minimize is :

$$\mathcal{L} = E \circ Pers \circ \Phi \circ C \circ F.$$

References

[Bel+22] Francisco Belchi et al. 'A Numerical Measure of the Instability of Mapper-Type Algorithms'. In: *Journal of Machine Learning Research* 21.1 (June 2022), pp. 1–45. ISSN: 1532-4435.

[BL08] Shai Ben-David and Ulrike Luxburg. 'Relating Clustering Stability to Properties of Cluster Boundaries.' In: *Proceedings of the 21st Annual Conference on Learning Theory* (Jan. 2008), pp. 379–390.

[Car+21] Mathieu Carriere et al. 'Optimizing persistent homology based functions'. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 1294–1303.

[CM21] Frédéric Chazal and Bertrand Michel. 'An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists'. In: *Frontiers in artificial intelligence* 4 (Sept. 2021). DOI: 10.3389/frai.2021.667963.

[CMO18] Mathieu Carrière, Bertrand Michel and Steve Oudot. 'Statistical Analysis and Parameter Selection for Mapper'. In: *Journal of Machine Learning Research* 19 (July 2018), pp. 1–39.

[Cos02] Michel Coste. 'AN INTRODUCTION TO O-MINIMAL GEOMETRY'. In: 2002.

[DW22] Tamal Krishna Dey and Yusu Wang. *Computational Topology for Data Analysis*. Cambridge University Press, 2022. DOI: 10.1017/9781009099950.

[NSW08] Partha Niyogi, Stephen Smale and Shmuel Weinberger. 'Finding the Homology of Submanifolds with High Confidence from Random Samples'. In: *Discrete & Computational Geometry* 39 (2008), pp. 419–441. DOI: 10.1007/s00454-008-9053-2.

[RB19] Raul Rabadan and Andrew J. Blumberg. *Topological Data Analysis for Genomics and Evolution: Topology in Biology*. Cambridge University Press, 2019. DOI: 10.1017/9781316671665.