Projected Barcodes : A New Approach to Multi-parameter Persistent Homology

Author : Luca Nyckees Supervisors : Nicolas Berkouk [1], Prof. Kathryn Hess Bellwald [1]

[1]Laboratory of Topology and Neuroscience at EPFL

Notation and Mathematical Framework

- Let K detote a simplicial complex, X a topological space and consider a fixed field $\mathbf{k}$.
- Write $\mathrm{Lip}_C(\mathbb{R}^d) = \{p : \mathbb{R}^d \to \mathbb{R} \mid p\ C\text{-Lipschitz}\}$
- $\mathbf{Vect_k}$ is the category of $\mathbf{k}$-vector spaces, and $\mathbb{R}^d$ is the category induced by the poset $(\mathbb{R}^d, \preceq)$
- $S_j(f) : \mathbb{R}^d \to \mathbf{Vect_k} : a \mapsto H_j(x \mid f(x) \preceq a)$ for a map $f : K \to \mathbb{R}^d$
- Persistence modules can be seen as sheaves for the euclidean topology.
- $\mathrm{D}^b(\mathrm{Pers}(\mathbb{R}^d))$ denotes its bounded derived category of n-persistence modules.
- A 1-Lipschitz map $p : \mathbb{R}^d \to \mathbb{R}$ induces a pushforward functor $Rp_* : \mathrm{D}^b(\mathrm{Pers}(\mathbb{R}^d)) \to \mathrm{D}^b(\mathrm{Pers}(\mathbb{R}))$.
- Key-point : for a d-persistence module M, one obtains a 1-persistence module Rp_*M.

Introduction

In topological data analysis (TDA), the algebraic framework used for multi-parameter persistent homology presents complex challenges for the community. The so-called class of projected barcodes, originally introduced in the language of sheaf theory, tackles common issues encountered in multi-persistence, presents nice theoretical properties and generalizes the fibered barcode. In this work, we concentrate on the linear projected barcode in the language of multi-parameter persistence modules. We propose an efficient way of computing the linear integral sheaf metric between two multi-parameter persistence modules, through topological optimization. Furthermore, we show how the linear projected barcode can integrate standard TDA pipelines and present a feature extraction application for MNIST classification. The complete implementation, together with different notebooks and visual features, are avaible at this GitHub account in the repository called `sheafmetric|`.

The Fibered Barcode

In general, multi-parameter persistence modules do not admit well-defined barcode decompositions. The quest for other computational invariants remains a central task in computational topology. In dimension $d \geq 2$, the interleaving distance is NP-hard to compute (Bjerkevik et al. 2019). The *fibered barcode* of a multi-parameter persistence module $M : \mathbb{R}^d \to \mathbf{Vect_k}$ is defined as the collection

$$\mathrm{Fib}(M) = \{\mathcal{B}(M_\mathcal{L})\}_{\mathcal{L}\in\Lambda},$$

where Λ is the set of lines in $\mathbb{R}^d$ with positive coefficients, and $M_\mathcal{L}$ is, intuitively, the restriction of M to the line $\mathcal{L}$. The *matching distance* between two persistence modules $M, N : \mathbb{R}^d \to \mathbf{Vect_k}$ is defined as

$$d_{\mathrm{match}}(M, N) = \sup_{\mathcal{L}\in\Lambda} w(\mathcal{L}) \cdot d_B(M_\mathcal{L}, N_\mathcal{N}),$$

where $w(\cdot)$ is a weight function that ensures the stability condition with respect to the interleaving distance : $d_{\mathrm{match}} \leq d_I$. In Lesnick and Wright 2019, M. Lesnick and M. Wright introduce the RIVET software, which computes the matching distance for sublevel set homology of functions $f, g : K \to \mathbb{R}^2$ in complexity $\mathcal{O}(|K|^{11})$.

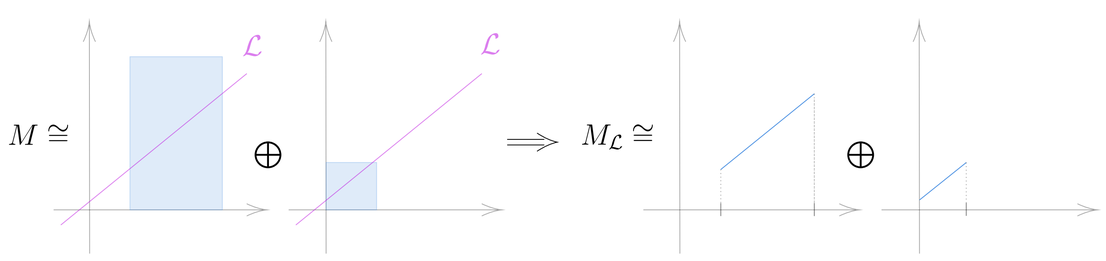

Figure 1. Restriction of a 2-parameter persistence module M to a line $\mathcal{L}$.

The Projected Barcode

In Berkouk and Petit 2022, N. Berkouk and F. Petit introduce a new class of invariants called *projected barcodes*, that successfully generalize the fibered barcode construction and present nice theoretical properties. Consider a class $\mathfrak{F} \subseteq \mathcal{C}^0(\mathbb{R}^d)$. The $\mathfrak{F}$-projected barcode of a continuous map $f : K \to \mathbb{R}^d$ is defined as

$$\mathrm{Proj}_\mathfrak{F}(f) = \{\mathcal{B}(Rp_*S(f))\}_{p\in\mathfrak{F}}.$$

The $\mathfrak{F}$-*integral sheaf metric* (ISM) between two continuous maps $f : K_1 \to \mathbb{R}^d$ and $g : K_2 \to \mathbb{R}^d$ is

$$d_\mathfrak{F}(f, g) = \sup_{p\in\mathfrak{F}} \sup_{j\in\mathbb{N}} d_B(R^jp_*S(f), R^jp_*S(g)).$$

Theoretical Properties

Consider continuous maps $f : X \to \mathbb{R}^d$ and $g : Y \to \mathbb{R}^d$. N. Berkouk and F. Petit show the following.

Theorem 1: Stability

Let $\mathfrak{F} \subseteq \mathrm{Lip}_C(\mathbb{R}^d)$ with $C \leq 1$. One has

$$d_\mathfrak{F}(f, g) \leq d_I(S(f), S(g)).$$

Theorem 2: Lower bound d_{match}

Let $\mathfrak{F} \subseteq \mathrm{Lip}_1(\mathbb{R}^d)$, with $\{\mathrm{push}_\mathcal{L}/w(\mathcal{L})\} \subseteq \mathfrak{F}$. One has

$$d_{\mathrm{match}}(S_j(f), S_j(g)) \leq d_\mathfrak{F}(f, g).$$

Theorem 3

If X is compact and $p \in \mathrm{Lip}_C(\mathbb{R}^d)$ is a linear form with positive coefficients, then $Rp_*S_\bullet(f) = S_\bullet(p \circ f)$.

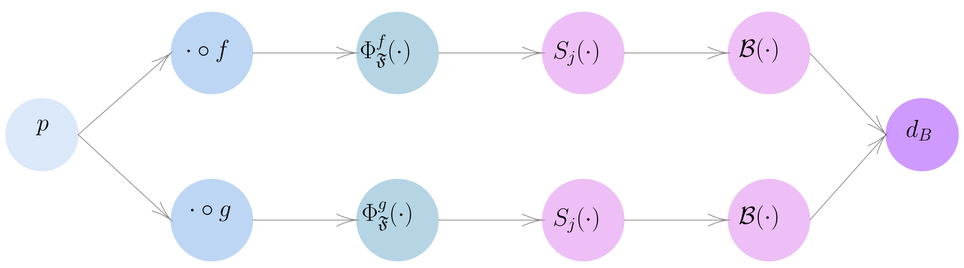

Figure 2. Auto-differentiation tree for LISM computation.

Topological Optimization

Provided that $\mathfrak{F}$ is compact, one may wish to maximize the functional

$$\mathcal{T}^j_{f,g} : \mathfrak{F} \to \mathbb{R} : p \mapsto d_B(\mathcal{B}(S_j(p \circ f)), \mathcal{B}(S_j(p \circ g)))$$

in order to compute $d_\mathfrak{F}(f, g)$.

Theorem 4: Differentiability

Let $f : K_1 \to \mathbb{R}^d$ and $g : K_2 \to \mathbb{R}^d$ be continuous. Let $\mathfrak{F} = \{p \in \mathbb{R}^d_+ \mid \|p\|_1 = 1\}$. Then the functional

$$\mathcal{T}^j_{f,g} : \mathfrak{F} \to \mathbb{R} : p \mapsto d_B(\mathcal{B}(S_j(p \circ f)), \mathcal{B}(S_j(p \circ g)))$$

is differentiable Lebesgue almost everywhere.

Hence, one can use a subgradient algorithm with iterations of the form

$$p_{t+1} = p_t - \eta_t(y_t + \zeta_t),$$

where $y_t \in \partial\mathcal{T}^j_{f,g}(p_t)$ ($\partial\mathcal{T}^j_{f,g}$ is the Clarke subdifferential of $\mathcal{T}^j_{f,g}$). Now, following the ideas of Theorem 4.2 in Carrière et al. 2021, one has the following convergence result.

Theorem 5: Convergence of subgradient algorithm

Let $\eta_t = \frac{1}{t}$ and $\zeta_t \sim \mathcal{N}(0, \sigma^2)$ with $\sigma \ll 1$. Then the sequence of projections $(p_t)_t$ in the subgradient algorithm above converges almost surely to a critical point of $\mathcal{T}^j_{f,g}$, and the sequence $(\mathcal{T}^j_{f,g}(p_t))_t$ converges.

Applications

Implementation

In the implementation, we look at the class $\mathfrak{F} = \{p \in \mathbb{R}^d_+ \mid \|p\|_1 = 1\}$, seen as a subset of $\mathcal{C}^0(\mathbb{R}^d)$ through the projection action $p(x_1, x_2) = p_1x_1 + p_2x_2$ for $(x_1, x_2) \in \mathbb{R}^2$. The optimization objective is implemented by considering the automatic-differentiation tree shown in Figure 2 (part of the code is built on top of Mathieu Carrière's topological optimization framework). Moreover, we consider a 2-Wasserstein version of the problem :

$$d_\mathfrak{F}(f, g) = \max_{p\in\mathfrak{F}} \max_{j\in\mathbb{N}} d_W(S_j(p \circ f), S_j(p \circ g)).$$

Image Analysis Pipeline

There are several ways to integrate the $\mathfrak{F}$-projected barcode to a machine learning pipeline. We focus on MNIST classification. For a choice of N images $\{I_k\}_{k=1}^N \subset \mathbb{R}^{n\times n}$ to classify and a family of image-induced multi-filtrations $\{f^k : [n] \times [n] \to \mathbb{R}^d\}_{k=1}^N$, one can compute a distance matrix $D = (d_{kl})_{k,l=1}^N$ with

$$d_{kl} = d_\mathfrak{F}(f^k, f^l).$$

By considering this matrix as input for a clustering algorithm, one may perform unsupervised learning (Fig. 3). We report some MNIST classification experiments below, where $R_c, H_v : \mathbb{R}^{n\times n} \to \mathbb{R}$ are c-radial and v-height filtrations (see Garin and Tauzin 2019 for more details). In brief, we compare multi-parameter persistence with single-parameter persistence. Figure 4 reports the clustering of three distance matrices for classifying $N = 100$ images with four labels : the 2-Wasserstein distance matrices of 1D sub-problems and the linear integral sheaf distance matrix.

metric	multi-filtration	N	digits	dimension	accuracy
d_W	$R_{(6,20)}$	100	$(0, 2, 8, 9)$	$0, 1$	56%
d_W	$H_{(0,1)}$	100	$(0, 2, 8, 9)$	$0, 1$	65%
$d_\mathfrak{L}$	$(R_{(6,20)}, H_{(0,1)})$	100	$(0, 2, 8, 9)$	$0, 1$	74%
d_W	$R_{(13,13)}$	100	$(3, 6, 8, 9)$	$0, 1$	60%
d_W	$H_{(0,1)}$	100	$(3, 6, 8, 9)$	$0, 1$	73%
$d_\mathfrak{L}$	$(R_{(13,13)}, H_{(0,1)})$	100	$(3, 6, 8, 9)$	$0, 1$	81%

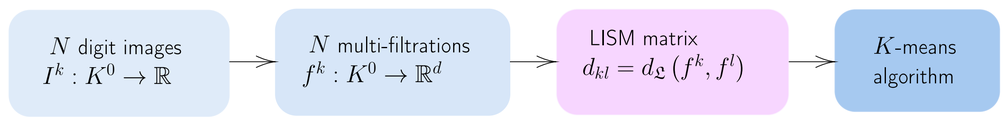

Figure 3. Projected barcode integrated to TDA pipeline.

Figure 4. Multi-dimensional embedding of distance matrices (last experiment in the table above.)

References

- Berkouk, N. and F. Petit (2022). *Projected distances for multi-parameter persistence modules*. DOI: 10.48550/ARXIV.2206.08818. URL: https://arxiv.org/abs/2206.08818.
- Carrière, M., F. Chazal, M. Glisse, Y. Ike, and H. Kannan (2021). *Optimizing persistent homology based functions*. arXiv: 2010.08356 [cs.CG].
- Bjerkevik, H., M. Botnan, and M. Kerber (Nov. 2019). "Computing the Interleaving Distance is NP-Hard". In: *Foundations of Computational Mathematics* 20. DOI: 10.1007/s10208-019-09442-y.
- Garin, A. and G. Tauzin (2019). "A Topological "Reading" Lesson: Classification of MNIST using TDA". In: *CoRR* abs/1910.08345. arXiv: 1910.08345. URL: http://arxiv.org/abs/1910.08345.
- Lesnick, M. and M. Wright (2019). *Computing Minimal Presentations and Bigraded Betti Numbers of 2-Parameter Persistent Homology*. DOI: 10.48550/ARXIV.1902.05708. URL: https://arxiv.org/abs/1902.05708.