Wasserstein Dictionaries of Persistence Diagrams

K.Sisouk[1] J.Delon[2] J. Tierny[1]

[1] CNRS, Sorbonne Université (LIP6); [2] CNRS, Université Paris Cité (MAP5)

[1]{keanu.sisouk, julien.tierny}@sorbonne-universite.fr; [2]julie.delon@parisdescartes.fr

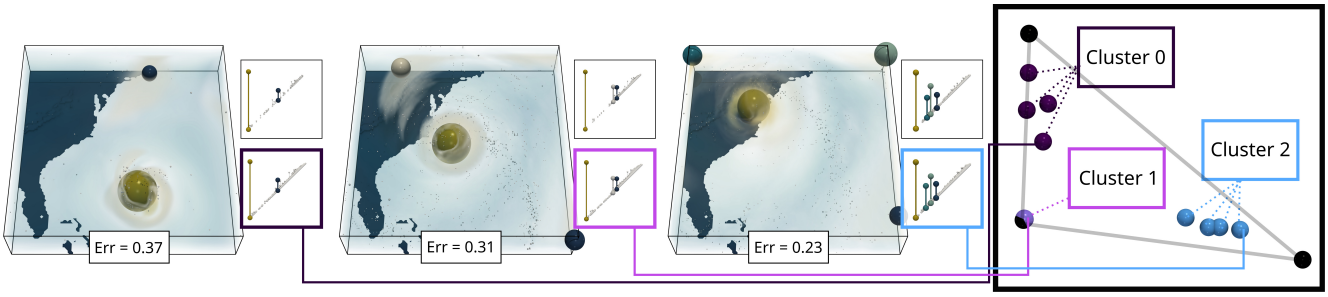

FIGURE 1 – *Visual comparison (left) between the input persistence diagrams (top insets, saddle-maximum persistence pairs only) and our compressed diagrams (bottom insets, saddle-maximum persistence pairs only) for three members of the* Isabel *ensemble (one member per ground-truth class). For each member, the sphere color encodes the matching between the input and the compressed diagrams (for the meaningful persistence pairs, above* 10% *of the function range). This visual comparison shows that the main features of the diagrams (encoding the main hurricane wind gusts in the data) are well preserved by the data reduction, especially for the members coming from the cluster* 2*, for which a lower relative reconstruction error (*Err*) can be observed. The planar overview of the ensemble (right) generated by our dimensionality reduction enables the visualization of the relations between the different diagrams of the ensemble. Specifically, this illustration shows a larger disparity for two clusters.*

Abstract

This paper presents a computational framework for the concise encoding of an ensemble of persistence diagrams, in the form of weighted Wasserstein barycenters [1, 2] of a dictionary of atom diagrams*. We introduce a multi-scale gradient descent approach for the efficient resolution of the corresponding minimization problem, which interleaves the optimization of the barycenter weights with the optimization of the* atom diagrams*. Our approach leverages the analytic expressions for the gradient of both sub-problems to ensure fast iterations and it additionally exploits shared-memory parallelism. Extensive experiments on public ensembles demonstrate the efficiency of our approach, with Wasserstein dictionary computations in the orders of minutes for the largest examples. We show the utility of our contributions in two applications. First, we apply Wassserstein dictionaries to* data reduction *and reliably compress persistence diagrams by concisely representing them with their weights in the dictionary. Second, we present a* dimensionality reduction *framework based on a Wasserstein dictionary defined with a small number of atoms (typically three) and encode the dictionary as a low dimensional simplex embedded in a visual space (typically in 2D). In both applications, quantitative experiments assess the relevance of our framework. Finally, we provide a C++ implementation that can be used to reproduce our results.*

Index Terms

Topological data analysis, ensemble data, persistence diagrams.

1 Introduction

As measurement devices and numerical techniques are becoming more and more advanced, datasets are becoming more and more complex geometrically. This geometrical complexity makes interactive exploration and analysis difficult, which challenges the interpretation of the data by the users. This motivates the creation of expressive data abstractions, capable of encapsulating the main features of interest of the data into simple representations, visually conveying the main information to the user.

Topological Data Analysis (TDA) [3] is a family of techniques which precisely addresses this issue. It provides concise topological descriptors of the main structural features hidden in a dataset. The relevance of TDA for analyzing scalar data, its efficiency and robustness have been documented in a number of visualization tasks [4].

Among the different topological descriptors studied in TDA , the Persistence Diagram (Fig. 2) is a particularly prominent example. It is a concise topological descriptor

which captures the main structural features in a dataset and which assesses their individual importance.
In addition to the challenge of increased geometrical complexity (discussed above), a new difficulty has recently emerged in many applications, with the notion of *ensemble dataset*. These representations describe a given phenomenon not only with a single dataset, but with a *collection* of datasets, called *ensemble members*. In that context, the topological analysis of an ensemble dataset consequently results in an ensemble of corresponding topological descriptors (e.g. one persistence diagram per ensemble member).

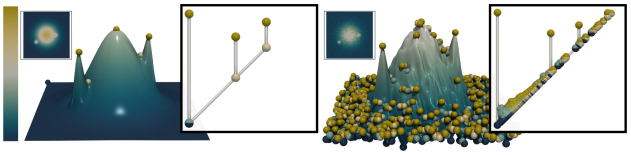

FIGURE 2 – *Persistence diagrams of a clean (left) and noisy (right) terrain (dark blue spheres : minima, dark yellow : maxima, other : saddles). The three main hills are clearly represented with long bars in the persistence diagrams. In the noisy persistence diagram, small bars encode noise.*

Then, a major challenge consists in developing practical tools for such an ensemble of topological descriptors, to facilitate its processing, analysis and visualization. Such tools include compression approaches (to facilitate the manipulation of the ensemble of descriptors) or visualization methods (for instance, with planar layouts, where each point encodes a descriptor and the distance between a pair of points encodes the intrinsic differences between the corresponding descriptors).
To enable the above tools, a key research question deals with the definition of a concise, yet informative, encoding of the ensemble of descriptors. A promising research direction consists in defining a *dictionary* (i.e. a set of reference descriptors, or *atoms*), such that the topological descriptors of the ensemble can be concisely encoded by expressing them as a specific *function* of the atoms (e.g. a linear combination). At a technical level, this requires to accurately capture and model the implicit relations (i.e. the possible functions) which link the different descriptors of the ensemble.
A series of recent works started the exploration of this overall direction, in particular with the notion of *average topological representation* [1, 5, 2, 6, 7]. These techniques can produce a topological descriptor which nicely summarizes the ensemble. However, they do not capture the implicit relations between the different topological descriptors.

2 Contributions

This paper addresses this issue by introducing a simple and efficient approach for the estimation of linear relations between persistence diagrams on their associated Wasserstein metric space. Inspired by previous work on histograms [8], our approach provides a linear encoding of the input ensemble, where each diagram is represented as a weighted Wasserstein barycenter [1, 2] of a *dictionary* of automatically optimized diagrams called *atom diagrams*. We introduce a novel multi-scale gradient descent algorithm for the efficient resolution of the corresponding minimization problem, for which we interleave the optimization of the barycenter weights with the optimization of the atom diagrams . Extensive experiments on public ensembles demonstrate the efficiency of our approach, with Wasserstein dictionary computations in the orders of minutes for the largest examples.

3 Applications

We illustrate the relevance of our contributions for the visual analysis of ensemble data with two applications, data reduction and dimensionality reduction :

1. *An application to data reduction :* We present an application to data reduction, where the persistence diagrams of the input ensemble are significantly compressed, by solely storing their barycentric weights as well as the atom diagrams.
2. *An application to dimensionality reduction :* We present an application to dimensionality reduction, by embedding each input diagram as a point within a 2D view, based on its weights relative to a Wasserstein dictionary composed of three atoms (thereby defining a 2-simplex).

Acknowledgments

This work is partially supported by the European Commission grant ERC-2019-COG *"TORI"* (ref. 863464, `https://erc-tori.github.io/`).

Appendix

Published in IEEE TVCG (Transactions on Visualization and Computer Graphics), Volume : 30, Issue : 2, 01 February 2024. DOI : 10.1109/TVCG.2023.3330262.
Paper available on arXiv :`https://arxiv.org/abs/2304.14852`

Références

[1] Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, et John Harer. Fréchet Means for Distributions of Persistence Diagrams. *DCG*, 2014.

[2] Jules Vidal, Joseph Budin, et Julien Tierny. Progressive Wasserstein Barycenters of Persistence Diagrams. *IEEE TVCG*, 2020.

[3] H. Edelsbrunner et J. Harer. *Computational Topology : An Introduction*. American Mathematical Society, 2009.

[4] C. Heine, H. Leitte, M. Hlawitschka, F. Iuricich, L. De Floriani, G. Scheuermann, H. Hagen, et C. Garth. A survey of topology-based methods in visualization. *CGF*, 2016.

[5] Théo Lacombe, Marco Cuturi, et Steve Oudot. Large Scale computation of Means and Clusters for Persistence Diagrams using Optimal Transport. Dans *NIPS*, 2018.

[6] Lin Yan, Yusu Wang, Elizabeth Munch, Ellen Gasparovic, et Bei Wang. A structural average of labeled merge trees for uncertainty visualization. *IEEE TVCG*, 2019.

[7] Mathieu Pont, Jules Vidal, Julie Delon, et Julien Tierny. Wasserstein Distances, Geodesics and Barycenters of Merge Trees. *IEEE TVCG*, 2022.

[8] Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, et Jean-Luc Starck. Wasserstein dictionary learning : Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 2018.