Comparing Distance Metrics on Vectorized Persistence Summaries

Brittany Terese Fasy*
School of Computing & Department of Mathematical Sciences
Montana State University
brittany.fasy@montana.edu

Yu Qin†
Department of Computer Science,
Tulane University
yqin2@tulane.edu

Brian Summa†
Department of Computer Science,
Tulane University
bsumma@tulane.edu

Carola Wenk†
Department of Computer Science,
Tulane University
cwenk@tulane.edu

Abstract

The persistence diagram (PD) is an important tool in topological data analysis for encoding an abstract representation of the homology of a shape at different scales. Different vectorizations of PD summary are commonly used in machine learning applications, however distances between vectorized persistence summaries may differ greatly from the distances between the original PDs. Surprisingly, no research has been carried out in this area before. In this work we compare distances between PDs and between different commonly used vectorizations. Our results give new insights into comparing vectorized persistence summaries and can be used to design better feature-based learning models based on PDs.

1 Introduction

Topological data analysis (TDA) is attracting increasing interest among researchers in machine learning due to the power of capturing shapes and structure in data [15]. One tool in TDA is persistent homology, which captures the connected components, tunnels, and holes – in particular, the homology – at various scales. Persistent homology can be represented in a structure called the persistence diagram (PD) with the Wasserstein distance (or bottleneck distance) is traditionally used to compare PDs due to their stability with respect to perturbations of the input [5, 7].

In order to apply PDs in machine learning tasks and statistical analyses, different functional and vectorized persistence summaries have been proposed and used [1, 2, 3, 13, 16]. The distance metrics employed for comparing these summaries are different from the Wasserstein distance between the original PDs. While their findings have focused on proving stability or on ascertaining predictive power for these vectorized persistence summaries, it is not widely understand how distances between these summaries correlate with the distances between the original PDs.

In this paper, we present a comparison of distances on vectorized persistence summaries with the bottleneck and Wasserstein distance. To the best of our knowledge, this is the first time the distance

Topological Data Analysis and Beyond Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

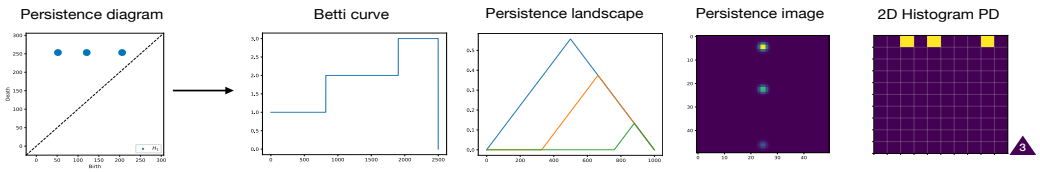


Figure 1: A persistence diagram and its four vectorized summaries. From left to right: the persistence diagram with three one-dimensional features; vectorization of PD as Betti curves; as persistence landscape, as persistence images in $\mathbb{R}^{50\times 50}$ and 2D Histogram.

relationship between PDs and their summaries has been studied. We believe this provides a new perspective on the use of TDA in machine learning.

2 Topological Descriptors

In this section, we begin with a brief introduction to persistent homology and vectorized summaries of persistence diagrams, but for a more thorough introduction to the topic, we refer the reader to [8, 11].

2.1 Topology and Persistent Homology

Given a dataset $\mathbb{X}$, we describe $\mathbb{X}$ by defining an associated topological space (e.g., a neighborhood graph, a Rips complex[19], or a sublevel set of a height function). However, a single topological space may not capture all relevant features of the data set. For that reason, persistent homology captures the data as some scale or time parameter changes (e.g., changing the distance parameter in a neighborhood graph or the height parameter in a sublevel set). Then, instead of a single topological space describing $\mathbb{X}$, we now have a *filtration*, or family of nested topological spaces indexed by an interval $[a, b] \subset \mathbb{R}$. If the changes are well-behaved, then a finite number of times witness a change in the *homology*, and the homology groups are always finitely generated. Moreover, features (or generators of homology groups) can be tracked through the time parameter, resulting in a finite set of birth-death pairs, which we call the persistence diagram. In particular, the diagram is a finite set of labeled points in the extended plane,[1] where a point $x = (x_1, x_2)$ represents the birth of a feature at time x_1 that dies going into time x_2. In addition, x has a label $\ell(x)$ corresponding to the dimension of the feature. In this paper, we focus on the comparison of diagrams and summaries of diagrams.

Two distances between persistence diagrams are the bottleneck (or interleaving) distance and the Wasserstein distance. Given two diagrams, D_1 and D_2, both distances can be defined by finding an optimal matching $M \subset D_1 \times D_2$, considering both matched points $(x, y) \in M$ as well as unmatched points $M^c \subseteq D_1 \cup D_2$. For $p > 0$, the p-Wasserstein distance between D_1 and D_2 is defined as:

$$W_p(D_1, D_2) := \inf_M \left(\sum_{(x,y)\in M} \|x - y\|_\infty^p + \sum_{x\in M^c} |x_2 - x_1|^p \right)^{\frac{1}{p}}, \tag{1}$$

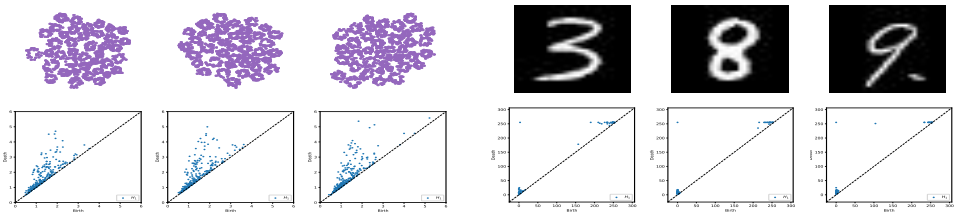
where $||\cdot||_\infty$ denotes the ∞-norm over the extended plane, and M ranges over all matchings between D_1 and D_2.[2] The bottleneck distance is the limit: $W_\infty(D_1, D_2) := \lim_{p\to\infty} W_p(D_1, D_2)$ [4, 7]. In this paper, we use distance metrics W_1 and W_∞.

2.2 Vectorized Summaries of Persistence Diagrams

Data in a PD is not amenable to many tasks; for example, the Fréchet mean is not unique [18]. One way to resolve this issue is to transform a PD into a vectorized summary, which can be easily used in machine learning tasks. In this section, we give an overview of several different vectorizations that can be used to summarize a persistence diagram D; see Figure 1. As these summaries are vectors, we

[1] The extended real line is the set: $\bar{\mathbb{R}} = \{-\infty, \infty\} \cup \mathbb{R}$, and the extended plane is $\bar{\mathbb{R}}^2$. For computational reasons, we often say that the diagram also includes all points on the diagonal (x, x) with infinite multiplicity.

[2] The second summation can be thought of as allowing points to match to the diagonal (the line $y = x$).



(a) Histology Simulation (b) MNIST

Figure 2: Examples and corresponding one-dimensional PDs (for visualization better to zoom in).Left: examples for Histology Simulation (top) and their PD via Rips filtration (bottom). Right: examples for MNIST (top) and their PD via sublevel filtration (bottom).

can compare two summaries using the L_p-distance; in particular, we use the L_2 or Euclidean distance between two summaries.

Betti Curves The *Betti curves* (BC's) are a $\mathbb{Z}$-indexed family of functions [10, 16, 17]. For each $z \in \mathbb{Z}$, the function $\beta_z \colon \mathbb{R} \to \mathbb{R}$ is defined by $\beta_z(t) = \#\{x = (x_1, x_2) \in D \mid \ell(x) = z \;\&\; x_1 \leq t \leq x_2\}$. To vectorize this, we choose a uniform grid over a closed interval and a finite set of dimensions.

Persistence Landscape The *Persistence Landscape* (PL) is a family of functions $\{\lambda_k \colon \mathbb{R} \to \mathbb{R}\}_{k \in \mathbb{Z}}$ defined by: $\lambda_k(t) := \sup\{m \geq 0 \mid \beta^{t-m,t+m} \geq k\}$, where $\beta^{i,j} := \#\{x = (x_1, x_2) \in D \mid x_1 \leq i \leq j < x_2\}$; see [3]. To vectorize the PL, we restrict these functions to a closed interval $[a, b] \subset \mathbb{R}$, and choose a uniform discretization to obtain a two-dimensional vector.

Persistence Image The *persistence intensity function* $\Phi(D)$ is a smoothing of the persistence diagram [1, 6]. One way to think of this is as a weighted kernel density estimate, or a density surface for the distribution of points in the diagram. Given a persistence diagram D, the *Persistence Surface* (same as *persistence intensity function*) uses $\Phi : \mathbb{R}^2 \to \mathbb{R}$ to turn the rotated D into a surface, that is defined as [1]: $\Phi(D) = \sum_{\mu \in D} \mathbf{w}(\mu)\phi_\mu(z)$, where $\phi_\mu(\cdot)$ is the Gaussian and $\mathbf{w}(\cdot)$ is a fixed piecewise linear weight function. The *Persistence Image* (PI) is a discretization of Φ, taking samples over a regular grid.[3]

2D Histogram Introduced in [12], the *2D Histogram* (2D-Histo) is an unsmoothed PI. It encodes the PD as a 2D histogram on an uniform with an extra diagonal cell, where each cell of the grid counts the number of points in D that fall in that cell.

In addition to an L_p-distance, we can also use an Earth mover's or optimal transport (OT) distance here. In particular, given two diagrams, we define this distance by snapping the points to the grid, then computing the p-Wasserstein distance between the snapped diagrams.

3 Results

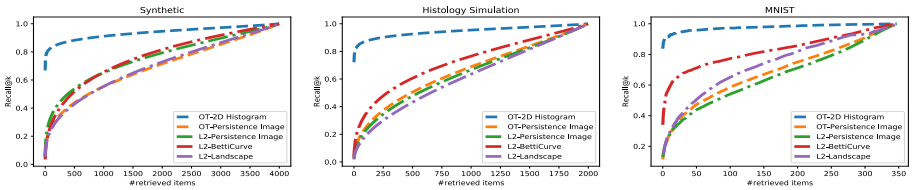
In order to compare distances between vectorized persistence summaries to distances between the original PDs, we consider three datasets: synthetic, histology simulation, and the MNIST dataset.

Synthetic Data This is a collection of 4,000 persistence diagrams with 20 random generated features (i.e. points), for each feature, we first create two random numbers in the range [0,1] then select the larger one as death, smaller one as birth for one point in the PD.

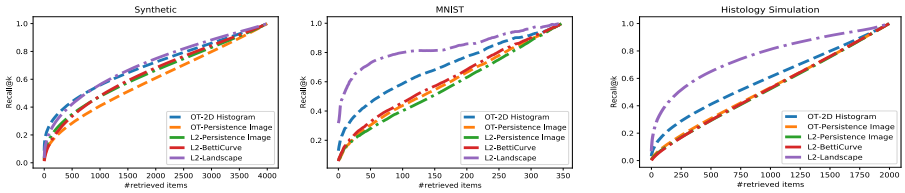
Histology Simulation This is a set of 2,000 persistence diagrams via Rips filtration (examples see Figure 2a). The original input data is from the gland generator [9], which is an application to create random 2-D points that form microscopic prostate-tissue images.

MNIST Data This dataset consists of 350 persistence diagrams via sublevel filtration of a subset of MNIST images of handwritten digits of size 28×28 [14]. Figure 2b shows examples of MNIST digits and their PDs.

[3]To be consistent with [1], we rotate the persistence diagram by $\pi/4$ before computing the persistence image.



(a) Comparison with Wasserstein Distance



(b) Comparison with Bottleneck Distance

Figure 3: Recall rate performance on different summaries distance and PD distance. Each line indicates the summaries with its metric. In (a), the 2D Histogram in OT metric has the best performance for three datasets. In (b), the persistence landscape in L_2 overcomes other summaries. The x-axis refers to the range of datasets, and the value of the y-axis aggregated by the mean of overall k, where k equals the x-axis(#retrieved items).

3.1 Evaluation Measure

Let Q be a given set of persistence diagrams, and let S be a function mapping each $q \in Q$ to a vectorized summary of q (e.g., a Betti curve). Then for any query $q \in Q$ we calculate the distance to each other $p \in Q$, $p \neq q$, using the vectorized summary's metric $d_s(S(q), S(p))$.

For any $k \in [1, |Q|]$, let $R_{d_s}(q, k) \subseteq Q$ consist of the k-nearest neighbors to q using $d_s(S(\cdot), S(\cdot))$. We compare this set to the set $R_{W_p}(q, k) \subseteq Q$ which consists of all k-nearest neighbors to q in $W_p(\cdot)$, here $p = 1, \infty$; using the original persistence diagrams (not summaries). We then evaluate the quality of d_s using the $Recall@k$:

$$Recall@k = \frac{1}{k|Q|} \sum_{q \in Q} |R_{d_s}(q, k) \cap R_{d_w}(q, k)| \ . \tag{2}$$

We chose this particular evaluation measure because it can directly reflect the k-nearest neighbor consistency of the vectorized persistence summaries.

3.2 Implementation Details

In all our experiments we only compare the 1-dimensional features of PD (i.e. cycle), and the test implemented by adapting the classical parameter settings based on the recommendations of the original papers (the size of all vectorized persistence summaries is 2500), using `Gitto-tda` [4] and `GUDHI`[5].

3.3 Performance Analysis

In Figure 3a, the 2D-Histo with Optimal Transport (OT) distance metric plays the best performance. It can be reasonably assumed that the extra diagonal cell of 2D-Histo influenced the comparison of PDs. To verify this assumption, we also compared PI with OT metric and the result is similar to PI in the L2 metric, which supported the observation that the special consideration of diagonal is non-trivial. Especially for PI takes the weight function to discard the points on the diagonal. To further understand the distance correlation of summaries and PDs, we apply the same test for bottleneck distance and the results are shown in Figure 3b.

[4] https://github.com/giotto-ai/giotto-tda

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

4 Discussion

This study is the first step towards enhancing our understanding of persistence diagram comparison. Unexpectedly, different vectorized persistence summaries have variant distance correlation in terms of original PD distance. In general, we have found how the diagonal role of PD comparison affects the distance correlation result in vectorized persistence summaries. This finding could be exploited in other real-world data where a distance of PDs is needed. We hope that our work is valuable for determining the vectorized persistence summaries that can be better represented original persistence diagram. Future work will focus on fast computing of the distance of PDs and applying it to machine learning tasks.

Acknowledgments and Disclosure of Funding

This material is based upon work supported by the National Science Foundation and the National Institutes of Health under Grant No. NSF DMS 1664848 and 1664858.

References

[1] ADAMS, H., EMERSON, T., KIRBY, M., NEVILLE, R., PETERSON, C., SHIPMAN, P., CHEPUSHTANOVA, S., HANSON, E., MOTTA, F., AND ZIEGELMEIER, L. Persistence images: A stable vector representation of persistent homology. *The Journal of Machine Learning Research 18*, 1 (2017), 218–252.

[2] BERRY, E., CHEN, Y.-C., CISEWSKI-KEHE, J., AND FASY, B. T. Functional summaries of persistence diagrams. *APCT 4* (2020), 211–262. Also available at arXiv:1804.01618.

[3] BUBENIK, P. Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research 16*, 1 (2015), 77–102.

[4] CHAZAL, F., DE SILVA, V., GLISSE, M., AND OUDOT, S. *The structure and stability of persistence modules*. Springer, 2016.

[5] CHAZAL, F., DE SILVA, V., AND OUDOT, S. Persistence stability for geometric complexes. *Geometriae Dedicata 173*, 1 (2014), 193–214.

[6] CHEN, Y.-C., WANG, D., RINALDO, A., AND WASSERMAN, L. Statistical analysis of persistence intensity functions. *arXiv preprint arXiv:1510.02502* (2015).

[7] COHEN-STEINER, D., EDELSBRUNNER, H., AND HARER, J. Stability of persistence diagrams. *Discrete & Computational Geometry 37*, 1 (2007), 103–120.

[8] EDELSBRUNNER, H., AND HARER, J. *Computational Topology: An Introduction*. AMS, Providence, RI, 2010.

[9] FASY, B. T., PAYNE, S., SCHENFISCH, A., SCHUPBACH, J., AND STOUFFER, N. Simulating prostate cancer slide scans. In preparation, 2020.

[10] GAMEIRO, M., MISCHAIKOW, K., AND KALIES, W. Topological characterization of spatial-temporal chaos. *Physical Review E 70*, 3 (2004), 035203.

[11] GHRIST, R. W. *Elementary Applied Topology*. CreateSpace, 2014.

[12] LACOMBE, T., CUTURI, M., AND OUDOT, S. Large scale computation of means and clusters for persistence diagrams using optimal transport. In *Advances in Neural Information Processing Systems* (2018), pp. 9770–9780.

[13] LAWSON, P., SHOLL, A. B., BROWN, J. Q., FASY, B. T., AND WENK, C. Persistent homology for the quantitative evaluation of architectural features in prostate cancer histology. *Scientific reports 9*, 1 (2019), 1–15.

[14] LECUN, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/* (1998).

[15] OBAYASHI, I., HIRAOKA, Y., AND KIMURA, M. Persistence diagrams with linear machine learning models. *Journal of Applied and Computational Topology 1*, 3-4 (2018), 421–449.

[16] RIECK, B., SADLO, F., AND LEITTE, H. Topological machine learning with persistence indicator functions. *arXiv preprint arXiv:1907.13496* (2019).

[17] ROBINS, V. Computational topology for point data: Betti numbers of α-shapes. In *Morphology of Condensed Matter*. Springer, 2002, pp. 261–274.

[18] TURNER, K., MILEYKO, Y., MUKHERJEE, S., AND HARER, J. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry 52*, 1 (2014), 44–70.

[19] VIETORIS, L. Über den höheren zusammenhang kompakter räume und eine klasse von zusammenhangstreuen abbildungen. *Mathematische Annalen 97*, 1 (1927), 454–472.