Persistent Homology and Machine Learning

Primož Škraba
Artificial Intelligence Laboatory[†], Jozef Stefan Institute
E-mail: primoz.skraba@ijs.si

Keywords: persistent homology, topological data analysis, overview

Received: March 27, 2018

In this position paper, we present a brief overview of the ways topological tools, in particular persistent homology, has been applied to machine learning and data analysis problems. We provide an introduction to the area, including an explanation as to how topology may capture higher order information. We also provide numerous references for the interested reader and conclude with some current directions of research.

Povzetek: V tem članku predstavljamo pregled topoloških orodij, predvsem vztrajno homologijo, ki je uporabna na področju strojnega učenja in za analizo podatkov. Začnemo z uvodom v področje in razložimo, kako topologija lahko zajame informacije višjega reda. Članek vsebuje tudi reference na pomembna dela za zainteresiranega bralca. Zaključimo s trenutnimi smernicami raziskav.

1 Introduction

Topology is the mathematical study of spaces via connectivity. The application of these techniques to data is aptly named topological data analysis (TDA). In this paper, we provide an overview of one such tool called persistent homology. Since these tools remain unfamiliar to most computer scientists, we provide a brief introduction before providing some insight as to why such tools are useful in a machine learning context. We provide pointers to various successful applications of these types of techniques to problems where machine learning has and continues to be used.

We begin with a generic TDA pipeline (Figure 1). The input is a set of samples, usually but not always embedded in some metric space. Based on the metric and/or additional functions (such as density), a multiscale representation of the underlying space of data is constructed. This goes beyond considering pairwise relations to include higher-order information. Persistent homology is then applied. This is a tool developed from algebraic topology, which summarizes the whole multiscale representation compactly in the form of a persistence diagram. This compact representation can then be applied to various applications.

The goal of this paper is to provide a brief overview and introduce the main components in the TDA pipeline.

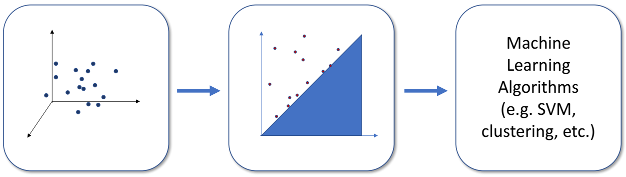

Figure 1: The TDA pipeline - taking in a points in in sime metric space along with potentially other information, the data is turned into a compact representation called a persistence diagram. This summary can then be input into machine learning algorithms rather than the raw point cloud.

2 Simplicial complexes

Representations of the underlying space are built up simple pieces glued together. There are many different approaches to this, however the simplest is perhaps the *simplicial complex*. A *simplex* is the convex combination of k points. A single point contains only itself, an edge is the convex combination of two points, three points make a triangle, four points a tetrahedron and so on (see Figure 2). More generally, a k-dimensional simplex is the convex combination of $(k+1)$ points. Just as an edge in a graph represents a pairwise relationship, triangles represent ternary relationships and higher dimensional simplices higher order relations. A graph is an example of a one-dimensional complex, as it represents all pairwise information - all higher order information is discarded. As we include higher dimensional simplices, we include more refined information yielding more accurate models. Note that these models need to not exist in an ambient space (i.e. may not be embedded), but rather represents connectivity information. The geometric realization of simplicial complexes has a long history of study in combinatorics but we do not address it here.

There are three main obstacles to this type of modeling. The first is lack of data. While it may be counterintuitive, in the age of big data we are often still faced with a lack of data. This is due to the non-uniformity and non-homogeneity of data. It may not make sense to consider 10-way relationships, if this data is only available for a small

[†] ARRS Project TopRep N1-0058

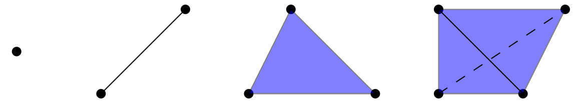

Figure 2: Simplicies come in different dimension. From left to right, a vertex is 0-dim, an edge is 1-dim, a triangle 2-dim and a tetrahedron is 3-dim.

subset of data. The second is computation. As we consider higher order relationships, there is often a combinatorial blow-up as one must consider all k-tuples, leading to preprocessing requirements which are simply not feasible. The final obstacle is interpretability. While we can understand a simplex locally, understanding the global structure becomes increasingly challenging.

This is the starting point for the tools we discuss below. Much of the effort of machine learning on graphs is trying to understand the qualitative properties of an underyling graph. This is often done by computing statistical features on the graph: degree distributions, centrality measures, diameter, etc. To capture higher order structure, we require a different set of tools. First, we note that a collection of simplices fit together. Just as in a graph, edges can only meet at an edge, simplices can only be glued together along lower dimensional simplices, e.g. triangles meet along edges or at a vertex. This represents a constraint on how simple building blocks (e.g. simplices) can be glued together to form a space. While this does not seriously limit the resulting spaces which can be represented, it does give us additional structure.

The starting point for the introduction is to describe the gluing map, called the *boundary operator*. For each k-simplex it describes the boundary as a collection of $k-1$ simplices. For example, the boundary of an edge consists of its two end points, the boundary of a triangle consists of its three edges (Figure 3). This can be represented as a matrix with the columns representing k-simplices and the rows $k-1$ simplices, which we denote ∂_k. The k-dimensional *homology* can be defined as

$$H_k = \frac{\text{ker } \partial_k}{\text{im } \partial_{k+1}}$$

The kernel is simply the collection of k-simplices which form the nullspace of the matrix which correspond to *cycles* (note that this agrees with the notion of graph-theoretic cycles). We the disregard all such cycles which bound regions filled-in by higher dimensional simplices. What remains is the numner of k-dimensional holes in the space. Specifically, 0-dimensional homology corresponds to the number of connected components, 1-dimensional homology the number of holes and so forth. The k-th Betti number, β_k is the number of independent such features. This is analogous to the rank of a matrix describing the number of basis elements a vector space has. This yields a qualitative description of the space. For a more complete introduction to homology, we recommend the book by Munkres [24] or the more advanced book by Hatcher[18]. An alternative introduction which also includes persistent homology (described in the following section) can be found in Edlesbrunner and Harer[13]. Our goal here is to point out the intuition behind simplicial complexes and one approach to describing them qualitatively. We do note that the algorithms and implementations are readily available [2, 19, 25, 23] and can often be interpreted through linear algebra.

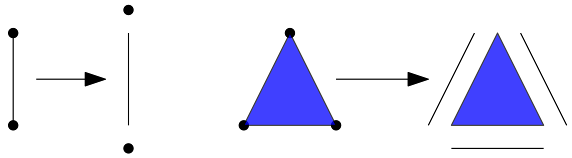

Figure 3: Simplicies are glued together in a specific way with each simplex is glued to lower dimensional simplices, called its boundary. Here we show an edge has 2 verticies as its boundary and a triangle has three edges as its boundary.

3 Persistent homology

One problem with homology and topological features in general is that they are unstable. Adding a point to a space changes the number of components and the correspoding Betti number. This would make it seems as though this technique were not suitable for the study of data. A key insight from [14, 39], is that we need not look at a single space but rather a sequence of spaces, called a *filtration*. This is an increasing sequence of nested spaces, which appears often when dealing with data.

$$\emptyset \subseteq X_0 \subseteq X_1 \subseteq \ldots \subseteq X_N$$

For example a weighted graph can be filtered by the edge weights. Perhaps the most ubiquitous example is a finite metric space, where the space is a complete graph and the weights are distances. This occurs whenever the notion of a “scale" appears, *Persistent homology* is the study of how qualitative features evolve over parameter choices. For example, the number of components is monotonically decreasing as we connect points which are increasingly far away. This is in fact precisely *single linkage clustering*. Higher dimensional features such as holes can appear and disappear at different scales.

The key insight is that the evolution of features over parameter choices can be encoded compactly in the form of a barcode or persistence diagram (Figure 4). We do not go into the algebraic reasons why this exists, rather we concentrate on its implications. An active research area has been to extend this to higher dimensional parameter spaces [6, 22, 34], but has remained a challenging area. We refer the reader to [13] for introductions to persistent homology and its variants. For the next section, rather than consider a persistence diagram rather than a barcode. Here each bar is mapped to a point with the starting point of the bar as the x-coordinate and the end point as the y-coordinate.

Consider a function on a simplicial complex, $f : K \to \mathbb{R}$ where we define the filtration as the *sublevel set* $f^{-1}(-\infty, \alpha]$. That is, we include all simplices with a lower function value. As we increase α, the set of simplicies with a lower function value only grows, hence we only add simplices. Therefore, we obtain an increasing sequence of topological spaces, i.e. a filtration. Define $X_\alpha := f^{-1}(-\infty, \alpha]$, then

$$X_{\alpha_1} \subseteq X_{\alpha_2} \subseteq \cdots \subseteq X_{\alpha_n} \qquad \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n$$

As another example in a metric space, we include all edges which represent a distance less than α. Consider a perturbed metric space, giving rise to a different function g. The following theorem establishes stability - that if the input (in this case, the function) does not change much, the output should not change much.

Theorem 1 ([11]). *Let K be two simplicial complexes with two continuous functions $f, g : X \to \mathbb{R}$. Then the persistence diagrams* $\mathrm{Dgm}(f)$ *and* $\mathrm{Dgm}(g)$ *for their sublevel set filtrations satisfy*

$$d_B(\mathrm{Dgm}(f), \mathrm{Dgm}(g))) \leq ||f - g||_\infty.$$

where $\mathrm{Dgm}(\cdot)$ represents the persistence diagram (i.e. a topological descriptor which is a set of points in $\mathbb{R}^2$) and $d_B(\cdot)$ represents bottleneck distance. This is the solution to the optimization which constructs a matching between the points in two diagrams which minimizes the maximum distance between matched points. While it is difficult to overstate the importance of this result, it does have some drawbacks. In particular the bound is in terms of the ∞-norm which in the presence of outliers can be very large. Recently this result has been specialized to Wasserstein stability, which is a much stronger result (albeit in a more limited setting).

Theorem 2 ([36]). *Let $f, g : K \to \mathrm{R}$ be two functions. Then $W_p(\mathrm{Dgm}(f), \mathrm{Dgm}(g)) \leq \|f - g\|_p$.*

Wasserstein distance is common in the machine learning and statistics literature as it is a natural distance between probability distributions. This recent result indicates that the distances between diagrams is indeed more generally stable and so suitable for applications. Stability has become an area of study in its own right and we now have a good understanding of the types of stabilty we can expect. The literature is too vast to list here so we limit ourselves to a few relevant pointers [3, 8].

4 Topological features

Here we describe some applications of persistence to machine learning problems. The key idea is to use persistence diagrams as feature vectors as input further machine learning algorithms, There are several obstacles to this. The most important is that the space of persistence diagrams is quite pathological. The first approach to move around this are *persistence landscapes* [4]. This lifts persistence diagrams into a Hilbert space which allows them to be fed into most standard machine learning algorithms. This has been followed up by rank functions [33], as well as several kernels [30], More recently, there has been work on learning optimal functions of persistence diagrams using deep learning [20].

There has also been significant work on the statistical properties of persistence diagrams and landscapes [16], including bootstrapping techniques [9].

These techniques have been applied to a number of application areas. Perhaps most extensive is in geometry processing. Combined with local features such as curvature or features based on heat kernels, different geoemtric structure can be extracted including symmetry [26], segmentation [35], and shape classification and retrival [7].

Another application area where persistence diagrams have been found to be informative are for biology, especially for protein docking [1] and modelling pathways in the brain [17]. The final application area we mention is material science. This is an area where machine learning has not yet been applied extensively. Partially due to the fact that the input is of a significantly different flavor than that which is typical in machine learning. For example, standard image processing techniques do not work well with scientific images such as electron microscope images. By using topological summaries, the relevant structure is well-captured [32, 21]. This area is still in the early stages with many more exciting developments expected.

We conclude this section by noting that persistence diagrams are not the only topological features which have been applied. Originally, the *Euler curve* was applied to fMRIs [38][1]. This feature has been extensively studied in the statistics literature, but is provably less informative than persistence diagrams - although it is far more computationally tractable. In addition to fMRI, it has been applied to various classification problems [31].

5 Other applications

In additon to providing a useful summary and features for machine learning algorithms, a second direction of interest is the map back to data. This inverse probelm is very difficult and can often be impossible in general. Nonetheless, the situation is often not as hopeless as it would seem. Some of the first work in this direction is re-interpreting single linkage clustering through the lens of persistence [10]. While it is well known that single linkage clusters are unstable, it is possible to use persistence to show that there exist stable parts of the clusters and a “soft" clustering algorithm can be developed to stabilize clusters, where each data point is assigned a probability that it is assigned to a given cluster. A current direction of research is to find similar stable representations in the data for higher dimensional structures (such as cycles).

[1] We note that this is where the term *topological inference* first used

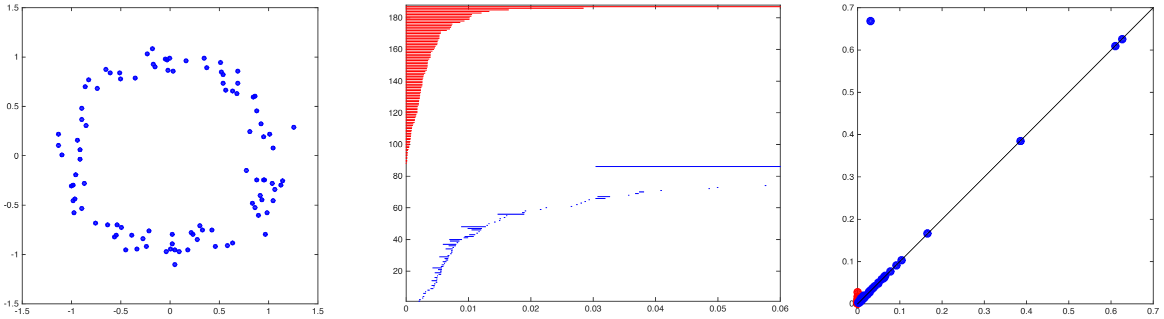

Figure 4: Persistence in a nutshell. Given input points (left), we compute a barcode (middle). which shows how long features live. The red shows the lifetimes of when components merge, while the blue bars show 1-dimensional holes. We can map each bar to a point by taking the start and end as the x and y coordinates respectively giving us the persistence diagram (right). Here we see that the big whole in the middle of the data set appears as a prominent feature (the blue dot far from the diagonal on the right).

A related problem is one of parameterization. That is, find intrinsic coordinates describing the data, extending successful techniques in dimensionality reduction, This includes linear methods such as PCA and MDS as well as non-linear methods such as ISOMAP and LLE. The first such work coordinizaed the space of textures using a Klein bottle as the underlying model [28] - a topological model found a few years prior [5]. This was however built by hand. The first class of general methods is first to map cicrular coordinates to data [12]. This is particularly useful when dealing with recurrence in time-varying systems. Recurrence (including periodicity) is naturally modeled by an angle, Combining persistence with least-squares optimization provides an automatic pipeline to finding such coordinates. This was applied to characterizing human motions such as different walks and other activities [37]. Further work has shown how to construct coordainte systems for higher dimensional structures based on the projective plane [27].

The final direction we consider is to encode topological constraints in machine learning algorithms. In [29] topological priors were used to aid in parameter selection. For example, the reconstruction of a racetrack should have one component and one hole (the main loop). Computing the persistence with respect to a reconstruction parameter (e.g. bandwith of a kernel) can allow us to choose a parameter value where the reconstruction has the desired topological “shape." The encoding of topological constraints is still in the very early stages but has the potential to provide a new type of regularization to machine learning techniques.

6 Discussion

Topological data analysis and applications of topology are still in their early stages. Various efforts to bridge the gap between algebraic topology and statistics (and probability) has made rapid progress over the last few years which has culminated in a dedicated R-package [15]. At the same time, increasingly efficient software exists for computing persistent homology exists, where now it is feasible to consider billions of points in low dimensions. This is increasingly bridging the gap between theory and practice.

The area has undergone rapid development over the last 10 years and is showing no signs of slowing down. In terms of theory, the primary question drinving the community is the notion of multi-dimensional or multi-parameter persistence, where the computational obstacles are much more daunting. Nonetheless, progress is being made. Success promises to further reduce the need and dependence on parameter tuning.

The combination of deep learning techniques with topological techniques promises to provide new areas of applications as well as potentially performance. These methods are primarily complementary allowing them to build on each other. In conclusion, while obstacles remain, the inclusion of topological techniques into the machine learning toolbox is rapidly making progress.

References

[1] Pankaj K Agarwal, Herbert Edelsbrunner, John Harer, and Yusu Wang. Extreme elevation on a 2-manifold. *Discrete &amp; Computational Geometry*, 36(4):553–572, 2006.

[2] U Bauer. Ripser. https://github.com/Ripser/ripser, 2016.

[3] Ulrich Bauer and Michael Lesnick. Induced matchings of barcodes and the algebraic stability of persistence. In *Proceedings of the thirtieth annual symposium on Computational geometry*, page 355. ACM, 2014.

[4] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16(1):77–102, 2015.

[5] Gunnar Carlsson, Tigran Ishkhanov, Vin De Silva, and Afra Zomorodian. On the local behavior of spaces of natural images. *International journal of computer vision*, 76(1):1–12, 2008.

[6] Gunnar Carlsson and Afra Zomorodian. The theory of multidimensional persistence. *Discrete &amp; Computational Geometry*, 42(1):71–93, 2009.

[7] Frédéric Chazal, David Cohen-Steiner, Leonidas J Guibas, Facundo Mémoli, and Steve Y Oudot. Gromov-hausdorff stable signatures for shapes using persistence. In *Computer Graphics Forum*, volume 28, pages 1393–1403. Wiley Online Library, 2009.

[8] Frédéric Chazal, Vin De Silva, Marc Glisse, and Steve Oudot. The structure and stability of persistence modules. *arXiv preprint arXiv:1207.3674*, 2012.

[9] Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. On the bootstrap for persistence diagrams and landscapes. *arXiv preprint arXiv:1311.0376*, 2013.

[10] Frédéric Chazal, Leonidas J Guibas, Steve Y Oudot, and Primoz Skraba. Persistence-based clustering in riemannian manifolds. *Journal of the ACM (JACM)*, 60(6):41, 2013.

[11] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete &amp; Computational Geometry*, 37(1):103–120, 2007.

[12] Vin De Silva, Dmitriy Morozov, and Mikael Vejdemo-Johansson. Persistent cohomology and circular coordinates. *Discrete &amp; Computational Geometry*, 45(4):737–759, 2011.

[13] Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

[14] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 454–463. IEEE, 2000.

[15] Brittany Terese Fasy, Jisu Kim, Fabrizio Lecci, and Clément Maria. Introduction to the r package tda. *arXiv preprint arXiv:1411.1830*, 2014.

[16] Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, Aarti Singh, et al. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339, 2014.

[17] Margot Fournier, Martina Scolamiero, Mehdi Gholam-Rezaee, Hélène Moser, Carina Ferrari, Philipp S Baumann, Vilinh Tran, Raoul Jenni, Luis Alameda, Karan Uppal, et al. M3. topological analyses of metabolomic data to identify markers of early psychosis and disease biotypes. *Schizophrenia Bulletin*, 43(suppl_1):S211–S212, 2017.

[18] Allen Hatcher. *Algebraic topology*. 2002.

[19] Gregory Henselman and Robert Ghrist. Matroid filtrations and computational persistent homology. *arXiv preprint arXiv:1606.00199*, 2016.

[20] Christoph Hofer, Roland Kwitt, Marc Niethammer, and Andreas Uhl. Deep learning with topological signatures. In *Advances in Neural Information Processing Systems*, pages 1633–1643, 2017.

[21] Yongjin Lee, Senja D Barthel, Paweł Dłotko, S Mohamad Moosavi, Kathryn Hess, and Berend Smit. Quantifying similarity of pore-geometry in nanoporous materials. *Nature Communications*, 8, 2017.

[22] Michael Lesnick. The theory of the interleaving distance on multidimensional persistence modules. *Foundations of Computational Mathematics*, 15(3):613–650, 2015.

[23] Dmitriy Morozov. Dionysus. *Software available at http://www. mrzv. org/software/dionysus*, 2012.

[24] James R Munkres. *Elements of algebraic topology*, volume 4586. Addison-Wesley Longman, 1984.

[25] Vidit Nanda. Perseus: the persistent homology software. *Software available at http://www. sas. upenn. edu/˜ vnanda/perseus*, 2012.

[26] Maks Ovsjanikov, Quentin Mérigot, Viorica Pătrăucean, and Leonidas Guibas. Shape matching via quotient spaces. In *Computer Graphics Forum*, volume 32, pages 1–11. Wiley Online Library, 2013.

[27] Jose A Perea. Multi-scale projective coordinates via persistent cohomology of sparse filtrations. *arXiv preprint arXiv:1612.02861*, 2016.

[28] Jose A Perea and Gunnar Carlsson. A klein-bottle-based dictionary for texture representation. *International journal of computer vision*, 107(1):75–97, 2014.

[29] Florian T Pokorny, Carl Henrik Ek, Hedvig Kjellström, and Danica Kragic. Topological constraints and kernel-based density estimation. *Advances in Neural Information Processing Systems*, 25, 2012.

[30] Jan Reininghaus, Stefan Huber, Ulrich Bauer, and Roland Kwitt. A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4741–4748, 2015.

[31] Eitan Richardson and Michael Werman. Efficient classification using the euler characteristic. *Pattern Recognition Letters*, 49:99–106, 2014.

[32] Vanessa Robins, Mohammad Saadatfar, Olaf Delgado-Friedrichs, and Adrian P Sheppard. Percolating length scales from topological persistence analysis of micro-ct images of porous materials. *Water Resources Research*, 52(1):315–329, 2016.

[33] Vanessa Robins and Katharine Turner. Principal component analysis of persistent homology rank functions with case studies of spatial point patterns, sphere packing and colloids. *Physica D: Nonlinear Phenomena*, 334:99–117, 2016.

[34] Martina Scolamiero, Wojciech Chachólski, Anders Lundman, Ryan Ramanujam, and Sebastian Öberg. Multidimensional persistence and noise. *Foundations of Computational Mathematics*, 17(6):1367–1406, 2017.

[35] Primoz Skraba, Maks Ovsjanikov, Frederic Chazal, and Leonidas Guibas. Persistence-based segmentation of deformable shapes. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 45–52. IEEE, 2010.

[36] Primoz Skraba and Katharine Turner. Wasserstein stability of persistence diagrams. submitted to the Symposium of Computational Geometry 2018.

[37] Mikael Vejdemo-Johansson, Florian T Pokorny, Primoz Skraba, and Danica Kragic. Cohomological learning of periodic motion. *Applicable Algebra in Engineering, Communication and Computing*, 26(1-2):5–26, 2015.

[38] Keith J Worsley, Sean Marrett, Peter Neelin, Alain C Vandal, Karl J Friston, Alan C Evans, et al. A unified statistical approach for determining significant signals in images of cerebral activation. *Human brain mapping*, 4(1):58–73, 1996.

[39] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete &amp; Computational Geometry*, 33(2):249–274, 2005.