1 Things I Have Done

The major reason I'm interested in TDA (topological data analysis), and, in particular, PH (persistent homology), is that they formalize our intuition about shape, resolution, invariance, and etc. into topological theories. Unlike other popular methodologies like DL (deep learning), with which many people tend to compare TDA, TDA is in some sense explicable and understandable. Besides the sole fact that such understandability itself is already extremely valuable and reassuring, TDA could possibly reveal features and accomplish tasks that are otherwise not achievable through other methodologies like DL.

However, one fundamental weakness of TDA is that it is not problem solving oriented. Most methods in DL are designed specifically for solving certain problems. For example, CNN (convolutional neural networks) was designed to simulate human perception and processing of images, and therefore, shows very good performance on image related problems. In contrast, as far as I'm concerned, TDA was not invented for a specific target problem. Thus, to put TDA into application is to find a question to a solution, which, to some, may seem inefficient, dubious, and bound to fail. In fact, I barely found any successful cases of TDA in any industry, except AYASDI, which, according to some unverified sources, has already turned to more mature methods like DL, and rebranded itself as "enterprise AI".

Nonetheless, the elegance of the theory, along with the flourishing academic research of the application of TDA to various fields (though far from actual use in industry), provides me with a motivation to learn more about it, and hopefully find a place where it can be naturally and effectively applied. Below is a list of things I have done in the last week. Everything can be found on this Github repository https://github.com/SiyangJ/Topological-Data-Analysis.

1. Learned about what TDA is in general, through searching on internet, and in particular, AYASDI's seminars on youtube.

2. Substantially learned TDA through course INF556 given by Steve Oudot, professor at École Polytechnique, and research scientist at Inria. In particular, I implemented all the algorithms by myself, and understood their solution algorithm as well. However, I did not delve into the theoretic details of certain certain characterization and proof, for example, the functoriality of homology, and the stable theories of Wasserstein and bottleneck distances.

3. TDA is more like a framework than a specific theory. Different approaches in TDA are listed below.

   (a) Mapper algorithm is the core of TDA products of AYASDI. However, it's not open source. Although some open source libraries and softwares do exist, I feel like it's not a good idea to compete with the founders of TDA. Besides, the Mapper algorithm of AYASDI and the related products seem to have very limited application and success. Often, Mapper provides certain insights that still need human interpretation.

(b) Persistent homology. The most popular and well-studied branch. Various theories on the generation of complexes, function used for filtration, stability of the results, and etc. have been discussed. Moreover, a wide variety of open source libraries are available for the computation. Therefore, my main focus will be PH.

(c) Euler calculus, cellular sheaves, etc. They are not as popular and as well-established.

4. Consolidate my understanding through a very useful paper "A roadmap for the computation of persistent homology"[1]. Together with its references, I basically figured out all the theories I think I need to know, and also acquainted myself with the basic and actual algorithms of PH.

5. Read some recent papers of the application of TDA and PH in different fields. Some interesting ones include:

   (a) Tumor segmentation. This 2017 paper [2] applies 1) persistent homology to extract features, 2) CNN to classify, and 3) Ensemble Random Forest to tumor segmentation problems. Specifically, they use CNN to divide each whole-slide image (WSI) to patches, then apply PH to create PH profiles, and finally use KLD (Kullback-Leibler divergence) as a distance function between the profiles to perform k-nearest neighbor (kNN).

   (b) Classification of histology images. This 2018 paper [3] uses persistence landscape (PL) and persistence image (PI) to train machine learning models. Persistence image is a very recent development, and it's still under active research. How to provide quantifiable statistical interpretation to PH and its typical representations, like barcode and persistence diagram, has always been a problem. This paper seems to successfully use PI and PL to interpret, and therefore provide further statistical analysis.

   (c) Brain artery tree (3D object). I haven't thoroughly read this one. This 2016 paper [4] tries to extract topological features from brain artery trees by PH. They analyze the result with PL and other methods, and draw conclusions on correlation between certain development and age/sex.

2 Notes

In this section, I give some selected notes I took during past week. All the notes are displayed on the Github repository.

2.1 Simplicial Complexes

General idea is that simplicial complexes extend the notion of graph to include higher dimensional components in it.
Turn simplicial complexes to topological spaces: "embed" it into Euclidean space.

Embedding: map points to coordinates and all points affinely independent.
Induced map: $\hat{f} : K \to 2^{\mathbb{R}^d}, \{v_0, v_1, ..., v_r\} \mapsto Conv(f(v_0), ..., f(v_r))$
All embeddings $f : K \to \mathbb{R}^d, g : K \to \mathbb{R}^{d'}$, $\hat{f}$ and $\hat{g}$ are homeomorphic.
Underlying space, $|K|$, the image of K through embedding, unique up to homeomorphism.
Triangulability: X is triangulable if $\exists K, h : X \to |K|$, h is homeomorphism.
Simplicial map: combinatorial equivalence of continuous map.
Topological realization of a map: simplicial $f : K \to L$ induces continuous $|f| : |K| \to |L|$
Reverse is not necessarily true.
Simplicial Approximation: continuous $f : |K| \to |L|$ is homotopic to $|f'| : |K| \to |L|$, where $f' : K' \to L$, for some simplicial subdivision K' of K.
Intuition: by dividing the simplicial complex sufficiently many times, we can approximate the topological space.

2.2 Simplicial Homology

Orientation: Ordering of vertices, unique up to even permutation, negative by odd permutation.

2.2.1 Chains

K finite simplicial complex and k a fixed field. Given $r \in \mathbb{N}$, we are interested in the $k-$linear combinations (formal sums) of $r-$simplices in K.
The linear space of all the formal sums for each r is a chain space of $k-$chains.
Note: Seems that the field k can be relaxed to a ring. Can study later.

2.2.2 Boundary Operator

Remove one vertex from a simplex, we get a face of the simplex. The linear combination together with interchanging orientation is the boundary.
$\partial_r : C_r(K, k) \to C_{r-1}(K, k)$
$[v_0, ..., v_r] \mapsto \sum_{j=0}^{r} (-1)^j [v_0, ...\hat{v_j}..., v_r]$
Clearly ∂_r is distributive.
Note:$\partial^2 = 0$, and therefore is nilpotent. (Actually it's $\partial_r \circ \partial_{r+1} = 0$.

2.2.3 Homology Groups

Important motivation: want to find "cycles modulo the boundaries".
Note: need to further think about this.
r-cycles: $Z_r(K, k) := \ker\partial$
r-boundaries: $Z_r(K, k) := \text{im}\partial_{r+1}$
Homology group: $Z_r(K, k)/Z_r(K, k)$

2.2.4 Algorithm for Homology

Since it's a vector space, it's isomorphic to k^{β_r},
Note: when relaxed to ring, it's a module, and therefore we still have the fundamental theorem for modules to decompose to a torsion part and a torsion-free part. Still have some kind of β_r.
Note: everything is linear space / linear transformation.
Matrix form M_r of ∂_r for each r:
$\#K_r$ columns and $\#K_{r-1}$ rows, $\beta_r = \#K_r - \text{rank}M_r - \text{rank}M_{r+1}$, just compute ranks to get the dimension, which I have learned and don't want to go into details.

2.2.5 Morphisms

Operator on spaces: $H_r : K \mapsto H_r(K, k)$, which we want to extend to maps as well.
Idea of a functor? Some category stuff?
Chain level: simplicial map induces a chain map. Details omitted.
The chain map commutes with boundary operator.
Functoriality: temporarily skipped.

2.3 From simplicial complexes to topological spaces

Theorem: X triangulable, then, $\forall K, L, H_r(K, k) \simeq H_r(L, k)$.
Conclusion: homology groups of triangulable spaces, and the morphisms between them, are uniquely defined, for different ways of triangulation. Moreover, morphisms are invariant under homotopy.
Corollary: $X \sim Y \implies H_r(X, k) \simeq H_r(Y, k)$.
However, homology does not completely characterize the topology of a space. Thus still much weaker than homeomorphism.

2.4 Computation of Persistence Homology

This part is largely cited from [1].
How to keep track of how one feature "merges" to another?
Boundary matrix: the matrix representation of boundary operator.
We also need a total ordering compatible with the "filtration" in the following sense:

- a face of a simplex precedes the simplex.
- a simplex in K_i precedes simplices in K_j for $j > i$, and not in K_i.
  - this essentially means that we place the simplices by the order of "appearing".

2.4.1 Standard Algorithm

- Form the boundary matrix from the ordering.

- Reduction, which is essentially Gaussian elimination.
- Reading off intervals.
    1. some details to do.
    2. degree: $\mathrm{dg}(\sigma) =$ smallest number l s.t.$\sigma \in K_l$
    3. pair (σ_i, σ_j) gives $[\mathrm{dg}(\sigma_i), (\sigma_j))$
    4. unpaired extends to infinity.
- Output the barcode, $[\mathrm{dg}(\sigma_i), (\sigma_j))$ indicating a feature birth at $\mathrm{dg}(\sigma_i)$ and death at $\mathrm{dg}(\sigma_j)$

2.4.2 Complexity

In the worst case, which does exist, the algorithm has cubic complexity. Note that when sparse, not cubic.

2.5 Statistical Interpretation

2.5.1 Problems

1. Compare outputs with null model.
   How to compare the different results?
   How to evaluate the significance of the data?
2. Average over multiple realizations of a random model.

2.5.2 Statistical Analysis

Statistical methods for PH addressed first time in [5]. Three approaches.

1. Topological properties of random simplicial complexes, viewed as null models.
2. Properties of a metric space, whose points are persistence diagrams. Key point: define an appropriate distance function between diagrams. The main object is the persistence diagram, which is in some sense isomorphic to a barcode. Wasserstein distance and bottleneck distance are popular ways to defined distance.
3. "Features" of persistence diagrams. Key point: map the space of persistence diagrams to analyzable spaces (e.g. Banach spaces). Persistence landscape[6], using space of algebraic functions[7], kernalized techniques.

3 Current Challenges

3.1 Construction of Filtration

Except some special data types which have a specific choice of how to construct the filtration, for example, cubical complexes for images, and WRCF for weighted networks. For general point clouds, among the numerous types of comlexes, such as Čech, VR, alpha, witness, and etc, which complex is best suited for what, and how to study and characterize such difference?

3.2 Statistical Interpretation

Two major challenges are described in [1].

1. Quantitatively assessing the quality of the barcodes. Specifically, one cannot just say I'll disregard the "shorter" ones, and the left are the features. How short? What other information can be exploited, e.g. the variation in length?

2. The space of barcodes lacks geometric properties to define basic concepts, e.g. mean, median, etc. Current approaches include mapping the space of barcode to a finite metric space, for example, persistent landscape and persistent image. However, more potential can be explored.

3.3 Persistence Diagram

Not many tools can be used in applications for computation of Wasserstein / bottleneck distance, and other distance functions between persistence diagrams. Some existing ones include: Dionysus, Hera, TDA Package.

4 Plan for Next Week

I haven't decided on a specific data set to analyze. Among the various fields relevant to TDA/PH, I'm most familiar with and interested in images 2D/3D. Other topics like networks and dynamical systems are also attractive, but I'm afraid I severely lack the knowledge to deal with them. Therefore, I will limit my attention to images, and in particular medical images, which as far as I'm concerned are the most widely studied type of images with PH. To decide on a data set, I plan to first try on relatively small and well-studied benchmark data sets. Many such data sets and benchmarking methodologies are suggested by [1]. I also found this website `digitalpathologyassociation.org/whole-slide-imaging-repository`, which provides an extensive accessible repository of medical images.

This week due to certain technical issues, I have not been able to install the PH libraries on my computer. After getting my computer ready, I will try the different libraries and different data sets to see what is the most proper one.

Moreover, I plan to read more recent papers on the application and development of PH to keep up with the progress. The 2018 paper [3] makes use of persistent images, a rather novel technique not seen in many older documents, and achieves interesting results. It makes good sense to try new methods on old problems.

In the meantime, I intend to seek more potential of TDA/PH in itself. Most of the papers I read regard TDA/PH as a way to extract features for machine learning. Although this is indeed useful and interesting, I wonder if we could do more than that.

5 List of Interesting Works

5.1 Introduction and Summary

1. [1] Roadmap for computation of PH, 2017. Thorough review of computation and some theories of PH.

5.2 TDA

1. [8] Foundation for TDA. Discusses why topology and functoriality are essential for data analysis.
2. [9] Mapper Algorithm
3. [10] The Natural Image paper.

5.2.1 Persistent Homology

1. [11] Discrete Morse theory
2. [5] Statistical approach to persistent homology

5.3 Computation

1. [12] Computation of PH for cubical data, e.g. images
2. [13] Computational homology. In particular, covers cubical complexes and other complexes
3. [14] Javaplex
4. [15] Intro to R TDA
5. [16] Intro of persistence landscapes w/ algorithms.

References

[1] N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, and H. A. Harrington, "A roadmap for the computation of persistent homology," *EPJ Data Science*, vol. 6, p. 17, Aug 2017.

[2] T. Qaiser, Y.-W. Tsang, D. Epstein, and N. Rajpoot, "Tumor segmentation in whole slide images using persistent homology and deep convolutional features," in *Medical Image Understanding and Analysis* (M. Valdés Hernández and V. González-Castro, eds.), (Cham), pp. 320–329, Springer International Publishing, 2017.

[3] D. R. Chittajallu, N. Siekierski, S. Lee, S. Gerber, J. Beezley, D. Manthey, D. Gutman, and L. Cooper, "Vectorized persistent homology representations for characterizing glandular architecture in histology images," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 232–235, April 2018.

[4] P. Bendich, J. S. Marron, E. Miller, A. Pieloch, and S. Skwerer, "Persistent homology analysis of brain artery trees," *Ann. Appl. Stat.*, vol. 10, pp. 198–218, 03 2016.

[5] P. Bubenik and P. T. Kim, "A statistical approach to persistent homology," *Homology Homotopy Appl.*, vol. 9, no. 2, pp. 337–362, 2007.

[6] P. Bubenik, "Statistical topological data analysis using persistence landscapes," *J. Mach. Learn. Res.*, vol. 16, pp. 77–102, Jan. 2015.

[7] A. Adcock, E. Carlsson, and G. Carlsson, "The Ring of Algebraic Functions on Persistence Bar Codes," *ArXiv e-prints*, Apr. 2013.

[8] G. Carlsson, "Topology and data," *Bull. Amer. Math. Soc. (N.S.)*, vol. 46, no. 2, pp. 255–308, 2009.

[9] G. Singh, F. Memoli, and G. Carlsson, "Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition," in *Eurographics Symposium on Point-Based Graphics* (M. Botsch, R. Pajarola, B. Chen, and M. Zwicker, eds.), The Eurographics Association, 2007.

[10] G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian, "On the local behavior of spaces of natural images," *Int. J. Comput. Vis.*, vol. 76, no. 1, pp. 1–12, 2008.

[11] K. Mischaikow and V. Nanda, "Morse theory for filtrations and efficient computation of persistent homology," *Discrete & Computational Geometry*, vol. 50, pp. 330–353, Sep 2013.

[12] H. Wagner, C. Chen, and E. Vuçini, *Efficient Computation of Persistent Homology for Cubical Data*, pp. 91–106. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[13] T. Kaczynski, K. Mischaikow, and M. Mrozek, *Computational homology*, vol. 157 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[14] A. Tausz, M. Vejdemo-Johansson, and H. Adams, "JavaPlex: A research software package for persistent (co)homology," in *Proceedings of ICMS 2014* (H. Hong and C. Yap, eds.), Lecture Notes in Computer Science 8592, pp. 129–136, 2014. Software available at http://appliedtopology.github.io/javaplex/.

[15] B. T. Fasy, J. Kim, F. Lecci, and C. Maria, "Introduction to the R package TDA," *CoRR*, vol. abs/1411.1830, 2014.

[16] P. Bubenik and P. Dłotko, "A persistence landscapes toolbox for topological statistics," *Journal of Symbolic Computation*, vol. 78, pp. 91 – 114, 2017. Algorithms and Software for Computational Topology.