Swansea University
Swansea University Research Excellence Scholarships (SURES)

Research Project Proposal

Name: Nicholas Sale
Application Number: 00202749

Start date: Available any date from 1st July onwards (flexible).

1 Project Title:

Topological Data Analysis (TDA)
Working with Prof. Jeffrey Giansiracusa and/or Dr. Paweł Dłotko

Topological Data Analysis is a relatively new field of mathematics, sitting somewhere between mathematical data science and algebraic topology, popularised by Carlsson's 2009 paper [1]. As a framework for applying the tools of algebraic topology and homological algebra to real world data problems, TDA has had some successes in recent years, particularly in analysing medical data and material sciences data (see for example [2, 3]). As a field of study, TDA has re-framed and motivated a broad variety of interesting theoretical questions. For example, the requirement for better computational tools has motivated the investigation of algorithms based on discrete Morse theory [4]. The two main tools are persistent homology, introduced by Edelsbrunner, Letscher and Zomorodian [5], and the Mapper algorithm, introduced by Singh, Mémoli and Carlsson [6]. I am primarily interested in the former.

Swansea University would be an excellent place to study TDA due to the mathematics department's recent commitment to the Centre for Topological Data Analysis, and the dedicated resources provided by the Computational Foundry.

2 Aims and Objectives:

I have identified three different strands of investigation which I would be interested in pursuing. These are not exhaustive.

2.1 Statistical approaches to TDA

At the Dragon Applied Topology Conference hosted in Swansea in September 2018, I had a conversation with Professor Wojciech Chachólski (KTH), who was keen to express the necessity of a statistical approach to TDA. In a statistical approach, we would consider data as being generated by some underlying distribution or process, and the inferences drawn from TDA methods as estimators of topological quantities. This formulation is important for the uptake of topological methods in the mainstream data science and statistics communities, but still has a variety of open problems.

In his 2015 habilitation thesis [7], Bertrand Michel summarises the goals of a statistical approach to TDA as the following list of problems:

1. Proving consistency and studying the convergence rates of TDA methods.

2. Providing confidence regions for topological features and discussing the significance of the estimated topological quantities.

3. Selecting relevant scales at which the topological phenomenon should be considered, as a function of observed data.

4. Dealing with outliers and providing robust methods for TDA.

I am particularly interested in problems 2 and 4 for the case of persistent homology, and in general about when probabilistic guarantees can be made about topological inferences. I would like to investigate these problems from a homology inference perspective, as in the line of work stemming from Niyogi, Smale, Weinberger [8].

A closely-related problem, important for the uptake of TDA in the data science community, is the adaptation of non-linear topological features for use in existing statistical and machine learning methods. In particular, finding vectorisations of persistence diagrams which are stable with respect to noise, efficient to compute and/or have other desirable properties. While there are a number of frameworks which have been proposed and found successes, there is still scope for improving these methods or for new approaches altogether. I would also like to understand and characterise when a given representation might be better than another.

Potential objectives range from giving learning-theoretic bounds to developing feature maps, and include:

- Proving stability results for different topological descriptors with respect to different metrics.
- Deriving lower bounds on the number of data samples required to effectively reconstruct the homology of certain classes of spaces.
- Determining hardness of learning TDA-related problems in computational topology and geometry.
- Constructing vectorisations of persistence barcodes with desirable properties.
- Characterising learning problems and datasets to determine an appropriate representation for the output of persistent homology.

2.2 Algorithms and implementation of TDA tools

The viability of TDA as a toolbox for solving modern data science problems relies heavily on the ability to efficiently construct complexes and compute persistent homology for potentially very large datasets. While the computation of persistence barcodes given a filtered complex is well-studied (see [9] for a summary), there are various ongoing lines of research into computing filtered complexes from the original data. These include finding representative reductions of the data, the development of new data structures, and the use of different types of complex for specific types of data. There are also algorithmic considerations to the adaptation of barcodes for use in statistics as mentioned in the previous section 2.1. Besides looking at specific examples, these may even be generally amenable to information-theoretic arguments to give bounds on computational complexity. These are things that I would be interested in investigating. Given the variety of software packages for computing persistent homology with different underlying algorithms, I would like to understand when a given algorithm is likely to be faster than another.

Potential objectives include:

- Developing ways to exploit known structure in data to reduce the size of complexes used in persistent homology calculations.

- Developing improved algorithms for the computation of popular feature maps for persistence diagrams.
- Deriving bounds for the computational complexity of potential algorithms for using barcodes for different learning problems.
- Characterising datasets to determine the most appropriate algorithm to apply.

2.3 Multiparameter persistence

One of the core theoretical problems in persistent homology is the stumbling block presented by multiparameter persistent homology, when a filtration has two or more scale parameters. This is due to the non-existence of any complete discrete invariants (like the barcode in the 1-dimensional case) [10]. I would be interested in studying algebraic approaches to obtaining different descriptive invariants, each providing a partial picture of what's happening.

The objective would be to develop stable invariants and/or feature maps for multiparameter persistent homology.

3 Literature Review:

3.1 Statistical approaches to TDA

As mentioned in section 2.1, Bertrand Michel's thesis [7] provides a summary of the problems of interest in a statistical approach to TDA, as well as recent contributions towards solving these problems. In particular, he gives a theorem of Chazal et al. [11] which bounds the convergence rate of persistence homology with respect to the bottleneck distance between diagrams for a broad class of probability measures, but notes the difficulty of using this result to calculate confidence regions. He also mentions the difficulty in providing tight lower bounds for convergence when there is additive noise. Another contribution is the distance to measure (DTM) of Chazal et al. [12], designed to improve the robustness of persistent homology in the presence of outliers. Besides convergence results, it is shown how the DTM can be 'de-noised', however this relies on knowing in advance the distribution of the noise which means it's not immediately useful for applications.

One of the more influential frameworks for featurising persistence modules is Peter Bubenik's persistence landscapes [13], provided with efficient algorithms for calculation in collaboration with Paweł Dłotko [14]. This gives a direct mapping from persistence diagrams to a Banach space, with all the nice statistical properties of Banach-valued random variables. More recently, persistence landscapes have been adapted for use in the framework of Chevyrev, Nanda and Oberhauser's persistence paths approach [15] to feature maps. In this framework persistence barcodes are first mapped to a path in a vector space of bounded variation via some embedding (for example the embedding whose k^{th} component at time t is the k^{th} landscape function evaluated at t). Then the path is mapped to the tensor algebra of the vector space by integration. It's shown that the resulting map from barcodes to the tensor algebra is universal and characteristic. Moreover, while the integration can be prohibitively costly in high dimensions, the map is kernelised making it viable for kernel classifiers. However a tradeoff between stability, computability and discriminative power emerges in the choice of path embedding. While a few examples with different balances in these properties are given, it is not clear how to determine a good path embedding to use for a given application.

3.2 Algorithms and implementation of TDA tools

In *A roadmap for the computation of persistent homology* [9], Nina Otter et al. introduce persistent homology from a computational perspective, and benchmark the speed and memory usage

of a subset of the open-source libraries for persistent homology computation available at the time (2017). They conclude that there is no single best software for all datasets or complexes, but suggest different software for different problems. These recommendations are based on the empirical benchmarks, but looking at the performance on the different datasets, it is not necessarily clear why certain implementations perform better on some datasets and worse on others.

3.3 Multiparameter persistence

In their recent paper *Stratifying multiparameter persistent homology* [16], Harrington et al. summarise the approaches taken to study multiparameter persistent homology: via the rank invariant originally proposed by Carlsson and Zomorodian in [10], which is a direct generalisation of the barcode in the one-parameter case; by restricting the multiparemeter persistence module to a line to obtain a one-parameter module; and by seeking to efficiently compute presentations of the modules. They themselves investigate the r-parameter persistence module from the perspective of $\mathbb{N}^r$-graded commutative algebra. They show that the Hilbert series and the associated primes of a module provide invariants that describe components that persist forever in all directions and in one given direction (respectively). This framework, covering all those others mentioned, seems like a good setting to investigate algebraic invariants. However, while there are code listings in the paper, there is scope to investigate the efficiency of the algorithms used.

In a very recent preprint [17], Oliver Vipond generalises Bubenik's persistence landscapes [13] to multiparameter persistence modules, which stably represent the multiparameter rank invariant. These inherit the statistical properties of the original persistence landscapes.

4 Research Methodology:

The original formulation of persistent homology for $(\mathbb{N}, \leq)$-indexed diagrams of finite dimensional vector spaces [5] may be a good setting for developing algorithms, however it is likely that Bubenik and Scott's categorical formulation [18] will be an easier setting in which to work on developing a statistical framework. As mentioned in section 3.3, the context of $\mathbb{N}^r$-graded commutative algebra used by Harrington et al. [16] seems to be a useful setting for investigating multiparameter persistence.

Investigation of the learning-theoretic aspects of TDA could naturally be approached via the PAC-learning framework introduced by Valiant [19]. This is a well-studied approach to defining the learnability of different classes of concepts. Reducing homological learning problems to more general learning problems may provide immediate results about the sample complexity or time complexity of potential algorithms.

Any invariants or vectorisations constructed could be evaluated for their discriminative power by experiment on datasets. Similarly, the time and memory requirements of any algorithms developed could be recorded when run on the standard datasets provided by Nina Otter [9].

5 Expected Outcomes:

By the end of this project, I expect to have contributed new proofs to the theory of persistent homology and its application to data science. Besides developing our understanding of these methods, I expect to provide constructions which have immediate practical use in data science applications, as well as results from experimental evaluation of these. These contributions should play a part in influencing the uptake of TDA as a reliable toolbox in the wider data science community.

References

[1] Carlsson G. *Topology and data.* Bull. Amer. Math. Soc. 2009;46(2):255-308

[2] Nicolau M., Levine A.J., Carlsson G. *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival.* Proc. Natl. Acad. Sci. 2011;108(17):7265-7270

[3] Buchet M., Hiraoka Y., Obayashi I. *Persistent homology and materials informatics.* In: Tanaka I. (eds) Nanoinformatics. Singapore: Springer; 2018. P.75-95

[4] Harker S., Mischaikow K., Mrozek M., Nanda V. *Discrete morse theoretic algorithms for computing homology of complexes and maps.* Found. Comput. Math. 2014;14(1):151-184

[5] Edelsbrunner H., Letscher D., Zomorodian A. *Topological persistence and simplification.* Discrete Comput. Geom. 2002;28:511–533

[6] Singh G., Mémoli F., Carlsson G. *Topological methods for the analysis of high dimensional data sets and 3D object recognition.* SPBG. 2007;91-100

[7] Michel B. *A statistical approach to topological data analysis.* UPMC Université Paris VI, 2015

[8] Niyogi P., Smale S., Weinberger S. *Finding the homology of submanifolds with high confidence from random samples.* Discrete Comput. Geom. 2008;39(1-3):419–441

[9] Otter N., Porter M.A., Tillmann U., Grindrod P., Harrington H.A. *A roadmap for the computation of persistent homology.* EPJ Data Sci. 2017;6:17, Springer Nature

[10] Carlsson G., Zomorodian A. *The theory of multidimensional persistence.* Discrete Comput. Geom. 2009;42(1):71–93

[11] Chazal F., Glisse M., Labruère C., Michel B. *Convergence rates for persistence diagram estimation in topological data analysis.* J. Mach. Learn. Res. 2015;16(1):3603-3635

[12] Chazal F., Cohen-Steiner D., Mérigot Q. *Geometric inference for probability measures.* Found. Comp. Math. 2001;11(6):733–751

[13] Bubenik P. *Statistical topological data analysis using persistence landscapes.* J. Mach. Learn. Res. 2015;16(1):77-102

[14] Bubenik P., Dłotko P. *A persistence landscapes toolbox for topological statistics.* J. Symb. Comp. 2017;78:91-114

[15] Chevyrev I., Nanda V., Oberhauser H. *Persistence paths and signature features in topological data analysis.* Preprint: arXiv:1806.00381 [stat.ML], 2018

[16] Harrington H.A., Otter N., Schenck H., Tillmann U. *Stratifying multiparameter persistent homology.* Preprint: arXiv:1708.07390 [math.AT], 2017

[17] Vipond O. *Multiparameter persistence landscapes.* Preprint: arXiv:1812.09935 [math.AT], 2018

[18] Bubenik P., Scott J.A. *Categorification of persistent homology.* Discrete Comput. Geom. 2014;51:600-627

[19] Valiant L.G. *A theory of the learnable.* Commun. ACM. 1984;27(11):1134-1142