RESEARCH STATEMENT

HAIBIN HANG

1. Research Interests

We are living in an age in which data are pervasive, big and complex data of all kinds are being collected at a brisk pace to enable advances in scientific, medical, industrial, financial and many other domains. Extracting information from data and turning data into actionable knowledge are fundamental challenges that we must address. Mathematics has much to contribute to these challenges in data science that opens up many new research perspectives. My research interests are in problems that involve data and mathematics, particularly algebraic topology, Riemannian geometry, optimal transport and statistics. I have worked on problems related to the theoretical foundations of geometric and topological data analysis, as well as applications to problems in ecology and materials science. I intend to continue to pursue some of these lines of investigation, but very importantly, I would like to broaden the scope of my research during my postdoctoral experience. Below, I provide some details about completed and ongoing projects in which I have participated, and also discuss some future research plans.

2. Topological Data Analysis

Persistent homology is one of the main tools in topological data analysis (TDA) that allows us to summarize and analyze complex features of data through computationally accessible barcodes or persistence diagrams. Thus far, I have worked on two main TDA projects: (i) examining ways of effectively using TDA to analyze random functional and structural data with persistence diagrams that are rich in information about their shapes and (ii) extending persistence homology to a more general setting that unifies such structures as persistence modules and zigzag modules.

2.1. **Topological study of functional data and structural data.** This is joint work with Washington Mio and Facundo Mémoli, published in the Journal of Applied and Computational Topology [8]. We investigated ways to use persistent homology to analyze functional data on compact topological spaces and structural data represented as compact metric measure spaces. The usual way of applying persistent homology to the sublevel (or superlevel) set filtration of functional data [4] may lead to highly innaccurate barcode expressions caused by the topology of regions where the signals are weak, often unrelated to the shape of the signal. In our paper, we introduced a cone construction, which has a "barcode trimming" effect that enhances the "actual" shape of functional data. We defined a metric on the space of compact functional spaces and proved a stability theorem for persistent homology, stronger than the stability theorem for standard barcodes. We also mapped metric-measure spaces to functional space via their Fréchet functions, proving consistency results and obtaining rates of convergence for barcodes derived from empirical distributions.

To make this approach amenable to computation, in an ongoing project with Woojin Kim and Facundo Mémoli at Ohio State University, we are studying a 2D persistence version of the problem using bi-filtered simplicial complexes. Associated invariants would summarize features which persist both along the scales of functional value and Vietoris-Rips scales for the domains of the functions. Most of the results in [8] have an analog in the 2D setting. The real difficulty is that there are no complete, computable invariants for 2D persistent homology [2]. We currently are working on obtaining meaningful and yet computable invariants of 2D persistent modules.

2.2. **Correspondence Modules and Persistence Sheaves.** This is the core material of my doctoral dissertation that I am writing under the supervision of Washington Mio. I introduce the notions of *correspondence modules* (c-modules) and *persistence sheaves* that not only provide a unifying framework for the study of various types of 1D algebraic persistence structures such as persistence modules and zigzag modules, but is also broader, including many additional structures. A manuscript on this material is at an advanced stage of preparation and a pre-print should be posted on arXiv shortly.

In its simplest form, a persistence module is a sequence a vector spaces (over field k) connected by (forward) linear maps: $V_1 \to V_2 \to V_3 \to \cdots \to V_n$. More general forms of these 1D modules are parameterized over $\mathbb{R}$. A zigzag module is a sequence of vector spaces connected by forward or backward linear maps: $V_1 \leftrightarrow V_2 \leftrightarrow V_3 \leftrightarrow \cdots \leftrightarrow V_n$, where each $\leftrightarrow$ can be either $\to$ or $\leftarrow$ [1]. However, it is less clear how to define zigzag modules parameterized over $\mathbb{R}$. In either case, if the module is "tame", it can be represented, uniquely up to isomorphism, as the direct sum of simple components called *interval modules* [5]. This gives us some indication that the directions of the arrows are not so relevant for the decomposability property. In my dissertation, I show that decomposition can be achieved for much more general structures in which linear mappings (forward or backward) are replaced with partial linear relations.

A persistence module over $\mathbb{R}$ assigns to each $s \in \mathbb{R}$ a vector space V_s, and to each pair $s \leq t$ a linear map $v_s^t : V_s \to V_t$ such that $v_s^s = I_{V_s}$ and $v_s^t \circ v_r^s = v_r^t$ for any $r \leq s \leq t$. In a *correspondence module*, v_s^t is only assumed to be a linear relation from V_s to V_t; that is, a vector subspace of $V_s \times V_t$. In this setting, the relations v_s^t are assumed to satisfy the following compatibility conditions: v_s^s is the diagonal subspace $\Delta_{V_s} \subseteq V_s \times V_s$ and

$$v_r^t = v_s^t \circ v_r^s := \{(u, w) \in V_r \times V_t \,|\, \exists v \in V_s \text{ s.t. } (u, v) \in v_r^s \text{ and } (v, w) \in v_s^t\},$$

for any $r \leq s \leq t$. Among many other things, this allows us to make sense of zigzag structures continuously parameterized over $\mathbb{R}$.

Our main result is a decomposition theorem for "tame" correspondence modules. Unlike persistence modules, to each interval $J \subseteq \mathbb{R}$, there are up to four non-isomorphic interval c-modules associated with J.

Theorem 1. *If a c-module $\mathbb{V}$ is tame, then $\mathbb{V}$ is isomorphic to a direct sum of interval c-modules.*

By passing from a c-module to its sections over intervals, we obtain a sheaf theoretical representation of a c-module that we call persistence sheaf (p-sheaf) of sections. Persistence sheaves have the technical advantage of placing us back in the category of vector spaces and linear mappings. The decomposition theorem stated above is a special case of a more general theorem we proved that applies to sufficiently "tame" p-sheaves. We also have proven an algebraic stability theorem in this more general setting for barcodes derived from interval decomposition of p-sheaves.

As already mentioned, c-modules and p-sheaves generalize persistence and zigzag modules to a broader framework that allows us to investigate a number of other problems. For example, they provide a more natural framework to formulate level-set persistent homology of a function (as opposed to using discrete zigzags), they enable us to define restrictions of a 2D persistence module to lines of negative slope, etc. In the future, I will continue to explore new applications of c-modules and p-sheaves, including ties with 2D persistence modules.

3. Covariance Fields on Riemanian Manifolds

The mean and covariance matrix are the most basic statistics of Euclidean data and crucial in principal component analysis (PCA). Nonetheless, the standard formulation of covariance uses the linear structure of the ambient space in an essential way, making it unclear how to extend the concept to data on Riemannian manifolds in a principled manner. For distributions with finite second moment, the Euclidean mean is the unique minimizer of the Fréchet function, but the Fréchet

function on more general metric spaces may have multiple minima. As such, it seems more natural to consider data covariation with respect to any point, not just the mean, and thus work with covariance fields. In recent work, we proposed a formulation for general covariance tensor fields on Riemannian manifolds. For a random variable distributed according to a Borel probability measure α on a Riemannian manifold (M, g) and a smooth symmetric kernel function $u : M \times M \to \mathbb{R}$, define a covariance k-tensor as

$$\Sigma^k_{\alpha,u}(x) := \int_M \otimes_k \nabla_x u(x, y)\alpha(dy).$$

If $M = \mathbb{R}^n$, $k = 2$, $u(x, y) = \frac{\|x-y\|^2}{2}$, and x is the mean of the distribution, this definition that only requires a locally linear structure coincides with the usual covariance matrix. Preliminary findings about the behavior of these covariance tensor fields have been published in an Oberwolfach report [7]. My future research plans along these lines include the formulation of discrete forms of covariance tensors for distributions on the vertex set of a weighted network, as well as investigation of relationships between covariance tensors and the Wasserstein distance between probability measures.

4. Detecting Carbon Nanotube Orientation with Topological Data Analysis of SEM Images

High-performance carbon nanotube (CNT) materials are in high demand as a result of their extraordinary mechanical, electrical and thermal properties. CNT alignment is an important property in the fabrication of ultra-strong CNT composites. Hence, it is fundamentally important to evaluate and quantify the degree of alignment using various characterization methods. In collaboration with researchers from FSU High-Performance Materials Institute, we developed a novel method to detect CNT orientation combining topological data analysis with scanning electron microscopy (SEM). We use barcodes derived from persistent homology of SEM images to quantify CNT alignment. The results we have obtained are highly consistent with those from polarized Raman spectroscopy and X-ray scattering. Our approach offers a simpler and more effective way of understanding the role that alignment plays in CNT properties. A manuscript reporting the results obtained is in preparation.

5. Characterizing Phenotypic Plasticity of Ginkgo Biloba Leaves with Topological Data Analysis

Phenotypic plasticity of living organisms can be seen in many guises and represents some of the raw material for the evolutionary process. For example, different morphologies have different functional properties and so can be favored, or not, under certain environmental conditions. In this project, joint with an ecologist from Open University, UK, we carried out a pilot study with *Ginkgo biloba* leaves that are known to produce leaves that are enormously diverse in their shape.

We employed persistent homology to develop models of morphological variation for leaves. This has allowed us to quantify morphological variation in leaves of the same tree and compare the morphology of leaves of extant plants and fossils. Such evolutionary mappings enable us to investigate evolution over geological time scales and potentially relate contrasts in phenotypes to ecological events. A manuscript reporting our findings is in preparation.

Once quantitative models of phenotypic plasticity are developed, interesting extensions become possible. For instance, one may address questions such as: How do the various adult leaf shapes relate to leaf development? Is it the case that all young leaves within a bud essentially have the same shape, and that the diverse morphologies are produced later following exposure to environmental conditions? Are the diverse morphologies related in any way to other aspects of morphology, such as the architecture of underlying vasculature? I intend to continue to explore some of these problems through this collaboration.

6. Other Research Interests

Deep learning has become a fascinating area which attracts researchers with various backgrounds. In the future, I would like to explore analysis of deep neural networks with geometric or topological methodology, which may give us some additional insights on the black box behind the training process. Some recent articles [3, 6] have addressed some questions related to the relationships between geometric or topological data analysis and deep neural networks. I think this will be a really promising research area in the next few years.

Optimal transport is another area I want to explore more in the future. My previous research projects have benefited a lot from ideas from optimal transport. Additionally, optimal transport offers mathematical foundations for the analysis of several classical algorithms in computer vision and shape analysis. There are also some some parallels between optimal transport and deep learning. This is an area that is presenting rich and promising grounds for research and I would like to devote some of my time to it.

I also intend to keep working on interesting scientific and practical applications of TGDA, as this can increase the relevance of the discipline and also drive further theoretical developments in TGDA.

References

[1] Gunnar Carlsson and Vin de Silva. Zigzag persistence. *Foundations of Computational Mathematics*, 10(4):367–405, Aug 2010.

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

[3] Gunnar E. Carlsson and Rickard Brüel Gabrielsson. Topological approaches to deep learning. *CoRR*, abs/1811.01122, 2018.

[4] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010.

[5] Peter Gabriel. Unzerlegbare darstellungen i. *manuscripta mathematica*, 6(1):71–103, Mar 1972.

[6] Rickard Brüel Gabrielsson, Bradley J. Nelson, Anjan Dwaraknath, Primoz Skraba, Leonidas J. Guibas, and Gunnar E. Carlsson. A topology layer for machine learning. *ArXiv*, abs/1905.12200, 2019.

[7] Haibin Hang, Facundo Mémoli, and Washington Mio. Covariance tensors on riemannian manifolds. *Mathematisches Forschungsinstitut Oberwolfach Report*, Mar 2018.

[8] Haibin Hang, Facundo Mémoli, and Washington Mio. A topological study of functional data and fréchet functions of metric measure spaces. *Journal of Applied and Computational Topology*, Aug 2019.