Research Statement

Joshua Mirth

My research interest is in topological data analysis. Topology is the field of mathematics which studies the qualitative aspects of geometry. Topological Data Analysis, or TDA, is a field based around the insight that many data sets possess topological structure. The field generally asks two questions:

1. What topological features of a set of data (and other spaces arising in applications) can be computed and how?

2. What does the topological and geometrical structure of a data set mean in application?

My work addresses both questions, which I describe in the three project areas below. Project 1 is concerned with theoretical guarantees of topological computations. I introduce ideas from optimal transport and Morse theory to determine the exact topological structure of certain *Vietoris–Rips simplicial metric thickenings*—spaces which are used as reconstructions of manifolds in applied topology. Project 2 is about how to compute topological features, in this case the *fractal dimension* of a data set. My coathors and I showed that persistent homology, a tool for computing large-scale topological features, can also be used to compute fractal dimensions, an inherently small-scale feature. Project 3 is about using topology for the analysis of a particular data set. Here our subject was a sample of optical flow data in which we found and interpreted a torus-shaped structure.

I am interested in pursuing each of these areas further, as well as exploring more connections between topology, machine learning, and optimal transport. I believe strongly in collaborative research, especially across mathematical disciplines and between STEM fields more generally. TDA is an emerging field with many open problems, but it is also based in the deeply theoretical field of algebraic topology. I think this combination makes it a uniquely exciting field in which to mentor undergraduate research projects.

Project 1: Morse Theory and Optimal Transport

The Wasserstein distance is a metric on probability measures. It is the namesake of Wasserstein GANs in machine learning [6], has been used to study the space of persistence diagrams in applied topology [7], and is extensively studied in partial differential equations and geometry [10] [5]. This project studied the topology and geometry of the Wasserstein space of probability measures $\mathcal{P}(M)$ on a manifold M in order to better understand manifold reconstruction techniques in topological data analysis.

Optimal transport asks this question: given a probability measure μ on a space M, is there a best way to transform μ into some desired measure ν? (See Figure 1.) Best is understood to mean

minimizing the product of mass and distance. Put rigorously, given probability measures μ and ν, the optimal transport problem is to find a joint probability distribution γ with marginals μ and ν and which minimizes the quantity

$$T_\gamma(\mu,\nu) = \int_{M\times M} d^2(x,y)\mathrm{d}\gamma(x,y). \tag{1}$$

Remarkably, the existence of a minimizing γ is guaranteed. The Wasserstein distance is then defined to be

$$W_2(\mu,\nu) = \inf_\gamma \left(\int_{M\times M} d^2(x,y)\mathrm{d}\gamma(x,y)\right)^{1/2}. \tag{2}$$

This is a true metric (i.e. it satisfies the triangle inequality) on the space $\mathcal{P}(M)$ of probability measures on M with finite second moment.

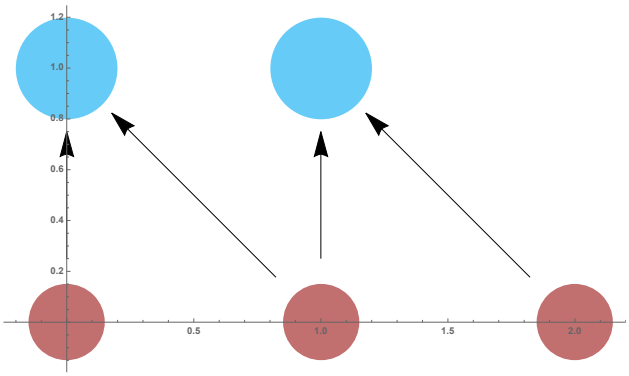

Figure 1: The optimal transport plan between $\mu = \frac{1}{3}\delta_{(0,0)} + \frac{1}{3}\delta_{(1,0)} + \frac{1}{3}\delta_{(2,0)}$ (in red) and $\nu = \frac{1}{2}\delta_{(1,1)} + \frac{1}{2}\delta_{0,1}$ (in blue) involves splitting the mass at $(1,0)$ between $(0,1)$ and $(0,2)$. (Here δ_p is the Dirac delta distribution centered at p.)

In topological data analysis, a data set is understood as a metric space, X. (For example, a finite collection of points in $\mathbb{R}^n$ with the Euclidean metric.) The topology can be estimated by constructing a simplicial complex, such as the Vietoris–Rips complex $\mathrm{VR}(X;r)$ which has vertex set X and where simplices are defined by

$$\sigma \in \mathrm{VR}(X;r) \iff \mathrm{diam}(\sigma) \leq r \tag{3}$$

for some $r \in [0,+\infty]$. Formally, a point in the geometric realization of the simplicial complex is a weighted sum of the collection of vertices. The weights in this sum must be positive and sum to unity, so this point is naturally interpreted as a probability measure. Since the vertex set of this simplicial complex is a metric space, the Wasserstein distance puts a metric on the simplicial complex. This is called a simplicial metric thickening [1].

A fundamental question is if X is sampled from a manifold M, does this reconstruction have the same topological structure as M? Adamaszek, Adams, and Frick proved in [1] that for sufficiently small r, and M a Riemannian manifold, the Vietoris–Rips simplicial metric thickening has the same homotopy type as M, which mirrors a classical result of Hausmann for Vietoris–Rips

simplicial complexes [8]. I gave a proof of the same fact using different techniques for X a subset of $\mathbb{R}^n$ with positive reach [4].

These results only hold when the amount of thickening is small. If the thickening parameter is allowed to be larger, the homotopy type is known to change, but it is not generally known how. Permitting the parameter to be unbounded recovers the space of all finitely-supported probability distributions, $\mathcal{F}(M)$, and the topological closure of that is the space $\mathcal{P}(M)$ of all probability distributions. I am studying the homotopy type of these spaces as both a first attempt toward understanding the topology of simplicial metric thickenings, and as interesting topological spaces in their own right.

To do so, I take inspiration from classical Morse theory. Morse theory takes a smooth function, f, on a manifold, M, and decomposes the manifold into cells using the sublevel sets of f. If the set $f^{-1}[a,b]$ contains no critical points of f and is compact, then the sublevel sets $f^{-1}(-\infty,a]$ and $f^{-1}(-\infty,b]$ are diffeomorphic, and if $f^{-1}[a,b]$ contains exactly one critical point, then $f^{-1}(-\infty,b]$ is obtained by gluing a k-cell to $f^{-1}(-\infty,a]$ where k is determined by the index of the critical point. This powerful theory formed the basis of Smale's proof of the generalized Poincaré conjecture.

The machinery necessary to prove the main theorems of Morse theory is a gradient of the function f and its associated flow, or one-parameter group of diffeomorphisms on M. Wasserstein space possesses much less structure than M (it is infinite dimensional), but it has a geodesic structure and for (geodesically) convex functions, there is a gradient and an associated flow. Using these, I show that the Wasserstein space on any convex subset of a Hilbert space is contractible by flowing along the gradient of a certain functional, and for the simplicial metric thickening of certain manifolds at small scales, one can flow along the gradient of the same functional to obtain a homotopy between the manifold and the thickening.

Project 2: Persistent Homology Fractal Dimension

The most popular tool in topological data analysis is persistent homology. This captures the topology of a data set in a barcode (see Figure 2). One usually assumes that the longest bars represent "real" features of the space and that short ones are a consequence of noise or artifacts of the reconstruction via simplicial complexes. In this project my coauthors from the Colorado State University Pattern Analysis Lab and I, expanding on early work such as [9], showed that those short barcodes actually do contain information, namely, the fractal dimension of the data [2].

Fractal dimensions are supposed to extend the concept of dimension beyond sets where it is obvious (such as manifolds). Often these sets have non-integer dimensions! There are many classical definitions of fractal dimensions: the Hausdorff, box-counting, and information dimensions being among the more widely-used. These all give the expected answer on submanifolds of Euclidean space, and they agree on nice fractals like the Cantor set, but they are not always equal.

Our definition of persistent homology fractal dimension looks at the scaling of the sum of the lengths of the barcodes in i-th dimensional persistent homology as more points are drawn from the data. One great benefit of our dimension is that it is computable, and computational experiments suggest that it gives the expected dimension for submanifolds and classical fractals, at least when $i = 0$ or 1. For higher dimensions the persistent homology becomes computationally inten-

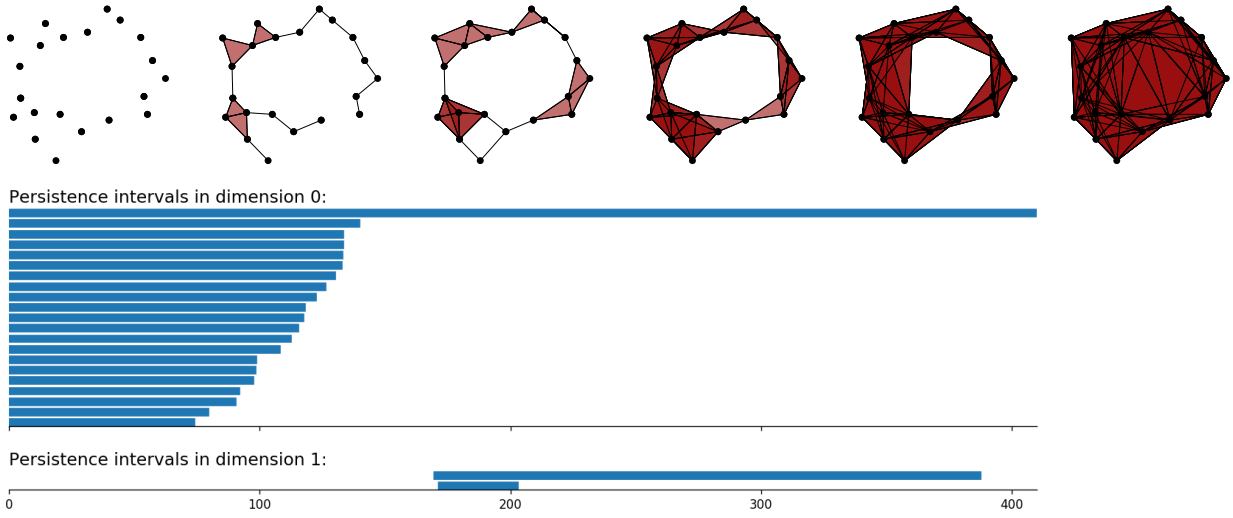

Figure 2: The Vietoris–Rips filtration (see Project 1) turns a data set into a sequence of simplicial complexes. The barcodes in dimension i record the homology in dimension i as the sequence progresses. Here the 0-dimensional persistent homology intervals show 21 connected components merging into a single connected component, and the 1-dimensional intervals show two 1-dimensional holes, one short-lived and the other long-lived.

sive. There are existing dimensions based on spanning trees, which correspond to the $i = 0$ case of our dimension, so the persistent-homology dimension can be thought of as a generalization of these.

Project 3: Optical Flow

The optical flow of a video is a vector field for each frame attaching to each pixel the velocity vector it appears to follow. If one knows the actual motion of each object in the scene, the optical flow can be computed; however, the inverse problem of determining optical flow from a video is difficult. In this project we determined the topology of an optical flow data set. Knowing the topology potentially helps in reconstructing the correct flow from a video.

We used the optical flow data from the animated movie *Sintel*. Being animated, the actual velocity vectors are known. In the space of flow images, there is typically a strong circle coming from the apparent difference in motion between foreground and background objects (see Figure 3. Using persistent homology, we showed that above each angle in this circle there lies another circle in the data. Moreover, these circles glue together to form a fiber bundle. We were able to determine through zig-zag persistence that this fiber bundle was a torus [3].

References

[1] Michal Adamaszek, Henry Adams, and Florian Frick. Metric reconstruction via optimal transport. *SIAM Journal on Applied Algebra and Geometry*, 2(4):597–619, 2018.

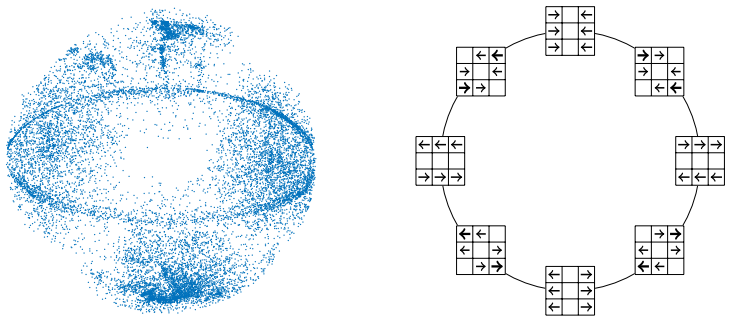

Figure 3: (Left) A projection of the optical flow data. The horizontal flow circle, is clearly visible. (Right) The optical flow types that produce the circle. Arrows indicate the direction of motion of the pixels in a 3×3 patch from the video.

[2] Henry Adams, Manuchehr Aminian, Elin Farnell, Michael Kirby, Chris Peterson, Joshua Mirth, Rachel Neville, Patrick Shipman, and Clayton Shonkwiler. A fractal dimension for measures via persistent homology. *arXiv preprint arXiv:1808.01079*, 2018.

[3] Henry Adams, Johnathan Bush, Brittany Carr, Lara Kassab, and Joshua Mirth. On the nonlinear statistics of optical flow. In *International Workshop on Computational Topology in Image Context*, pages 151–165. Springer, 2019.

[4] Henry Adams and Joshua Mirth. Metric thickenings of euclidean submanifolds. *Topology and its Applications*, 254:69–84, 2019.

[5] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[6] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

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

[8] Jean-Claude Hausmann. On the vietoris-rips complexes and a cohomology theory for metric spaces. *Annals of Mathematics Studies*, 138:175–188, 1995.

[9] Robert MacPherson and Benjamin Schweinhart. Measuring shape with topology. *Journal of Mathematical Physics*, 53(7):073516, 2012.

[10] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.